Phd Program in Transportation

Transport Demand Modeling

Filipe Moura

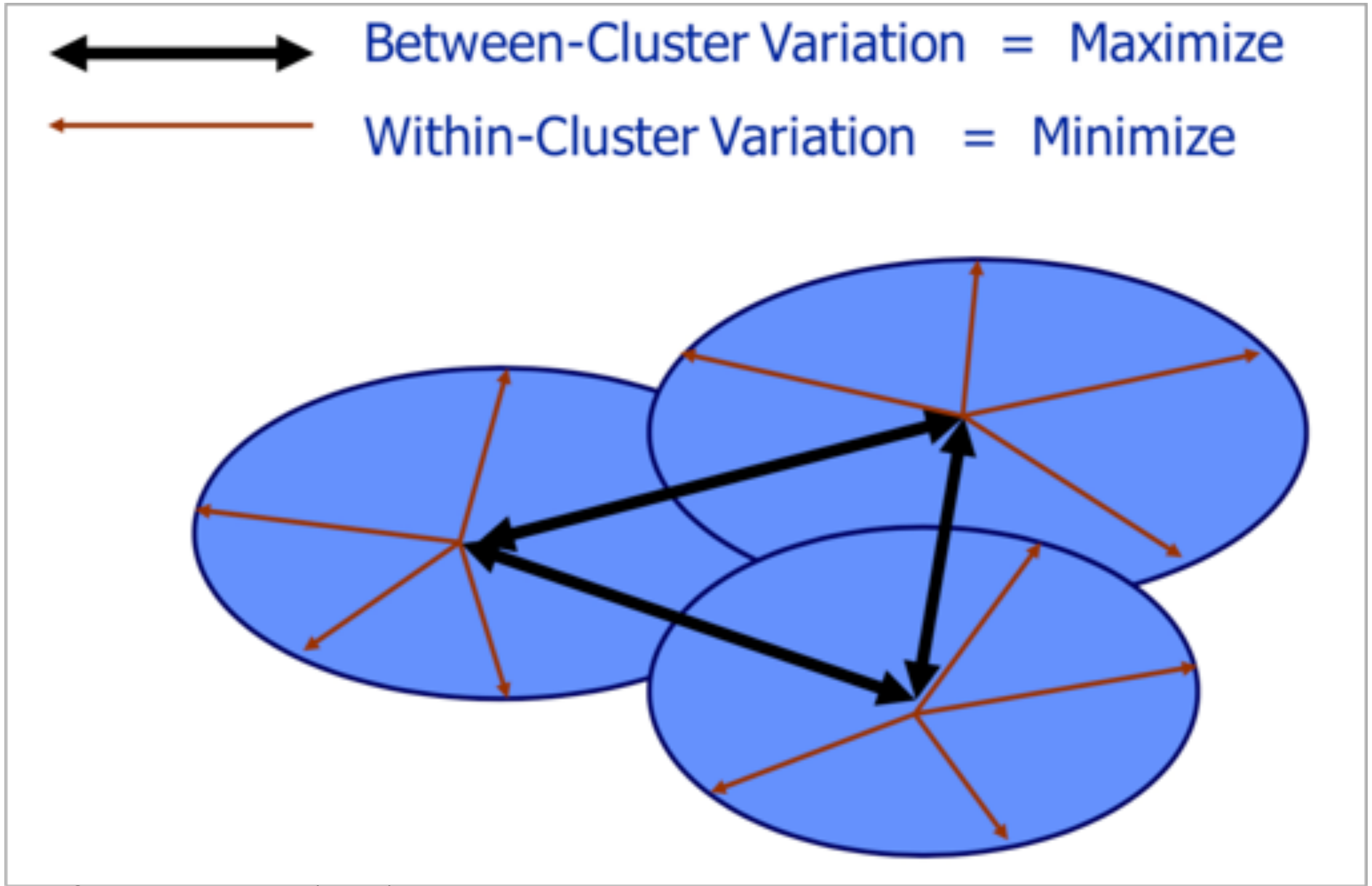
Session 5

Cluster Analysis

What is cluster analysis?

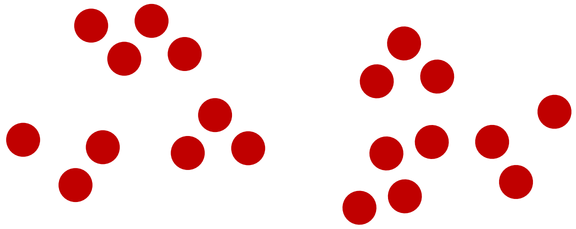
- Cluster analysis is a **exploratory technique of multivariate analysis**
- It allows **to group observations in homogeneous or compact groups** relative to one or more common characteristics
- **Each observation belonging to one cluster is similar to the other ones** belonging to it and different from all the other ones belonging to other clusters
- Basically it does **pattern recognition and grouping**
- The clusters should exhibit **high internal homogeneity** and **high external heterogeneity**
- It differs from factor analysis in that **cluster analysis groups objects** whereas **factor analysis mainly groups variables**

Objectives in Cluster Analysis

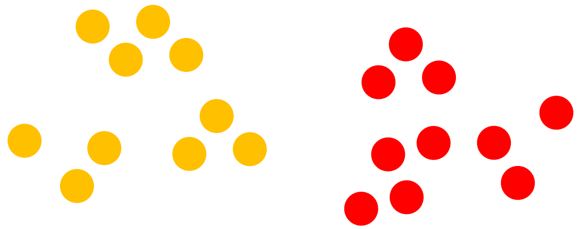


Source: Hair et al (2010)

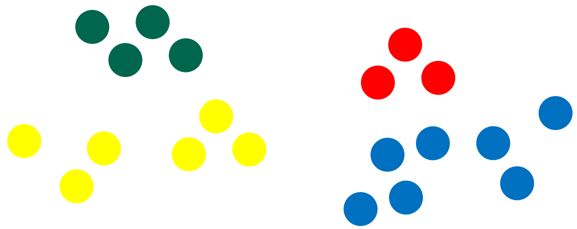
Different ways of clustering the same set of points



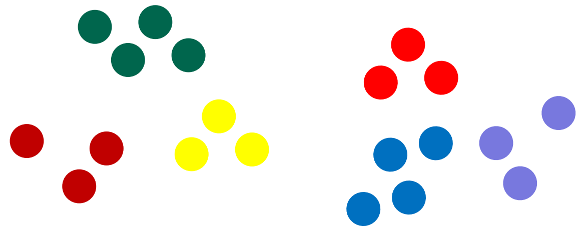
Original points



2 Clusters

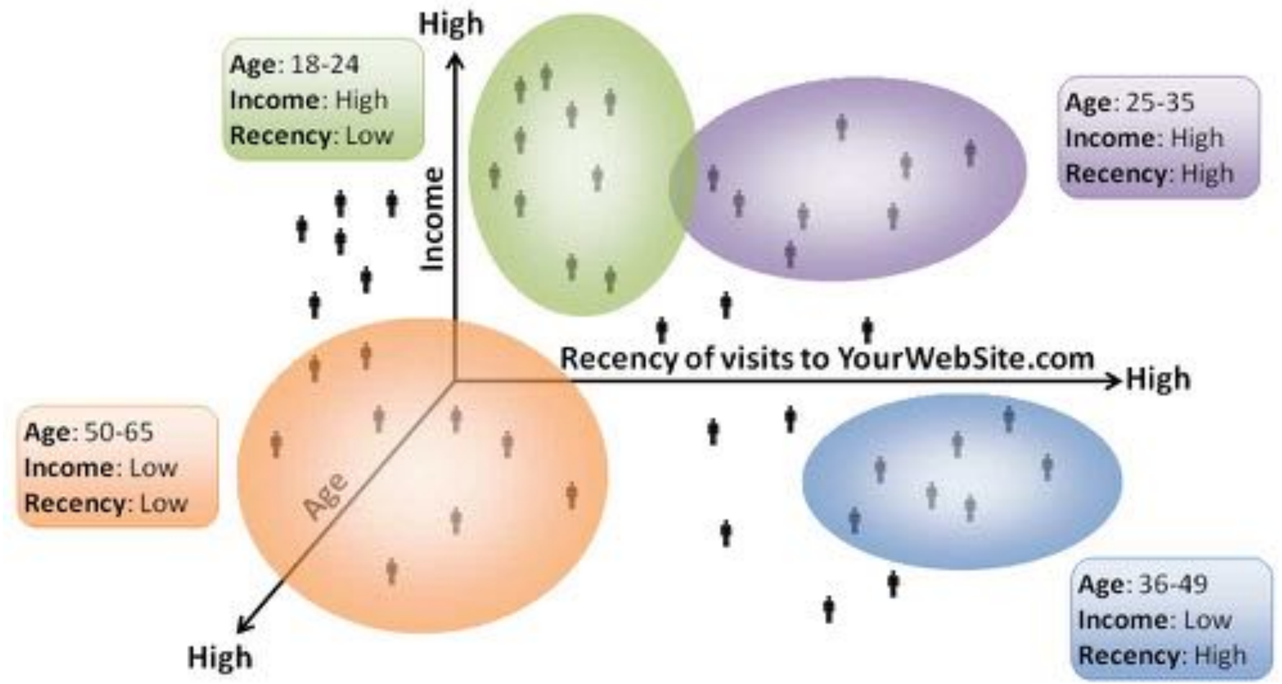


4 Clusters



6 Clusters

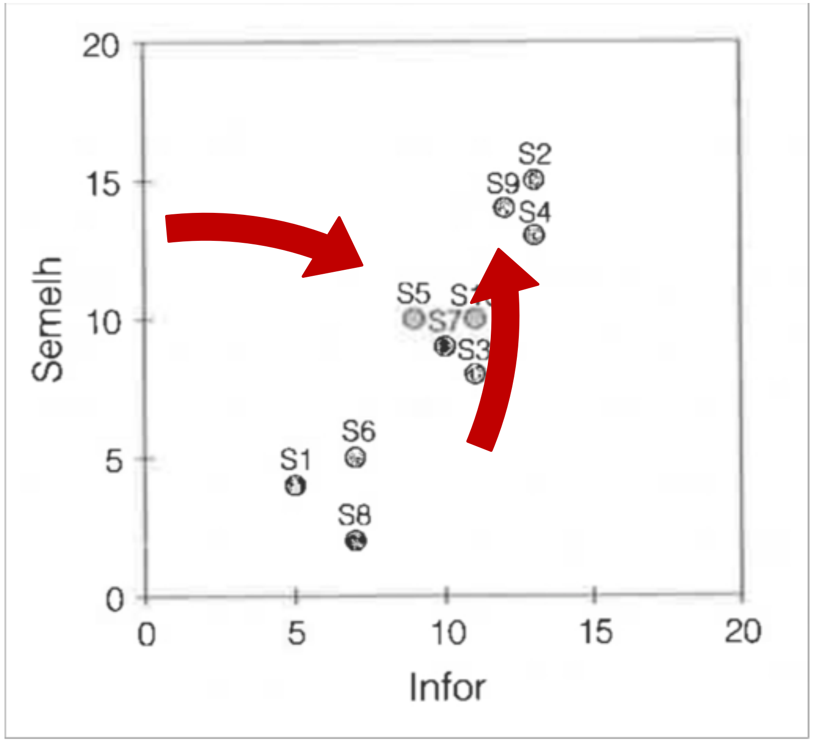
3D example



High
Age: 18-24
Income: High
Recency: Low
Income
Age: 25-35
Income: High
Recency: High
Recency of visits to YourWebSite.com
High
Age: 50-65
Income: Low
Recency: Low
Age
High
Age: 36-49
Income: Low
Recency: High

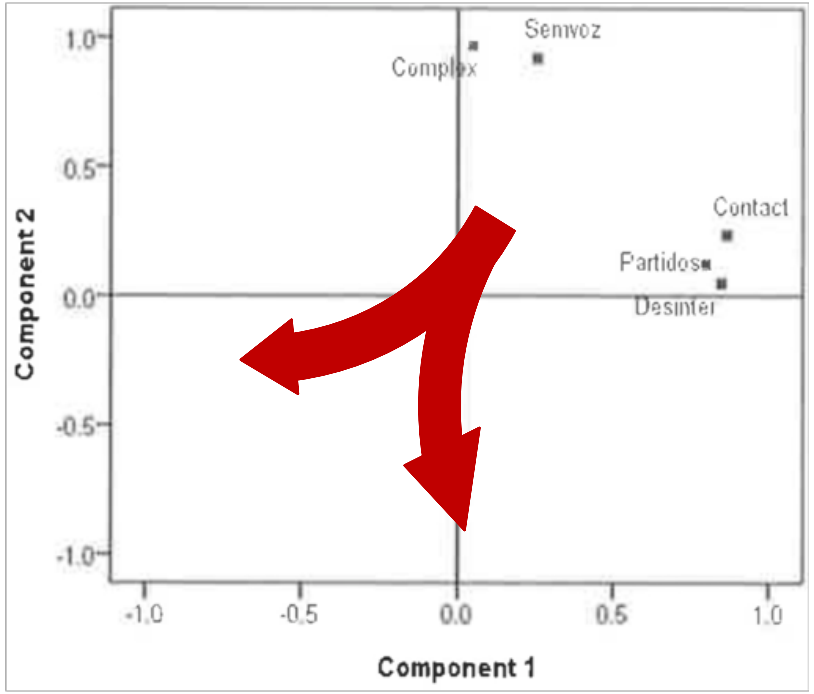
Cluster versus Factorial Analysis

Clusters



Explanatory variables form clusters of observations (individuals)

Factors



Explanatory variables inform Factors (components)

Uses of cluster analysis

- **Applications in many fields**
  - Its uses range from the derivation of taxonomies in biology to psychological classifications, to segmentation analysis of markets
- **Data reduction**
  - When a large number of observations are meaningless unless classified into manageable groups.
  - Cluster analysis can perform this data reduction
    - E.g. Understand the attitudes of population regarding public transport by identifying major groups (profiles) within the population
- **Hypothesis generation**
  - If we wish to develop hypothesis concerning the nature of data or confirm previously stated hypothesis
    - E.g. Attitudes towards transport modes could be used to separate individuals into segments or logical groups.
    - The resulting clusters could be profiled for demographic similarities and differences

Research questions in cluster analysis?

- **Taxonomy description**
  - Empirical **classification of objects**.
  - In these cases a proposed typology could be compared with the one resulting from the cluster analysis
- **Data simplification**
  - Can give a **simplified perspective** by grouping observations for further analysis
  - **Factor** analysis attempts to provide dimensions or structure to **variables**, **cluster** analysis performs the same task for **observations**
  - Instead of viewing all of the observations as unique they can be viewed as **cluster members and profiled** by their general characteristics
- **Relationship identification**
  - The underlying structure of the data represented in the clusters provides means to **reveal relationships among the observations**

Conceptual issues and critiques

- **Strong conceptual framework**
  - There should be always a strong conceptual analysis
    - Why do groups exist?
    - What variables logically explain why objects end up in the groups they do?
- **Critiques**
  - **Cluster analysis is descriptive, atheoretical and non-inferencial**.
  - It has **no statistical basis** upon which to draw inferences from the sample to the total population.
  - **Nothing guarantees a unique solution**.
  - Cluster membership is **dependent upon many elements** in the procedure, thus many solutions could be obtained by varying one or more elements

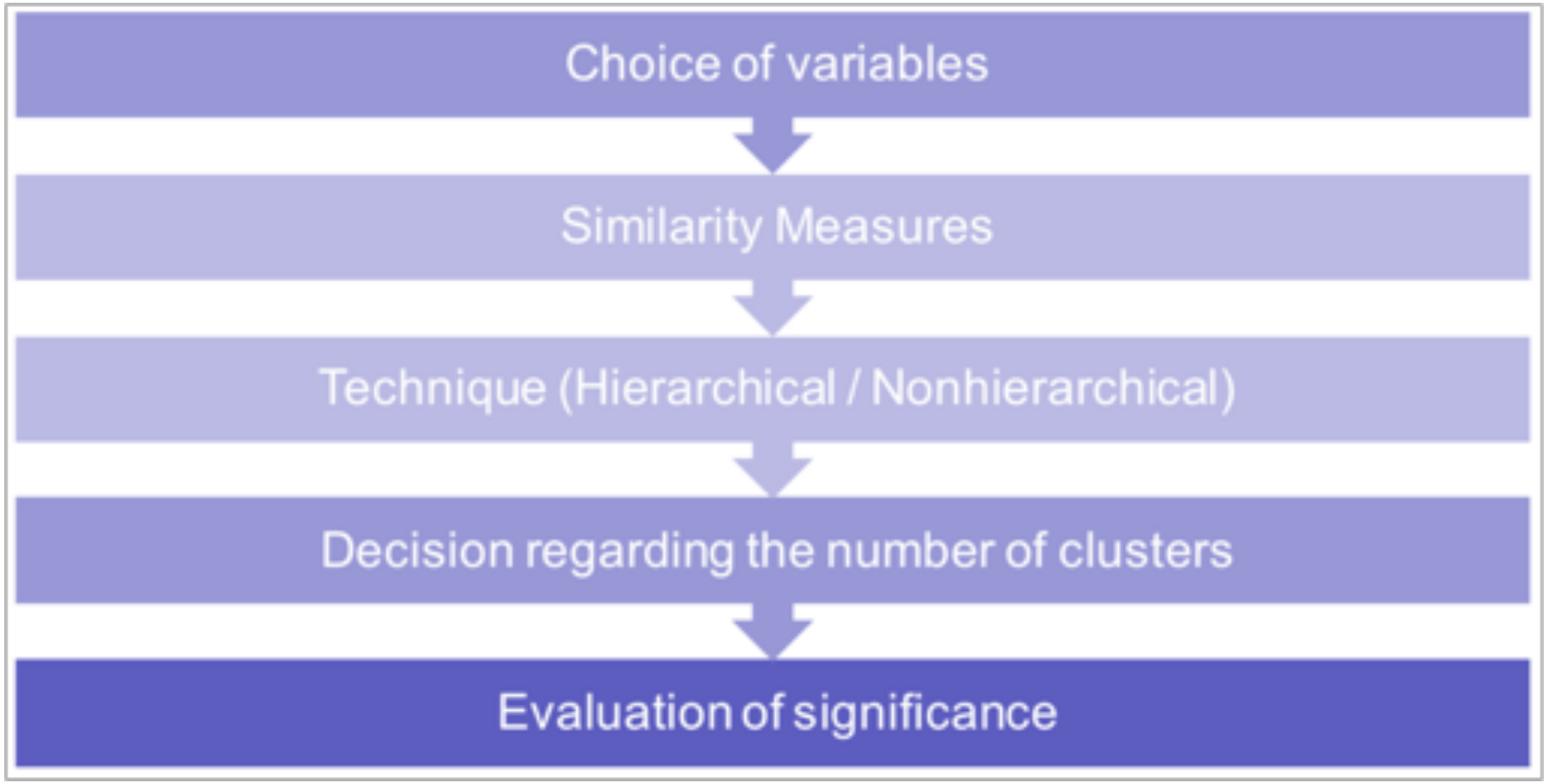
Critiques

- **Cluster analysis will always create clusters, regardless of the actual existence of any structure in the data**.
  - Just because clusters can be found it does not validate their existence.
  - Only with strong conceptual support and then validation are the clusters potentially meaningful and relevant.
- **The cluster solution is not generalizable because it is totally dependent upon the variables used as the basis for the similarity measure**.
  - It can be generalized against any statistical technique but cluster analysis is more dependent on the measures used to characterize the objects than any other multivariate technique.
  - Spurious variables or the deletion of relevant variables can have a strong impact on the resulting solution

