Effectiveness of a Kanji Graph from the Point of View of Clustering

(1997年４月２日受理)

福森　護
Mamoru Fukumori

Key words: Graphical method, Kanji graph, Clustering

abstract

In this paper, a new graphical method called kanji graph is proposed, and its effectiveness is examined from the point of view of clustering. The kanji graph describes the value of the data in the height or width of a kanji. Thus two variables can be represented by one character of kanji, and multi-variables are represented by different kanji, preferably by characters of kanji whose meanings correspond to the variables, arranged in a row. In order to examine the effectiveness of the kanji graph, data from the restaurant business (Japan Almanac, 1985) is utilized. An experiment is carried out to compare the kanji graph with dendrogram, face graph and principal component plot.

1. Introduction

According to Osumi (1979) it is possible to classify clustering techniques as follows:

1) Principal component analysis and related techniques
2) Hierarchical clustering techniques: Agglomerative type (e. g., Word method), Divisive type (e. g., Association Analysis, AID)
3) Non-hierarchical clustering techniques such as K-means and ISODATA
4) Mode method
5) Fuzzy clustering techniques
6) Diagrammatic techniques

The techniques which contain graphical representation are (1), (2) and (6). Individual clustering methods in (1) include principal component analysis, discriminant analysis, factor analysis and various types of multi-dimensional methods. These methods derive new variables of smaller numbers than the original set without losing important information. The results are shown by way of graphical plots. Hierarchical technique use the so-called dendrogram to display the results of analysis. Finally, diagrammatic techniques (6) used in clustering include radar chart, face graph (Chernoff, 1973) , constellation graph (Wakimoto and Taguri, 1978) , Andrews plot (Andrews, 1973), kanji (letter) graph (Hirai, Fukumori and Wakimoto, 1988).

In this paper, the effectiveness of kanji graph is examined from the point of view of clustering.

Generally, graphical methods are roughly divided into two types: the descriptive type and the analytical type. Of these two types one includes methods such as radar chart, face graph, constellation graph, tree graph and linked vector graph, and the other includes methods such as biplot, non linear mapping, triangle polynomial graph, probability plot and statistical ellipse. The kanji graph is the former type. This graph can be applied to cluster analysis, and can present the features of data clearer than other graphical methods in the descriptive category.

Wakimoto et al. (1979) discussed the relationship between multivariate statistical analysis and graphical methods. They arranged the graphical methods into groups in accordance with multivariate methods. According to Wakimoto et al. (1979) the graphical methods concerned with clustering can be summarized as scatter plot, radar chart, face graph, constellation graph, biplot, non-linear mapping and dendrogram. Further work by Goto et al. (1986) arranged graphical representation in accordance with usage of the statistical method. In their study, the following graphical methods can be used for the purpose of cluster analysis or discriminant: scatter plot, histogram, constellation graph, face graph, dendrogram, SHADE, AID, tree and castle graph.

Fukumori et al. (1994, 1995) discussed the effectiveness of kanji graph from the cognitive aspect. In their study, three kinds of graphical representation, i, e., radar chart, kanji graph and face graph, are selected and compared. The results show that those three graphical methods have different properties such that it is easier to classify the kanji graphs when the number of variables are small while the case of five variables is the easiest for the other two kinds of graphical representation.

To evaluate the effectiveness of the kanji graph, this paper shows its application to clustering and discusses the effectiveness of the kanji graph by comparing it to other graphical techniques.

2. Producing the kanji graph

In the kanji graph, the value of the multivariate data is denoted by the size of the character. The graph is produced as shown below:

1) The first step is to prepare the data which consists of n observations on 2 × p variables. According to the situation of data, appropriate transformations performed on x_{ij}, y_{ij} $(i=1, 2, \cdots, n; j=1, 2, \cdots, p)$ such as

$$x'_{ij} = \frac{x_{ij} - \overline{x_j}}{S_{xj}} + b,\ y'_{ij} = \frac{y_{ij} - \overline{y_j}}{S_{yj}} + b$$

where, b is a positive constance, x_j, y_j is the sample mean of X_j, Y_j respectively and S_{xj}, S_{yj} is sample standard deviation of each variable.

