Segmentation Using Cluster Analysis

Dr Robert Mayer

Istanbul Technical University

Air Transportation Management, M.Sc. Program

Strategy Module

April 2016

Lecture Overview

- Market Segmentation
- What is “Cluster Analysis”?
- 6 Steps to Cluster Analysis

Market Segmentation

- "… *one of the most widley held theories in strategic marketing*."
  (Piercy and Morgan, 1993)

- "*Many markets are significantly, but not completely, heterogeneous regarding consumers' **needs**, **wants**, **use requirements**, **tastes**, and **preferences**, and, therefore, can be dividied into **smaller**, **meaningful**, **realtively homogenous** segments of consumers*."
  (Hunt and Arnett, 2004)

Market Segmentation

- Theory is good…

  ... but how can we do it?

- Based on experience and wisdom
  - More subjective

- One of the most common “statistical” techniques is “***Cluster Analysis***”
  - More objective (but not completely)

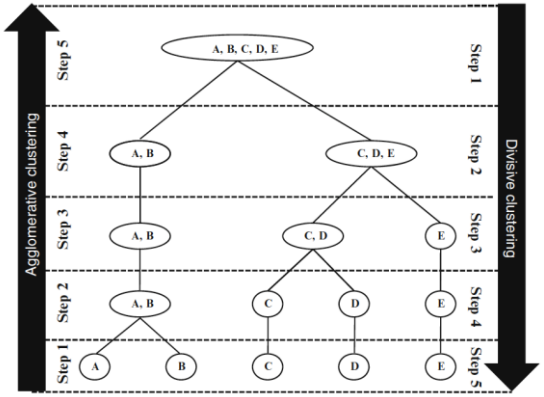
- Cluster Analysis can help to identify homogenous groups of air passengers

What is “Cluster Analysis”?

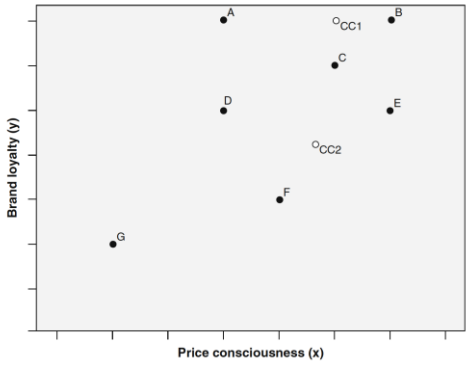
- Statistical method for classification
- **Data driven** rather than marketer driven
  - No prior assumptions with regards to the clusters (number of clusters)

Approaches to Cluster Analysis

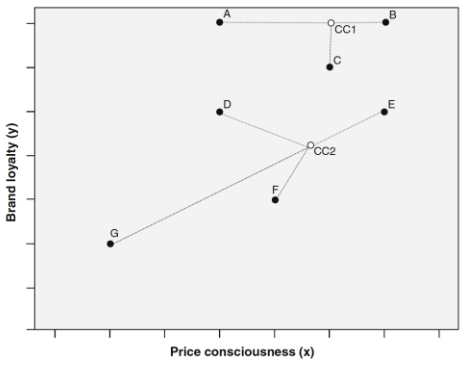
- **Hierarchical Clustering**



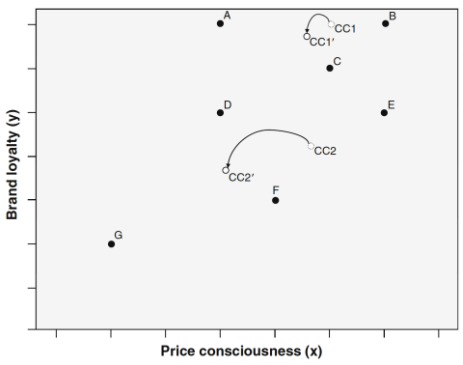
- ***k*-means Clustering**



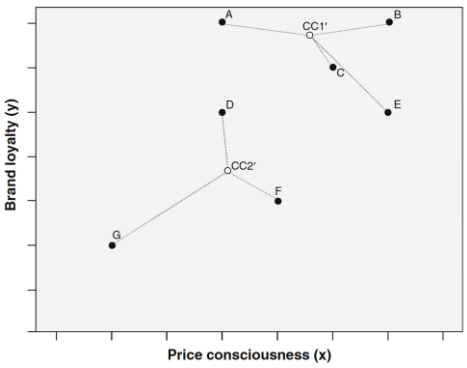
Step 1



Step 2



Step 3



Step 4

- **Two-step Clustering**

Source: Mooi and Sarstedt, 2011

6 Steps to Cluster Analysis

1. Objectives of Cluster Analysis
2. Research Design in Cluster Analysis
3. Assumption in Cluster Analysis
4. Deriving Clusters and Assessing Overall Fit
5. Interpretation of Clusters
6. Validation and Profiling of Clusters