Basic questions of cluster analysis

- **Measuring similarity**
  - Need for a method for simultaneously comparing the clustering variables.
  - Several methods are possible
    - Correlation between objects, measure of their proximity (e.g. distance between observations)
- **Cluster formation**
  - The observations whose similarity is higher should be grouped into a cluster (cluster membership of each observation)
- **Number of groups to be formed**
  - Fewer clusters implies less homogeneity within clusters
  - Larger number of clusters has more “within group homogeneity” but is less parsimonious
  - Achieving a balance between the most basic structure and an acceptable level of within cluster heterogeneity

Phases of clustering



How does cluster analysis works? (I)

❒ Similarity

- It is the **degree of correspondence among objects** across all characteristics used in the analysis (dissimilarity measures)
- Similarity is determined among each of all observations to enable **each observation to be compared to each other (proximity)**
- **Dissimilarity** will separate observations from each other (**distance**)

❒ Forming clusters

- **Hierarchical Procedure**
  - Each observation is started as is own cluster and then combining the two closest clusters until all observations are in one cluster.
  - It is also an agglomerative method since clusters are formed by combining existing clusters

How does cluster analysis works? (II)

❐ Final number of clusters

- The hierarchical method leaves several solutions, which one should be chosen?
- **Measuring heterogeneity**
  - Any measure of heterogeneity should represent the **overall diversity among observations in all clusters**.
  - The measure of heterogeneity starts with a **zero value** (each cluster is one observation) and **increase to show the level of heterogeneity as clusters are combined**
- **Select a final cluster solution**
  - By examining the **changes in the homogeneity measure** to identify large increases which are an indication of merging dissimilar clusters

How does cluster analysis works? (III)

1

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0.0	13.6	7.2	12.0	7.2	2.2	7.1	2.8	12.2	8.5
S2	13.6	0.0	7.3	2.0	6.4	11.7	6.7	14.3	1.4	5.4
S3	7.2	7.3	0.0	5.4	2.8	5.0	1.4	7.2	6.1	2.0
S4	12.0	2.0	5.4	0.0	5.0	10.0	5.0	12.5	1.4	3.6
S5	7.2	6.4	2.8	5.0	0.0	5.4	1.4	8.2	5.0	2.0
S6	2.2	11.7	5.0	10.0	5.4	0.0	5.0	3.0	10.3	6.4
S7	7.1	6.7	1.4	5.0	1.4	5.0	0.0	7.6	5.4	1.4
S8	2.8	14.3	7.2	12.5	8.2	3.0	7.6	0.0	13.0	8.9
S9	12.2	1.4	6.1	1.4	5.0	10.3	5.4	13.0	0.0	4.1
S10	8.5	5.4	2.0	3.6	2.0	6.4	1.4	8.9	4.1	0.0

2

	S1	(2,9)	S3	S4	S5	S6	S7	S8	S10
S1	0.0								
(2,9)	12.2	0.0							
S3	7.2	6.1	0.0						
S4	12.0	1.4	5.4	0.0					
S5	7.2	5.0	2.8	5.0	0.0				
S6	2.2	10.3	5.0	10.0	5.4	0.0			
S7	7.1	5.4	1.4	5.0	1.4	5.0	0.0		
S8	2.8	13.0	7.2	12.5	8.2	3.0	7.6	0.0	
S10	8.5	4.1	2.0	3.6	2.0	6.4	1.4	8.9	0.0

3

	S1	(2,9)	(3,7)	S4	S5	S6	S8	S10
S1	0.0							
(2,9)	12.2	0.0						
(3,7)	7.1	5.4	0.0					
S4	12.0	1.4	5.0	0.0				
S5	7.2	5.0	1.4	5.0	0.0			
S6	2.2	10.3	5.0	10.0	5.4	0.0		
S8	2.8	13.0	7.2	12.5	8.2	3.0	0.0	
S10	8.5	4.1	1.4	3.6	2.0	6.4	8.9	0.0

4

	S1	(2,9,4)	(3,7)	S5	S6	S8	S10
S1	0.0						
(2,9;4)	12.0	0.0					
(3,7)	7.1	5.0	0.0				
S5	7.2	5.0	1.4	0.0			
S6	2.2	10.0	5.0	5.4	0.0		
S8	2.8	12.5	7.2	8.2	3.0	0.0	
S10	8.5	3.6	1.4	2.0	6.4	8.9	0.0

5

	S1	(2,9,4)	(3,7,5)	S6	S8	S10
S1	0.0					
(2,9,4)	12.0	0.0				
(3,7,5)	7.1	5.0	0.0			
S6	2.2	10.0	5.0	0.0		
S8	2.8	12.5	7.2	3.0	0.0	
S10	8.5	3.6	1.4	6.4	8.9	0.0

6

	S1	(2,9,4)	(3,5,7,10)	S6	S8
S1	0.0				
(2,9,4)	12.0	0.0			
(3,5,7,10)	7.1	3.6	0.0		
S6	2.2	10.0	5.0	0.0	
S8	2.8	12.5	7.2	3.0	0.0

7

	(1,6)	(2,9,4)	(3,5,7,10)	S8
(1,6)	0			
(2,9,4)	10.0	0		
(3,5,7,10)	5.0	3.6	0	
S8	2.8	12.5	7.2	0

8

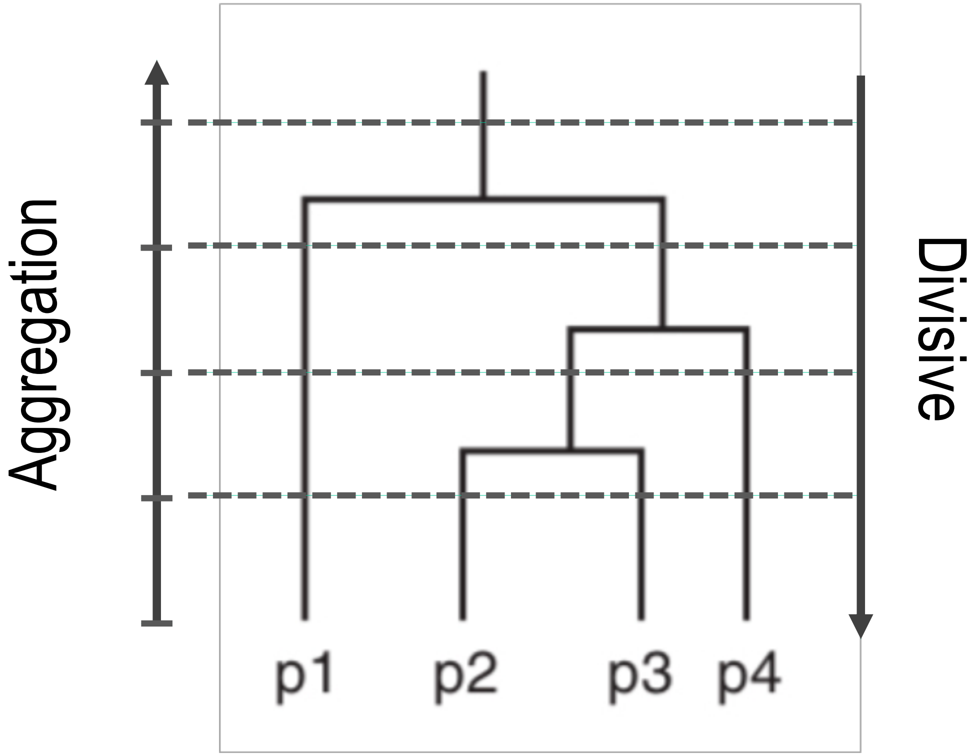
	(1,6,8)	(2,9,4)	(3,5,7,10)
(1,6,8)	0.0		
(2,9,4)	10.0	0.0	
(3,5,7,10)	5.0	3.6	0.0

9

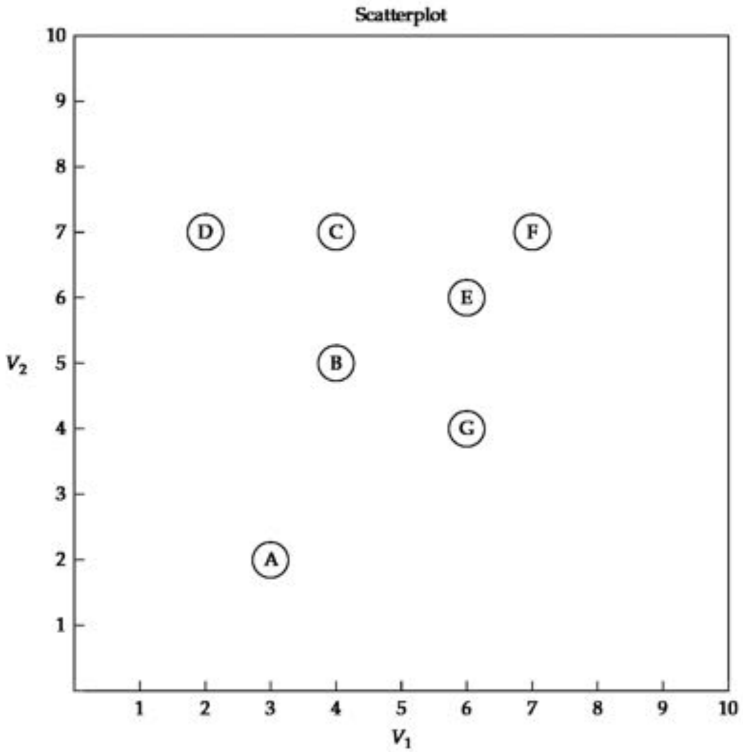
	(1,6,8)	(2,9,4,3,5,7,10)
(1,6,8)	0.0	
(2,9,4,3,5,7,10)	5.0	0.0

Hierarchical procedure - Dendrogram

- A hierarchical clustering is often displayed graphically using a **tree-like diagram called a dendrogram**, which displays both the cluster-subcluster relationships and the order in which the clusters were merged (agglomerative view) or split (divisive view).



Example of proximity matrix calculation



Data Values

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

TABLE 1 Proximity Matrix of Euclidean Distances Between Observations

Observation	A	B	C	D	E	F	G
	Observation						
A	—						
B	3.162	—					
C	5.099	2.000	—				
D	5.099	2.828	2.000	—			
E	5.000	2.236	2.236	4.123	—		
F	6.403	3.606	3.000	5.000	1.414	—	
G	3.606	2.236	3.606	5.000	2.000	3.162	—

Euclidean distance: $$d_{ij} = \sqrt{\left[\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2\right]}$$

$$d_{AB} = \sqrt{[(3-4)^2 + (2-5)^2]} = 3.162$$

Practical considerations

- Only the **relevant and meaningful variables** should be included for cluster analysis
  - That characterize the objects being clustered
  - Relate specifically to the objectives

AGAIN… GARBAGE – IN – GARBAGE OUT!

- Cluster analysis could be **dramatically affected** by the inclusion of:
  - Only one or two **inappropriate variables**
  - Variables that are not distinctive (do not differ significantly across the derived clusters)

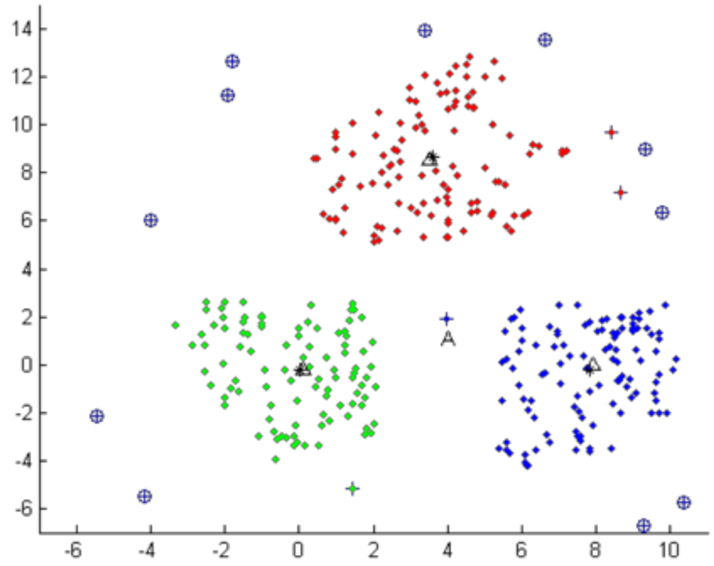
Sample size and outliers

- ❒ **Sample size**
  - ➢ Large enough to provide sufficient representation of small groups within the population and represent the underlying structure
- ❒ **Outliers**
  - ➢ An outlier is a representative element of a small but substantive group? Small samples make it difficult to answer this question
- ❒ Sample size also depends on the research objectives:
  - ➢ Does it requires **the identification of small groups** within the population? => **Larger sample**
  - ➢ Is the interest only focusing in **larger groups** (major segments)?

    => **Smaller sample**

Outliers

- **Cluster analysis is sensible to outliers**
  - **Truly aberrant** observations should be **removed**
  - Representative observations of small segments could be removed but noticing that the analysis will only accurately represent the important segments



- **Graphic Profile diagram** lists the variables along the x-axis and the variable values along the y-axis
- Outliers could also be identified through **measures of similarity** (e.g. each observation against overall group centroid)

Measuring similarity (I)

- Inter-object similarity
  - Empirical measure of correspondence or resemblance between objects to be clustered
- Correlation Measures
  - Correlating pairs of objects based on several variables.
  - High correlations indicate similarity.
    - It doesn´t look at the observed mean value but instead looks at the patterns of movement over the variables measured – **Similarity of profiles**
    - Correlation measures are rarely used because most applications put emphasis on the magnitudes of the objects instead of on the patterns
    - They could instead be used when the objective is the grouping of variables and not of observations. In this case they are more appropriate.

Measuring similarity (II)

Distance measures

- Measures similarity as the **proximity of observations to one another** across the variables in the cluster variate.
- They are also a measure of **dissimilarity (Distance)**.

Euclidean distance

- Straight line distance

$$d_{ij} = \sqrt{\left[\sum_{l=1}^{q} (x_{il} - x_{jl})^2\right]}$$

Squared (absolute) Euclidian distance

- Better in computational aspects

$$d_{ij} = \sum_{l=1}^{q} (x_{il} - x_{jl})^2$$

Measuring similarity (III)

❒ Minkowski distance

- Generalization of the Euclidian Distance

$$d_{ij} = \sqrt[m]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{m}}$$

❒ City-block (Manhattan) distance

- Special case of the Minkowski distance were m=1

❒ Mahalanobis Distance

- Accounts for the correlation among variables (statistical distance between objects) – Not available in SPSS for Cluster Analysis

$$d_{ij} = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$$

, where S is an estimate of the Variance-Covariance matrix of cluster groups

Measuring similarity (IV)

❒ Cosine Similarity Measure

- Measures the proximity between two objects for *p* vectors (at least interval variables)

$$CoSIN\ (i,j) = \frac{\sum_{k=1}^{p} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{ij}^2}}$$

❒ Jaccard, Russel & Rao and Measures of binary association

- When there are nominal variables, the measures of metrical distance cannot be applied

Measuring similarity (V)

- If two objects are characterized by *p* nominal dichotomous variables (binary)
  - ***a*** is the number of attributes **present in both** objects
  - ***b*** is the number of attributes **present in object *i*** and **absent in object *j***
  - ***c*** is the number of attributes **absent in object *i*** and **present in object *j***
  - ***d*** is the number of attributes **absent in both** objects

- **Jaccard coefficients**
  - Similarity $$s_{ij} = \frac{a}{a+b+c}$$
  - Dissimilarity $$d_{ij} = \frac{b+c}{a+b+c}$$

Measuring similarity (VI)

- ❒ **Russel & Rao**
  - ➢ Similarity

$$s_{ij} = \frac{a}{a+b+c+d}$$

- ❒ **Johnson and Wichern**
  - ➢ Similarity

$$s_{ij} = \frac{a+d}{a+b+c+d}$$

  - ➢ Dissimilarity

$$d_{ij} = \frac{b+c}{a+b+c+d}$$

Standardization

- **Different distance measures** or a change in scale of the variables may lead to **different cluster solutions**.
  - It is advisable to test different measures
- The **distance measures** are generally the preferred ones because they **represent more accurately the concepts of proximity** (fundamental to cluster analysis)
- **Standardization**
  - Distance measures are quite sensitive to different scales or magnitudes among the variables.
    - In general the **variables should be standardized**
  - Usually the most common standardization is the **z score**

Assumptions in cluster analysis

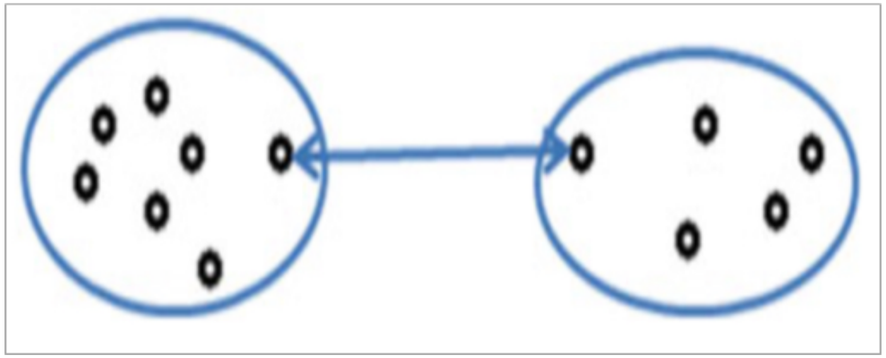
- **No requirements of normality, linearity and homoscedasticity**
  - Cluster analysis is not influenced by the requirements of normality, linearity and homoscedasticity
- **Sample Representativeness**
  - The sample used should be truly representative of the entire population.
  - The results are only as good as the representativeness of the sample
- **Multicollinearity**
  - It acts as a weighting process not apparent to the observer but affecting the analysis.
  - Thus research about substantial multicollinearity should be performed prior to the cluster analysis, take measures against it (e.g. reducing the number of variables)

Hierarchical clustering procedures

- Series of n-1 clustering decisions (n observations) that **combine observations into a hierarchy structure** (***dendrogram***)
  - **Agglomerative methods**
    - Each object starts as is own cluster and is successively joined with the closest one until only a single cluster remain – most commonly used
  - **Divisive methods**
    - Departs from a single cluster which is successively divided
- Clustering algorithms **hierarchical procedure** that determines **how similarity is defined between clusters** in the process
  - When we have more than one element in each cluster how do we do?

