United States
Environmental Protection
Agency

Environmental Monitoring Systems
Laboratory
Research Triangle Park NC 27711

Research and Development

EPA-600/S4-80-048 August 1982

EPA

# Project Summary

# Application of Cluster Analysis to Aerometric Data

Harold L. Crutcher, Raymond C. Rhodes, Maurice E. Graves, Beth Fairbairn, A. Carl Nelson, and Michael Symons

**The NORMIX data analysis program, which incorporates cluster analysis and multivariate statistical analysis routines, was modified and revised for use in a UNIVAC 1110 computer. The revised program was tested on three sample data sets and produced results in agreement with those from the original program. The NORMIX program was then used to evaluate and analyze eight sets of aerometric data from various sources. Comparison of the performance of NORMIX with two other cluster analysis algorithms, MIKCA and SAS CLUSTER, revealed that all three programs produce similar results in terms of hierarchical clustering, but NORMIX produces considerably more statistical evalu ation and information to the user. Thus NORMIX is recommended as the most useful cluster analysis program of these three.**

***This Project Summary was developed by EPA's Environmental Monitoring Support Laboratory, Research Triangle Park, NC, to announce key findings of the research project that is fully documented in a separate report of the same title (see Project Report ordering information at back).***

## Introduction

Pollutants in the environment pose an enduring threat to all living organisms and inanimate structures. It is continually necessary to assess this threat and to take remedial action. For such assessment, it is essential to monitor environmental conditions in order to provide information bases about the possible threats and their variation over time. Such comparisons permit reassessment of the average conditions, their expected variabilities, and any significant changes in those average conditions and variabilities.

To produce credible models and assessments of atmospheric pollution requires extensive, reliable aerometric data. Production of valid data bases requires adequate instrumentation and maintenance, representative exposure, competent personnel, excellent communications, and sufficient quality assurance and control systems in an ongoing updated observational program. To aid in the development of valid data bases and to extend methods of data analysis, five specific goals of this study on aerometric data clustering were defined:

1. Develop and document a validated and calibrated digital computer program for cluster analysis.
2. Extend the theory for clustering data.
3. Validate the data obtained.
4. Classify the data.
5. Demonstrate the application of the computer program to various types of data.

As a result of this project, an extensively developed computer program for cluster analysis is available to all users of the UNIVAC 1110 computer at the National Computer Center at the Environmental Protection Agency, Research Triangle Park (NCC-EPA/RTP), North Carolina.

Several clustering techniques that separate heterogeneous aerometric data sets into homogeneous groups were reviewed. It is sometimes difficult to state categorically that a datum belongs to a specified group or cluster, hence assignment or classification to a group of related data is made in a probabilistic sense. This report series illustrates the use of these clustering algorithms to grouped data from a large data base of observed aerometric and meteorological information. These grouped data can be interpreted by researchers and presented to decision-makers. For example, conditions accompanying pollutant episodes can be identified, and oftentimes specific pollutant sources can be identified.

Data from the Community Health Air Monitoring Program (CHAMP) were analyzed by the NORMIX clustering algorithm[1], as described in the three-volume Project Report. Volume I presents a detailed examination of the historical development of the NORMIX algorithm and its application to the CHAMP data set, plus descriptions of cluster analyses of data from the Los Angeles Catalyst Study (LACS) by the SAS[2] and MIKCA[3] programs. Volume II contains the modified NORMIX program with complete documentation, and Volume III contains more detailed examples and discussion of the application of the NORMIX algorithm to the CHAMP data.

Until the advent of clustering techniques and their algorithms, analyses of multivariate data were generally of the multiple regression type, factor analysis, or principal component analysis. Clustering analysis involves a hierarchical grouping of data. Some cluster analysis programs (notably NORMIX) also provide statistical estimates of the multimodal, multivariate distribution. Cluster analysis has been used on air pollution data to reveal cyclic patterns (over days, weeks, or seasons) and to identify local source effects. All of these relations are reflected in different combinations of values of the variables that cluster together. Because of the clustering in aerometric data, multivariate cluster analyses are useful as preliminary analytical and data validation techniques. The results of the report support the well-known inverse concentration relationship between ozone and the oxides of nitrogen, presumably due to their production timing and to their chemical interaction. These and many other relationships are indicated in tabular form and illustrated in some examples by diagrams in which data are plotted with overlying distribution ellipses in bivariate form, or by profile models of means and three-sigma limits for given measurements.

