



## Phd Program in Transportation

# **Transport Demand Modeling**

**Filipe Moura** 

# Session 5

**Cluster Analysis** 

Phd in Transportation / Transport Demand Modelling

# What is cluster analysis?



- Cluster analysis is a exploratory technique of multivariate
- □ 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



analysis

# **Objectives in Cluster Analysis**









Source: Hair et al (2010)



# **3D example**









Phd in Transportation / Transport Demand Modelling



# 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**

- > There should be always a strong conceptual analysis
  - Why do groups exist?

**Strong conceptual framework** 

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



# How does cluster analysis works? (I)

#### **G** 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



**FEUP** 

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**



FEUP







# Example of proximity matrix calculation



Phd in Transportation / Transport Demand Modelling

# **Practical considerations**



- **FEUP**
- 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

# 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





Outliers

FEUP

- 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
- □ Squared (absolute) Euclidian distance
  - Better in computational aspects

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

q

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







# 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

Phd in Transportation / Transport Demand Modelling







IJÎ

UPERIO

#### □ 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



# SUPERIOR

FEUP



**Measuring similarity (V)** 

- > *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



**FEUP** 



#### □ Russel & Rao

> Similarity 
$$S_{ij} = \frac{a}{a+b+c+d}$$

#### Johnson and Wichern

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

Phd in Transportation / Transport Demand Modelling







- **FEUP**
- 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**



- > 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?







#### Phd in Transportation / Transport Demand Modelling

# 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**.







#### Phd in Transportation / Transport Demand Modelling

#### 31/70

### □ 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 (II)** 







# **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)









**FEUP** 

# **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



abcdefg

b c d e f g

а

# 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



37/70







# 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





FEUP

- □ 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





**FEUP** 

- Use the excel file "Dados\_Aeroportos\_Clusters" to build an hierarchical and a k-means cluster analysis.
- $\hfill\square$  Use only the metric variables.



**FEUP** 

E (# 🗱

\*

Visible: 37 of 37 Versibles

20 13 25

18

9

9

23 12

13

5 12

MainairlineLowCo Maximumpercenta Domest geofauficpercount

Ó

14

#### Contribed 2 (DataSet1) - SPSS Statistics Data Editor

Code	NOT	Opportative Statistics						
	Code Ordern	Tages Cogpare Means General Livear Model	:	Passengers	Movements	Numbersfarines	Manairinefightsp ercertage	MainairlineLo st
-	CON .	Ceneralized Linear Moders		963090/	119522	04	10	
	0.00	Mged Models		9/42300	132200		- 30	
	UPA NC	Converse		9150665	001007		17	
	ALC .	Segression		91394				
6	W	Logineer		83009	Select	the tyr	he of cl	uster
7	FAO	Casalt	BR Developer Charter	50619	001001			aotor
8	040	Conversion Reduction	· Di K. Means Chatter	45000	analya	ic to na	orform	
9	STN	Scale	+ C) theather Outer	199670	anaiyə	15 IU PE		
10.	CPH	Storpersentric Tests		19715451	2361/2		- 25	5
11	MAN	Firelading	P CT	10724989	172495	84	15	
12	VE	Saven	M Decrement	10114103	243430	64	29	
13	OSL.	Multiple Response	. Sevent Heightor	16/960892	211000	36	38	
14	OUS	Moong Value Analysis.		17793493	214004	70	.31	
15	MIP	Multiple imputition		17551635	227718	311	19	
95 -	BRU	Comptex Samples		16999000	231668	63	29	
17	us	Quelty Control		13277969	136286	60	31	
18	LHR	ROC Curye		67054745	460026	99	32	
19	CDO	Amon 17		60874681	518018	136	24	
20	FRA	20 Fighton	_	50933000	453111	1,28	26	
21	MAD	21 Madrid		49437547	435107	82	27	
22	AMS	22 Schipel		46299000	418672	301	.10	
23	FCO	23 Leonardo Da Vinci		33723000	319000	122	13	
24	MUC	24 Munich		32681000	396805	91	27	
36	I GW	25 London Gatwick		32401000	261679	90	36	

