COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

ICAI – GITI/GITT

Clustering
Estadística II

Eugenio Sánchez Úbeda

January 2024

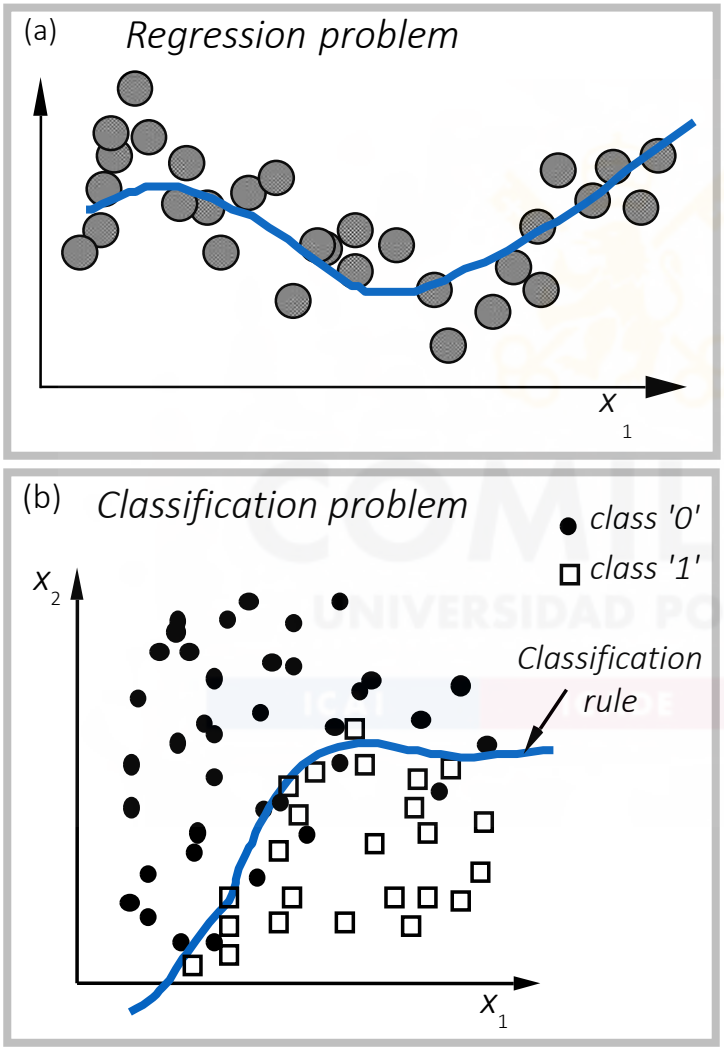
comillas.edu

1

Introduction

Clustering Introduction

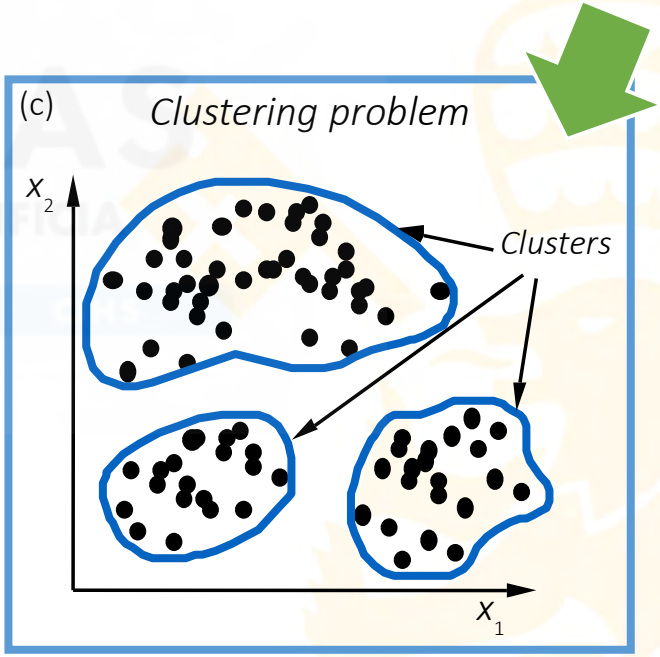
Supervised learning



Both clustering and PCA seek to simplify the data via a small number of summaries, but their mechanisms are different:

- PCA looks to find a low-dimensional representation of the observations that explain a good fraction of the variance
- Clustering looks to find homogeneous subgroups among the observations

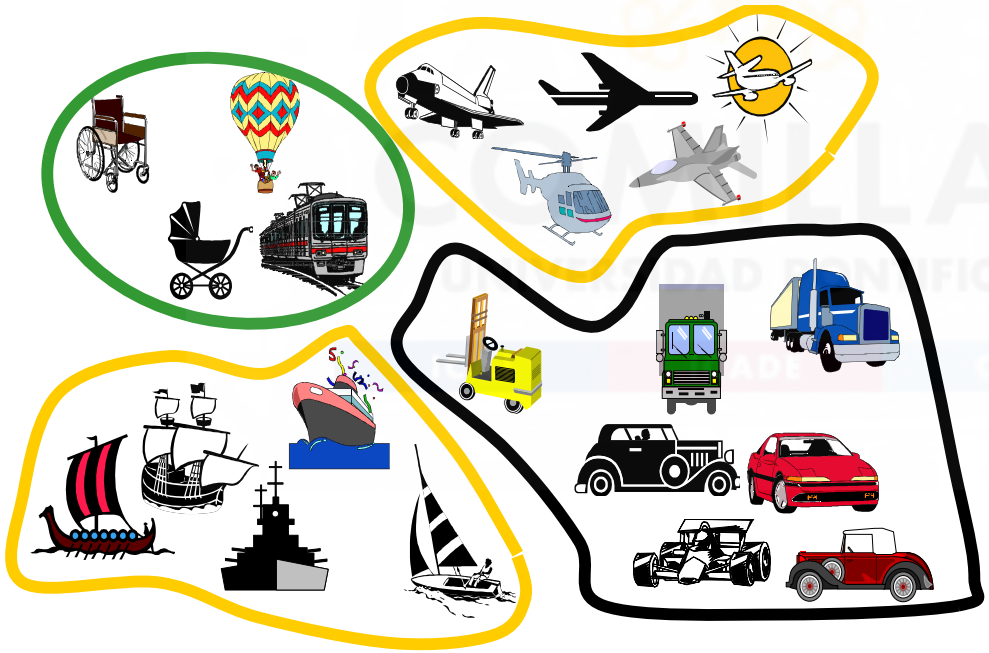
Unsupervised learning



Clustering
Introduction

- Clustering refers to a comprehensive set of techniques for finding subgroups, or clusters, in a data set

Each observation is a vehicle, described by a set of input variables (features) such as the weight, mean speed, and number of wheels.



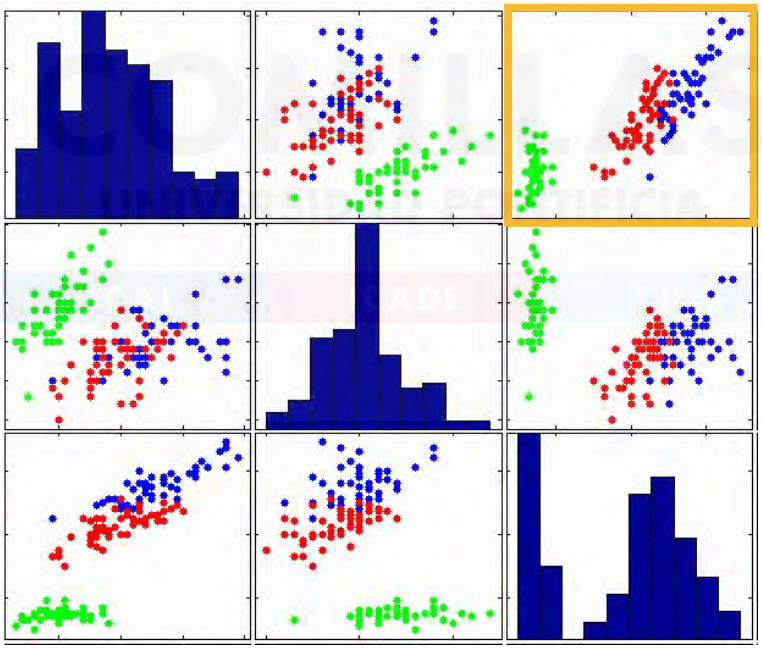
When we cluster the observations of a data set, we seek to partition them into different groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.

What does *similar* or *different* mean?

- Unlabeled observations (no output variable)

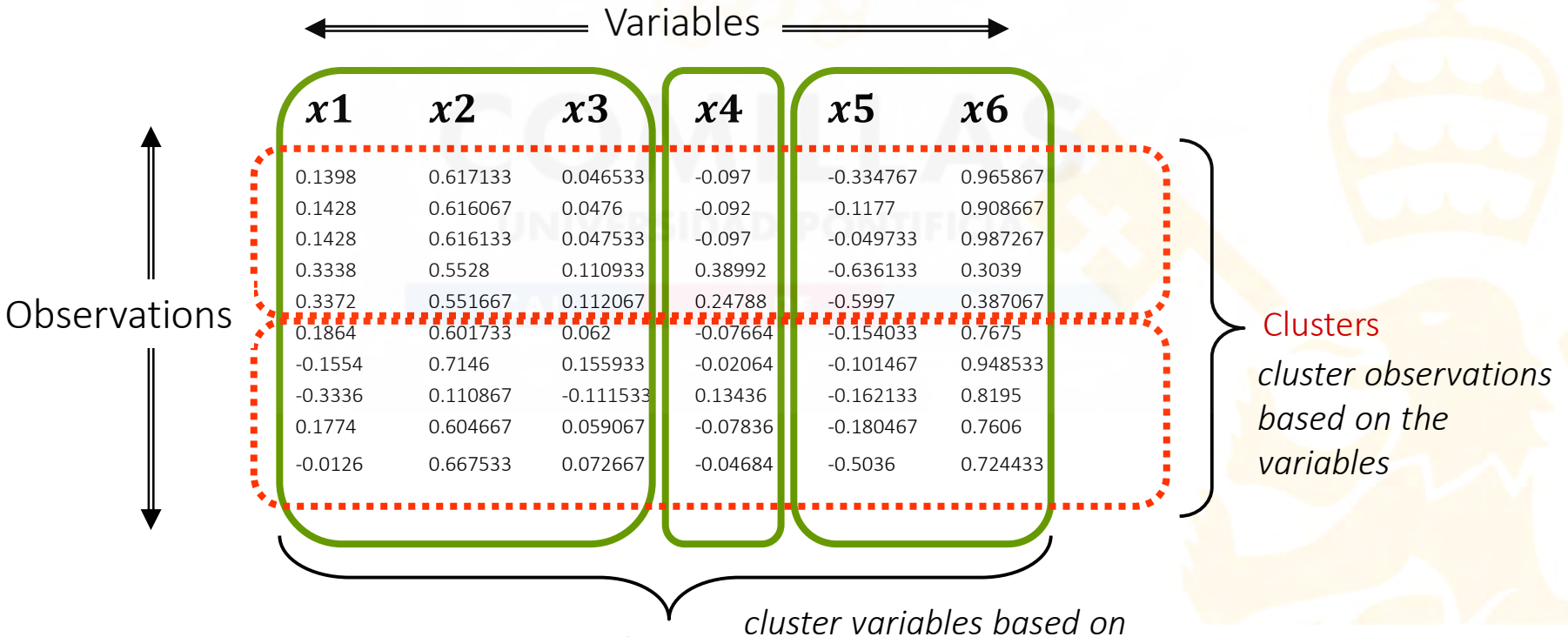
Clustering
Introduction

- Why clustering?
  - To gain some insight into the structure of the data (interpretation)
  - Discovery of patterns
  - Grouping highly correlated attributes



Clustering
Types of clustering problems

- Clustering of observations (data reduction):
  - Identify homogeneous groups of similar observations
- Clustering of variables (dimensionality reduction):
  - Identify similarities among (input) variables



Clustering
Introduction: A real case

- Find daily patterns of demand, i.e., typical daily consumption profiles

The consumption along the weeks and days has different patterns in winter and summer.

