An Interactive Graphical Interface for Hierarchical Classification of Data Generated through Individual Particle Analysis

Thomas, K. E., O’Brien & Gere Engineers, Inc., Syracuse, NY
Alexander, A., SAS Institute, Inc., Cary, NC; Raza, A., O’Brien & Gere Engineers, Inc., Syracuse, NY

ABSTRACT

Hierarchical cluster analysis has been found to be an effective means of discerning joint distributions in large volumes of data. The generation and interpretation of often hundreds of plots and descriptive statistics, however, can make such analyses extremely labor intensive. O’Brien & Gere has developed a program using SAS/AF® and SAS Screen Control Language to facilitate the hierarchical classification of data generated through individual particle analysis (IPA).

The program guides you through a series of frames to perform the following:

- select a data set for analysis by “pointing and clicking” on a pop-up directory
- select a field to begin the analysis by examining preliminary normal quantile plots
- proceed to hierarchical classification by successively “clicking” on break points in normal quantile plots of data set fields, then repeating the analysis, as necessary, on subsequent data subsets
- examining descriptive statistics for data subsets by clicking on appropriate tools from a tool bar.

Sample data from contaminated environmental media can then be attributed to identified source materials by comparing data from environmental samples to descriptive dendrograms developed through the hierarchical cluster analysis. This paper is directed at SAS users with a knowledge of SAS/AF FRAME Entry and SAS/AF Screen Control Language.

INTRODUCTION

Individual particle analysis (IPA) using scanning electron microscopy (SEM) and X-ray energy spectroscopy (EDX) is a technique which has been applied in the environmental field to characterize materials with respect to known source materials. The analysis generates multiple observations for each particle analyzed. Thus, analysis of many particles results in large volumes of data.

In environmental applications, IPA represents a valuable tool for identifying sources of environmental contamination. Characteristics of contaminated media can be compared with characteristics of known materials to assess whether contamination can reasonably be attributed to the source material used for comparison. In order to perform such a comparison, however, source materials must first be characterized. Hierarchical classification of IPA data based on normal quantile plots can be used to develop characterizations of source materials with respect to a number of observation variables.

Hierarchical classification using normal quantile plots is a statistical technique for distinguishing joint distributions in data (Johnson, Watt, and Hunt, 1991). The advantage of the technique is that it retains the individual feature information gathered; data are not subjected to transformation or compression. The classification procedure is begun by generating a normal quantile plot of particle data with respect to the values for a particular observation variable (ex. “angle”). Where the plot departs from normality (that is where the plot deviates from a straight line), the presence of joint distribution(s) is indicated. Subsets are created based on the estimated distributions. Each subset is then plotted, and the process repeated until the distribution of each subset appears to approximate normality. That is, until breaks or shifts no longer appear in the normal quantile plot of each data subset. Data subsets created through analysis of a selected observation variable (in this case, angle) can then be classified with respect to other observation variables (such as concentrations of manganese, iron, calcium, aluminum) in the data set. Upon completion of this analysis for a given known source material, the characteristics of that source material are defined with respect to each variable in the data set. The breakdown of the data into data sets and subsets can be graphically represented as a dendrogram.

A customized computer code was developed by Hunt and Johnson to perform classification using divisive hierarchical cluster analysis to identify compositionally similar source types for lead-bearing particles. Descriptive statistics were then used to evaluate data subsets created through the cluster analysis. Performing hierarchical cluster analysis using this approach is a labor intensive process. The procedure requires the creation of numerous normal probability plots, then the creation of a large number of data subsets, followed by repeated performance of statistical analyses to examine the characteristics of each subset. In addition to the significant amount of time required to develop one dendrogram, the arduousness of this process also discourages reiterations of classifications that can often optimize the classification of data. The purpose of developing this SAS® application was to facilitate the analysis by enabling a user, on a computer equipped with SAS for Windows and a mouse, to perform each step of the hierarchical classification by “pointing and clicking” using the mouse.