Data from monitoring activities cannot be considered as random samples from a single universe, but rather, such result from sampling mixtures of distributions, usually internally correlated within each group. As expected, these studies show that pollutant data depend on, or are highly correlated with, meteorological conditions. At a given site and with a given set of pollutant sources, the pollutant concentration at the site is heavily dependent upon the meteorological regime. Thus, the pollutant distributions will be a mixture of the distributions that result from the mixture of the meteorological regimes and the interaction of the pollutants over the time of the monitoring. Solar radiation is an effective agent in the meteorological regime, but these data were not available for inclusion in this study. This discussion suggests that the pollutant data distributions first be separated into meteorological regimes by cluster analysis and that these subsets then be evaluated individually by other approporiate techniques. Moreover, analysis involving prediction of future pollutant concentration distributions for each meteorological regime should also consider the probabilities of occurrence of each of the meteorological regimes. This will enhance the clustering and the classification of data.

Normality of distribution is not required for simple hierarchical clustering, but if statistical significance statements are to be made, or if statistical characteristics of the clusters are to be used, the normal distribution is quite useful. It is not necessary that exact normality be achieved, as the techniques used are sufficiently robust to accommodate considerable departure from normality.

The results are more reliable if individual element (variate) distributions are assumed to be normal or near normal during the application of clustering techniques. If the distributions are distinctly non-normal in appearance, various mathematical operations are available to transform the individual data so that the distributions of the transformed data may be described by the normal distribution prior to their entry into the heirarchical clustering scheme. If no prior information is available, the heirarchical clustering may be the only product of the operation, or it may be a prelude to further information extraction.

Because the NORMIX program has a substantial statistical basis with corresponding statistical assumptions and tests of significance, it was selected for further development in this study. The program, originally written in IBM FORTRAN IV language, was converted to the ASCII FORTRAN language requirement of the UNIVAC 1110 at NCC-EPA/RTP.

Figure 1 shows the general configuration of the expanded NORMIX program, detailing the NORMIX pre-processing options, the central NORMIX core algorithm, and the post-processing flow: Figure 2 presents the NORMIX flow chart. Documentation is available in supplementary and complementary reports, Volumes II and III, which are discussed later.

## Calibration and Validation of the Expanded Normix Program

The ability of the present revised version of the NORMIX program to produce results equivalent to those of Wolfe[1] (for the same data in the older program version and with a different computer) indicated that the revised version has been adequately calibrated. Program validation consisted of application over several types of data sets, not necessarily all aerometric. Data validation consisted of the isolation of outliers, if any, for examination and treatment. Both single and clustered outliers are identifiable in the hierarchical clustering as well as in the NORMIX processing. Since the NORMIX program uses the same hierarchical clustering algorithms as several of the other programs, it is not necessarily more useful for this purpose.

In order to demonstrate that the present version of the NORMIX program is available and works properly, 11 sets of data were used. These were:

1. A classical data set composed of measurements of petal and sepal lengths and widths (four variates) made by Anderson[4] and used by Fisher[5] to illustrate clustering and classification.
2. A synthetic bivariate data set from Wolfe[1].
3. A set of synthetic data consisting of three predetermined three-element data sets that could be expanded both in variances and distances between the centroids.