In the case of examination marks at school, it is suitable to transform the data into deviation values as

$$x'_{ij} = 10\left(\frac{x_{ij} - \overline{x_j}}{S_{xj}}\right) + 50,\ y'_{ij} = 10\left(\frac{y_{ij} - \overline{y_j}}{S_{yj}}\right) + 50$$

2) The next step is to decide the kanji character. As 2 variables can be represented by a kanji, p types of kanji must be prepared in the case of 2 × p variables.

3) For step 3, the length and width of the kanji is assigned to the values of x_i, y_i. In observation i, length and width of the first kanji are transformed as

$$l\left(\frac{x'_{i1}}{b}\right),\ l\left(\frac{y'_{i1}}{b}\right)$$

or

$$l\left(\frac{x'_{i1}}{50}\right),\ l\left(\frac{y'_{i1}}{50}\right)$$

respectively, where *l* is a length of a kanji.

Similarly, the p-th kanji is drawn by the values of x'_{ip}, y'_{ip}.

4) For the final step, characters of kanji are drawn and arranged from left to right.

3. Example

Table 1 shows information, pertaining to the restaurant business, from 47 prefectural areas of Japan in 1982. (Japan Almanac, 1985). This information is in 5 restaurant type categories.

- Simple Restaurants (食堂)
- Japanese noodle shops (そば・うどん)
- Susi shops (すし屋)
- Japanese style restaurants (料亭)
- Coffee shops (喫茶店)

The table shows annual sales figures and the number of establishments for each category making a total of 10 variables of data. Since the value of each variable is influenced by prefectural populations, a standardized procedure is applied. Firstly for each category the total annual sales is divided by the total number of establishments in that category. Further the total number of establishments in each category is divided by the respective prefectural population to give an establishments per 1000 population figure. Thus the number of sales per establishment type and the number of establishments (in each category) per 1000 population in each prefecture is found.

These 10 adjusted variables concerning restaurants and eateries are used as raw data.

Table 1. Restaurant data of Japan in 1982 (Japan Almanac,1985)