DISCUSSION

System Features

A frame-oriented application using SAS/AF software and screen control language was developed. The frame application that was developed allows the analyst, by using a mouse, to:

- select a data set to be examined

- perform exploratory analysis using normal probability plots on any or all variables within the selected data set
- subset a selected database into subsets based on breakpoints shown on normal probability plots, then construct normal probability plots on the subsets
- present a graphical depiction of the progression of data sets created (the dendrogram)
- provide descriptive statistics on any selected data set created during the classification procedure
- generate classification criteria from the minimum and maximum values contained in the descriptive statistics display
- save data subsets to electronic files.

System Operation

Upon invocation of the program, a captured directory listing (of the default directory) appears in a frame (see Figure 1). From the listing, you can select a data set for analysis by clicking on the desired data set with the mouse pointer. After selection of a data set, the list of observation variables from the data set appears in a frame (see Figure 2). Using this list, you can, by previewing normal quantile plots of any of the observation variables in the data set, select an observation variable with which to begin the analysis. After an observation variable has been so selected, a push-button prompt allows you to proceed to the classification (cluster) analysis. Two frames appear in the screen after you have progressed to the classification analysis component of the program. The right portion of the screen again presents a normal quantile plot of the selected observation variable. This screen allows you, by pointing and clicking, to select points on the graph where deviations from linearity occur as “break points” in the data. Data subsets are then created and named as temporary data sets. The left portion of the screen presents a dendrogram which depicts the evolution of the data sets. The naming convention consists of taking the first two characters from the selected observation variable (in this case, “AN” for the observation variable “angle”) then appending either an “L” or and “H,” representing “Low” or “High” depending on whether the data are below or above the break point selected by you. By clicking on a subset in the dendrogram, you can view a normal quantile plot of the subset in the right hand screen. For example, in Figure 3 you have clicked on the “ANH” data subset, and the plot of the ANH data subset appears in the right-hand screen. If appropriate, a break point can be selected for this subset in the manner described above. If a break point were selected from the plot of the ANH data set in Figure 3, the dendrogram would be updated to show two subsets attached to the ANH node in the dendrogram in the left-hand frame, and named “ANHH” and “ANHL,” using a continuation of the data set naming convention described above.

The tool bar in the upper left hand corner of the screen depicted in Figure 3 allows you to:

- examine descriptive statistics of the data generated using PROC UNIVARIATE
- view, in tabular form, the data from a selected data set on-screen
- print the table of summary statistics
- print the data set.

Figure 4 presents an example of univariate statistical output frame generated by clicking on the appropriate icon in the toolbar.

After you have concluded that the data set has been satisfactorily classified based on the first selected observation variable, the process can then be repeated for other observation variables in the data set.

CONCLUSION

The frame-based program described in this paper is a fully functional SAS application that reduces manyfold the labor required to perform hierarchical classification of IPA data. By significantly lowering the cost of data interpretation, the use of this program increases the feasibility of applying IPA technology in the field of environmental science.

