A Technique for Selection and Drawing of Scatterplots for Multi-Dimensional Data Visualization

*

Asuka Nakabayashi
Grad. School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan
g1520533@is.ocha.ac.jp

Takayuki Itoh
Grad. School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan
itot@is.ocha.ac.jp

***Abstract*—Scatterplot matrix and parallel coordinate plots are well-used multi-dimensional data visualization techniques. These techniques have a problem that they need a very large screen space when an input dataset has an enormous number of dimensions. To solve this problem, we propose a method for selecting important scatterplots from all scatterplots generated from input datasets and for drawing the scatterplots as "outliers" and "regions enclosing non-outlier plots." The technique is useful for users to determine whether to delete outliers from the datasets and form mathematical models of non-outlier plots. This paper introduces an example of visualization using this technique with a retail transaction dataset and climate values.**

***Index Terms*—Multi-Dimensional Data, Visualization, Scatterplot**

I. Introduction

There have been many multi-dimensional datasets in daily life and application domains. Features and regularity of multi-dimensional data are important knowledge for understanding and utilizing the data. We can easily discover the features and trends of multi-dimensional data by using effective visualizations.

Scatterplot matrix (SPM) and parallel coordinate plots (PCPs) [3] are well-known multi-dimensional data visualization techniques. To visualize n-dimensional data, SPM generates scatterplots with arbitrary pairs of dimensions and arranging them as a $n \times n$ matrix. Meanwhile, PCPs displays n parallel axes, plots the values on the axes, and connects them by polygonal lines. Although these techniques can visualize all dimensions of input datasets without deficiency of information, it is problematic that they need a very large screen space when the input dataset has an enormous number of dimensions. Also, we cannot expect that interesting features and trends are necessarily discovered in all dimensions of the input dataset. As a result of this discussion, there have been many visualization techniques which selectively display only a limited number of meaningful dimensions in recent years.

Mathematical modeling of multi-dimensional data is another interesting and important issue. We may need to discuss what types of models can be applied to the multi-dimensional data after removing outliers. Machine learning is a typical task which users may need to select appropriate mathematical models for multi-dimensional data. We expect visualization techniques can contribute to assist the model selection with multi-dimensional datasets.

We present a multi-dimensional data visualization technique addressing the above problems in this paper. This technique consists of the following two processing steps.

- Selection and display of important scatterplots from all possible scatterplots.
- Drawing of scatterplots as "outliers" and "regions enclosing non-outlier plots."

This paper describes the processing flow of the presented technique and introduces a case study with a retail transaction dataset.

II. Related Work

A. Multi-dimensional data visualization with dimension selection

There have been many multi-dimensional data visualization techniques which select a set of low-dimensional subspaces that are meaningful to visualize. For example, Zheng et al. [11] presented a technique which selects scatterplots that satisfy user-specified criteria from all scatterplots generated from input datasets and arranging them based on the dissimilarity among the scatterplots. Suematsu et al. [8] presented a technique which generates low-dimensional PCPs from meaningful groups of dimensions and arranging them based on their dissimilarity. However, these techniques do not the interactive control of the number of PCPs or scatterplots to be displayed.

Itoh et al. [4] presented a visualization technique for multi-dimensional data to solve this problem. This technique displays a set of low-dimensional subspaces as a set of low-dimensional PCPs on the left side of the screen, which are

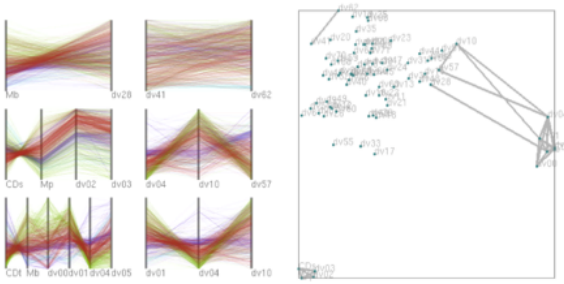

Fig. 1. A snapshot of visualization by Itoh et al. [4].

selected interactively with the dimension graph displayed on the right side, as shown in Figure 1. Note that Figure 1 is excerpted from Itoh et al. [4], which represents the data other than what will be described later. Watanabe et al. [9] presented an extended technique which visualizes as a combination of low-dimensional PCPs and scatterplots. This technique applies scatterplots for pairs of dimensions if the pairs have interesting numeric distributions, but it is difficult to represent by PCPs effectively.

