A STATISTICAL LOOK AT FAMILY WELFARE DATA - APPLICATION OF SOME MULTIVARIATE TECHNIQUES

Selvanayagan Ganesalingam and Siva Ganesh,
Institute of Information Sciences and Technology, Massey University, New Zealand
Kuldeep Kumar, School of Information Technology, Bond University, Australia

In view of the rapid population growth during the last four decades and resulting economic problems and social tensions, the family welfare programmes are gaining momentum in developing countries. Governments of such countries take various measures to uplift the social structure and these measures include programmes such as family planning services, child welfare, immunisation and other health programmes. In this study we have considered statistical analysis of some family welfare data from India where the social structure is heavily linked with religion and ethnic diversity.

INTRODUCTION

The role of statistics is to summarise, to simplify and eventually to explain the behaviour of data. It is a fact that a picture is worth a thousand numbers. A graphical description is more easily assimilated and interpreted than a numerical one. The graphical display can assist in summarising a large mass of numerical data, simplifying the aspects of the data by appealing to our natural ability to absorb visual images and providing a global view of the information, thereby stimulating possible explanations. In this paper we have used two different multivariate techniques, namely, the *Cluster Analysis* and the *Canonical Discriminant Analysis* (see Gnanadesikan (1977) for details), in the context of exploring a family welfare programme data from India. The family welfare programmes include family planning services, child welfare immunisation and various other health programmes. To achieve the set goal of birth rate of 21 per thousand population and the death rate of nine per thousand population by year 2000 AD in India, is a challenging task. Family welfare programmes in operation in the country are to contribute significantly in the attainment of the above goals.

In this paper we examine the data collected on various socio-economic and demographic indicators in Uttar Pradash, the largest state in India. This state has a total population of 139 million and consists of 63 districts grouped into 5 regions (namely, Hill(H), Western(W), Bundelkhand(B), Central(C) and Eastern(E)). Data were collected from all districts. (Total population of India is 844 million as per 1991 census). Each district varies in size and with respect to the socio-economic and demographic variables.

Each district may require a different approach to family welfare programmes but the administration would be very difficult; for instance, we need 63 administrators which may cause substantial (financial) burden to the Government. This calls for an in-depth study to identify and critically evaluate the factors that have bearings on family welfare programmes.

On the one hand, we may examine the similarities (or dissimilarities) between the five regions, and on the other, we may look for general patterns among the districts to form clusters where some uniformity exists. For example, if there are 3 or 4 distinct clusters then we could study each cluster more closely and administrate accordingly. In this situation, districts in each cluster would have similar characteristics and therefore an administrator could be appointed for each cluster and we only need three or four of them. These few administrators could be trained individually at a high cost (usually abroad) to run the programmes different to each cluster, but the same programme administered to all the districts within a cluster. Although this paper discusses an example from India, the techniques can be used in any family welfare programme in general and their administration in particular. This type of study will be quite useful to the policy makers, administrators, and other functionaries associated with such family welfare programmes, in aiding the administrative structure set up of such programmes.

DESCRIPTION OF DATA

The data has been collected from the 1994 Annual Report of Uttar Pradesh Health Director's Office. It consists of the following socio-economic and demographic indicators collected from each of the 63 districts: X1: district population as per 1991 census; X2: decennial growth rate (1981-1991); X3: population density per square km; X4: gender ratio; X5: area of the district in square km; X6: crude birth rate; X7: total fertility rate; X8: infant mortality rate; X9: couple protection rate (CPR) using all methods (effectively as on March, 1994); X10: CPR using sterilisation; X11: CPR using intra uterus device (IUD); X12: CPR using conventional contraceptives; X13: CPR using oral pills; X14: female age at marriage; X15: percentage of Muslim population; X16: percentage of scheduled caste and scheduled tribe population; X17; percentage of urban population to total population; X18: total crude literacy rate; X19: total male literacy rate; X20: total female literacy rate; X21: per capita income at current price; X22: number of schools per 100,000 population; X23: female work participation; X24: percentage of village with electricity; X25: percentage of villages with drinking water; X26: length of concrete road; X27: number of medical hospitals per 100,000 population.

These variables contribute substantially towards family welfare programmes. For example, Kumar and Srivastava (1988) studied the profiles of the acceptors of family planning programmes and observed that in the context of India, religion (including caste), literacy and per capita income play a significant role. Kumar and Sahai (1993) applied discriminant analysis on the Indian family planning data and noticed that profiles of acceptors using sterilisation, IUD, conventional contraceptives and oral pills were completely different.

