J Vis (2023) 26:199–210
https://doi.org/10.1007/s12650-022-00871-6

REGULAR PAPER

Takayuki Itoh · Asuka Nakabayashi · Mariko Hagita

Multidimensional data visualization applying a variety-oriented scatterplot selection technique

Received: 18 January 2022 / Revised: 1 June 2022 / Accepted: 13 August 2022 / Published online: 29 August 2022

Abstract Multidimensional data visualization is one of the most active research topics in information visualization since various information in our daily life forms multidimensional datasets. Scatterplot selection is an effective approach to represent essential portions of multidimensional data in a limited display space. Various metrics for evaluating scatterplots, such as scagnostics, have been applied to scatterplot selection. One of the open problems of this research topic is that various scatterplots cannot be selected if we simply apply one of the metrics. In other words, we may want to apply multiple metrics simultaneously in a balanced manner when we want to select a variety of scatterplots. This paper presents a new scatterplot selection technique that solves this problem. First, the technique calculates the scores of scatterplots with multiple metrics and then constructs a graph by connecting pairs of scatterplots that have similar scores. Next, it uses a graph coloring algorithm to assign different colors to scatterplots that have similar scores. We can extract a set of various scatterplots by selecting them that the specific same color is assigned. This paper introduces two case studies: the former study is with a retail transaction dataset while the latter study is with a design optimization dataset.

Keywords Multidimensional data visualization · Scatterplot selection · Graph coloring algorithm · Retail transaction data · Design optimization data

1 Introduction

Various information in our daily life forms Multidimensional datasets. There have been various multidimensional data in science, engineering, business, and social research and industry fields. Multidimensional data visualization is therefore one of the most important issues in information visualization. In addition to geometric methods with explicit coordinate axes such as ScatterPlot Matrix (SPM) and Parallel Coordinate Plots (PCP), icon-based and pixel-based methods are known as multidimensional data visualization methods.

Dimension selection Itoh et al. (2017), Yuan et al. (2013), Zhang et al. (2012) is a hotspot for visualizing high-dimensional data. It is unreasonable to represent every dimension in a limited display space; therefore, it is essential to remove noisy or meaningless dimensions and focus on visualizing informative dimensions.

In multidimensional data methods using scatterplots, dimension selection is equivalent to selecting scatterplots that are worth viewing. Automatic selection of scatterplots is therefore one of the interesting issues in multidimensional data visualization. Scagnostics Wilkinson et al. (2005) is a set of typical metrics applied to scatterplot selection problems Watanabe et al. (2017), Nakabayashi and Itoh (2019), Zheng et al.

T. Itoh (✉) · A. Nakabayashi · M. Hagita
Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
E-mail: itot@is.ocha.ac.jp

(2015). We can selectively display a set of similarly featured scatterplots by applying one of the metrics. Meanwhile, it is not always suitable to apply a single metric for a scatterplot selection to capture all the characteristics of multidimensional data. For instance, interesting correlations are observed from some pairs of dimensions, while interesting clusters are observed from some other pairs of dimensions. It is bothering if we need to switch the metrics to display various types of scatterplots. On the other hand, automatic selection of scatterplots that have a variety of characteristics by applying multiple metrics simultaneously is a reasonable approach to understand various characteristics of the dataset just observing a single display space that shows the selected scatterplots.

This paper presents a fast scatterplot selection technique that automatically selects a variety of scatterplots applying multiple metrics. First, this technique generates scatterplots with every pair of dimensions. Then, it calculates multiple scores based on multiple metrics for each scatterplot and forms a vector from the scores. Next, it constructs a graph by connecting pairs of scatterplots if it determines that at least one of them can be eliminated. Further, it assigns colors to the vertices corresponding to the scatterplots while complying with a rule that different colors are assigned to a pair of vertices connected by an edge. In other words, the same color is assigned to a set of significantly different scatterplots. The technique selectively displays a constant number of scatterplots that have the same color.

Figure 1 shows an example of scatterplot selection by this technique. We currently implement four metrics presented in Sect. 3.3: correlation, thinness, clumping, and separateness, to select a variety of scatterplots that show various characteristics of the input dataset.

This paper is an extended version of a conference paper presented by the authors Itoh et al. (2021). The paper additionally introduces a case study with a design optimization dataset.

2 Related work

This section introduces recent multidimensional data visualization techniques applying dimension selection and scatterplot evaluation.