	食堂・レストラン		そば・うどん		すし屋		料 亭		喫茶店	
	年間販売額	店数	年間販売額	店数	年間販売額	店数	年間販売額	店数	年間販売額	店数
北海道	153.941	10032	20.190	1383	54.986	2525	5.411	137	59.743	6935
青森	36.009	2941	3.114	459	8.068	582	3.580	73	10.408	1650
岩手	35.462	2671	3.211	285	7.784	451	3.477	94	7.115	982
宮城	61.106	3495	9.427	858	17.118	933	3.485	91	15.259	1478
秋田	30.756	2116	2.341	266	6.236	448	7.960	197	5.999	777
山形	29.508	1982	5.109	591	7.818	432	4.322	114	6.508	903
福島	57.932	4188	3.879	358	11.965	683	4.789	139	12.542	1526
茨城	78.708	5227	10.927	983	16.954	1074	9.530	262	16.187	1783
栃木	63.899	4416	8.001	844	14.283	824	6.138	170	13.685	1681
群馬	61.767	4232	11.247	992	15.403	862	3.933	109	11.716	1612
埼玉	161.037	8895	37.212	2642	43.755	2428	9.264	230	38.345	3210
千葉	171.460	8919	26.941	1798	43.394	2170	6.961	196	37.947	3249
東京	997.834	34952	154.645	7053	209.694	7956	33.379	629	341.865	20216
神奈川	316.727	13006	46.087	2350	67.499	2951	8.243	153	75.935	5600
新潟	72.887	4324	5.645	438	17.378	890	23.810	483	17.206	1731
富山	26.865	1446	4.430	431	7.888	422	4.611	175	10.435	1335
石川	36.824	1855	5.897	457	10.322	557	7.601	197	17.720	2071
福井	22.382	1227	3.522	287	6.266	318	3.521	130	9.513	1306
山梨	28.151	2359	2.378	252	7.770	475	1.585	61	6.007	883
長野	76.102	4863	7.151	570	14.787	735	5.735	165	15.283	1911
岐阜	64.498	3766	6.811	603	14.847	773	7.032	222	39.187	3751
静岡	125.736	7420	15.007	1107	29.108	1734	8.782	185	39.918	4711
愛知	257.870	10844	43.725	2767	60.476	3132	13.421	243	170.236	14730
三重	49.056	2683	5.420	470	9.464	559	4.656	112	21.306	2745
滋賀	34.883	1447	2.795	211	5.541	250	5.603	121	9.004	1001
京都	132.095	5118	14.582	1044	20.326	963	16.851	562	51.514	5046
大阪	395.369	17519	59.345	3370	95.017	4446	21.785	619	229.860	23124
兵庫	184.827	8260	26.178	1621	39.492	2090	9.069	229	105.248	10033
奈良	30.286	1433	2.490	208	8.382	385	1.641	53	10.627	1166
和歌山	32.725	1933	2.754	316	6.129	437	1.683	128	17.457	2052
鳥取	19.376	1172	0.930	104	1.780	121	1.002	45	8.973	1064
島根	20.753	1375	1.634	129	3.113	187	3.968	114	7.003	742
岡山	46.132	3095	6.808	660	10.485	501	2.615	88	23.712	3071
広島	81.991	4833	8.053	643	14.879	741	5.717	108	44.775	4259
山口	43.880	3069	3.447	283	6.673	346	5.302	164	14.199	1773
徳島	20.367	1609	2.629	330	2.990	234	2.015	67	10.252	1220
香川	28.772	1582	6.512	645	3.357	200	5.045	104	14.871	1949
愛媛	36.138	2724	4.921	545	6.981	471	3.122	132	19.288	2671
高知	20.133	1574	1.211	200	2.139	138	3.045	63	15.719	2563
福岡	171.605	10102	19.215	1373	30.170	1667	17.567	320	44.068	4946
佐賀	22.344	1725	2.119	184	4.051	235	4.347	115	5.224	661
長崎	50.262	3213	2.648	216	8.045	599	3.681	85	9.520	1359
熊本	52.609	3823	3.043	271	9.912	492	6.305	142	11.041	1333
大分	33.282	2520	2.114	223	4.738	302	5.560	148	8.796	1217
宮崎	28.045	2626	3.217	303	6.288	359	3.142	81	8.112	1207
鹿児島	44.207	3418	2.875	325	7.904	483	2.857	86	11.434	1443
沖縄	33.952	2820	0.781	94	2.875	264	1.406	56	14.124	1420

Example shows the kanji graph of data from the restaurant business.

Kanji graph was used as a means of representing data from the Japanese restaurant business. Ten variables were expressed by use of 5 characters of kanji each character being a composite of two variables. The width and height of each character was compiled as follows.

- The width of '食' corresponds to sales (per establishments) of restaurants ('食堂') in the respective prefectural area.

- The height of '食' corresponds to the number of '食堂' per 1000 population in the respective area.
- The width of '麺' corresponds to sales (per establishment) of Japanese noodle shop ('そば・うどん') in the respective prefectural area.
- The height of '麺' corresponds to the number of 'そば・うどん' shops per 1000 population in the respective area.
- The width of '寿' corresponds to sales (per establishment) of Susi shop ('すし屋') in the respective prefectural area.
- The height of '寿' corresponds to the number of 'すし屋' per 1000 population in the respective area.
- The width of '亭' corresponds to the sales (per establishment) of Japanese style restaurant ('料亭') in the respective prefectural area.
- The height of '亭' corresponds to the number of '料亭' per 1000 population in the respective area.
- The width of '茶' corresponds to sales (per establishment) of coffee shop ('喫茶店') in the respective prefectural area.
- The height of '茶' corresponds to the number of '喫茶店' per 1000 population in the respective area.

