COMPUTER GRAPHICS FOR MULTIVARIATE DATA

Robert Cléroux[1], Yves Lepage[2] and Normand Ranger[1]

(1)Département d'informatique et de rech. opér., (2)Département de mathématiques et de stat., Univ. de Montréa

ABSTRACT

The exploration of multidimensional data involves the use of a set of empirical techniques which aid in the discovery of interesting avenues to be pursued in later statistical analysis. Data exploration often directs this analysis. The availability and power of computers has changed the nature of statistical work and has made the exploration of multidimensional data more accessible. Graphical methods constitute one of the main tools for data exploration, and they are therefore of primary importance.

In this article, four graphical representation methods are presented and applied to atmospheric pollution data for the Montreal region. These representations enable the data from each monitoring station to be visualized and grouping may then be formed from observed similarities.

RESUME

L'exploration de données multivariées nécessite l'utilisation de techniques empiriques qui font souvent découvrir des avenues intéressantes pour l'analyse subséquente. L'exploration guide alors l'analyse des données. La disponibilité de la puissance informatique a changé la nature du travail statistique en rendant possible l'exploration de données multivariées. Les représentations graphiques constituent l'un des principaux outils de l'exploration de données et sont en conséquence, d'une grande importance.

Dans cet article, quatre méthodes de représentation graphiques de données multivariées sont introduites et appliquées à des données de pollution atmosphérique de la région de Montréal. On peut ainsi visualiser les postes de pollution et effectuer certains regroupements à partir des similarités observées.

INTRODUCTION

Graphical representations have been in use for a long time in several disciplines. The usefulness of these methods lies in the capacity of the human eye to recognize shapes and to identify similarities and aberrations. For example, a two dimensional cloud of points reveals at a glance the essential relationship between the two variables. However, in several dimensions, even projections on different planes are insufficient to simplify the graphical representation, especially with a fairly large number of dimensions. Data in more than two or three dimensions is therefore difficult to grasp. However, the advent of the computer has facilitated the exploration of multivariate data in general and has, in particular, produced more accessible graphical representations for such data.

Data exploration is defined as the set of empirical techniques applied to data before the statistical analysis proper begins. The purpose of these techniques is basically to discover interesting avenues to be pursued in more detail. The techniques usually involve summarizing the information through elementary calculations, tables, diagrams, histograms and graphs. They provide a way of establishing intimate contact with the raw data and of learning about it. Data exploration often orients statistical analysis, and graphical methods are therefore of primary importance.

There are two main classes of graphical methods for multivariate data. Some methods are mainly used in the context of a geometric structure or statistical model following data analysis. All factor analysis methods belong to this class. Other methods are used mainly at the exploratory stage to represent raw data. This paper discusses only this latter class of methods.

Any graphical representation should have certain of the following qualities: it should communicate information easily and rapidely, help the reader to understand the information, produce a greater impact when the information is more important, have a mnemonic

effect, in the sense that important information should be retained by the reader, be simple, compact and attractive, be clear, precise and without distorsion, be quickly and easily comprehensible, be constructed using standard forms, allow the representation of a large number of dimensions simultaneously, allow comparisons and groupings.

There are many graphical representation methods for raw data. A sampling is listed here: glyphs (Anderson, 1960), stars (Siegel et al, 1971A), faces (Chernoff, 1973; Chernoff and Rizvi, 1975), curves (Andrews, 1972), constellations (Wakimoto and Tagun, 1978), profiles (Bertin, 1967), triangles (Pickett and White, 1966), draftsman's display (Chambers et al., 1983), weathervanes (Cleveland and Kleiner, 1974), boxes (Hartigan, 1975), trees (Wakimoto, 1977), trees and castles (Kleiner and Hartigan, 1980). The interested reader will find other references on this subject in the bibliography. All are not, however, cited in the text.

In the next sections, we study four of these methods in more detail, applying them to atmospheric pollution data for the Montreal region.

THE DATA

The data to be considered consists of measures of the sulphur dioxide concentration in parts per hundred million (pphm), collected at various monitoring stations in the Montreal region during the year 1975. The data are collected by the "Services de Protection de l'Environnement du Québec" which also provides summary statistics. The results are published in the pamphlet "Qualité de l'air" by the "Editeur officiel du Québec" in the form of monthly tables of hourly or bi-hourly mean concentrations. The tables also provide 24 hour averages, monthly means for each hour and their respective maxima.

