United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-81-031e
September 1981

Air

EPA

Evaluating Simple Oxidant Prediction Methods Using Complex Photochemical Models

Cluster Analysis Applied To Urban Ozone Characteristics

EPA-450/4-81-031e

Evaluating Simple Oxidant Prediction Methods Using Complex Photochemical Models

Cluster Analysis Applied To Urban Ozone Characteristics

EPA Project Officer: Edwin L. Meyer, Jr.

Prepared by

U.S. Environmental Protection Agency
Office of Air, Noise and Radiation
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

September 1981

EXECUTIVE SUMMARY

Control of urban ozone pollution poses a unique problem because ozone is not directly emitted into the atmosphere by anthropogenic sources. Rather, it results from atmospheric photochemical reactions involving hydrocarbons and nitrogen oxides. Because the reactions involved in ozone formation take several hours to produce maximum ozone levels, the dispersion during such periods results in ozone's being a regional, rather than a local, problem. Thus, the processes involved in determining the location and severity of ozone concentrations include the temporal and spatial characteristics of the emission rates of the two precursor species, transport and dispersion by meteorological and topological effects in the region, and photochemical reactions dependent on solar intensity, temperature, and the like.

Currently, the only satisfactory way to characterize the temporal and spatial nature of ozone pollution in an urban area is through the use of complex mathematical models. Because of the complexity of these models, the costs associated with data gathering and their exercise on a computer can be substantial. Moreover, considerable expertise is needed to successfully mount a full-scale study of an urban area. It is thus fruitful to search for ways in which knowledge gained in one application can be transferred to application in another urban area. Specifically, if two urban areas can be shown to have sufficiently similar characteristics with respect to their ozone problems, a control strategy developed for one through application of a complex model could also be applied to the other, thus potentially obviating the need for a second costly study.

This report covers two exploratory studies that apply multivariate clustering techniques to the identification of similarities between urban areas. The results showed promise in assigning urban areas to distinct, relatively homogeneous classes; however, no clear-cut classification was achieved. The more qualitative of the two techniques showed a greater ability to classify areas into well-defined groups. The results indicate that further work is needed to refine the choice of classificatory variables, and that other techniques might be applied with greater success.

CONTENTS

ILLUSTRATIONS

TABLES

I INTRODUCTION

Ozone is unique among regulated air pollutants in that it is not directly emitted into the atmosphere from anthropogenic sources. Rather it results from atmospheric photochemical reactions involving hydrocarbon (HC) and nitrogen oxide (NO_x) precursors, which are emitted in varying amounts by industrial, utility, and automotive sources. Levels of atmospheric ozone thus depend not only on the usual factors of atmospheric transport and dispersion and on the amount of pollutant emitted, but also on the relative amounts and spatial distribution of emissions of two precursors and on the level of solar radiation necessary to initiate the ozone-producing reactions. The speed as well as the extent of the reactions depend on the ratio of HC to NO_x and on the level of solar radiation.

Because the ozone-producing reactions take several hours to produce the maximum amounts of ozone, by the time this maximum has been reached much atmospheric dispersion has taken place. Thus, ozone tends to be a regional, rather than a local, problem. Furthermore, since a period of maximum ozone concentration depends on the amount of solar radiation available to sustain the photochemical reactions leading to its production, the highest concentrations of ozone are reached during summer and early fall when higher insolation is observed.

The regional nature of the problem and the lack of a direct source-receptor relationship make the direct control of ozone concentrations, using emissions control strategies, more difficult. Currently, the concentration level at which the NAAQS is set is 0.12 ppm, and many urban areas exceed this level more than the prescribed one time per year. However, control of HC or NO_x emissions, or both, does not necessarily produce corresponding reductions in ozone (EPA, 1977a). In fact, the benefit to be derived from controlling a given set of sources in an urban area depends on current levels of HC and NO_x as well as on the location of those sources relative to the location of observed ozone maxima. To assess potential benefits of different strategies, many methods have been developed and applied in recent years.

The simplest of these methods is that of proportional rollback, which assumes that a reduction in HC emissions will result in a proportional reduction in ozone. As pointed out above, this approach does not work,

not only because of the influence of NO_x levels on ozone-producing reactions, but also because the relationship between HC and ozone is nonlinear. Other rollback methods that take nonlinearity into account (e.g., the Appendix J method, 40 CFR) also fail because of their lack of consideration of NO_x.

More complex methods, that account for the dependence of ozone concentrations on both hydrocarbon and nitrogen oxides (EPA, 1977a), are more successful in describing or predicting the results of potential ozone control strategies. However, these models concentrate on the chemistry of the problem and do not account for the spatial relationships between emissions sources and transport and dispersion phenomena in the region of interest. To include all aspects of the problem, simulation models have been developed that account for emissions and their spatial and temporal relationships, atmospheric photochemical reactions, and atmospheric transport and dispersion. Such models are required to solve a large and complex set of equations describing all of the atmospheric phenomena listed above. Moreover, these equations are solved many times through a series of time steps to yield temporally- and spatially-averaged ozone concentrations.

Such photochemical simulation models are computationally extremely complex, and for use in simulating a major urban area they require access to a large computer. In addition, because of the complexity of the computer programs, a knowledge of atmospheric pollution processes and computer programming is necessary. An additional characteristic of these models that discourages their use for many potential ozone problems is the cost of setting up and running the program. Before applying such a model to an urban area, an extensive data base containing emissions and meteorological data on a temporally- and spatially-resolved basis must be developed.

Because of the cost and the difficulties involved in applying complex photochemical models to a large number of urban areas, it seems fruitful to investigate ways in which knowledge and experience gained by model application to one city could have transfer value when a second city is being considered. To that end, we have applied multivariate clustering techniques to pertinent emissions, meteorological, and ozone-level data from several cities in the United States. The objective of the study was to determine whether urban areas could be objectively classified by characteristics relevant to photochemical pollution. Identification of city groups with similar characteristics would permit a small number of prototypical cities to be subjected to detailed analysis using complex photochemical models. Results obtained for one city within a group could be used in evaluating possible control strategies for the rest of the group. In addition, the performance of a model when applied to a prototypical city could be used as an indicator of its likely performance when applied to other cities in the same group.

The work described in this report covered two essentially exploratory studies using different clustering techniques: first, a study employing profile analysis to identify similarities between 29 urban areas (chapter II), and second, a further analysis that applies a hierarchical clustering technique to data from 45 urban areas (chapter III).

Cluster analysis comprises a set of mathematical techniques that are used to examine and develop underlying structure in multivariate data. The techniques range from fairly mathematical to almost purely descriptive, and have been applied to an extremely wide range of data sets (Hartigan, 1975). Clustering techniques can be thought of as a qualitative analog of regression analysis in terms of describing structural content of multivariate data. In carrying out regression analysis a well-defined mathematical model of the data structure is used; however, in cluster analysis one frequently has no preconceived notions about an appropriate model. Thus, whereas the goal of regression analysis is to estimate the parameters of the model, the purpose of cluster analysis is often merely to see whether or not the data fall into any reasonably well-defined groups.

II CLASSIFICATION OF URBAN AREAS BY PROFILE ANALYSIS

In this study the urban areas chosen were those where the ozone NAAQS was exceeded by more than 100 percent during the period 1974 to 1976 (EPA, 1977b). The areas included are shown in table 1. For classification variables, five meteorological and two emissions-related variables were chosen. These were:

- Summer morning mixing height. This quantity was chosen as a measure of the volume into which morning emissions are injected. These emissions are mainly responsible for ozone formation in the afternoon.

- The difference between summer afternoon and summer morning mixing height. This variable represents the degree to which morning emissions are diluted by the increase in the depth of the mixed layer.

- Summer afternoon wind speed. This measure represents the dilution of the pollutant cloud in the afternoon, when ozone reaches its peak in some areas.

- Normal daily July maximum temperature. Ozone formation is a strong function of temperature, and areas having high maximum temperatures are expected to have a greater potential for ozone formation.

- Mean daily July solar radiation. Ozone is formed photochemically, and the amount of solar radiation is an important measure of its formation potential.

- Ratio of hydrocarbon to nitrogen oxides emissions. This ratio affects the rates of atmospheric photochemical reactions and the amount of ozone that can be formed.

- Percentage of nitrogen oxides from transportation sources. This variable was chosen as a surrogate for the mix of point and mobile sources in the area. This factor can have important effects on ozone impacts.

TABLE 1. URBAN AREAS INCLUDED IN THE PROFILE ANALYSIS

Hartford-New Haven, Connecticut
Philadelphia, Pennsylvania
Chicago, Illinois
Milwaukee, Wisconsin
Houston, Texas
Denver, Colorado
San Francisco Bay Area, California
Fresno, California
Boston, Massachusetts
Northern New Jersey
District of Columbia
Erie, Pennsylvania
Richmond, Virginia
Newport News, Virginia
Huntsville, Alabama
Tampa, Florida
Louisville, Kentucky
Nashville, Tennessee
Kingsport, Tennessee
Detroit, Michigan
Minneapolis-St. Paul, Minnesota
Cincinnati, Ohio
Cleveland, Ohio
Baton Rouge, Louisiana
Dallas, Texas
Wichita, Kansas
St. Louis, Missouri
Salt Lake City, Utah
Phoenix, Arizona

The values of the variables for the different urban areas (mixing heights and wind speeds) were taken from Holzworth (1972). the Climatic Atlas of the United States 1968 and the 1973 Emissions Trends report (EPA, 1974). The meteorological data were interpolated from the maps given in the two compilations.

Profile analysis, as a cluster analysis technique, falls into a more qualitative class and is very simple in both concept and execution. First a set of axes, one for each variable, is laid out. Scales on these axes are chosen so that the ranges of the variables over the complete set of urban areas (or "cases") cover approximately equal lengths. Each case is then plotted on each axis according to its value for that variable. A "profile" for each case can then be constructed by joining all of the plotted points for that case. The similarities between the profiles can then be used to judge similarities between urban areas. In addition, points of difference can be identified by discrepancies between profiles. Using this technique, we identified five clusters of cities, which are illustrated in figures 1 through 5.