Though the technique presented in this paper also selects interesting pairs of dimensions as our previous techniques [4], [9], the technique displays the selected pairs of dimensions only by scatterplots, not by PCPs.

B. Multi-dimensional data visualization using scatterplots

Wilkinson et al. [10] and Dang et al. [6] presented scatterplot-based multi-dimensional data visualization techniques. Wilkinson et al. proposed a method to quantitatively evaluate nine features called Scagnostics based on the appearance of the scatterplots generated from input datasets. Dang et al. proposed a method to cluster scatterplots characterized by Scagnostics generated from input datasets, select scatterplots and display applying a force-directed layout algorithm. These techniques do not need a very large screen space even if an input dataset has an enormous number of dimensions.

Scagnostics can be applied to the technique presented in this paper while selecting the scatterplots; however, our current implementation applies different criteria as described in Section 3.

C. Density-aware scatterplots

Continuous Scatterplots [1] can map the density of multi-dimensional items into m-dimensional scatterplots in consideration of an arbitrary density defined on an input field of the n-dimensional domain. This method combines statistical visualization such as scatterplots, with scientific visualization such as volume or flow visualization. Our technique presented in this paper also continuously paints the scatterplots; however, we do not directly consider density while painting the scatterplots.

III. Proposed Technique

This section describes our technique for selecting important scatterplots generated from input multi-dimensional datasets and for drawing the scatterplots as "outliers" and "regions enclosing non-outlier plots."

Our current implementation provides the same two types of algorithms for scatterplot selection as Itoh et al. [4]. The first algorithm calculates correlation coefficients between arbitrary pairs of dimensions and selects scatterplots formed from pairs of dimensions which have larger absolute values of correlation coefficients. Let d_1 and d_2 be dimensions, the absolute value of the correlation coefficient d_{d_1,d_2} between d_1 and d_2 is defined as

$$d_{d_1,d_2} = |1.0 - f_c(d_1, d_2)| \tag{1}$$

where $f_c(d_1, d_2)$ denotes Spearman's rank correlation coefficients.

The second algorithm calculates the entropy of positions of individuals on a scatterplot treating values of a categorical dimension as labels: this mechanism is useful to select scatterplots those specific labels of individuals are distantly positioned from others. In particular, we compute the entropy H for all pairs of dimensions

$$H(d_1, d_2) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} p(y_i = c|x_i^{d_1,d_2}) \log p(y_i = c|x_i^{d_1,d_2}) \tag{2}$$

where x_i represents the i-th vector from an n-dimensional input dataset Ds defined as $Ds = \{x_1, \ldots, x_N\}$, y_i represents the label assigned to the i-th vector, N and C represent the number of vector and the number of labels, and $p(y_i = c|x_i^{d_1,d_2})$ represents the probability that the c-th class Y_c is assigned to the i-th vector x_i. Moreover, $x_i^{d_1,d_2}$ is a two-dimensional vector containing the dimensions d_1 and d_2 of x_i. This value represents the separation of labels in a scatterplot generated by d_1 and d_2.

Our technique also draws scatterplots as "outliers" and "regions enclosing non-outlier plots." The current implementation applies Delaunay triangulation to connect adjacent dots in the scatterplots and detect outliers from the edges of the triangular mesh. Delaunay triangulation is a generic method which connects vertices scattered in a 2D space and generates a triangle mesh. We apply an incremental triangulation algorithm which firstly generates a large rectangle surrounding every vertex in a scatterplot as the initial mesh, then adds the vertices one-by-one, and connects the vertices to refine the triangular mesh.

After constructing the mesh, our implementation deletes edges which are longer than a user-specified threshold and also deletes triangles which contain such edges. Here, we extract vertices that are not connected to any vertices as outliers as shown in Fig 2. Users can control the threshold to adjust the number of outliers. Then, the technique represents the regions enclosing non-outlier plots by drawing the region boundary of the triangles that connect the non-outlier plots by a dark color and filling the triangular mesh by a light color.

