# Contents

# 1 Introduction to Cluster Analysis

- Cluster analysis is a major technique for classifying a large volumes of information into manageable meaningful piles. Cluster analysis is a data reduction tool that creates subgroups that are more manageable than individual data items. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist. Like factor analysis, it examines the full complement of inter-relationships between variables. Both cluster analysis (and later, discriminant analysis) are concerned with classification.

- However, the latter requires prior knowledge of membership of each cluster in order to classify new cases. In cluster analysis there is no prior knowledge about which elements belong to which clusters. The grouping or clusters are defined through an analysis of the data. Subsequent multivariate analyses can be performed on the clusters as groups.

  ***A cluster is a group of relatively homogeneous cases or observations.***

- The term cluster analysis encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories.A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop ***taxonomies***.

## 1.1 Applications of Cluster Analysis

We deal with clustering in almost every aspect of daily life. For example, a group of diners sharing the same table in a restaurant may be regarded as a cluster of people. In food stores items of similar nature, such as different types of meat or vegetables are displayed in the same or nearby locations. There is a countless number of examples in which clustering plays an important role. Clustering techniques have been applied to a wide variety of scientific research problems. For example, in the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy. In archeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques. According to the modern system employed in biology, man belongs to the primates, the mammals, the amniotes, the vertebrates, and the animals.

Note how in this classification, the higher the level of aggregation the less similar are the members in the respective class. Man has more in common with all other primates (e.g., apes) than it does with the more "distant" members of the mammals (e.g., dogs), etc.

In general, whenever we need to classify a "mountain" of information into manageable meaningful piles, cluster analysis is of great utility.

# 2 Cluster Analysis Techniques

- Cluster analysis (CA) is an exploratory data analysis tool for organizing observed data into meaningful taxonomies, groups, or clusters, based on combinations of independent variables, which maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown.
- In this sense, Cluster analysis creates new groupings without any preconceived notion of what clusters may arise, whereas ***discriminant analysis*** classifies people and items into already known groups.
- Cluster analysis provides no explanation as to why the clusters exist nor is any interpretation made. Each cluster thus describes, in terms of the data collected, the class to which its members belong. Items in each cluster are similar in some ways to each other and dissimilar to those in other clusters.
- Cluster analysis is a tool of discovery revealing associations and structure in data which, though not previously evident, are sensible and useful when discovered. Importantly, CA enables new cases to be assigned to classes for identification and diagnostic purposes; or find ***exemplars*** to represent classes.

## 2.1 Types of Cluster Analysis

There are three main types of cluster analysis.

- Hierarchical Clustering Analysis
- Non-hierarchical Clustering Analysis (K-means clustering)
- Two Step Clustering Analysis

Within hierarchical clustering analysis there are two subcategories:

- Agglomerative (start from n clusters, to get to 1 cluster)
- Divisive (start from 1 cluster, to get to $n$ cluster)

## 2.2 Hierarchical cluster analysis

- This is the major statistical method for finding relatively homogeneous clusters of cases based on measured characteristics.
- Agglomerative clustering starts with each case as a separate cluster, i.e. there are as many clusters as cases, and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left.
- The clustering method uses the dissimilarities or distances between objects when forming the clusters. The SPSS programme calculates ***distances*** between data points in terms of the specified variables.
- A hierarchical tree diagram, called a ***dendrogram*** on SPSS, can be produced to show the linkage points. The clusters are linked at increasing levels of ***dissimilarity***. The actual measure of dissimilarity depends on the measure used.

### 2.3 Distance measures

- Distance can be measured in a variety of ways. There are distances that are Euclidean (can be measured with a ruler) and there are other distances based on similarity.
- For example, in terms of geographical distance (i.e. Euclidean distance) Perth, Australia is closer to Jakarta, Indonesia, than it is to Sydney, Australia.
- However, if distance is measured in terms of the cities characteristics, Perth is closer to Sydney (e.g. both on a big river estuary, straddling both sides of the river, with surfing beaches, and both English speaking, etc).
- A number of distance measures are available within SPSS. The ***squared Euclidean distance*** is the most widely used measure.

### 2.4 Statistical Significance Testing

- Note that the previous discussions refer to clustering algorithms and do not mention anything about statistical significance testing. In fact, cluster analysis is not as much a typical statistical test as it is a collection of different algorithms that "*put objects into clusters according to well defined similarity rules.*"
- The point here is that, unlike many other statistical procedures, cluster analysis methods are mostly used when we do not have any ***a priori hypotheses***, but are still in the exploratory phase of our research. In a sense, cluster analysis finds the "most significant solution possible." Therefore, statistical significance testing is really not appropriate here, even in cases when p-values are reported.

### 2.5 Dendrograms

- The dendrogram is a tree-structured graphical representation, used to visualize of the results of ***hierarchical cluster analysis*** . This is a tree-like plot where each step of hierarchical clustering is represented as a joining (or fusion) of two branches of the tree into a single one.
- The branches represent clusters obtained on each step of hierarchical clustering.
- The result of a clustering is presented either as the ***distance*** or the similarity between the clustered rows or columns depending on the selected distance measure.

