Kansai University Review of Economics
No. 4 (March 2002), pp. 57–69

Hierarchical Structure in Food Consumption

Takaharu Araki
Noriko Hashimoto

Summary

In recent years, a group of physicists have started to research financial systems by using tools and methodologies which are specific to physics. Mantegna (1998) proposed a new method detecting the existence of the economic information stored in the time series of stock prices. He used correlation coefficient to evaluate the distance between stocks and reveal the spatial structure of stocks traded in a financial market.

In this paper, we apply his method to consumption data and explain a hierarchical structure of food consumption.

keywords : correlation coefficient, food consumption, hierarchical structure, minimal spanning tree (MST), dendrogram, cluster analysis

1 Introduction

Mantegna (1998) proposed a new method to detect the existence of the economic information stored in the time series of stock prices. He used correlation coefficient to evaluate the distance between stocks and reveal the hierarchical and spatial structure of stocks traded in a financial market. His idea is simple and adopted tools and methods are well known in data analysis and graph theory, but he obtained substantial results from applying his method to data in different field.

In this paper, we apply his method to food consumption data and disclose the hierarchical structure of food consumption by calculating correlation coefficient and evaluating distance between food items. In other words, we can show latent spatial locations among foods' classification.

This approach is quite novel and different from traditional methods in the study of consumption analysis. We can expect to extract the information which cannot be seen by the conventional techniques. This method can be considered as a complementary technique for the conventional methods.

2 Method of Mantegna

Mantegna (1998) found a topological arrangement of stocks traded in a financial market by detecting the economic information stored in the time series of stock prices. He used the empirical correlation coefficients computed between all pairs of stocks in a portfolio and obtained a hierarchical tree called dendrogram and a minimal spanning tree (MST) by considering the synchronous time evolution of the difference of the logarithm of daily stock prices. The dendrogram and the MST provide useful information to investigate the nature of the common economic factors affecting the time evolution of logarithm of stock prices.

Mantegna (1998) reported the results obtained by researching the portfolio of the stocks used to compute the Dow Jones industrial average index and the portfolio of the stocks used to compute the Standard and Poor's 500 index in the time period from July 1989 to October 1995. He used following methods to detect the hierarchical structure which is latent in changes of stock prices and disclosed the spatial locations among stocks.

For scales to measure the relation of the price change between two stocks i and j, he selected the Pearson's correlation coefficient r_{ij} defined in Eq. (1)

$$r_{ij}=\frac{<Y_iY_j>-<Y_i><Y_j>}{\sqrt{(<Y_i^2>-<Y_i>^2)(<Y_j^2>-<Y_j>^2)}} \qquad (1)$$

where, $Y_i=\ln(P_i(\mathrm{t}))-\ln(P_i(t-1)=\ln(P_i(t)/P_i(t-1))$, $P_i(t)$ denotes the closing price of stock i on day t, and $<\cdots>$ denotes a time average over the period studied. For instance, $<x>=\sum_{i=1}^{T} x_i/T$ where $x=\{x_1, x_2, \ldots, x_T\}$.

Pearson's correlation coefficient r_{ij} measures the degree of the similarity on the pattern of the change between time changes of two stock prices i and j. By definition, r_{ij} takes the value between -1 (perfect negative linear relationship) and 1 (perfect positive linear relationship). A value r_{ij} close to 0 indicates the lack of linear association between two stocks.

Calculating correlation coefficients of all pairs of stocks, Mantegna got the $n\times n$ matrix of correlation coefficients for daily logarithm price differences. As $r_{ij}=r_{ji}$, the correlation coefficient matrix is symmetric and all the diagonal elements (r_{ii}'s) are 1, the matrix is characterized by $n(n-1)/2$ correlation coefficients which are supposed to carry information about change of two stocks.

Next, he adopts the technique of minimal spanning tree (MST) of graph theory and the cluster analysis of statistics. Therefore, it is necessary to define the distance or the degree of dissimilarity between stocks. However, the correlation coefficient itself cannot be used as a distance between the two stocks because it does not fulfill the three axioms that define an Euclidean distance: (i) $d(i,j)=0$ if and only if $i=j$; (ii) $d(i,j)= d(j,i)$ and (iii) $d(i,j) \leqq d(i,k)+d(k,j)$.

