Application Of The Parallel Coordinate Plot For Ranking Comparison Between Two Groups

Fujino, Tomokazu
Fukuoka Women's University, Department of Environmental Science
1-1-1, Kasumigaoka, Higashi-ku
Fukuoka, 813-8529, Japan
E-mail: fujino@fwu.ac.jp

Yoshiro, Yamamoto
Tokai University, Department of Mathematics
4-1-1, Kitakaname
Hiratsuka, Kanagawa, 259-1292, Japan
E-mail: yama@toka-u.ac.jp

1 Introduction

Wegman [1] showed that a parallel coordinate plot [2] can be used for data analysis. This is one of the statistical graphics for the visualization of multivariate data and has become an important tool for exploratory data analysis. Swayne et al. [3] developed software with interactive functions such as linking, highlighting and brushing for various statistical graphics including parallel coordinate plots. Furthermore, Edsall [4] proposed tools for the interactive and exploratory analysis of complicated spatial data by linking parallel coordinate plots with maps. This paper presents and illustrates a ranking comparison plot through an example of actual POS data analysis, which is a visualization method for detecting the relationship between order fluctuation for two variables and other contributing factors on the basis of a parallel coordinate plot. Moreover, we developed tools for generating ranking comparison plot as an interactive statistical graphics.

2 Ranking Comparison Plot

For the ranking comparison plot, we compared the product sales of two shops, based on POS data. The basic ranking comparison plot was then plotted for the data shown in Table 1. Let

$r_A(i)$: the item number where the ranking of unit sales for Shop A is i
$r_B(j)$: the item number where the ranking of unit sales for Shop B is j

where $i, j = 1, 2, \ldots, n$. In addition, the parameters of the unit prices for the items are divided into K applicable classes and the boundary value is decided as $c_1 < c_2 < \cdots < c_{K+1}$. Furthermore, colors on the basis of the appropriate color scheme are assigned to each divided classes. In this way, color of item ℓ based on unit price is expressed as

$$\text{col } p_\ell = \text{color of class } k \text{ if } c_k \leqq p_\ell < c_{k+1}.$$

The ranking comparison plot thus resembles Figure 1. First, on the axis of the parallel coordinate corresponding to Shop A and Shop B, the items are aligned according to their sales ranking at each shop. Furthermore, placed next to each item is a rectangle that is colored according to the unit price. Then, for $u = 1, 2, \ldots, n$, a line is drawn between the rectangles corresponding to the items where

$$r_A(s) = r_B(t) = u$$

Table 1: data format for ranking comparison plot

item	unit sales of shop A	unit sales of shop B	unit price of item
1	a_1	b_1	p_1
2	a_2	b_2	p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	a_i	b_i	p_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	a_n	b_n	p_n