In figure 1, Los Angeles and San Francisco are identified as forming a fairly good cluster of coastal California urban areas. Some discrepancy is noted in July maximum temperatures, but the value obtained from the climatic atlas has more marine influence than do the values obtained in the area having the highest ozone concentration. In figure 2, a cluster of midwestern urban areas consisting of Chicago, Milwaukee, Minneapolis-St. Paul, Detroit, and Cleveland is identified. Although their climates are predictably similar, the emissions-related variables also reinforce the clustering. Figure 3 shows a group of urban areas, all of which are on the eastern seaboard of the United States, stretching from Connecticut to Virginia. In figure 4 we have a cluster of eastern United States cities consisting of Washington, D.C., Louisville, Nashville, Kingsport, Cincinnati, and St. Louis. These cities can be differentiated from those in the previous cluster on the basis of lower wind speeds, which are presumably related to greater distance from the ocean. This cluster seems to be closer in terms of meteorology than it is in terms of emissions-related variables. Figure 5 shows profiles of three urban areas located near the Gulf of Mexico: Houston, Tampa, and Baton Rouge. While these cities do not form a particularly tight cluster, the similarities between them are evident.

In figure 6 we show an example of three profiles that, though they might be expected to show similarities, in fact show large discrepancies. These profiles are for Denver, Phoenix, and Salt Lake City. The largest discrepancies are for July maximum temperature, July insolation, and the ratio of HC to NO_x emissions.