So, Mantegna proposed a conversion of correlation coefficient r_{ij} with certain function f to make a generalized metric $d(i, j) = f(r_{ij})$ which can be defined as distance. For example, Mantegna (1998) used conversion

$$d(i,j)= \sqrt{2(1-r_{ij})} \tag{2}$$

which satisfies three axioms of distance. This definition is obtainable from the Pythagorean relation of the normalized Y_i and Y_j (Montegra and Stanley (2000)). (Mantegna (1999) used another definition $d(i,j)= 1-r_{ij}^2$.) Using this distance between two stocks, we can construct a distance matrix $D=[d(i,j)]$. This matrix D is used to produce the minimal spanning tree (MST).

MST is the technique to choose the nodes (stocks, in the analysis of Mantegna) related most to each other in a set and provides the spatial cluster. The strongest relationship means the minimum total distance between items. With the assistance of MST, we can characterize the spatial distribution patterns of stocks from the distance matrix.

Mantegna (1998, 1999) and Bonanno *et al*. (2000) applied the above method to stock price and obtained results of great interest from an economic point of view.

We will adopt above methods to consumption data study and show spatial structure of food items by using only the correlation information of time series data. Moreover, we will display their spatial locations hierarchically by drawing the dendrogram.

3 Hierarchical structure of food consumption

3.1 Data

We use expenditure data on food of "family-income and expenditure survey" by Statistics Bureau, Ministry of Public Management, Home Affairs, Posts and Telecommunications in this analysis. In this survey, food is classified into following 12 items.

1.1 Cereals
1.2 Fish and shellfish
1.3 Meat
1.4 Daily products and eggs
1.5 Vegetables and seaweeds
1.6 Fruits
1.7 Oils, Fats and seasonings
1.8 Cakes and candies
1.9 Cooked food
1.10 Beverages
1.11 Alcoholic beverages
1.12 Eating out

We use the monthly data of workers' households in all Japan. The total number of months is 376 in the time period from January 1970 to April 2001.

3.2 Correlation

Let $x_i(t)$ denote expenditure of item i $(i=1,\ldots,12)$ on the tth month $(t=1,\ldots,376)$. Calculating $Y_i=\ln(x_i(t))-\ln(x_i(t-1))$ for all i on all months, we obtain 375 Y_i's. Fig.1 shows the sequence plot of each item. It is very difficult to see the relation between 12 items from this graph. The big spike observed in every year-end (from November to December) is caused by the year-end food expense growth.

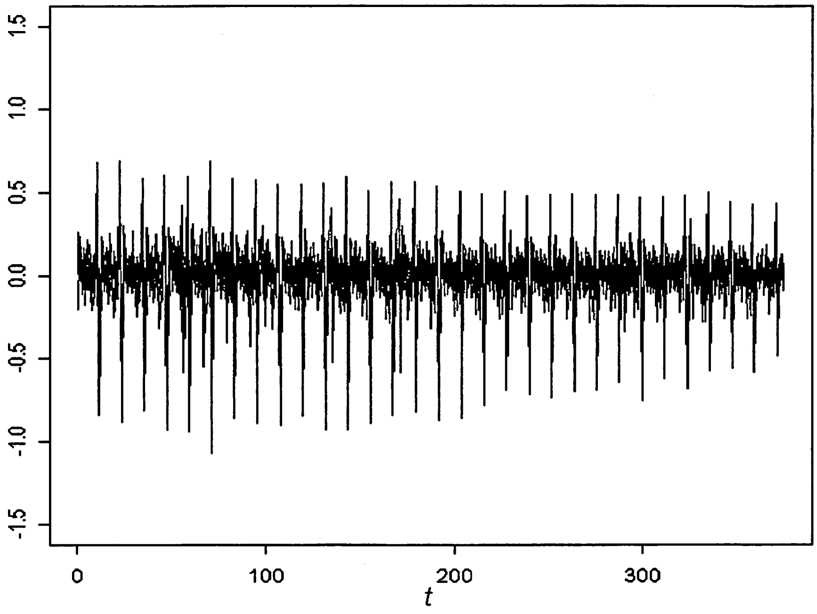


Figure 1 Sequence plot

Fig.2 shows histograms of all food items overlaid with estimated density using a normal kernel. We can see some outliers but it turns out that the distribution of Y_i's is almost symmetric and close to a normal distribution.