Clustering algorithms (I)

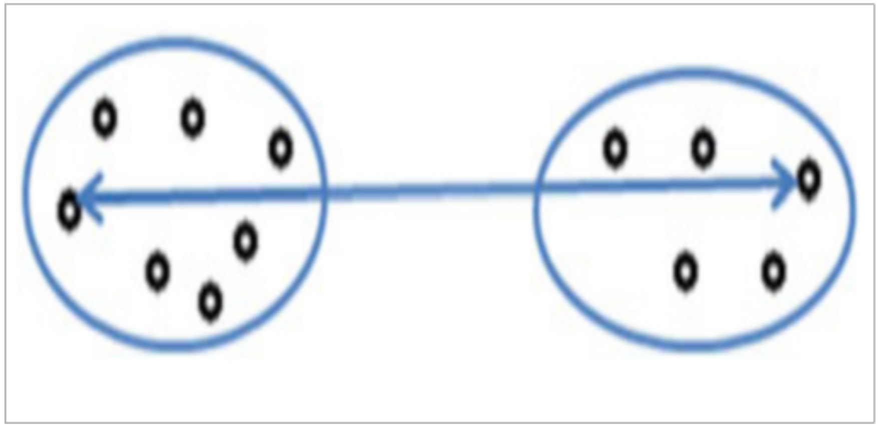
- **Single linkage our nearest neighbor**
  - The distance between two-clusters is represented by **the minimum of the distance between all possible pairs of subjects in the two groups**
    - The similarity between clusters is the **shortest distance between any object** in one cluster and any object in the other cluster.
  - It is the most commonly used and its very flexible.
  - It can define a wide range of clustering patterns.
  - When clusters are **poorly delineated**, it could **create problems**.



Clustering algorithms (II)

❒ Complete linkage or farthest neighbor

- In this approach, the cluster similarity is based on the **maximum distance between observations in each cluster**.
- Similarity between the clusters is the smallest circle that could encompass both of them.
- Eliminates some of the problems of earlier method and has been found to generate the most compact clustering solutions



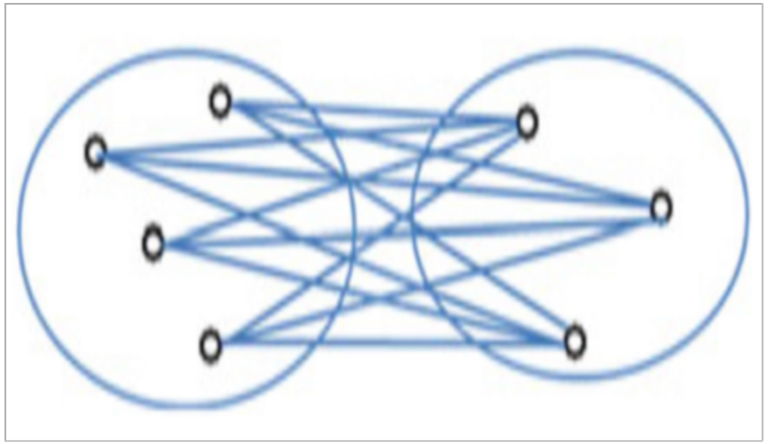
Clustering algorithms (III)

❒ Average linkage between groups

- The distance between clusters is **the average of the distances** between observations in **one cluster** to **all the members in the other cluster**.

❒ Average linkage within groups

- Similar to the previous method but here the clusters are united in a way to minimize the sum of squared errors (**minimize variability inside the clusters**)



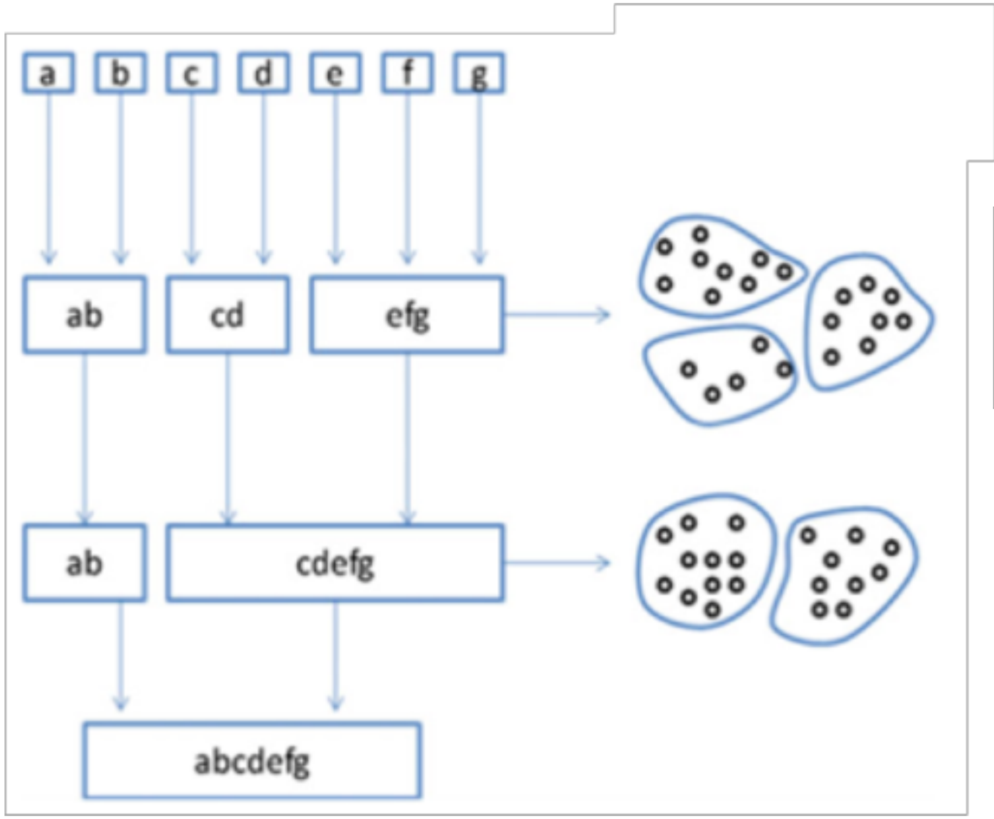
Clustering algorithms (III)

❒ Ward's Method

- The measures of similarity are the **sum of squares within the cluster** summed over all variables.
- The retained **clusters** are the ones with the smallest values
- Easily distorted by outliers

❒ Centroid method

- The similarity between two clusters is the **distance between its centroids**.
- Less affected by outliers
- They could produce confusing results



Non-hierarchical cluster procedures

- **Main difference between non-hierarchical from hierarchical**
  - Do not involve tree-like construction procedures
  - Assign objects to a **predetermined number of clusters**

- **Two steps approach**
  - Specify cluster seeds
    - Starting points for each cluster could be pre-specified by the analyst.
  - Assignment
    - Assign each observation to one of the cluster seeds based on similarity.
    - Each observation is assigned to the most similar cluster seed.

K-means

❒ **K-means**

1. Cluster partition in the *k* clusters (***k* defined by the analyst**)
2. **Estimate the centroids** of each one of the k clusters and calculation of the **Euclidean distance from each centroid to each object**
3. **Group the observations in to the clusters which have its centroid closest to each observation**, return to the previous step until the point in which there is no significant variation in the minimum distances (or until the number of iteration or the convergence criteria have been reached)

Advantages of Hierarchical Methods

- **Simplicity**
  - Simple and comprehensive image of the entire clustering solutions.
  - One can evaluate any of the possible clustering solutions
- **Measures of similarity**
  - Several similarity measures.
  - Could be applied to almost any type of research questions
- **Speed**
  - Hierarchical methods generate an entire set of solutions efficiently

Disadvantages of Hierarchical Methods

- **Misleading**
  - Can be misleading due to undesirable early combinations. Sensible to outliers
- **Outliers are very influential**
  - The reduction of the number of outliers (deletion) might distort the solution
- **Large samples**
  - Not appropriate to analyze large samples
    - Solution: extract a random subsample

Advantages and disadvantages of Non-hierarchical methods

- **Nonhierarchical methods – Advantages**
  - The results are not so susceptible to **outliers**, the **distance measure** used, and the inclusion of **irrelevant or inappropriate variables**.
  - Can analyze extremely **large datasets**
    - It doesn´t require the calculation of similarity matrices but only the similarity of each object to each cluster centroid.
- **Nonhierarchical methods – Disadvantages**
  - It does **not** guarantees **optimal solutions**.
  - Not suitable to explore a wide range of solutions based on similarity measures, observations included and potential seed points.

Best way to proceed

- **Combination approach**
  - First use a **hierarchical** technique to **generate a complete set of cluster** solutions and **establish the appropriate number of clusters**
  - After the **elimination of outliers**, use a **nonhierarchical method**

- One should analyze and **examine the rational behind the clusters** defined.
  - Clusters with small number of observations should be fully examined
    - Do they represent valid components or simply outliers?

Number of clusters

- It is one of the **most critical aspects of cluster analysis**
- Since there is no statistical inference, **several methods** have been developed.
  - *Ad-hoc* procedures that are sometimes complex and must be calculated by the analyst
  - Specific to particular software packages

Stopping rules

- ❒ **Measures of heterogeneity change**
  - Percentage of changes in heterogeneity
  - Measures of variance change
    - Root mean square standard deviation
  - Statistical measures of heterogeneity change
    - Pseudo F-test
- ❒ **Direct Measures of heterogeneity**
  - **Cubic clustering criterion** (in SAS)
    - Measure of the deviation of the clusters from an expected distribution of points (multivariate uniform distribution)

Interpretation

- The **profiling and interpretation** provide
  - A way to assess the **correspondence** of the derived clusters to those proposed by **prior theory or experience**.
  - When used in a **confirmatory mode**, cluster analysis provides a mean to assess this correspondence.
- **The analyst compares the derived clusters to a preconceived typology**

Validation

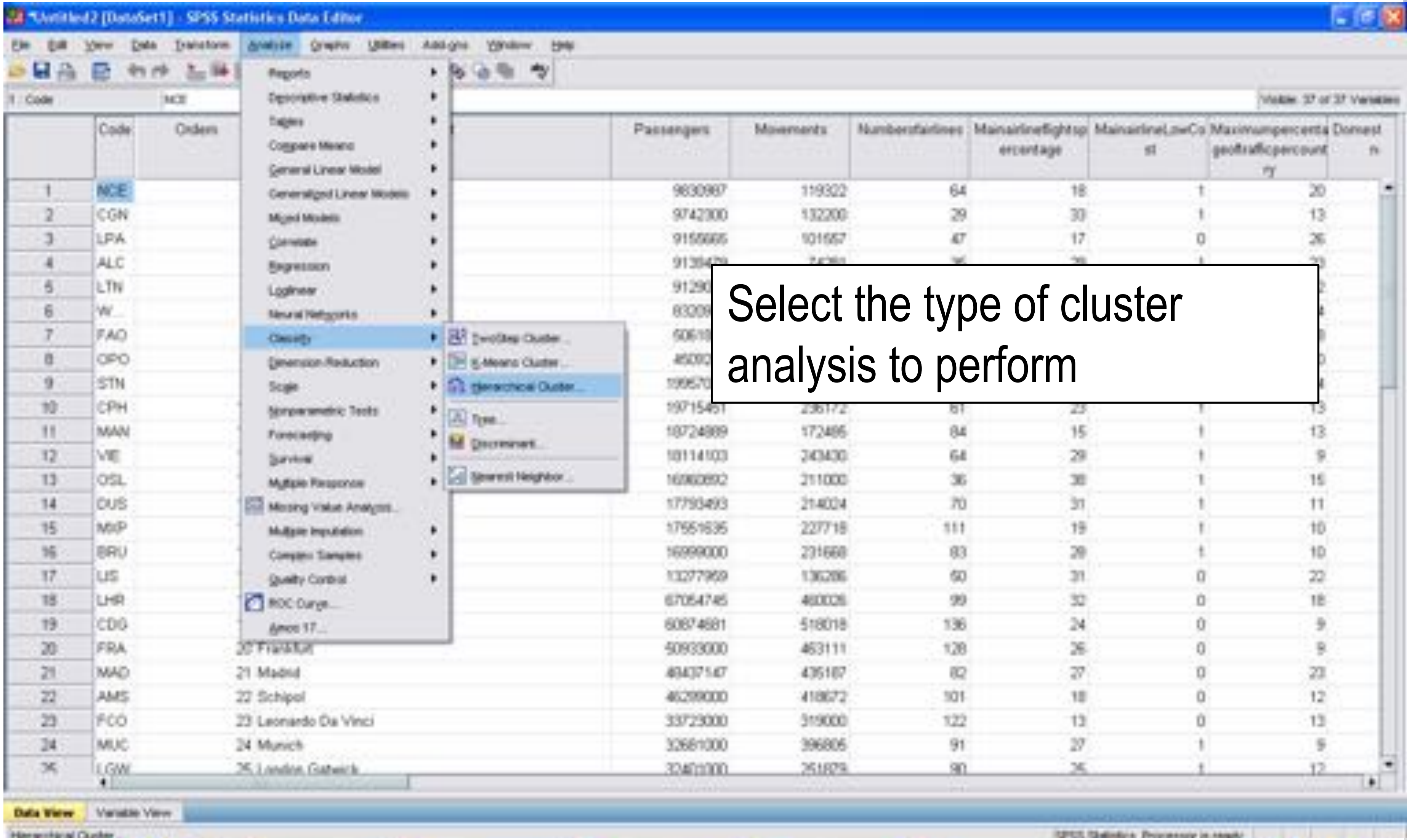
- ❒ **Validating** the cluster solution
  - ➢ Ensure that the cluster solution is representative of the general population

- ❒ **Perform cross-validation ALWAYS**
  - ➢ Perform cluster analysis on separate (re)samples and assess the correspondence of the results

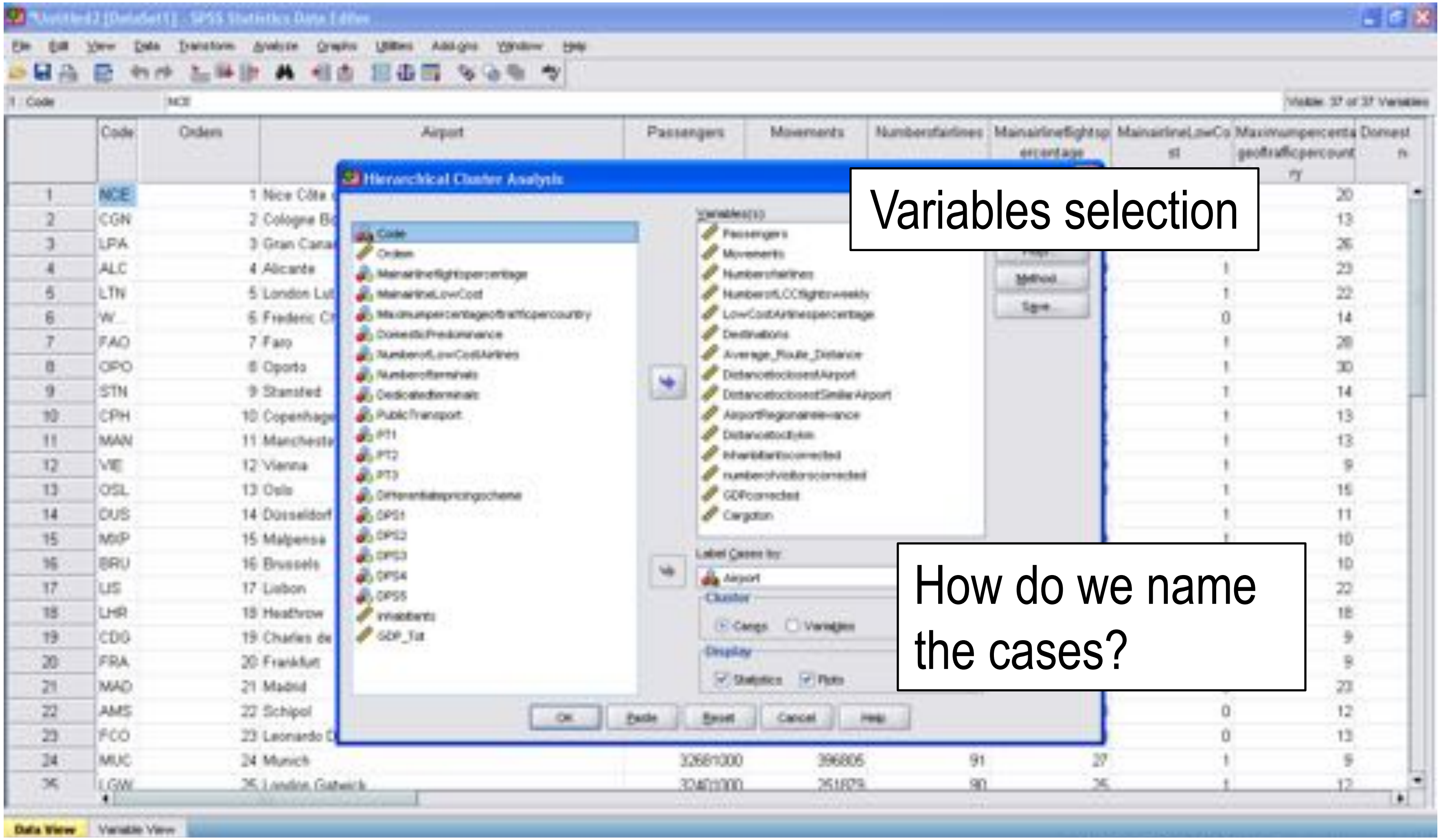
- ❒ **Criterion validity**
  - ➢ Using variables not selected to the cluster analysis but for which there are theoretical and relevant reasons that lead to the expectation of variation across the clusters