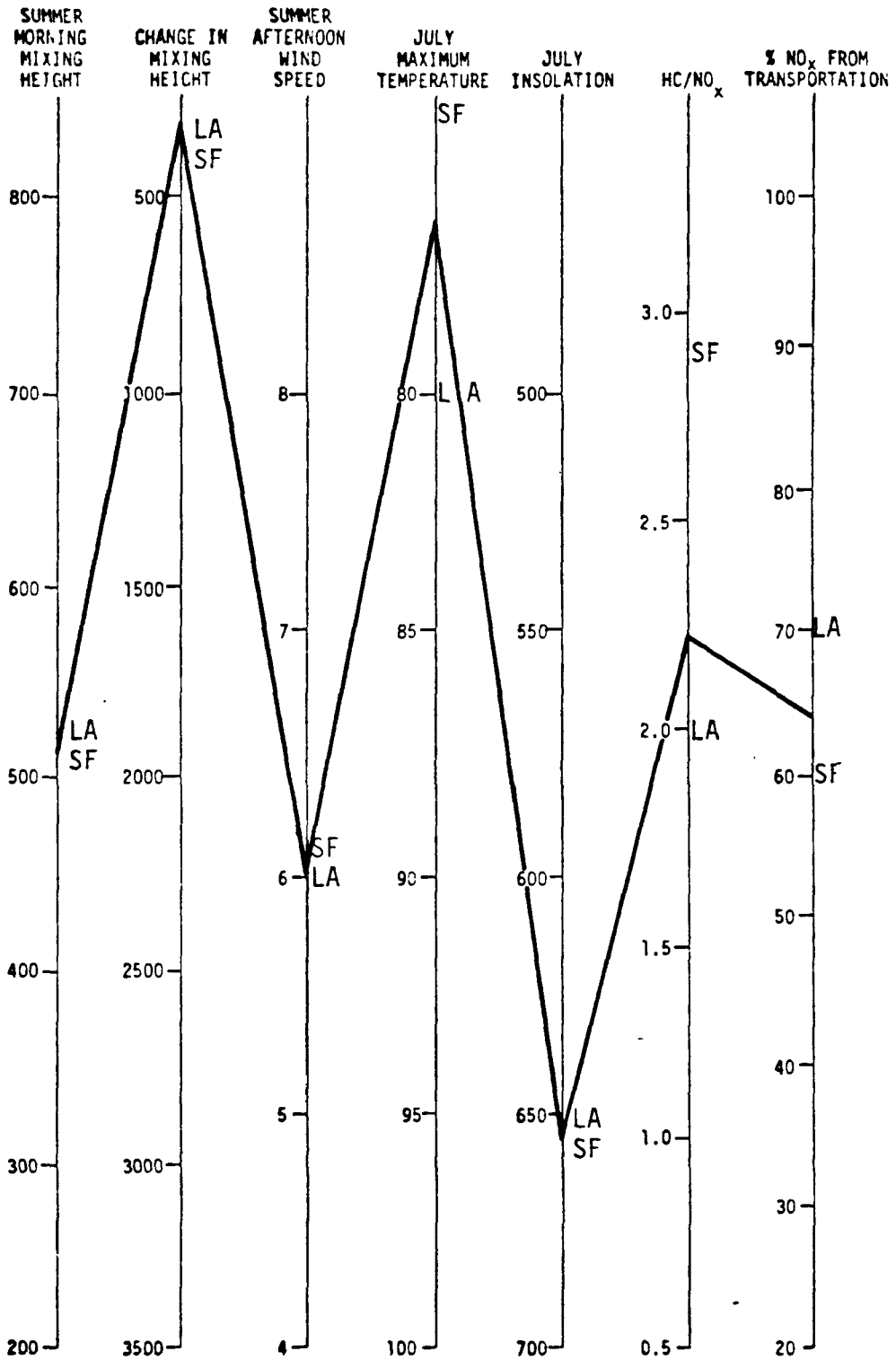

FIGURE 1. PROFILE OF CLUSTER 1

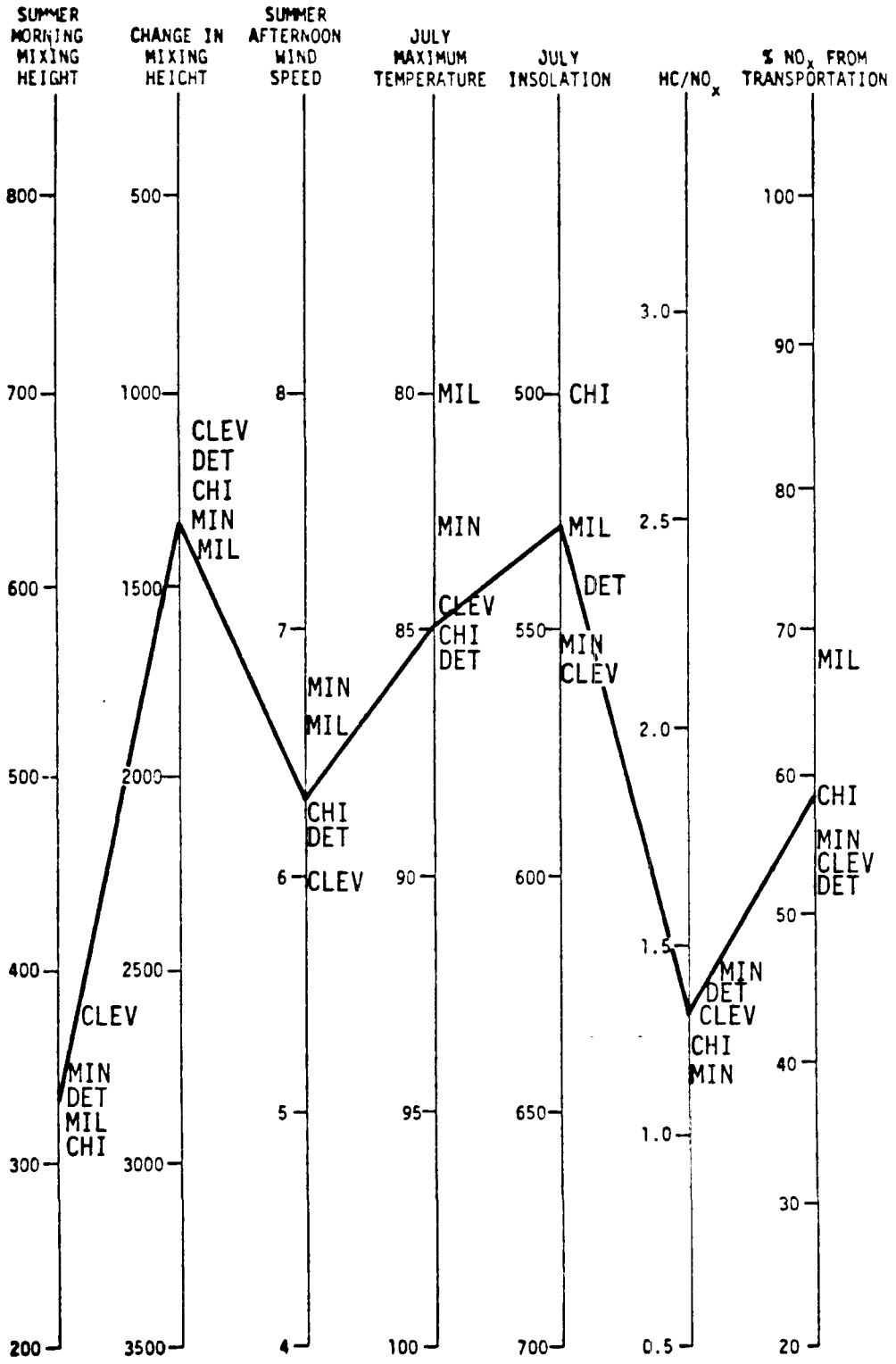

FIGURE 2. PROFILE OF CLUSTER 2

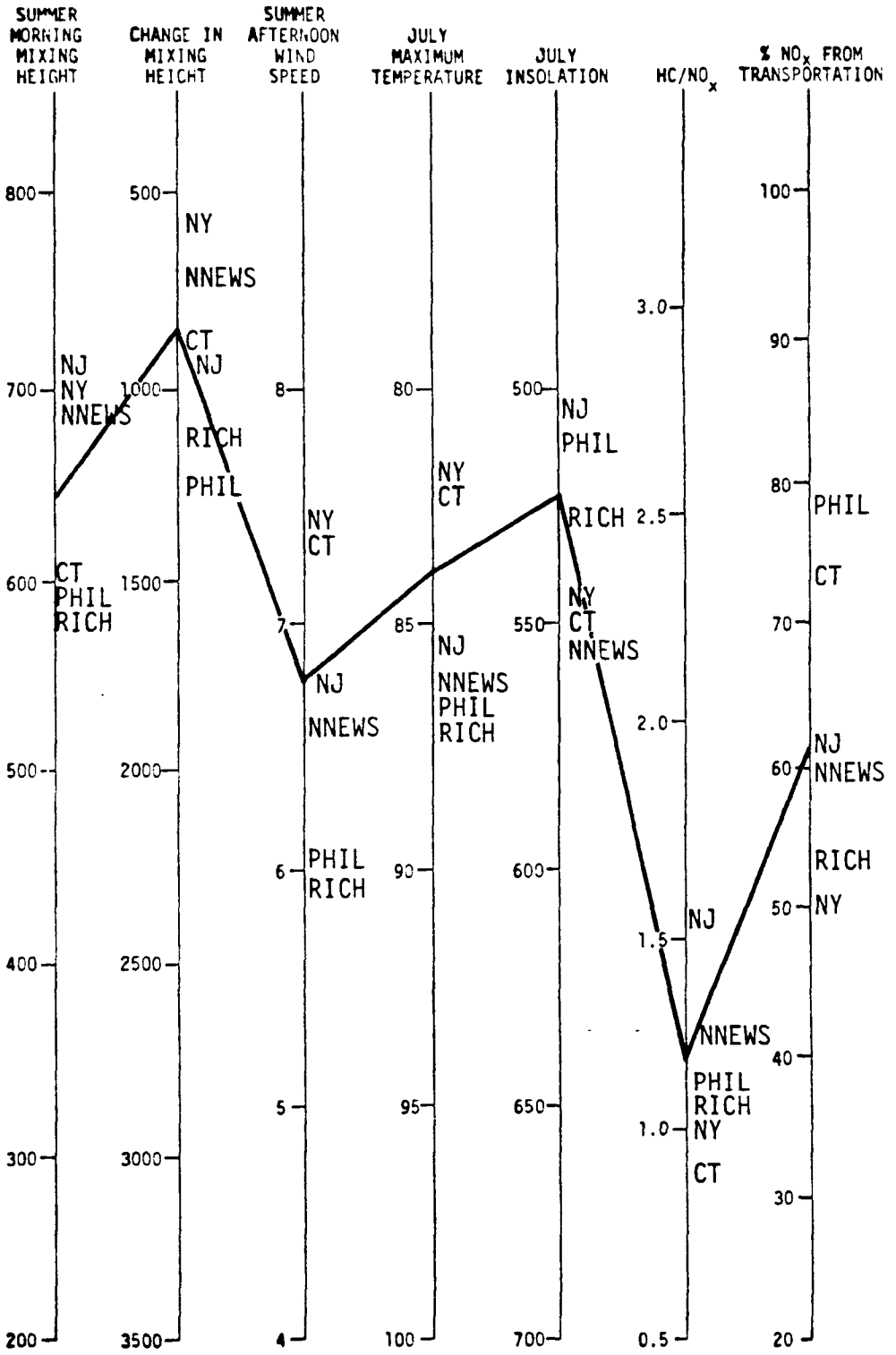

FIGURE 3. PROFILE OF CLUSTER 3

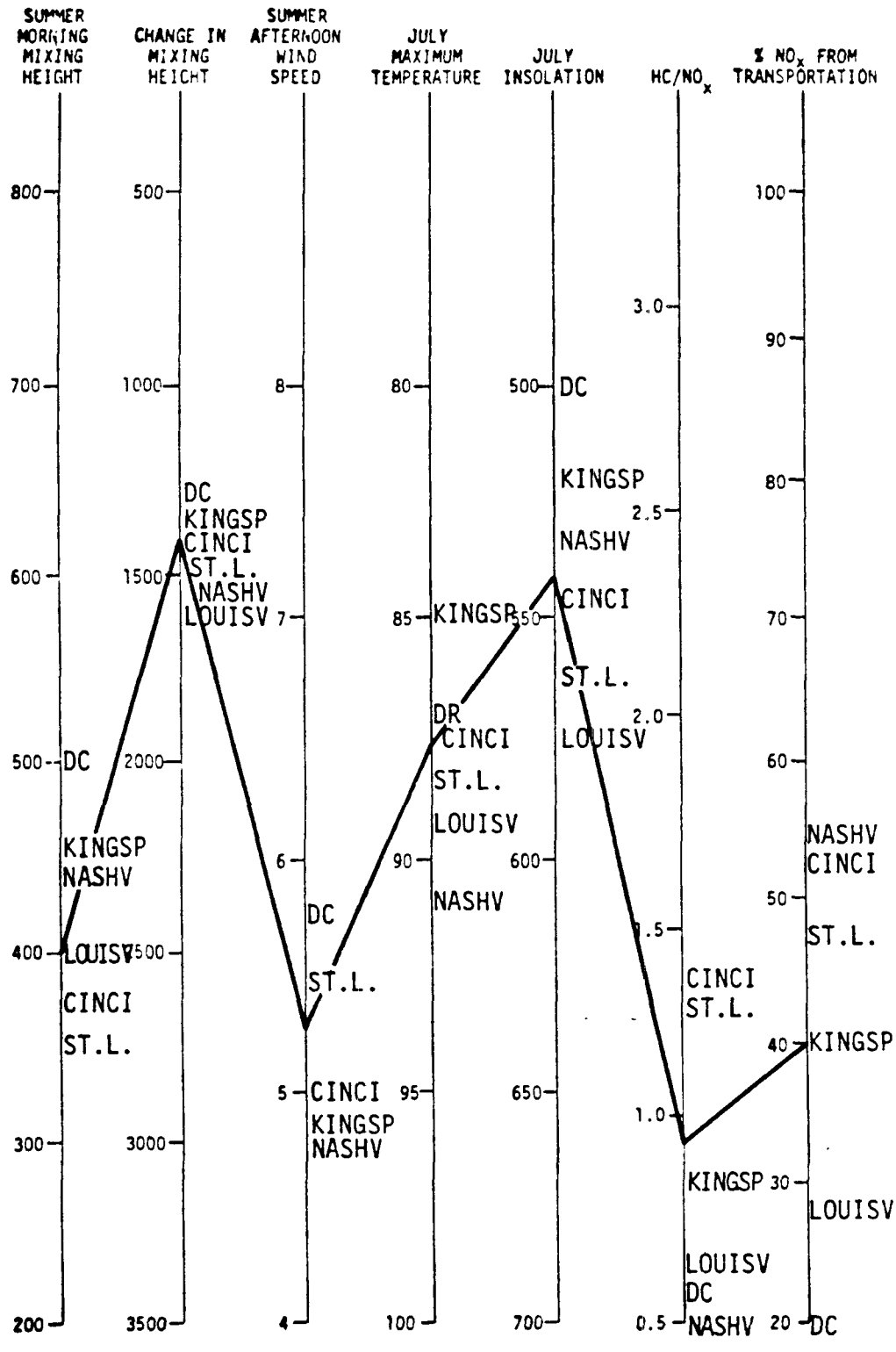

FIGURE 4. PROFILE OF CLUSTER 4

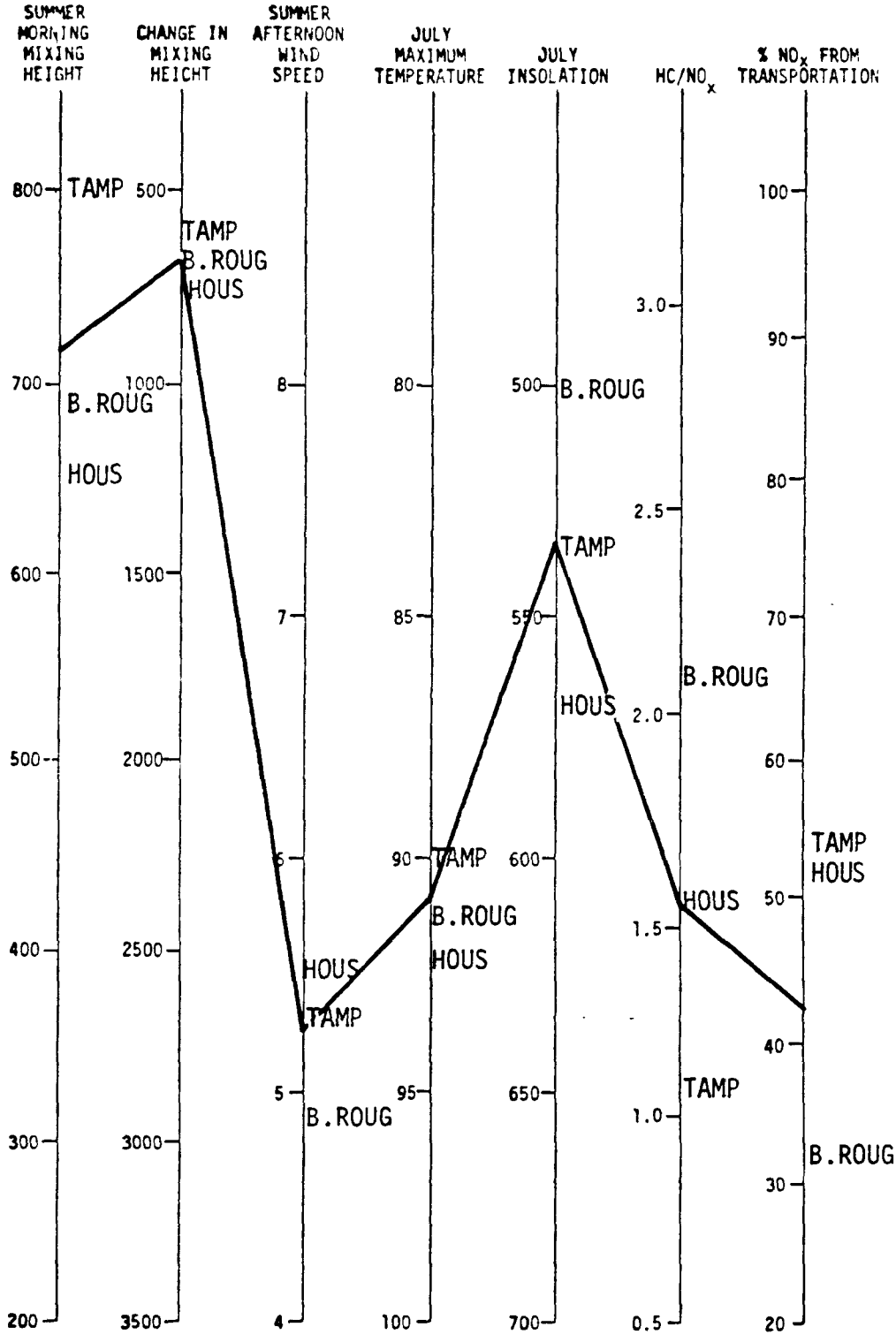

FIGURE 5. PROFILE OF CLUSTER 5

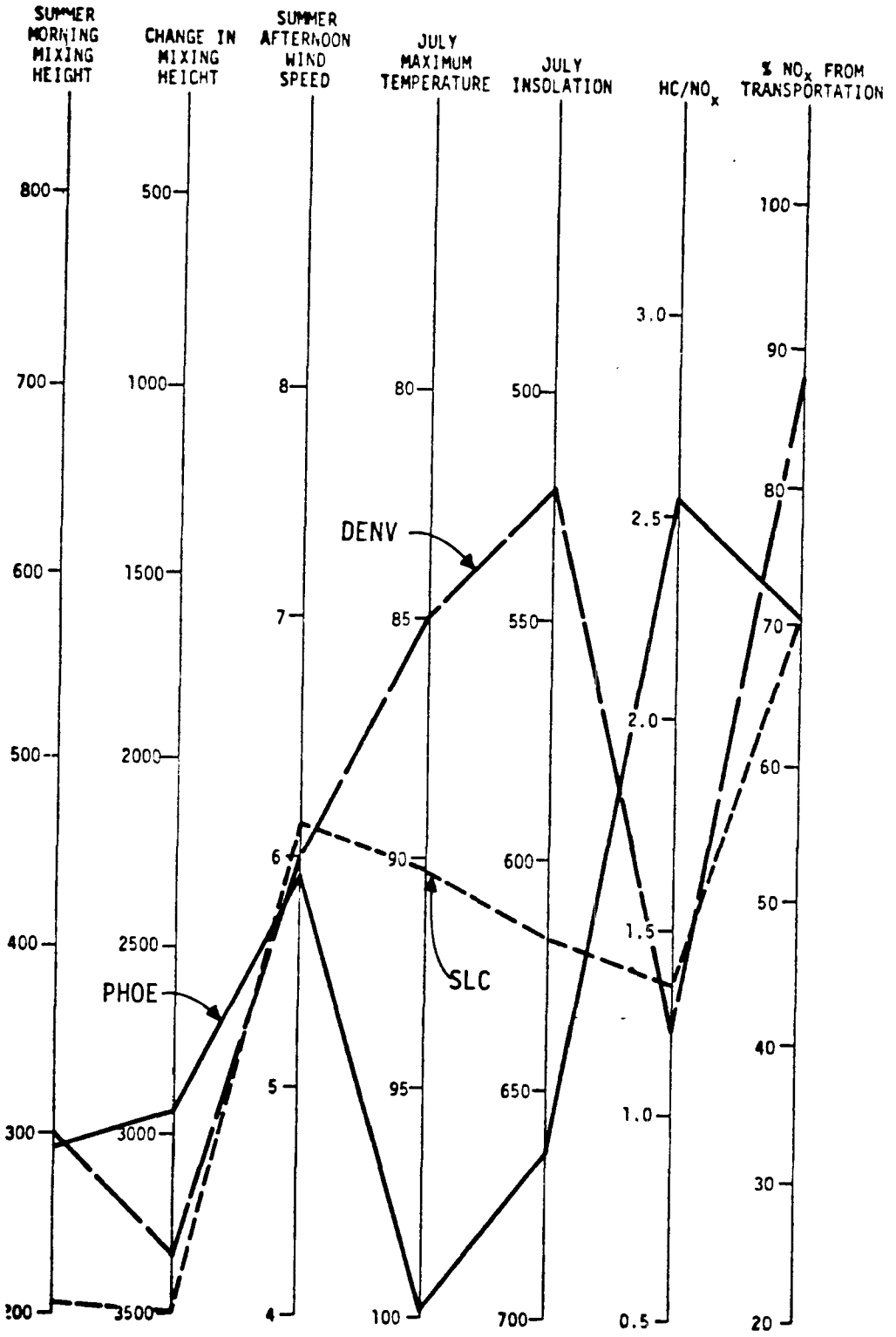

FIGURE 6. PROFILES OF DENVER, PHOENIX, AND SALT LAKE CITY

This preliminary analysis demonstrated the feasibility of grouping urban areas according to the factors contributing to oxidant problems. However, profile analysis gives no information of a quantitative nature about the degree to which cases in a given cluster resemble each other and differ from cases in other clusters. Moreover, since the clusters are identified by visual inspection of profiles, there is an arbitrary element in the selection of cases. We therefore carried out a further study in an attempt to achieve a more quantitative clustering and to consider more variables in the clustering process.

III CLASSIFICATION OF URBAN AREAS BY HIERARCHICAL CLUSTERING

For this further analysis, we examined data for 45 urban areas. They were selected as follows: First, we took those major urban areas that requested an extension to 1987 of their attainment date for the ozone NAAQS (Federal Register, 44, 65667). Of these, we eliminated Wilmington, Delaware, because some of the required data were not available. To this list of urban areas we added six more, to have more comprehensive geographical coverage of the United States. The urban areas included are shown in table 2 and figure 7.

Three types of data were compiled for each urban area: emissions, climatological, and ozone levels. Emissions data for each area, obtained from the National Emissions Data System, were taken for each county within that area. Three emissions variables were included: total HC emissions, the ratio of HC to NO_x emissions, and CO emissions from transportation sources. The HC and NO_x emissions influence the ozone-producing chemical reactions as detailed above and, thus, should be important in classifying urban ozone problems. We included CO emissions from transportation sources as a surrogate for vehicle miles traveled in an area. The amount of transportation-related emissions is an important facet of urban ozone problems.