Substituting rates of change between the food items (Y_i's) in Eq.(1), we compute correlation coefficients r_{ij} and construct the correlation matrix $R=[r_{ij}]$. The calculated correlation coefficients and scatter diagrams are shown in the pairs plot (Fig.3).

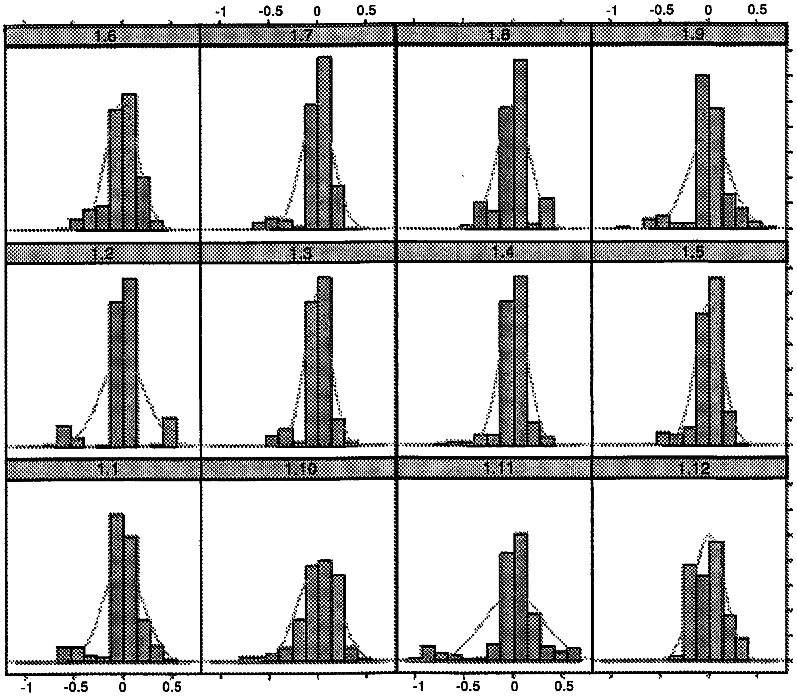


Figure 2　Histograms of log increasing rates for food items

In Fig.3, the label (item number) of food items are shown in the diagonal boxes, the correlation coefficients are shown in the lower left triangle and scatter diagrams are in the upper right triangle.

From scatter diagrams, we can detect a possibility for stratification. Especially, the graph of fish and shellfish (1.2) indicates strong need of stratification. Some items show the nonlinear relationships. Although most items show high correlations, eating out (1.12) does not show much relation with other items.

We show the histogram of the 66 correlation coefficients in Fig.4 ($n(n-1)/2=66, n=12$). Curved line shows the estimated density curve based on a normal kernel. Histogram and estimated density are negatively skewed. Most correlations are strong and their coefficients are over 0.7.

Representing the correlation strength, gray scale table can be shown in Fig. 5. The gray scale denotes correlation strength and deeper color indicates smaller correlation. Namely, distance decreases from very dark gray to light gray. From this figure, we can confirm overall strong relationship among food items and there seems to be overall relation between items.