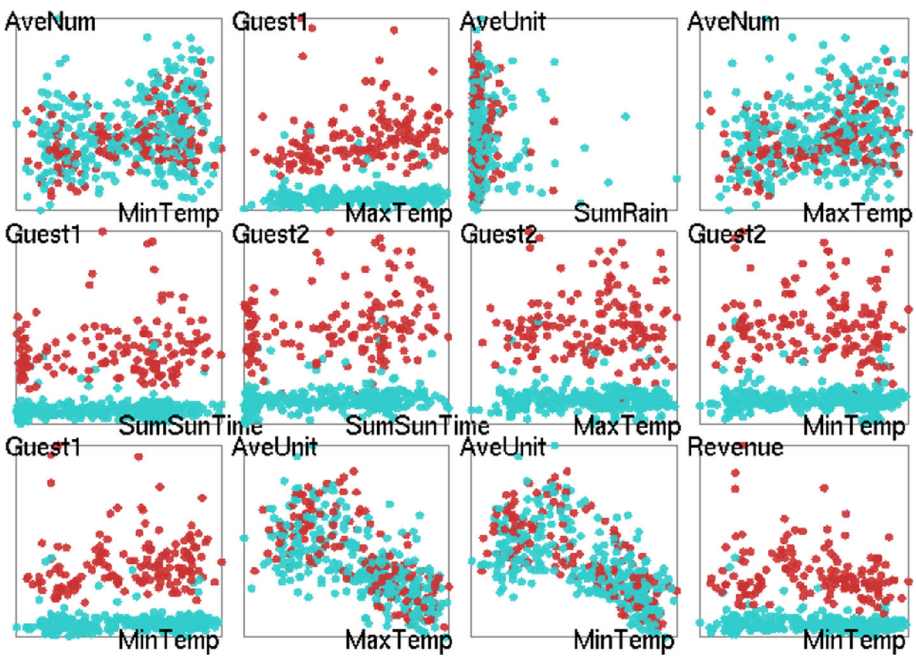

Fig. 1 An example of our multidimensional data visualization applying a variety-oriented scatterplot selection technique. Several scatterplots show strong correlations between dimension pairs; some scatterplots clearly show clusters or outliers; several scatterplots show how two labels drawn in red and blue are separated. The presented technique selects a variety of scatterplots to show the various characteristics of the input multidimensional dataset in a single display space

2.1 Dimension selection for multidimensional data visualization

Dimension selection has been one of active topics for multidimensional data visualization to effectively represent essential subsets of dimensions. Claessen et al. (2011) visualized high-dimensional datasets by representing a set of low-dimensional subspaces as a combination of PCPs (parallel coordinate plots) and scatterplots. Suematsu et al. (2013) and Zheng et al. (2015) also converted high-dimensional datasets into low-dimensional subsets and visualized these subsets using multiple PCPs or scatterplots, respectively. These techniques did not provide rich interaction mechanisms to freely select the numbers of dimensions.

Several studies have demonstrated interaction mechanisms to freely visualize interesting low-dimensional subspaces. Lee et al. (2013) and Liu et al. (2014) applied dimension reduction schemes to interactively select subsets of high-dimensional data. Nohno et al. (2014) presented a technique to interactively contract highly correlated dimensions to adjust the number of axes displayed in PCPs. Itoh et al. (2017), Watanabe et al. (2017), and Nakabayashi et al. (2019) presented a series of techniques that easily control the number of dimensions displayed in the PCPs or the number of dimension pairs represented by scatterplots.

It is also important to understand relationships among dimensions while extracting low-dimensional subspaces. Dimension spaces have been visualized by applying scatterplots or graphs by several recent studies Itoh et al. (2017), Yuan et al. (2013), Zhang et al. (2015). This is an effective approach to interactively select reasonable sets of dimensions.

Despite many studies on multidimensional data visualization employing dimension selection techniques, there have been few studies to automatically select various limited number of informative scatterplots. We address this problem and present a new technique in this paper.

2.2 Evaluation of scatterplots

Numeric evaluation of the informativeness of scatterplots has been an active research topic. Scagnostics is a remarkable method to quantitatively evaluate the informativeness of scatterplots. Wilkinson et al. (2005) proposed nine features of scagnostics based on the appearance of scatterplots. Wang et al. (2020) proposed an improved scagnostics by considering the human perception to several metrics, including ”Outlying” and ”Clumpy.” There have been several more studies that focus on specific metrics of scatterplots, including correlation Harrison et al. (2014), Shao et al. (2017) and class separation Aupetit and Sedlmair (2016), Sedlmair et al. (2012, Sips et al. (2009).

There have been several visualization studies on the overview and exploration of a large number of scatterplots. Dang et al. (2014) presented an exploration mechanism for finding similarly featured scatterplots and filtering scagnostics. Matute et al. (2017) presented another approach to represent the distribution of characteristics of scatterplots. The goal of our study is somewhat similar to the above studies since we also focus on representing various scatterplots; however, our focus is different from these studies in that we aim to selectively display the user-defined number of various scatterplots.

3 Scatterplot selection applying a graph coloring algorithm

This section presents a processing flow of the presented scatterplot selection technique. The technique calculates the scores of scatterplots based on multiple metrics and stores it as vector values. Figure 2 illustrates the concept of scatterplot selection. Scatterplots are depicted as vectors in the metric space. The requirements for scatterplot selection in this study are summarized as follows.

R1: Avoid selecting sets of close vectors to avoid selecting scatterplots that have similar features.
R2: Avoid selecting short vectors to avoid selecting less informative scatterplots.

The technique applies a graph coloring algorithm to satisfy the above requirements and displays various informative scatterplots.

The technique sets R1 and R2 in order to prioritize selecting a variety of scatterplots rather than selecting important scatterplots without missing any of them while selecting a user-specified number of scatterplots. Therefore, even if there are multiple important scatterplots, one of them may not be selected if both of them have similar characteristics. We compromise such unavoidable results that overlook some important scatterplots, especially in case a small number of scatterplots is specified.