(Hair et al., 1998)

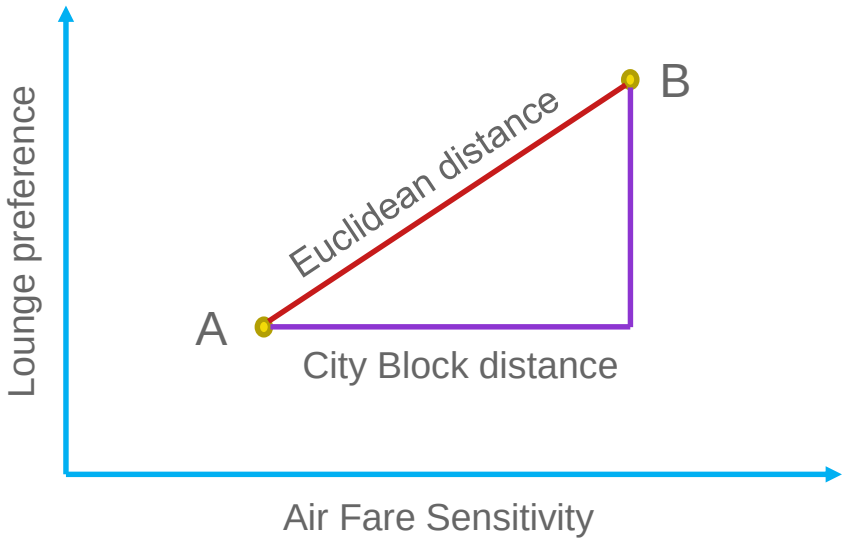
Objectives

- **Why** are we doing this analysis?
- **Data** requirements?
- Selection of variables
  - Dependent on the objectives (and the data)
  - **Psychographic variables preferred** over
  - Demographic variables (but easier to measure)
  - Often a mix of psychographic and demographic variables

Research Design

- Outliers?
  - *k*-means is very sensitive to outliers

- For Hierarchical Clustering
  - How to **measure similarity/dissimilarity**?
    - E.g. Euclidean distance, city block



Research Design

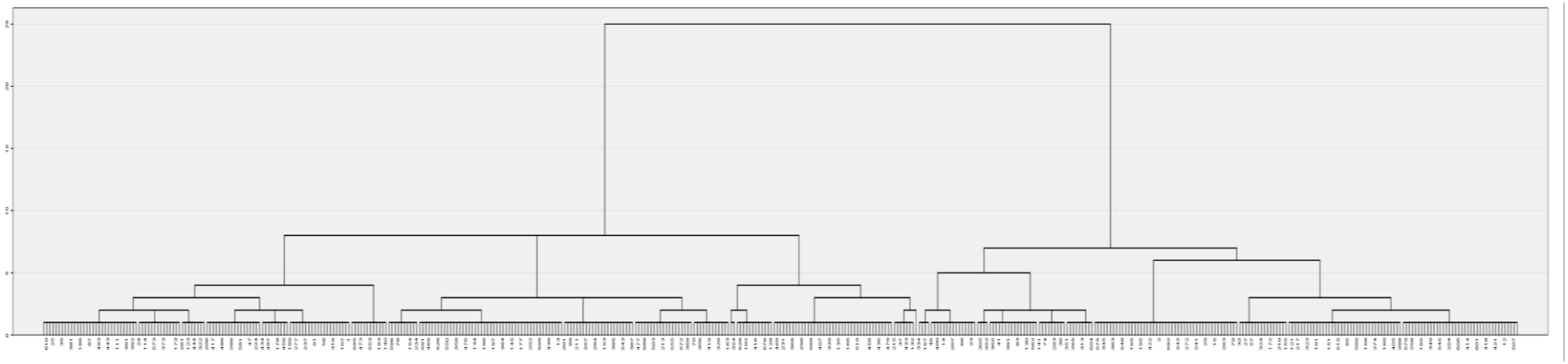
- For Hierarchical Clustering
  - **Cluster algorithm**
    - I.e. from where do we measure the distance between clusters
    - E.g. “nearest neighbour” (single linkage)
    - Frequently used: **Ward’s method** (particularly when equally sized clusters are expected and no outliers; often used with squared Euclidean distance)
  - **Standardisation**
    - What if the input data is measured on different scales?
    - E.g. Air fares in € and Seat Pitch in centimetres