Example

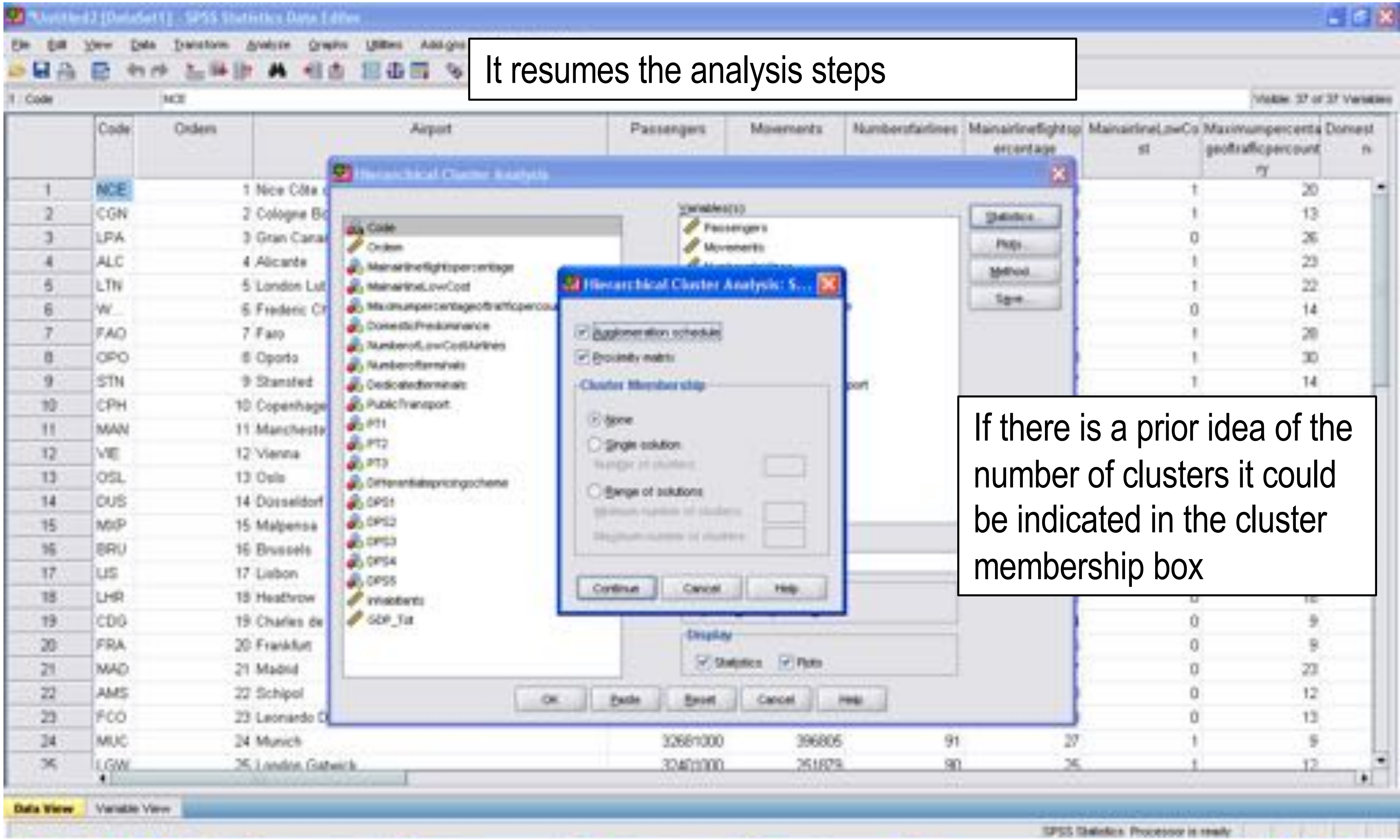
- Use the excel file "Dados_Aeroportos_Clusters" to build an hierarchical and a k-means cluster analysis.
- Use only the metric variables.



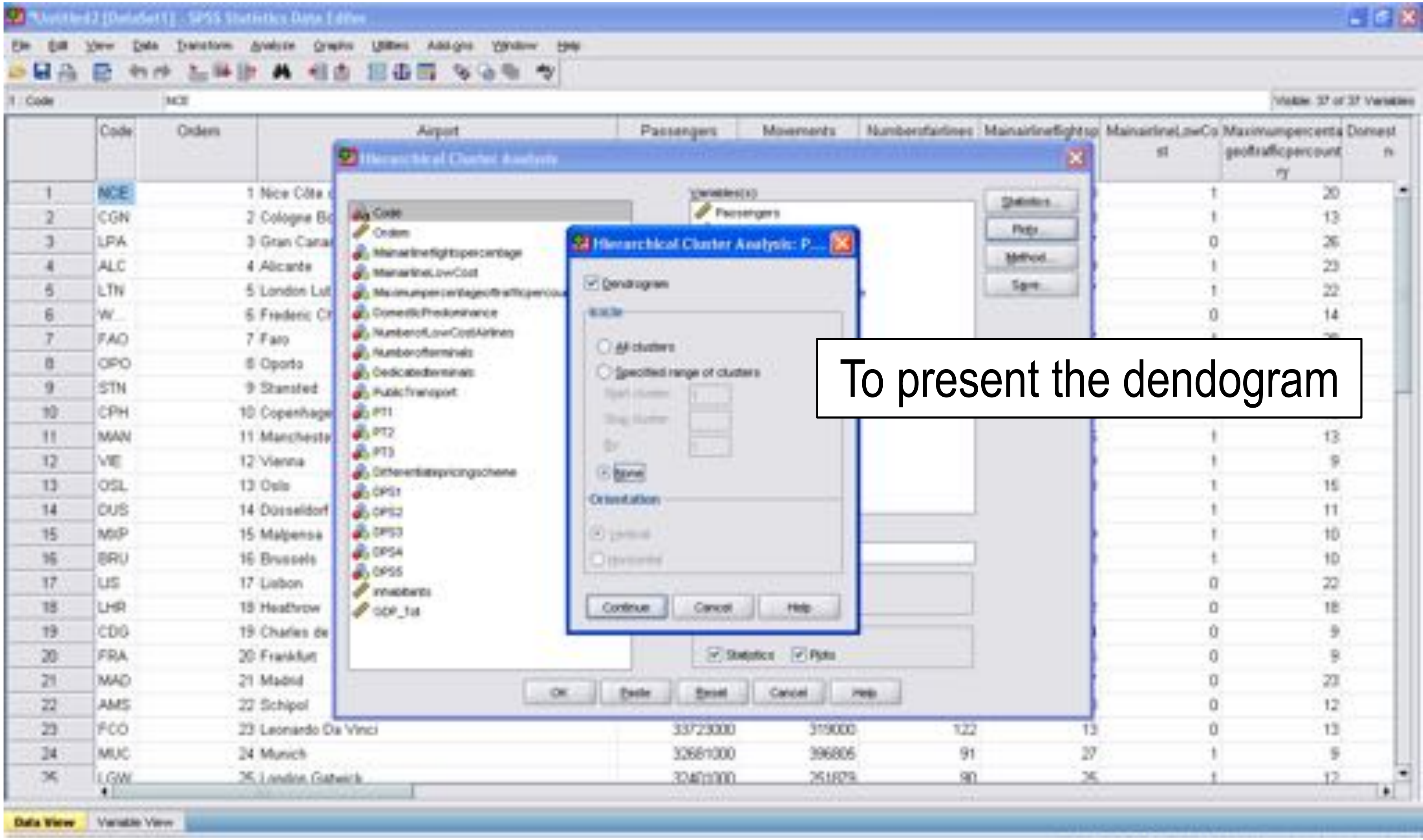
Select the type of cluster analysis to perform


Hierarchical Cluster Analysis
Variables selection
How do we name the cases?

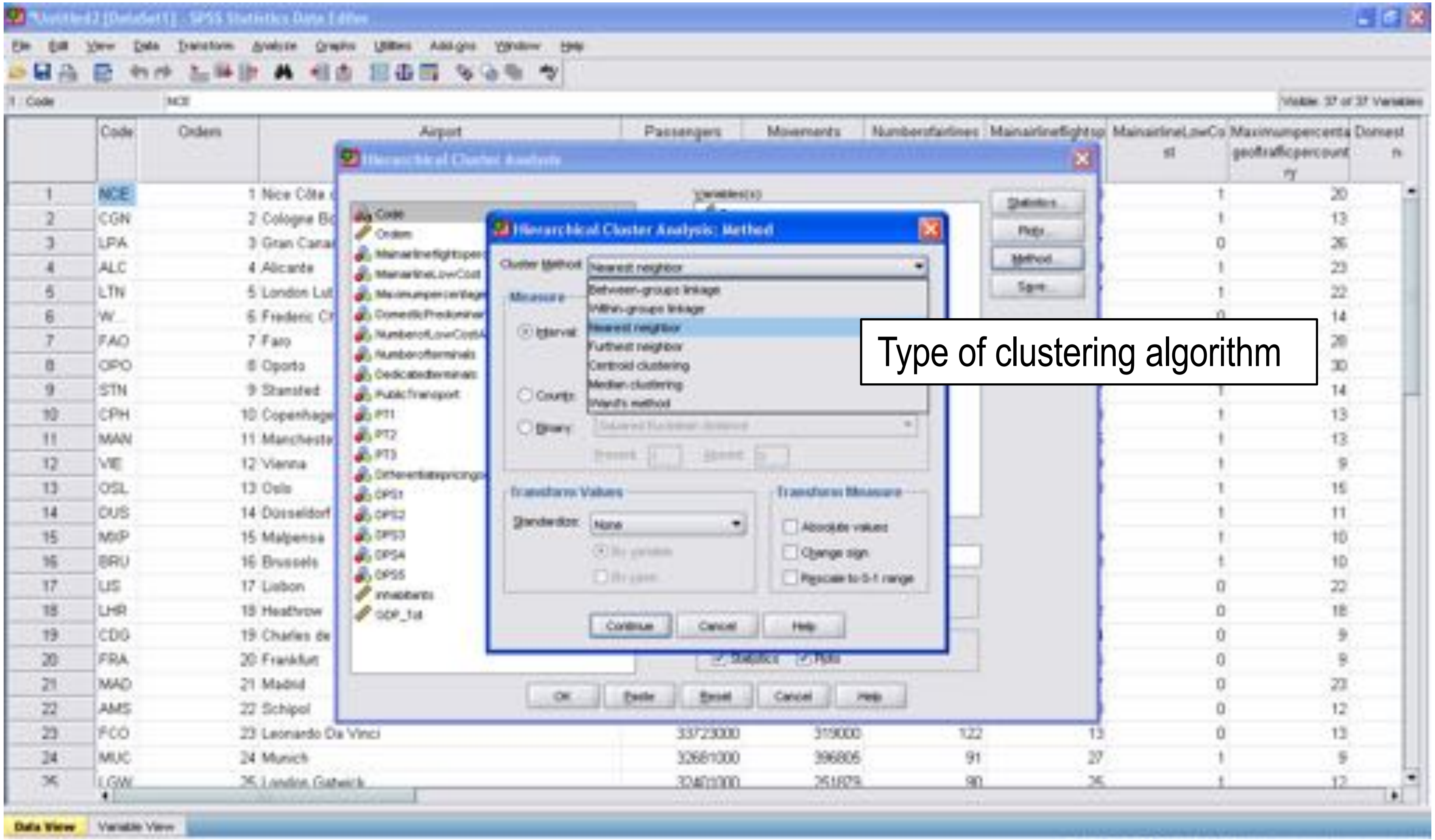
It resumes the analysis steps

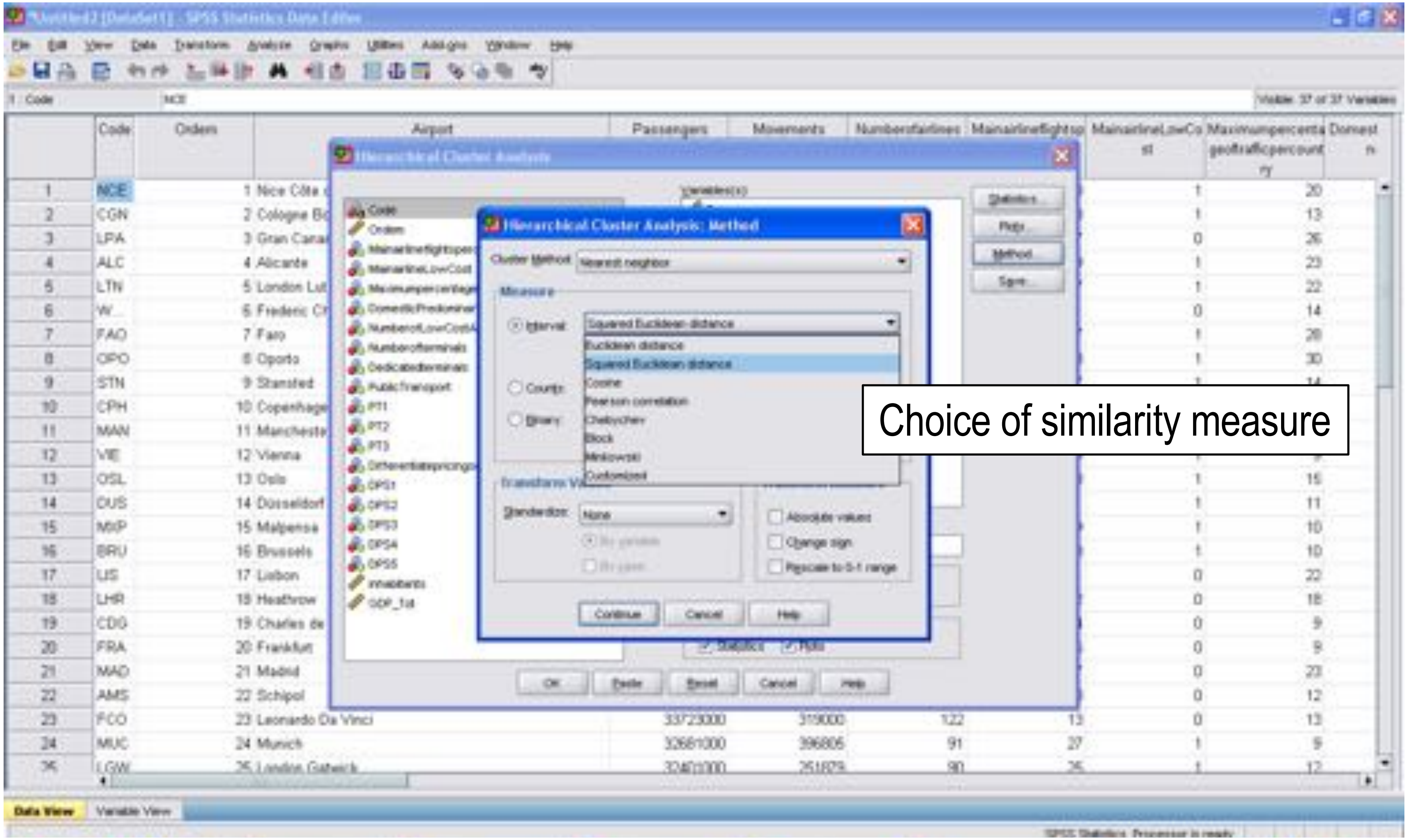


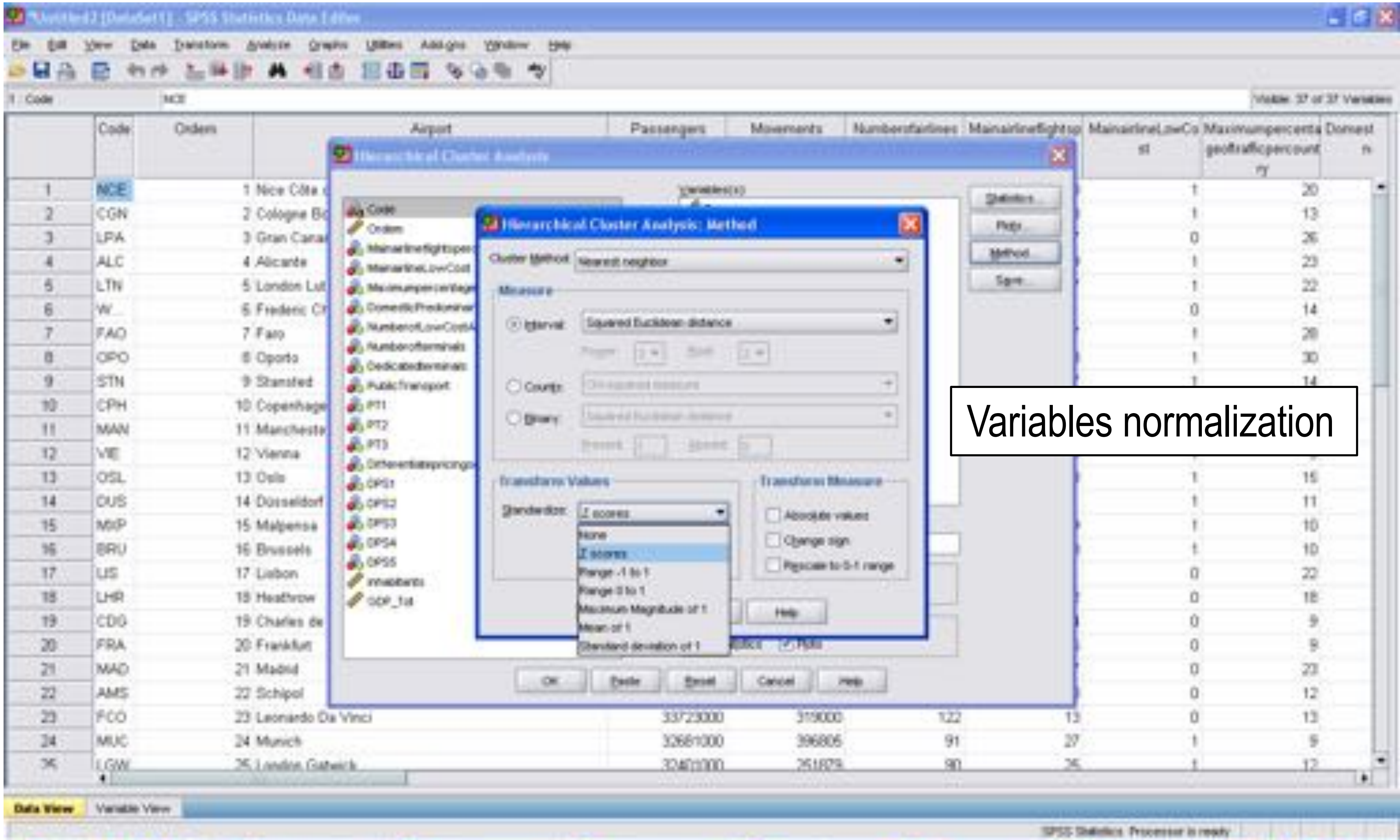
If there is a prior idea of the number of clusters it could be indicated in the cluster membership box