The monitoring stations chosen are sufficiently well geographically distributed to both adequately cover the territory of Montreal, and to allow the study of local pollutant effects. Since the purpose of this article is not to study the problem of sulphur dioxide pollution in Montreal but rather to present certain graphical representation methods, only monitoring stations 1, 12, 13 and 20 will be represented in the following discussion.

These stations are:

Station		Height above		Type of equipment	
No.	Address	Sea Level	Ground Level	Continuous	Sequential
1	Botanical Garden Montreal	55m	4m	Titrilog	
12	1125 Ontario East Montreal	23m	13m	Technicon	
13	1212 Drummond Montreal	35m	12m		Sequential
20	525, 9th Avenue Pointe-aux-Trembles	9m	6m	Beckman 906	

The presence of sulphur dioxide in the air is determined by several methods: some stations are equipped with automatic machines which register atmospheric gas concentration continuously, while others have sequential apparatus which takes samples every two hours which then have to be analyzed in a laboratory.

Each pollution monitoring station is represented by a vector of dimension 17. The first 12 components give the monthly averages of sulphur dioxide concentrations for the year 1975, and the other 5 components are determined by the relative frequencies of the daily maxima; component 13: % daily maxima of 15 pphm or over, component 14: % daily maxima between 10 and 15 pphm, component 15: % daily maxima between 5 and 9 pphm, component 16: % daily maxima of 3 or 4 pphm, component 17: % daily maxima of 2 pphm or less.

THE STARS METHOD

Let p variables be measured on n individuals to obtain n vectors of the form $X' = (X_1, \ldots, X_p)$. Using polar coordinates, the circle is divided into p equal angles. This division defines p planar vectors whose origins are located at the centre of the circle, and whose directions are determined by the angles. For i=1, ..., p, variable X_i is placed on the ith vector at a distance from the origin proportional to its size. The p variables are thus placed in the polar plane. The extremities of the vectors are then joined together to obtain a polygonal figure called a star (see Siegel et al, 1971A,B). We obtain n stars, one for each vector of observations. Similarities and differences between the stars may then be detected.

In certain practical situations, the initial variables must first be transformed to ensure a certain compatibility between them. Sometimes, a variant of this method may be useful. For example, a star corresponding to the

vector of means might be constructed and then tranformed into a circle by multiplying each component of the observation vectors by an appropriate factor. Then, the standard deviations might be traced on each vector of the circle to indicate the dispersion of the variable with respect to the mean.

Figure 1 indicates the arrangement of variables around a conceptual star, while Figure 2 shows the stars corresponding to the 4 stations described above. For each representation, components 13 to 17 were

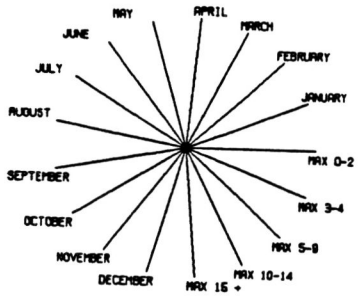

Figure 1: Arrangements of variables around a conceptual star

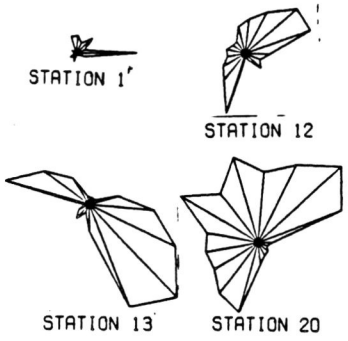

Figure 2: Representation of certain pollution monitoring stations by stars

multiplied by 5 to make them move compatible with the others.

It can be seen that pollution levels are relatively low at station 1 (botanical garden), high at station 20 (refineries to the East of Montreal) while the structures of pollution are somewhat similar for stations 12 (Ontario Street East) and 13 (Drummond Street) although the overall level is higher at station 13 than station 12.

THE CURVES METHOD

When n observations of the vector $X' = (X_1, \ldots, X_p)$ are measured, or in other words, p variables are measured for n individuals, each of the n observation vectors is represented by a function of the form

$$f_x(t) = \frac{X_1}{\sqrt{2}} + X_2 \sin t + X_3 \cos t + X_4 \sin 2t + X_5 \cos 2t + \ldots$$

and its graph is drawn over the interval $-\pi < t < \pi$. Thus, n different curves will be traced. It would also be possible to draw the graph corresponding to the mean of a group, and then to repeat this for each group studied to compare these visually. This method has interesting statistical properties (Andrews, 1972).