## 3 Cluster Methods

Having selected how we will measure distance, we must now choose the clustering algorithm, i.e. the rules that govern between which points distances are measured to determine cluster membership. There are many methods available, the criteria used differ and hence different classifications may be obtained for the same data. This is important since it tells us that, although cluster analysis may provide an objective method for the clustering of cases, there can be subjectivity in the choice of method.

The linkage distances are calculated by SPSS. The goal of the clustering algorithm is to join objects together into successively larger clusters, using some measure of similarity or distance. SPSS provides seven clustering algorithms, the most commonly used one being ***Ward's method***.

## 3.1 Nearest neighbour method

(*Also known as the single linkage method*).
In this method the distance between two clusters is defined to be the distance between the two closest members, or neighbours. This method is relatively simple but is often criticised because it doesnt take account of cluster structure and can result in a problem called chaining whereby clusters end up being long and straggly. However, it is better than the other methods when the natural clusters are not spherical or elliptical in shape.

## 3.2 Furthest neighbour method

(*Also known as the complete linkage method*).
In this case the distance between two clusters is defined to be the maximum distance between members i.e. the distance between the two subjects that are furthest apart. This method tends to produce compact clusters of similar size but, as for the nearest neighbour method, does not take account of cluster structure. It is also quite sensitive to outliers.

## 3.3 Average (between groups) linkage method

(*sometimes referred to as UPGMA*).
The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. This is considered to be a fairly robust method.

## 3.4 Centroid method

Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged. This method is also fairly robust.

## 3.5 Wards method

In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.

# 4 Simple Case Studies

## 4.1 Market Segmentation

Suppose a market research company wants to undertake direct mail advertising with specific advertisements for different groups of people. You could use a variety of independent variableless like ***family income*** , ***age***, ***number of cars per family***, ***number of mobile phones per family***,***number of school children per family*** etc., to see if different postal or zip codes are characterized by particular combinations of demographic variables which could be grouped together to create a better way of directing the mail out.

This firm might in fact find that postal codes could be grouped into a number of clusters, characterized as "the retirement zone", "nappy valley", "the golf club set", the "rottweiler in a pick-up" district, etc. This sort of grouping might be valuable in deciding where to place several new wine stores, or 'Tummy to Toddler" shops.

Using cluster analysis, a customer "type" can represent a homogeneous market segment. Identifying their particular needs in that market allows products to be designed with greater precision and direct appeal within the segment. Targeting specific segments is cheaper and more accurate than broad-scale marketing. Customers respond better to segment marketing which addresses their specific needs, leading to increased market share and customer retention.

This is valuable, for example, in banking, insurance and tourism markets. Suppose four clusters or market segments in the vacation travel industry. They are:

(1) The high spending elite - they want top level service and expect to be pampered;

(2) The escapists - they want to get away and just relax;

(3) The educationalist - they want to see new things, go to museums, have a safari, or experience new cultures;

(4) the sports person - they want the golf course, tennis court, surfing, deep-sea fishing, climbing, etc.

Different brochures and advertising is required for each of these.

Brand image analysis, or defining product 'types' by customer perceptions, allows a company to see where its products are positioned in the market relative to those of its competitors. This type of modelling is valuable for branding new products or identifying possible gaps in the market. Clustering supermarket products by linked purchasing patterns can be used to plan store layouts, maximizing spontaneous purchasing opportunities.

## 4.2 A Banking example

Banking institutions have used hierarchical cluster analysis to develop a typology of customers, for two purposes, as follows:

- To retain the loyalty of members by designing the best possible new financial products to meet the needs of different groups (clusters), i.e. new product opportunities.

- To capture more market share by identifying which existing services are most profitable for which type of customer and improve market penetration.

One major bank completed a cluster analysis on a representative sample of its members, according to 16 variables chosen to reflect the characteristics of their financial transaction patterns. From this analysis, 30 types of members were identified. The results were useful for marketing, enabling the bank to focus on products which had the best financial performance; reduce direct mailing costs and increase response rates by targeting product promotions at those customer types most likely to respond; and consequently, to achieve better branding and customer retention.

This facilitated a differential direct advertising of services and products to the various clusters that differed inter alia by age, income, risk taking levels, and self-perceived financial needs. In this way, the bank could retain and win the business of more profitable customers at lower costs.

## 4.3 Steps to conduct a Cluster Analysis

1. Select a distance measure
2. Select a clustering algorithm
3. Determine the number of clusters
4. Validate the analysis

Because we usually don't know the number of groups or clusters that will emerge in our sample and because we want an optimum solution, a two-stage sequence of analysis occurs as follows:

1. We carry out a hierarchical cluster analysis using Ward' method applying squared ***Euclidean Distance*** as the distance or similarity measure. This helps to determine the optimum number of clusters we should work with.
2. The next stage is to rerun the hierarchical cluster analysis with our selected number of clusters, which enables us to allocate every case in our sample to a particular cluster.