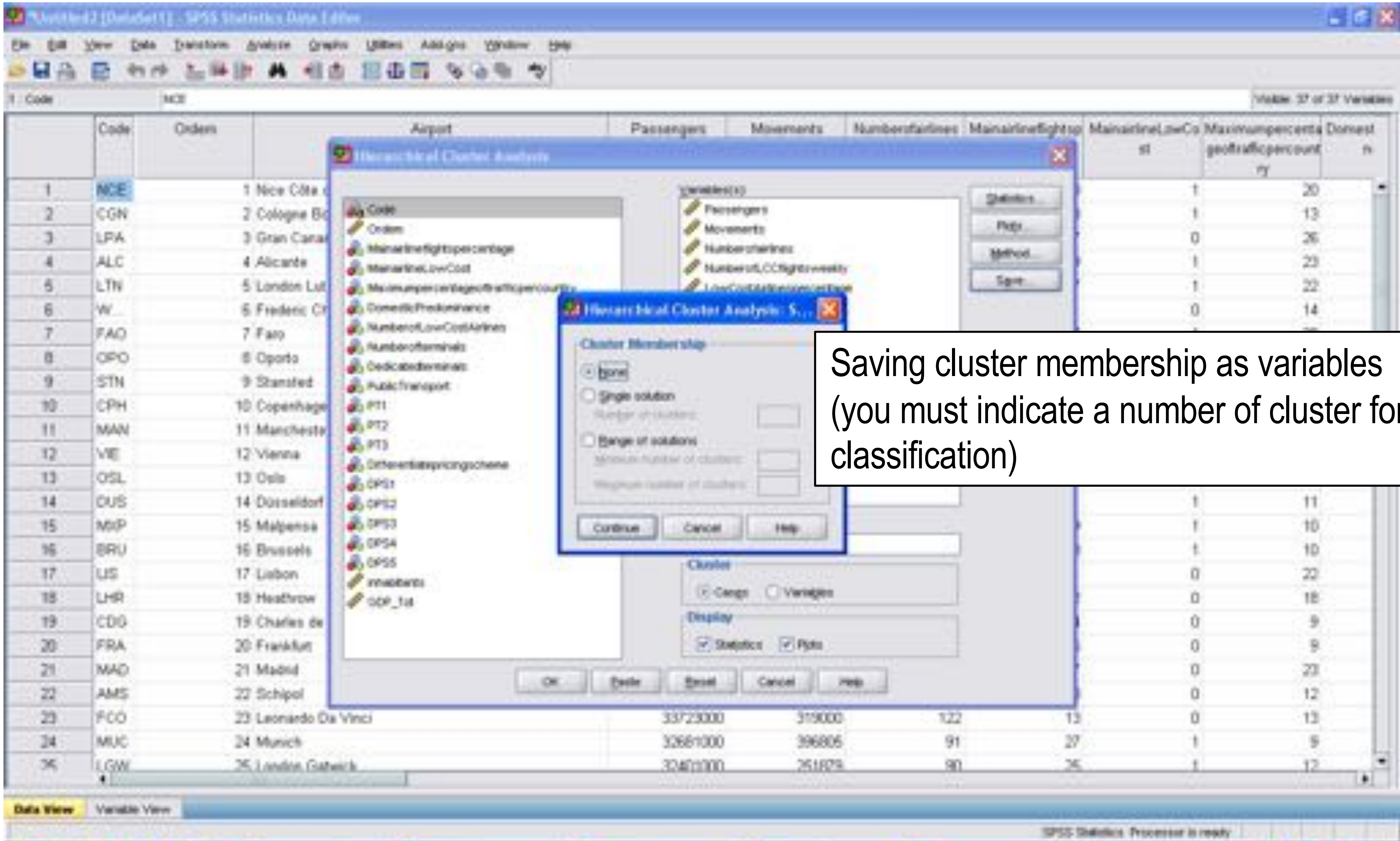


To present the dendogram


Hierarchical Cluster Analysis: Method
Nearest neighbor
Between-groups linkage
Within-groups linkage
Nearest neighbor
Furthest neighbor
Centroid clustering
Median clustering
Ward's method
Type of clustering algorithm


Hierarchical Cluster Analysis: Method
Cluster Method
Nearest neighbor
Measure
Interval
Squared Euclidean distance
Euclidean distance
Squared Euclidean distance
Cosine
Pearson correlation
Chebychev
Block
Minkowski
Customized
Counts
Binary
Standardize
None
Absolute values
Change sign
Rescale to 0-1 range
Continue
Cancel
Help
Choice of similarity measure


Variables normalization
Hierarchical Cluster Analysis
Hierarchical Cluster Analysis: Method
Cluster Method:
Nearest neighbor
Measure
Interval:
Squared Euclidean distance
Counts:
Binary:
Transform Values
Standardize:
Z scores
None
Range -1 to 1
Range 0 to 1
Maximum Magnitude of 1
Mean of 1
Standard deviation of 1
Transform Measure
Absolute values
Change sign
Rescale to 0-1 range
Help
OK
Paste
Reset
Cancel
Statistics...
Plots...
Method...
Save...
Data View
Variable View



Saving cluster membership as variables (you must indicate a number of cluster for classification)

Proximity matrix

Case	1:Nice Côte d'Azur	2:Cologne Bonn	3:Gran Canaria	4:Alicante	5:London Luton	6:Frederic Chopin	7:Faro	8:Oporto	9:Stansted	Cop
1:Nice Côte d'Azur	.000	6.645	9.039	3.313	19.418	19.872	15.253	7.247	12.459	
2:Cologne Bonn	6.645	.000	8.688	3.826	9.119	25.313	14.738	9.705	2.695	
3:Gran Canaria	9.039	8.688	.000	4.290	17.066	20.029	7.321	5.358	13.262	
4:Alicante	3.313	3.826	4.290	.000	13.791	21.387	8.273	3.763	8.107	
5:London Luton	19.418	9.119	17.066	13.791	.000	44.462	25.190	15.809	7.325	
6:Frederic Chopin	19.872	25.313	20.029	21.387	44.462	.000	11.659	16.550	32.313	
7:Faro	15.253	14.738	7.321	8.273	25.190	11.659	.000	4.726	21.134	
8:Oporto	7.247	9.705	5.358	3.763	15.809	16.550	4.726	.000	15.487	
9:Stansted	12.459	2.695	13.262	8.107	7.325	32.313	21.134	15.487	.000	
10:Copenhagen	8.237	10.593	15.192	10.790	29.498	20.649	22.368	20.165	13.721	
11:Manchester	11.071	10.127	14.196	12.638	28.508	21.908	23.908	22.674	12.971	
12:Vienna	11.085	11.397	11.572	10.248	31.561	24.604	21.994	21.447	13.407	
13:Oslo	10.938	15.960	14.134	12.656	28.500	13.117	17.013	15.720	17.764	
14:Düsseldorf	11.689	5.879	11.562	9.874	24.733	29.861	23.406	21.279	8.489	
15:Malpensa	10.840	11.875	17.179	16.108	27.960	25.145	28.833	25.136	13.911	
16:Brussels	7.426	6.535	12.541	9.669	24.908	26.856	24.922	20.624	10.032	
17:Lisbon	17.749	23.699	14.847	18.708	43.135	2.220	10.668	16.063	30.938	
18:Heathrow	72.948	65.818	68.186	74.705	88.337	99.080	96.716	96.893	64.718	
19:Charles de Gaulle	66.721	66.601	72.230	72.536	93.777	85.612	93.116	94.353	65.060	
20:Frankfurt	51.474	50.796	50.826	53.853	80.273	53.291	63.301	69.336	50.912	
21:Madrid	19.950	28.565	30.639	27.171	51.216	31.380	41.298	37.350	29.231	
22:Schipol	32.476	33.465	38.615	36.271	63.015	44.814	51.489	54.302	34.373	
23:Leonardo Da Vinci	23.832	31.189	34.377	31.669	54.279	34.048	45.422	44.383	32.205	
24:Munich	29.783	27.951	34.633	33.253	56.836	30.541	41.758	47.163	28.409	
25:London Gatwick	22.443	12.764	25.154	23.297	26.700	41.899	40.715	37.483	10.650	
26:Barcelona	10.630	14.178	16.220	12.944	36.380	22.624	22.582	22.555	18.734	
27:Skavsta	30.291	24.485	25.117	23.275	9.885	50.356	29.307	22.377	19.154	
28:Girona	26.526	19.194	20.762	18.826	5.606	43.748	23.308	16.637	14.879	
29:Orly	15.206	8.982	16.182	14.653	17.396	45.869	33.435	25.057	11.393	
30:Euroairport Basel Mulhouse Freiburg	5.455	2.981	6.955	3.993	8.528	21.818	11.535	4.541	8.639	
31:Kaunas	27.562	28.571	17.368	19.395	28.159	21.642	5.890	8.249	34.984	
32:Beauvais-Tille	30.326	20.552	31.114	27.147	9.413	47.345	33.083	25.540	17.675	

- Squared Euclidean distance between cases
- How would this matrix look like if the variables were not standardized?

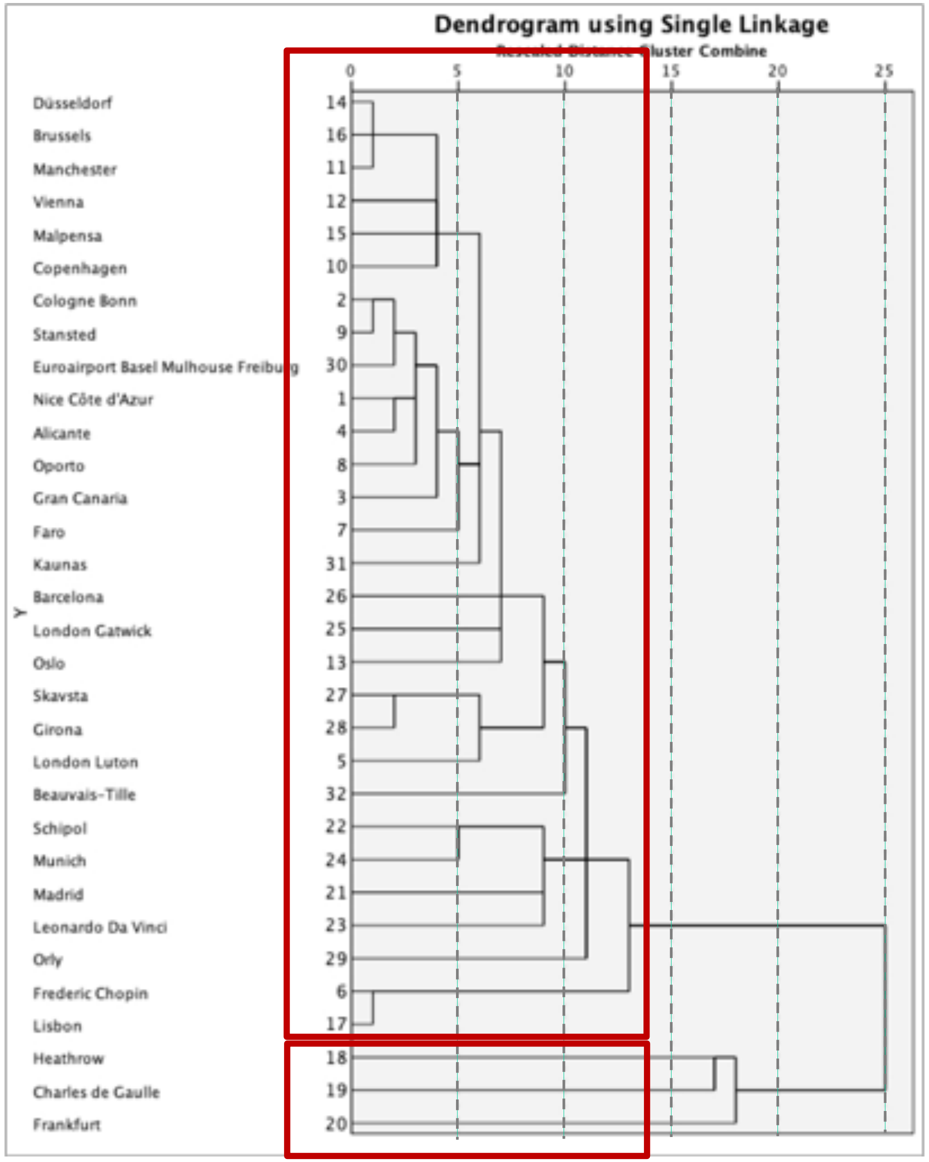
Agglomeration schedule

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	16	2.138	0	0	3
2	6	17	2.220	0	0	28
3	11	14	2.689	0	1	10
4	2	9	2.695	0	0	5
5	2	30	2.981	4	0	9
6	1	4	3.313	0	0	8
7	27	28	3.340	0	0	16
8	1	8	3.763	6	0	9
9	1	2	3.826	8	5	12
10	11	12	4.201	3	0	11
11	11	15	4.225	10	0	13
12	1	3	4.290	9	0	14
13	10	11	4.543	0	11	17
14	1	7	4.726	12	0	17
15	22	24	5.140	0	0	24
16	5	27	5.606	0	7	22
17	1	10	5.879	14	13	18
18	1	31	5.890	17	0	19
19	1	26	6.156	18	0	20
20	1	25	6.373	19	0	21
21	1	13	6.504	20	0	22
22	1	5	7.325	21	16	25
23	21	23	7.433	0	0	24
24	21	22	7.642	23	15	26
25	1	32	8.231	22	0	26
26	1	21	8.965	25	24	27
27	1	29	8.982	26	0	28
28	1	6	10.279	27	2	31
29	18	19	12.751	0	0	30
30	18	20	13.445	29	0	31
31	1	18	17.914	28	30	0

- Show the agglomeration order of the observations
- Cases 14 and 16 are the first to be agglomerated
  - In step 3 the case 11 joins that cluster
  - In step 10, 12 joins the cluster,
  - In step 11, 15 joins
  - In step 13, 10 joins
  - etc…

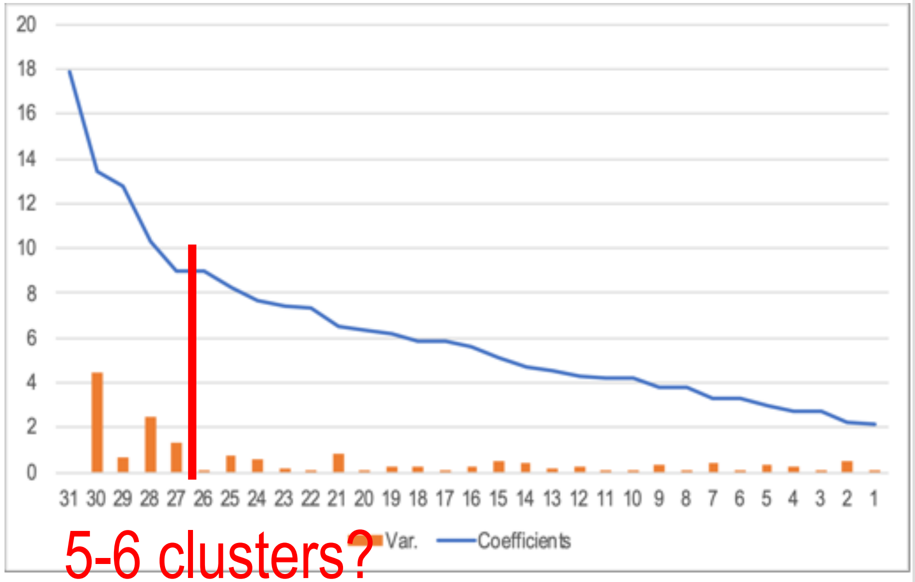
Dendrogram



- ❒ What is observed in the previous slide, could be also graphically seen here in the dendrogram

How many clusters should be retained

- We can test the possible number of clusters to retain by using two indicators
- Distance between clusters
  - Obtained from the "Agglomeration Schedule" directly in SPPS
- When the curve starts elbowing, agglomeration values don't change much and is a good indicator for the number of clusters to retain