Harmonia al Carden

OPEN Challengian Decisionary in campbi-

0 ò

Û

0

ū

0

Ú



Code		MOT							Weble: 37 of 3	17 Metallele
	Code	Ordern	Aippt	Passe	ngers Movements	Numbersfa	dives Mainairing	elightsp MainairlineLowC	o Maximumpercente	Domest.
	1574	200				Carda Marca	ercard	lage si	geofisatiopercount	n
1	NOF	1 Mice Citta	Hierarchical Cluster Asalysis			\ / ·		1 (1	-77 - 20	
2	CON	2 Colone Rd		14 16	(generatives)	Vari	ables	selection	13	
3	LPA	3 Gran Carta	Come		Passengers				25	
4	ALC	4 Alicante	A Manual Andread Andread Andread		Movements			-	1 20	
5	LTN	5 London Lut	A Manaritel.ov/Cod		/ NumberofLCCSgterase	Adv.	Marcol		1 22	
6	W.	5 Frederic Cr	a Malanumpercentageoffranticpercountry		& LowContAltimespercent	tage.	10-4	-	0 14	See 11
7	FAO	7 Faio	DokesticPredonnance		Destinations				1 20	
8	040	6 Oporto	Austorof.awCostArines	10000	Average PixRe Dietav	04			1 30	
9	STN	9 Stansfed	A. Cedicatediaminais		/ DetancetociosentSmile	Arport		1	1 14	
10.	CPH :	10 Copenhage	Rubic Transport		Asportingunaree-and	*			1 13	
11	MAN	11 Mancheste	#D #71		/ Delancetocitykm				1 13	
12	VE	12 Vienna	(a) PT2		# hteriofartscorrected				1 9	
13	OSL.	13 Osla	A Orterentiatepromytcheme		/ GDPconected	100			1 15	
14	CUS	14 Düsseldorf	A 0P01		& Cargotan				1 11	
15	MIP	15 Malpensa	a) 0912		- Automation				1 10	
95 -	BRU	16 Brussels	al cente	14	Caller Caller By				10	
17	US	17 Liebon	A. 0455		Charles		ow do	) we nam	1e 22	
18	UHR	18 Heathrow	/ whetheres		Reason Change				18	
19	CDG	19 Charles de	# SDP_TA		Charles Coverage	l th		002	9	
20	FRA	20 Frankfurt			Company of the local division of the local d	U	10 Las	C9 (	9	
21	MAD	21 Madrid			Southers Southers				21	
22	AMS	22 Schipel	04	Bathe	Sevel Cancel	Peer .			0 12	
23	FCO	23 Leonardo D	Internet In	and and a second second	and the second s				0 13	
24	MUC	24 Munich		3.	2681000 3968	06	91	27	1 9	100
35	LGW	25 London Gab	eich .	3	240:000	79.	.90	x	1 12	-

88	2000 Dat 100 An	* Datation goaless ()* ** 2_14 ()* # 4()	tres	sumes the an	alysis step	)S					TÉCNIC
ode	le al	NCI		-					Vielaine 37 o	137 Vehilles	
	Code	Orden	Arpot	Passengers	Movements No.	unberstannes	ercentage	Manarine(.com	geofisiaficpercount	Domest	TING OF
	-		😰 the an obtain the over the adjust				8		14		2/10/
-	COM	T Note Cola o		Contraction of the second seco	enia.		(		1 20	- ñ	Constr
1	LPA	3 Gran Care	dia Code	100	sergers		Danice.		0 20		
4	ALC	4 Alcerte	Orden	all May	eteriti		Page		1 23		
6	LTN	5 London Lut	A Manarinet propertientage	M Historchical Cluster	Analysis: S		Mend		1 22		1 TH
6	W.	S Frederic Cr	An unsurpresentage of atticperor	Contraction of the second second second			104		0 14		A A A A A A A A A A A A A A A A A A A
7	FAO	7.Faio	Doirestic/Healownance	Regimention schedule					1 20		FEU
8	OPO	6 Oporto	Aunkerof, au CodiArities	Provinely matter					1 30		
9	STN	9 Stansted	Codestediaminate	-Chaster Member ship		8			1 14		
10.	CPH :	10 Copenhage	Public fivenaport.				1.4 .1				
11	MAN	11 Mancheste	Ø2 PTI	(c) (gene			If there	is a prio	r idea of t	he I	
12	VE	12 Vienna	40, PT2	Sude approx.							
13	OSL.	13 Osla	2. Orterentiatepricingschene	Constant			number	of clust	ers it coul	d	
14	OUS .	14 Düsseldorf	A 0PS1	C Seide of storgood			la a tradita				
15	MIP	15 Malpensa	0,0912	The second se			pe inaic	ated in		r	
95 .	BRU	16 Brussels	a) cerca	A STREET STREET AND STREET			momho	rohin ha	X		
17	us	17 Liebon	CPSS	Comme II cane	1	1		isuih no	X		
18	LHR	15 Heathrow	/ moters	Contra 1 Carlos	all and the second		1		U 16		
19	CDO	19 Charles de	\$ 500 Ja	- Parada					0 9		
20	FRA	20 Frankfurt			and Dime				0 9	3	
21	MAD	21 Madrid		3.9	where the stress	1			0 23		
22	AMS	22 Schipel	100	OK Baste Broot	Cancel Heal				0 12		
23	FCO	23 Leonardo D			-terminer disables				0 13		
24	MUC	24 Munich		32661000	396805	91	27		1 9		
35	LGW	25. London Gat	eich .	3240000	251829		36		1 12		