Figure 3 shows the curves $f_x(t)$ for the stations studied, as a function of the parameter t, $-\pi < t < \pi$. The values of the $f_x(t)$ are somewhat arbitrary but a large value of $f_x(t)$ implies a high level of pollution. The vertical axis can be taken as representing an axis of pollution.

It can be seen here, also, that the pollution level is low at station 1 (botanical garden) and high at station 20 (refineries to the East of Montreal). The pollution levels of stations 12 and 13 appear to be roughly similar.

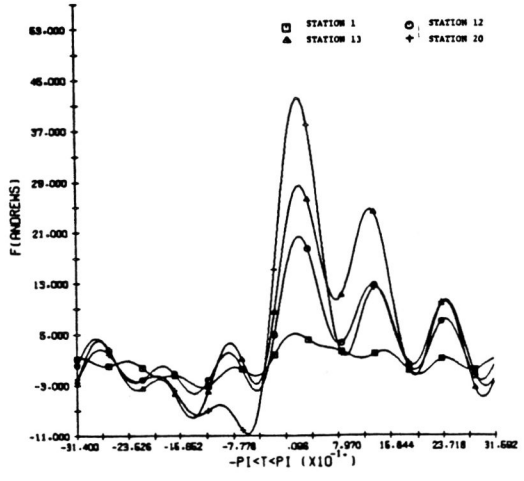

Figure 3: Andrew's curves corresponding to stations 1, 12, 13 and 20

THE FACES METHOD

Chernoff (1973) proposes the representation of a vector $X' = (X_1, \ldots, X_p)$ of observations using a human face whose characteristics are determined by the values of the components. Each component corresponds to a part of the face. More precisely, the parameters used to represent a vector of 20 components are as follows: parameter 1: width of the face, 2: level of ear, 3: height of face, 4: excentricity of upper face, 5: excentricity of lower face, 6: length of nose, 7: level of mouth, 8: curvature of mouth, 9: length of mouth, 10: level of eyes, 11: distance between eyes, 12: angle of eyes, 13: excentricity of eyes, 14: size of eyes, 15: position of pupil, 16: vertical position of eyebrows, 17: angle of eyebrows, 18: length of eyebrows, 19: diameter of ear, 20: width of nose.

The nose corresponds to a triangle, the ears and pupils are circles. Ellipses are used for the face outline and the eyes while the arc of a circle describes the mouth and straight lines are used for the eyebrows.

To use this method, each component must be brought to within a precise interval to control the dimensions of the face. In addition, when the dimension of a vector is less than 20, some pre-assigned values are used. Thus, n faces are obtained, one for each vector of observations.

Chernoff's faces suffer from the fact that extreme values of certain parameters diminish the effect of others (see Chernoff and Rivzi, 1975, and Bruckner, 1978). Conscious of these methodological limits, Flury and Riedwyl, 1981, presented a new "Chernoff face" which allows the appropriate representation of all observations and also increases the size of the vector considered.

The different parameters used are as follows: parameter 1: size of eye, 2: size of pupil, 3: position of pupil, 4: angle of eye, 5: horizontal position of eye, 6: vertical position of eye, 7: curvature of eyebrow, 8: density of eyebrow, 9: horizontal position of eyebrow, 10: vertical position of eyebrow, 11: upper outline of hair, 12: lower outline of hair, 13: outline of face, 14: density of hair, 15: angle of hair, 16: nose, 17: mouth size, 18: curvature of mouth.

The eyes and pupils are drawn as arcs of circles. The hair, nose and mouth, and outlines are drawn using parameterized curves (polynomials) (see Flury, 1980). This method allows the representation of pairs of vectors of dimension 18; asymmetry in the face obtained reflects changes in each component. Obviously, this also allows the representation of vectors of dimension up to 36. If the dimension is less than 18 or 36, as the case may be, some default values are used. Hence, n new faces are obtained, one for each observation vector.