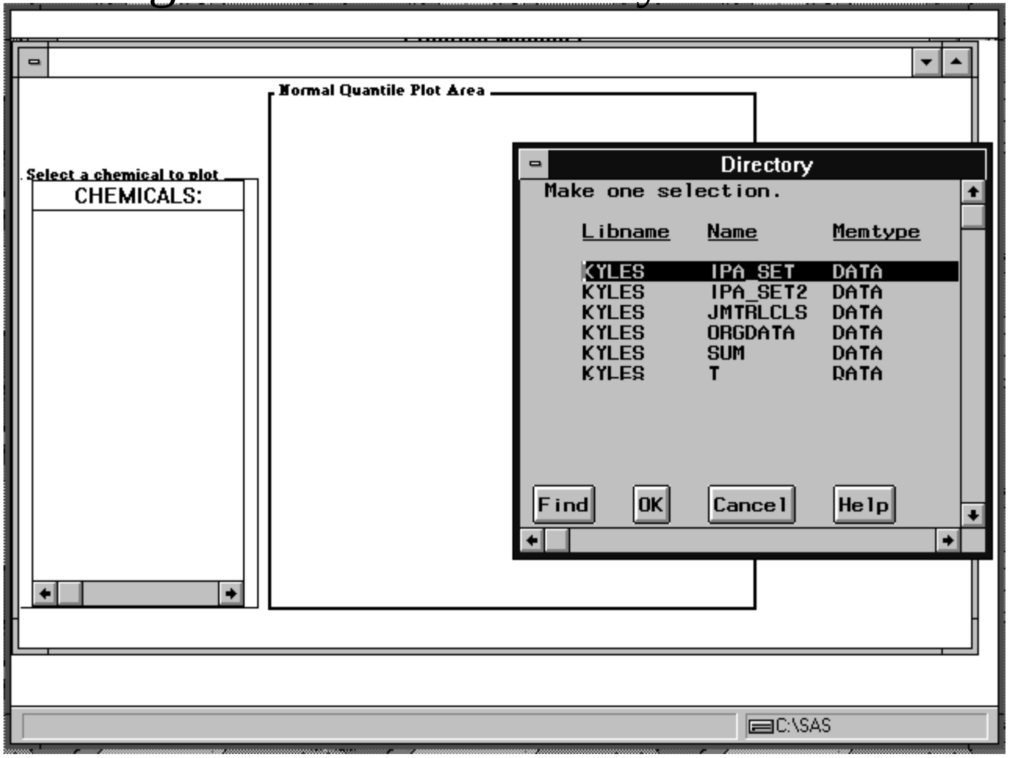
REFERENCES

Johnson, D. L., Watt, J. M., and Hunt, A. (1991), “The Advantages of Normal Quantile Plots for Classification of Multivariate Data,” unpublished, November 1991.

Johnson, D. L., and Hunt, A. (1995), “Analysis of Lead in Urban Soils by Computer Assisted SEM/ED--Method Development and Early Results,” *Lead in Paint, Soil and Dust: Health Risks, Exposure Studies, Control Measures, Measurement Methods, and Quality Assurance, ASTM STP 1226*, Michael E. Beard and S. D. Allen Iske, Eds., American Society for Testing and Materials, Philadelphia, 1995.

Hunt, A., and Johnson, D. L. (1992), “Characterizing the Sources of Particulate Lead in House Dust by Automated Scanning Electron Microscopy,” *Environmental Science & Technology*, 26(8), 1513-1523.

Figure 1 - Data directory screen



Make one selection.

Libname	Name	Memtype
KYLES	IPA_SET	DATA
KYLES	IPA_SET2	DATA
KYLES	JMTRLCLS	DATA
KYLES	ORGDATA	DATA
KYLES	SUM	DATA
KYLES	T	DATA

