Demograpliy India Vol. 28. ,No. I (1999}. pp. 111-122

K. Navaneetham* and Prem C. Saxena**

Multivariate Graphical Methods for Characterizing Development: An Application of Chernoff-Type Faces'

Introduction

IN recent years, with rapid development of computers and software technology, there has been a considerable interest in using graphical methods as an analytical tool for data analysis. The use of computer graphics has made it possible to transform data into a number of different types of geometrical patterns. There was a renaissance of activity in graphical display with many of the fascinating techniques developed for representing multivariate data in the past. Examples of these graphics are glyphs and metroglyphs (Anderson, 1957), function plots (Andrews, 1972), trees and castles (Kleiner and Hartigan, 1981) and human-like faces (Chernoff, 1973; Turner and Tidmore, 1977 and Flury and Riedwyl, 1981). These innovations in the methods of graphical displays are yet to be exploited fully for their potentialities through suitable applications; however, they have certainly marked the beginning of the new era in applied multivariate research. This paper highlights the use of graphical methods in statistical data analysis, which may find useful to overcome some of the limitations in the formal statistical procedures. In particular, an attempt has been made in this paper to illustrate the potential use of Chernoff-type faces for characterizing the development of regions. This paper also examines the way in which 'face' method would be effective for communicating

* Associate Fellow. Centre for Development Studies. Thiruvananthapuram. Kerala. India.
** Professor and Head. Department of Population Studies. American University of Beirut. Beirut, Lebanon

[1]An earlier version of this paper was presented at an informal session on Advances in Methods for the Analysis of Demographic Data of the XXIII General Population Conference of the IUSSP held in Beijing. China, October 11-17. 1997.

information of the data and therefore helpful to derive additional information which otherwise would have been missed from the analytical methods.

The next section highlights the use of multivariate graphical display for statistical data analysis in the applied field. We then introduce the Chernoff-type faces and its application in different fields. Subsequently we present the advantage of the face method for characterizing the development of region comparing with the results obtained from formal statistical technique. The last section gives summary and conclusions.

Multivariate Graphics for Statistical Data Analysis

Graphical methods are useful in many situations where formal statistical procedures fail to arrive at meaningful conclusions. For example, when the variables are more, the application of most of the multivariate statistical procedures for analysis becomes quite cumbersome. However, the dimensionality of data may not pose problems in the analysis through newly developed methods of graphical displays. Moreover, several multivariate techniques are not applicable when the number of variables exceeds the number of observations. This limitation could be overcome to some extent using graphical techniques. The use of multivariate techniques for survey data analysis in social science research has increased in the past. It is known that all the multivariate techniques have built-in assumptions. For example, some clustering methods assume that the cluster represents mixture of multivariate normal populations. Similarly, the discriminant function and its testing procedure require the populations to follow multivariate normal probability laws. Analogously, the multiple regression, a frequently used technique, assumes that the residuals of the estimated equations follow the normal distribution. Therefore, the applicability of these techniques becomes invalid when the assumption of normality is not met. In such situations, the graphical methods may prove useful to understand the pattern and nexus of relationships between the variables.

Several statistical procedures involve computation of composite index from a set of multivariate observations. For example, discriminant scores are computed from the magnitude of the variables taken into account that are believed to discriminate between the objects. A metric used in cluster analysis is an index computed from the magnitude of the variables. Analogously, in principal component analysis, the component score is computed by the weights of the variables. The limitation in the composite index derived from analytical methods is that while computing them, a linear additive model is generally assumed. The assumptions of linearity may not hold true in many practical situations. These problems could be overcome to a certain extent through graphical methods. The other situation where graphical methods can play a vital role in social science research is that when the explanatory variables are not independently distributed (problem of multicollinearity)—a mandatory condition for application of most of the formal statistical procedures. In addition, the problem of autocorrelation is being faced in the case of time series analysis. Although statistical tests have proposed to detect the existence of the presence of multicollinearity in data, the procedures available to circumvent the multicollinearity are quite cumbersome and are not easy to apply. In contrast, the analysis of data through graphical methods does not pose such problems; they rather facilitate to preserve the information of each variable and enable studying the effect of their individuality.