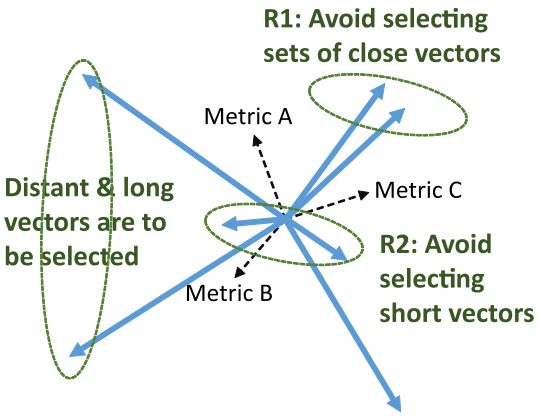

Fig. 2 Concept of scatterplot selection in the metric space. Blue arrows illustrate the vectors of metrics. Our technique selects various scatterplots while satisfying **R1** and **R2**

3.1 Data structure

This paper formalizes the problem as follows. An input multi-dimensional dataset A has n individuals as $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n\}$. The i-th individual $\boldsymbol{a}_i$ has the m-dimensional values as $\boldsymbol{a}_i = (a_{i1}, a_{i2}, ..., a_{im})$. A set of scatterplots formed from every pair of dimensions is described as $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_N\}$, where N is the total number of scatterplots. Each scatterplot has a set of scores calculated based on predefined metrics. This section describes the score of the j-th scatterplot as $\boldsymbol{s}_j = (s_{j1}, s_{j2}, ..., s_{jM})$, where M is the number of metrics. Here, the IDs of the scatterplots in S are simply determined by the IDs of the two variables that the scatterplots have. We would like to reconsider on how to determine the IDs of the scatterplots since it affects the search order in the breadth-first search described below.

3.2 Graph coloring algorithm

This technique applies a graph coloring algorithm to select various scatterplots with different characteristics. Suppose a graph $G = (S, E)$, where S is a set of vertices corresponding to the scatterplots, and E is a set of edges connecting pairs of scatterplots. Here, we select a pair of distant and long vectors, as shown in Fig. 2. In other words, we would like to select a pair of the i-th and j-th vectors if the area of the triangle d_{ij} constructed by these vectors is large. Here, the technique constructs the graph by generating edges between the i-th and j-th scatterplots if the area of the triangle d_{ij} is smaller than the predefined threshold d_{thres}. Remark that we suppose all scores are zero or positive and therefore no pairs of vectors forms opposite directions.

Then, the technique assigns colors to the scatterplots while complying to a rule that different colors are assigned to a pair of vertices connected by an edge. In other words, the same color is assigned to a set of significantly different scatterplots. Figure 3 illustrates the process. First, the process selects the scatterplot that has the largest $|\boldsymbol{s}_k|$ and assigns the color identification $c_k = 0$. Then, adjacent vertices connected by edges are traversed in the breadth-first order. The search order for adjacent vertices is simply based on the IDs of the scatterplots in the current implementation, but other criteria (e.g., descending order of $|\boldsymbol{s}_k|$) are worth experimenting with.

While visiting the k-th vertex, the process specifies the minimum color identification that is assigned to none of the adjacent vertices connected with the k-th vertex and assigns it to the k-th vertex. For instance, if color identifications 0, 1, and 3 have been assigned to the vertices adjacent to c_k, the process specifies c_k as 2. The breadth-first search is repeated until color identifications are assigned to all vertices.

Finally, we select a predefined number of scatterplots to be displayed. The technique extracts a set of scatterplots in which the same color is assigned. We calculate the sums of the length of the vectors $|\boldsymbol{s}_k|$ for each color and select the color that brings the largest sum. The extracted set of scatterplots excludes similarly looking or less informative pairs because such pairs of scatterplots are connected and therefore have different colors. In other words, it satisfies **R1** because the extracted set comprises various differently looking scatterplots. If the number of extracted scatterplots is larger than the user-defined number, the technique selects the scatterplots in descending order of $max_k(s_{k,j})$, the maximum value of the scores s_{k1} to

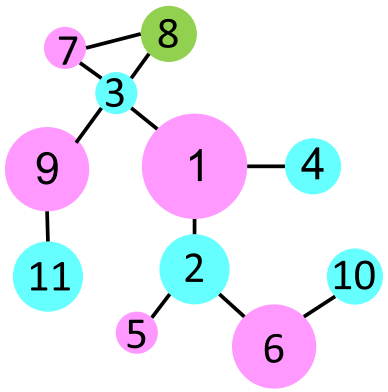

Fig. 3 Graph coloring. The process assigns different colors to the vertex pairs connected by edges. The numbers in this figure denote the order of the breadth-first search

s_{kM} of the k-th scatterplot, to satisfy **R2**. Our implementation provides a user interface to interactively specify the number of scatterplots to be displayed.

The processing flow is as follows.

1. Initialize the vertices S and calculate the scores of the k-th scatterplot as $|\boldsymbol{s}_k|$.
2. Construct the graph and generate an edge between the i-th and the j-th scatterplots if d_{ij} is smaller than the pre-defined threshold.
3. Select the scatterplot that has the largest $|\boldsymbol{s}_k|$ as the starting vertex.
4. Traverse the connected vertices by the breadth-first search. Assign color identifications to the traversed vertices. Repeat this traverse until the color identifications are assigned to all the vertices.
5. Collect the vertices that have the same color identification. Select the user-defined number of vertices in the descending order of the maximum value of the scores of each scatterplot.

The problem solved using the above algorithm is similar to the maximum independent set problem. The presented algorithm is better for our study because it prioritizes to select "long" vectors and "distant" vectors.

3.3 Selection of metrics

Based on the discussion with the owner of the retail transaction dataset introduced in Sect. 4, we focused on finding the following scatterplots.

S1: Scatterplots with the variables that can contribute to the regression for predicting transaction values from climate values.
S2: Scatterplots representing isolated clusters.
S3: Scatterplots that separate different attributes (e.g., weekdays and weekend) of the plots.

We implemented the following four metrics to assist finding the above scatterplots.