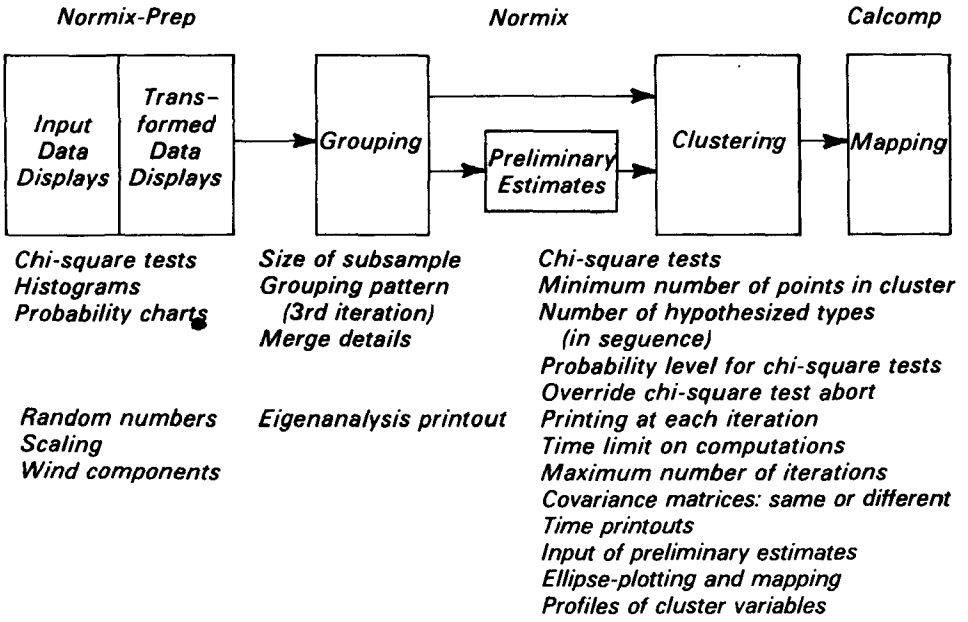



*Figure 1. NORMIX flow and options.*

4. A bivariate set of maximum and minimum temperatures for June and July at the Raleigh-Durham, North Carolina, airport, supplied by Professor Jerry Davis of the North Carolina State University at Raleigh.
5. A five-variate set of health-related data for trace metals.
6. A univariate set of river discharge data supplied by the U.S. Geological Survey.
7. A 12-variate set of data on new filters, 12 impurities, and trace metals.
8. A six-variate set of data on the physical and chemical characteristics of new filters.
9. A univariate set of acid precipitation data for each of several locations within the United States.
10. The LACS data set.
11. The CHAMP data set.

The first three sets of data mentioned above were processed by the NORMIX algorithm to calibrate the program