Climatological data were again obtained from Holzworth (1972) and the Climatic Atlas of the United States, 1968. Data were interpolated from maps or taken from tabular compilations in these documents. Since ozone is a regional problem, regional climatology is likely to be more apposite; it therefore seems appropriate to use data interpreted on a large-scale rather than local-scale, climatology. The climatological temperature data used for the analysis were June, July, August, and September maximum temperatures, the annual maximum temperature, and the average maximum for June through September. This late-summer period generally produces the highest ozone concentrations. Three variables related to the amount of sunlight at each location were obtained: the total hours of insolation in the summer months, the average daily summer insolation, and the average percent of cloud cover.

Since the ozone-producing reactions are initiated and sustained by sunlight, ozone formation should be sensitive to the amount of sunlight incident at a specific location. Average summer morning and afternoon

TABLE 2. URBAN AREAS INCLUDED IN CLUSTER ANALYSIS

1	Allentown, PA	25	New York, NY
2	Baltimore, MD	26	Philadelphia, PA
3	Boston, MA	27	Phoenix, AZ
4	Bridgeport CT	28	Pittsburgh, PA
5	Butte, MT	29	Portland, OR
6	Chicago, IL	30	Providence, RI
7	Cincinnati, OH	31	Richmond, VA
8	Cleveland, OH	32	Sacramento, CA
9	Dallas, TX	33	Salt Lake City, UT
10	Dayton, OH	34	San Bernardino, CA
11	Denver, CO	35	San Diego, CA
12	Detroit, MI	36	San Francisco, CA
13	Fresno, CA	37	Scranton, PA
14	Hartford, CT	38	Seattle, WA
15	Houston, TX	39	Springfield, MO
16	Indianapolis, IN	40	St Louis, MO
17	Kansas City, KN	41	Trenton, NJ
18	Los Angeles, CA	42	Ventura - Oxnard, CA
19	Louisville, KY	43	Washington, DC
20	Miami, FL	44	Worcester, MA
21	Milwaukee, WI	45	Youngstown, OH
22	Minneapolis, MN		
23	New Haven, CT		
24	New Orleans, LA		

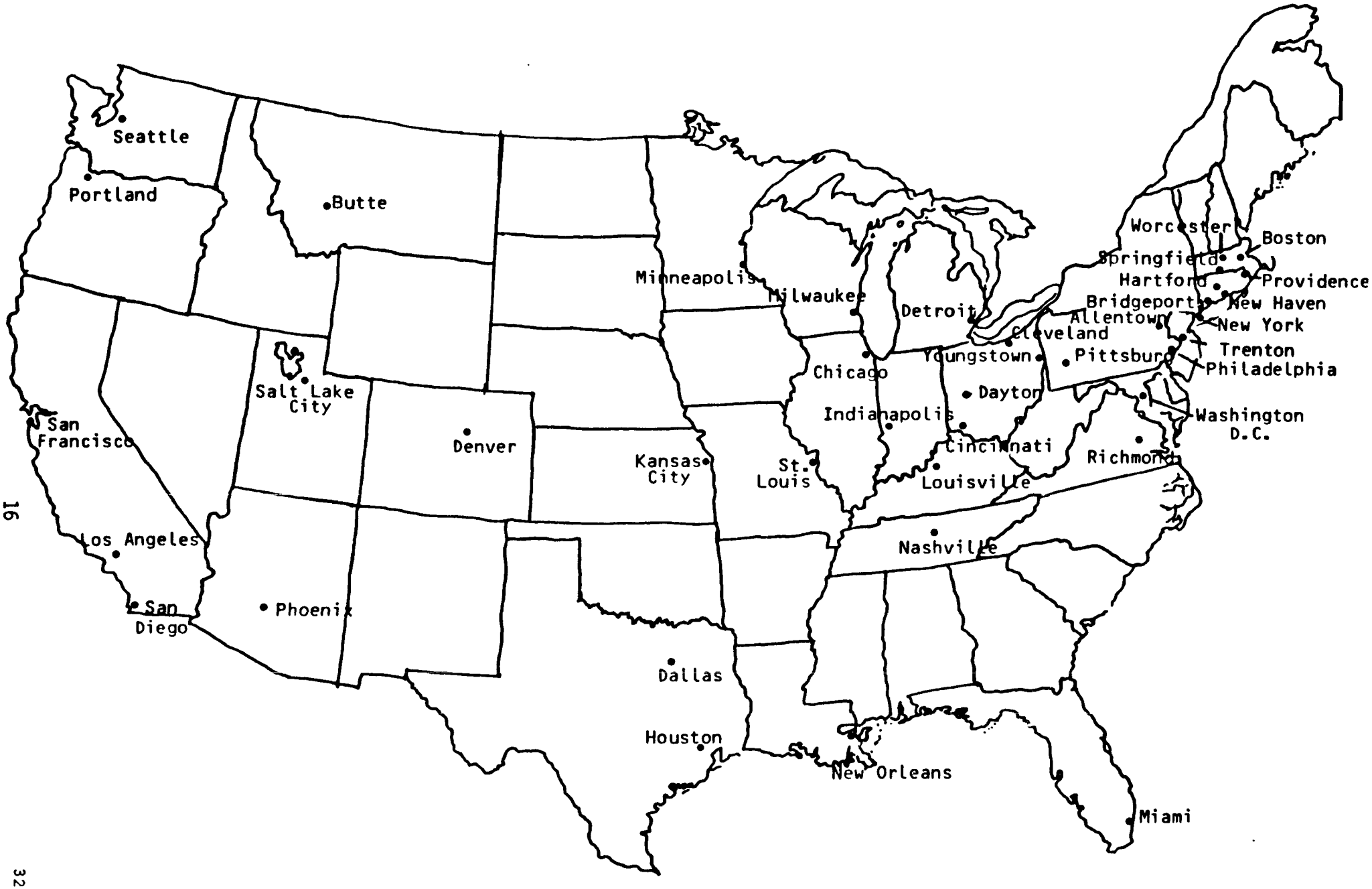

FIGURE 7. URBAN AREAS INCLUDED IN HIERARCHICAL CLUSTER ANALYSIS

wind speeds were obtained because a higher average wind speed should favor dispersion of precursor emissions and thus limit the concentrations of ozone that can be formed. Average summer morning and afternoon mixing heights were also recorded, as well as the change in average mixing height from morning to afternoon. The height of the mixing layer gives a measure of the effective volume into which emissions are discharged, and the concentrations reached are to a first-order approximation inversely proportional to this volume. Moreover, the change in mixing height is a measure of the dilution of morning precursor emissions. In some hot, interior locations a low morning inversion is largely dissipated by afternoon, whereas in a coastal location an inversion can persist into the afternoon, trapping pollutants into a concentrated layer near the ground.