The transformation of each variable into a kanji character is with respect to formula (1) mentioned above. In this case b=3. The width and height of 5 characters are decided based on formula (3). For each prefectural area 5 separate characters display the data.

Figure 1 shows the kanji graph for the restaurant business data in the 47 prefectural areas. The results indicate that the four urban areas of Tokyo, Kanagawa, Aichi and Osaka each have large sales figures in all categories of the restaurant business. However these four areas have a relatively small '亭' character indicating few Japanese style restaurants. Kyoto,on the other hand has a large '亭' character representing relatively abundant numbers of '料亭'. Further, Yamanashi shows a different pattern to other areas in that the numbers of '食堂' and 'すし屋' are abnormally high.

Additionally as pairs, further matching patterns can be observed. Aichi and Osaka, Fukui and Toyama, Gunma and Tochigi, Miyagi and Saitama all show similar characteristics as kanji graphs. A final feature is that while Fukuoka, Miyazaki and Kagoshima all show comparatively large numbers of '食堂', they also exhibit almost identical patterns in the others 4 categories of each establishment type. In this way, it is possible to treat the kanji graph in the same manner as a bar graph in interpreting the data.

Below, I attempted to compare and contrast techniques (1), (2) and (6) from the view of application to clustering.

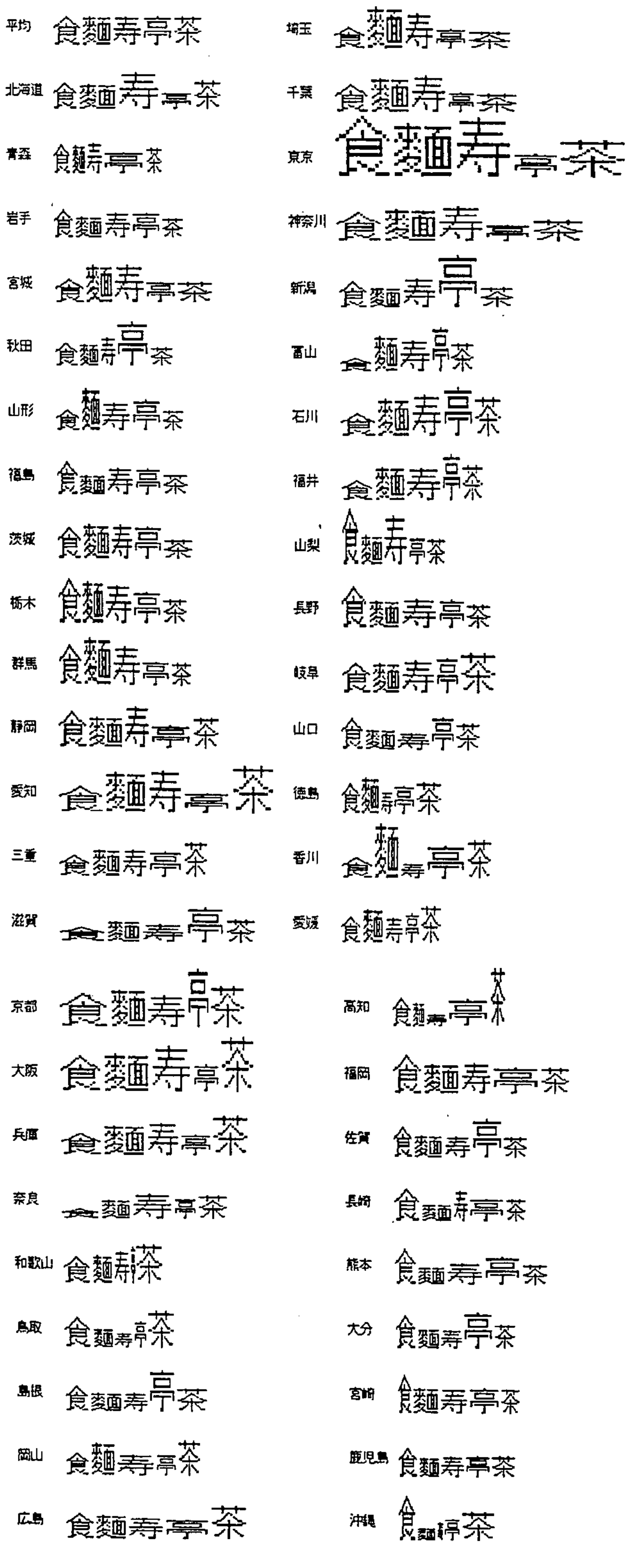