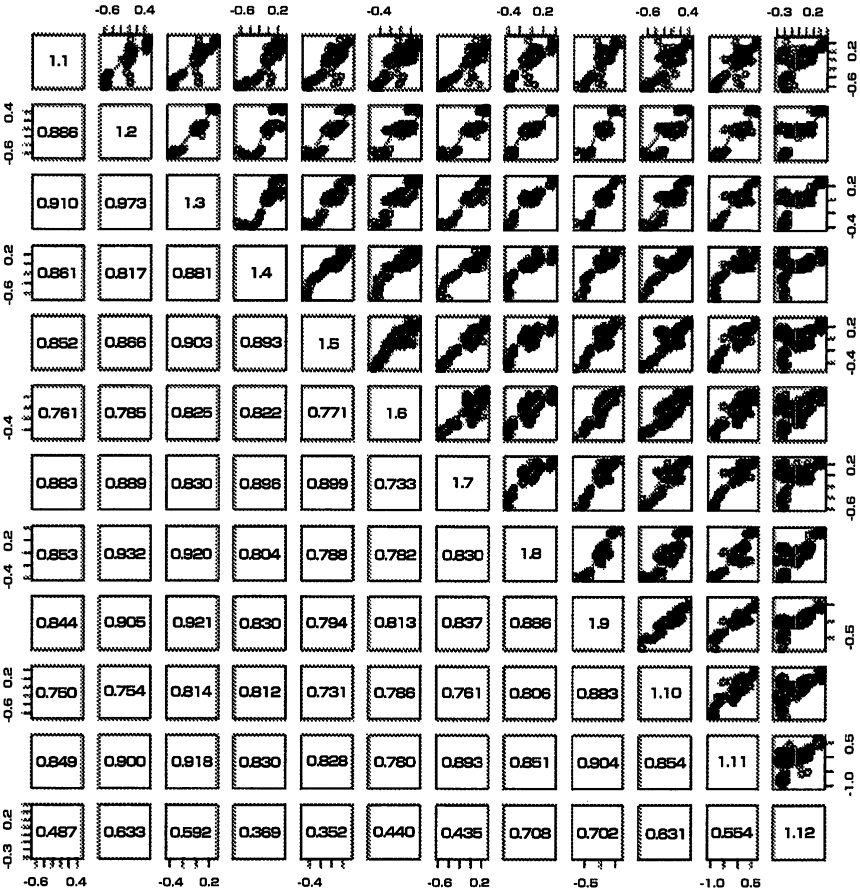


Figure 3 Pairs plot

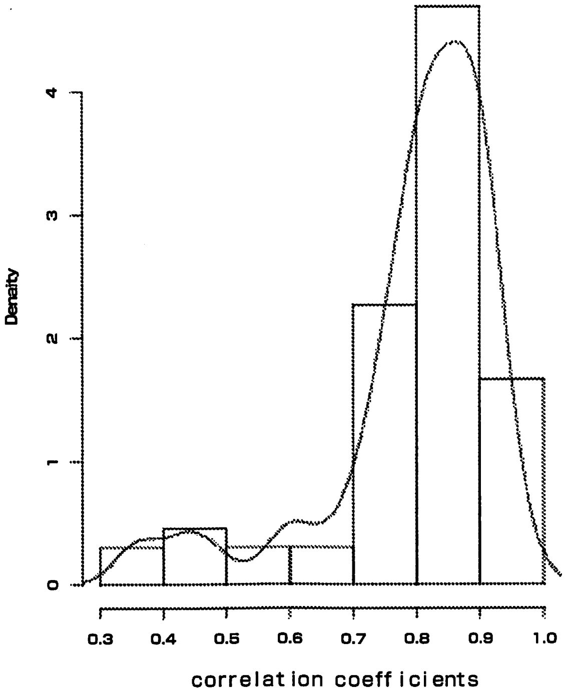


Figure 4 Histogram and estimated density of correlation coefficients

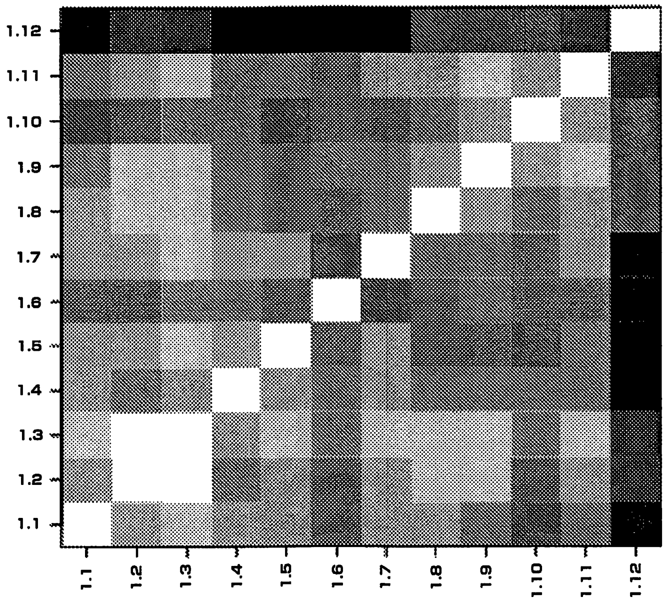


Figure 5 Gray scale table of the distance matrix of correlation coefficients

3.3 Minimum Spanning Tree

Using the distance formula $d(i,j)=\sqrt{2(1-r_{ij})}$, we calculate the distance (dissimilarity) matrix D and produce MST. When the distance between n nodes are defined, the MST connects all these nodes with n-1 lines so that the total of distance becomes the minimum.