Researchers in social science are also being confronted with the sensitivity of outliers or wild points while analysing multivariate data. Detecting outliers in a set of multivariate observations is not trivial; especially when there are several outliers. The formal statistical procedures used for identifying the outliers do not always detect them because the methods are based on sample mean and co-variance, which are themselves affected by the outliers (Rousseeuw and van Zomeren, 1990). Consequently, the solution to this problem is cumbersome in the formal statistical analysis of multivariate observations. It is well known that the presence of outliers in the analysis may distort the results. It is, therefore, mandatory to remove the outliers before any meaningful statistical analysis of data is carried out. Again, the graphical methods may enable us to detect the multivariate outliers present in the data set easily. In a more standard setting, no analysis would be considered complete without the detailed graphical examination of data. The limitations in the analytical procedures and the utility of graphical methods in statistical data analysis made it possible to increase the attention paid to graphics as an intellectual discipline (Chernoff, 1978).

Chernoff-type Faces

Among the various methods of statistical graphics for displaying and analyzing multivariate data, the use effaces seems to be growing in popularity due to their potential useful medium for communicating information from complex data and to its more perceptual integrity. Also, the representation effaces may present a *gestalt* that recalls psychologically meaningful objects or ideas. In fact, our familiarity and our ability to perceive and remember even the small changes in the structure of human face are inculcated from the childhood. According to Chernoff (1978)

> ". . .We perceive the faces as a *gestalt* and our built in computer is quick to pick out the relevant information and to filter out the noise when looking at the limited number of faces."

Thus, the use efface representation of multivariate data may be effective in revealing complex relations not always visible from simple correlation based on two-dimensional linear analysis. The faces can be useful for identifying outliers; clustering the points; discriminating the variables and to detect substantial changes. Besides this, the faces may give much of information and insight into the structure of complex data.

Chernoff (1973) proposed a graphical technique representing a point in 18— dimensional space by drawing a face whose 18 characteristics (shape of the face, length of the nose, curvature of mouth, size of eyes etc.) are determined by the coordinates or position of the point. Several changes were made to Chernoff s original faces. Turner and Tidmore (1977) introduced the asymmetric Chernoff-type faces that could be generated on a line printer. Flury and Riedwyl (1981) made some real improvements in the construction effaces (hereafter referred as F&R faces) where paired measurements are mapped separately to the left and right half of a face which double the number of representable variables from 18 to 36. These faces have less implicit co-variation among the features than in Chernoff s faces (Wainer, 1983). Also, F&R faces give more 'human' look than the earlier ones.

Attempts were made to validate the Chernoff-type faces for clustering multivariate observations (Chernoff and Rizvi, 1975; Tidmore and Turner, 1983; Saxena and Navaneetham, 1986, 1991 and 1993; Navaneetham, 1994). Nevertheless, a few studies have attempted to exploit the Chernoff-type faces for its applications in different field. For instance, Chernoff (1973) applied the face technique for the analysis of geological data. Jacob (1979) used the Chernoff faces to understand a multidimensional spatial relationship by studying a plot of spatial location against other attributes of the data. Flury and Riedwyll (1981) applied the technique effaces (F&R faces) to distinguish between real and forged Swiss bank notes. Saxena (1983) has used Flury and Riedwyl faces to classify 13 selected countries in Asia and the Far East into groups according to their overall development and stages of demographic transition. Navaneetham (1994) exploited the face technique to various demographic data analyses.

Characterizing Development of Regions: An Illustration

This section presents how the Chernoff-type faces will be quite constructive for characterizing the regional development. The results obtained from face technique are compared with the conclusions derived from the statistical procedures.

Composite Indexes

Social scientists construct a composite index for studying the regional development or the physical quality of life or index for human development using a set of multivariate characteristics (see, Morris, 1979; UNDP, 1990; lyengar and Sudarshan, 1982). While constructing these indexes, the practice is to weigh each characteristic equally (Morris, 1979). Therefore, this method has a limitation, which assumes constant variance for each characteristic. Other objective methods, which are widely used for constructing a composite index based on a set of multivariate observation are Principal Component and Hedonic Method (Ram, 1982; Slottje, 1991). The Principal Component method relies solely on variance and covariance matrix to derive the weights for constructing an index. Hedonic method uses instrumental variables and gives weights to the attributes by the regression coefficients. Also this method assumes a linear additive model. Although, other methods which are considered to be non-conventional in approach also depend on the criteria of selection of weights. For instance, lyengarand Sudarshan (1982) assumed that the weights vary inversely as the variation in the respective variable. If the weights are not suitably selected, some variables will dominate the value of the rest of the variables and distort comparison between the units. Moreover, different methods used for constructing composite indexes may vary according to their rank order performance and thus may distort the results. For example, Slottje (1991) has studied the physical quality of life index (PQLI) for 126 countries. The author has used six methodologies for constructing physical quality of life index. All these six methods gave a different ranking of PQLI across countries. The author arrived at the final ranking of PQLI by averaging the ranks of the six indices of PQLI constructed based on different methods. Therefore, the overall indexes of these six methods are related to each other in a statistical sense—their differences among them by rank order, to conclude that choice of weighting technique is important in the construction of an overall index. In these situations, social scientists are confused to determine which method should be used for constructing composite index. Also the composite index provides an aggregate value and not the overall impression of the data. In fact, one can derive at the same value of a composite index by several combinations of the weights of the selected variables. Thus, any composite index may not reveal a true structure of the data and thus may mislead the researchers.