3.3.1 Correlation

Correlation is one of the most common metrics used to determine the relationship between a pair of dimensions. It is an effective metric used to find tightly correlated pairs of variables and find **S1**. Our current implementation just calculates the score of the k-th scatterplot as follows:

$$s_{k1} = |S_{pear}(i,j)| \tag{1}$$

where $S_{pear}(i,j)$ is the Spearman's rank correlation between the i-th and j-th dimensions. A dimension pair gets a higher score if they have a strong positive/negative correlation. Instead of applying the Spearman's rank correlation, recent approaches Harrison et al. (2014), Shao et al. (2017) can also be useful.

3.3.2 Thinness

It is easier to adopt a mathematical model to a set of individuals if they form thin regions in a scatterplot. Such scatterplots correspond to **S1**. We measure the thinness of the region where individuals are placed in the scatterplot as Wilkinson et al. (2005) did. Our implementation generates a Delaunay triangular mesh T connecting the individuals in a scatterplot and then removes all triangles that have at least one edge that is longer than a predefined threshold. Then, we calculate the score as follows:

$$s_{k2} = 1 - \sqrt{4\pi A_{rea}(T)}/P_{erimeter}(T) \tag{2}$$

where $A_{rea}(T)$ is the total area of T, and $P_{erimeter}(T)$ is the total length of the boundary of T.

3.3.3 Clumpy

It is remarkable if the individuals in a scatterplot are well-separated into several isolated clusters. Such scatterplots correspond to **S2**. Our current implementation simply applies the metric "Clumpy" presented by Wilkinson et al. (2005) defined as follows:

$$s_{k3} = 1 - length(e_{maxr})/length(e_{mind}) \tag{3}$$

Here, our implementation generates a Delaunay triangular mesh, as described in the previous section, and deletes the edges longer than e_{mind}. e_{maxr} is the longest remaining edge. Newer approaches on clumping Wang et al. (2020) can also be applied.

3.3.4 Separateness

Suppose that one of the labels is assigned to each individual. It is remarkable if individuals that have a particular same label are well-separated in a scatterplot. Such scatterplots correspond to **S3**. We measure the separateness of a particular label by calculating the entropy of the labels. Particularly, we compute the entropy of the labels in the scatterplot generated with the i-th and j-th dimensions as follows:

$$H(i,j) = -\frac{1}{n}\sum_{k=1}^{n}\sum_{c=1}^{C} p(y_k = c|(a_{ki}, a_{kj}) \log p(y_k = c|(a_{ki}, a_{kj}))) \tag{4}$$

where y_k is the label of the k-th individual, (a_{ki}, a_{kj}) is the position in the scatterplot of the k-th individual, and C is the number of labels. Our implementation divides the scatterplot into L subareas and calculates the entropy at the l-th subarea $H(i,j)_l$ using the above equation, and finally calculates the score of the k-th scatterplot as follows:

$$s_{k4} = \left(H_{max} - \sum H(i,j)_l\right)/H_{max} \tag{5}$$

where H_{max} is the maximum value of $\sum H(i,j)_l$.

Instead applying the above-mentioned technique, other approaches Aupetit and Sedlmair (2016), Sedlmair et al. (2012) can be also applied to determine the class separateness.

4 Case study 1: retail transaction data

This paper introduces an example of visualization by the presented technique applying a retail transaction and climate dataset. Table 1 shows the explanatory variables (climate values) assigned to the horizontal axis and the objective functions (retail transaction values) assigned to the vertical axis in this dataset. We clarified how the retail transaction values can be estimated from the climate values by visualizing them. The dataset contained the records of 457 days from May 1, 2016, to July 31, 2017, corresponding to 457 data points in the scatterplots. We generated 35 scatterplots consisting of five horizontal axes and seven vertical axes. The data points are drawn in red or blue; red denotes holidays, while blue denotes weekdays.

Table 1 The explanatory variables and the objective functions

Explanatory variables (climate values)	
MinTemp	Minimum temperature
MaxTemp	Maximum temperature
SumRain	Precipitation
SumSunTime	Sunshine duration
MaxWind	Maximum wind speed
Objective functions (retail transaction values)	
Revenue	Revenue
Guest1	Number of customer
Guest2	Number of visitor
Ratio	Conversion rate
PerGuest	Average revenue per customer
AveUnit	Average price of purchased items
AveNum	Average number of purchased items

4.1 Visualization result

Figure 1 shows an example of a scatterplot selection using our technique. Here, several scatterplots show correlations between dimension pairs, some others show clusters or outliers, while several others show how two labels drawn in red and blue are separated. This figure demonstrates that our technique successfully selects various scatterplots to show various characteristics of the dataset.

Figures 4, 5, and 6 show top four scatterplots that achieved the highest scores on correlation, separateness, and clumpy. The horizontal axes of scatterplots are MinTemp or MaxTemp, while the vertical axes are PerGuest or AveUnit (Fig. 4). This implies that the average revenue or price correlates well with the temperature. Meanwhile, the vertex axes of scatterplots in Fig. 5 are Revenue, Guest1, or Guest2. It implies that revenue and the number of guests significantly differ between holidays and weekdays.

The scatterplot selection result shown in Fig. 1 is well-balanced because it represents various characteristics of the input dataset by selecting various scores of scatterplots. Meanwhile, Fig. 7 shows examples of scatterplots that have no higher scores with all metrics. These scatterplots do not look characteristic or informative. The presented technique does not aggressively select such scatterplots.

By the way, Fig. 6 shows the four scatterplots with the highest clumpy scores, but unfortunately, there are no scatterplots with clearly separated multiple clusters, instead, several scatterplots that contain distant isolated points have been selected. This result indicates that clumpy may not be an important metric for this particular dataset used in this study. It might have been better to adopt instead a metric that directly evaluates the presence of outliers, for example.