6 H 4	E +1	아 노동한 취 세종	1 II II II 🤟	100							
Code		NOT								Weble 37 a	ar 37 Versities
	Code	Orden	Aquit 2)	er danstreite	Passengers	Movements	Numbersfairline	e Mainairinefighte	o MainairtineLowCo st	Maximumpercent geofraficpercours ry	a Domest K n
1	NCE	1 Nice Côte c			(wontes)	0		Date:	1	20	
2	CGN	2 Cologne Bo	All Code	(a)	Charles Sectors into	-	100	Party in the local division of the local div		13	6
3	LPA	3 Gran Cana	Coden		Contra analysis, soli		<u>1991</u>	rido	(	28	£3.
4	ALC :	4 Alicante	A Manarine, pwCod	Cluster Mathical In	ewest reighter		•	Mered	1	23	£
5	LTN	5 London Lut	Animumper certage	Manage	fvider-groups linkapi			59-0	1	. 22	2
6	W.	5 Frederic Cr	Conette Predoninar	-	Rivi-groups lekkige			1.1	1	14	£
7	FAO	7 Faio	R Number of Low CostA	Citarvat Pu	ethent reighter		Tuno	faluatari	na alaar	ithm 🔤	£
0	OPO	6 Oporto	Autorofermies	0	ettroid clustering		гуре с	n ciusten	ny alyon		5
9	STN	9 Stansted	Public frameport	Courts	ndien-clustering				8 0	14	1
10.	CPH :	10 Copenhage	a) #11	Contraction Provide State	and's method	-			1	13	8
11	MAN	11 Marcheste	A 1912	OBeak 1					1	13	£
12	VE	12 Vienna	619 6	1	anne There	Berlin			1	9	£
13	OSL	13 Opio	a dest	r fransfarrs Val	Les .	Transform the	Anna		1	15	6
14	OUS .	14 Dosseldorf	8, CP12	Generative La				1	1	11	í.
15	MIP	15 Malpensa	a) 2453	Saccasor 1	ens	Abarolite vi	9.01			10	<u>855 – 1</u>
95 -	BRU	16 Brussels	CPS4	18	C By garana	Chaude sid	n			10	1
17	us	17 Liebon	0955	1. 18	O Mrgann	Pageciae to	0-1 range	T I	1	22	2
18	LHR	18 Heathrow	\$ 509.14	1		-		-		18	É.
19	CDG	19 Charles de	15.6.6		Continue Cancel	100		1	1	9	6 I
20	FRA	20 Frankfurt			2.34	IRE PAR	_	1	(	9	£
21	MAD	21 Madrid	÷	[]		1			1	21	6
22	AMS	22 Schipel		0	Casts Date	Cancer 2				12	£.
23	FCO	23 Leonardo Di	Vinci		33/23000	319000	1	12 1	1	13	6
34	MUC	24 Musich			32681000	396805		91 25	1		£ 1
35	LGW	25 London Gab	dith .		32401000	251829		0 2	c	12	



- inter	and a second sec	and the second se	Construction of the other states of the state	and the same strength of the						States STOREY	And and and a state of the local division of
	Terrer of				The second se					1	
	Code: 0	Adem	2 Illerachical (Park	e daadaata	-stinder	Movements	noorstanores : Mar	N STATES	si geofia	Mopercent Unit 17	n
1	NCE	1 Nice Côte c	Care and the second sec		Ymosteria			pantes	. 1	20	
2	CGN	2 Cologne Bo	dia Com	20 Bernetike	Chater Andreis: Meth	60°	193	Dete	1	13	
3	LPA	3 Gran Cana	A Managine Soltware						0	26	
4	ALC	4 Alicante	A ManarineLowCost	Clutter Method	New and neighbor			100	1	23	
5	LTN	5 London Lut	a Muleumpercerbage	Measure				50.4	1	22	
6	W	5 Frederic Cr	(Conetty Predunition	(Thereas)	Transmit Furnisheer, Balance		-	1.0	0	14	
7	FAQ	7.#aro	R NumberotLowCost4	O GRAN	Euclidean distance		-		1	20	
0	040	6 Oporto	Cedicatedeminais		Squered Bucktean distance				1	30	
9	STN	9 Stansted	B Public frameport	Courte	Cooke				1	14	-
10.	CPH	10 Copenhage	Ø3 #71	Oter	Casholiner		Choico	ofsimil	larity may	acura	
11	MAN	11 Mancheste	40, PT2	- mar .	Block		CHUICE	01 21111	anty mea	วอนเย	
12	VIE	12 Vienna	a otherstatement	-	Mekowski	L	14			1. To	
13	OSL.	13 Osla	CPS1	Transform V	Outonized				1	15	
14	OUS	14 Düsseldorf	B, 0P12	Standardize.	Non +	Demonstra			1	11	
15	MIP	15 Malpensa	<ul> <li>(a) 0P(3)</li> </ul>	and the second	(Distances)	C Change and			1	10	
95 -	BRU	16 Brussels	A. 0454		Danuar	Condeside			4	10	
17	US	17 Liebon	1 remewes	1		C report to 0.1 rs			0	22	
18	LHR	15 Heathrow	\$ 509,34		Course I course				0	18	
19	CDO	19 Charles de			Carde   Carde	100		1	0	9	
20	FRA	20 Frankfurt			2.54	KI (CHU			0	9	
21	MAD	21 Madrid		C.Or.	Party Rend	Canval 2 man	1		0	23	
22	AMS	22 Schipel		-	A Start Street 1	and a second	-		0	12	
23	FCO	23 Leonardo Di	Vinci		33/23000	319000	122	13	0	13	
24	MUC	24 Munich			32681000	396805	91	27	1	5	1.00
36	LGW	25.1 ondos Gab	with .		32401000	251629	90	36	. 1	12	•