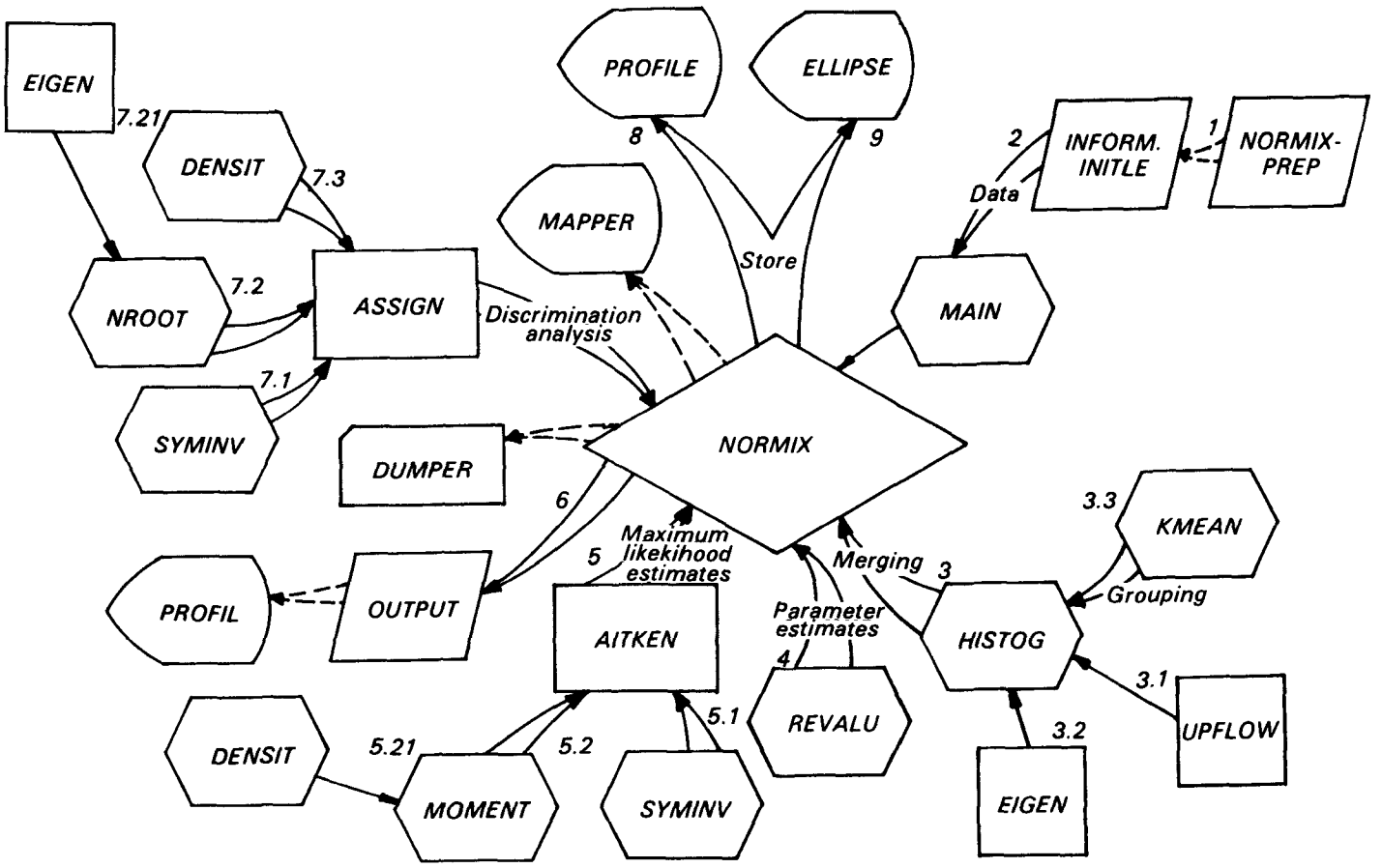



*Figure 2. NORMIX flow chart.*

conversion. The output of Sets 1 and 2 agreed with those of Wolfe. The output of Set 3 returned the original stipulated clusters prior to their being mixed.

The NORMIX program produces hierarchical mapping of the data, as do most of the other programs discussed in this report, although the metrics used may differ. Tree (branching or dendritic) diagrams may be prepared from the maps, which show the coalescence of data into clusters. The report contains such tree diagrams, which are not reproduced here. From such tree diagrams, outliers (either singletons or small groups) can be identified easily, as they are the last to enter a larger cluster or the last to join the total group. A lengthy discussion on the reading and interpretation of the diagrams is also available.

The presence of extreme outliers, as singletons or as minimum sized clusters, creates near-singularities in the data matrices, which halt a running computer program, produce slow convergence, or do not allow the program to converge to a solution. This phenomenon is characteristic of any program that uses convergence routines involving matrix calculations.

The extensive environmental data bases, LACS and CHAMP, were processed by means of cluster algorithms; SAS CLUSTER and MIKCA were used for the LACS data and NORMIX was used for the CHAMP data.

In order to study the effect of the use of automobile catalytic converters on aerometric parameters, the period of the LACS data necessarily had to include the periods before and after the 1975 introduction of these devices; the period selected was 1974 through 1978. The pollutant elements (variates) observed were suspended particulates (SP), ozone ($O_3$), nitrogen oxide (NO), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$), carbon monoxide (CO), and lead (Pb). The meteorological variates were temperature, wind speed, wind direction and traffic count. For this analysis, measurements of the variates SP, CO, Pb, wind speed, wind direction, and traffic count from two selected sites were used. These sites were on either side of the San Diego Freeway between the intersection of the freeway with the two boulevards, Sunset and Wilshire.

When more than one element is being observed and recorded, one of the elements may not be obtained for a particular observation time for various reasons. Effectively, in the multivariate sense, the omission of a single element requires the rejection of the entire observation from the data set under consideration. In some cases, it may be reasonable to merge one incomplete multivariate vector (observation) with another complementary incomplete vector from a nearby locale to obtain a usable complete vector. If this is done, this factor must be considered when interpreting the results

Here is an example of the two effects mentioned above: using only two sites from the LACS data bank, the number of available and useful hourly observations for the 1977 through 1978 period is about 3031, as compared to about 25,000 observations originally available from all eight sites of the LACS. The reader should consult Part 2 of Volume I of the Project Report for further details.