4.2 Statistics of the result

Figure 8 shows the statistics of areas d_{ij} and maximum score values $max_i(s_{ij})$ of the scatterplot selected/unselected in Fig. 1. This figure demonstrates that the presented technique tends to select scatterplots that have larger $max_i(s_{ij})$ values and pairs of scatterplots that have larger d_{ij} values preferentially.

The result of scatterplot selection strongly depends on the choice of d_{thres}. The smaller d_{thres} brings a larger number of edges and consequently a larger number of scatterplots groups corresponding to the

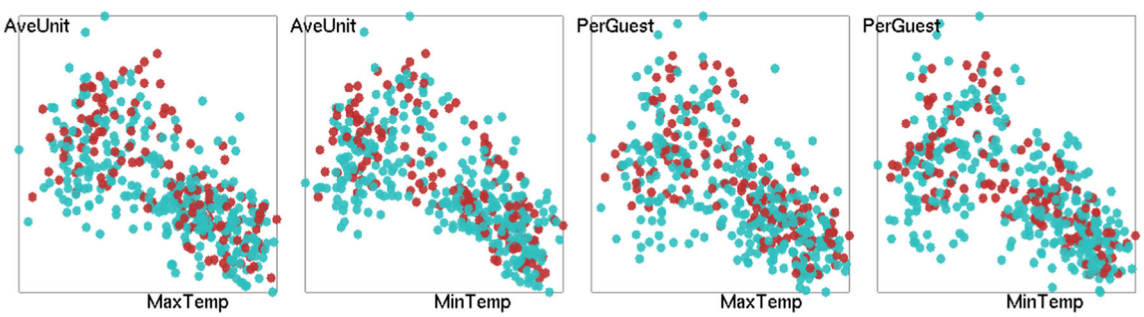

Fig. 4 Scatterplots that achieved the highest scores on correlation

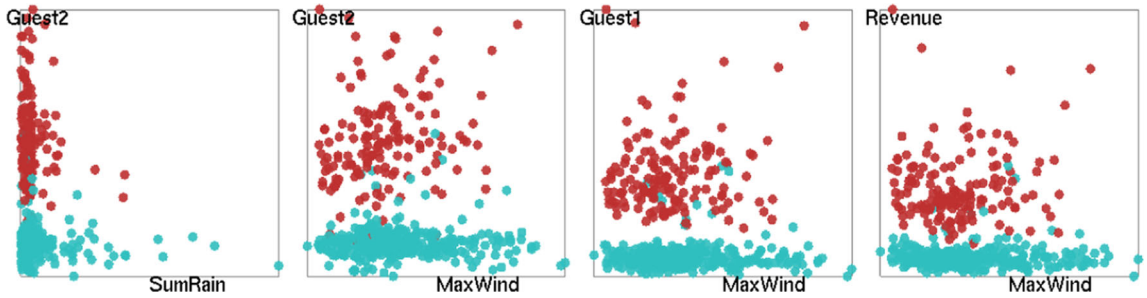

Fig. 5 Scatterplots that achieved the highest scores on separateness

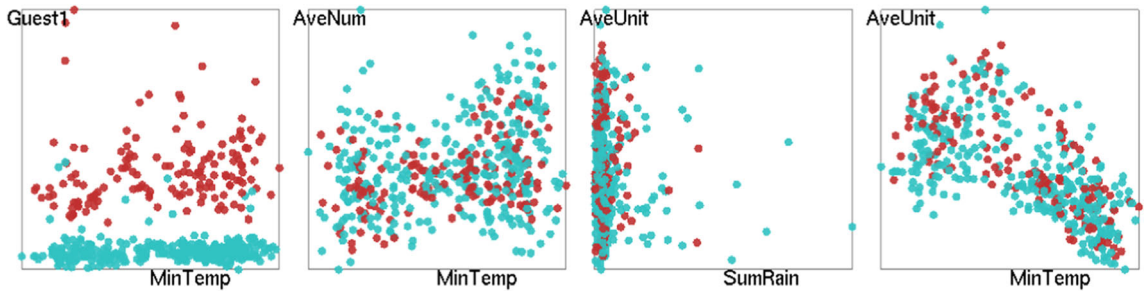

Fig. 6 Scatterplots that achieved the highest scores on clumpy

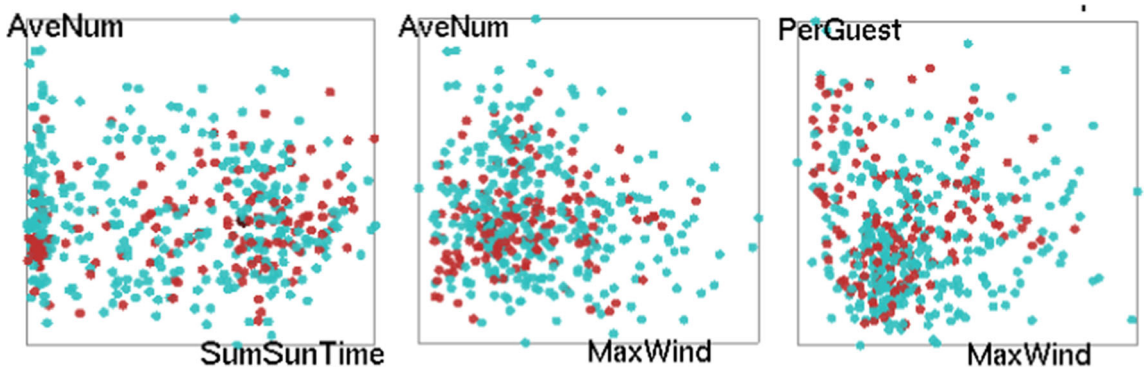

Fig. 7 Example of scatterplots that have no higher scores with all metrics

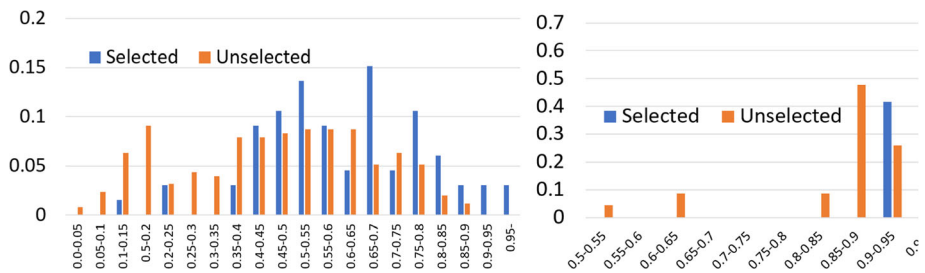

Fig. 8 Statistics of scatterplots selected/unselected in Fig. 1. Vertical axes denote the ratios of the number of corresponding scatterplots. (Left) Statistics of areas d_{ij}. (Right) Statistics of maximum score values $max_i(s_{ij})$

number of colors in Fig. 3. Table 2 shows the numbers of edges and colors, the number of scatterplots belonging to the selected color. The table also shows the minimum d_{ij} and $max_i(s_{ij})$ values among the displayed scatterplots supposing twelve of them are displayed. The result shows that selection of similarly looking or less informative scatterplots would be avoided when a larger number of colors are made and the

Table 2 Trade-off between the minimum d_{ij} and $min(s_{ij})$ values

d_{thres}	0.45	0.5	0.55	0.6	0.65
Num. edges	26	49	124	228	299
Num. colors	7	9	10	14	19
Num. scatterplots	29	26	26	20	15
minimum d_{ij}	0.5158	0.5158	0.5549	0.6115	0.6512
minimum $max_i(s_{ij})$	0.9289	0.9289	0.9247	0.9189	0.8935

minimum d_{ij} value gets larger. But simultaneously, the informativeness of the selected scatterplots may be decreased because the minimum $max_i(s_{ij})$ values get smaller. In other words, one of the features of the presented technique is that users can easily deal with this trade-off problem just by adjusting the d_{thres} value.

5 Case study 2: design optimization data

This section introduces the second case study using a dataset of the optimization process of an aircraft wing shape design. In this case study, the wing shape was designed with 72 explanatory variables, and 4 objective functions were calculated by hydrodynamic simulation. This process was iterated using a multi-objective genetic algorithm to obtain 776 Pareto solutions Sasaki et al. (2002). In other words, this optimization process yielded 776 different design results. The designer can choose the blade shape by making a decision among these design results.