Figure 2 - Variable selection screen

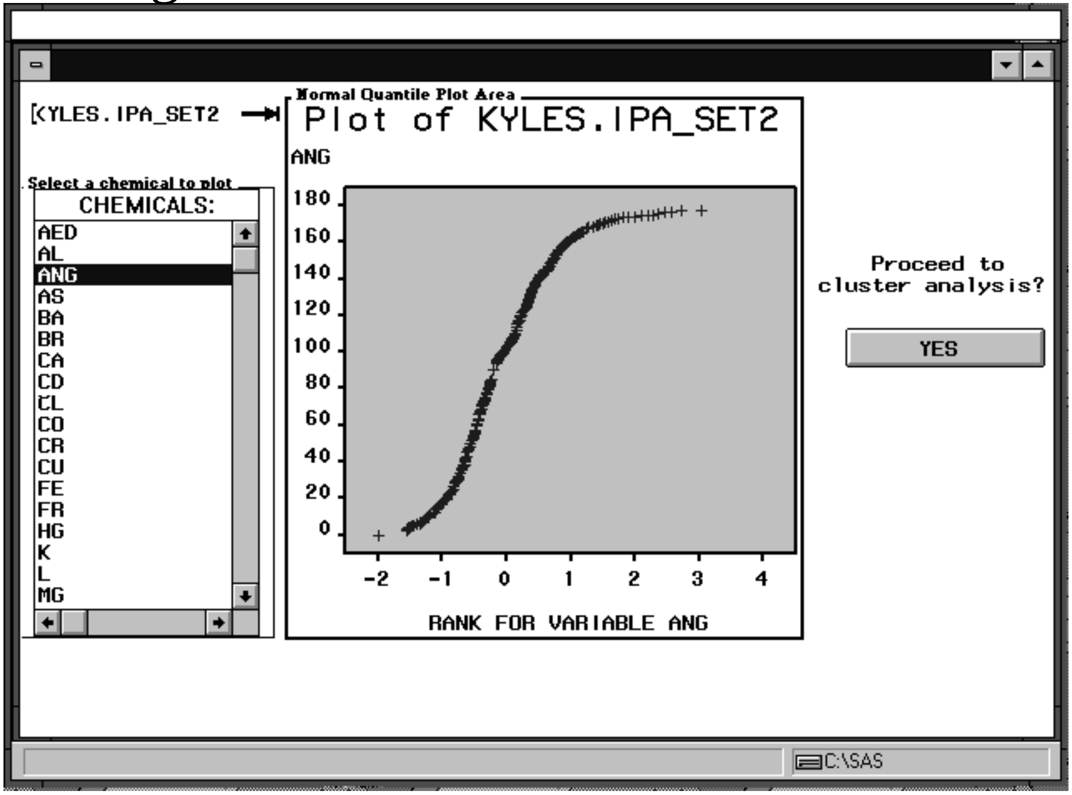


Figure 3 - Classification screen

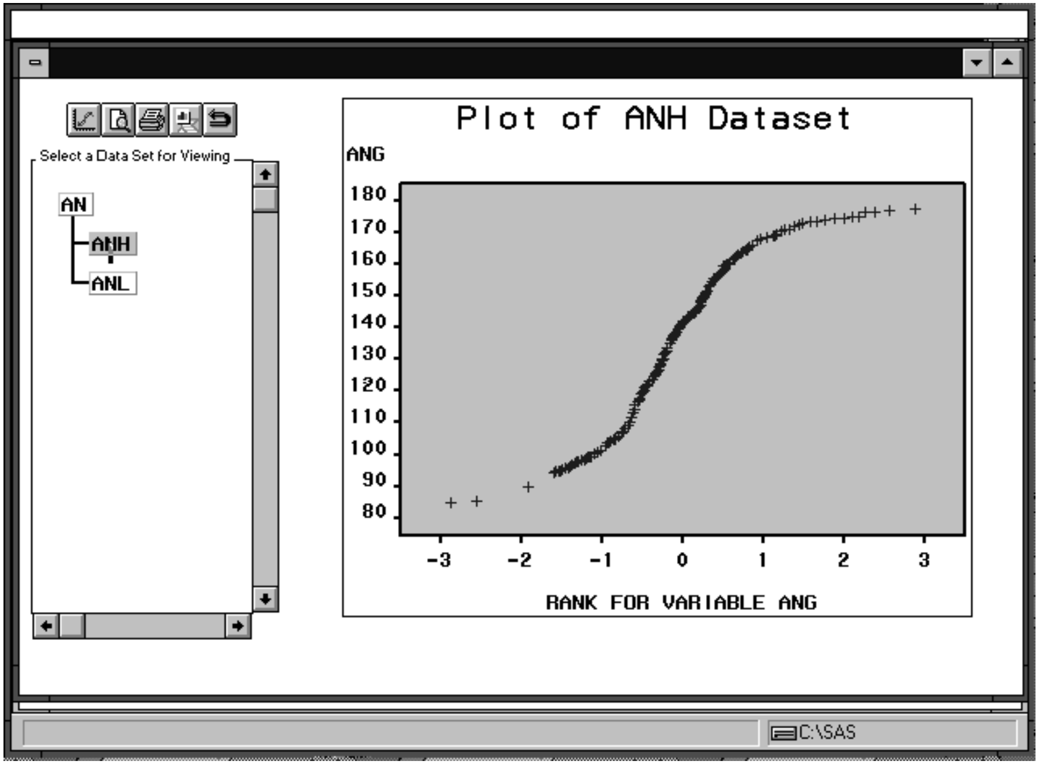
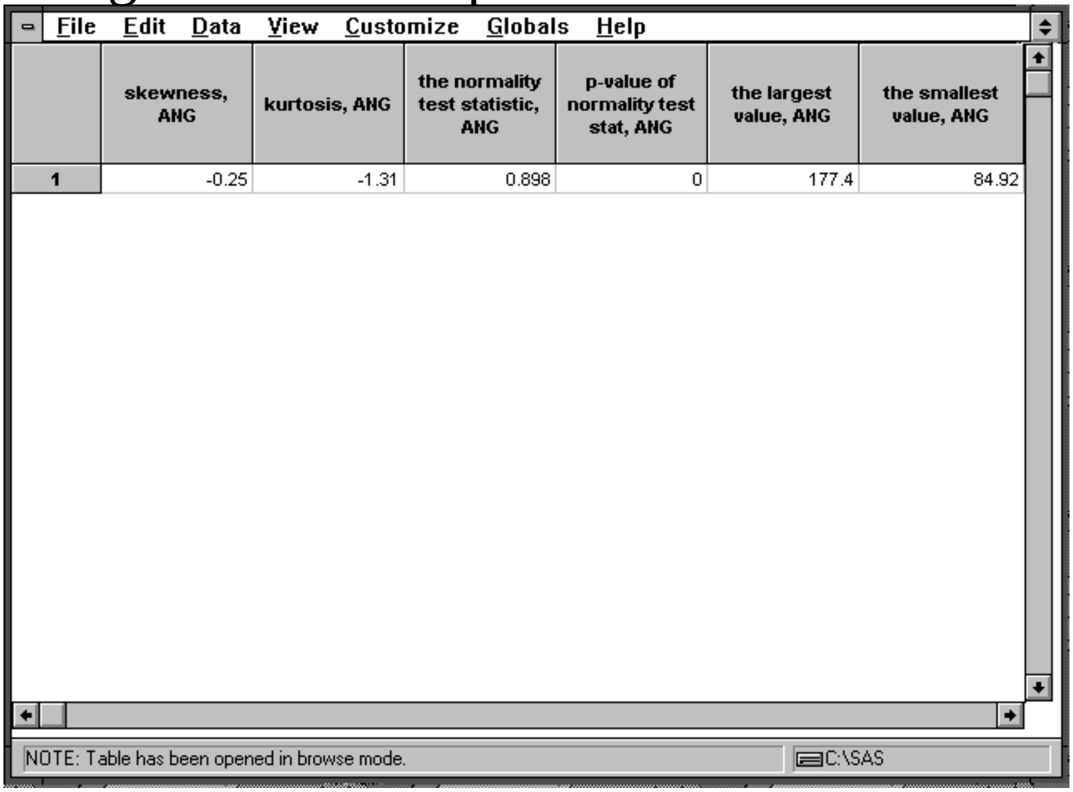


Figure 4 - Descriptive statistics screen



	skewness, ANG	kurtosis, ANG	the normality test statistic, ANG	p-value of normality test stat, ANG	the largest value, ANG	the smallest value, ANG
1	-0.25	-1.31	0.898	0	177.4	84.92

NOTE: Table has been opened in browse mode.