The periods of record at the CHAMP stations were relatively short, i.e. from September 1975 through November 1976 for Angwin, California, and Loma Linda, California, and from August 1974 through September 1976 for Magna, Utah. The data selected for use are for certain hours of the day, days of the week, weeks, and seasons.

The pollutant and meteorological data consisted of oxides of nitrogen ($NO_x$), calculated nitrogen oxide (NO), nitrogen dioxide ($NO_2$), sample nitrogen oxide (SNO), ozone ($O_3$), sulfur dioxide ($SO_2$), total hydrocarbons (THC), non-methane hydrocarbons (NMHC), temperature (T), dew point, winds, and atmospheric pressure (P). For this study, all winds were transformed to east-west and north-south components, positive from the west and south. An option in the program permits transformation of polar coordinates (wind direction and speed) to rectangular coordinates, along and at right angles to any preselected direction. The default option is the east-west and north-south configuration.

## Comparison of Three Clustering Programs.

Table 1 compares three clustering algorithms, SAS CLUSTER, MIKCA, and NORMIX. All three algorithms select clusters by an agglomerative rather than a divisive procedure, and the number of clusters to be examined must be stipulated. For each program, the recommended maximum number of cluster configurations is seven.

The reader and user of the Project Report will find a rather extensive discussion of clustering and data validation problems and uses for computer programs of clustering techniques. No one program satisfies all users. Some of the limitations of each program are discussed. The NORMIX program, being the most complex, produces much more informational output than do the SAS CLUSTER and MIKCA programs.

## Examples of Processing Output

Table 2 shows an intercomparison of selected pollutant data throughout the year for NO, $NO_x$, and $O_3$ at Angwin, California, Loma Linda, California, and Magna, Utah.

The information in Table 2 reveals that Angwin, California probably has the lowest levels of oxides of nitrogen and highest level of ozone, and the lowest variability of these three variates at the three stations. This is, of course, one reason why the Angwin, California, site was selected for monitoring and for this study. The large standard deviation and negative mean value for the Loma Linda, California $NO_x$ data reflect that the number added to low observed values (to ensure a minimum value of two before logarithms are taken was insufficiently large.

Figure 3, developed from NORMIX output information from Magna, Utah data, illustrates the 0.50 probability ellipses for wind speed and direction and the associated pollutant means, standard deviation, proportions, and the number of observations. Cluster numbers for each point are included to help assess the clustering efficacy. It must be remembered that the pollutant variables have been logarithmically transformed and numbers refer to such transformed data. The 0.50 probability ellipses are centered on the centroids of the plots of east-west wind components versus north-south wind components. The ellipses are for the wind components, but the cluster classifications are in terms of the eight variates. As previously noted, it is in the overlapping regions that errors of misclassification may occur for an individual datum. However, the statistical estimates are generally expected to provide the best estimates of the cluster configurations. This type of presentation is a projection of the total multidimensional ellipsoid onto the plane of the two selected variates. Any two variates can be selected by options provided in the program.

The program option that developed Figure 4 arranged the variate output in terms of the largest cluster with variate

*Table 1.* *Comparison of Capabilities of the SAS CLUSTER, MIKCA and NORMIX Algorithms*

| Property | SAS | MIKCA | NORMIX |
|---|---|---|---|
| Complexity | Low | Medium | High |
| Output Quantity | Minimum | Moderate | Extensive |
| Number of Input Data | 250 | 500 | 2000* |
| Limit to Number of Variables | 10 | 20 | 20* |
| Maximum Number of Clusters | 250 | 15 | 150 |
| Distance Options | 1 | 3 | 1 |
| Criteria Options | 1 | 9 | 1 |
| Hierarchical Clustering | Yes with Maps | No | Yes with Maps |

**May be increased if computer memory space permits. Computation time increases exponentially with the numbers of variates and observations.*