We describe the explanatory variables as dv_{00} to dv_{71} in this section. Among these, the following six explanatory variables are well-known to be particularly important in finding the optimal solution.

dv_{00}, dv_{01}: Span lengths of the inboard/outboard wing panels.
dv_{02}, dv_{03}: Leading-edge sweep angles.
dv_{04}, dv_{05}: Root-side chord lengths.

Other explanatory variables include the following:

dv_{06} to dv_{25}: Variables to define the inner surface connecting corresponding points on upper and lower surfaces of the wing.
dv_{26} to dv_{32}: Variables to design the twist of the wing.
dv_{33} to dv_{71}: Variables to design the thickness of the wing.

The following are four objective functions applied to the optimization process.

CD_t: Drag coefficient during transonic cruise.
CD_s: Drag coefficient during supersonic cruise.
M_b: Bending moment at the wing root during supersonic cruise.
M_p: Pitching moment during supersonic cruise.

We visualized this dataset as a set of 776 points that constitute 76-dimensional real values.

By the way, in this data, each of the 776 samples does not have its own class. Instead, we attempted to color the point clouds displayed in the scatterplot with a real variable of the k-th dimension. Specifically, we divide the interval $[min_k, max_k]$ indicated by the minimum and maximum values of the k-th real values a_{1k} to a_{nk} of each sample into N intervals, identify to which interval each value of a_{1k} to a_{nk} belongs, and color the points in the interval to which each sample belongs. The points are colored in the interval to which each sample belongs. The above process corresponds to the conversion of the k-th real variable into N classes. Then, the i-th and j-th real variables are assigned to the two axes of a scatterplot. The scatterplot can represent the distribution of the three real variables, i, j, and k.

We expect that the following trends can be observed from the point cloud that constitutes the Pareto solution by applying our technique:

- Strong correlation between two variables
- Correlation across three or more variables
- Clusters separated from other point clouds

We verified whether the above trends can be discovered from the dataset.

This section introduces the visualization results with color-coding the scatterplots by the explanatory variable dv_{05} to explain the trends of the optimization dataset. Figure 9 shows the scatterplots selected by

this technique. These scatterplots have one of the explanatory variables as the horizontal axis and one of the objective variables as the vertical axis. The interval of the explanatory variable dv_{05} is divided into three parts, and one of the colors yellow, magenta, or cyan is given to the points. Among the scatterplots displayed in Fig. 9, M_b is assigned to the vertical axis of most of the scatterplots in which the three colors are well-separated, and the three colors are separated in the vertical direction. The result suggests that the explanatory variable dv_{05} has a particularly strong correlation with M_b.

