CHAPTER 2

NUMERICAL TAXONOMY

2.1 Definition

Biological classification is the concept restricted to the grouping of organisms by their structural attributes into taxa, from phylum (division) down to genus and species (Clifford and Stephenson, 1975). While, numerical taxonomy is the grouping of taxonomic units into taxa on the basis of their character states by numerical methods (Sneath & Sokal, 1973). Previously, Mayr (1966) designed the term "taxometrics", while Blackith and Reyment (1971) created the term "multivariate morphometrics" and Jardine and Sibson (1971) coined the term "mathematical taxonomy". The term includes the drawing of phylogenetic inferences from the data by statistical or other mathematical methods or other method, for example serology or paper chromatography, to the extent to which this is possible.

In fact that it has approach consists of a variety of numerical techniques, but they are not included in numerical taxonomy if their techniques cannot apply to problems of classification.

2.2 Principles of Numerical Taxonomy

The fundamental position of numerical taxonomy that are frequently called neo-Adansonian was originated from a french botanist, Michel Adanson (1727-1806). The followings are summary of Adason's opinions by Sneath and Sokal (1973).

- The greater the content of information in the taxa of a classification and the more characters on which it is based, the better a given classification will be.

- A priori, every character is of equal weight in creating natural taxa.
- Overall similarity between any two entities is a function of their individual similarities in each of the many characters in which they are.
- Distinct taxa can be recognized because correlations of characters differ in the groups of organisms under study.
- Phylogenetic inferences can be made from the taxonomic structures of a group and form character correlation, given certain assumptions about evolutionary pathways and mechanism.
- Classifications are based on phenetic similarity.
- Taxonomy is viewed and practiced as an empirical science.

  Organisms and characters are chosen and recorded.

These successive sequences are routine of the operation of numerical taxonomy.

- The resemblances between organisms are calculated. Estimation of resemblance is the most important and fundamental step in numerical taxonomy.
- Taxa are based upon these resemblances.
- Generalizations are made about the taxa.

2.3 Kind of Character

Characters employed in numerical taxonomy can be morphological, physiological, chemical, ecological as well as distributional characters.

2.4 The Advantages of Numerical Taxonomy

Sneath & Sokal (1973) has briefly cited the diverse advantages of numerical taxonomy such as:-

- Numerical Taxonomy has the power to integrate data from many sources: morphology, physiology, chemistry, amino acid

sequences of protein, and more which is very difficult to do by classical taxonomy.

- The less highly skilled worker can be done due to numerical taxonomy are promoted the greater efficiency automation taxonomic process.
- Being quantitative, the methods provide greater discrimination along the spectrum of taxonomic differences and are more sensitive in delimiting taxa. Thus they should give better classifications and keys than can be obtained by the conventional methods.
- It is easily to use the data, which are coded in numerical form, for the creation of descriptions, keys, catalogs, maps, and other documents.
- The creation of explicit data tables for numerical taxonomy has already forced workers in this field to use more and better-described characters. This necessarily will improve the quality of conventional methods as well.
- Numerical taxonomy can reexamine the principles of taxonomy and of the proposes classification. This has benefited taxonomy in general, and has to lead to the posing of some taxonomic questions.
- A number of biological concepts are reinterpreted by numerical taxonomy and it can be used to solve the new biological and evolutionary problem.

2.5 The numerical techniques

In this thesis Factor Analysis (FA), Cluster Analysis (CA) and Canonical Discriminant Analysis (CDA) were used to solve the classification problem in Cassia s. l. in Thailand. Details of each technique are summarized as below (Anonymous,1997).

2.5.1 The Factor Analysis (FA)

The factor analysis was introduced by Charles Spearman who published "Two Factors Theory" in 1904. It was also called 'component analysis'. It is a statistical technique used to catalogue a relatively small number of factor that can be used to represent relationships among sets of many interrelated variables.

2.5.1.1 The goal of factor analysis

- To identify factors that are substantively meaningful.
- To reduce a large number of variables to a smaller number of factors.
- To test and confirm the accuracy of the measurement.

2.5.1.2 Step in a factor analysis

In general, four steps are usually processed in factor analysis.