Fig. 1　Kanji graph in the 47 prefectural areas.

3.1 Comparison with hierarchical analysis

In this section, Kanji graph is compared with dendrogram. Figure 2 shows the dendrogram by group average method for the restaurant business data showed in table 1. This figure shows that Kochi, Osaka, Aichi, Yamanashi and Tokyo are somewhat diagramatically separated from the rest of the prefectures. Also it suggests that I can classify these other 42 prefectural areas into 4 groups.

Further more, a cluster of 12 prefectures is also apparent in the dendrogram (Kyoto, Kagawa, Tottori, Hiroshima, Okayama, Mie, Fukui, Ehime, Wakayama, Gifu, Hyogo and Ishikawa). The kanji graphs for each of these 12 prefectures were compared with the dendrogram. The results shows that a different pattern was established in Kyoto, Kagawa and Tottori as compared to the remaining 9 areas. Word analysis can be seen to be essential as a means of evaluation and comparison but by itself the dendrogram is not always simple to interpret nor easy to understand. The kanji graph on the other hand illustrates clearly the differences in size and value between variables. Cluster analysis is really only a useful way to confirm this.

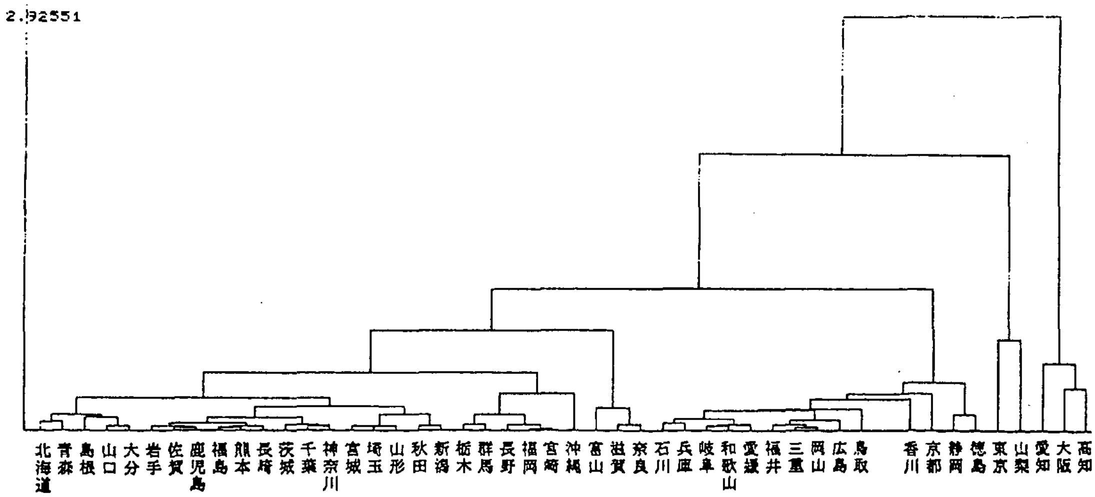

Fig. 2 Dendrogram (Group average method).

3.2 Comparison with graphical method (face graph)

I compared the standardized data with method (6) mentioned in the introduction. The data for the 12 prefectural areas refer to in table 1 were applied to the face graph. The results being shown in figure 3.