To characterize the different stages of development of the region using composite indexes, it has been assumed that the index follows some statistical distribution. lyengar and Sudarshan (1982) assumed that their development index follows the Beta distribution for characterizing the different stages of development. Vidwan (1983) demonstrated that assumption of normal distribution will give a better classification than the Beta distribution. If it does not follow the assumed distribution, the classification based on this assumption may be misleading. Also, researchers may require *a priori* information on the number of groups they wish to classify.

In the following sections, we will demonstrate how the face analysis may prove to be quite promising in overcoming some of the issues mentioned above. For an illustration, we have used the databases for Indian states given in Rao (1985). The lyengar and Sudarshan (1982) method is applied for constructing a composite index[2].

Results Based on Composite Indexes

Table 1 shows the composite index derived by lyengar and Sudarshan method (IS

For details about the method and description, please see lyengar and Sudarshan, 1982.

method) for the major states in India[3]. The states have been arranged in a descending order of the level of development. It is revealed from Table 1 that Punjab ranked first with respect to overall development followed by Kerala and Maharashtra. Further, it is important to note that the index for Kerala (0.6572) and Maharashtra (0.6413) is almost the same. The state Orissa was put in the lowest position followed by Madhya Pradesh and Bihar.

TABLE 1: LEVEL OF DEVELOPMENT OF INDIAN STATES

States	*IS Method*	
	Index	*Rank*
Punjab	0.8270	1
Kerala	0.6572	2
Maharashtra	0.6413	3
Tamil Nadu	0.5560	4
Haryana	0.5388	5
Gujarat	0.4542	6
Karnataka	0.3715	7
West Bengal	0.3669	8
Andhra Pradesh	0.2371	9
Jammu & Kashmir	0.2163	10
Uttar Pradesh	0.1853	11
Rajasthan	0.1547	12
Assam	0.1530	13
Bihar	0.1462	14
Madhya Pradesh	0.1418	15
Orissa	0.1229	16
India	0.3325	—

The IS development index has been used to classify the Indian States into different stages of development. It has been grouped into five stages of development as presented in Table 2.

It is observed that the classification of states is sensitive to the assumption made regarding the distribution of the development indexes. For example, if we assume the index follows Beta distribution the state of Punjab stood as a distinct group. When it was assumed to follow normal distribution, the states Maharashtra, Kerala and Punjab grouped together. Thus, it is clear that states grouped differently with respect to these two assumptions from the same index value. The other problem is with the states that belonged to border line cases. For example, when we assumed that the index follows beta distribution, the state Gujarat (0.4542) appears to be under borderline

[3]We have also applied Principal Component Method for constructing composite index. This method gave similar ranking, though not identical, as that of IS method.

TABLE 2: CHARACTERIZATION OF STATES IN INDIA BY THEIR LEVEL OF DEVELOPMENT

Category	*IS Index*	
	Beta Distribution Assumption	*Normal Distribution Assumption*
Very Backward	Nil	—
Backward	Orissa Madhya Pradesh Bihar Assam Rajasthan	Orissa Madhya Pradesh Bihar Assam Rajasthan Uttar Pradesh Jammu & Kashmir Andhra Pradesh
Developing	Uttar Pradesh Jammu & Kashmir Andhra Pradesh West Bengal Karnataka	West Bengal Karnataka Gujarat
Developed	Gujarat Haryana Tamil Nadu Maharashtra Kerala	Haryana Tamil Nadu
Highly Developed	Punjab	Maharashtra Kerala Punjab

case[4]. Similarly, the two states namely, Maharashtra (0.6413) and Kerala (0.6572) fell under the borderline cases when we classified the states using normal distribution assumption. In that situation, researchers will be in dilemma with the border line cases about their position with respect to development.