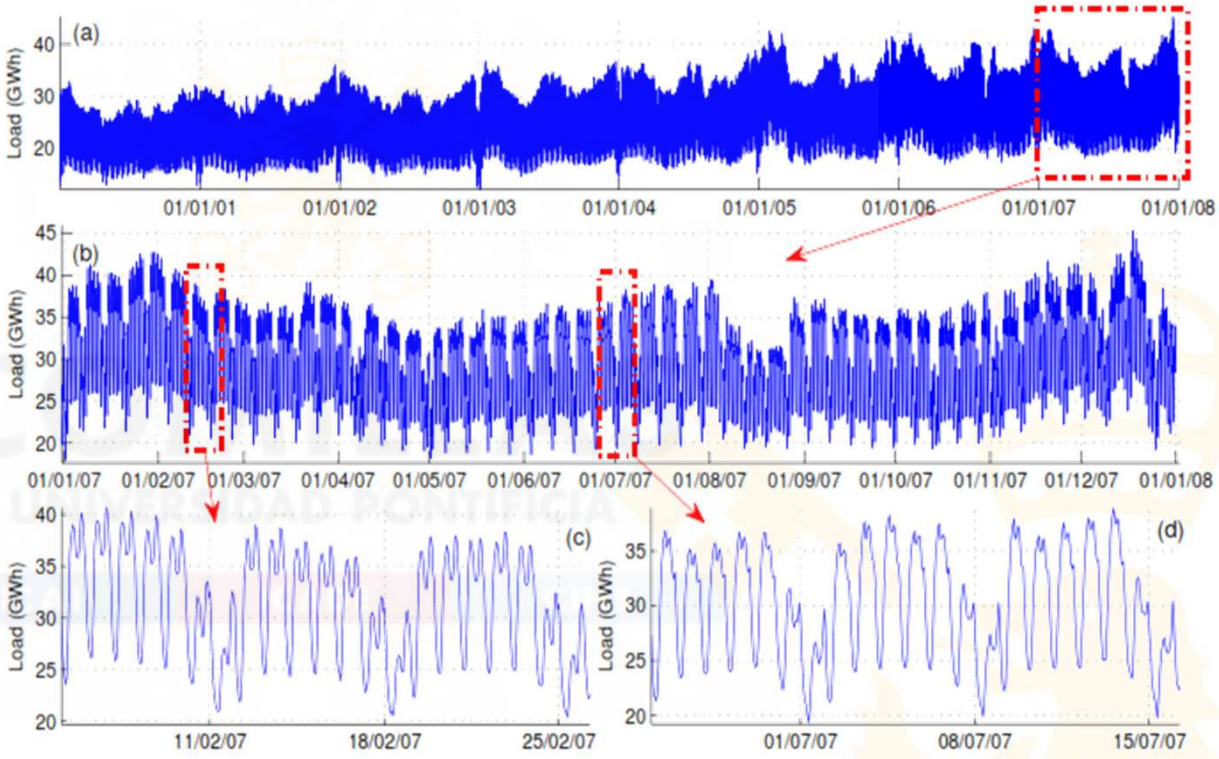


Fig. 1 Hourly electricity demand in Spain:(a)From January 1, 2000 to December 31, 2007; (b) From January 1, 2007 to December 31, 2007; (c) Three winter weeks (2007); (d) Three summer weeks (2007).

Source: A. Muñoz, E.F. Sánchez-Úbeda, A. Cruz, J. Marín, "Short-term Forecasting in Power Systems: A guided Tour", Handbook of Power Systems II. Eds. Pardalos, P.M.; Rebennack, S.; Pereira, M.V.F and Iliadis, N.A.. Ed. Springer. Berlin, Germany, 2010.

Clustering of observations

Which are the variables?

Clustering
Introduction: A real case

- Using a clustering technique (there are a lot), these are the cluster prototypes, i.e., the typical consumption profiles

There are 8 clusters, each with a pattern representing the cluster's center. In this example, each prototype is defined by 24 variables (the demand of each hour)

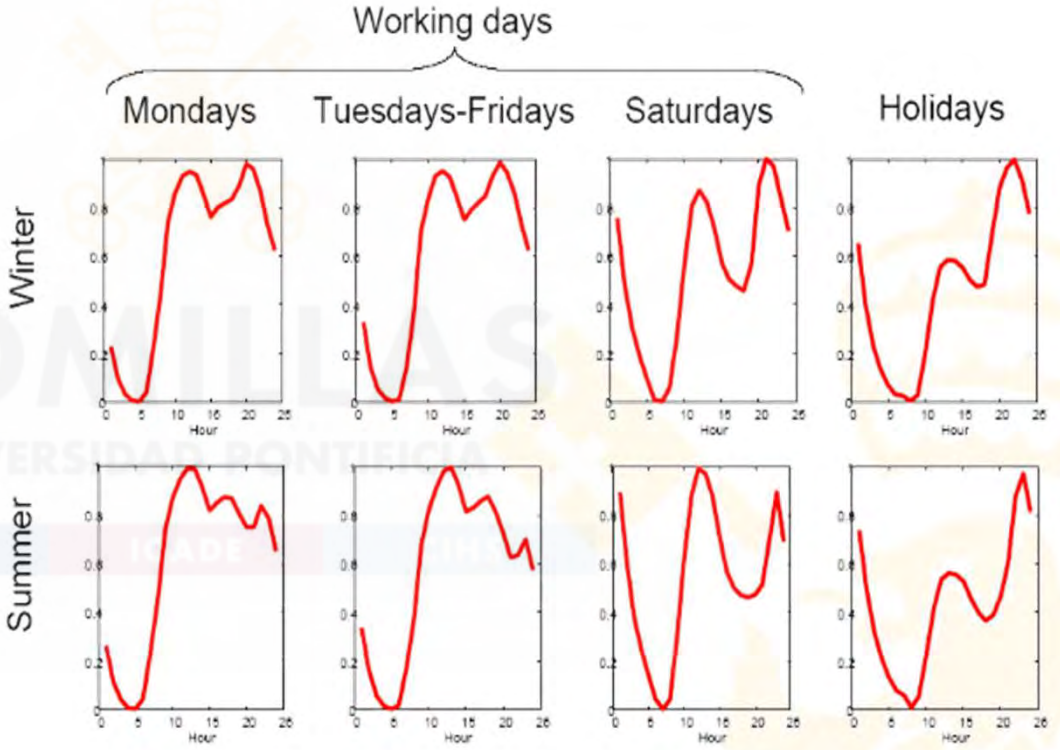
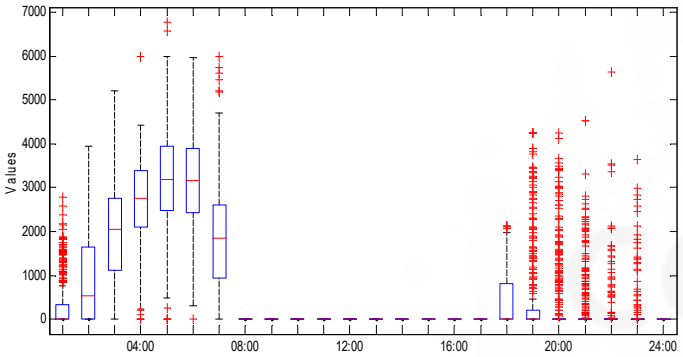


Fig. 2 Normalized intra-day load profiles for the Spanish electricity load.

Source: A. Muñoz, E.F. Sánchez-Úbeda, A. Cruz, J. Marín, "Short-term Forecasting in Power Systems: A guided Tour", Handbook of Power Systems II. Eds. Pardalos, P.M.; Rebennack, S.; Pereira, M.V.F and Iliadis, N.A.. Ed. Springer. Berlin, Germany, 2010.

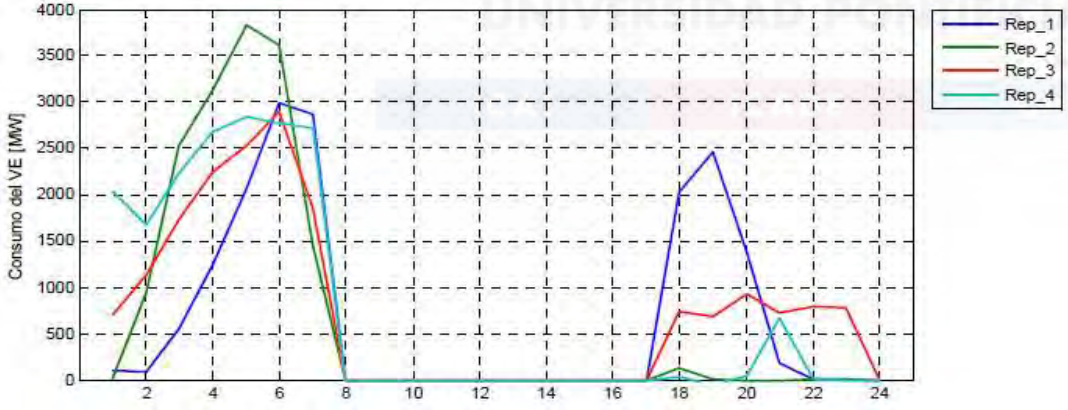
Clustering
Introduction: Another real case

- Optimal charge of 1.7 MEV depending on the type of day
  - Result of a simulation and optimization model

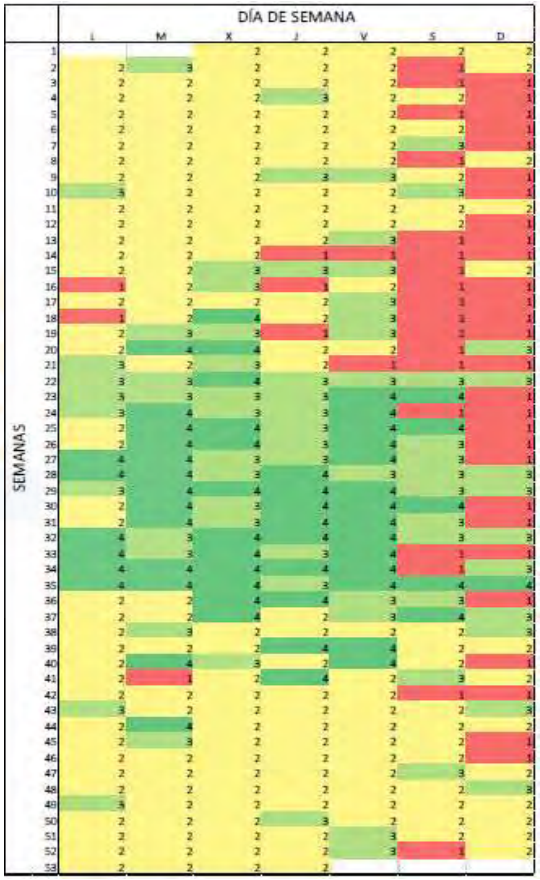


One year of hourly profiles of optimal charge of EVs

Four patterns of optimal EV charge (obtained using clustering techniques)

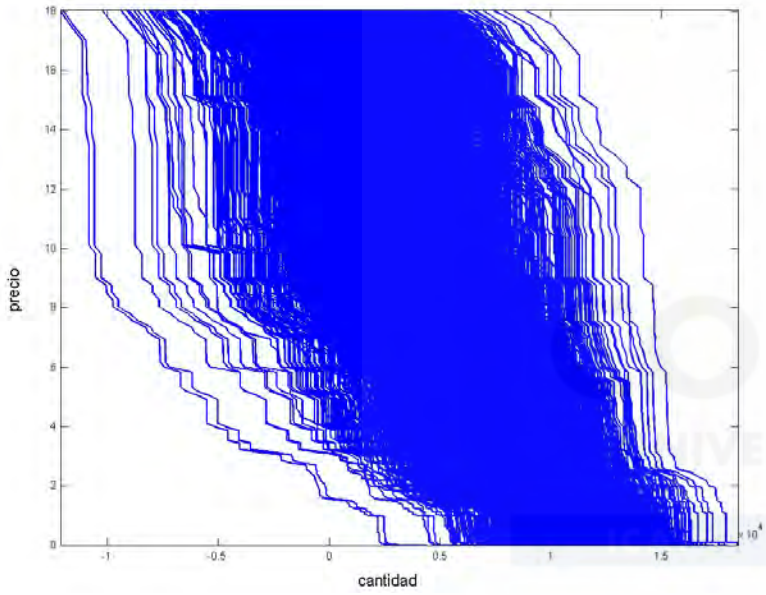


Source: CENIT-VERDE. Charging strategies for using and integration of renewable sources.



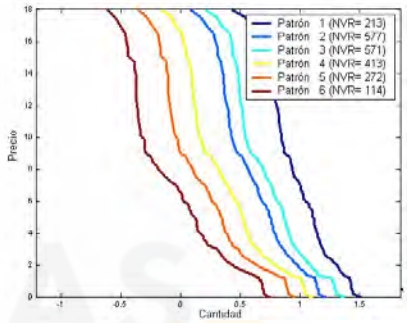
Clustering
Introduction: Another real case

- Bidding curves (Spanish day-ahead electricity market)
  - One curve for each hour

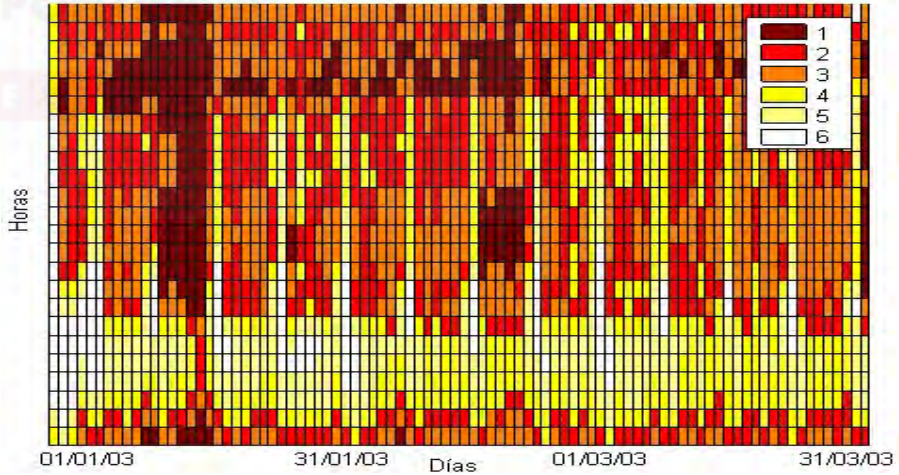


Aggregated bidding curves (1/1/2003 - 31/3/2003)

Patterns (obtained using clustering techniques)



Activation of the bidding-curves patterns



E.F. Sánchez-Úbeda, A. Muñoz, J. Villar, "Minería y visualización de datos del mercado eléctrico español", *Inteligencia Artificial - Revista Iberoamericana de Inteligencia Artificial* . vol. 10, no. 29, pp. 79-88, May 2006.