From figure 3 it is understood that there are quite large differences between the prefectures of Ehime,Wakayama, Gifu, Hyogo and Ishikawa in the face graph despite being in a cluster according to the dendrogram method. The face graph unlike the dendrogram allows us to see, at a glance, differences in the variables. Similarly, principal component analysis cannot facilitate such observations. In this sense the face graph is a visually straight forward classification.

However, the interpretation of the value of the variable is far from simple. This is because of the problem of which variable should be assigned to which facial feature is inherent in the representation. Altering variable assignment will accordingly alter the facial expression of the face graph and thus change the interpretation of results. Furthermore, problems arise since the face graph cannot cope with more than 18 variables at once.

The kanji graph and face graph can be analyzed by looking at it in it's entirely or in a holistic way. However, the kanji graph has the advantages of being able to express variable size and value simply, not affecting the application to clustering even if variable order is altered and is able to express an unlimited number of variables.

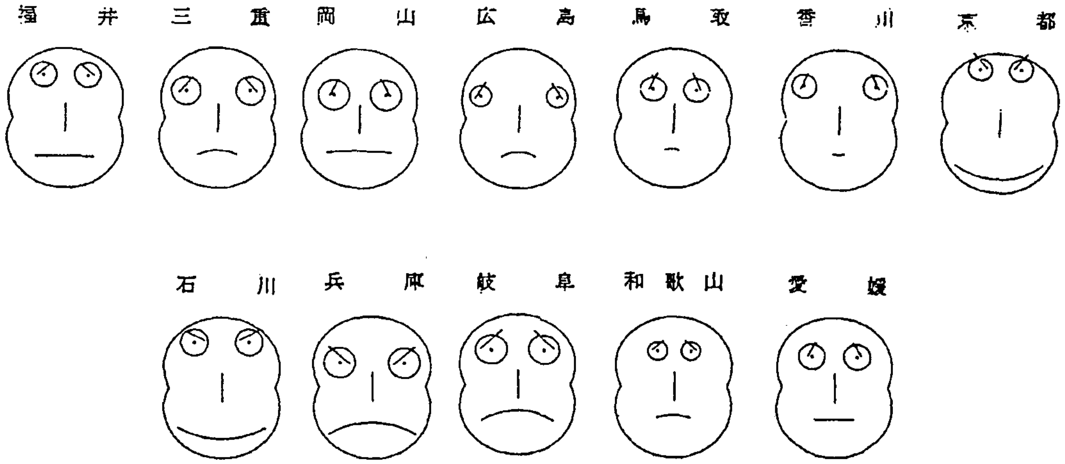

Fig. 3 Face graph for restaurant business data.

3.3 Comparison with principal component analysis

The standardized restaurant business data was further applied to principal component analysis. As table 2 shows the data was taken to the third principal component and produced a plot as shown in figure 4.

Table 2 Principal Component Loadings

Items	Component 1	Component 2	Component 3
Sales of '食堂'	0.429	−0.334	−0.039
The number of '食堂'	0.084	0.609	0.389
Sales of 'そば・うどん'	0.457	0.021	−0.159
The number of 'そば・うどん'	0.276	−0.047	0.453
Sales of 'すし屋'	0.392	−0.151	−0.216
The number of 'すし屋'	0.346	0.257	0.375
Sales of '料亭'	0.233	0.169	−0.428
The number of '料亭'	−0.081	−0.519	0.248
Sales of '喫茶店'	0.424	0.055	−0.094
The number of '喫茶店'	0.110	−0.353	0.421
Eigenvalue	4.037	1.429	1.273
Contribution ratio	0.404	0.143	0.127

The first principal component was (for overall sales) the separation of those prefectures with large sales figures and those with small sales figures. The second principal component split those prefectures with many '食堂' and 'すし屋' but low sales and those with many '料亭' and '喫茶店' but low sales. The third principal component separated prefectures with numerous shops but low sales and those with few shops but high sales figures. Figure 4 shows the principal component plot for the first and second components, having been connected up by MST.

