1. Data for experiment

Sample data for the presentation is from 1995 Korean Census data. The data consists of 12,958,181 ordinary households with 31 variables. The data was aggregated and averaged over 7,148 dongs (administrative units consisting of around 1,500 households). Out of these 7,148 dongs, we randomly selected 400. We also selected 7 variables for our study. They are average age of the household members (the variable name is "age"), information about the household head, and information about living quarters. Variables related to the household head are: percentage of those with education of junior college and above(coll.%); percentage of those in service-related occupations, such as banking, insurance, real estate and business service (service%). Variables related to living quarters are: percentage of those in apartments (apt.%), percentage of those in single-family homes(ownerH%), percentage of those in living quarters of a size less than approximately 46 square meters(14 pyong)(smallH%), percentage of those in living quarters of a size larger than approximately 230 square meters(69 pyong)(largeH%). Table 1 gives the mean value of these 7 variables for the selected sample.

2. Classical Methods

First approach for visualization is to draw a multivariate scatterdiagram. Figure 1 is the multivariate scatterdiagram of the census data. It is not easy to see the structure of the data from the figure. It only informs us of the relationships between the two pair-wise variables.
Second approach could be a dendrogram. Dendrogram for the data is given in Figure 2a. This is with Euclidean distance and single linkage method. Another dendrogram with Euclidean distance and complete linkage method is given in Figure 2b. It is well known that dendrogram is very sensitive to the outliers as can be observed from the plots.

Third method is Grand Tour. With Grand Tour, we can explore the structure of multi-dimensional data projected in a two dimensional window. Grand Tour, however, shows us only a series of projections of data points on a plane that interpolates a sequence of randomly generated planes. Hence with Grand Tour, we can observe only the projections of data points on the current interpolating plane. Also, when the sample size of the data set is small, Grand Tour may lose its power to explore data structure.

3. Tracking Grand Tour

Tracking Grand Tour (TGT) proposed by Huh and Kim (1999) relieves the deficiencies of Grand Tour. TGT enables viewers to watch the tracking of the touring process of each data point. It has been found to be also useful in exploring the structure of data of small size. A similar idea to TGT can be found in Buja et al. (1996, section 8).

Figure 3a gives the Grand Tour and Figure 3b shows the two snapshots of TGT for the Census data. The first snapshot strongly suggests the existence of data grouping, and the second snapshot suggests there are a few points that are behaving differently from the majority of the data points. This suggests that TGT can be used to visually classify the multi-dimensional data into an appropriate number of groups, and to visually identify multivariate outliers.

Since the plots suggest Census data consists of two groups, our process is to classify the data set into two groups. How should we classify the data set into two groups? And how should we verify whether the grouping is well done or not? A solution to this is by linking the dendrogram, scattergram, and TGT, and visually explore the inherent structure of the multidimensional data. For this, we may first group the data according to some clustering method. Then we divide the data set into these groups, and apply TGT for each divided group. If the grouping is well done, TGT will show no anomalies. If the grouping is ill-formed, we can visually detect the ill-formation.

Figure 4 shows the snapshot of TGT when the data set is divided into two groups by k-means clustering method, and TGT is applied to each group. The TGT representation shows that the grouping is appropriate for most of the data sets accept for a few outliers. We may use dendrogram for the initial grouping, and apply TGT to verify the result of dendrogram. We may interactively select a portion of data sets from the scatterdiagram, or from the dendrogram, or even from the Grand Tour, and visualize the structure of the selected data sets by applying TGT. Our presentation will show these various types of visualization methods.

All the above techniques for visualization are implemented with JAVA, and hence anyone with WWW access on the globe can use the system on the network freely.
Figure 1. Multivariate scatterdiagram for Census data.

Figure 2a. Dendrogram of Census data with Euclidean distance and single linkage method after dividing into two groups by the K-means method.

Figure 2b. Dendrogram of Census data with Euclidean distance and complete linkage method after dividing into two groups by the K-means method.
**REFERENCES**


**RÉSUMÉ**

Notre équipe a récemment développé un système pour visualiser la structure de multidimensionel data. Cet outil pour le système est JAVA, et dès maintenant le résultat est disponible sur “WWW browser” comme “applet”. Dans cette présentation, on introduira une nouvelle méthode de visualisation appelée TGT, proposée par deux des auteurs, et démontrera l’efficacité de SIVON.