Clustering
Introduction

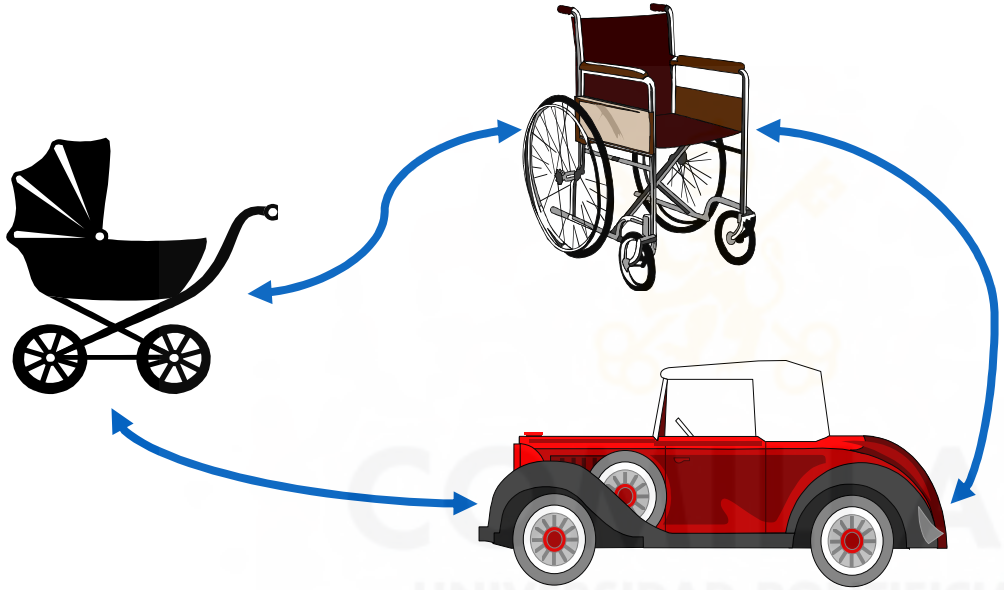
- There exist a significant number of clustering methods
- The two best-known clustering approaches:
  - K-means clustering
  - Hierarchical clustering
- There exist other well-known and broadly used methods
  - Kohonen self-organizing maps (KSOM)
  - Neural Gas

The critical aspect of all clustering methods is the way the similarity between observations (or variables) is measured.

2

Similarity distances

Similarity distances
Types



What is more similar?

- There are many different distances (a lot!)
- Classified according to the type of data vectors:
  - Distances for clustering of observations
  - Distances for clustering of variables

Similarity distances
Distances for clustering of observations

- Data set (p input variables, N observations)

$$\{(x_{1e}, \cdots, x_{pe})\}_{e=1,N}$$

- Distance between two points $\boldsymbol{u}$ and $\boldsymbol{v}$

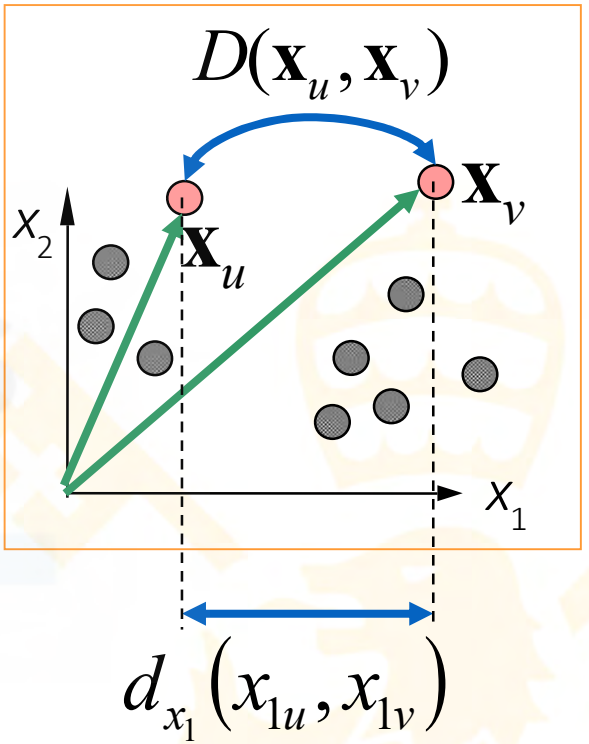
$$D(\mathbf{x}_u, \mathbf{x}_v) = \sqrt[L]{\sum_{j=1,p} \left| d_{x_j}\left(x_{ju}, x_{jv}\right) \right|^L}$$

  - L=2 Euclidean distance
  - L=1 Manhattan distance

- The distance measured along a numerical variable

$$d_{x_j}\left(x_{ju}, x_{jv}\right) = w_{x_j}\left(x_{ju} - x_{jv}\right)$$

- For categorical variables, a pairwise difference table is used

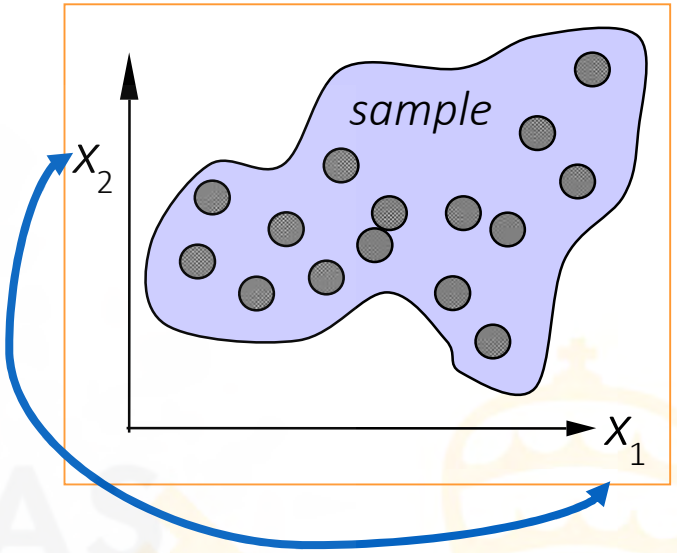


Distance between two variable values

Similarity distances
Distances for clustering of variables

- Distance between two variables

$$D(x_u, x_v)_{sample}$$



- For real-valued variables:
  - Linear correlation coefficient

$$D(x_u, x_v)_{sample} = \frac{\text{cov}(x_u, x_v)}{\sqrt{\text{var}(x_u)\,\text{var}(x_v)}}$$

Similarity distances
List of distance metrics

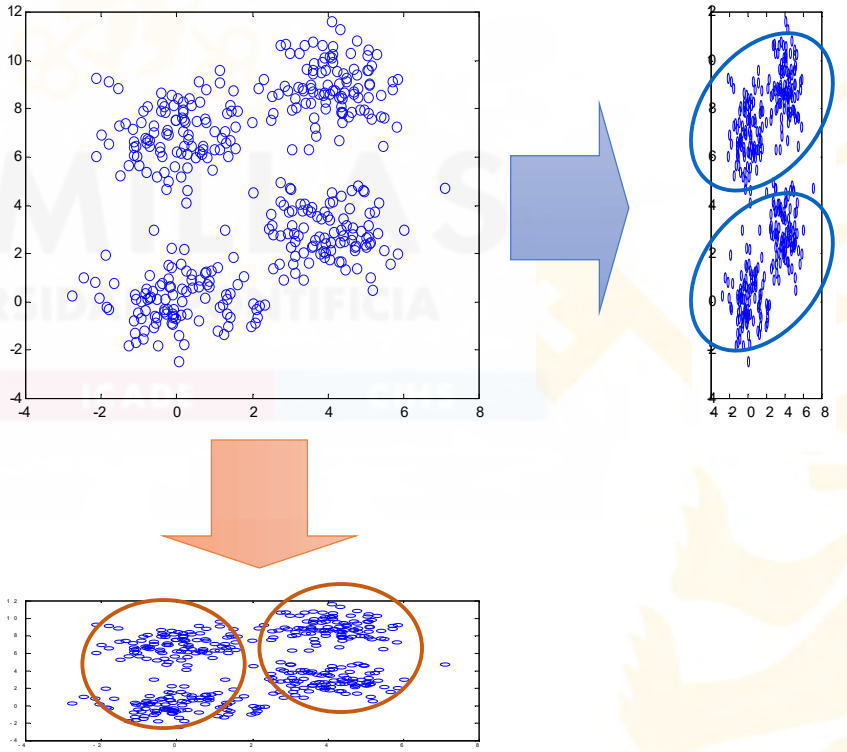
- Matlab

Metric	Description
'euclidean'	Euclidean distance (default).
'squaredeuclidean'	Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality.)
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between rows in X is scaled by dividing by the corresponding element of the standard deviation S=nanstd(X). To specify another value for S, use D = pdist(X,'seuclidean',S).
'cityblock'	City block metric.
'minkowski'	Minkowski distance. The default exponent is 2. To specify a different exponent, use D = pdist(X,'minkowski',P), where P is a scalar positive value of the exponent.
'chebychev'	Chebychev distance (maximum coordinate difference).
'mahalanobis'	Mahalanobis distance, using the sample covariance of X as computed by nancov. To compute the distance with a different covariance, use D = pdist(X,'mahalanobis',C), where the matrix C is symmetric and positive definite.
'cosine'	One minus the cosine of the included angle between points (treated as vectors).
'correlation'	One minus the sample correlation between points (treated as sequences of values).
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).
'hamming'	Hamming distance, which is the percentage of coordinates that differ.
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.

Similarity distances
Main difficulty

- Similarity between two vectors involves combining its components
- Usually, the components are non-comparable
- Simple rescaling of the input variables can result in different clusters!
- Idea

By default, people use the Euclidean distance and are very sensitive to the scale of the inputs when selecting clusters in a scatterplot by eye.



Similarity distances
Main difficulty

- Illustrative example using K-means clustering

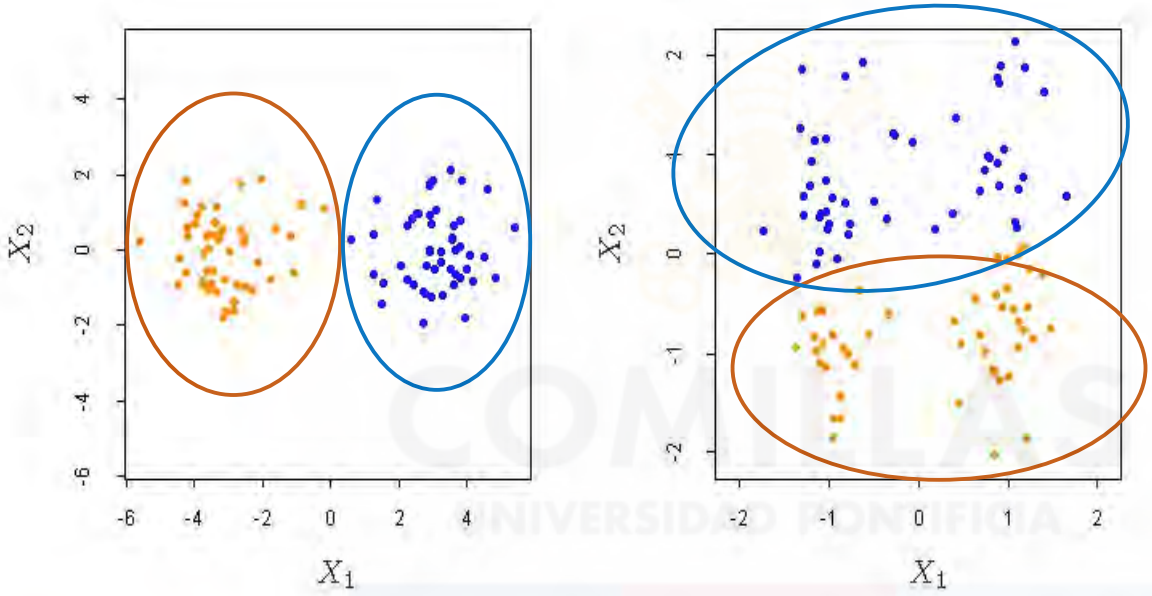


FIGURE 14.5. *Simulated data: on the left, K-means clustering (with K=2) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights* $1/[2 \cdot \mathrm{var}(X_j)]$. *The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.*

Whether or not it is an excellent decision to scale the variables before computing the dissimilarity measure depends on the application at hand