Missing values in the data were replaced by the mean values of the respective variables. Indicator X9 was omitted from the analysis as it is the sum of X10 - X13.

STATISTICAL ANALYSIS

Since the data were collected from five *known* regions (H, W, B, C and E), a preliminary interest was focused on the variables (indicators) or linear combinations of variables that distinguish the five known regions adequately. A canonical discriminant analysis to maximally separate these known groupings of districts, not only revealed that the 1st two canonical dimensions account for at least 89% of the variation among the regions, but also that most of the 70% of the separation along the 1st dimension can be attributed to the distinction of the Hill region from the others (see Figure 3.1). The corresponding discriminant functions indicated that the 1st dimension produces large canonical scores with large X6, X12, X22 and X27, and small X3 values. This indicates that the Hill region is more prosperous with better medical and school facilities and awareness of conventional contraceptive methods and the consequences.

These findings prompted us to see whether there exists another natural grouping of districts. If the new grouping is different from the grouping of the five regions then the administration may need to be changed accordingly. To identify another natural grouping, we performed a cluster analysis using *Ward's minimum variance* approach. Here, we treated all 63 districts as an unclassified set and performed the clustering technique (with *Euclidean distance* criterion, based on the 26 indicators). The resulting *dendrogram* is

Figure 3.1. Plot of canonical discriminant scores of 63 districts identified by the region they belong to (H, W, etc.)

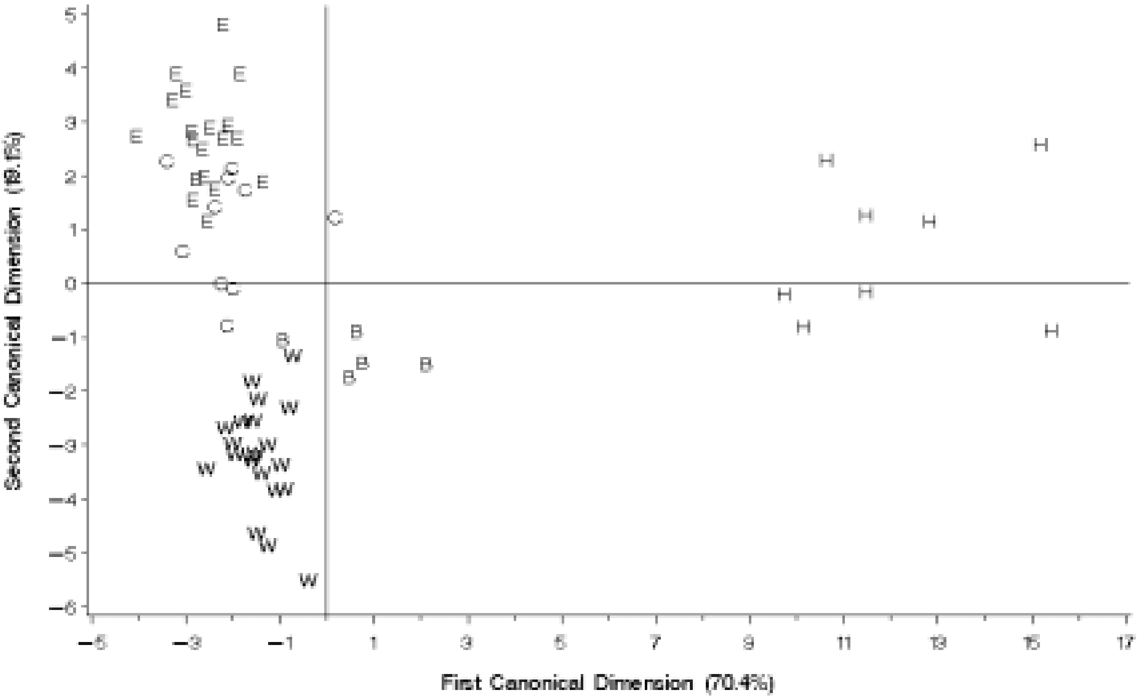
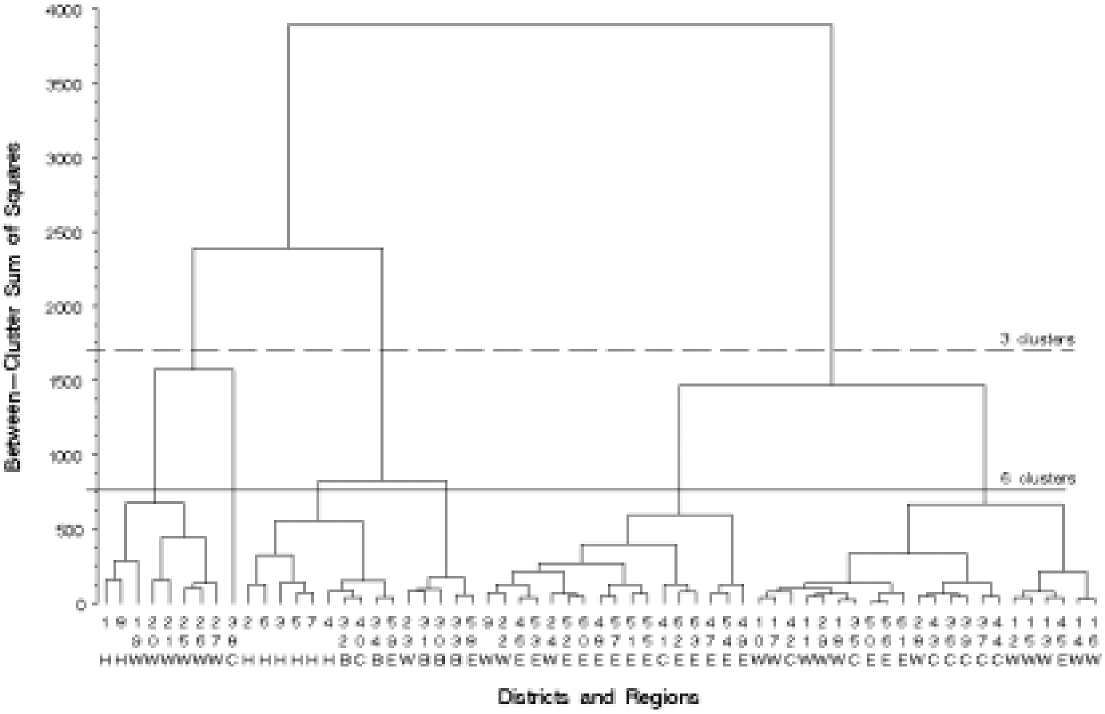