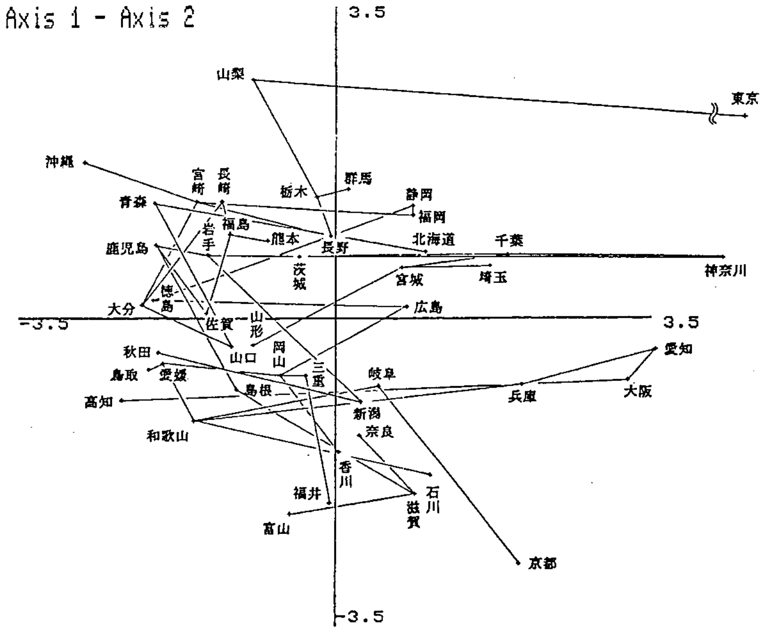

Fig. 4 PCA plot with MST

Looking at figure 4, I can see and overall 'U' shape however from the MST connect up I can establish that Hiroshima, Kochi, Ishikawa, Kanagawa and Akita are in rather unlikely positions. Following from this, analysis of Akita and Niigata for example in figure 1 show characters of differing sizes but an overall pattern can be understood. Principal component analysis not always a suitable method to decipher such patterns since the influence of the first component on the other variables can be too heavy. Thus I can say that the kanji graph (figure 1) is superior at displaying detail in the data and for observing patterns than principal component analysis.

4. Concluding remarks

This paper propose the kanji graph and discuss the effectiveness of the kanji graph. The character of kanji have a meaning. So, kanji graph is easy to grasp the feature of the data by assigning reasonable kanji to the data.

Having compared the kanji graph with a variety of other techniques I can see that it has several important qualities and advantages. Not only does the kanji graph enable simple interpretation of the data but also allows many variables to be expressed. Additionally the size of the expressed variable is rapidly understood, clustering can be visually verified with ease and since it can be reproduced on a personal computer, it's practical application to statistical analysis is clearly potentially extensive. The kanji graph may for these reasons be thus considered 'user friendly' and 'Japanese original' since the Japanese observer can soon feel familiar with the data representation.

References

〔1〕 Andrews, D. F. (1972). Plots of high-dimensional data. Biometrics, 28, 125-136.

〔2〕 Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. J. Amer. Statist. Assoc., 68, 361-368.

〔3〕 Fukumori, M. and Tanaka, Y. (1994). Evaluation of multivariate graphs from the aspect of human recognition. Bulletin of The Computational Statistics of Japan, 7, 1, 37-45.

〔4〕 Fukumori, M. (1995). Evaluation of Multivariate Graphs. Journal of Chugoku Junior College., 26, 53-66.

〔5〕 Hirai, Y., Fukumori, M. and Wakimoto, K. (1988). Kanji graph representation for multivariate data and its application to cluster analysis. Bulletin of The Computational Statistics, 1, 1, 11-21.

〔6〕 Wakimoto, K. and Taguri, M. (1978). Constellation graphical method for representing multidimensional data. Ann. Inst. Statist. Math., 30, Part A, 77-84.