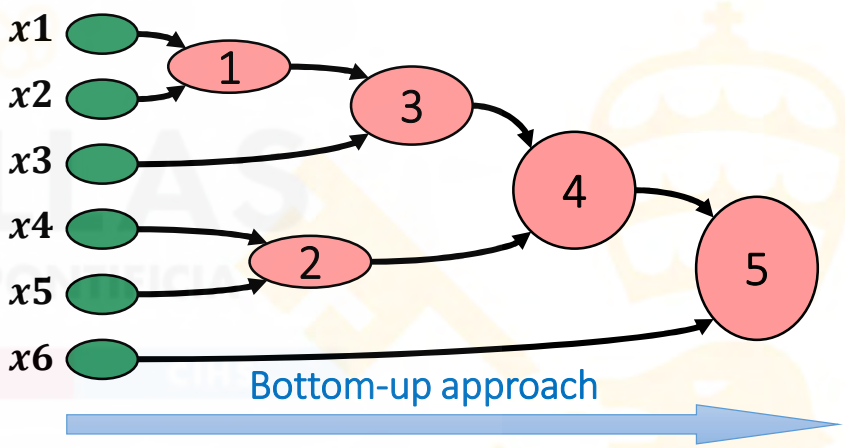
3

Hierarchical clustering

Hierarchical clustering
Overview

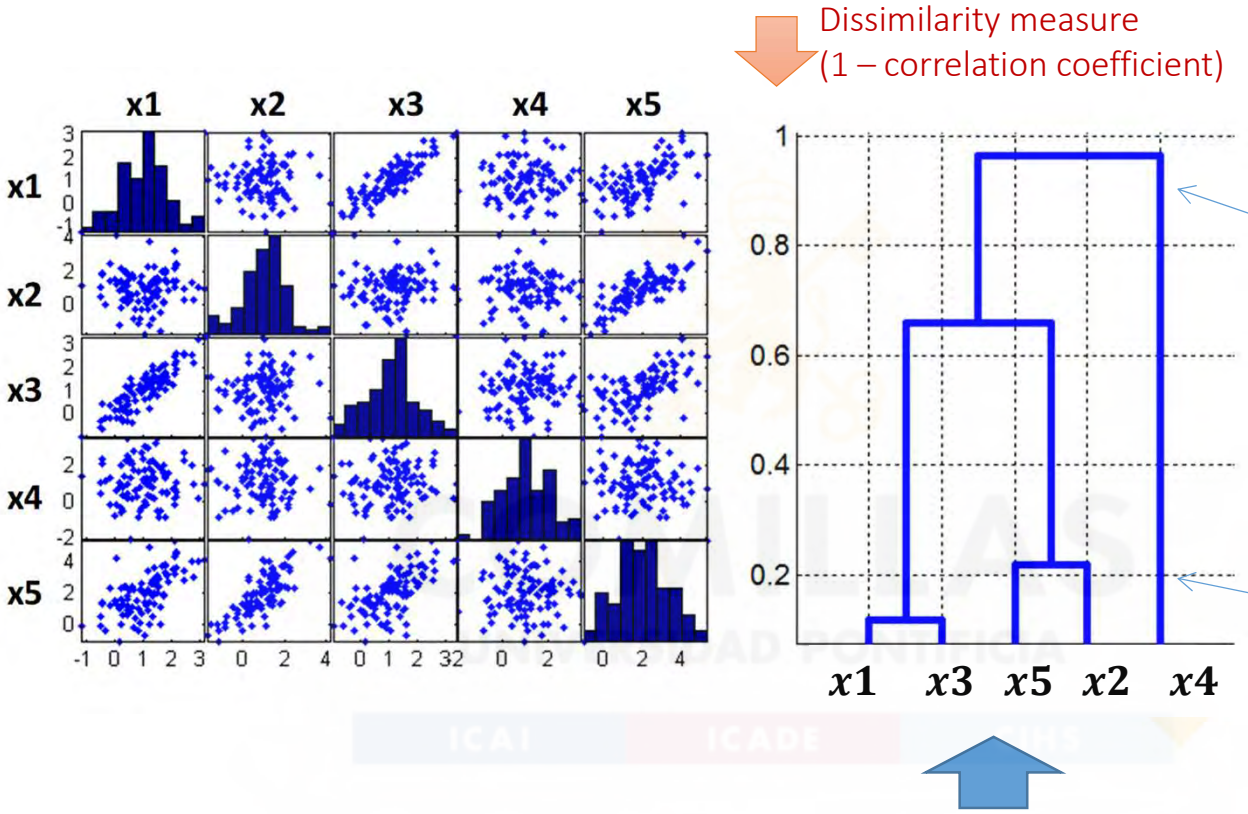
- Usually for clustering of real-valued variables
  - Similarity measure: the linear correlation coefficient
- Agglomerative approach (idea): Identify the two most similar clusters and fuse them

Starting at the bottom, each of the n variables is treated as its cluster. The two clusters most similar are fused, so there are now $n-1$ clusters. Next, the two clusters most similar are fused again, so there are now $n-2$ clusters. The algorithm proceeds in this fashion until all the variables belong to one single cluster



- Dendrogram: An attractive tree-based representation of the clustering process

Hierarchical clustering
Interpreting dendrograms



Variables that fuse close to the top of the tree will tend to be quite different (e.g., $x4$ and rest)

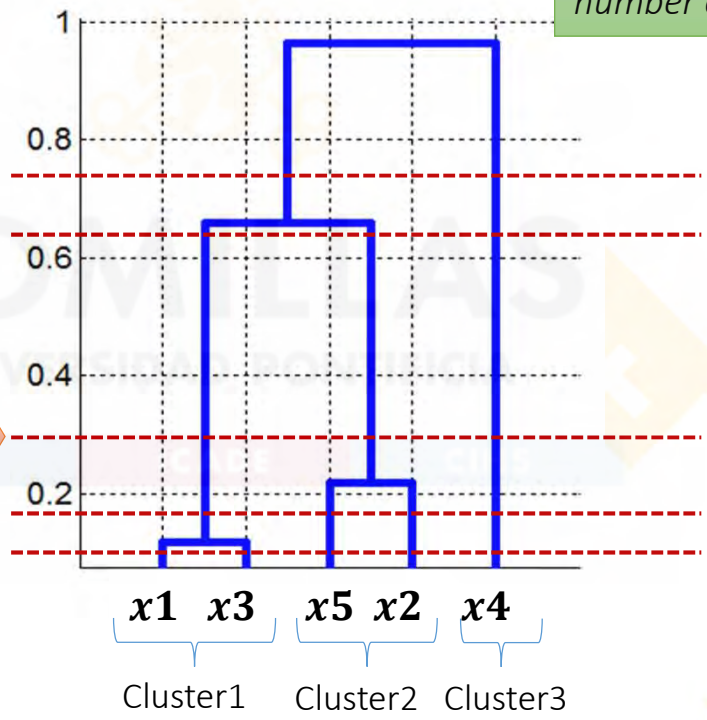
Variables that fuse at the very bottom of the tree are pretty similar to each other (e.g., x1 and $x3$)

The positions of the two fused branches could be swapped without affecting the meaning of the dendrogram. Therefore, we cannot conclude the similarity of the two variables based on their proximity along the horizontal axis. Instead, we conclude the similarity of two variables based on the location on the vertical axis where branches containing those two features first are fused.

Hierarchical clustering
Identifying clusters by a dendrogram

- Select by eye a sensible number of clusters based on the heights of the fusion and the number of clusters desired

The height of the cut to the dendrogram *controls the number of* clusters obtained



2 clusters

3 clusters

Same clusters

3 clusters

4 clusters

5 clusters

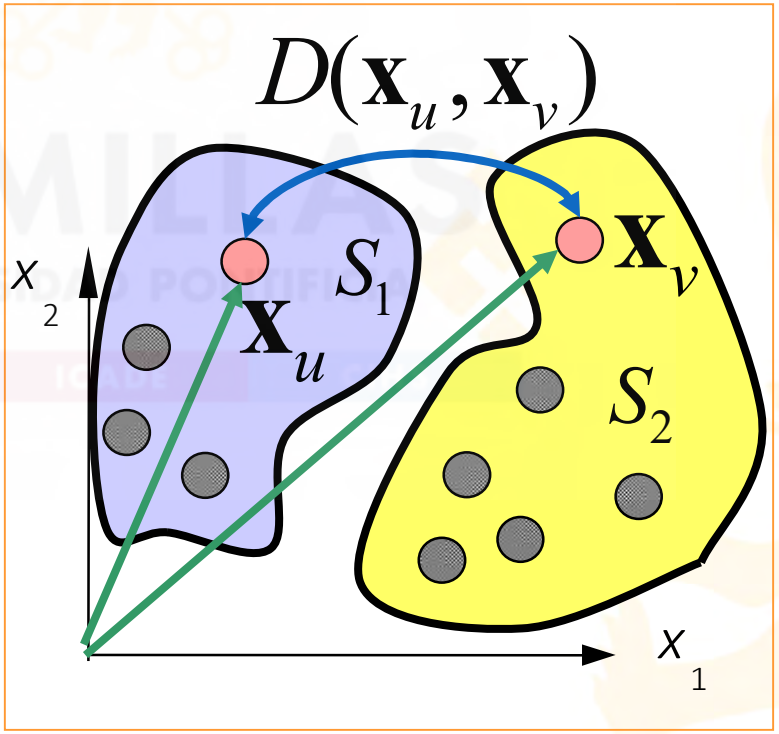
We make a horizontal cut across the dendrogram

The different sets of variables below the cut can be interpreted as clusters

Hierarchical clustering
Distance between two clusters: linkage

- The concept of dissimilarity between a pair of variables needs to be extended to a pair of groups of variables
- This extension is achieved by developing the notion of linkage
- The three most common types of linkage are *complete*, *average*, and *single*

Linkage defines the dissimilarity between two groups of variables



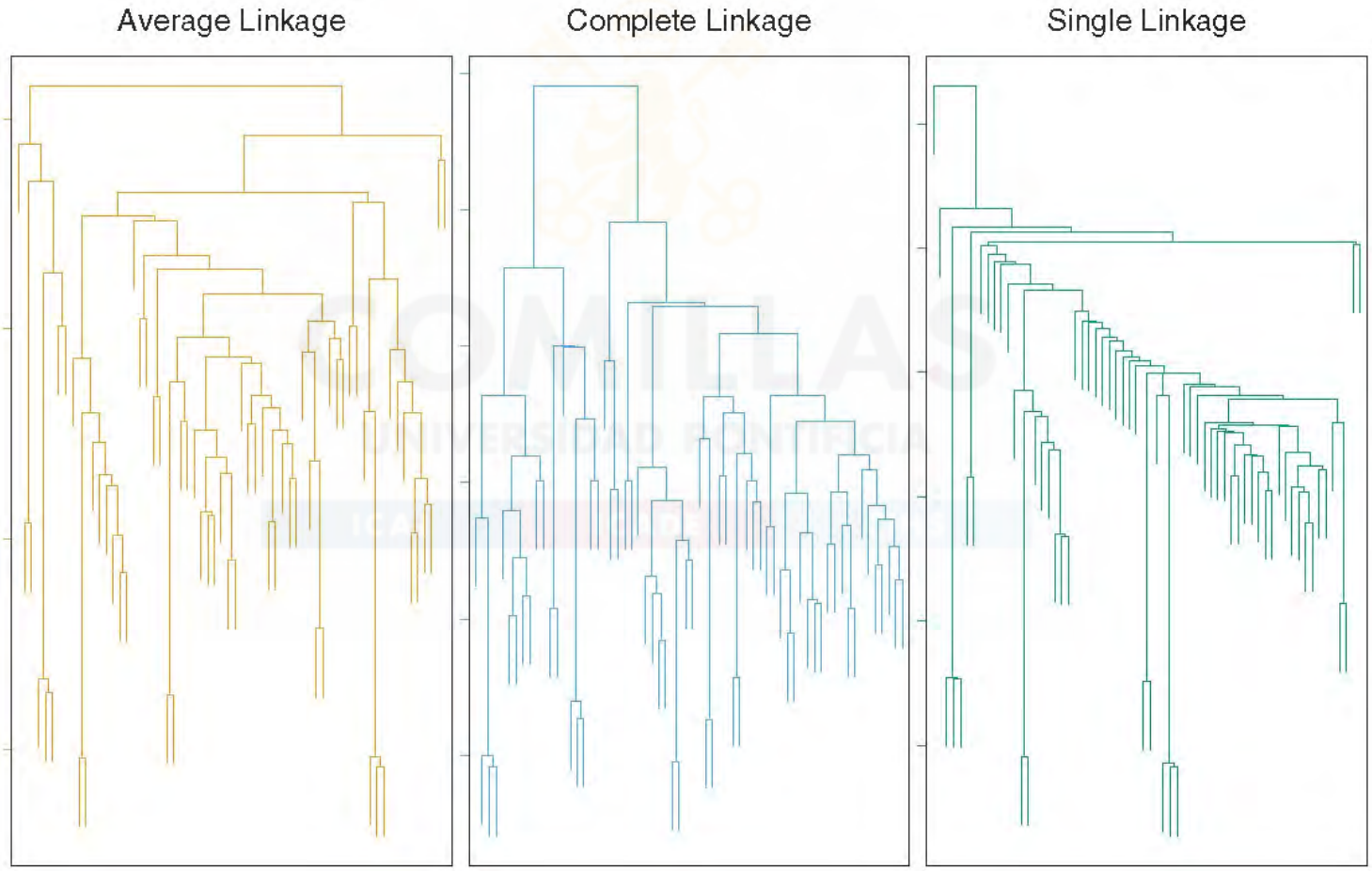
Hierarchical clustering
Distance between two clusters: linkage

- Main types

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.