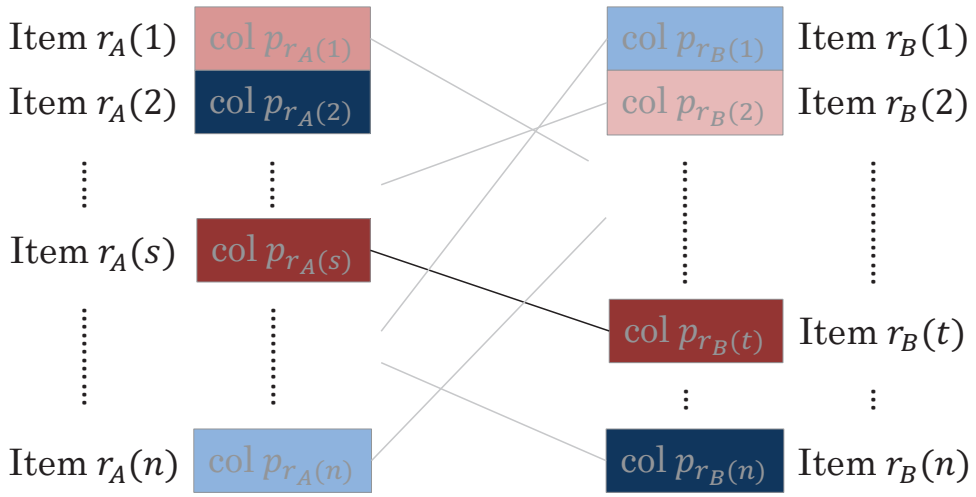

Figure 1: ranking comparison plot

From the ranking comparison plot, it is possible to intuitively and simultaneously understand the sales ranking of products in each shop, price distribution per ranking and variation in sales ranking between the two shops. However, if there are a large number of items, it gets more difficult to grasp the ranking fluctuations of each item. Thus, it is effective to highlight items according to category or pricing.

3 Applications

Here we have shown an example of a ranking comparison plot created with actual POS data proposed in the "2010 Data Analysis Competition" organised by the Japanese Federation of Council for Management Science Research. The data is arranged into four tables, as seen in the Figure 1. The POS table lists the actual sales data, and each record in the POS table lists the data equivalent to the each item in the receipt.

Our group calculated the frequency and cost of purchases for each shop according to age, gender and date of purchase in order to understand the characteristics relating to the sales of each shop. After calculating the unit prices for each shop, it was found that there was a difference of 84 JPY in the unit prices between Shop 10 and Shop 12. However, because these two shops belonged to the same chain, the calculated results and the shops' attributes (locations, floor space, number of parking

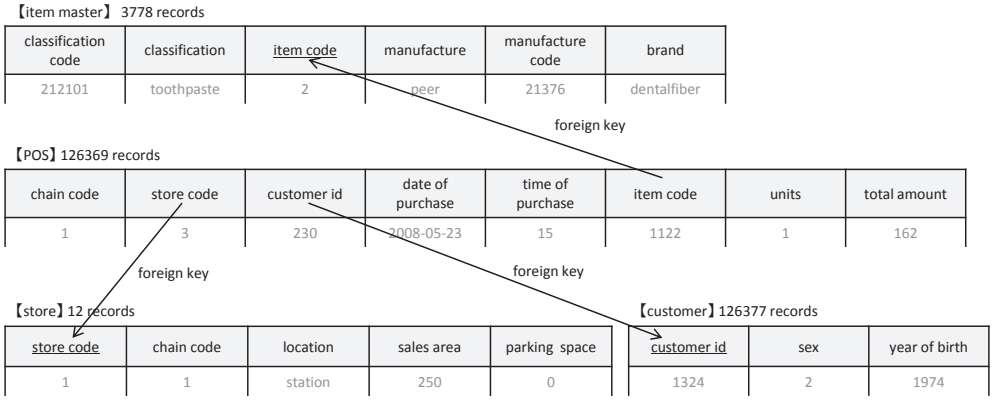

Figure 2: table format of POS data

spaces) were almost similar. Therefore, it was decided that a more detailed analysis of the levels of products could be gathered, specifically from these shops. In order to find differences between the unit prices of the two shops, we developed a ranking comparison plot that could simultaneously perceive the variations in price distribution between two shops, the price distribution according to ranking and the sales ranking for products in each shop. Figure 3 and 4 indicate ranking comparison plots for the sales ranking of toothpaste in Shops 10 and 12. Although they are both on the same plot, the emphasized parts are different. In the figure 3, segments showing rather high-priced products and variations in their ranking are emphasized, while in the figure 4, those of the lowest-priced products are emphasized.

Looking at the overall price distribution for toothpaste, lower-priced products are at most of the top positions. Some high-priced products can be seen at the middle positions and mid-priced products are at the bottom positions. In the figure 3, which emphasizes somewhat higher-priced products, we find that products selling relatively well in Shop 10 with specific results show a large drop in sales ranking in Shop 12. Furthermore, in the figure 4, which emphasizes the lowest-priced products, the majority of the products have a high ranking in Shop 12.