Figure 4 shows Chernoff's faces and Figure 5 shows the new faces according to Flury and Riedwyl for stations 1, 12, 13 and 20. Again it is seen that the pollution level is low at station 1 (botanical garden) and high at station 20 (refineries to the East of Montreal). The levels at stations 12 and 13 are comparable although it is slightly higher at station 13.

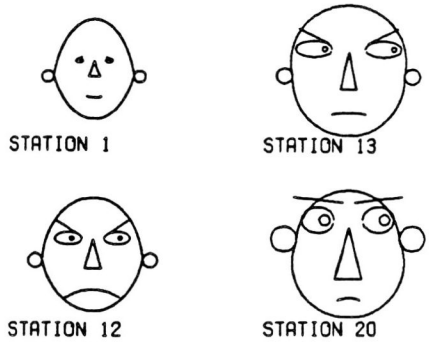

Figure 4: Chernoff's faces for stations 1, 12, 13 and 20

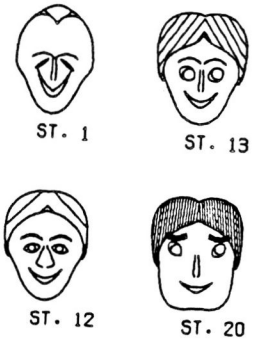

Figure 5: New Chernoff's faces as suggested by Flury and Riedwyl for stations 1, 12, 13 and 20

THE TREES AND CASTLES METHOD

In the graphical representations illustrated so far, the order of variables in the vector plays an important role. In fact, each graphic would look quite different if the variables under study were permuted. Furthermore, these graphics contain a certain inherent correlation structure because of their design. For example, in Chernoff's faces, the length of the eyebrows is highly correlated with the length of the eyes. Both of the problems mentioned above are discussed by Chernoff and Rizvi, 1975. To eliminate these difficulties, Kleiner and Hartigan, 1981, developed a method which uses a hierarchical clustering algorithm on the variables to make them basically independent of permutation effects. This technique is known as the "trees and castles" method.

First, suppose the variables X_1, ..., X_p are normalized with mean 0 and standard deviation 1. The most frequently used hierarchical clustering algorithm is the complete linkage method with Euclidean distances between the variables (see Hartigan, 1975). This algorithm is preferred because it tends to divide the variables into two clusters of the same size. The two variables which are closest to one another are joined to form the first cluster. The distance between this cluster and each of the other variables is defined as the maximum of the distances between each variable of the cluster and the other variable. The process is repeated by joining together the pair of clusters or of variables with the smallest distance between them, with distance then being defined as the maximum distance between pairs. At the final step, two clusters are joined together to form a single cluster containing all the variables.

For the 17 variables observed at the 14 monitoring stations, the results of the complete linkage hierarchical clustering algorithm are shown in the dendogram of Figure 6. The first variables to be clustered together are June and September; February and March form the next group, followed by max 15+ (component 13) and July. The process continues until the cluster made up of max 0-2 (component 17) and max 3-4 (component 16) joins up with the cluster containing all the other variables.

The trees method represents each vector of observations by a tree. All the trees have the same typology as the dendogram of variables obtained from the complete linkage hierarchical clustering algorithm and each leaf of the tree corresponds to a variable.

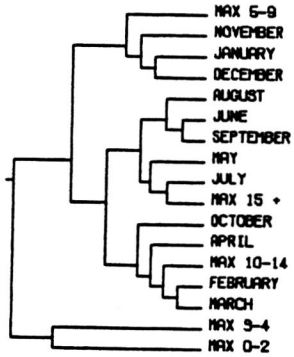

Figure 6: Dendogram obtained by the complete linkage hierarchical clustering algorithm for the 17 variables observed at monitoring stations

Before drawing the graph, the width and length of the branches, and the angle between branches must be chosen. The width of a branch is proportional to the number of variables above that branch. A variable is said to be above a branch, if the path from the leaf representing this variable to the foot of the tree includes this branch. Thus, in our example, as each leaf has a width of 1, the base of each tree will have a width of 17, while the branch supporting the leaves for July and max 15+ is of width 2 increasing to width 3 when May is added (see Figure 7). The angle between two branches at a division point is a linear function of the maximum of the logarithm of the distance between the variables above these two branches; the range of variation of these angles is a parameter which may be determined in advance. In our example, the minimum angle has been set at 5° while the maximum angle permitted is 95°.