. . . Trainations Available Utilities Assigns Tests. Ormann. Yman-**HM** 27.00 40 10 14 1 (2) **4** -\* NOT Value 17 or 37 Venue 1 Code Code Orders Passengers Movements Numbersfairines: Mainairinefightsp: MainairineLowCo Maximumpercenta Domest Airport geofauficpercount -Contract Marel Contract Association 1 NCE: 1 Nice Cite. (white state 20 Saintes. Passangers dig Code 2 13 CGN 2 Cologne Bo # Movements Pogs. J Coden 3. LPA. 26 3 Gran Cana Mainterline fights per centerer / National International Metrod. 23 4 ALC: 4 Alicarde **FEUP** ManarineLowCost Automat, CChiptoweekty Sare: 5 LTN. 22 5 London Lut Molecumper certrage offer attrager co. Contraction 6 2. Conect/Predoninance M History blant Closter Analysis: 5 14 W. 5 Fraderic Cr NumberofLowCostAirlines 7 FAO 7 Fato **Chatter Hembership** Anteroferminals Saving cluster membership as variables 10 OPO 6 Oporto Cedicatedevninait bone 9 STN 9 Stansted Busicfraneport. Single solution (you must indicate a number of cluster for 10 CPH 10 Copenhage a. 171 Repar of Golden PT2 11 MAN 11 Mancheste Bange of solutions A.PT5 classification) 12 VIE. 12 Vienna Conterentiatepricingschere 13 OSL 13 Oala 3. OPS1 14 Dosseldorf 11 14 DUS 8. CP12 A 1993 10 15 MIP. 15 Malpensa Continue. Cancel. 164 CPS4 16 BRU 16 Brunnels 10 A. 0PS5 Cluster 37. US. 17 Liebon Ö, 22 / readeres C-Cangel Variables Ó 18 LHR 18 Heathrow 18 # 00P\_14 Display 000 Û 19 19 Charles de 9 States Pate 20 FRA 20 Frankfurt 0 9 ū 21 MAD 21 Madrid 23 OK. Diele. Same. Catical 200 Û. 22 AMS 22 Schippi 12 23 FCÓ 23 Leonardo Da Vinci 33/23000 319000 122 ú 13 13 24 MUC 24 Munich 32681000 396806 91 27 5 36 35 LGW 25.1 and/or. Gatwick 3240:000 251879 30 12 • Dalla Water Variable Verve 19755 Shideling Processor is ready

# **Proximity matrix**







	1:Nice Côte	2:Cologne	3:Gran	d Alexand	5:London	6:Frederic	7.5.00	8-0	0.5mand	
Case	U AZUF	Borin	Canana	4:Alicante	LMON	Chopin	7:Faro	a:uporto	9:scansted	0
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:Cirona	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	
MALE 1 10 10 10 10										-

Squared Euclidean
 distance between
 cases

How would this matrix look like if the variables were not standardized?

# **Agglomeration schedule**







		,	Agglomeratio	n Schedule			
	Cluster C	ombined		Stage Cluster	First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage	
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	
5	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

previous slide, could be

in the dendrogram

also graphically seen here









## 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





FEUP



# Analysis of variance



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







**FEUP** 

### Analysis of variance (II)



Fair fit (no strong overlapping between profiles)