We show the MST in Fig.6. Each node shows the food item labeled by family survey item number. Numerical numbers labeling the line denotes the distance $d(i, j)$ between items.

Generally speaking, the items directly connected by a straight line can be considered to have strong relationship. Fig. 6 shows that the nodes about 12 food items constitute one cluster and its central node is meat (1.3). Meat and fish and shellfish (1.2) have especially strong connection. Subsequently, oils, fats and seasonings (1.7) and cooked food (1.9) have strong relation with meat. Eating out (1.12) has the weakest relation with other items. This result looks reasonable since other items are consumed at home.

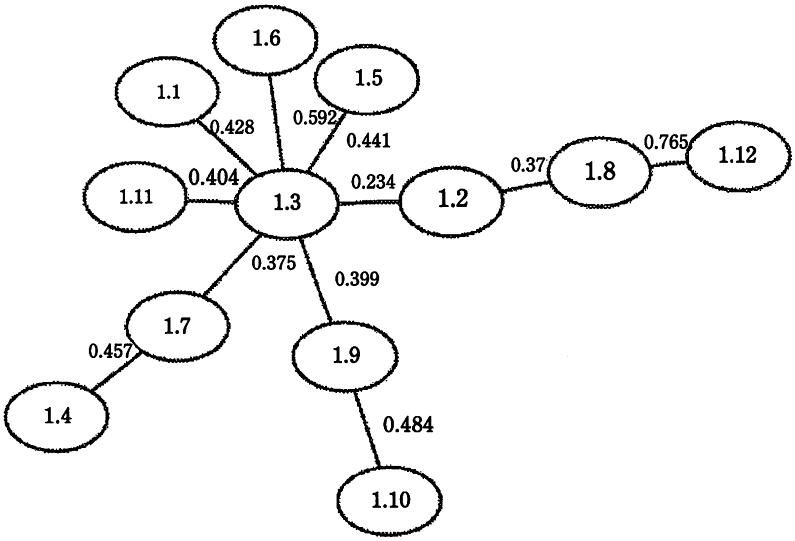


Figure 6 Minimal spanning tree (MST)

The dendrogram obtained by the single linkage method have close relationship with the MST. It gives us the information about the process of clustering. Fig.7 shows the dendrogram. The close relation between meat (1.2) and fish and shellfish (1.3) can be seen from this figure.

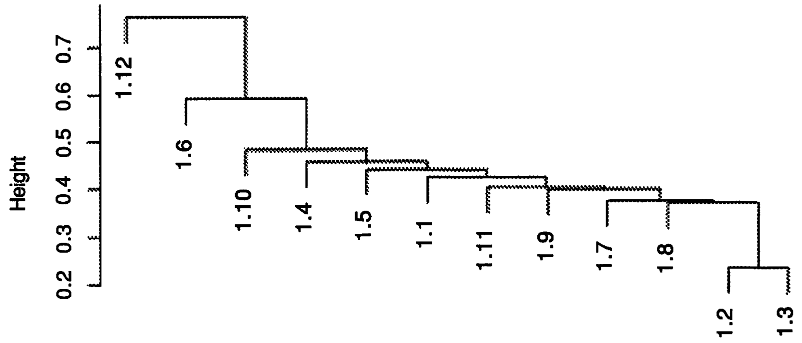


Figure 7 Dendrogram

4. Analysis of time evolution

The analysis in section 3 is based on a long period data about 30 years from 1970 to 2001. In this section, we study the time evolution of the hierarchical structure between consumption items.

