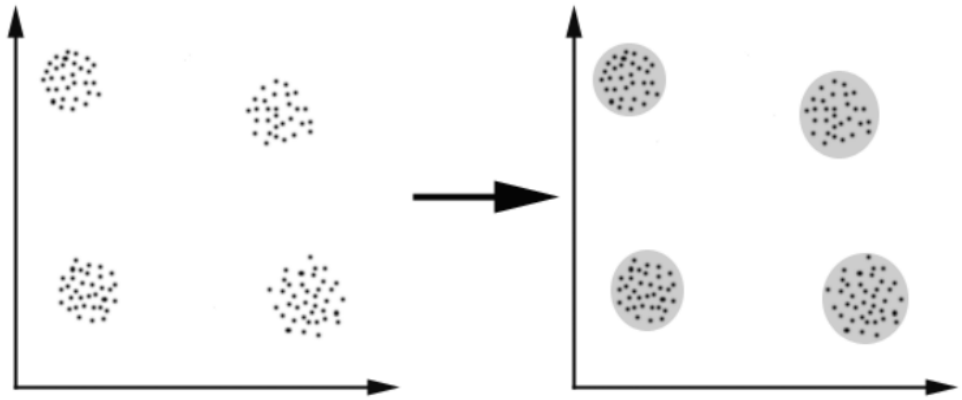
Clustering Implementation Issues

Clustering or cluster analysis is the task of grouping a set of objects in such a way that objects in the dame group (called a cluster) are more like each other than those in other groups (clusters). Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers, and for use in market segmentation, product positioning, new product development, and selecting test markets. Clustering can be used to group all the shopping items available on the web into a set of unique products. For example, all the items on eBay can be grouped into unique products.

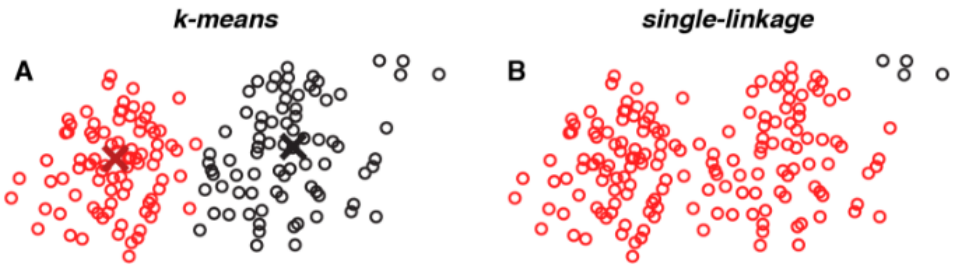
The following are the implementation issues associated with clustering:

- The different methods of clustering usually give very different results. This occurs because of the different criterion for merging clusters (including cases). It is important to think carefully about which method id best for what you are interested in looking at.
- With the exception of simple linkage, the results will be affected by the way in which the variables are ordered.
- The analysis is not stable when cases are dropped. This occurs because selection of a case (a merge of clusters) depends on similarity of one case to the cluster. Dropping one case can drastically affect the course in which the analysis progresses.



The above figure, which is an example of distance-based clustering, shows how to divide data into 4 clusters with distance as the similarity criterion.

- In cluster analysis there is no official guidelines or conventional approaches to identifying or defining clusters.
- Different initial clustering can lead to different final clustering. It is thus advisable to run the procedure several times with different (random) initial clustering.
- The resulting clustering, in the k-means algorithm, depends on the units of measurement. If the variables are of different nature or are very different with respect to their magnitude, then it is advisable to standardize them.
- The variables, in the k-means algorithm, must be Euclidean (real) vectors, so that we can calculate centroids and measure the distance from centroids; it is not enough to have only the matrix of pairwise distances or "dissimilarities".



The figure above depicts a dataset where k-means outperforms single-linkage. Single-linkage tends to fuse overlapping groups of points (red). Small groups of outliers (black) are clustered together based on small pairwise distances

- Hierarchical clustering is relatively unstable and unreliable. The first combination or separation of objects, which may be based on a small difference in the criterion, will constrain the rest of the analysis.
- It is not possible to undo the previous step in hierarchical scaling. Once the instances have been assigned to a cluster, they can no longer be moved around.
- In nonhierarchical clustering, the series of cluster is usually a mess and difficult to interpret. Further, you will have to choose the number of clusters priori, which could be a difficult task (Aaker, Kumar, Day).
- In nonhierarchical clustering, it is often difficult to know how many clusters you are likely to have and therefore the analysis may have to be repeated several times and it can be very sensitive to the choice of initial cluster centers.
- K-means clustering is sensitive to scale i.e. rescaling your datasets (normalization or standardization) will completely

change results. While this itself is not bad, not realizing that you have to spend extra attention to scaling your data might be bad.

- Clustering is computationally intensive and can only be used for relatively small datasets. It is very hard to scale effectively to large datasets not fitting in the computer main memory (Andrienko, Andrienko, Rinzivillo, 2009).

Aaker, D. Kumar, V. Day, G. Marketing Research. Retrieved from: https://books.google.co.in

Andrienko, G. Andrienko, N. Rinzivillo, S. Interactive Visual Clustering of Large Collection of Trajectories. 2009. Retrieved from: https://ieeexplore.ieee.org