Hierarchical clustering
Distance between two clusters: linkage

- Impact of the selected linkage method
- Average and Complete tend to yield more balanced clusters



Hierarchical clustering
Average linkage with correlation metric

- Three variables with 5 observations

```
X =
    0.5377   -1.3077   -1.3499
    1.8339   -0.4336    3.0349
   -2.2588    0.3426    0.7254
    0.8622    3.5784   -0.0631
    0.3188    2.7694    0.7147
```

Basic pairwise dissimilarities

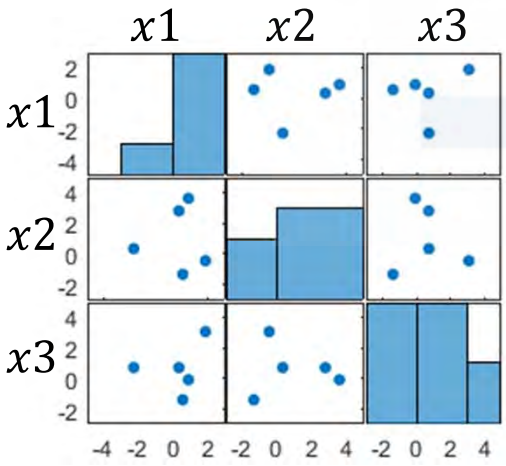
```
1-corr(X)=
    0.0000    0.9675    0.7343
    0.9675    0.0000    1.0433
    0.7343    1.0433    0.0000
```

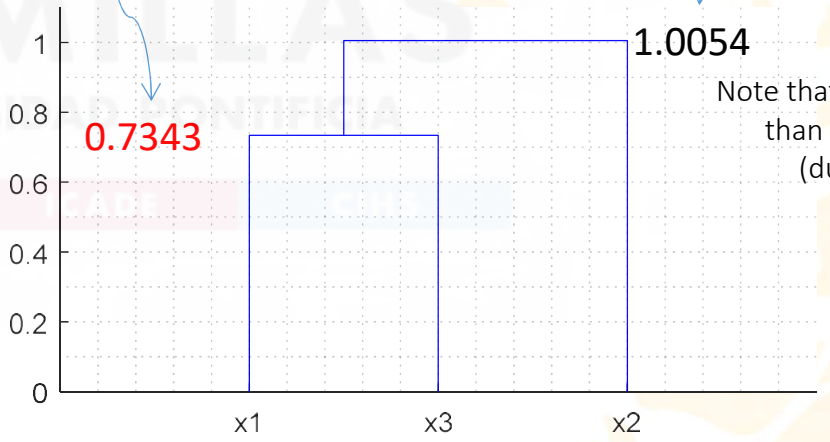
The first cluster consists of $x1$ and $x3$ because their dissimilarity 1-corr($x1$,$x3$) 0.7343 is minimum

The first cluster is fused with $x2$, with dissimilarity given by the average

mean([0.9675 1.0433])

```
corr(X) =
    1.0000    0.0325    0.2657
    0.0325    1.0000   -0.0433
    0.2657   -0.0433    1.0000
```



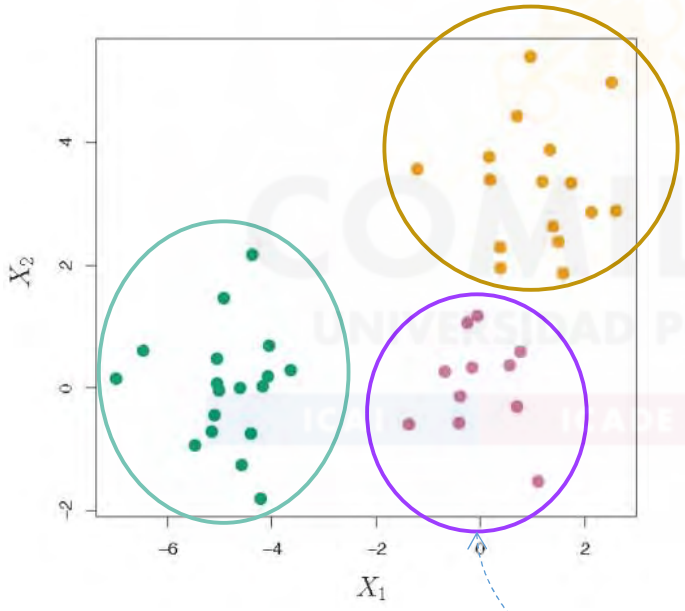


Note that values larger than 1 are possible (due to negative correlations)

```
tree = linkage(X', 'average', 'correlation');
dendrogram(tree, 'labels', {'x1' 'x2' 'x3'});
ylim([0 1.1]); grid minor;
```

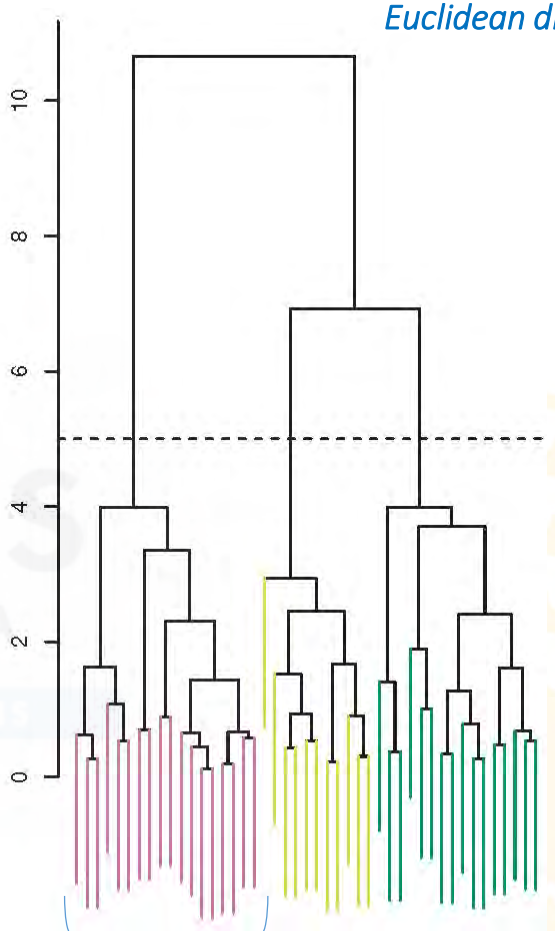
Hierarchical clustering
Clustering of observations

- It can be used for clustering observations
  - It can be performed by simply transposing the data matrix
- Example (two variables)



In this case, we seek to cluster the observations to discover groups of them

Dendrogram obtained from hierarchical clustering with complete linkage and Euclidean distance



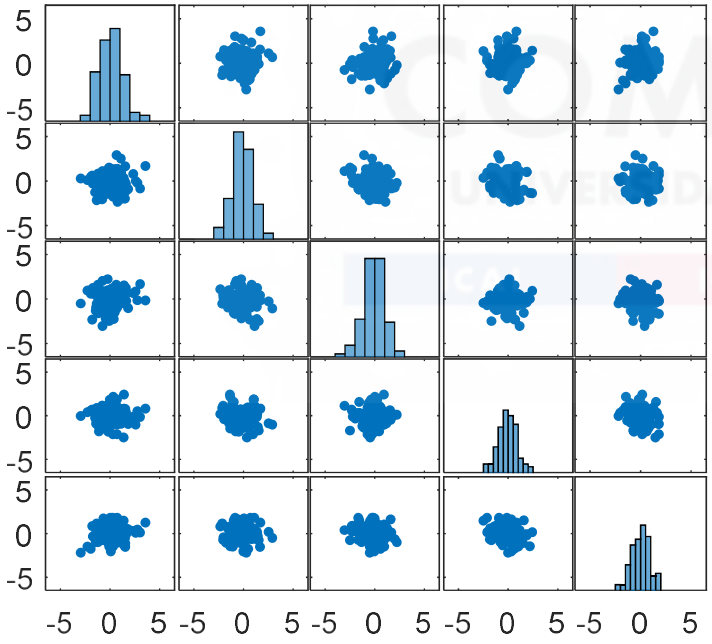
Hierarchical clustering
Illustrative synthetic cases

- C1: Uncorrelated input variables

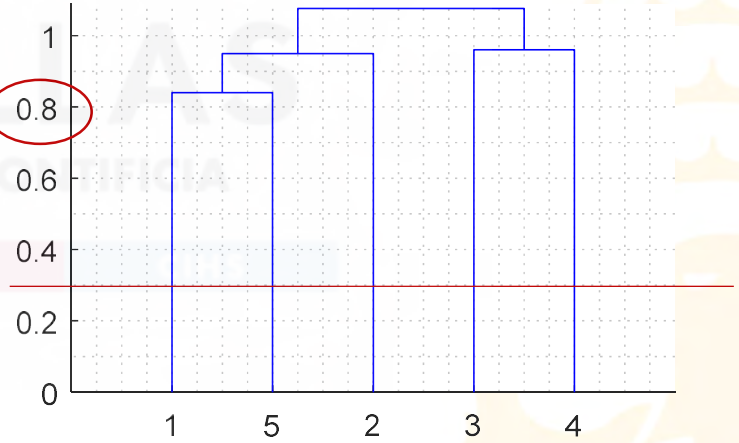
```
n = 100; p = 5;
X = normrnd(0,1,n,p);
```

```
corr(X)=
    1.0000    0.0754    0.1317    0.0740    0.1596
    0.0754    1.0000   -0.1829   -0.1360    0.0248
    0.1317   -0.1829    1.0000    0.0400   -0.1177
    0.0740   -0.1360    0.0400    1.0000   -0.2270
    0.1596    0.0248   -0.1177   -0.2270    1.0000
```

Rows of X correspond to observations and columns correspond to variables




```
tree = linkage(X', 'average', 'correlation');
dendrogram(tree);
```



According to the dendrogram, the five variables are uncorrelated (there are 5 clusters, one for each variable)

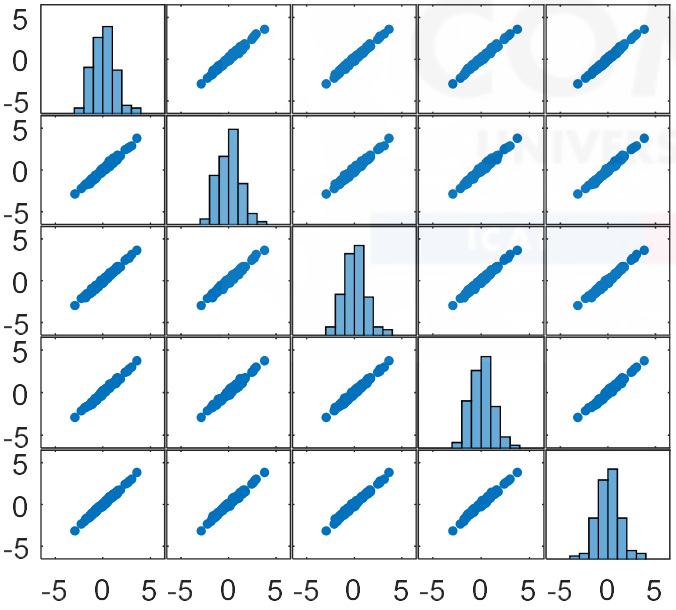
Hierarchical clustering
Illustrative synthetic cases

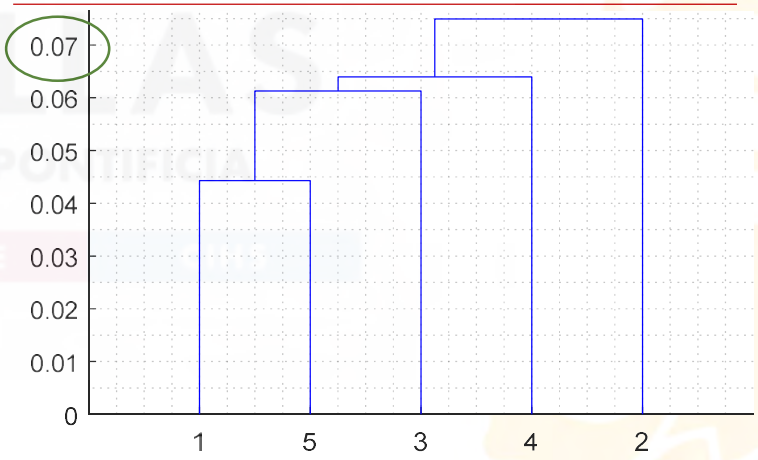
- C2: Very correlated input variables

```
n = 100; p = 5;
mu = zeros(p,1);
Sigma = eye(p);
Sigma(Sigma==0) = 0.9;
X=mvnrnd(mu,Sigma,n);
```

```
corr(X)=
    1.0000    0.9269    0.9407    0.9445    0.9557
    0.9269    1.0000    0.9137    0.9186    0.9411
    0.9407    0.9137    1.0000    0.9350    0.9367
    0.9445    0.9186    0.9350    1.0000    0.9286
    0.9557    0.9411    0.9367    0.9286    1.0000
```

```
tree = linkage(X', 'average', 'correlation');
dendrogram(tree);
```