The two scatterplots in the lower left corner of Fig. 9 (Fig. 10) are ones that dv_{00} and M_b, and dv_{02} and M_p are assigned to the two axes, respectively. Our previous study Itoh et al. (2017) showed that the five variables dv_{00}, dv_{04}, dv_{05}, CD_t, and M_b are strongly correlated, and the four variables dv_{02}, dv_{03}, CD_s, and M_p are also strongly correlated. Figure 10 (left) shows the strong correlation between dv_{00} and M_b from the distribution of the point cloud, and the correlation with dv_{05} from the color separation of the point cloud. Meanwhile, Fig. 10 (right) also shows the strong correlation between dv_{02} and M_p, but the correlation with dv_{05} is not so strong because the colors of the point clouds are not well-separated. Furthermore, comparing the two scatterplots in Fig. 10, we can see that the correlation between dv_{02} and M_p is sharper than that between dv_{00} and M_b.

Figure 11 is another two scatterplots included in Fig. 9, where dv_{57} and M_b, dv_{07} and M_p are assigned to the two axes, respectively. The upper right part of the scatterplot in Fig. 11 (left) and the upper left part of the scatterplot in Fig. 11 (right) are almost blank. This represents the coarseness in the distribution of the Pareto solution. We expect that the optimization process can be made more efficient by analyzing the coarseness and denseness that can be observed only when specific variables are assigned to the horizontal axis.

On the other hand, we could not discover any outliers or clusters were in any of the scatterplots that were clearly separated from the rest of the point cloud. In other words, the 776 Pareto solutions in this dataset are distributed continuously in the 76-dimensional space, and therefore no fragmentations among the Pareto solutions are found.

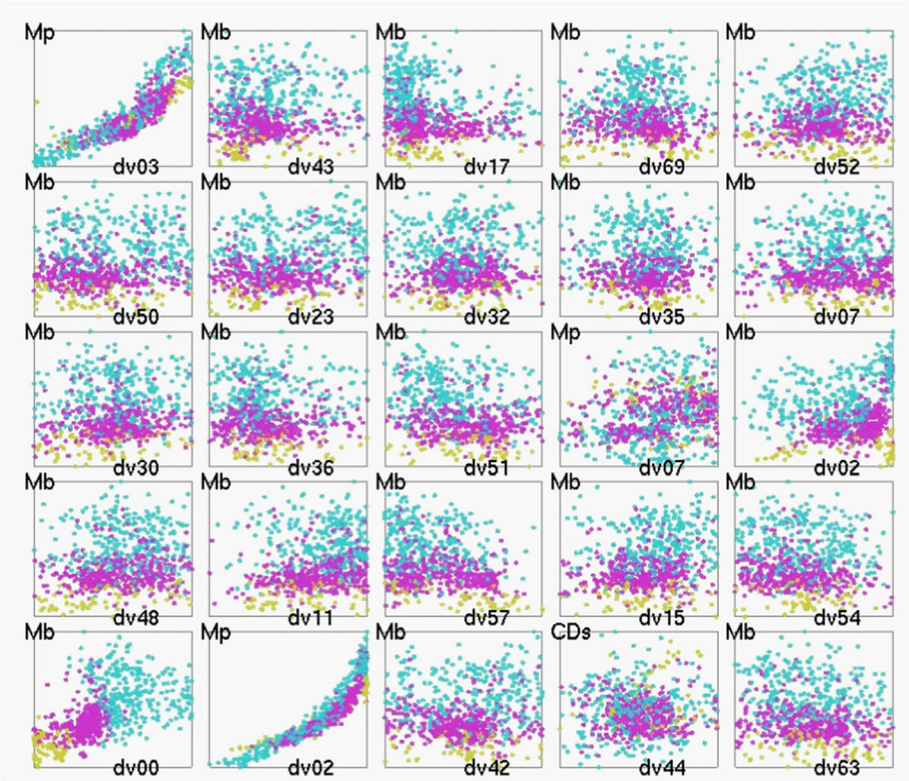

Fig. 9 Scatterplot selection result with color-coding by the explanatory variable dv_{05}

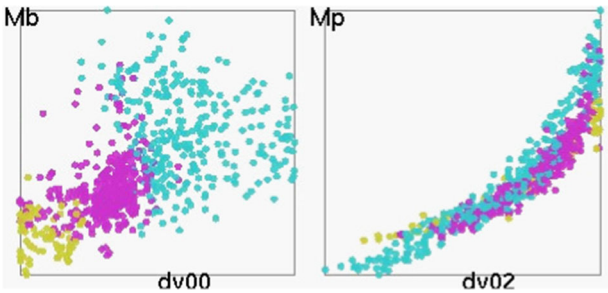

Fig. 10 (Left) dv_{00} and M_b are assigned to the axes. (Right) dv_{02} and M_p are assigned to the axes

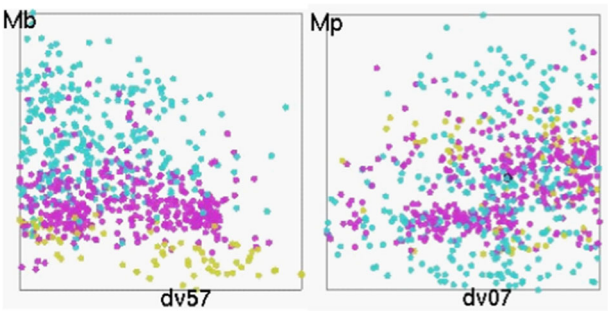

Fig. 11 (Left) dv_{57} and M_b are assigned to the axes. (Right) dv_{07} and M_p are assigned to the axes

6 Conclusion and future work

This paper presented a multidimensional data visualization technique that represents various characteristics of input datasets as a variety of scatterplots. The technique automatically selects scatterplots using a graph coloring algorithm. The technique calculates scores based on several independent metrics for each scatterplot. Then, it constructs a graph by connecting vertex pairs corresponding to scatterplot pairs if these scores are similar. The graph coloring algorithm is used for the graph, and scatterplots that have the user-specified color are extracted. The paper introduced a case study with a design optimization dataset as well as another case study with a retail transaction dataset presented in the authors' conference paper Itoh et al. (2021).