# **Calculating the ANOVA**

<b>H</b> 19	B 41	CP 1. 04	Pagota	1	0042							
0.00	Inc.	M.T	Eggorgeve statemen								THERE AT OF A	2 Vendare
	Code	Orders	Compare Means	1	M stears	percendant.	Movements	Numbersfairlines :	Mainainineflightsp	MainairlineLowCo	Maximumpercents D	Internet
	2		General Linear Model		One-Sanale T Test				ercercage		- M	
1	NCE		Generalized Linear Modern		A Independent Samples 7 Test	9630967	119322	64	18	1	20	
2	CGN		Miged Models		A Careti-Samples 7 Test	9742300	132200	29	30	1	13	
3	LPA		Converse		C DIVERSI AND VA	9155665	101657	47	17	0	26	
4	ALC :		Segretaion			9136479	74261	35	29	1	23	
5	LTN		Logineer			9129053	83013		37	-1	22	
6	W.		Neural Netgorius			8300927	115934	36	31	0	14	
7	FAQ		Owenty			5061901	37328	23	27	1	20	
8	OPO		(Jerenske-Reduction			4509250	54107		30	1	30	
9	STN		Scale	٠	r	19967077	195000	18	37	1	14	
0.	CPH		Sprowweek: Tests			Roford	a calcula	ting the /		vo should	13	
1	MAN		Porecarging	٠		DEIDIG					13	
2	ME		Same	٠		save t	ha clueta	ar estima	ited in th	e nrevioi	19	
3	OSL.		Mytple Response			Javel					15 15	
14	CUS		Mooing Value Analysis.			hierar	chical clu	ister ana	alvsis for	6 cluster	11	
15	MOP		Mulgate Impublication								10	
16.	BBU		Coregon Sanates			(classi	ifving all	31 cluste	ers)		10	_
17	us		Quelty Control	•		(0.0.00)	.,				22	
15	CHR		ROC Ourge			67054745	400006	99	30	0	18	
9	CDO		Amos 17	_	]	606/4681	518018	136	24	0	9	
10	PRA.		20 Fragature			90933000	453111	1,28	26	0		
1	MAD		21 Madrid			49437547	435167	82	27	0	21	
	446		22 schipel			40,20000	410672	101	.10	0	12	
	PCO .		23 Leonardo Dia Vinci			33/23000	319000	122	13	0	- 13	_
14 : 	NOC N		24 Million			3,3681000	206605	91	27			-
-	u.um		A Landor, Isaberch			3.061100	-718/a	901			12	



### **Calculating the ANOVA**





# Calculating the R squared (I)

		Sum of	df	Mean Square	F	Sia
Passengers	Between Groups	5,275E15	5	1,055E15	6,335	,001
	Within Groups	4.330E15	26	1.665E14		
	Total	9.605E15	31	.,		
Maria anda	Patrona Orana	0.040544	51	5 007540	0.500	010
wovements	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	Between Groups	3926,621	5	785,324	,845	,530
Airlinespercentage	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	Between Groups	336248,471	5	67249,694	2,467	,059
Airport	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			
numberofvisitorscorrecte	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		1,100210		
	iotai	1,092012	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