- Firstly, the correlation matrix for all variables is computed. Variables that don't appear to be related to other variables can be identified from the matrix and associated statistic.
- The second step, factor extraction-the number of factor necessary to represent the data and the method for calculating them must be determined. The goal of factor extraction is to determine the factor. In the factor extraction phase, the number of common factors needed to adequately describe the data is determined. This decision based on eigenvalues and the percentage of the total variance accounted for by difference numbers of factors. A plot of the eigenvalues (the scree plot is helpful in determining the number of factors.) To identify the

factors, it is necessary to group the variables that have large loading for the same factors.

- The third step, rotation will be made with focusing on transforming the factor to make them more interpretable because the unrotated factor matrix is difficult to interpret. The goal of rotation is to transform complicated matrices into simpler matrices. If a rotation has achieved a simple structure, cluster of variable should occur near the end if the axes and at their intersection when its were plotted graph.
- Finally, scores for each factor can be computed for each case. The factor score can be used in subsequent analyses to represent the values of the factors. Plot of factor scores for pairs of factors are useful for detecting unusual observations.

2.5.2 The Canonical Discriminant Analysis (CDA)

The discriminant analysis was firstly introduced by Sir Ronald Fisher. It was the statistical technique most commonly used to investigate the problem in classification. The linear combinations of the independent variable are calculated and served as the basis for classifying cases into one of group. Thus, information contained in multiple independent variables is summarized in a single index. In discriminant analysis, the weights are estimated so that they resulted in the best separation between the groups.

The linear discriminant equation is as follow.

$$D = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p \quad \ldots\ldots\ldots\ldots\ldots \text{(equation 2.1)}$$

The D is discriminator variable. The X's are the independent variable ($p \geq 1$) and the B's are coefficients estimated from the data. In discriminant analysis, the equation 2.1 was called Discriminant function or Fisher Discriminant

function. If a linear discriminant function is to distinguish, the two groups must differ in their D values.

Therefore, the B's are chosen so that the values of the discriminant function differ as much as possible between the groups, or so that for the discriminant scores are a maximum. The discriminant score is the following ration.

$$\frac{\text{between-groups sum of square}}{\text{within-groups sum of square}} \quad \text{..... (equation 2.2)}$$

2.5.2.1 The goal of canonical discriminant analysis

- To find the discriminant function which showed the relationship between discriminator variable (D) and independent variable (X's).
- To test the differentiation between two groups (Multivariate) by comparison of group centroid.
- Use the discriminant function in 1 to predicted or classified new case.

2.5.2.2 Step in a canonical discriminant analysis

Typically, five steps are carried out as follows.

- The independent variables which showing tendency to differentiate between group were selected.
- Sampling the representative of population or use the whole population.
- Accumulating data of independent variables, which were chosen in the first step.
- Discriminant function was created from data from step 2 and 3. The values of the discriminant function should be differed as much as possible between the groups, or so that for the discriminant scores are a maximum.

- Predicting or classifying new case using the discriminant function from step 4.

2.5.3 The cluster analysis (CA)

A statistical procedure employed to gather similar objects or cases and place them into groups is called a cluster analysis. The cluster analysis was previously used to classify various organisms in biology. Although both the discriminant analysis and the cluster analysis do the same thing in sorting objects or cases into group. However, the discriminant analysis do requires to know group membership for each case before processing the classification. In contrast, the cluster analysis doesn't need to know group membership of each case beforehand, but it arranges objects or cases by calculating the distance and similarity of objects or cases before classifying them into groups. In fact, selecting the variables to include in an analysis is always crucial. Poor or misleading findings may occur if important variables are excluded. In cluster analysis, the initial choice of variables determines the characteristics that can be used to identify subcategories.

The concepts of distance and similarity are basic to many statistical techniques. A measuring of how far apart of two objects are distance and the similarity is the assessments of closeness. The similarity values are large, but the distance values are small for cases that are similar.

There are many methods for calculating distances between objects and for grouping objects into a cluster. A commonly one is a sequential, agglomerative, hierarchical and nested (SAHN) clustering (Sneath and Sokal 1973). In this method, clusters are formed by grouping cases into bigger and bigger cluster until all cases are members of single cluster.

The outcome of the cluster analysis can be demonstrated with a display called a dendrogram. It is a diagramatic illustration of relationship based on degree of similarity morphology or otherwise (Clifford and Stephenson, 1975). The researcher will assigns a phenon line to divide a cluster on dendrogram. The number of phenon line on a dendrogram depends on a decision of the researcher.

2.5.3.1 The goal of cluster analysis

The goal of cluster analysis is to identify homogenous groups or clusters on concepts of distance and similarity.