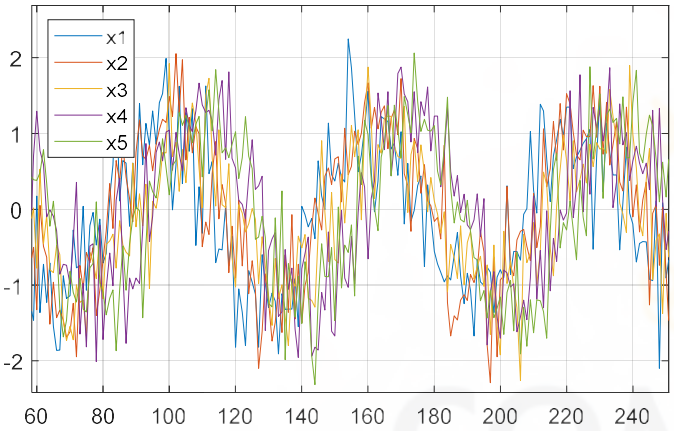


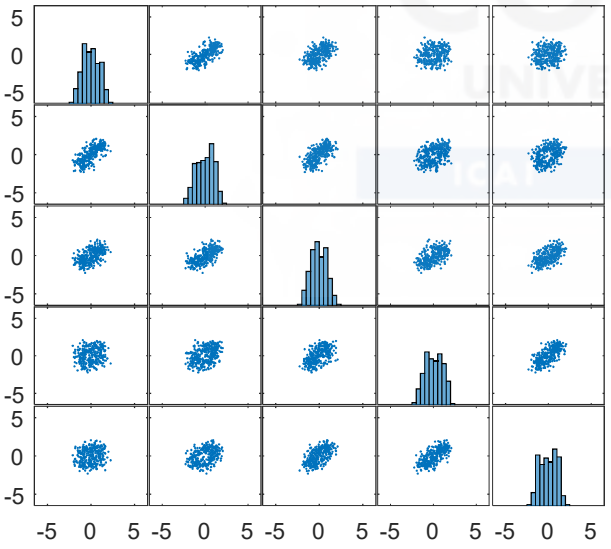


According to the dendrogram, the five variables are in the same cluster, i.e., they are highly correlated)

Hierarchical clustering
Illustrative synthetic cases

- C3: Mixture of correlated input variables

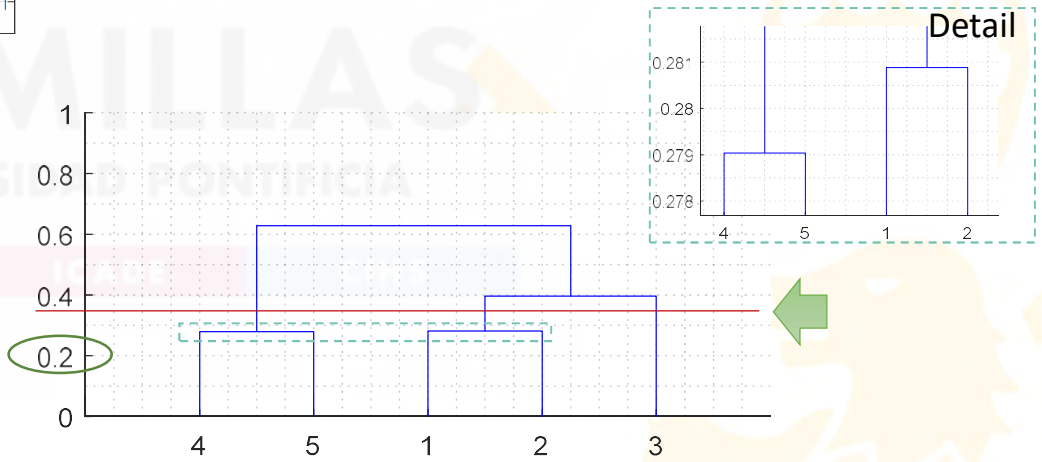




corr(X)=

1.0000	0.7191	0.5406	0.1637	0.1757
0.7191	1.0000	0.6677	0.3698	0.3783
0.5406	0.6677	1.0000	0.5493	0.5913
0.1637	0.3698	0.5493	1.0000	0.7210
0.1757	0.3783	0.5913	0.7210	1.0000

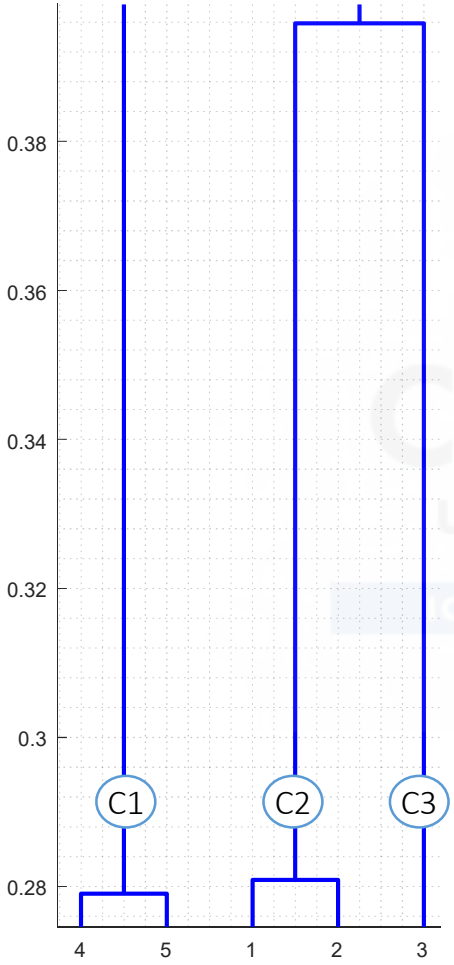
```
tree = linkage(X', 'average', 'correlation');
dendrogram(tree);
```



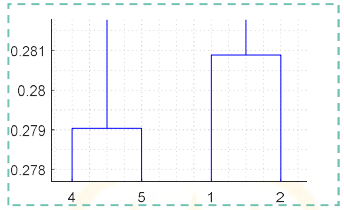
According to the dendrogram, $x4$ and $x5$ form a cluster, $x1$ and $x2$ are in a different cluster, and $x3$ has its own cluster

Hierarchical clustering
Illustrative synthetic cases

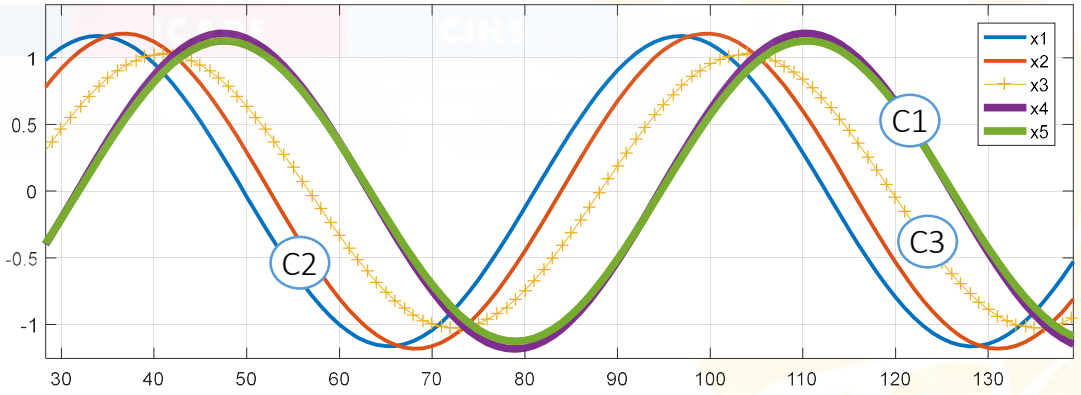
- C3: Mixture of correlated input variables



- Possible clusters according to the dissimilarity measure:
  - 5 clusters: $(x4)$, $(x5)$, $(x1)$, $(x2)$ and $(x3)$
  - 4 clusters: $(x4, x5)$, $(x1)$, $(x2)$ and $(x3)$
  - 3 clusters: $(x4, x5)$, $(x1, x2)$ and $(x3)$
  - 2 clusters: $(x4, x5)$ and $(x1, x2, x3)$
  - 1 cluster: $(x4, x5, x1, x2, x3)$

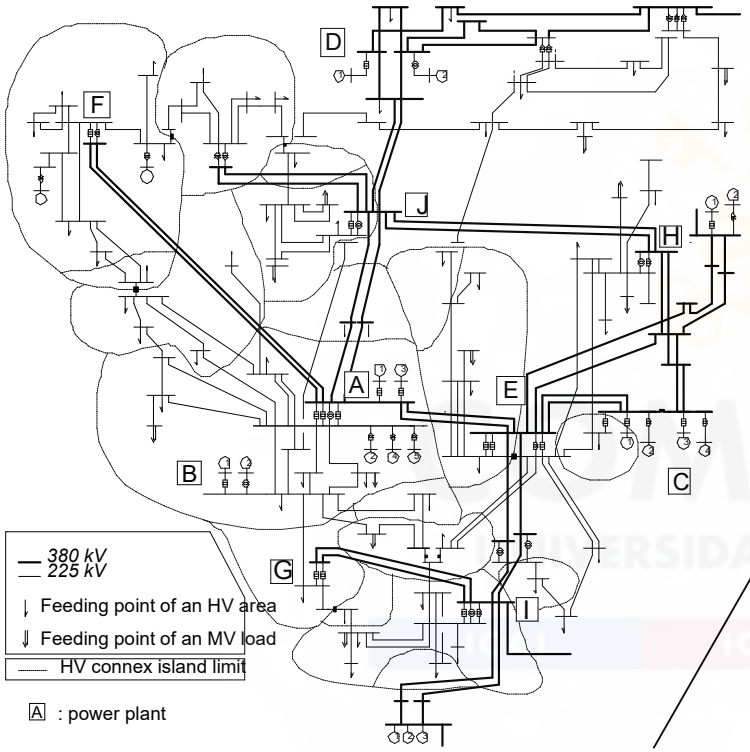


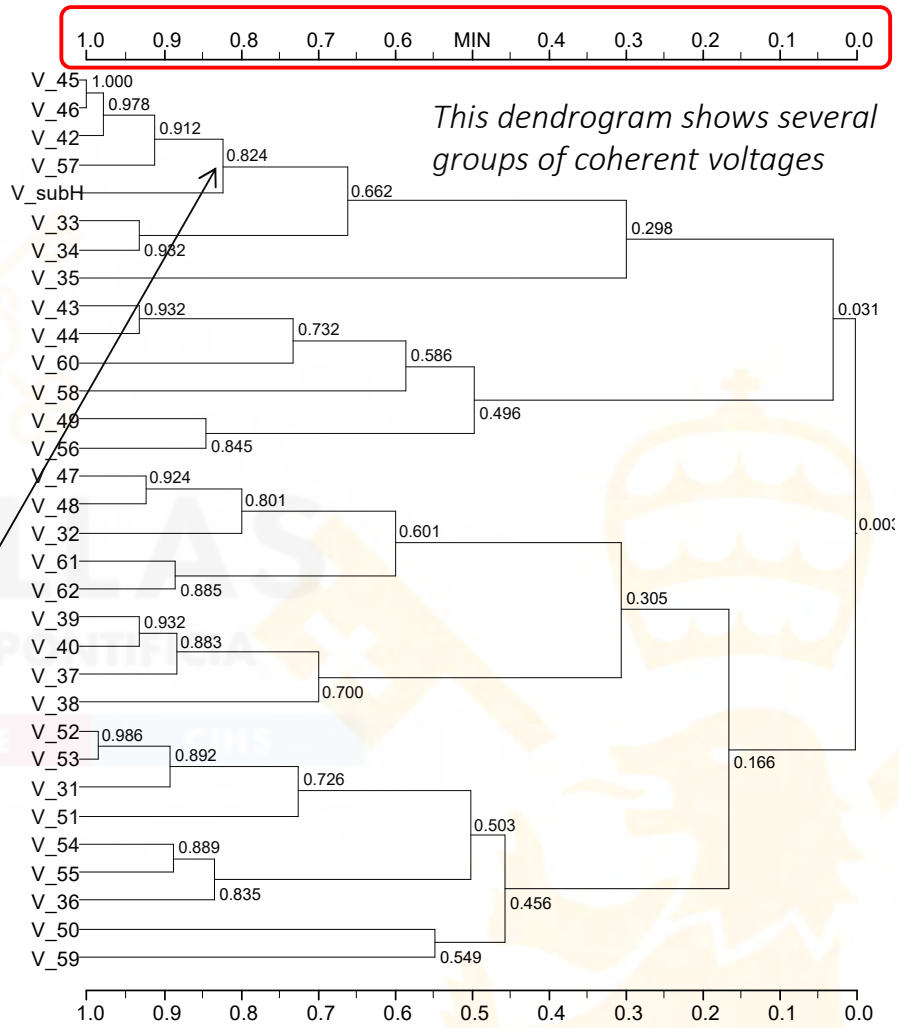
- If you accept that $x1$ and $x2$ are in the same cluster, then you must also accept that $x4$ and $x5$ are in a cluster (a different one)
- The original signals without the additive Gaussian noise:



Hierarchical clustering
Real example: EdF power system

- 5000 observations





This dendrogram shows several groups of coherent voltages

V_45, V_46, V_42, V_57, and V_subH form a rather homogeneous group of similar input variables, the correlation being at least equal to 0.824 for each pair. This is physically sound because voltages are usually controlled variables of the electric power systems

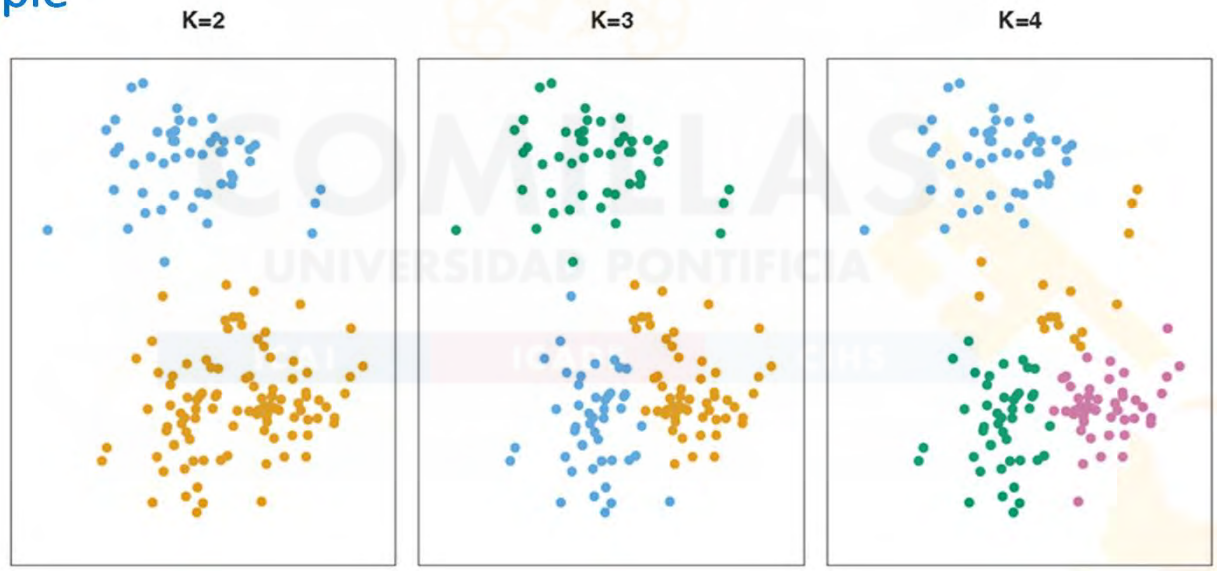
Source: E. F. Sánchez-Úbeda (1999), Models for data analysis: contributions to automatic learning, Ph.D. thesis, Universidad Pontificia Comillas, no. 255/1999.