[4]The classifications of different stages of development from the index are as follows:

Stages of Development	*IS method*	
	Beta distribution	*Normal distribution*
Very backward	$0 < IS \leq 0.040$	$0 < IS \leq 0.08$
Backward	$0.040 < IS \leq 0.180$	$0.080 < IS \leq 0.246$
Developing	$0.180 < IS \leq 0.419$	$0.246 < IS \leq 0.476$
Developed	$0.419 < IS \leq 0.724$	$0.476 < IS \leq 0.641$
Highly developed	$0.724 < IS \leq 1.000$	$0.641 < IS \leq 1.000$

The F&R faces for 16 major states as well as for India were drawn using eight sectoral variables^. The sectoral variables and their assignments to the face parameters are given in Table 3.

TABLE 3: ASSIGNMENT OF SECTORAL INDICATORS TO THE FACE PARAMETERS

.Name *of the sector*	*Face parameter*
1. Agriculture	density of eyebrow
2. Industry (general)	eye size
3. Industry (small scale)	darkness of hair
4. Banking	curvature of eyebrow
5. Power	upper hair line
6. Transport	lower hair line
7. Health	face line. nose
8. Education	curvature of mouth, size of mouth

Note: The other face parameters were kept constant and given a default value.

An experiment was conducted with 25 subjects and they were asked to classify the faces into some groups, which look alike by looking at the *gestalt* of the faces. The results of the experiment are used to develop a perception matrix indicating number of subjects classified by any pair of faces together. From the perception matrix, a true characterization of development of the states has been obtained[6]. Figure 1 shows the states that have been characterized into five stages of development. Faces have been arranged in ascending order of the value of the IS index within the groups. It was observed that the states Kerala and Punjab stood as outliers when we looked at *the gestalt* of the faces. The states such as Orissa, Madhya Pradesh, Bihar, Assam, Rajasthan and Uttar Pradesh are grouped together and they appear more homogeneous regarding the *gestalt* of the faces. From the face analysis, just by merely looking at the faces, it may be possible to conclude that these states were under same degree of development with respect to all sectoral indicators. In the second group three states namely, Jammu and Kashmir, Andhra Pradesh and West Bengal are classified together. If we look at these faces, the pattern of development was not the same between the sectoral indicators. For example, the bigger eye size for West Bengal indicates that the industrial (general) development in the state was higherthan in Jammu and Kashmir. This type of information would have escaped notice, had we examined the level of development through the composite indexes only. Further, the states namely Karnataka, Gujarat, Haryana, Tamil Nadu and Maharashtra were grouped together. These states were under uniform development

[5]*For* details about the method of constructing faces of this type, see Flury, 1980.
[6]For details about the method of recovering true cluster structure from perception matrix, see Saxena and Navaneetham, 1993.