The ozone data used in this study were obtained from the Monitoring and Data Analysis Division of the Office of Air Quality Planning and Standards. They consisted of the maximum and second highest ozone level, the average ozone level, and the number of exceedances of the ozone standard for 1978. In cases where data from more than one station were available for an urban area, the stations with the readings most representative of the area's ozone problems were chosen. The areas for which a differently located monitor was used are shown in table 3.

A. STEPWISE DISCRIMINANT ANALYSIS

We first attempted to reduce the number of variables to be considered by ascertaining which of the total number were most effective in discriminating between levels of severity of ozone problems. To do this we applied stepwise discriminant analysis, using the variables related to ozone concentration level to classify the cases. The cases were classified into five groups of approximately equal size using the variable values shown in table 4, which also shows the urban areas in each group, it may be seen that the groups do vary according to the variable used for classification.

Since we carried out the discriminant analysis in a stepwise manner, those variables entered early in the analysis should be the most influential in discriminating between the groups shown in table 4 (an analogy can be drawn using stepwise regression). Ideally, the results of the three classifications would show the same variables to be important, but the results obtained allowed only general conclusions to be drawn.

Table 5 shows the order of entry of variables for the three cases run and the percentage of cases correctly classified, for the first 14 steps. Entry of variables was halted when an entering variable had a squared multiple correlation coefficient (R^2), with the other variables, of more than 0.99. At this stage, more than 60 percent of the cases were

TABLE 3. OZONE MONITORS CORRESPONDING TO CERTAIN URBAN AREAS

Area	Monitor Used
New York	Richmond County
Philadelphia	Morristown
Springfield	Amherst
Cleveland	Painesville
San Diego	Escondido
Ventura	Simi Valley
New Haven	Derby
Bridgeport	Greenwich
San Francisco	San Jose
Dallas	Arlington

TABLE 4. URBAN AREA CLASSIFICATIONS

Group No.	Second Highest Ozone Values (pphm)	Second Highest Ozone No. of Cases	Average Ozone Concentration Values (pphm)	Average Ozone Concentration No. of Cases	Number of Exceedances Values (pphm)	Number of Exceedances No. of Cases
1	Less than 12	6	Less than 6	9	Less than 5	13
2	12-16	11	6-7	8	5-10	8
3	16-18	10	7-8	10	10-15	9
4	18-20	9	8-9	1	15-20	6
5	More than 20	9	More than 9	6	More than 20	9

TABLE 5 IDENTIFICATION OF URBAN AREA CLASSIFICATIONS

(a) Based on Second Highest Ozone Concentration

Group No.				
1	2	3	4	5
Portland	Boston	Washington	New York	Houston
Miami	Springfield	Pittsburgh	Philadelphia	Los Angeles
New Orleans	Worcester	Detroit	Baltimore	Cleveland
Dallas	Trenton	San Diego	Chicago	Ventura
Minneapolis	Youngstown	Providence	St. Louis	New Haven
Butte	Dayton	Hartford	Cincinnati	Bridgeport
	Indianapolis	Allentown	Milwaukee	Richmond
	Denver	Scranton	Sacramento	Salt Lake City
	Phoenix	San Francisco	Louisville	San Bernardino
	Fresno	Kansas City		
	Seattle			

(b) Based on Average Ozone Concentration

Group No.				
1	2	3	4	5
Boston	Springfield	New York	Philadelphia	Houston
Worcester	Providence	Washington	Baltimore	St. Louis
Trenton	Hartford	Cincinnati	Chicago	Los Angeles
Seattle	Denver	Detroit	Pittsburgh	Ventura
Portland	Phoenix	Milwaukee	Cleveland	Salt Lake City
Miami	San Francisco	Sacramento	San Diego	San Bernardino
New Orleans	Fresno	New Haven	Bridgeport	
Minneapolis	Dallas	Youngstown	Allentown	
Butte		Dayton	Scranton	
		Kansas City	Richmond	
			Indianapolis	
			Louisville	

TABLE 5 (Concluded)

(c) Based on number of exceedances

Group No.				
1	2	3	4	5
Springfield	Boston	New York	Washington	Philadelphia
Worcester	Baltimore	Detroit	Chicago	Houston
Trenton	Cincinnati	Milwaukee	Pittsburgh	St. Louis
Youngstown	Providence	Sacramento	Bridgeport	Los Angeles
Denver	Dayton	San Diego	Allentown	Cleveland
Phoenix	Indianapolis	Hartford	Louisville	Ventura
Seattle	San Francisco	New Haven		Richmond
Portland	Fresno	Scranton		Salt Lake City
Miami		Kansas City		San Bernardino
New Orleans				
Dallas				
Minneapolis				
Butte				

correctly classified. However, with 5 to 6 variables, over 50 percent could be correctly classified. The data in table 6 show that somewhat different variables are important in discriminating between groups based on the three criteria, as would be expected given the different composition of the groups for different classification variables.

Some general conclusions can be drawn from these discriminant analyses, however: First, the effect that appears to be most important overall is the insolation; cloud cover and total and average insolation are among the first variables to be entered in each case. Next most important appear to be precursor emissions, since all three of these variables are brought in among the first eight or so. After these two factors, it appears that some measure of ventilation (that is, a wind speed or a mixing height, or both) is brought in. Summer temperatures do not appear to have great importance in the classifications; they are only used after many other variables have been brought in.

B. CLUSTER ANALYSIS

We had hoped that the discriminant analysis would give us a clear picture of the most influential variables to include in a cluster analysis. Because this did not happen, we tried clustering the cases on the basis of several different sets of variables. According to Hartigan (1975), this method can be used to test the stability of the clustering process; that is, clusters that persist for different combinations of variables have a greater probability of representing a real effect. Accordingly, we carried out clustering using the program BMDP2M (Dixon and Brown, 1979), with the following sets of variables:

1) All variables (ozone levels, meteorological variables, emissions).

2) Meteorological and emissions variables.

3) Meteorological variables excluding temperature variables.

4) Ozone level variables.

5) Meteorological variables.

6) Emissions variables.

The clustering based on all variables should give an indication of overall effects. Analyses 2 and 3 give a clustering based on ozone formation potential, analysis 3 being restricted by eliminating temperatures, which were shown to be relatively unimportant by the discriminant analy-

TABLE 6. VARIABLES ENTERED AND PERCENT OF CASES CLASSIFIED CORRECTLY AT EACH STEP OF DISCRIMINANT ANALYSIS

Step No.	Variable on Which Classes Were Based: Second Highest Ozone Concentration		Average Ozone Concentration		Number of Exceedances	
	Variable Entered	Percent	Variable Entered	Percent	Variable Entered	Percent
1	Morning mixing height	24.4	Average insolation	31.1	Cloud cover	33.3
2	Average insolation	26.7	Afternoon wind speed	37.8	Total summer insolation	24.4
3	Total summer insolation	37.8	HC emissions	44.4	Afternoon mixing height	40.0
4	Cloud cover	44.4	HC/NO_x emissions ratio	40.0	Average insolation	46.7
5	Afternoon wind speed	51.1	CO emissions	48.9	HC emissions	42.2
6	HC emissions	53.3	Cloud cover	51.1	CO emissions	48.9
7	CO emissions	55.6	Total summer insolation	51.1	HC/NO_x emissions ratio	48.9
8	HC/NO_x emissions ratio	55.6	Change in mixing height	46.7	Morning mixing height	55.6
9	August max temperature	57.8	Morning mixing height	53.3	Afternoon wind speed	60.0
10	June max temperature	62.2	June max temperature	55.6	June max temperature	64.4
11	Average max temperature	60.0	Average max temperature	64.4	September max temperature	57.8
12	Morning wind speed	64.4	August max temperature	62.2	August max temperature	62.2
13	Afternoon mixing height	68.9	Morning wind speed	66.7	Morning wind speed	64.4
14	Summer max temperature	64.4			July max temperature	60.0

sis. Analyses 4, 5, and 6 should show groups of cities that have similar overall ozone problems, similar meteorology, and similar emissions, respectively.

Output from BMDP2M includes a dendogram (Everitt, 1977). The dendogram based on all variables is given in figure 8. Identification of clusters is still, to a degree, a matter of judgment, but the dendogram gives quantitative information on the similarities between cases. The distance measure used is the Euclidean distance between cases:

$$D_{ij} = \sum_{k=1}^{n} \left[(x_{ik} - x_{jk})^2 \right]^{1/2} ,$$

where x_{ik} is the value of the i-th variable in the k-th case. The program standardizes the variables to z-scores (subtract mean and divide by standard deviation) so that distances are comparable for all variables. The dendogram produced by the program is based on the single-linkage algorithm (Everitt, 1977). In this method, cases are joined according to the distance between them, with the closest being joined first. (The separation between cases is read from the distance scale on the figure.) Clusters are identified visually, and several can be seen in figure 8, though the appearance of this dendogram is indicative of little group structure (Everitt, 1977). Five clusters can be tentatively identified:

1) Boston, Hartford, Bridgeport, New Haven, Providence, Worcester, Springfield, Scranton.

2) St. Louis, Kansas City.

3) Cleveland, Detroit, Milwaukee.

4) Baltimore, Washington, Allentown, Richmond, Youngstown, Dayton, Indianapolis, Louisville, Cincinnati, Philadelphia.

5) Fresno, Sacramento.

The first cluster would represent urban areas in New England, and the second, the midcontinent. The third cluster has cities in the Great Lakes area, and the fourth includes the Ohio river valley and the East Coast. Cluster 5 has warm, dry, interior California areas. Thus, these clusters can be interpreted mainly on a geographical basis.