Figure 3.2. Dendrogram of 63 districts, identified by the district number and the region they belong to (H, W, etc.).



given in Figure 3.2 which shows the existence of five prominent clusters and a cluster with district #38 (Kanpur U) only. The horizontal solid line drawn across the dendrogram identifies these clusters. The dendrogram can also be used for determining a smaller number of clusters, say, three, by cutting the tree by the dotted line as shown in Figure

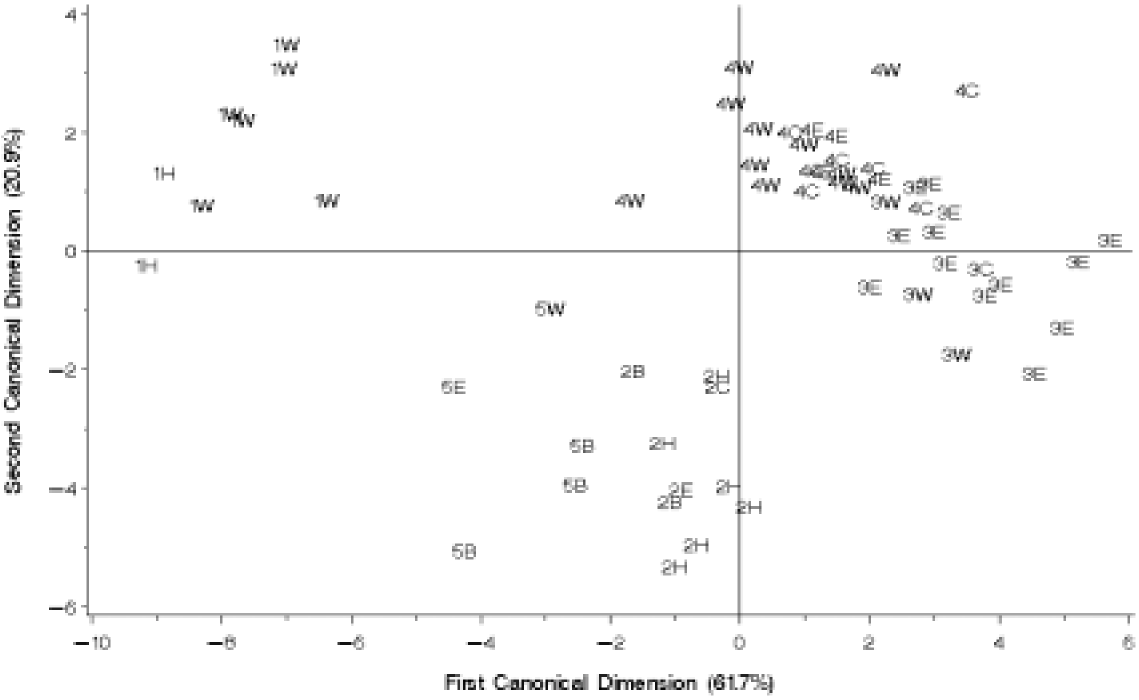
3.2. An obvious disadvantage here is that the sizes of clusters may become large, which in turn may impose administrative problems.