*Table 2.* *Intercomparison of Selected Pollutant Data Throughout the Year for NO, $NO_x$ and $O_3$ at Three Locations*

| | Transformed variates* | | | | | | |
|---|---|---|---|---|---|---|---|
| | NO | | $NO_x$ | | $O_3$ | | |
| Locations | Mean | Std. dev. | Mean | Std. dev. | Mean | Std. dev. | Number of observations |
| Angwin, California | 0.6981 | 0.0017 | 0.7002 | 0.0045 | 0.7127 | 0.0087 | 288 |
| Loma Linda, California | 0.7036 | 0.0130 | -3.0442 | 0.9576 | 0.7081 | 0.0198 | 122 |
| Magna, Utah | 0.7055 | 0.0164 | 0.7123 | 0.0216 | 0.7081 | 0.0090 | 324 |

**Values are in logarithmically (base e) transformed data originally in ppm.*

means of increasing in sequence. Again, as in all presentations such as this, the values of the other variates follow the sequence established by the largest cluster. The numbers below the minimum three-sigma limit, ranging from one through eight, identify the variates in order of their entry into each observational vector. The other numbers identify the three-sigma levels.

In Figures 3 and 4, it may be noted that weak southeast winds with a mean speed of approximately 4 km/h are associated with ozone readings lower than average and oxides of nitrogen and sulfur readings higher than average. Strong winds from the northwest, approximately 9 km/h, and from the south-southeast, approximately 12 km/h, again show the inverse relationship of ozone with oxides of nitrogen and sulfur. Clusters 1 and 3, with southeast and south-southeast winds, respectively, show the greatest temperature difference between the means, namely, 17.54°C. Further investigation might yield the reason for this temperature difference. Speculatively, this feature might be associated with seasonal characteristics or synoptic episodes.

## Conclusions

The applications of three clustering algorithms to aerometric data bases were compared in order of investigation: SAS CLUSTER, MIKCA, and NORMIX. The three routines produce similar results through the processing steps of hierarchical clustering and output of cluster means. Beyond that point, MIKCA appears to provide slightly more information than SAS CLUSTER.

NORMIX, as modified, produces considerably more information and guidance than either MIKCA or SAS CLUSTER. NORMIX is the recommended clustering program; a calibrated and tested program with full documentation, available as Volume II of this three-volume report series.

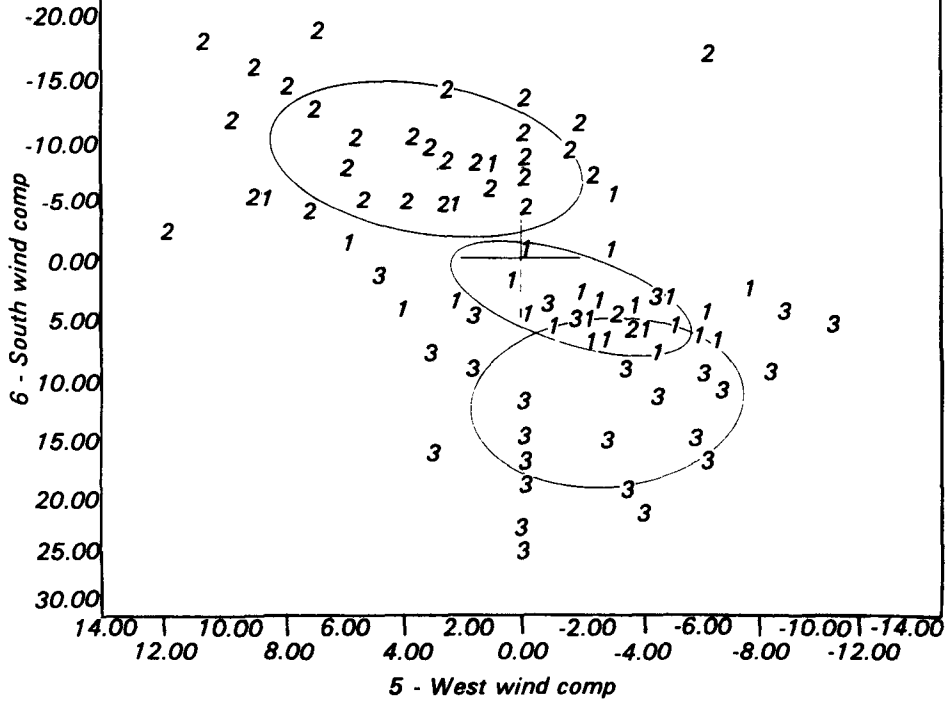
Many other clustering programs may be used, but only the aforementioned three have been examined in this study, and only NORMIX has been examined in detail. Of the three, only NORMIX provided complete statistical estimates of the multimodal, mulitvariate distributions. SAS CLUSTER is strictly hierarchical in grouping and mapping and uses this information as initial statistical estimates for further iterations to achieve maximum likelihood estimates.