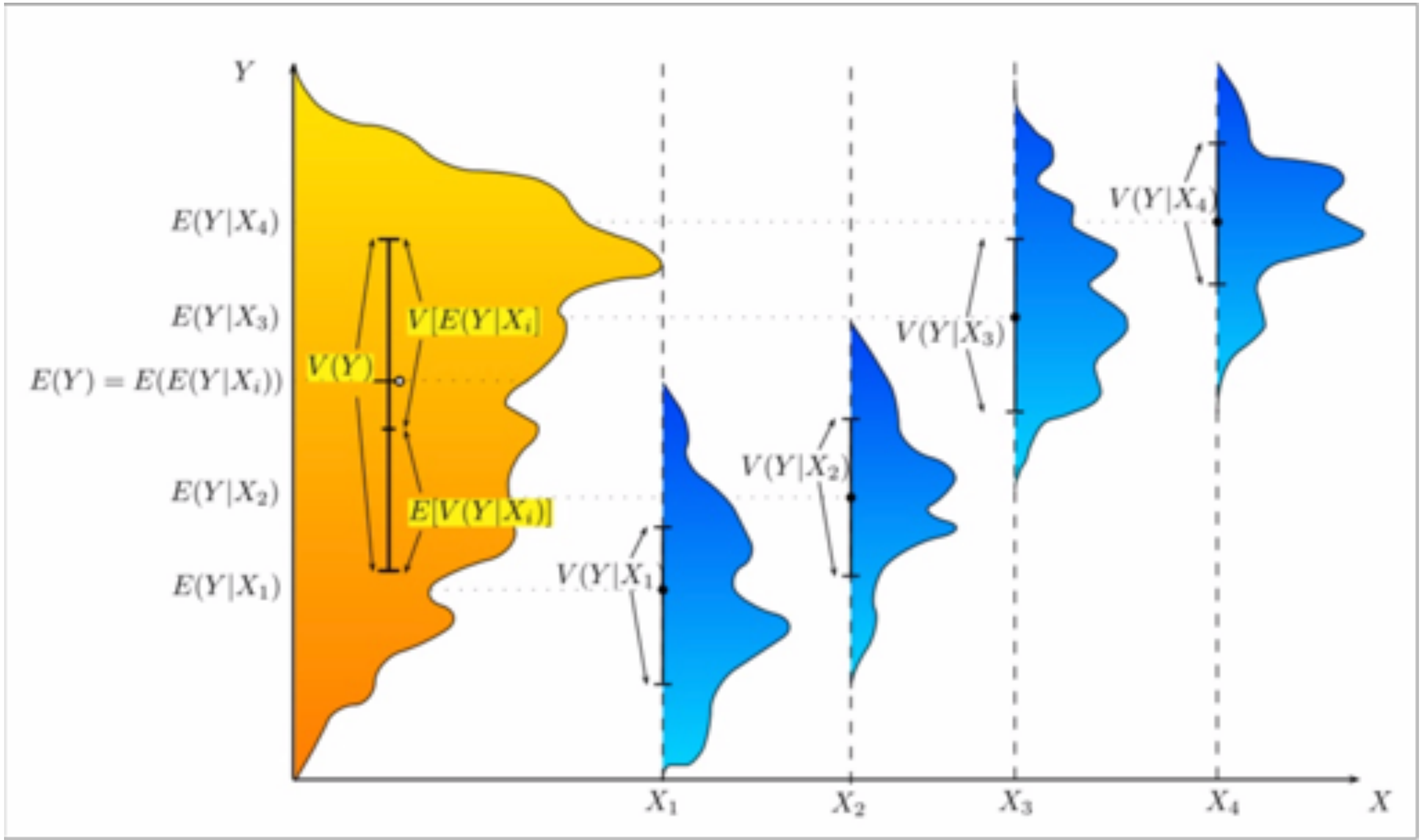


Analysis of variance

- The objective is to **compare differences between two or more groups for single metric dependent variable**.
- Do the means between the different groups 1 to *k* differ?
- Test of Hypothesis
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
  - H_a: $one\ or\ more\ of\ the\ groups\ has\ a\ differente\ mean$

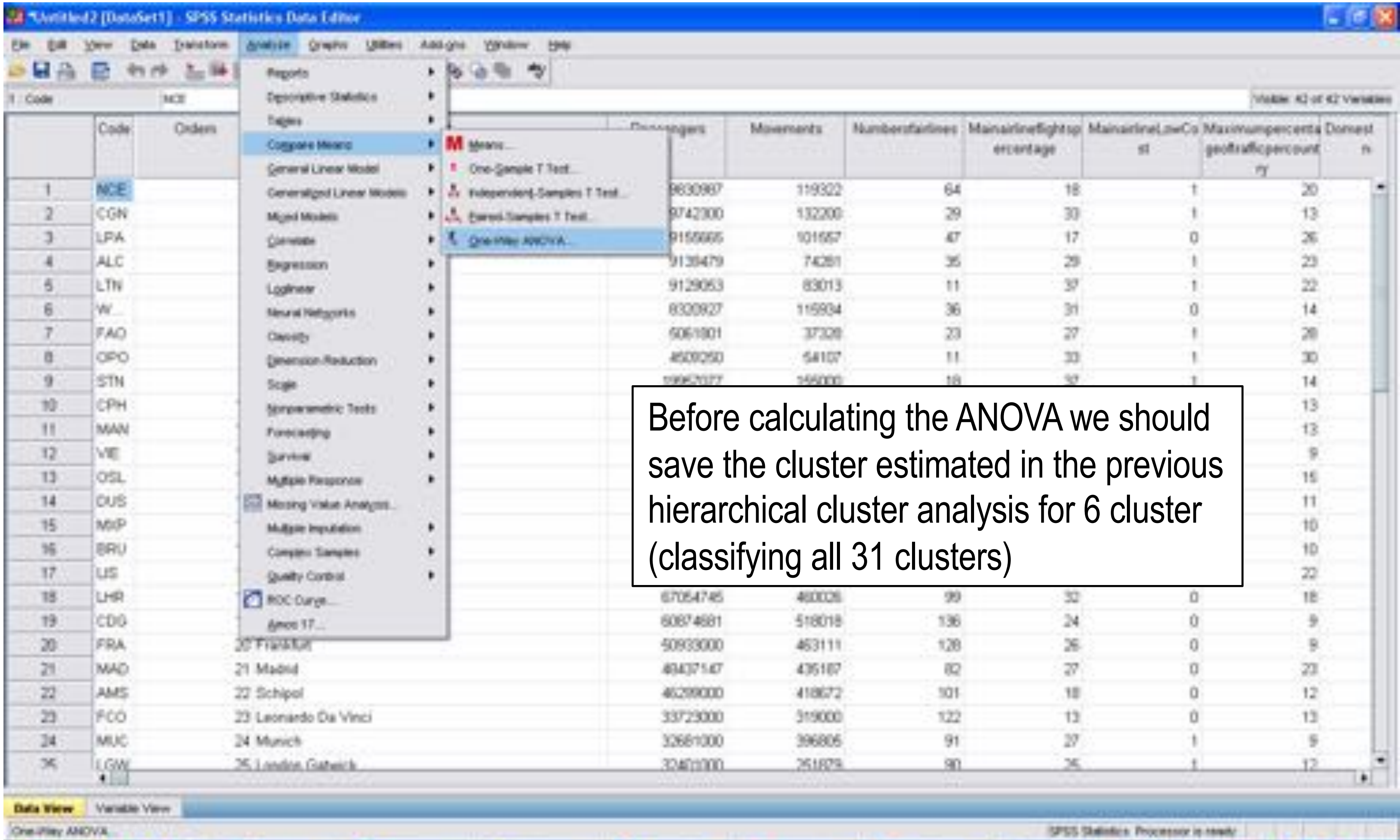
  - We want a low p-value in order to reject the null hypothesis that there are no differences between groups/clusters/profiles
  - This is calculated for each variable for clustering

Analysis of variance (II)



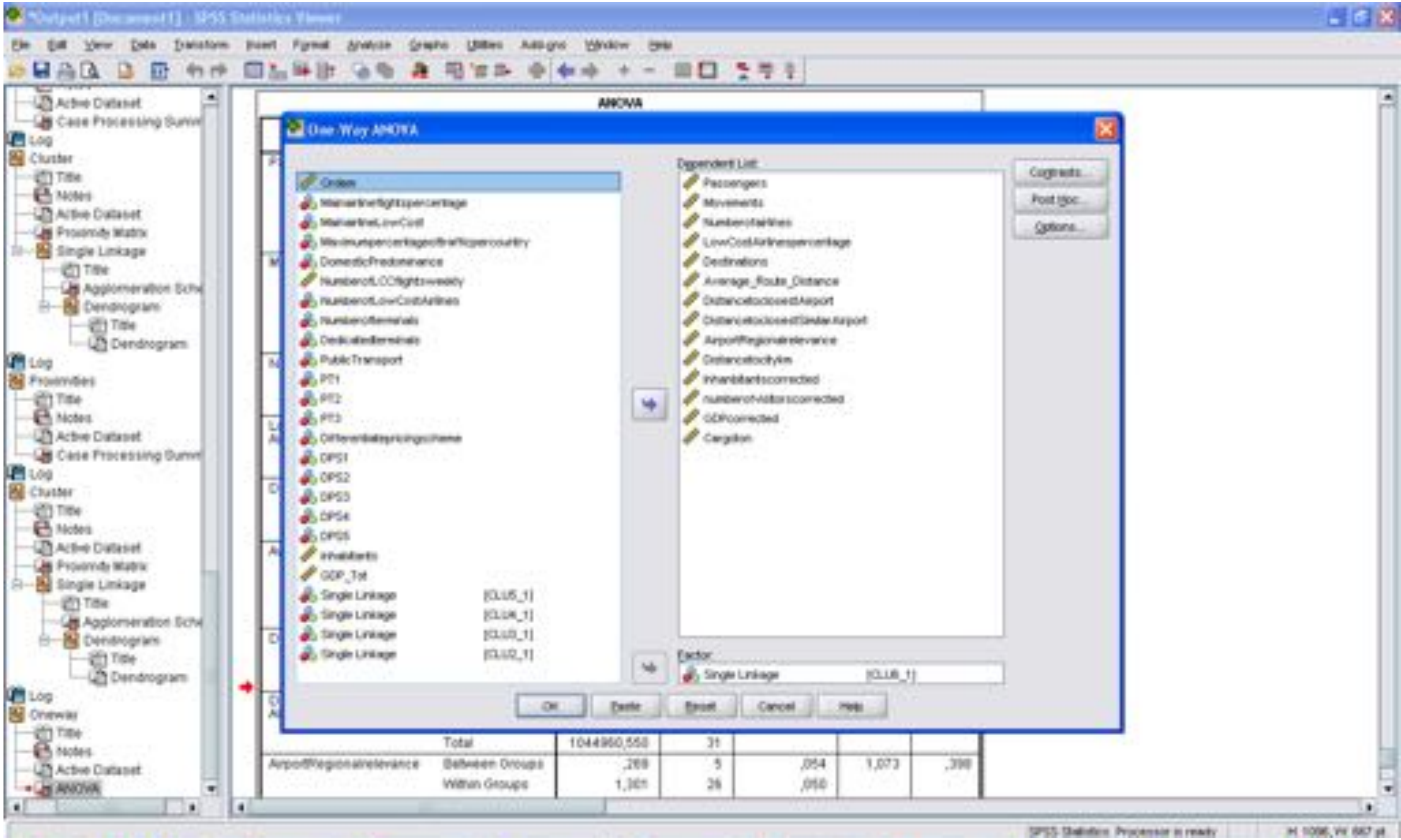
Fair fit (no strong overlapping between profiles)

Calculating the ANOVA



Before calculating the ANOVA we should save the cluster estimated in the previous hierarchical cluster analysis for 6 cluster (classifying all 31 clusters)

Calculating the ANOVA



Calculating the R squared (I)

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Passengers	Between Groups	5,275E15	5	1,055E15	6,335	,001
	Within Groups	4,330E15	26	1,665E14		
	Total	9,605E15	31			
Movements	Between Groups	2,613E11	5	5,227E10	3,599	,013
	Within Groups	3,776E11	26	1,452E10		
	Total	6,389E11	31			
Numberofairlines	Between Groups	14245,337	5	2849,067	2,035	,107
	Within Groups	36399,538	26	1399,982		
	Total	50644,875	31			
LowCost Airlinespercentage	Between Groups	3926,621	5	785,324	,845	,530
	Within Groups	24159,359	26	929,206		
	Total	28085,980	31			
Destinations	Between Groups	40642,654	5	8128,531	1,334	,281
	Within Groups	158390,846	26	6091,956		
	Total	199033,500	31			
Average_Route_Distance	Between Groups	1,716E7	5	3431191,675	9,224	,000
	Within Groups	9671910,500	26	371996,558		
	Total	2,683E7	31			
DistancetoclosestAirport	Between Groups	58556,061	5	11711,212	4,310	,005
	Within Groups	70645,176	26	2717,122		
	Total	129201,238	31			
DistancetoclosestSimilar Airport	Between Groups	336248,471	5	67249,694	2,467	,059
	Within Groups	708712,079	26	27258,157		
	Total	1044960,550	31			
AirportRegionalrelevance	Between Groups	,269	5	,054	1,073	,398
	Within Groups	1,301	26	,050		
	Total	1,570	31			
Distancetocitykm	Between Groups	1136,490	5	227,298	,312	,901
	Within Groups	18926,385	26	727,938		
	Total	20062,875	31			
Inhanbitantscorrected	Between Groups	2,843E13	5	5,685E12	,823	,545
	Within Groups	1,796E14	26	6,908E12		
	Total	2,080E14	31			
numberofvisitorscorrected	Between Groups	9,351E13	5	1,870E13	4,501	,004
	Within Groups	1,080E14	26	4,155E12		
	Total	2,015E14	31			
GDPcorrected	Between Groups	1,849E9	5	3,697E8	6,099	,001
	Within Groups	1,576E9	26	6,062E7		
	Total	3,425E9	31			
Cargoton	Between Groups	6,718E12	5	1,344E12	93,460	,000
	Within Groups	3,738E11	26	1,438E10		
	Total	7,092E12	31			

$$R^2 = \frac{SQC}{SQT} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{k} n_{ij}(\bar{X}_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{p}\sum_{j=1}^{k}\sum_{l=1}^{n_i}(X_{ijl} - \bar{X})^2}$$

, Where SQC is sum square between clusters and SQT is the total sum squares of ALL variable for all possible cluster (in this case, 2 to 31)

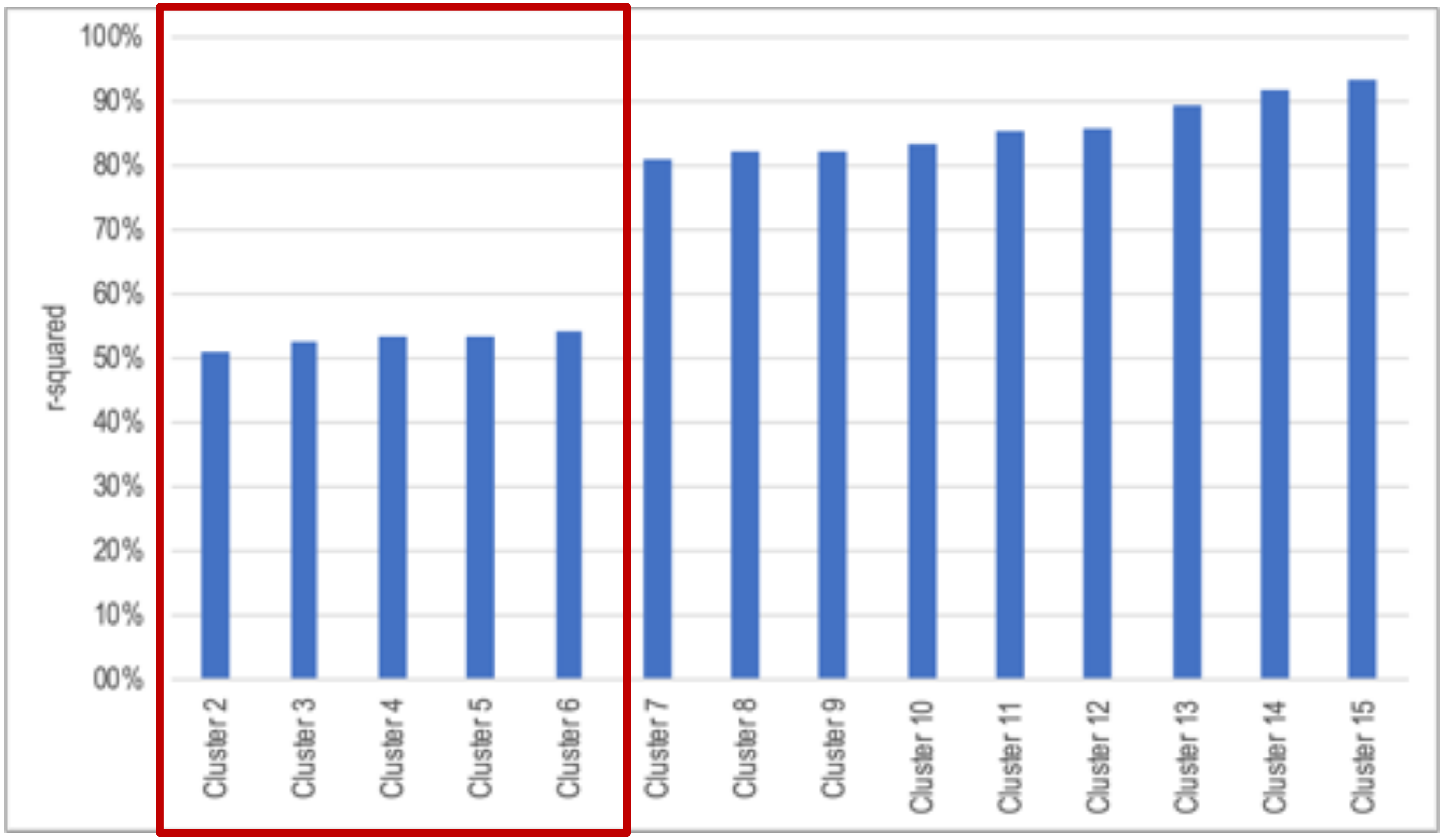
- When the slope in this curve starts to decrease we can use that value as the number of clusters to be retained