At each division point, the orientation of the branches must be chosen. The stem of the tree is defined as the path which begins at the foot of the tree and at each division follows the widest branch until the division point with two branches of the same width. The stem alternates directions at each division point. At a division point not touching the stem, the widest branch will bear towards the right for division points to the right of the stem, or towards the left for division points to the left of the stem. In other words, the widest branches are furthest from the stem. When two branches have the same width, the choice is arbitrary.

The angle between a branch and the vertical is proportional to the width of the branch except on the stem where the angle is inversely proportional to the width of the branch. However, the sum of angles with respect to the vertical corresponds to the angle determined previously. Thus, the stem will tend to be roughly vertical. Finally, the length of a branch is proportional to the average value of the variables above this branch.

In Figure 7, trees corresponding to the four pollution stations studied are presented. Those corresponding to stations 1, 12 and 13 have been doubled in size to facilitate the visual perception. Note that in general the level of pollution appears low

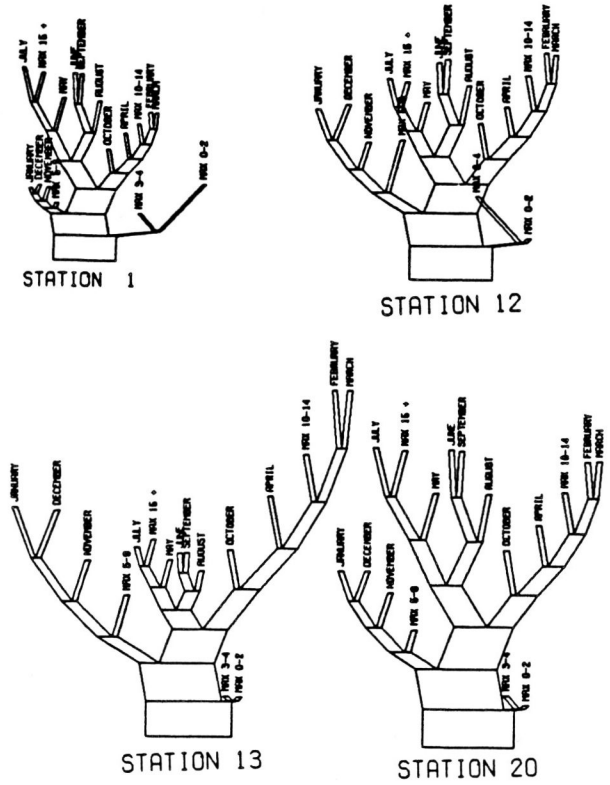

Figure 7: Representation of stations 1, 12, 13 and 20 by trees

for station 1 (botanical garden) and high for station 20 (refineries to the East of Montreal). Although not as high as station 20, the level of pollution at station 13 (Drummond Street) can be seen to be higher than that at station 12 (Ontario Street East).

The graphical representation using trees is very useful when the variables studied can be divided into clusters of highly correlated variables. It is also desirable that the measures of the variables be comparable. However, this method makes it difficult to compare variables of the same tree even if they are next to one another. Kleiner and Hartigan, 1981 suggest representing each point by a castle. This method is a mixture of the trees method and the profiles method (see Bertin, 1967) and it allows variables of the same tree to be compared. It also enables variables for the same observation to be compared more easily than in the profiles method.

Suppose that the variables take positive values which are comparable. Complete linkage hierarchical clustering using Euclidean distances on the variables is used to generate a tree. This tree forms the basis for the construction of castles. The width of each branch is set to be proportional to the number of branches above it, the angle between all branches is fixed at zero, and the order of the variables is that obtained by the complete linkage clustering algorithm.

In the following way, each vector is represented by a castle: the upper extremity of a branch is at a distance v from the base where v is the minimum value of all variables above the branch minus qd, where q is the number of branches joining the branch to the variable with the minimal value, and d is a value to be chosen.

The distance between the base and the extremity of the branch containing a single variable corresponds to the value of that variable, so this branch gives the same information as the profiles method. The position of the other branches reflects the information presented by the trees method. The value of d must be strictly positive to ensure that the tree structure is retained.

Castles for the 4 pollution stations studied are presented in Figure 8. Note once more that station 20 (refineries to the East of Montreal) shows the highest pollution levels while the lowest levels appear at station 1 (botanical garden). The pollution level at station 13 (Drummond Street) is higher than at station 12 (Ontario Street East).