Figure 9 shows the dendogram based on meteorological and emissions variables. Again, there is a lack of obvious clusters, though more group

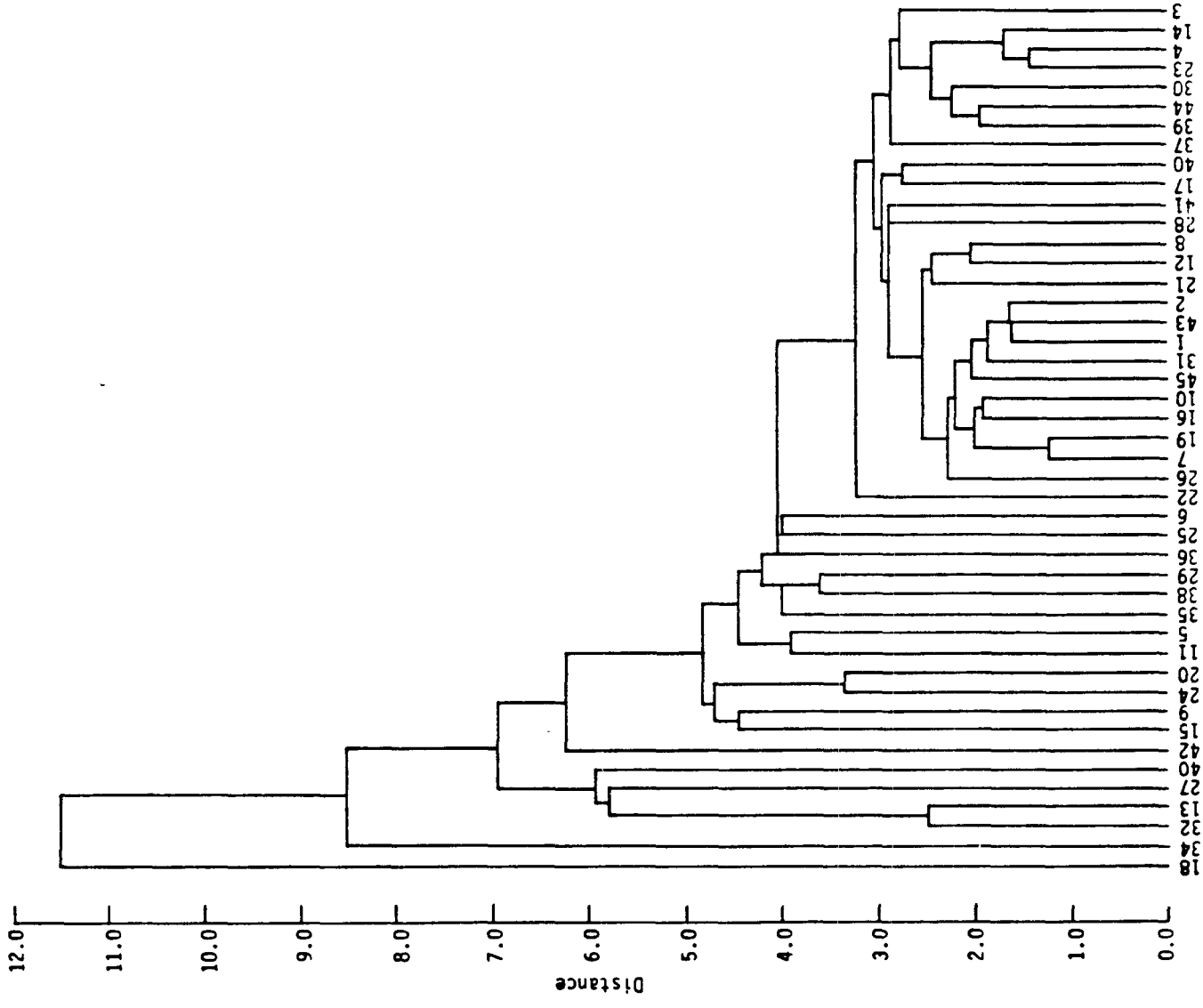

FIGURE 8. DENDOGRAM BASED ON ALL VARIABLES

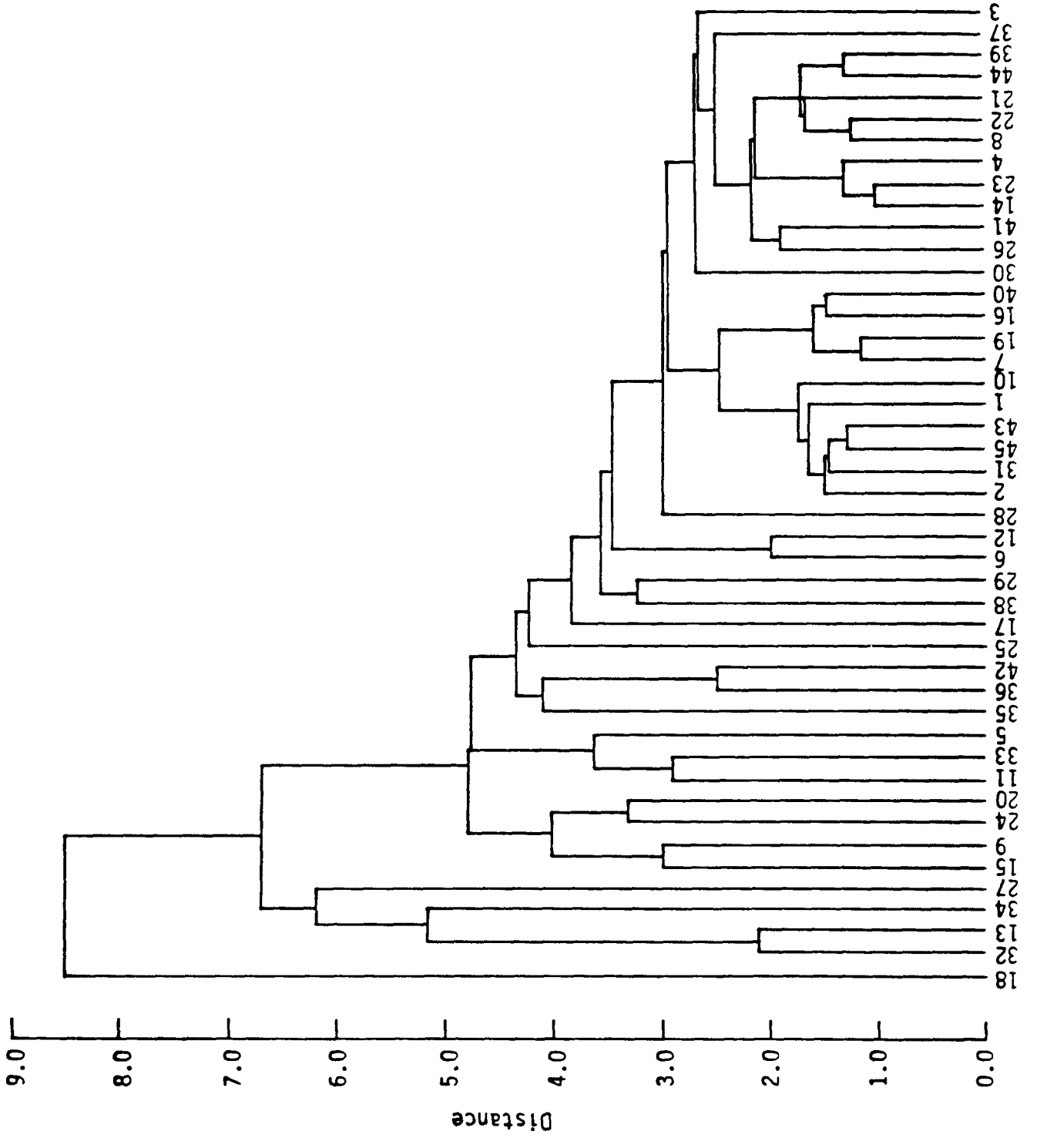

FIGURE 9. DENDOGRAM BASED ON METEOROLOGICAL AND EMISSIONS VARIABLES

structure is apparent than in figure 8. However, it is hard to identify many clear clusters. Possible clusters are:

1) Boston, Scranton, Springfield, Worcester, Milwaukee, Minneapolis, Cleveland, Bridgeport, New Haven, Hartford, Trenton, Philadelphia, Providence.

2) St. Louis, Indianapolis, Louisville, Cincinnati.

3) Dayton, Allentown, Washington, Youngstown, Richmond, Baltimore.

After these three clusters are identified, the remainder show several pairs of similar cases:

4) Detroit, Chicago.

5) Ventura, San Francisco.

6) Butte, Salt Lake City, Denver.

7) Miami, New Orleans, Dallas, Houston.

8) Fresno, Sacramento.

Again, there are the obvious geographical connotations to the clusters, except for cluster 1, which consists mostly of the New England area but also includes Milwaukee, Minneapolis, and Cleveland.

When temperatures are left out of the analysis, we obtain the dendogram in figure 10. There is a little more structure in this diagram, and we identify these clusters:

1) Springfield, Worcester, Bridgeport, New Haven, Hartford, Providence.

2) Scranton, Trenton.

3) St. Louis, Indianapolis, Louisville, Dayton, Cincinnati, Milwaukee, Minneapolis, Cleveland.

4) Baltimore, Richmond, Youngstown, Washington, Allentown.

5) Seattle, Philadelphia.