Calculating the R squared (II)

ANOVA								
		Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Passengers	Between Groups	5,00155E+15	5,13386E+15	5,20941E+15	5,20941E+15	5,27486E+15	7,86807E+15	7,97311E+15
	Within Groups	4,60312E+15	4,47081E+15	4,39526E+15	4,39526E+15	4,32981E+15	1,7366E+15	1,63156E+15
	Total	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15
Movements	Between Groups	2,50844E+11	2,52973E+11	2,58498E+11	2,58498E+11	2,61337E+11	4,77948E+11	4,82128E+11
	Within Groups	3,88036E+11	3,85969E+11	3,80477E+11	3,80477E+11	3,77603E+11	1,60985E+11	1,56805E+11
	Total	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11
Numberofairlines	Between Groups	13217,082	13675,082	14122,208	14122,208	14245,337	24551,602	26297,065
	Within Groups	37427,793	36969,793	36522,667	36522,667	36399,538	26093,273	24347,81
	Total	50644,875	50644,875	50644,875	50644,875	50644,875	50644,875	50644,875
NumberofLCCflightsweekly	Between Groups	375371,742	383076,408	388183,422	388183,422	412290,103	598594,378	665194,397
	Within Groups	1146417,977	1138713,31	1133606,296	1133606,296	1109496,615	923195,341	856595,321
	Total	1521789,719	1521789,719	1521789,719	1521789,719	1521789,719	1521789,719	1521789,719
LowCostAirlinespercentage	Between Groups	2610,639	2628,345	3455,873	3455,873	3926,621	6939,692	9870,596
	Within Groups	25475,341	25457,635	24630,107	24630,107	24159,359	21146,288	18115,424
	Total	28085,98	28085,98	28085,98	28085,98	28085,98	28085,98	28085,98
Destinations	Between Groups	29692,695	36353,362	40574,333	40574,333	40642,654	81376,795	91550,44
	Within Groups	169340,805	162680,138	158459,167	158459,167	158390,846	117656,705	107483,06
	Total	199033,5	199033,5	199033,5	199033,5	199033,5	199033,5	199033,5
Average_Route_Distance	Between Groups	13536244,21	16269984,88	16346213,41	16346213,41	17155958,38	18890152,31	19428136,48
	Within Groups	13291624,67	10557984	10481655,46	10481655,46	9671910,5	7929716,566	7399732,393
	Total	26827868,88	26827868,88	26827868,88	26827868,88	26827868,88	26827868,88	26827868,88
DistancetoclosestAirport	Between Groups	737,264	7542,464	54810,777	54810,777	58556,061	61335,099	61726,088
	Within Groups	128463,974	121658,774	74390,461	74390,461	70645,176	67866,139	67475,15
	Total	129201,238	129201,238	129201,238	129201,238	129201,238	129201,238	129201,238
DistancetoclosestSimilarAirport	Between Groups	3102,437	84857,775	301164,374	301164,374	336248,471	700949,813	713716,543
	Within Groups	1041858,113	960102,775	743796,177	743796,177	708712,079	344010,737	331244,007
	Total	1044960,55	1044960,55	1044960,55	1044960,55	1044960,55	1044960,55	1044960,55
AirportRegionalrelevance	Between Groups	0,006	0,044	0,178	0,178	0,269	0,488	0,704
	Within Groups	1,564	1,526	1,392	1,392	1,301	1,081	0,866
	Total	1,57	1,57	1,57	1,57	1,57	1,57	1,57
Distancetocitykm	Between Groups	125,036	217,703	909,708	909,708	1136,49	1374,534	4585,625
	Within Groups	19937,839	19845,172	19153,167	19153,167	18926,385	18688,341	15477,25
	Total	20062,875	20062,875	20062,875	20062,875	20062,875	20062,875	20062,875
Inhabitantscorrected	Between Groups	1,96918E+13	2,13285E+13	2,79542E+13	2,79542E+13	2,84262E+13	8,54278E+13	9,83589E+13
	Within Groups	1,88347E+14	1,8671E+14	1,80084E+14	1,80084E+14	1,79612E+14	1,22611E+14	1,09679E+14
	Total	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14
numberofvisitorscorrected	Between Groups	6,0153E+13	8,87765E+13	8,92389E+13	8,92389E+13	9,35054E+13	1,42211E+14	1,4385E+14
	Within Groups	1,41389E+14	1,12763E+14	1,12303E+14	1,12303E+14	1,08036E+14	5,93311E+13	5,76916E+13
	Total	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14
GDPcorrected	Between Groups	841025182,6	1063645183	1564001374	1564001374	1848700900	1924127954	1931097394
	Within Groups	2583840355	2361223055	1860866864	1860866864	1576167338	1500740284	1493770844
	Total	3424868238	3424868238	3424868238	3424868238	3424868238	3424868238	3424868238
Cargoton	Between Groups	6,54216E+12	6,71801E+12	6,71815E+12	6,71815E+12	6,71839E+12	6,80184E+12	6,80653E+12
	Within Groups	5,50036E+11	3,74187E+11	3,74048E+11	3,74048E+11	3,73805E+11	2,90359E+11	2,85663E+11
	Total	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12
	Between Groups	5,08819E+15	5,25094E+15	5,33358E+15	5,33358E+15	5,40377E+15	8,10296E+15	8,22361E+15
	Within Groups	4,93379E+15	4,77104E+15	4,6884E+15	4,6884E+15	4,61821E+15	1,91899E+15	1,79837E+15
	Total	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16
	r_square	**50,8%**	**52,4%**	**53,2%**	**53,2%**	**53,9%**	**80,9%**	**82,0%**

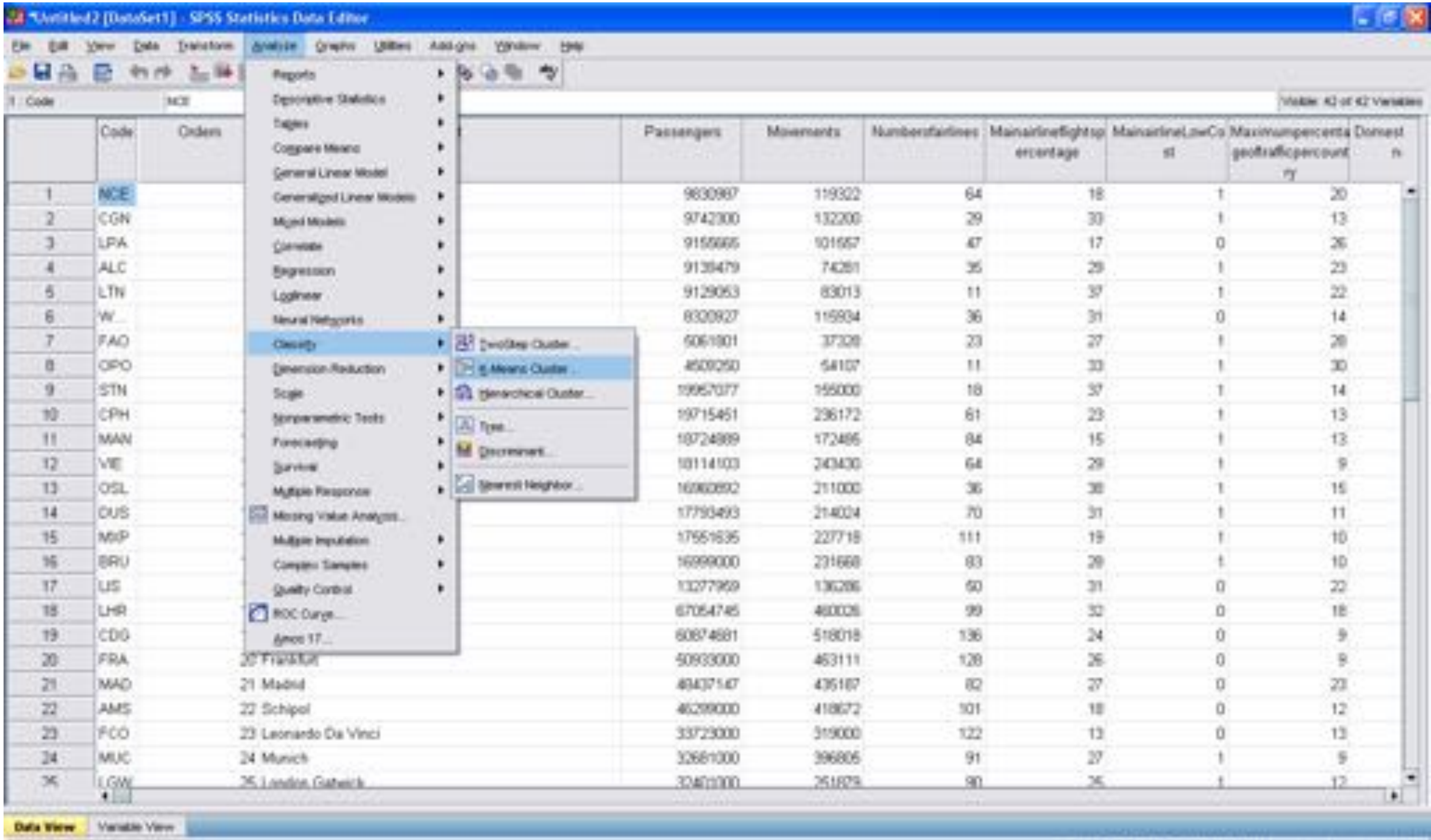
		Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Cluster 14	Cluster 15
Passengers	Between Groups	7,97564E+15	8,08389E+15	8,2778E+15	8,28517E+15	8,60842E+15	8,85942E+15	9,00387E+15
	Within Groups	1,62903E+15	1,52077E+15	1,32686E+15	1,31949E+15	9,96248E+14	7,45249E+14	6,00794E+14
	Total	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15	9,60467E+15
Movements	Between Groups	4,83067E+11	4,89817E+11	5,17113E+11	5,20182E+11	5,30363E+11	5,48898E+11	5,67779E+11
	Within Groups	1,55866E+11	1,49116E+11	1,2182E+11	1,18751E+11	1,08575E+11	90235478221	71159438298
	Total	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11	6,38803E+11
Numberofairlines	Between Groups	26333,065	27133,065	32966,597	32950,983	34377,938	36437,142	38817,375
	Within Groups	24311,81	23511,81	17678,278	17693,892	16266,937	14207,733	11827,5
	Total	50644,875	50644,875	50644,875	50644,875	50644,875	50644,875	50644,875
NumberofLCCflightsweekly	Between Groups	742200,647	747200,647	883689,608	889496,434	984223,552	989791,619	1102914,838
	Within Groups	779589,071	774589,071	638100,111	632293,284	537566,167	531998,1	418874,881
	Total	1521789,719	1521789,719	1521789,719	1521789,719	1521789,719	1521789,719	1521789,719
LowCostAirlinespercentage	Between Groups	9877,027	9877,468	17794,058	18076,427	18458,283	18658,772	22667,247
	Within Groups	18108,953	18108,512	10291,922	10009,553	9627,697	9427,208	5418,733
	Total	28085,98	28085,98	28085,98	28085,98	28085,98	28085,98	28085,98
Destinations	Between Groups	98860,69	99041,19	121574,722	123153,235	128527,063	131702,6	151753,571
	Within Groups	100172,81	99992,31	77458,778	75880,265	70506,438	67330,9	47279,929
	Total	199033,5	199033,5	199033,5	199033,5	199033,5	199033,5	199033,5
Average_Route_Distance	Between Groups	19860442,73	19881743,23	20870139,93	21090200,18	21454179,96	21559432,78	22080033,71
	Within Groups	6967426,143	6946125,643	5957728,944	5737668,696	5373688,917	5268436,1	4748835,167
	Total	26827868,88	26827868,88	26827868,88	26827868,88	26827868,88	26827868,88	26827868,88
DistancetoclosestAirport	Between Groups	66015,973	66314,376	66718,979	70013,404	71784,176	71918,566	96176,7
	Within Groups	63185,265	62886,862	62482,259	59187,834	57417,062	57282,672	33024,538
	Total	129201,238	129201,238	129201,238	129201,238	129201,238	129201,238	129201,238
DistancetoclosestSimilarAirport	Between Groups	753558,979	755178,195	784319,717	905575,364	924377,034	924811,107	927832,343
	Within Groups	291401,572	289782,355	260640,833	139385,187	120583,516	120149,443	117128,207
	Total	1044960,55	1044960,55	1044960,55	1044960,55	1044960,55	1044960,55	1044960,55
AirportRegionalrelevance	Between Groups	0,704	0,71	1,008	1,065	1,16	1,181	1,186
	Within Groups	0,866	0,86	0,561	0,505	0,41	0,388	0,384
	Total	1,57	1,57	1,57	1,57	1,57	1,57	1,57
Distancetocitykm	Between Groups	4615,875	4857,875	15317,431	15627,444	16528,271	16555,608	16555,78
	Within Groups	15447	15205	4745,444	4435,431	3534,604	3507,267	3507,095
	Total	20062,875	20062,875	20062,875	20062,875	20062,875	20062,875	20062,875
Inhabitantscorrected	Between Groups	1,01838E+14	1,0434E+14	1,26556E+14	1,27126E+14	1,42787E+14	1,4518E+14	1,55829E+14
	Within Groups	1,062E+14	1,03698E+14	8,14821E+13	8,09127E+13	6,52516E+13	6,28582E+13	5,2109E+13
	Total	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14	2,08039E+14
numberofvisitorscorrected	Between Groups	1,43852E+14	1,52886E+14	1,58464E+14	1,58464E+14	1,58514E+14	1,8212E+14	1,8443E+14
	Within Groups	5,76898E+13	4,86534E+13	4,30781E+13	4,30781E+13	4,3028E+13	1,94214E+13	1,71111E+13
	Total	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14	2,01542E+14
GDPcorrected	Between Groups	1972057394	1973337394	2035999210	2169607748	2249707196	2252067238	2602839428
	Within Groups	1452810844	1451530844	1388869028	1254930490	1175161042	1172791000	822028809,5
	Total	3424868238	3424868238	3424868238	3424868238	3424868238	3424868238	3424868238
Cargoton	Between Groups	6,88026E+12	6,89402E+12	6,8941E+12	6,89813E+12	6,89768E+12	6,90422E+12	6,90862E+12
	Within Groups	2,05938E+11	1,98179E+11	1,98012E+11	1,96069E+11	1,94519E+11	1,87979E+11	1,83579E+11
	Total	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12	7,09226E+12
	Between Groups	8,22875E+15	8,34851E+15	8,57024E+15	8,57818E+15	8,94715E+15	9,19417E+15	9,35171E+15
	Within Groups	1,79328E+15	1,67347E+15	1,45179E+15	1,4438E+15	1,07483E+15	8,27808E+14	6,7027E+14
	Total	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16	1,0022E+16
	r_square	**82,1%**	**83,3%**	**85,5%**	**85,6%**	**89,3%**	**91,7%**	**93,3%**

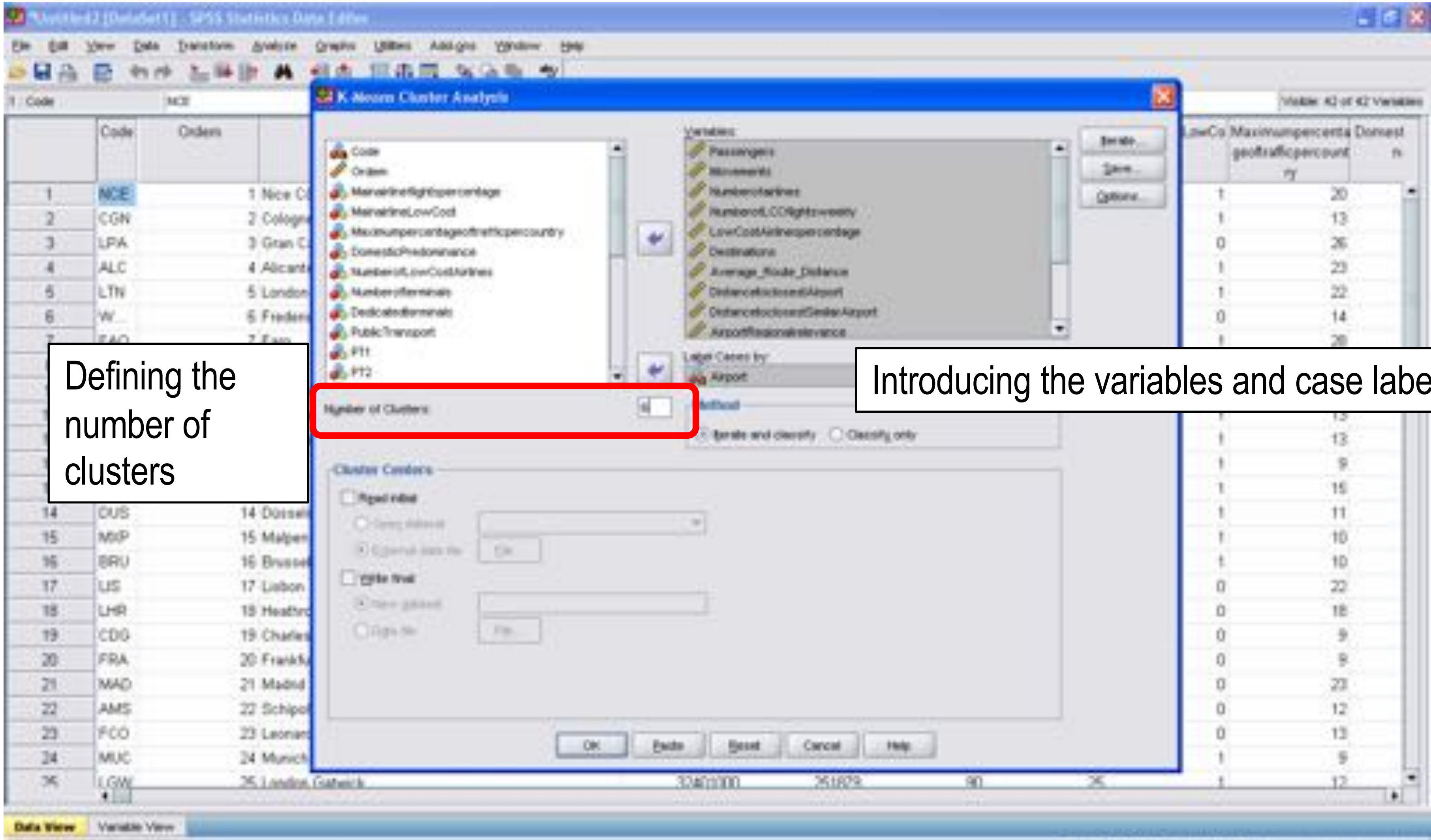
Calculating the R-Squared (III)