It is evident from the dendrogram that, though the data were obtained from five separate regions, there is a substantial overlap among the districts of the regions. For example, in the 5 (or 6) cluster situation, the districts 32 and 34 from Bundelkhand, district 40 from Central and district 59 from Eastern regions, all fall in a cluster which mostly consists of the districts from the Hill region. As noted earlier, district 38 (Kanpur-U) from the Central region has been singled out. A closer look at the various indicators revealed a marked difference between this district and others. A notable difference is the population density of Kanpur U (2390 per sq.km), which is almost five times that of the other districts in the same Central region. It is very clear that there are dissimilarities among the districts within each region, and consequently these clusters demand different approaches of policy implementation.

This lead us to see which variables/indicators are responsible for this distribution of district grouping. This question is addressed below for the case of the five clusters obtained (ignoring the district Kanpur U) by means of a canonical discriminant analysis to maximally separate the clusters. The 1st two canonical discriminant functions account for about 83% (61.7% and 20.9% respectively) of the between group variation and hence expected to discriminate the clusters reasonably well. The third canonical dimension explains a further 11% of the total between cluster variation providing a substantial dimensionality reduction (from 26 to 3). If we can find meaningful description for these canonical dimensions and identify the indicators that influence to these dimensions, then special attention could be focussed on such new variates (and the influential indicators) for efficient administration. The canonical scores (for the 1st two dimensions) are displayed in Figure 3.3, where the cluster membership (1, 2, 3, 4 and 5) and the regional membership (H, C, B, E and W) have been utilised as symbols.

It can be seen from Figure 3.3 that the 1st canonical dimension separates the clusters reasonably well, although there is a notable overlap between clusters 3 and 4, and a close proximity between clusters 2 and 5. As expected, the considerable overlap among the regions is very prominent. This also reveals the similarities or dissimilarities between the different districts in terms of their natural clustering against their actual regional membership. Examination of the canonical coefficients indicated that the 1st dimension is a contrast between the indicators such as the *population density*, *crude birth rate*, *infant* *mortality rate*, *female literacy rate* and *medical facilities* and *CPR-IUD, CPR-Conventional*, *female age at marriage*, *per capita income*, *number of schools* and *length of concrete road*. This means that the districts of cluster #1 have small values for the 1st set of indicators above and large values for the 2nd set (resulting in large negative 1st dimensional canonical scores). The opposite appears to be the case with cluster #3 and to a lesser extent with cluster #4. A similar interpretation can be found for the behaviour of clusters 5, 2 and 4. Therefore it is reasonable to conclude that there exists three main clusters (1, 2and5 and 3and4) of districts (also identified in Figure 3.2).

Figure 3.3. Plot of canonical discriminant scores of 63 districts identified by their cluster membership (1, 2 …) and the region they belong to (H, W, …)



CONCLUSIONS

As mentioned in the introduction, the pictorial representations of the data gave us a clear and better understanding of the problem. Using techniques such as the cluster analysis and canonical discriminant analysis we could partition the 63 districts into three main clusters. It is very clear that the level of female literacy, the contraceptive methods they adopt and the available medical facilities and its awareness play an important role in their welfare. This in turn reflects the major responsibilities and the counter steps, the administration must undertake to increase the female literacy and awareness of the various contraceptive methods among the masses. For administrative purposes one could use the three clusters as a guide to maintain the uniformity of the family welfare programmes. It may be mentioned here that the health administration of UP is centrally located at its capital in Lucknow. The different policies or approaches for different clusters can be implemented and monitored from the capital despite the geographical disparities in various clusters.

ACKNOWLEDGEMENT

Authors are grateful to Mr Sanjay Srivastava for the collection of the data.

REFERENCES

Gnanadesikan, R. (1977). Statistical data analysis of multivariate observations. John Wiley and Sons, New York.

Kumar, K. and Srivastava, S. (1989). An analysis of profiles of acceptors of family welfare programme in India. Int. Union of the Scientific Study of the Population Conf., New Delhi.

Kumar, K. and Sahai, A. (1993). Application of discriminant analysis to the family planning data. *Biometrical Journal, 35*, 869-875.