4

K-means clustering

K-means clustering
Overview

- Usually for clustering of real-valued observations
  - Similarity measure: Euclidean, cityblock (Manhattan), etc.
- K-means clustering is a simple approach for partitioning a data set into K different, non-overlapping clusters
- Example



A simulated data set with 150 observations in a two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned

K-means clustering
Algorithm

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

- Find K clusters minimizing the quadratic quantization error $\boldsymbol{QE}$:

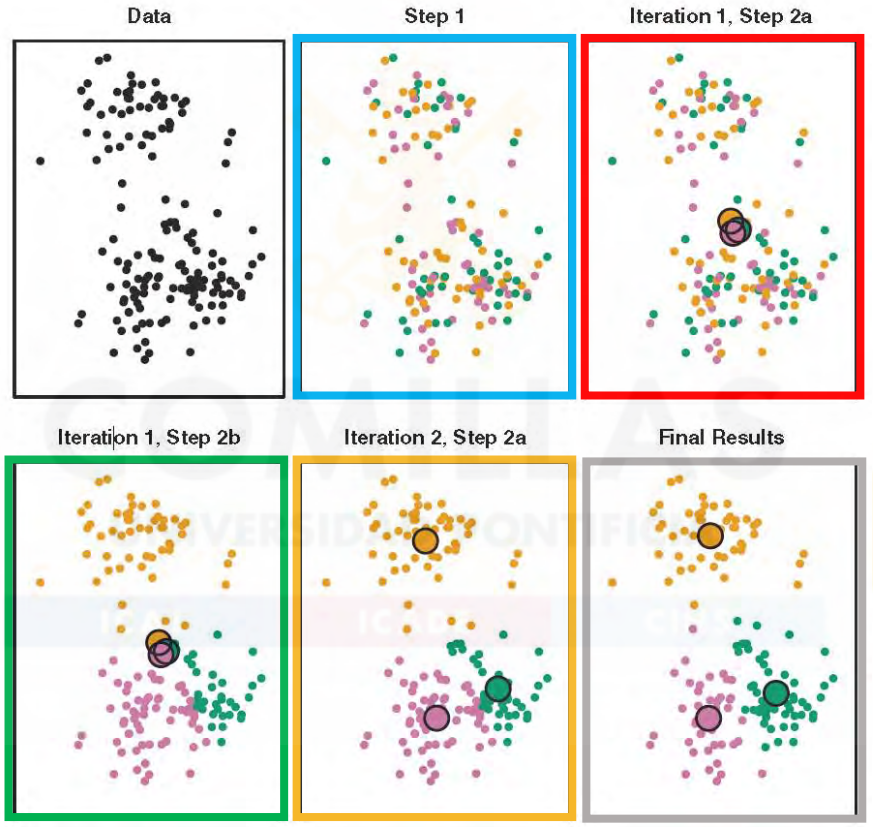
$$QE = \sum_{i=1,K} QE_i = \sum_{i=1,K} \sum_{\substack{e \in LS \\ e \in Clusteri}} \left\| \mathbf{x}_e - \mathbf{c}_i \right\|^2$$

$\mathbf{c}_i$: cluster prototype (centroid)

K-means clustering Algorithm

- Example with K=3

Step 1, each observation is randomly assigned to a cluster



In Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially, the centroids are almost entirely overlapping because the initial cluster assignments were chosen at random

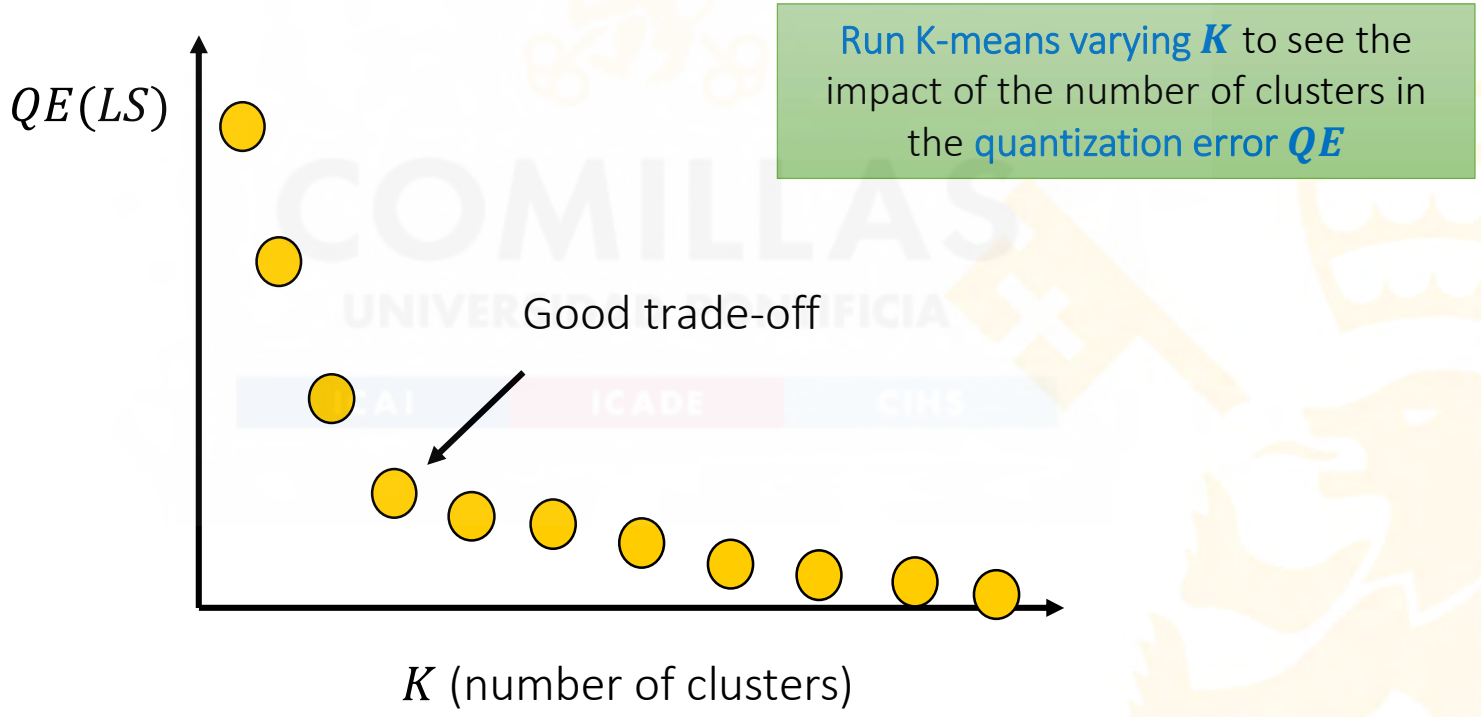
Step 2(b), each observation is assigned to the nearest centroid

Results obtained after ten iterations

Step 2(a) is once again performed, leading to new cluster centroids

K-means clustering
Selecting K

- To perform K-means clustering, we must first specify the desired number of clusters $\boldsymbol{K}$
  - The K-means algorithm will assign each observation to exactly one of the K clusters



Run K-means varying $\boldsymbol{K}$ to see the impact of the number of clusters in the quantization error $\boldsymbol{QE}$

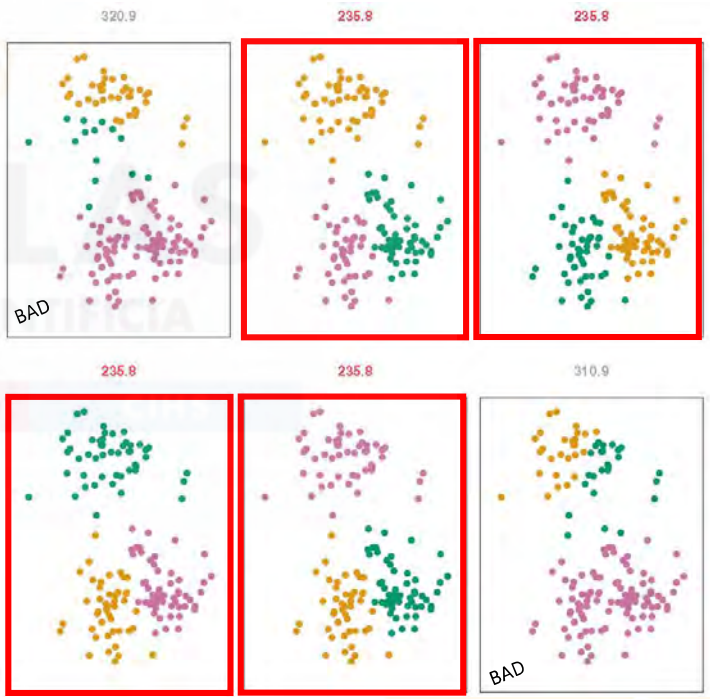
K-means clustering Algorithm

- Because the K-means algorithm finds a local *rather than a global optimum,* the results obtained will depend on the initial (random) cluster assignment

It is essential to run the algorithm multiple times from different random initial configurations (called replications)

One selects the best solution, *i.e., that for which* the quantization error is smallest

K-means clustering performed six times on the data with $K = 3$

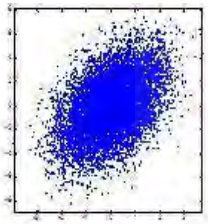


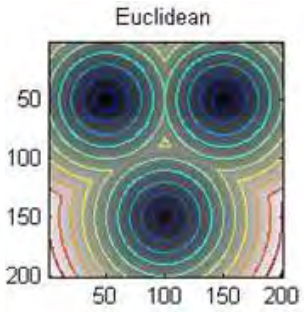
Three different local optima were obtained, one of which resulted in a smaller value of the objective and provided better separation between the clusters, with an objective value of 235.8

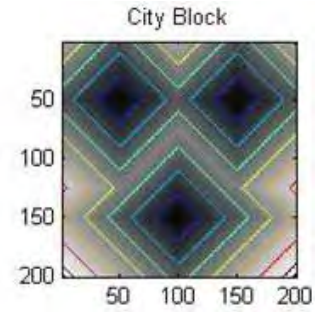
Those labeled in red all achieved the same best solution

K-means clustering
Impact of the distance metric

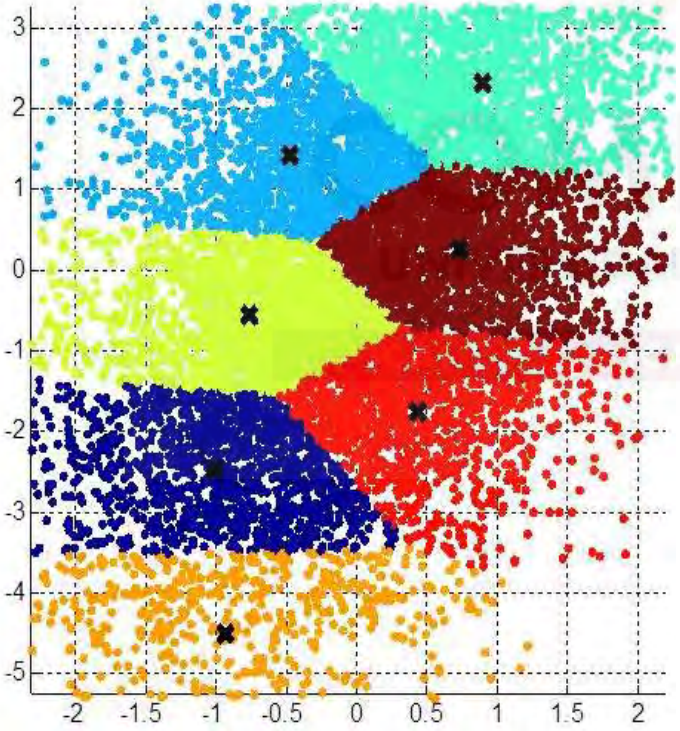
- Synthetic data, with not evident clusters