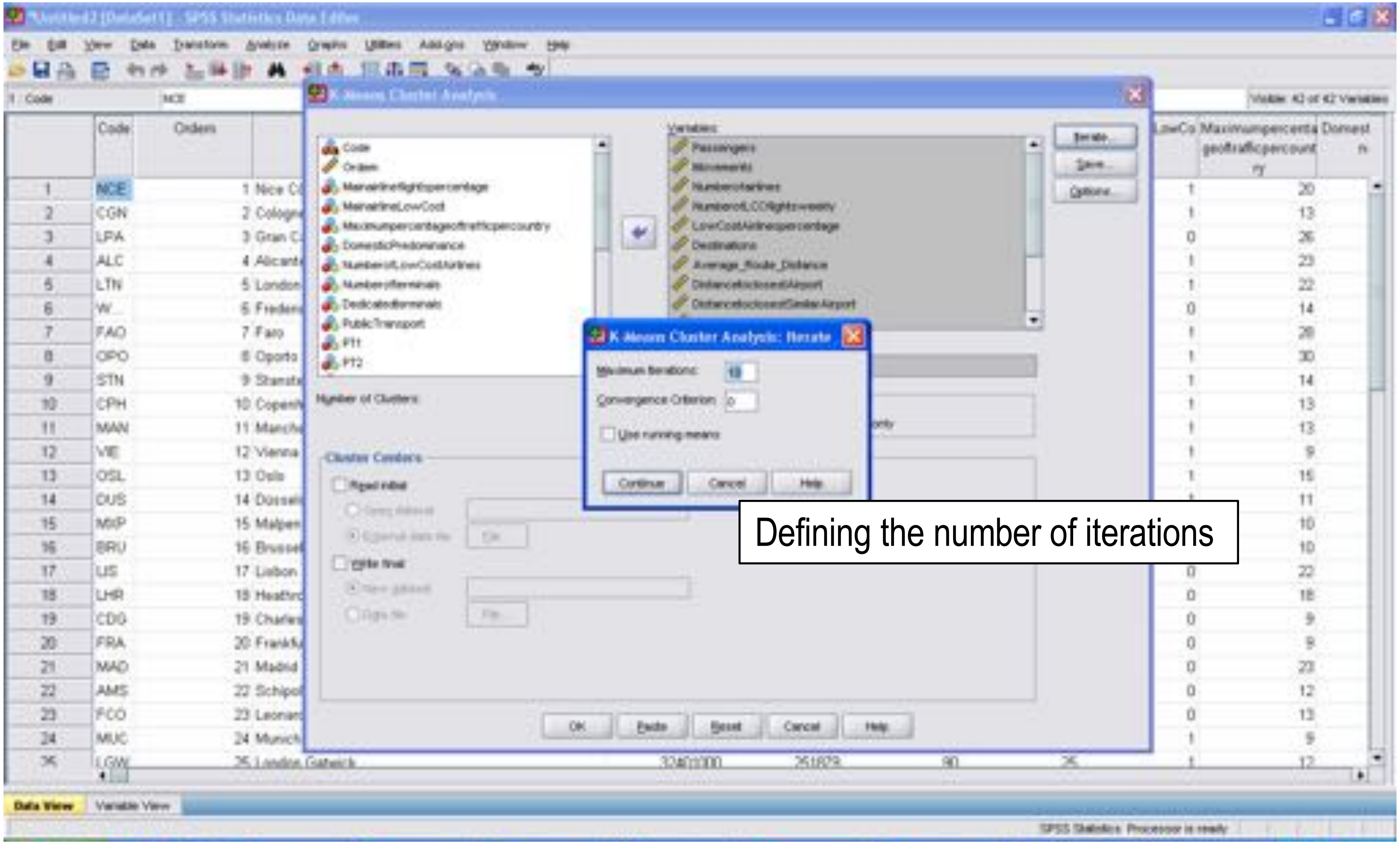


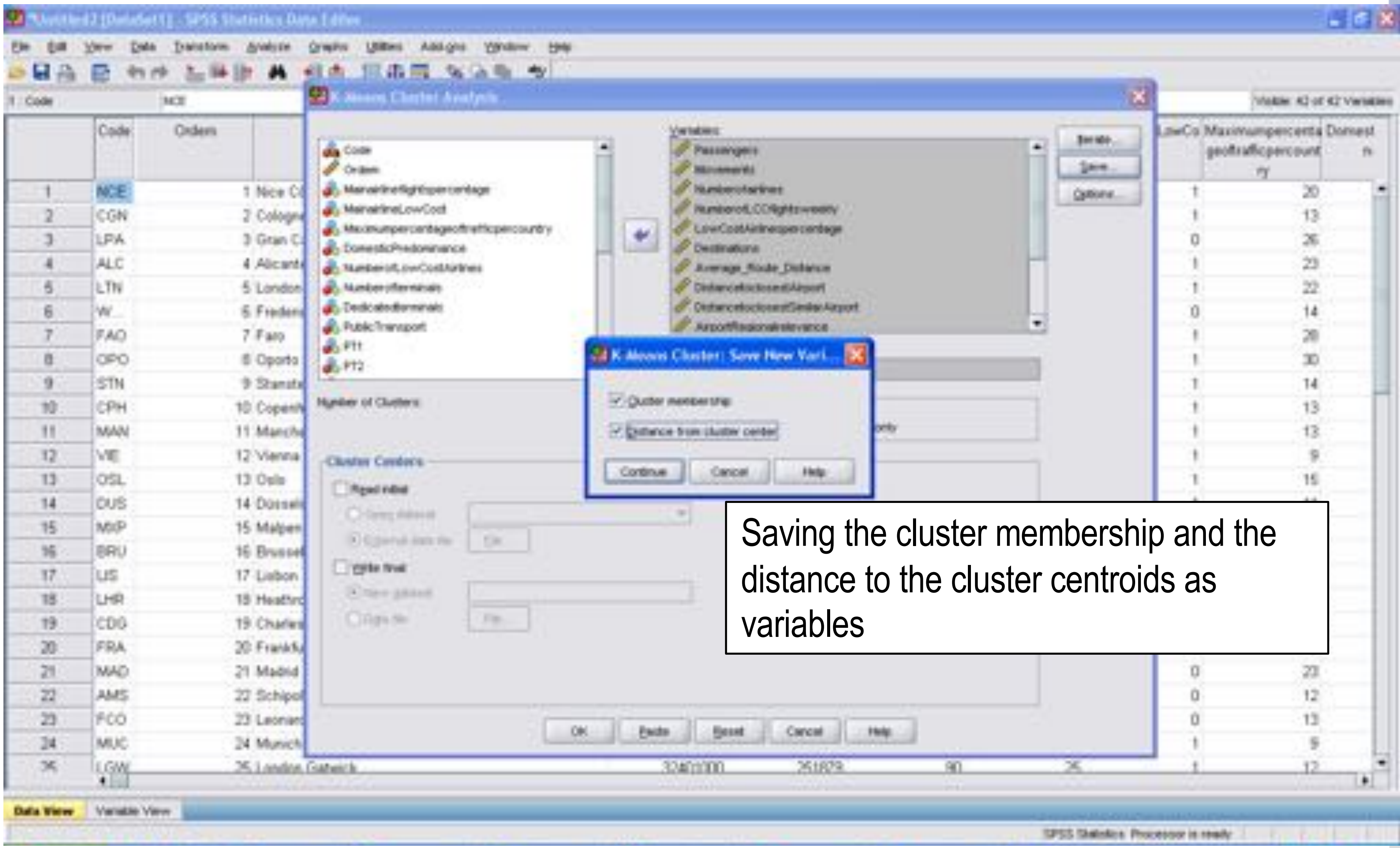
- When the slope in this curve starts to decrease we can use that value as the number of clusters to be retained

K-Means Cluster Analysis

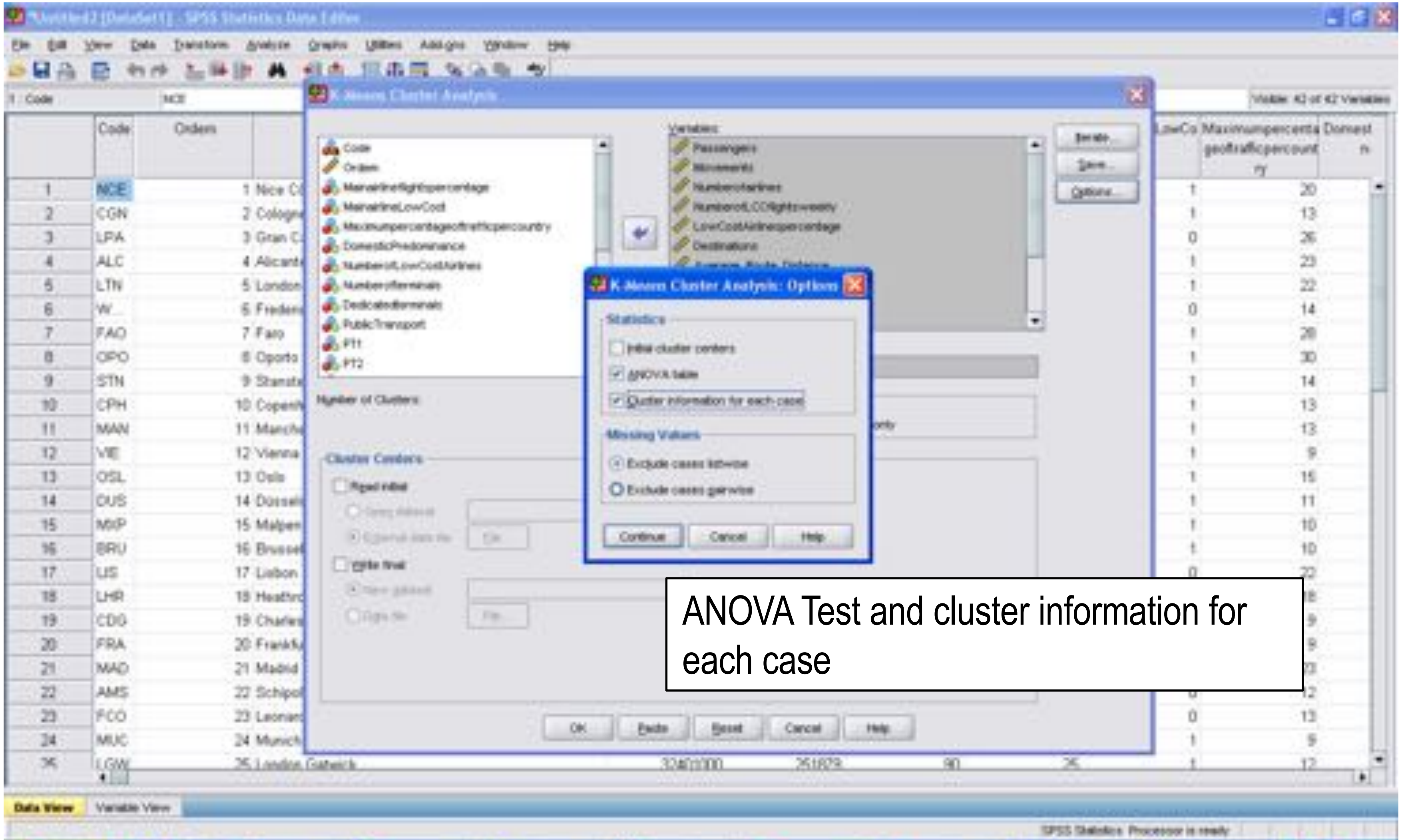



Defining the number of clusters
Introducing the variables and case labels


K-Means Cluster Analysis: Iterate
Maximum Iterations:
Convergence Criterion:
Use running means
Continue
Cancel
Help
Defining the number of iterations



Saving the cluster membership and the distance to the cluster centroids as variables



ANOVA Test and cluster information for each case

Iteration history

Iteration History[a]

Iteration	Change in Cluster Centers					
	1	2	3	4	5	6
1	2627040.47	2465953.12	3145745.99	4796145.13	3448153.30	5568562.14
2	.000	.000	.000	1519993.27	.000	1247734.55
3	.000	.000	.000	880925.328	.000	1132612.67
4	.000	583063.645	.000	.000	.000	714269.454
5	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 5. The minimum distance between initial centers is 7616767.215.

- In each iteration we can see the changes in the cluster centers.
- It takes five iterations to achieve stability in the cluster centers

Cluster membership

Cluster Membership

Case Number	Airport	Cluster	Distance
1	Nice Côte d'Azur	6	633575.348
2	Cologne Bonn	6	905977.842
3	Gran Canaria	6	2741793.99
4	Alicante	6	2012775.99
5	London Luton	6	2361196.87
6	Frederic Chopin	6	3661620.00
7	Faro	4	1548420.06
8	Oporto	4	1235193.79
9	Stansted	2	2061153.47
10	Copenhagen	2	1258789.16
11	Manchester	2	1578103.98
12	Vienna	2	2338111.06
13	Oslo	2	2575736.28
14	Düsseldorf	2	1630939.32
15	Malpensa	2	3754740.59
16	Brussels	2	2207785.26
17	Lisbon	6	4285616.72
18	Heathrow	3	3145745.99
19	Charles de Gaulle	3	3145745.99
20	Frankfurt	5	2851901.20
21	Madrid	5	1106910.45
22	Schipol	5	3448153.30
23	Leonardo Da Vinci	1	3396777.68
24	Munich	1	2627040.47
25	London Gatwick	1	3503480.52
26	Barcelona	1	5081897.12
27	Skavsta	4	2270076.58
28	Girona	4	2397493.97
29	Orly	2	6618034.89
30	Euroairport Basel Mulhouse Freiburg	4	1400337.74
31	Kaunas	4	3726858.87
32	Beauvais-Tille	4	5285551.97

- This table allow us to see to which cluster each airport belongs, and how far from the cluster center it is (Distance).

Final cluster centers

Final Cluster Centers

	Cluster					
	1	2	3	4	5	6
Movements	311666	213832	489022	47828	438990	108942
Passengers	31556671	18991616	63964713	3525226	48556382	9799481
Numberofairlines	100	63	118	9	104	39
NumberofLCCflightsweekly	575	466	720	129	624	289
LowCostAirlinespercentage	18.5113981	21.0234578	9.4177065	81.5573663	13.2320097	39.0757326
Destinations	242	194	243	69	267	126
Average_Route_Distance	2375	2354	4711	1508	2996	1879
Distancetoclosest Airport	78.687948	65.192021	41.646632	127.634463	69.905070	114.012213
Distancetoclosest SimilarAirport	301.160832	227.991590	216.046139	151.568404	493.947213	246.425329
AirportRegionalrelevance	.79083269	.72968233	.69882575	.61286704	.94346327	.73175601
Distancetocitykm	30	22	24	44	11	17
Inhanbitantscorrected	7617844.75	4618520.16	7472517.75	2217498.04	6278199.17	3367686.10
numberofvisitors corrected	5030589.60	2133872.11	9190522.00	1032241.19	4248534.17	1547505.75

- This is the average distance of each variable to every cluster center

Distance between cluster centers

Distances between Final Cluster Centers

Cluster	1	2	3	4	5	6
1		13239225.7	32674743.1	28826758.7	17070813.5	22441303.3
2	13239225.7		45613559.8	15691247.8	29687580.9	9295953.06
3	32674743.1	45613559.8		61215186.1	16225560.1	54856924.1
4	28826758.7	15691247.8	61215186.1		45329812.2	6399878.48
5	17070813.5	29687580.9	16225560.1	45329812.2		38961172.2
6	22441303.3	9295953.06	54856924.1	6399878.48	38961172.2	

- Distances between each cluster centers.

Number of Cases in each Cluster

Cluster	1	4,000
	2	9,000
	3	2,000
	4	7,000
	5	3,000
	6	7,000
Valid		32,000
Missing		,000

Variables and clusters

- The objective is to evaluate which variables allow the cluster separation
- If one variable discriminates well the clusters then its variability between clusters is high and its variability within the clusters is small
- The F-test null hypothesis is "variance within cluster is equal to variance between clusters"
- F =QMC/QME
  - QMC – Cluster mean square
  - QME – Error means square
- Higher F means higher contribution to the clusters definition

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Movements	1.219E+11	5	1.139E+9	26	106.974	.000
Passengers	1.893E+15	5	5.340E+12	26	354.501	.000
Numberofairlines	7907.850	5	427.139	26	18.514	.000
NumberofLCCflightsweekly	223132.430	5	15620.291	26	14.285	.000
LowCostAirlinespercentage	4159.735	5	280.281	26	14.841	.000
Destinations	29930.563	5	1899.257	26	15.759	.000
Average_Route_Distance	3746221.58	5	311413.884	26	12.030	.000
DistancetoclosestAirport	5177.546	5	3973.596	26	1.303	.293
DistancetoclosestSimilarAirport	52703.433	5	30055.515	26	1.754	.158
AirportRegionalrelevance	.050	5	.051	26	.983	.447
Distancetocitykm	741.823	5	628.991	26	1.179	.346
Inhanbitantscorrected	2.232E+13	5	3.710E+12	26	6.016	.001
numberofvisitorscorrected	2.894E+13	5	2.187E+12	26	13.233	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Exercise

- Standardize the variables and run again the K-means cluster, compare the obtained results and analyze which variables should be removed

Recommended Readings

- Hair, Joseph P. et al (1995) “Multivariate Data Analysis with Readings”, Fourth Edition, Prentice Hall - Chapter 9
- Maroco, João (2003) “Análise Estatística com utilização do SPSS”, Ed. Sílabo– Capítulo 11