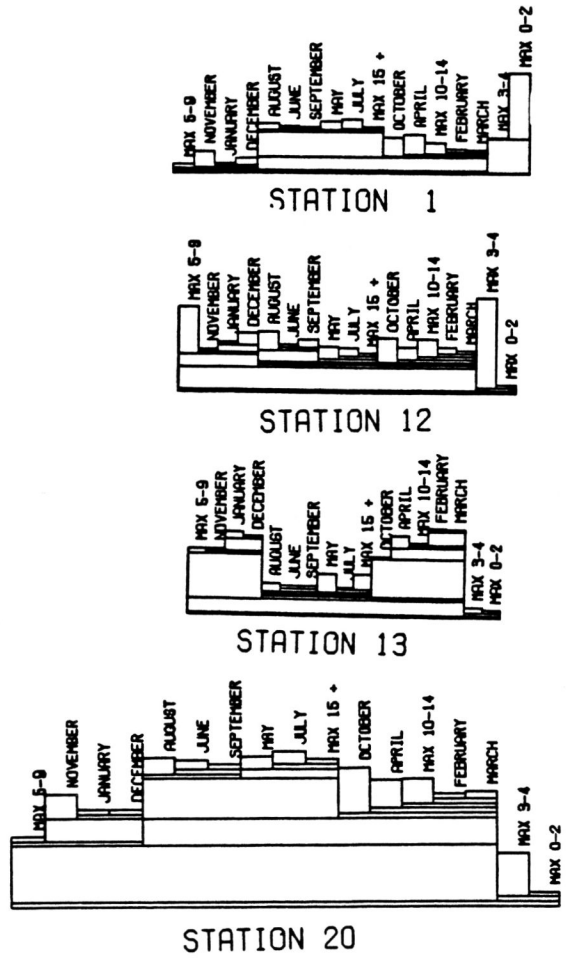

Figure 8: Representation of stations 1, 12, 13 and 20 by castles

CONCLUSION

In this article, we have sketched several graphical representation methods for multivariate data. Great progress has been made since the work of Playfair, 1801, one of the fathers of graphical representation in statistics. However all the problems have not yet been completely solved. Development of and experimentation with new methods must continue, always keeping in mind that graphics should provide a simple way of communicating the information contained in a set of data.

BIBLIOGRAPHY

Anderson, E. (1960): A Semi-Graphical Method for the Analysis of Complex Problems, Technometrics 2, 287-292.

Andrews, D.F. (1972): Plots of High-Dimensional Data, Biometrics, 28, 125-136.

Banfield, C.F. and Gower, J.C. (1980): A Note on the Graphical Representation of Multivariate Binary Data, Applied Stat. 29, 238-245.

Beniger, J.R. and Robyn, D.L. (1978): Quantitative Graphics in Statistics: A Brief History, Amer. Statistician 32, 1-11.

Bertin, T. (1967): Sémiologie graphique, Gauthier-Villars, Paris.

Bruckner, L.A. (1978): On Chernoff Faces, in Graphical Representation of Multivariate Data, ed. P.C.C. Wang, Academic Press, New York.

Caporal, P.M. and Hahn, G.H. (1979): Computer Offerings for Statistical Graphics, An Overview, Proc. Computer Science and Statistics, 13th Symposium on the Interface.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tuckey, P.A. (1983): Graphical Methods for Data Analysis, Wadsworth Int. Group, California.

Chernoff, H. and Rizvi, M.H. (1975): Effect on Classification Error of Random Permutations of Features in Representing Multivariate Data by Faces, JASA, 70, 548-554.

Chernoff, H. (1973): Using Faces to Represent Points in k-dimensional Space Graphically, JASA, 68, 361-368.

Cléroux, R. (1982): L'impact présent et futur de l'informatique en statistique, CIPS Review, 6, 18-21.

Cléroux, R., Roy, R. and Fortin, N. (1980): Air Pollution in Montreal: A Statistical Analysis of Sulphur Dioxide Data, Water, Air and Soil Pollution, 13, 143-156.

Cleveland, W.S. and Kleiner, B. (1974): The Analysis of Air Data Pollution from New Jersey and New-York, Annual Meeting ASA, St-Louis, Miss.

Everitt, B.S. (1978): Graphical Techniques for Multivariate Data, Heinemann Education Books, London.