4 Implementing Interactive Functions

According to the ranking comparison plot, the greater the range and number of products, the greater their output. In order to extract products requiring attention, it is desirable to be able to output a statistical graphics equipped with interactive functions. In this study, we developed a script that can generate output from the statistical software R, allowing the user to use a mouse to designate the pricing, product type and scope of variation in ranking, and then devise a ranking comparison plot which is able to emphasize the applicable items. Output from this script is in the Scalable Vector Graphics (SVG) format, the standard for vector graphics on the web. By reading these files with leading web browsers, it is possible to use these interactive functions of ranking comparison plots. As SVG is a text file, it is easy to dynamically generate data, and it is possible to implement interactive functions using JavaScript. Examples of statistical graphics with these types of functions developed with SVG, therefore, are extremely common [5].

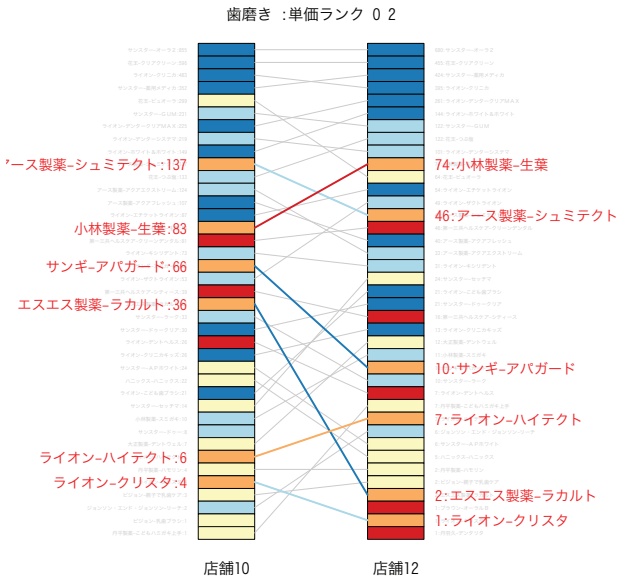

Figure 3: ranking comparison plot which emphasizes the higher-priced products

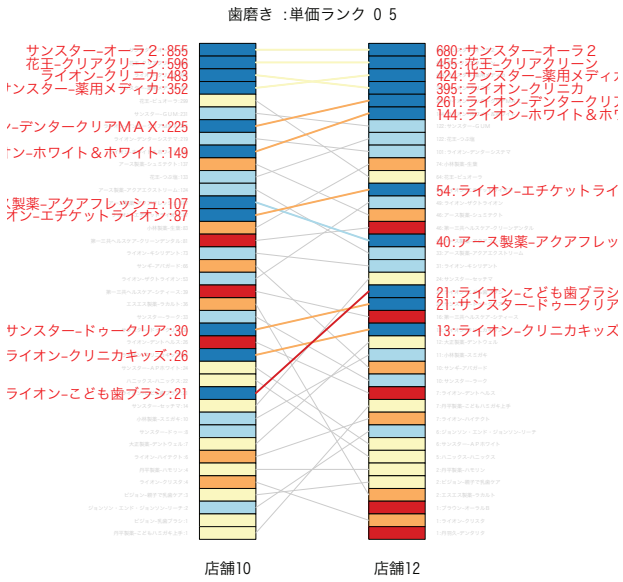

Figure 4: ranking comparison plot which emphasizes the lowest-priced products

5 Discussion

Ranking comparison plots are believed to be a powerful tool for comparing sales trends of products among various shops. In the future, we plan to implement the ability to dynamically swap numbers, display three or more stores and provide links to other statistical graphics.

References

[1] E.J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85:664–675, 1990.

[2] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985.

[3] D. F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998.

[4] R. M. Edsall. The parallel coordinate plot in action: design and use for geographic visualization. *Computational Statistics & Data Analysis*, 43:605–619, 2003.

[5] Y. Yamamoto, M. Iizuka, and T. Fujino. In C. Chen, W. Härdle, and A. Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, chapter Web-Based Statistical Graphics using XML Technologies, pages 757–789. Springer, 2007.