**FEUP** 

# Calculating the R squared (II)

ANOVA															
		Cluster 2	Cluster 3	Outler-4	Ouster 5	Charler 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Ouster 14	Cluster 15
Passengers	Between Groups	5,001980+15	5,100800+15	5.209410+15	5,209410+15	5.27486E = 15	7.86807E = 15	7.97311E+15	7.97564E+15	8.0K389E+15	8.27780+15	8.285170+15	8.608420+15	8.859420+15	\$1003870+12
	Within Groups	4,60312E+15	4,470810+15	4,305262+15	4,305282+15	4.32961E = 15	1,7366E +15	1.82156E+15	1.62903E+15	1.52677E+15	1.326866+15	1,319496+15	9/882480+14	7,45249E+14	6,007948+1
	1004	\$804675+15	9.604675+15	1.60467E+15	1,604676+15	9.60467E+15	9.80467E+15	9.80467E+15	9.80467E+15	9.50467E+15	9.804676+15	9.80467E+15	9.804675+15	\$.604675+15	1.604678+18
Movements	Between Groups	2.50844E+11	2,529736+11	2,584962+11	2,58496E+11	2.6133/1E+11	4,77948E+11	4.82138E+11	4.80067E+11	4.89817E+11	8.17113E+11	5.20162E+11	5,30363E+11	5,488198E+11	5,677792+11
	Within Groups	3.8800E+11	3.8596E+11	3.804775+11	3.80477E+11	3,77602E+11	1.80985E+11	1.56805E+11	1.558665+11	1.49119E+11	1,2182E+11	1.18791E+11	1.08575+11	90235479221	71198438298
	Total	6.38833E+11	6.388336+11	6.386838E+11	6,388338 +11	6.38933E+11	6.389336+11	6.389833E+11	6.388336+11	6.388336+11	6.38833E+11	6.388336+11	6.388336+11	6.38833E+11	6.386836+11
Numberofairlines	Between Groups	10017-002	13675.082	14102-208	541022-208	14245,337	24551,602	26297,005	26330.005	27130,005	32006.567	32990.963	34377.838	36437,542	30017,375
	Within Groups	37427,783	36009.793	36522.667	36522;667	36399,538	20090,273	24547,81	24311,81	23671,81	17578,278	17090.862	16206.307	94207,733	11627,5
	Total	50644,875	50044,875	50044,875	50644,875	50644,875	\$0644,875	50644,875	50644,875	50644,875	50644,875	50044,875	50044,875	50644,875	50044,875
NumberofLCCRightsweekly	Between Groups	375371,742	383676,438	388163,422	388183,422	412290,100	598504,378	660194,397	742200,647	747200,647	883680,608	889496,434	984223.552	989791,519	1102914,808
	Within Groups	1148417.977	1136713.31	1133606.296	1130606.296	1109406,615	923195.341	896006.321	779680.071	774580.071	608100.111	632290.284	537966.167	531998.1	478674,881
	Total	1521780.719	1521789.719	1521789.719	1521780.719	1521790.719	1521780.719	1521780.719	1521780.719	1521780.719	1521780.719	1521780.719	1521780.719	1521780.719	1521780.718
LowCostArinespecentage	Between Groups	2610,639	2628.345	3496.873	3495.873	3026.621	0000.002	9670,996	9977,027	9617,468	17794.058	18076.427	18458,283	18036.772	22967.247
	Within Groups	25475.341	25457.635	24630.107	24630.107	24159.309	21146.388	18115.434	18108.963	18108.512	10294,822	10009.953	9627,697	9427,208	5418,733
	Total	28085.98	28085.98	28065.96	28065.98	28085.98	28085.98	28085.98	28085.98	28085.98	28085.98	28085.98	28085.98	28085.98	28065.96
Destinations	Between Groups	29082-695	36363,362	40574.303	40574.303	40642.654	81376,795	91550,44	96660.09	19041,19	121574,722	123153.236	128527.063	131702.6	151753.571
	Within Groups	168340.805	162680.138	158459.167	158459.167	198390.846	117056,705	107483.06	100172.81	99902.31	77458.778	75880.205	70506.438	67300.9	47279-305
	Total	199000.5	199000.5	199000.5	199000.5	1990003.5	199000.5	199000.5	1990003.5	199000.5	199000.5	199000.5	199000.5	199000.5	199000.5
Average Route Distance	Between Groups	13636244.21	16240884.88	16346213.41	16346213.41	17155058.38	18800152.31	19438138.48	19860442.73	19861743.23	20870130.93	21090200.18	21464179.96	21559432.78	22083033.71
	Within Groups	10291624.67	10557984	10481655.46	10481935.48	9671910.5	7928716.968	7396732.385	8067428,143	6060125.643	5057728.944	5737068.696	5073688.917	5268436.1	4744535.187
	Total	25827968.88	25827908.88	25827968.88	25827968.88	20617968.88	20027968.08	20027568.08	25827568.68	25827968.88	25827968.88	25827968.88	25827968.88	25827968.88	26827868.88
DistancetoclosestAirport	Between Groups	737,264	7542.464	54810.777	54810.777	58556.001	61335.099	61726.088	66015.573	66314.376	66718.879	70013.404	21784.176	21918.506	96179.7
	Within Groups	128403-974	121806.774	74390.401	7436.461	70645.176	67906,139	67475.15	63185,205	621046.002	63462.359	59187.834	57417.002	57362-673	33824.538
	Total	125001.238	125001.238	125001.238	129(201.238	128201.238	128201.238	1250201.238	125001.236	128201.238	125001.238	125001.236	125001.238	125001.238	125001.238