Also, users can delete arbitrary outliers from input datasets and redraw the scatterplots as shown in Fig 3. When the outlier at the upper right in Fig 3 (left) is deleted, the scale of the vertical axis is updated, and the non-outlier region is also

updated as shown in Fig 3 (right). The non-outlier regions can be enlarged when we delete the distant outliers. Therefore, users can discover new features, trends, and potential outliers that are previously undiscovered.

IV. Example

We implemented the presented technique with Java Development Kit (JDK) 1.8.0, reusing the implementation of Itoh et al. [4] for scatterplots selection and user interfaces.

This paper introduces an example of visualization by the presented technique applying a retail transaction dataset and climate values. Table I 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. These values are recorded in the dataset day-by-day. Data points are 457 days from May 1, 2016 to July 31, 2017, and 35 scatterplots consisting of 5 horizontal axes and 7 vertical axes can be analyzed. Remark that this dataset is perturbated by adding random real values to each column of the original dataset.

TABLE I
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

Fig 4 shows an example of a numerical distribution in which one attribute (weekday) is shaped to enclose the other attribute (holiday). This figure depicts that the average revenue per customer and the average price of purchased items are more dispersed on weekdays than on holiday.

Fig 5 shows the numerical distribution of the average price of purchased items and the average number of purchased items from September to November. As the winter approaches, the average price of purchased items tends to increase. This is presumably because the thicker clothes are, the higher the prices of items are. On the other hand, the average number of purchased items tends to decrease. This is also presumably because the prices of items are higher than those of clothes for spring or summer, and therefore customers get more careful while purchasing items.

Figure 6 shows the numerical distribution of the conversion rate in February, July, and November. This figure shows that there are some periods in which the conversion rate is prominently high only in early February and late July. One of the reasons for this phenomena is that a much larger number of customers than visitors for window shopping came there due to special events such as over-stock sales. The figure also depicts that there is only one day with an unusually high conversion rate in November.

V. Conclusion

This paper proposed a multi-dimensional data visualization technique which selects important scatterplots from all possible scatterplots generated from input datasets and draws the scatterplots as ”outliers” and ”regions enclosing non-outlier plots.” The technique presented in this paper is useful for users to determine whether to delete outliers from the datasets and form mathematical models of non-outlier plots.

As future work, it is necessary to draw the regions enclosing non-outlier plots more clearly. When the three or more colors of polygons overlap each other, our current implementation may vaguely paint the overlapping portions, and therefore comprehensibility of the visualization results may be worse. We need to improve the implementation so that many overlapped regions are clearly drawn.

We found experimentally that the two types of algorithms of our current implementation for scatterplot selection do not necessarily select all scatterplots that we subjectively determine important. Thus, we would like to implement new criteria so that such scatterplots are automatically selected.

Another issue is how to extract outliers. Our current implementation to extract outliers is just based on distances on scatterplots. Data items are regarded as outliers if they are distant from all other data items on one of the scatterplots because we extract outliers using Delaunay triangular meshes. There have been various techniques for extracting outliers from multi-dimensional data [2] [5] [7]. We would like to develop our implementation of extracting outliers.

After implementing these functions, we would like to apply more diverse datasets to this technique and further continue with development to be versatile implementation.

Acknowledgment

We appreciate ABEJA, Inc. for providing the dataset.

References

[1] Sven Bachthaler and Daniel Weiskopf. Continuous Scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, pp. 1428–1435, 2008.

[2] Denis Cousineau and Sylvain Chartier. Outliers detection and treatment: a review. *International Journal of Psychological Research*, Vol. 3, No. 1, pp. 58–67, 2010.

[3] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *IEEE Visualization (VIS1990)*, pp. 361–378, 1990.

[4] Takayuki Itoh, Ashnil Kumar, Karsten Klein, and Jinman Kim. High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots. *Journal of Visual Languages and Computing*, Vol. 43, pp. 1–13, 2017.