Los Angeles Catalyst Study (LACS) data were analyzed by use of the two algorithms, SAS CLUSTER and MIKCA. The results were similar. Community Health Air Monitoring Program (CHAMP) data also were analyzed by means of the NORMIX program.

## References


1. Wolfe, John J. (1971) NORMIX 360 Computer Program. Naval Personnel and Training Research Laboratory, San Diego, California. Research Memorandum SRM 72-4. 125 pp.
2. Barr, A.J., J.H. Goodnight, J.P. Sale, and J.T. Helwig (1976) A User's Guide to SAS. Spanks Press.
3. McRae, D.J. (1973) MIKCA: A FORTRAN IV Iterative K-Means Cluster Analysis Program. CTB/McGraw Hill, Del Monte Research Park, Monterey, California. Revised by M.J. Symons, October 1973.
4. Anderson, Edgar (1953) The Irises of the Gaspe Peninsula. Bull. Amer. Iris Soc. 59:2-5.
5. Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugen. VII:11:179-188.

| | Cluster | | |
|---|---|---|---|
| Variable | 1 | 2 | 3 |
| 1 - NO | P = .354 | P = .333 | P = .313 |
| Mean | 0.72 | 0.70 | 0.70 |
| Standard deviation | 0.03 | 0.00 | 0.00 |
| 2 - NOX | | | |
| Mean | 0.74 | 0.70 | 0.70 |
| Standard deviation | 0.03 | 0.00 | 0.01 |
| 3 - Ozone | | | |
| Mean | 0.70 | 0.71 | 0.71 |
| Standard deviation | 0.01 | 0.00 | 0.00 |
| 4 - TS | | | |
| Mean | 0.78 | 0.72 | 0.71 |
| Standard deviation | 0.07 | 0.03 | 0.01 |
| 5 - West wind comp | | | |
| Mean | -1.70 | 3.13 | -2.86 |
| Standard deviation | 3.41 | 4.42 | 3.86 |
| 6 - South wind comp | | | |
| Mean | 3.40 | -8.32 | 12.12 |
| Standard deviation | 4.12 | 5.56 | 6.00 |
| 7 - Temperature | | | |
| Mean | 0.47 | 14.85 | 18.01 |
| Standard deviation | 3.49 | 5.59 | 6.22 |
| 8 - Dew point | | | |
| Mean | -5.60 | -1.31 | -3.83 |
| Standard deviation | 3.36 | 2.95 | 3.02 |




West vs South wind - probability level = 0.50

*Figure 3. Magna, Utah, Day 3, 0.50 probability ellipses of the west-east and south-north wind components for three cluster types.*