Our future studies include the following. First, we would add and modify the metrics. The metrics presented in this paper are selected based on the requirements of the data owner of the first case study, and therefore, other metrics may be necessary while applying other fields of datasets. In addition, there have been various improved metrics for scagnostics. We will apply them and explore the best combination of the metrics for this study. Then, we will test the scalability of the presented technique. Particularly, we suppose it is necessary to test datasets with a large number of dimensions; therefore, a large number of scatterplots can be generated. It is also important to test datasets with a large number of individuals. Finally, we would like to conduct user evaluations to verify the satisfaction of users with the scatterplot selection results.

Acknowledgements We appreciate ABEJA, Inc. for providing the retail transaction dataset. Also, we appreciate Prof. Shigeru Obayashi of Tohoku University, Japan, for providing the design optimization dataset.

References

Aupetit M, Sedlmair M (2016) Sepme: 2002 new visual separation measures. In: IEEE Pacific visualization symposium 2012:43–52

Claessen JHT, van Wijk JJ (2011) lexible linked axes for multivariate data visualization. IEEE Trans Visual Comput Graphics 17(12):2310–2316

Dang TN, Wilkinson L (2014) Scagexplorer: Exploring scatterplots by their scagnostics. In: IEEE Pacific visualization symposium 2014:73–80

Harrison L, Yang F, Franconeri S, Chang R (2014) Ranking visualizations of correlation using weber's law. IEEE Trans Visual Comput Graphics 20(12):1943–1952

Itoh T, Kumar A, Klein A, Kim J (2017) High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. J Vis Lang Comput 43(1):1–13
Itoh T, Nakabayashi A, Hagita M (2021) Scatterplot selection applying a graph coloring algorithm. In: The 14th International symposium on visual information communication and interaction (VINCI)
Lee JH, McDonell KT, Zelenyuk A, Imre D, Muller K (2013) A structure-based distance metric for high-dimensional space exploration with multidimensional scaling. IEEE Trans Comput Graph 20(3):351–364
Liu S, Wang B, Bremer P-T, Pascucci V (2014) Distortion-guided structure-driven interactive exploration of high-dimensional data. Comput Graph Forum 33(3):101–110
Matute J, Telea AC, Linsen L (2017) Skeleton-based scagnostics. IEEE Trans Comput Graph 24(1):542–552
Nakabayashi A, Itoh T (2019) A technique for selection and drawing of scatterplots for multi-dimensional data visualization. In: 23rd international conference on information visualisation (IV2019), pp 62–67
Nohno K, Wu H-Y, Watanabe K, Takahashi S, Fujishiro I (2014) Spectral-based contractible parallel coordinates. In: 18th international conference on information visualisation, pp 7–12
Sasaki D, Obayashi S, Nakahashi K (2002) Navier-stokes optimization of supersonic wings with four objectives using evolutionary algorithm. J Aircr 39(4):621–629
Sedlmair M, Tatu A, Munzner T, Tory M (2012) A taxonomy of visual cluster separation factors. Comput Graph Forum 31(3):1335–1344
Shao L, Mahajan A, Schreck T, Lehmann DJ (2017) Interactive regression lens for exploring scatter plots. Comput Graph Forum 36(3):157–166
Sips M, Neubert B, Lewis JP, Hanrahan P (2009) Selecting good views of high-dimensional data using class consistency. Comput Graph Forum 28(3):831–838
Suematsu H, Zheng Y, Itoh T, Fujimaki R, Morinaga S, Kawahara Y (2013) Arrangement of low-dimenional parallel coordinate plots for high-dimensional data visualization. In: 17th international conference on information visualisation, pp 59–65
Wang Y, Wang Z, Liu T, Correll M, Cheng Z, Deussen O, Sedlmair M (2020) Improving the robustness of scagnostics. IEEE Trans Visual Comput Graphics 26(1):759–769
Watanabe A, Itoh T, Kanazaki M, Chiba K (2017) A scatterplots selection technique for multi-dimensional data visualization combining with parallel coordinate plots. In: 21st international conference on information visualisation (IV2017), pp 78–83
Wilkinson L, Anand A, Grossman R (2005) Graph-theoretic Scagnostics. In: IEEE symposium on information visualization, pp 157–164
Yuan X, Ren D, Wang Z, Guo C (2013) Dimension projection matrix/tree: interactive subspace visual exploration and analysis of high dimensional data. IEEE Trans Visual Comput Graphics 19(12):2625–2633
Zhang Z, McDonnel KT, Zadak E, Muller K (2015) Visual correlation analysis of numerical and categorical data on the correlation map. IEEE Trans Visual Comput Graphics 21(2):289–303
Zhang Z, McDonnell KT, Mueller K (2012) A network-based interface for the exploration of high-dimensional data spaces. In: IEEE Pacific visualization symposium 2012:17–24
Zheng Y, Suematsu H, Itoh T, Fujimaki R, Morinaga S, Kawahara Y (2015) Scatterplot layout for high-dimensional data visualization. J Visualization 18(1):111–119