Distancelocioses/SimilarAirport	Between Groups	3102.407	\$4857.775	301184.374	301104.374	336248.471	700949.813	7(37%.543	753458 979	750178,195	704219.717	908075.304	924377.034	924811.107	927832.343
	Within Groups	1041858.113	960102.775	743756.177	743796.177	706712.079	344010.737	301044.007	291401.572	286782 365	280640.833	138385.187	120583.516	120149-443	117138,207
	104	1044960.95	1044960.55	1044960.55	1044960.55	1044060.55	1044060.95	1044060.95	1044060.95	1044060.95	1044960 55	1044960 195	1044960 55	1044960 55	1044960.55
ArportRegionalitelevance	Batween Groups	0.006	0.044	0.178	0.178	0,299	0.488	0.704	0.704	0.71	1.008	1.005	1.96	1.181	1.186
	Within Groups	1.564	1.526	1.382	1.302	1.301	1.081	0.896	0.000	0.86	0.981	0.505	0.41	10.300	0.384
	Total	1.57	1.57	1.67	1.57	1.67	1.62	1.67	1.67	1.57	1.67	1.67	1.67	1.57	1.57
Distancelocitykm	Between Groups	126.036	217 703	909.708	909.708	1136.49	TEIN SIM	4585.625	4015.875	487.6%	1017.401	19627-444	16508-271	16005-604	16005.78
	Within Groups	19657 839	19845.172	19153.167	19153.167	1805.365	10000.341	19477.25	15447	15205	4745,444	405.401	2034.604	3607.267	3607 098
	Total	20062-075	20062.875	20062-875	2002.875	20062 875	20062 875	20062 875	20062.875	20062-875	20062.875	20062.875	20062-875	20062-875	2002.875
Inharibitaritacurracted	Between Groups	1.989180+13	2102892+13	2.795428+13	276420+13	2.642626+13	8.54278E = 13	9.80989E = 13	1.018385+14	1.04548+14	1203002+14	1.27136E+14	1.427878+14	1.40180+14	1.558290+54
	Within Groups	1.80472+14	1.80718+14	1.800842+14	1.800842+14	1.796128+14	1.236118+14	1.096796+14	1.0628+14	1.036982+14	8.146215+13	8.001272+13	6.525192+13	6.285828+13	5.210E+13
	Total	2.080398+14	2.080386+14	2.080399+14	2.080389+14	2.060385+14	2.080385+14	2.080388+14	2.080308+14	2.080398+14	2.080398+14	2.080308+14	2.080398+14	2.080398+14	2.080308+14
surface feature consciout	Batanen Grauss	6.07536+13	8.877896+13	8.1073896+13	8.1273896+123	9.390546 + 13	1.400116+14	1.43896+14	1.438035+14	1.528886 +14	1.588/45+14	1.588/45+14	1.585145+14	1.82136+14	1.8468-+14
	Within Canada	14/0896+14	1.127636+14	1.123036+14	112308-114	1.090395+14	5 82215+13	5,700166+13	1, 202005+13	4.805346+13	4.307915+13	4.307916+13	4.30396+13	19016-13	1.71118-11
	Total	2015426+14	2.015426+14	3-015436+14	2.015436+14	2.015426+14	2.015426+14	2.015426+14	2.015426+14	2015426+14	2015426+14	2.015426+14	2015426+14	2015426+14	2015436+14
GOPconected	Between Groups	AVVED1ABD 6	100345183	1564001374	1564001374	104720300	1004107964	1011007304	1672057394	1973007394	20100000210	2109007748	2349707196	2050067238	20030408
	Within Groups	2583443085	2361223086	1000003004	100000004	158/67136	1500740254	1400770844	1452610844	1411520844	1366600028	1254000490	1075363042	1172784000	822128800.5
	Total	343498738	3434490798	3434W827W	MANAGEMENT	MINNETH	MINNETH	MINNETH	MINNER	MANUTA	3434990794	MINNETH	363498078	343498078	MANAGEM
Caracter	Batagen Graune	6.5(71)(2+12	6.718048+12	6.718798+12	6.718792+12	6,716308 + 12	6.801AUE+12	6.806526+12	6.8800982+12	6.894000+12	6.894195+12	6.806136+12	6.807980+12	6.904739+11	6.9086281+12
april 1	Within Caract	5.500398-+11	3.762879+11	3 747495+11	3.74249 + 11	3,738095+11	2 102006+11	2.850026+11	2.050395+11	1.001706+11	1.060126+11	1.060000+11	1.001106+11	1.870700+11	1.80628-+11
	5 mail	7.00206.472	1.00006-4-10	T-DECK-ATT	7.00206-412	2 08226 + 12	7 00006 + 11	7.00006+10	7.00206-110	7.00006.410	7.00006-4-10	7.00006-110	1.00006-412	1.00206-4-12	7.00206-412
	Batanan Groups	5.000100-100	6.06/64E + 10	5. 10000E + 10	5.10058E-110	5.41177.+15	# 10000E + 10	# 37364E+44	# 70475++4	# SARIAE + H	# 57/54E + H	8.578185.445	8-947492.446	S 1941 WAYE	6-501742-440
	Within General	A 9000796 + 10	A TOMAE + 10	4.48847.410	4.68647.410	4.040040.+44	1.010005.+11	1 300175+40	4 2012062 + 40	1474/044	1.45(192.+4	1.4/100.41	1.024896.446	8-129006-4-14	4 NUM- 14
	Total	1.00207-10	4.00000.446	4.00706.4.90	4.000700.+10	1 00007 - 14	1 00000 - 44	1.00000-44	4.00000 + 40	4.000002.440	4.000000.444	4.00000440	4.000000440	4.00792.444	4.00700.410
/		100001730	1,000,000,7,00	1,000,000,7.00	1,000000 * 30		1,00448.7 70	1,00448.7 70	1,00048.7 10	1,000,000 * 70	1,000,000 * 70	1,000,000 * 70	1,044645 * 20	1,000,000,7.00	1,000,000,7120