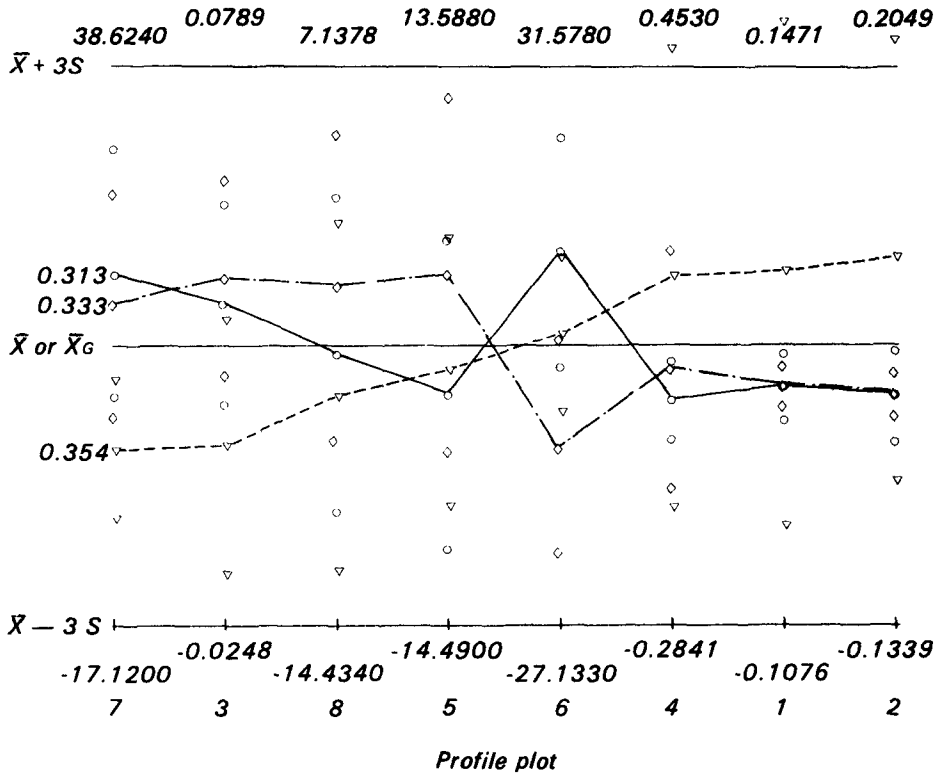



***Figure 4.*** *The means and three-sigma limits for each variate of the three data clusters of Figure 3 are represented by triangles (cluster 1), diamonds (cluster 2) and circles (cluster 3), respectively. The variate numbers along the abscissa refer respectively to: 1, NO; 2, NOx; 3, O3; 4, TS; 5, W component of wind; 6, S component of wind; t, temperature; and 8, dew point. The sequence of variates is determined by the value of their logarithms for cluster 1 (lowest to highest). The other numbers refer to the three sigma limit values for each variate.*

☆USGPO: 1982 — 559-092/0494

*Harold L. Crutcher is a private consultant at 35 Westall Avenue, Asheville, NC 28804; the EPA author* ***Raymond C. Rhodes*** *(also the EPA Project Officer, see below) is with the Environmental Monitoring Systems Laboratory, Research Triangle Park, NC 27711; Maurice E. Graves is with Northrop Services, Inc., Research Triangle Park, NC 27709; Beth Fairbairn and A. Carl Nelson are with PEDCo Environmental, Inc., Durham, NC 27701; Michael J. Symons is with the University of North Carolina, Chapel Hill, NC 27514.*

*The complete report consists of three volumes, entitled "Application of Cluster Analysis to Aerometric Data:"*

*"Volume I. Part 1—Clustering, Validation, and Classification of Data; Part 2—Investigation and Report of Cluster Analysis," (Order No. PB 82-226 432; Cost: $13.50, subject to change)*

*"Volume II. Part 3—Modifications and Options Applied to Wolfe's NORMIX 360 Cluster Analysis Program," (Order No. PB 82-226 440; Cost: $16.50, subject to change)*

*"Volume III. Part 4—Separation of Environmental Data Into Clusters by the NORMIX Program," (Order No. PB 82-226 457; Cost: $10.50, subject to change)*

*The above reports will be available only from:*

*National Technical Information Service*
*5285 Port Royal Road*
*Springfield, VA 22161*
*Telephone: 703-487-4650*

*The EPA Project Officer can be contacted at:*

*Environmental Monitoring Systems Laboratory*
*U.S. Environmental Protection Agency*
*Research Triangle Park, NC 27711*

United States
Environmental Protection
Agency

Center for Environmental Research
Information
Cincinnati OH 45268

Postage and
Fees Paid
Environmental
Protection
Agency
EPA 335

Official Business
Penalty for Private Use $300