We split the entire period into three 125-months periods: from February 1970 to June 1980 (period A), from July 1980 to October 1990 (period B) and from November 1990 to April 2001(period C). Roughly speaking, period A corresponds to the 1970s, period B does to the 1980s and period C does to the 1990s. We use these divided data to detect whether there is any change in consumption pattern with the times.

Fig.8 shows the estimated density curve based on the normal kernel of the elements of correlation in each period. The shapes are almost the same among periods. However, from the following statistics, it turns out that the averages and medians decreased and variances increased a little with the times. We can presume from this that the relationships between patterns of the change among food items become weak.

Period A	Mean: 0.813 ,	median: 0.848,	variance: 0.017
Period B	Mean: 0.809 ,	median: 0.851,	variance: 0.021
Period C	Mean: 0.781 ,	median: 0.808,	variance: 0.022

Next, MST and dendrogram are shown in Fig.9 and Fig.10. From Fig.9, two items, fish and shellfish (1.2) and meat (1.3) have significant relationship throughout the whole period. From the centrality notion of the graph theory, in period A (the 1970s), there is only one large cluster of meat (1.3). In period B, the cluster of cereals (1.1) and fish and shellfish (1.2) are joined to the central position. However, in period C (the 1990s and after), oils, fats and seasonings (1.7) departs from the cereals (1.1) to the next position of the meat (1.3).

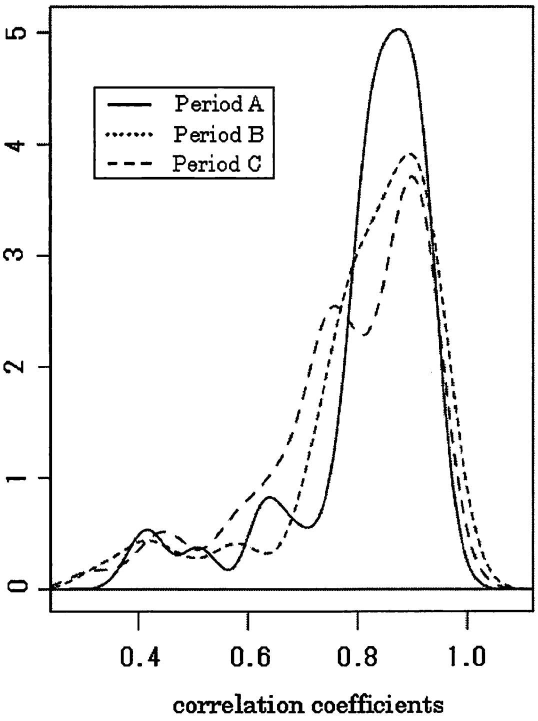
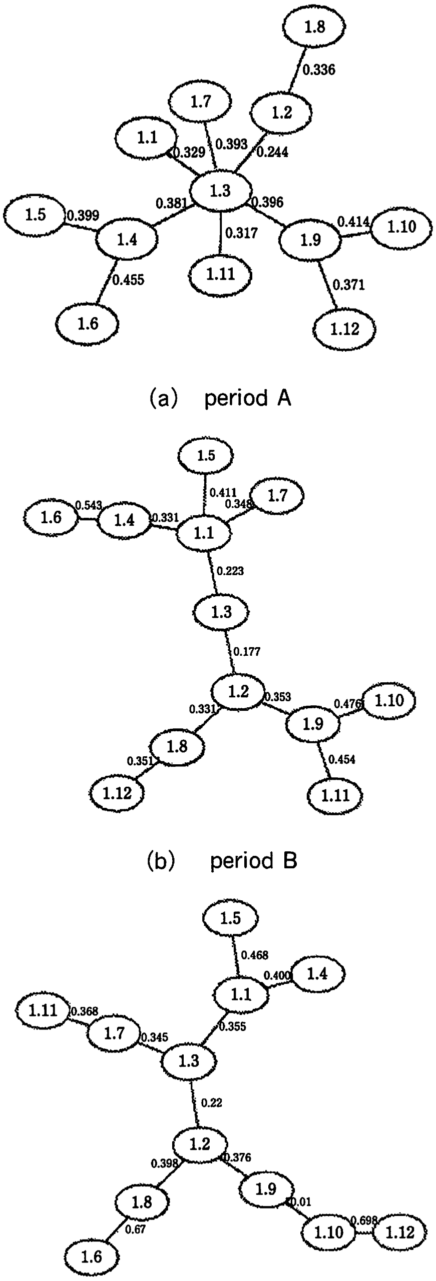