**FEUP** 

# **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**

- H - A	E 41#	2.14	Pegoris		5 0 0 V							Sal 5
Code	3	ICE .	Opporative Statistics								Vielate: A2 of K	Vention
	Code	Ordern	Tagers Cogpare Means General Livear Model	:		Passengers	Maxements	Numbersfairlines	Mainairlineflightsp ercentage	MainairlineLowCo st	Maximumpercenta D geofisaticpercount ry	n n
1	NCE		Ceneralized Linear Wodells			9630967	119322	64	18	1	20	- 6
2	CGN		Migrel Models			9742300	132200	29	- 30	1	13	
3	LPA		Corvease			9155665	101657	47	17	0	26	8
4	ALC		Segression			9139479	74281	35	29		23	E F
5	LTN		Logineer			9129053	83013		- 39	1	- 22	
6	W.		Neural Netgoria		1	8320927	115934	36	31	0	14	
7	FAQ		Cessity	• 2	Two Des Chater	6061901	37328	23	27		20	
0	OPO		(Jeension-Reduction	21	Means Chater .	4509250	54107		30		30	_
9	STN		Scale	. 1	Benarchicel Ouster_	19957077	155000	18	37	1	14	12
10.	CPH		Sprawanetric Tests		A) Tree	19715451	296172	61	23	1	13	
.11	MAN		Forecarging	1	Correspond.	18724989	172495	84	15	1	13	
12	VIE		Savera .			10114103	243430	64	29		9.	
13	OSL:		Mytale Response	1	Several Heightor	16960892	211000	- 36	38	1	. 15	
14	OUS		Mooing Value Analysis.	3.6		17793493	214024	70	. 31		11	
15	MIP		Multiple Imputition	•		17651635	227718	511	19	1	10	
95 -	BRU		Congen Sangles			16999000	231668	83	29		10	
17	US		Quelty Control			11277969	136286	60	31	0	22	
18	LHR		ROC-Durye			67054745	460026	99	32	Ó	18	
19	CDO		Amos 17			60674681	518018	136	24	0	9	
20	FRA		C Field of			50933000	463111	128	- 26	0	9	
21	MAD	53	Maded P			49437547	405107	82	- 27	0	21	
22	AMS		2 Schipol			46299000	418672	101	.10	0	12	
23	FCO	33	3 Leonardo Da Vinci			33723000	319000	122	13	0	13	
24	MUC		4 Munich			32681000	396805	. 91		1	5	
35	LGW		5. London, Gatwick			32401000	251829	90			12	

























			Iteration Histor	(y <sup>a</sup>												
		Change in Cluster Centers														
Iteration	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.															
- C	nemes a chieved		II also and in also he	The second second second	and second a back of	a second leases										

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			
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

### **Cluster membership**

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
NumberofLCCflig htsweekly	575	466	720	129	624	289
LowCostAirlinesp ercentage	18.5113981	21.0234578	9.4177065	81.5573663	13.2320097	39.0757326
Destinations	242	194	243	69	267	126
Average_Route_D istance	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
AirportRegionalre levance	.79083269	.72968233	.69882575	.61286704	.94346327	.73175601
Distancetocitykm	30	22	24	44	11	17
Inhanbitantscorre cted	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

Phd in Transportation / Transport Demand Modelling







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	

**Distance between cluster centers** 

□ 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

Error

Mean Course

- 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

		mean square	ui	Mean Square	ui	F
	Movements	1.219E+11	5	1.139E+9	26	106.974
	Passengers	1.893E+15	5	5.340E+12	26	354.501
	Numberofairlines	7907.850	5	427.139	26	18.514
es	NumberofLCCflig htsweekly	223132.430	5	15620.291	26	14.285
	LowCostAirlinesp ercentage	4159.735	5	280.281	26	14.841
rs is nin	Destinations	29930.563	5	1899.257	26	15.759
	Average_Route_D istance	3746221.58	5	311413.884	26	12.030
	Distancetoclosest Airport	5177.546	5	3973.596	26	1.303
sis	Distancetoclosest SimilarAirport	52703.433	5	30055.515	26	1.754
	AirportRegionalre levance	.050	5	.051	26	.983
	Distancetocitykm	741.823	5	628.991	26	1.179
n	Inhanbitantscorre cted	2.232E+13	5	3.710E+12	26	6.016
	numberofvisitors	2.894E+13	5	2.187E+12	26	13.233

Cluster

Maan Causes



Sig.

.000. 000.

.000. .000

.000

.000

.000

.293

.158

.447

.346

.001

.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.

ANOVA

Phd in Transportation / Transport Demand Modelling

## Exercise





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

## **Recommended Readings**





FEUP

- 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