[5] Chang-Tien Lu, Dechang Chen, and Yufeng Kou. Algorithms for Spatial Outlier Detection. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*, 2003.

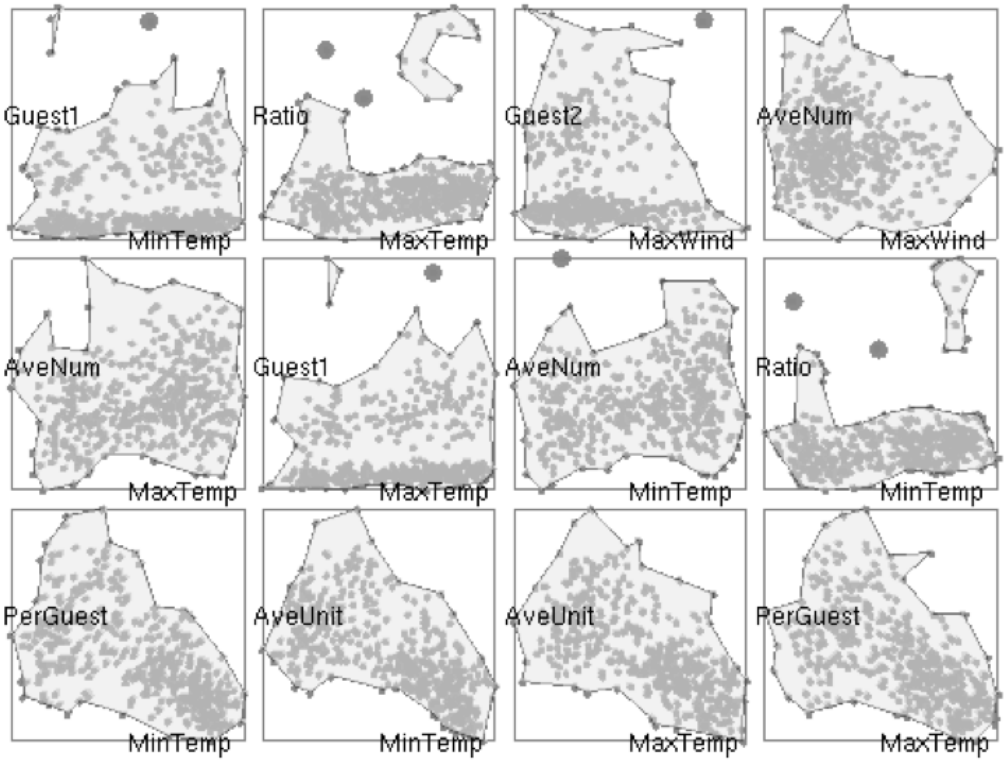

Fig. 2. An example of drawing the scatterplots as ”outliers” and ”regions enclosing non-outlier plots.”

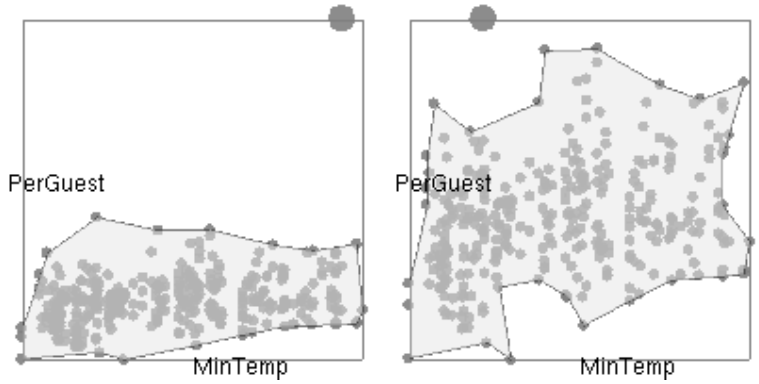

Fig. 3. An example of deleting an outlier. (left) before deleting an outlier. (right) after deleting.

[6] Dang Tuan Nhon and Leland Wilkinson. ScagExplorer: Exploring Scatterplots by Their Scagnostics. In *IEEE Pacific Visualization Symposium (PacificVis 2014)*, pp. 73–80, 2014.