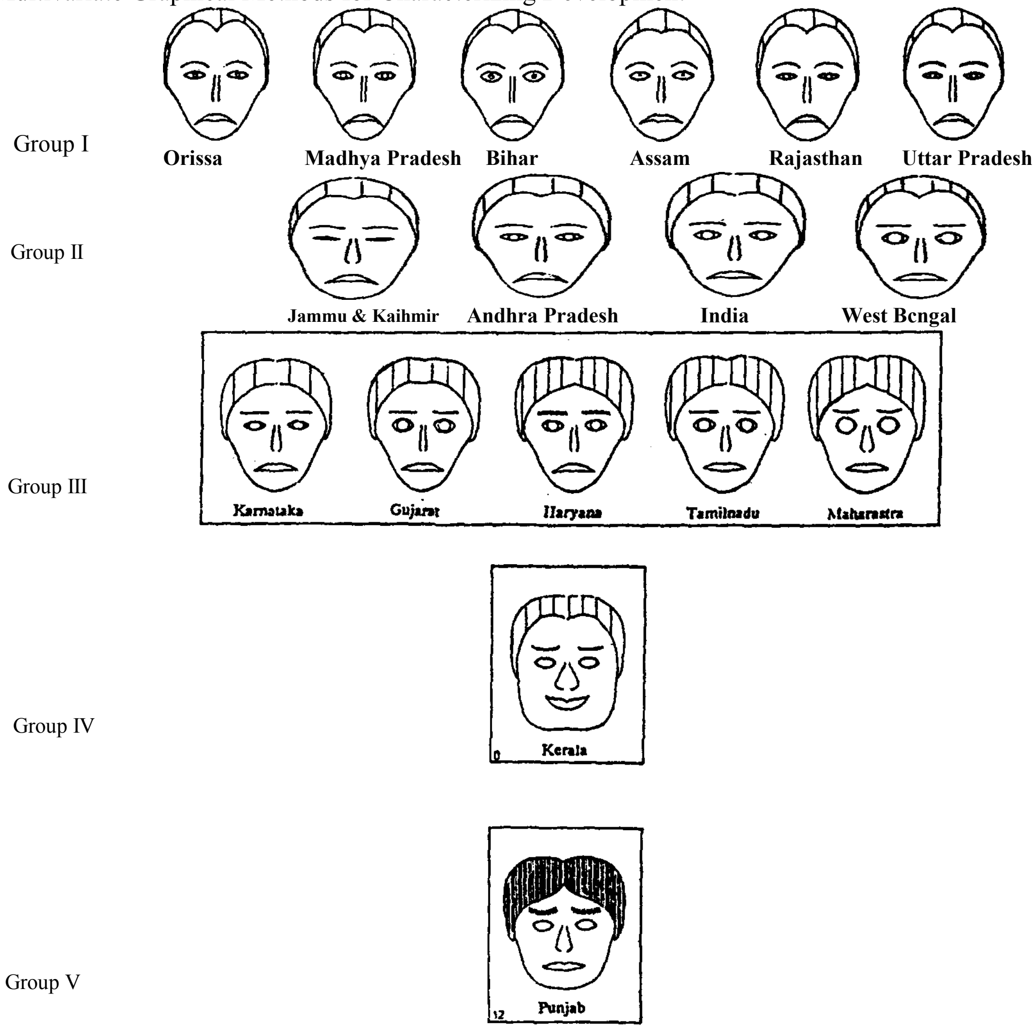

Fig. 1. Faces characterizing the development of states in India

with respect to industry (small scale), power, transport health and education sectors. Nevertheless, one can also observe by looking at the faces that the degree of development is not same in some sectors within the group. For example, Maharashtra had highly developed state concerning industrial *(eye size)* and banking *(curvature of eyebrow)* sectors whereas Haryana showed higher development with respect to agriculture *{density of eyebrow).*

Kerala occupied a unique place as an outlier since the process of development in the state was not uniform in all the sectors of the economy. For instance, the state has experienced development in the educational *(curvature of mouth and size of mouth)* and health *(face line and nose)* sectors; whereas it lagged behind with respect to other sectors namely agriculture, industrial, banking, transport and power. Further, the state of Punjab was characterized as a highly developed state using composite index, however, it lagged behind with respect to education and health sectors as compared to Kerala. Moreover, it stood behind the development regarding industrial (general) sector as compared to Maharashtra, Gujarat and Tamil Nadu. This type of phenomena may escape notice if we use only composite indexes for studying regional development. Consequently, if the process of development is not uniform concerning all the indicators of development in a state, the composite indexes may fail to reveal it. The use ofF&R faces may facilitate to study those aspects as well. It would be important to show how the composite indexes may be unable to give solution to the researchers about the true characterization of development. For instance, it may be worthwhile to note that the values of the composite index for the states of Maharashtra and Kerala were found to be 0.6413 and 0.6572, respectively. From these values, we may conclude that the two states are nearly at the same level of development. However, when we look at the faces for Maharashtra and Kerala from Figure 1, it clearly indicates the distinct pattern of development. This crucial and important observation would have escaped notice if we studied only through composite indexes.

Summary and Conclusions

In this paper, an attempt is made to show the potential application ofChernoff-type faces for characterizing the development of regions in India. For an illustration, data on developmental characteristics for 16 major states of India were used and F&R faces were drawn. The results obtained from the faces were compared with the same from a formal statistical procedure namely lyengar and Sudarshan method. From the analysis, it was observed that the use of F&R faces, is not only confined to characterize the development of states, but also facilitate the researchers to find out which sector(s) is (are) responsible for classifying them into various stages of development. This may not be possible from the composite index constructed by formal statistical procedure. Although the composite index is useful to rank the States as well as to group them according to the level of development, the face analysis may help to understand the differences with respect to sectoral development within the group. Thus, the use efface analysis would facilitate to confirm or contrast the findings observed from the composite index. Further, if there are any outliers present in the data, the face technique may bring out this fact in the analysis. Therefore the development index constructed by applying the formal statistical procedures may be used for studying the overall development of the regions, but it may not be sufficient to contemplate the sectoral development. The use of F&R faces along with the composite index may supplement the information and may ameliorate the researcher's ability to interpret the data in a way that is much more meaningful. Analogously, F&R faces may find a useful application in characterizing State/Country on the basis of the physical quality of life index, poverty index and human development index.