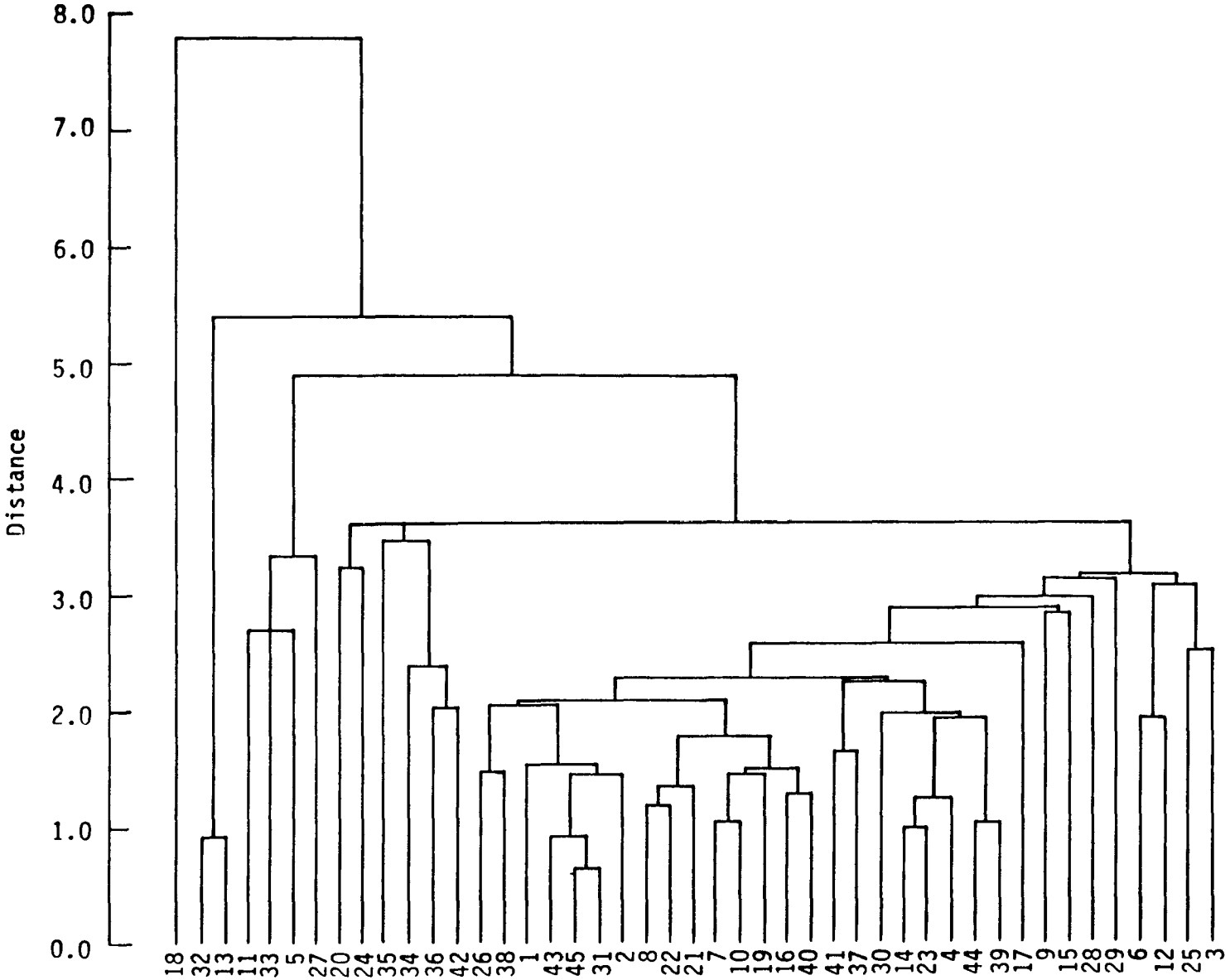

FIGURE 10. DNEDOGRAM BASED ON METEOROLOGICAL VARIABLES EXCLUDING TEMPERATURE VARIABLES

6) Fresno, Sacramento.

On the basis of the analyses performed to this point, no obvious pattern emerges. As with the discriminant analysis, the results obtained appear to depend more on the details of the analysis than on any underlying structure in the data. The problem may lie with the algorithm used in BMDP2M, which does not deal effectively with noisy data even when there is clear structure (Everitt, 1977). Possibly a different algorithm, or use of a divisive rather than an agglomerative technique, would be more successful.

Dendograms based on ozone levels, meteorological variables, and emissions variables alone are given in figures 11, 12, and 13. In these cases the algorithm has been more successful in identifying clusters, and these clusters are listed in table 7. As would be expected, clustering based on meteorology alone produces geographically close groups. The other two types of variables, however, produce clusters that do not have any geographical component to them at all. For instance, it appears that Boston and Seattle resemble each other in terms of their ozone levels. The values of the variables for these two cities are, respectively, maximum ozone, 16.9 and 16.0 pphm; second highest ozone, 13.8 and 14.0 pphm; average ozone, 5.7 and 4.3 pphm; and numbers of exceedances, six and four. Similarly, based on emissions, St. Louis and San Diego are in the same cluster. Their emissions are, respectively: HC, 127,000 and 137,000; HC/NO_x, 1.42 and 1.42; and CO, 495,000 and 1,000.