[7] Kay I. Penny and Ian T. Jolliffe. A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, Vol. 50, No. 3, pp. 295–308, 2001.

[8] Haruka Suematsu, Yunzhu Zheng, Takayuki Itoh, Ryohei Fujimaki, Satoshi Morinaga, and Yoshinobu Kawahara. Arrangement of Low-Dimensional Parallel Coordinate Plots for High-Dimensional Data Visualization. In *17th International Conference on Information Visualisation (IV2013)*, pp. 59–65, 2013.

[9] Ayaka Watanabe, Takayuki Itoh, Masahiro Kanazaki, and Kazuhisa Chiba. 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, 2017.

[10] Leland Wilkinson, Anushka Anand, and Robert Grossman. Graph-Theoretic Scagnostics. In *IEEE Symposium on Information Visualization*, pp. 157–164, 2005.

[11] Yunzhu Zheng, Haruka Suematsu, Takayuki Itoh, Ryohei Fujimaki, Satoshi Morinaga, and Yoshinobu Kawahara. Scatterplot layout for high-dimensional data visualization. *Journal of Visualization*, Vol. 18, No. 1, pp. 111–119, 2015.

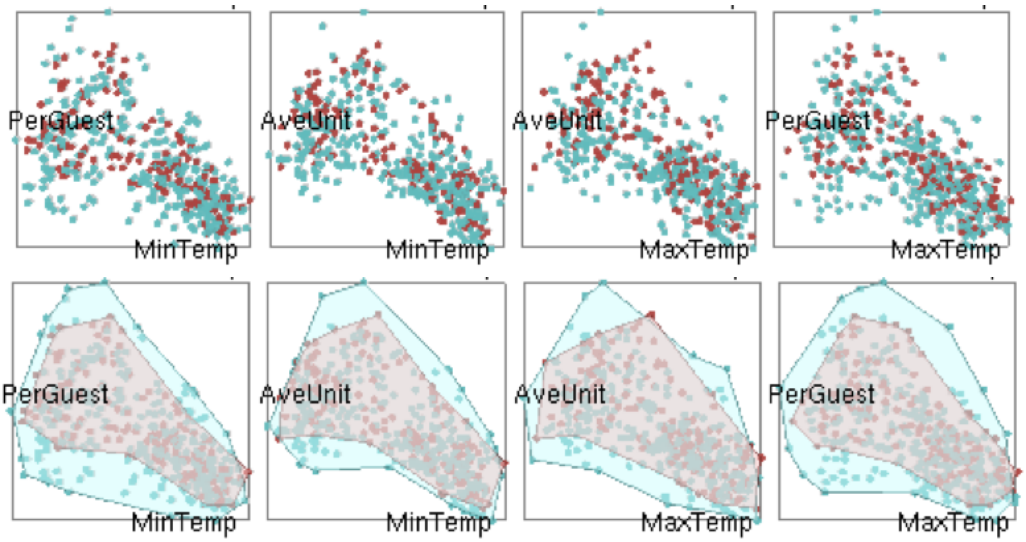

Fig. 4. An example of a numerical distribution in which a non-outlier region of a specific attribute shapes to enclose the non-outlier region of the other attribute.
The light blue region represents the distribution on weekdays while the red region represents the distribution on holidays.
(upper) before enclosing. (lower) after enclosing.