Fienberg, S.E. (1979): Graphical Methods in

Statistics, Amer. Statistician 33, 165-178.

Flury, B. and Riedwyl, H. (1981): Graphical Representation of Multivariate Data by Means of Asymmetrical Faces, JASA, 76, 757-765.

Flury, B. (1980): Construction of an Asymmetrical Face to Represent Multivariate Data Graphically, Technical Report No. 3, Université de Berne, Dép. de Statistique.

Friedman, H.P., Farrell, E.J., Goldwyn, R.M., Miller, M. and Siegel, J.H. (1972): A Graphic Way of Describing Changing Patterns, Proc. Comp. Sci. and Statist., 6th. Annual Symposium on the Interface, Berkeley, Calif., 56-59.

Friedman, J.H., and Rafsky, L.C. (1981): Graphics for the Multivariate Two-Sample Problem, JASA, 76, 277-295.

Gascon, A. (1978): Méthodes graphiques d'analyse de données multidimensionnelles, Masters' thesis, Dép. d'informatique et de recherche opérationnelle, Université de Montréal.

Gnanadesikan, R. (1980): Graphic Data Analysis: Issues, Tools and Examples, Ann. Meeting Amer. Assoc. Adv. Sci., San Francisco.

Gnanadesikan, R. (1977): Methods for Statistical Data Analysis of Multivariate Observations, Wiley, New York.

Hartigan, J.A. (1975): Printer Graphics for Clustering, Journal Statist. Comput. Simul., 4, 187-213.

Hartigan, J.A. (1975): Clustering Algorithms, Wiley, New York.

Jacob, R.J.K. (1980): Correspondence on Fienberg (1979) Amer. Statistician 34, 252-253.

Kalence, K.W. and Kiviat, P.J. (1973): Software Unit Profiles and Kiviat Figures, ACM Perf. Eval. Rev. (sept.).

Kent, P. (1982): An Efficient New Way to Represent Multidimensional Data, Doctoral Thesis, Ecole Polytechnique Fédérale de Lausanne.

Kleiner, B. and Hartigan, J.A. (1981): Representing Points in Many Dimensions by Trees and Castles, JASA, 76, 260-276.

Kruskal, W. (1977): Visions of Maps and Graphs, Proc. Int. Symp. on Comput. Assisted Cartography, 25-36.

Kruskal, J.B. (1964): Non-Metric Multidimensional Scaling: a Numerical Method, Psychometrika 29, 115-129.

Lee, R.C.T., Slagle, J.R. and Blum, H. (1977): Triangulation Method for the Sequential Mapping of Points from N-Space to Two-Space, IEEE Trans. Comput. 26, 288-292.

Lin, C.H. and Chen, H.F. (1977): Representation-Space Transformation for the Display of Multivariate Chemical Information, Anal. Chem. 49, 1357-1363.

Pickett, R. and White, B.W. (1966): Constructing Data Pictures, Proc. 7th. Nat. Symp. on Information Display, 75-81.

Playfair, W. (1801): The Statistical Breviary, London.

Siegel, J.H., Goldwyn, R.M. and Friedman, H.P. (1971A): Iteration and Interaction in Computer Data Bank Analysis: A Case Study in the Physiologic Classification and Assessment of the Critically Ill, Comput. and Biomed. Res., 4, 607-622.

Siegel, J.H., Goldwyn, R.M. and Friedman, H.P. (1971B): Pattern and Process in the Evolution of Human Septic Shock, Surgery, 70, 232-245.

Wainer, H. (1974): The Suspended Rootogram and Other Visual Displays: an Empirical Validation, Amer. Statistician 28, 143-145.

Wainer, H. and Thissen, D. (1981): Graphical Data Analysis, Ann. Rev. Psychol., 32, 191-241.

Wakimoto, K. (1977): Tree Graph Method for Visual Representation of Multidimensional Data, Jour. Japan Statist. Soc., 7, 27-34.

Wakimoto, K. and Taguri, M. (1978): Constellation Graphical Method for Representing Multidimensional Data, Ann. Inst. Stat. Math. 30, 97-104.

Wang, P.C.C. (1978): Graphical Representation of Multivariate Data, Academic Press, New York.

Welsch, R.E. (1973): Graphics for Data Analysis, Comput. & Graphics, 2, 31-37.