Figure 8 Estimated densities for each period



(c) period C

Figure 9 MST of each period

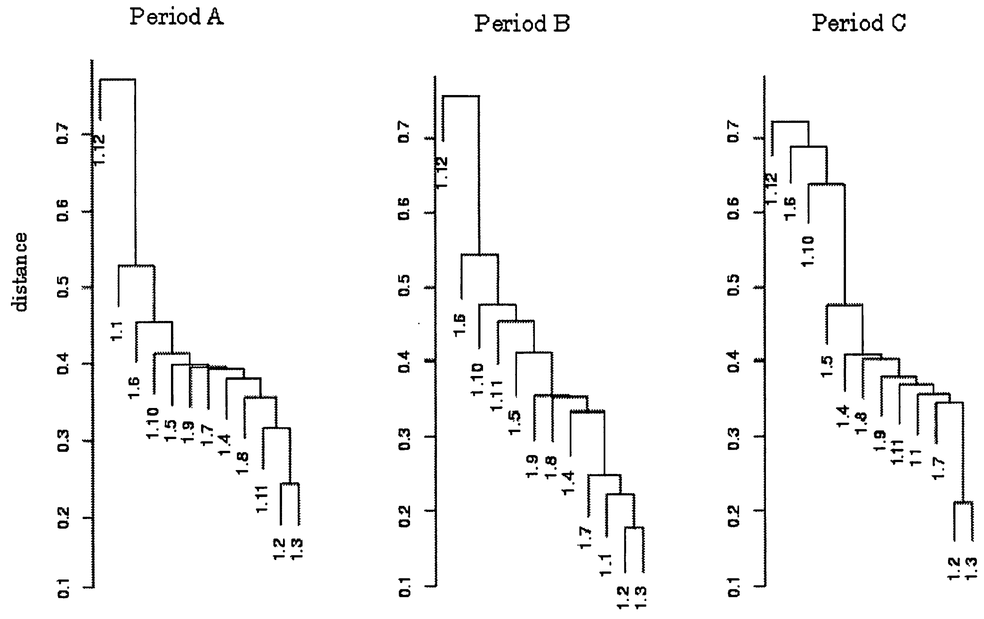


Figure 10 Dendrogram of each period

From Fig.10, eating out (1.12) has always the weakest linkage with other items. However, in period C, the diameter of the graph decreases and fruits (1. 6) and beverages (1.10) approach to the upper part. On the other hand, cooked food (1.9) located middle position of the graph throughout.

Takaharu Araki
(Professor of Business Statistics, Faculty of Commerce)
Noriko Hashimoto
(Professor of Econometrics, Faculty of Economics)

References

[1] Bonanno, G., N. Vandewalle and R. N. Mantegna (2000). Taxonomy of stock market indices. Physical Review E 62, R7615-R7618.

[2] Bouchaud, J-P, and M. Potters (2000). *Theory of financial risks - From statistical physics to risk management*. Cambridge University Press.

[3] Mantegna, R. N. (1998). Hierarchical structure in financial markets. European Physical

Journal B 11, 193-197.

[4] Mantegna, R. N. (1999). Information and hierarchical structure in financial markets. *Computer Physics Communications* 121-122, 153-156.

[5] Mantegna, R. N., and H. E. Stanley (2000). *An introduction to econophysics - Correlation and complexity in finance*. Cambridge University Press.

[6] Stephen C. North (1992). neato User's Guide.
http://www.research.att.com/sw/tools/graphviz/

[7] Struyf, A., Hubert, M. and Rousseeuw, P.J. (1997). Integrating robust clustering techniques in S-PLUS. *Computational Statistics* and Data Analysis 26, 17-37.

[8] Statistics Bureau, Ministry of Public Management, Home Affairs, Posts and Telecommunications. JAPAN. *Annual Report on the Family Income and Expenditure Survey*.

[9] Venables, W.N. and B.D. Ripley (1999). *Modern applied statistics with S-PLUS 3rd Ed*. Springer.