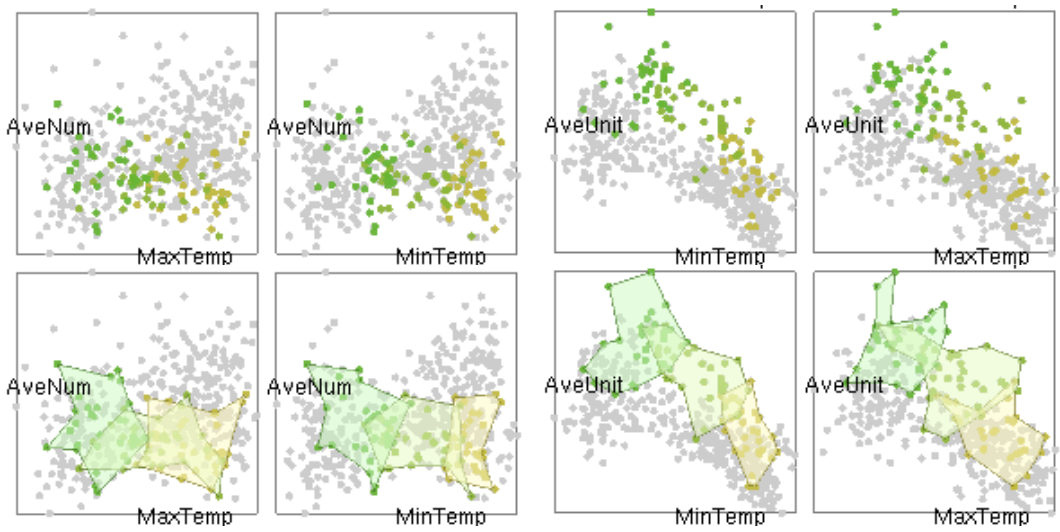

Fig. 5. The numerical distribution of the average price of purchased items and the average number of purchased items from September to November.
The yellow region represents the distribution in September, the yellow green region represents in October and the green region represents in November.
The gray dots represent the distribution in the other months.
(upper) before enclosing. (lower) after enclosing.

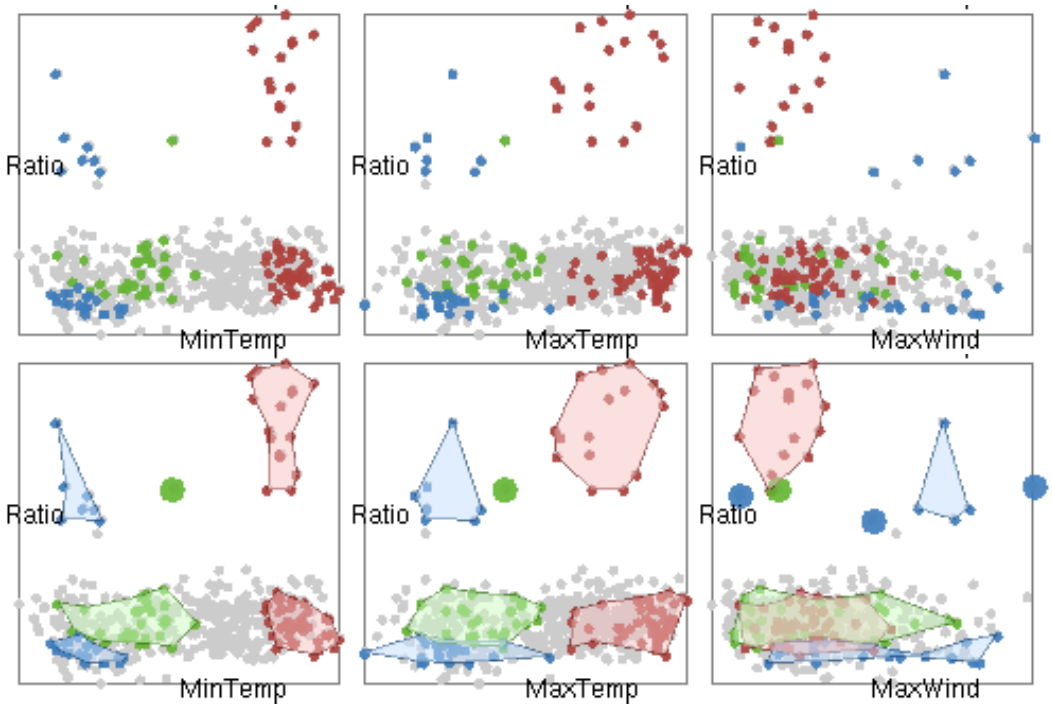

Fig. 6. The numerical distribution of the conversion rate in February, July, and November.
The blue region represents the distribution in February, the red region represents in July and the green region represents in November.
The gray dots represent the distribution in the other months.
(upper) before enclosing. (lower) after enclosing.