Assumptions

- Representativeness of Sample
- Interdependence between Variables

Deriving Clusters

- For Hierarchical Clustering
  - **Number of Clusters**?
  - No “standard” as to the “ideal” number
  - Needs to be practical (see objectives) – common sense!
  - Software packages can provide guidance (Agglomeration Schedule)
  - Dendrogram can be useful:



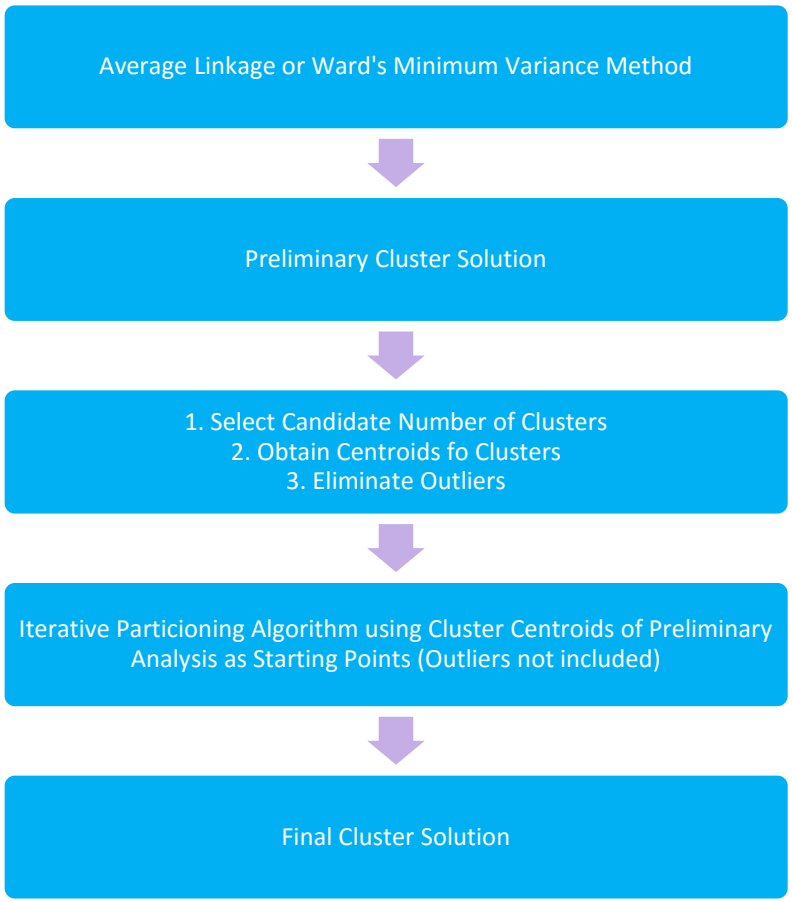
Source: Mayer, 2013

Deriving Clusters

- For *k*-means
  - **Number of Clusters** needs to be defined
- Therefore…

Deriving Clusters

- **Two-step Clustering**
- Either Hierarchical Clustering followed by *k*-means
- Statistical software (e.g. SPSS) has a two-step function



Source: Punj and Stewart, 1983

Assessing Overall Fit

- How good is the cluster solution?
  - How much do the individual variables contribute to the cluster identification?
  - How big are the individual clusters?

Interpretation

- Can the clusters be distinguished?
- Often based around the **cluster centroids** (cluster variables' average values in a certain cluster)
- Developing "**labels**" for each cluster

Validation

- Internal Validity
  - Are there (statistically) significant differences between the clusters based on **variables used** in the cluster analysis?
- External Validity
  - Are there (statistically) significant differences between the clusters based on **variables not used** in the cluster analysis?
- Replicability
  - Can the results be replicated? How "**robust**" is the result?
  - E.g. Splitting the sample
- Operational Validity
  - Is the result **practical**?

Profiling

- Previously not included variables are used to characterise the clusters
  - What are the **main characteristics** of the clusters?
  - Practical relevance
  - E.g. Demographic and behavioural variables

Cluster Analysis Workshop

- Cluster analysis is widely used in marketing
- Many air transport researchers use cluster analysis
- So it’s time to look at an example