Acknowledgement

The authors are grateful to an unknown referee for his valuable comments, which was very useful to improve the earlier version of the paper.

References

Anderson. E., 1957, A semigraphical methods for the analysis of complex problems. *Proceedings of the National Academy of Science,* **13:** 923-27. Andrews. D. F., 1972, Plots of high dimensional data. *Biometrics,* 28: 125-36. Chernoff. H.. 1973. The use to represent points in K-dimensional space graphically. *Journal of the American Statistical Association,* **68:** 361-68. Chernoff. H. and M. H. Rizvi., 1975, Effect on classification error of random permutations of features in representing multivariate data by faces. *Journal of the American Statistical Association,* 70: 548-59. Chernoff. H.. 1978. Graphical representations as a discipline. *In:* P.C. Wang(ed.), *Graphical Representation of Multivariate Data.* Academic Press, New York. Dillon. W. R. and M. Goldstein, 1984, *Multivariate Analysis: Methods and Applications.* John Wiley & Sons. New York. l-'lury. I}.. 1980. Construction of an Asymmetrical Face to Represent Multivariate Data Graphically. Technical Report No. 3, University of Berne, Dept. of Statistics. Plury. 1). and 11. Riedwyl, 1981, Graphical representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association,* 76: 757-65, Kleiner. B. and ,1. A. llartigan. 1981, Representing points in many dimensions by trees and castles. *Journal of l he American Statistical Association,* **16:** 260-76. lyengar. N. S. and Sudarshan. 1982, A method of classifying regions from multivariate data. *Economic and Political Weekly.* Dec. 18. Jacob. R. .1. K..1979. Mapping geographical relationships with faces. Paper Presented at the Harward Computer Graphics Week Conference, Cambridge. Massachusetts. July 15. Morris. M. D.. 1979, *Measuring the Condition of the World Poor: The Physical Quality of Life Index* Pcrgamon. Oxford. Navaneetham. K... 1994. Statistical Graphics for the Analysis of Data of Multivariate Observations: Some Theoretical Issues and Demographic Applications. Unpublished Ph.D. Thesis, University of Bombay. Bombay. Kam. Rati. 1982. Composite indices of physical quality of life, basic needs fulfillment and income. *Journal of Development Economics,* **II:** 227-47.

Rao, Hemalata. 1985. Inter-state disparities in development in India. *In:* G. P. Mishra (ed.). *Regional Structure a/Development and Growth in India.* Ashish Publishing House, New Delhi. Rousseeuw. P. J. and b.C. vanZomeren, 1990, Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association,* 85: 633-39. Saxena. P. C.. 1983. Use of faces as an iconic display for multidimensional data in regional analysis. *In:* K. Srinivasan and S. Mukerji (eds.). *Dynamics of Population and Family Welfare.* Himalaya Publishing House. Bombay. Saxena. P. C. and K. Navaneetham, 1986, The validation of Chernoff faces as a clustering algorithm. *Proceeding of the VIII Annual Conference of the Indian Society for Probability and Statistics,* Shivaji University. Kolhapur, pp. 179-93. Saxena, P. C. and K. Navaneetham, 1991, The effect of cluster size, dimensionality, and number of clusters on recovery of true cluster structure through Chernoff-type faces. *The Statistician,* **40:** 415-25. Saxena, P. C. and K. Navaneetham, 1993, Comparison of Chernoff-type face and non-graphical methods for clustering multivariate observations. *Computational Statistics and Data Analysis,* **15:** 63-79. Slottje, D. J., 1991, Measuring the quality of life across countries. *The Review of Economics and Statistics,* **LXXII:** 684-93. Tidmore. F. E. and D. W. Turner, 1981, On clustering with Chernoff-type faces. *Communications in Statistics.* **A 12:** 381-96.

Tukey. J. W., 1977. *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass. Turner. D. W. and F. E. Tidmore, 1977. Clustering with Cheinoff-type faces. *Proceedings of the American Statistical Association,* Statistical Computing Section, pp. 372-77. UNDP. 1990, *Human Development Report 1990.* Oxford University Press, New York. Vidwan. S. M., 1983. Discussion on A method of classifying regions from multivariate data. *Economic and Political Weekly,* December, 17. Wainer. H.. 1983. On multivariate display. *In:* M. Z. Rizvi (ed.), *Recent Advances in Statistics.* Academic Press. London.