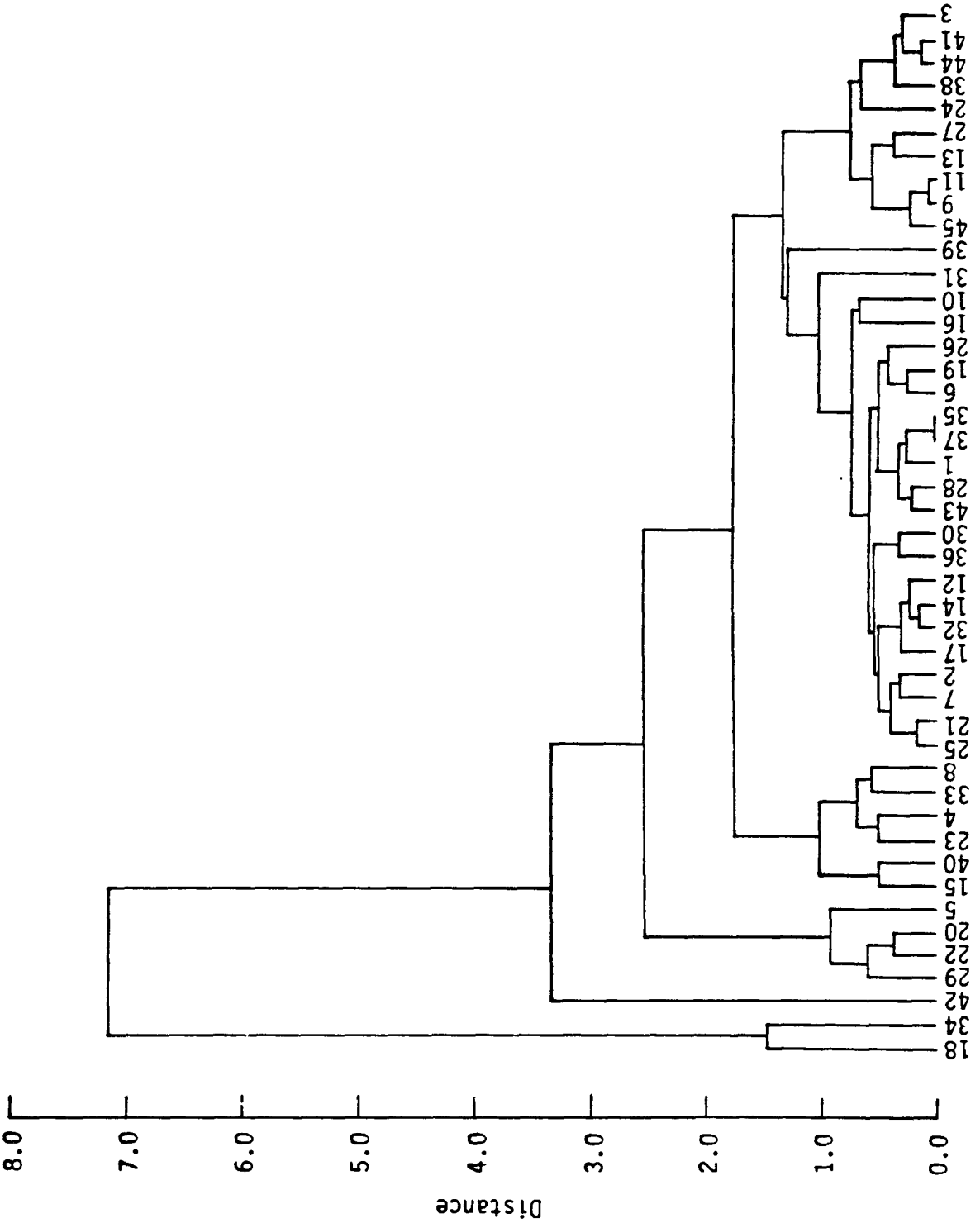

FIGURE 11. DENDOGRAM BASED ON OZONE LEVEL VARIABLES

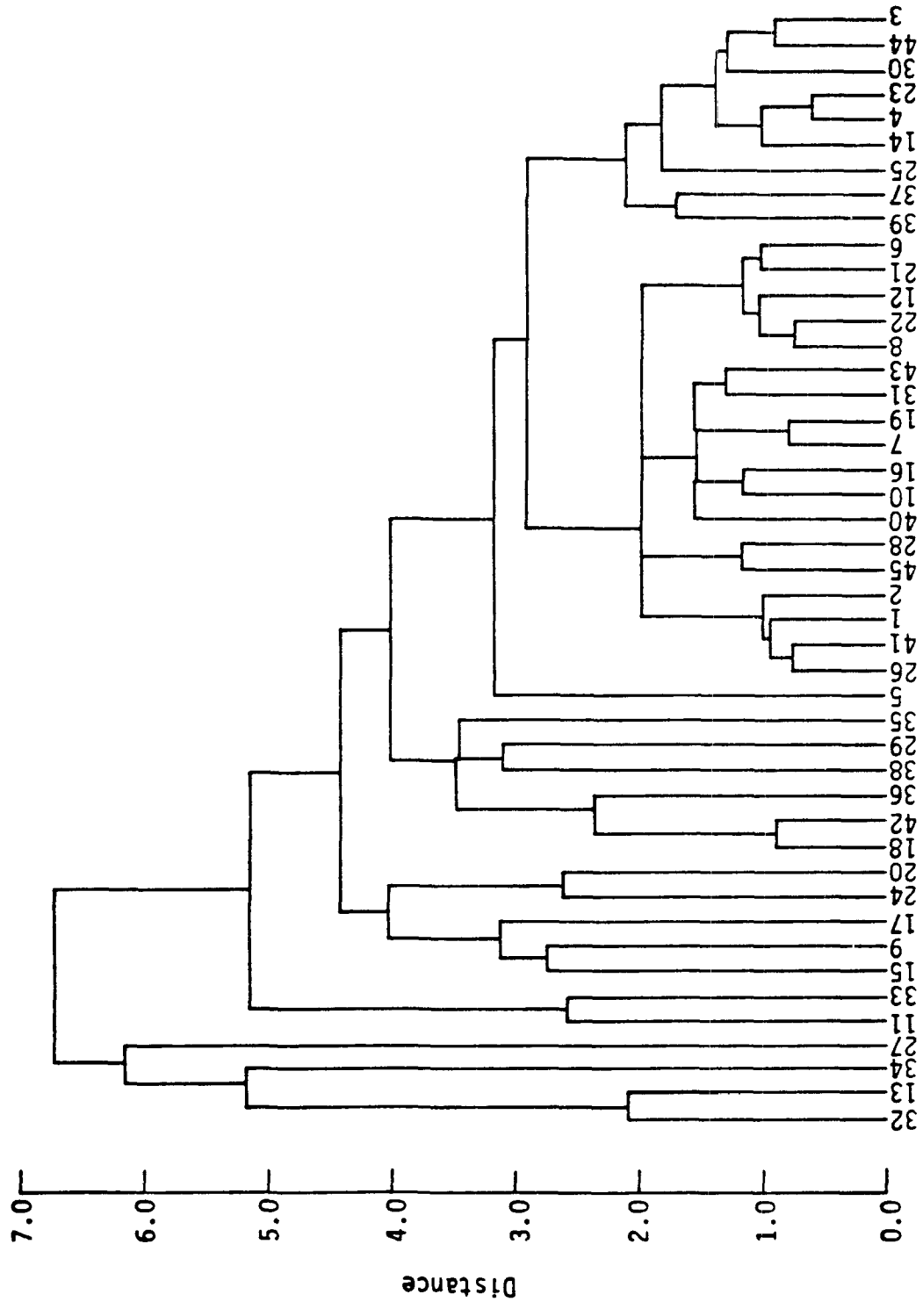

FIGURE 12. DENDOGRAM BASED ON METEOROLOGICAL VARIABLES

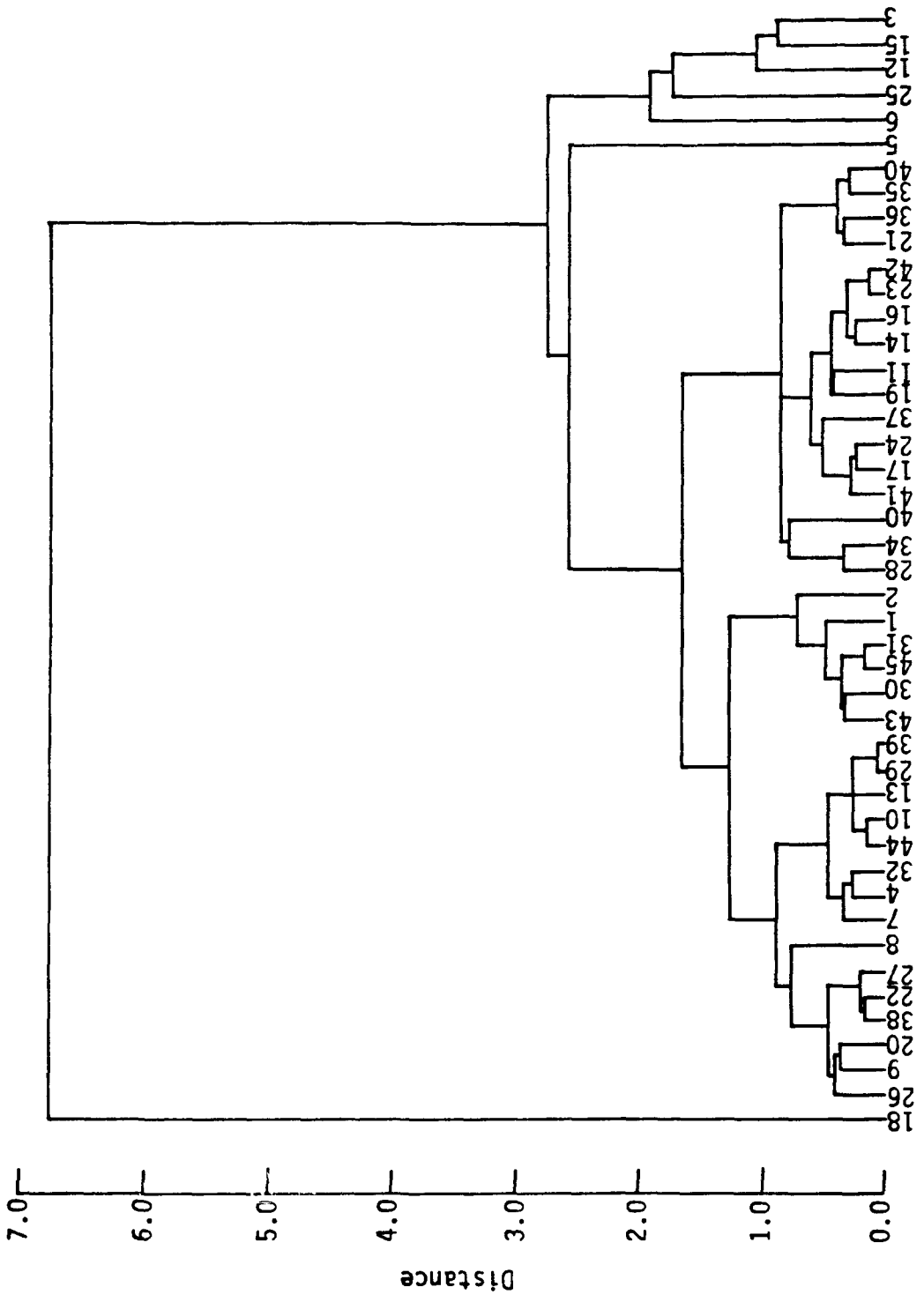

FIGURE 13. DENDOGRAM BASED ON EMISSIONS VARIABLES

TABLE 7. SUMMARY OF CLUSTERS BASED ON OZONE METEOROLOGICAL AND EMISSION VARIABLES

(a) Clusters Based on Ozone Variables (figure 5)

1) Boston, Trenton, Worcester, Seattle, Phoenix, Fresno. Denver, Dallas, Youngstown.

2) Philadelphia, Louisville, Chicago, San Diego, Scranton. Allentown, Pittsburgh, Washington, Providence, San Francisco, Detroit, Hartford, Sacramento, Kansas City, Baltimore, Cincinnati, Milwaukee, New York.

3) Cleveland, St. Louis, Bridgeport, New Haven, St. Louis, Houston.

4) Miami, Minneapolis, Butte, Portland.

(b) Clusters Based on Meteorological Variables (figure 6)

1) Boston, Worcester, Providence, New Haven, Bridgeport, Hartford.

2) Chicago, Milwaukee, Detroit. Minneapolis, Cleveland.

3) Washington, Richmond, Louisville, Cincinnati, Indianapolis, Dayton, St. Louis.

4) Pittsburgh, Youngstown.

5) Baltimore, Allentown, Trenton, Philadelphia.

6) Ventura, Los Angeles.

(c) Clusters Based on Emissions Variables (figure 7)

1) St. Louis, San Diego, San Francisco, Milwaukee.

2) Ventura, New Haven, Indianapolis, Hartford, Denver, Louisville, Scranton, New Orleans, Kansas City, Trenton.

3) St. Louis, San Bernardino, Pittsburgh.

4) Baltimore, Allentown, Richmond, Youngstown, Providence, Washington.

5) Springfield, Portland, Fresno, Dayton, Worcester, Sacramento, Bridgeport, Cincinnati.

6) Phoenix, Minneapolis, Seattle, Miami, Dallas, Philadelphia.

IV SUMMARY AND RECOMMENDATIONS

The analyses carried out for this study do not lead to a definite conclusion about the possibility of classifying cities by using combinations of characteristics such as we have used here. On the one hand, the profile analysis appears qualitatively to show that there are definite resemblances and differences, and the discriminant analysis was reasonably successful in classifying the ozone problems of the cities on the basis of a set of variables that included both meteorology and emissions. On the other hand, the agglomerative hierarchical clustering algorithm with which we attempted some quantitative clustering failed to achieve a clear-cut classification. This technique is known to be susceptible to failure in the presence of noisy data, and it is possible that a different agglomerative algorithm (e.g., Ward, 1963) or a divisive technique such as the Automatic Interaction Detector (A.I.D.) method (Sonquist and Morgan, 1963, 1964) could give better results. These methods are more robust in the presence of noisy data.

We believe that the results presented here indicate that classification techniques can be used to identify urban areas with similar ozone problems. However, more work is necessary to determine the best groupings. One possible approach would be to apply principal components or factor analysis to identify groups of variables that best account for variations in the data. An alternative to this approach would be to apply insights into the physical nature of the problem. Once an appropriate set of variables has been identified, clustering algorithms could be applied to the data; many of these algorithms can be found in the work of Hartigan (1975).

REFERENCES

Dixon, W. J., and M. B. Brown, eds. (1979), Biomedical Computer Programs P-Series, Systems, Program and Statistical Development (University of California Press, Berkeley, California).

EPA (1977a), "Uses, Limitations and Technical Basis of Procedures for Quantifying Relationships Between Photochemical Oxidants and Precursors," EPA-450/2-77-021a, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

EPA (1977b), "National Air Quality and Emissions Trends Report, 1976," EPA-450/1-77-002, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

EPA (1974), "Monitoring and Air Quality Trends Report, 1973," EPA-450/1-74-007, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Federal Register (1979), Vol. 44, No.1 221, Nov. 14, 1979.

Everitt, B. (1977), Cluster Analysis (Heinemann Educational Books, London, England).

Hartigan, J. A. (1975), Clustering Algorithms (John Wiley & Sons, New York, New York).

Holzworth, G. C. (1972), "Mixing Heights, Wind Speeds, and Potential for Urban Air Pollution Throughout the Contiguous United States," AP-101, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

The Climatic Atlas of the United States, 1968 (U.S. Department of Commerce, Washington, D.C.).

Sonquist, J. A., and J. N. Morgan (1964), "The Determination of Interaction Effects," Survey Research Centre, Institute of Social Research, University of Michigan.

Sonquist, J. A., and J. N. Morgan (1963), "Problems in the Analysis of Survey Data and a Proposal," J. Am. Stat. Assoc., Vol. 58, pp. 415-435.

Ward, J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," J. Am. Stat. Assoc., Vol. 58, pp. 236-244.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-450/4-81-031e		

4. TITLE AND SUBTITLE	5. REPORT DATE
Evaluating Simple Oxidant Prediction Methods Using Complex Photochemical Models: Cluster Analysis Applied to Urban Ozone Characteristics	August 1981
	6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Martin J. Hillyer	SAI No. 81176

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Systems Applications, Incorporated 101 Lucas Valley Road San Rafael, California 94903	11. CONTRACT/GRANT NO. 68-02-2870

12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
U.S. Environmental Protection Agency Office of Air Quality Planning and Standards Research Triangle Park, North Carolina 27711	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This report describes efforts to classify cities observing ozone levels greater than 0.12 ppm into distinct subgroups. Cluster analysis, using such factors mixing height, wind speed, temperature, NMOC/NO_x ratio and type of precursor sources, is used to identify subgroups of cities. Identification of a limited number of such subgroups could provide a means for more convincingly demonstrating the general applicability of complex photochemical models by conducting validation exercises in cities representative of each subgroup. The report indicates that the technique shows promise but, nevertheless, requires some further refinement before it can be used to identify most appropriate subgroups.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Photochemical models Ozone Cluster analysis		

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
Unlimited		42
	20. SECURITY CLASS (This page)	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE