Chapter IX
The Distance and Cluster Procedure

ABSTRACT

This chapter describes the distance and cluster procedure of the SAS system. SAS version 9 introduced the proc distance procedure. All previous versions of SAS used two programs (xmacro.sas and distnew.sas) to process a transposed cocitation matrix (input) to produce a distance matrix (output). Cluster analysis is a data reduction technique for grouping various entities (individuals, variables, objects) into clusters so that the entities in the same cluster have more similarity to each other with respect to some predetermined selection criteria. The first section of this chapter explains the creation of a distance matrix, which is the input to the cluster procedure. The second part of this chapter focuses on the PROC CLUSTER statement which sets out the CLUSTER procedure steps. This chapter also includes the discussions of interpreting results of cluster analysis.

INTRODUCTION

SAS version 9 introduced the proc distance procedure. All previous versions of SAS used two programs (xmacro.sas and distnew.sas) to process a transposed cocitation matrix (input) to produce a distance matrix (output). The input to the cluster and multi-dimensional scaling analysis is a proximity matrix. The cocitation frequency counts matrix must be converted into a distance or similarity matrix. SAS version 9 created a new procedure, the distance procedure, to compute various measures of distance, dissimilarity, or similarity between the authors under investigation. The distance matrix is the input to the CLUSTER and MDS procedures.

There are many different ways of measuring inter-object similarity, including distance measures (proximity/difference between each pair of objects) and the correlation coefficient between a pair of objects. The higher cocitation frequencies between a pair of authors represent a higher level of cognitive linkages or similarities between them. In ACA, the cocitation frequency count matrix, correlation coefficient matrix, and distance matrix represent three different outputs in the same transformation process (see Table 1). Understanding input and output relations in the process helps us select the correct options in the distance and MDS procedures.

Table 1 highlights the input/output relationships in many PROC statements. The cocitation frequency counts matrix is the original input to all other procedures in ACA. The PROC CORR statement processes the cocitation frequency counts matrix to produce the correlations matrix. The third column heading of output/input indicates that the correlations matrix is the output of PROC CORR and it is also the input to the PROC FACTOR. The bold faced outcomes (factor pattern, clusters, two and three dimensional MDS maps) are the final outputs which are not going to be used as the input to other procedures. Therefore, distance matrix and MDS configuration coordinates are the outputs of previous stages as well as the inputs to the following stages.

The following section begins with the discussion of the four levels of measurements. The variables of the cocitation frequency matrix are authors, represented by labels (Alter, Keen, Scott Morton, etc.) or numerical algebraic expressions (x_1-x_{100}). Variables have certain characteristics that define the type of statistical analysis to be performed. These characteristics are referred to as the level of measurement of the variables. Understanding the four levels of measurement are critical to conduct ACA accurately. Some PROC statements such as PROC distance and PROC MDS ask ACA analysts to specify the measurement level of the data. The remaining sections of this chapter discuss the following topics.

- Creating permanent distance matrix
- The distance procedures
- The cluster procedures
- Interpreting results of cluster analysis

THE FOUR LEVELS OF MEASUREMENT

Variables have certain characteristics that define the type of statistical analysis to be performed. These characteristics are referred to as the level of measurement of