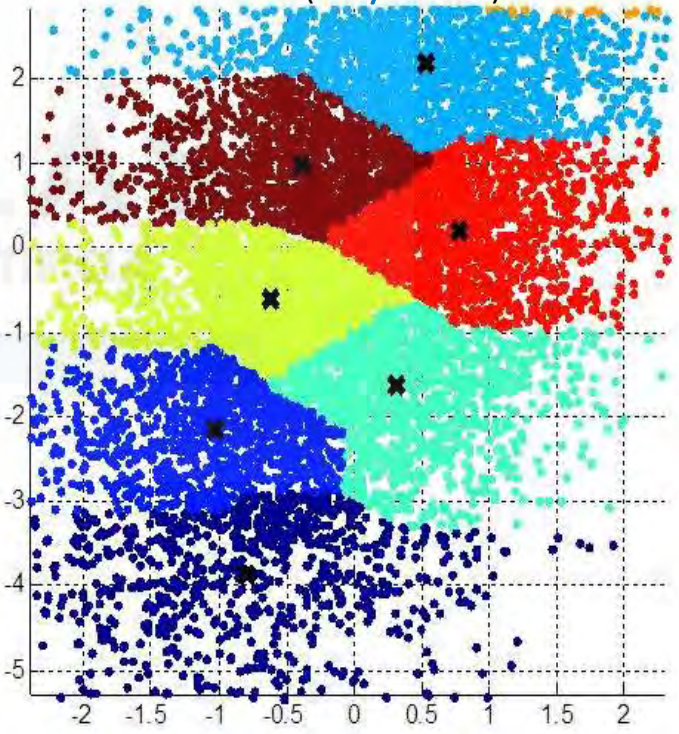




Euclidean K=7

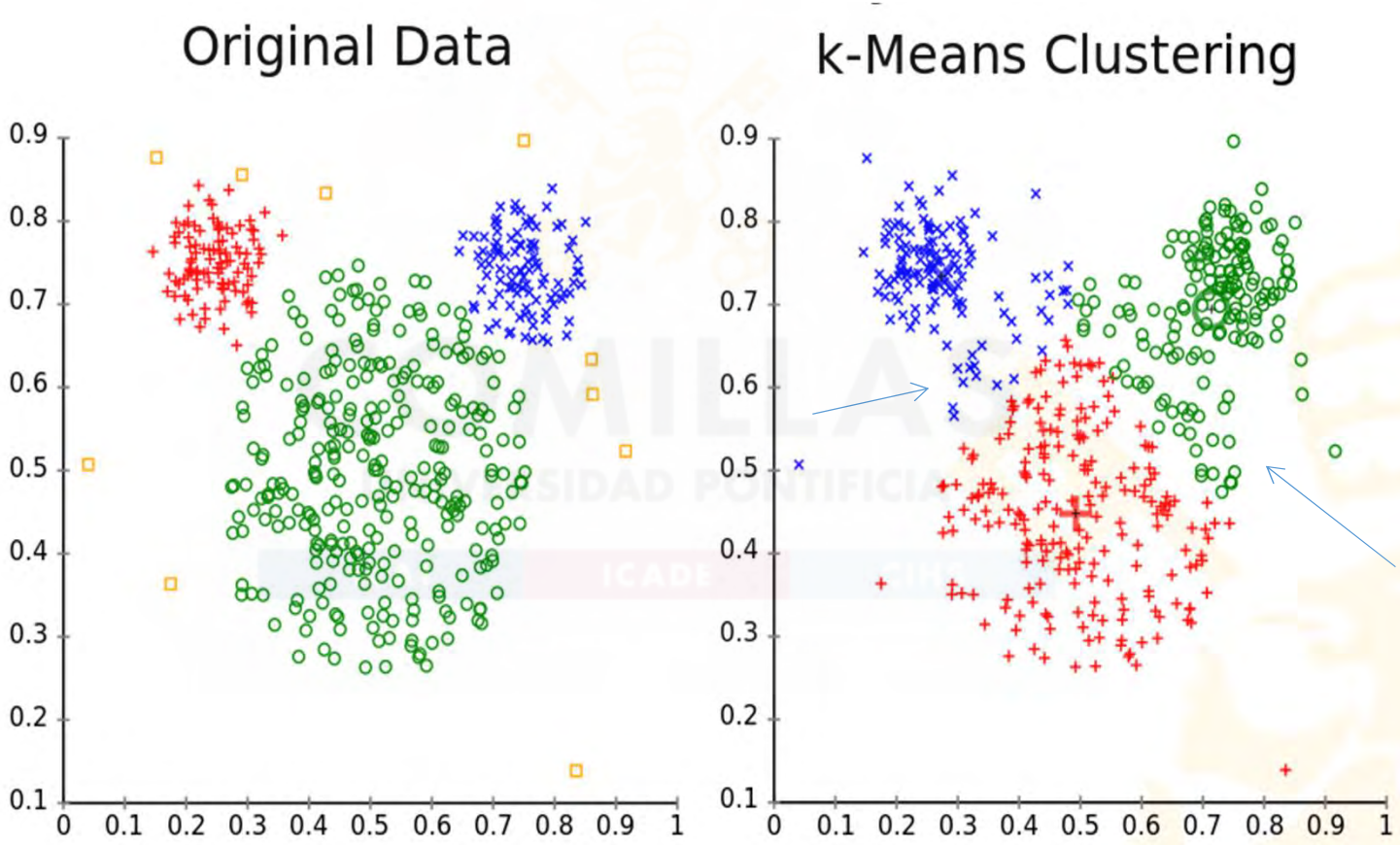


Manhattan (cityblock) K=7



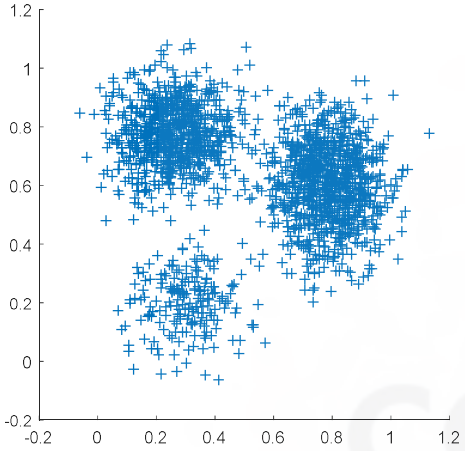
K-means clustering
Illustrative synthetic cases

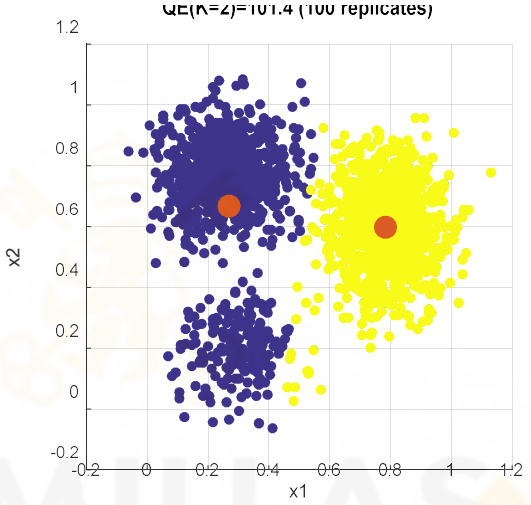
- C1: Mouse data

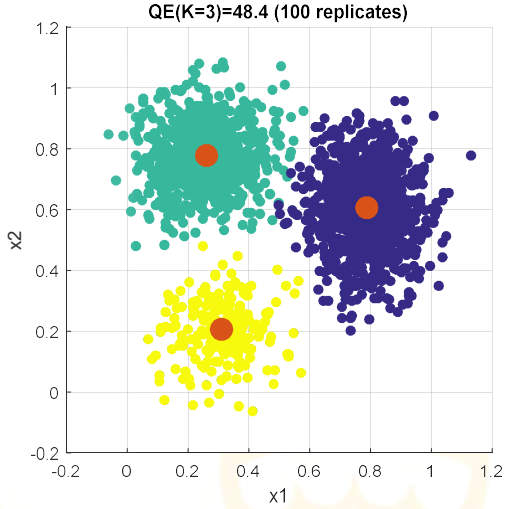


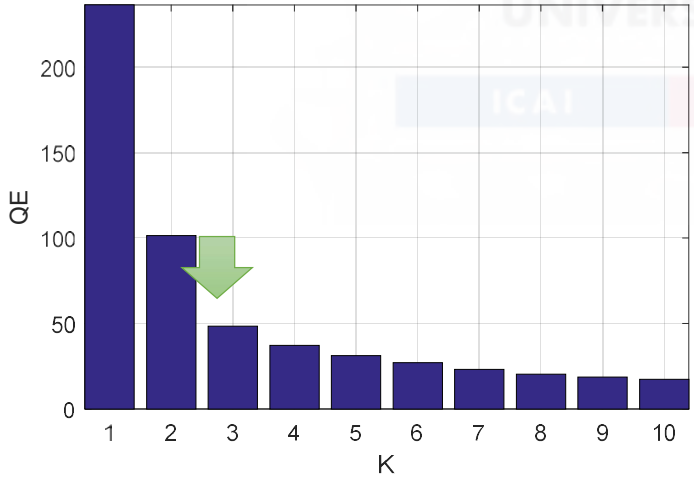
K-means clustering
Illustrative synthetic cases

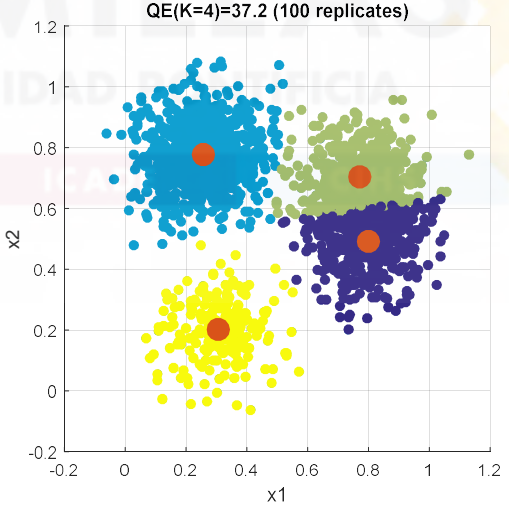
- C2: 3 clusters







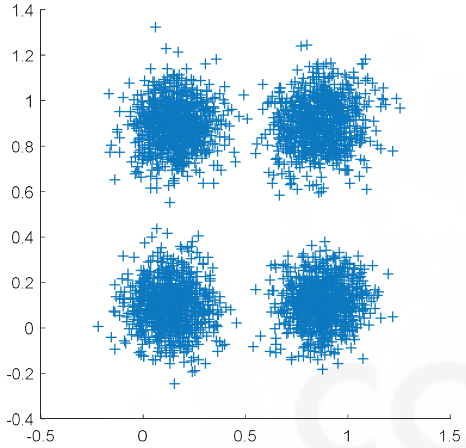


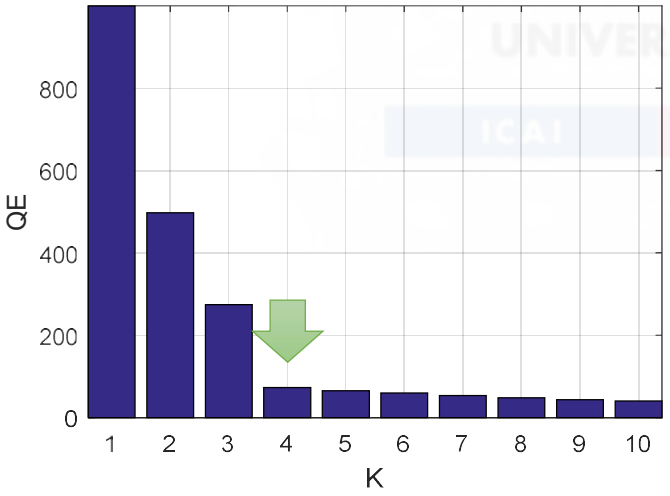


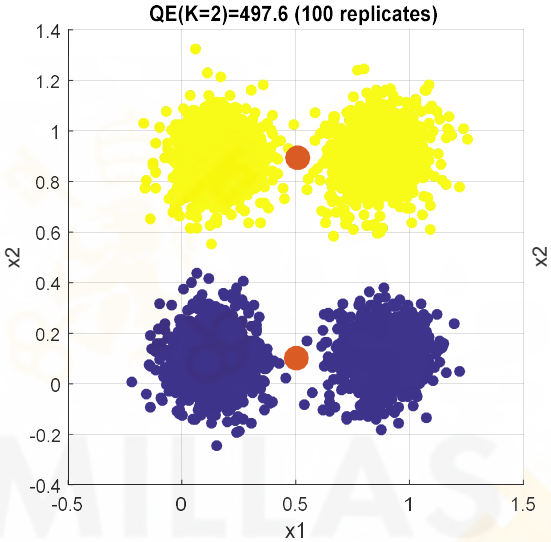
- From 2 to 3 clusters the decrease in QE is clear due to the improvement by incorporating an apparent cluster
- From 3 to 4 clusters, the decrease in QE is marginal due to an artificial split of one of the existing clusters

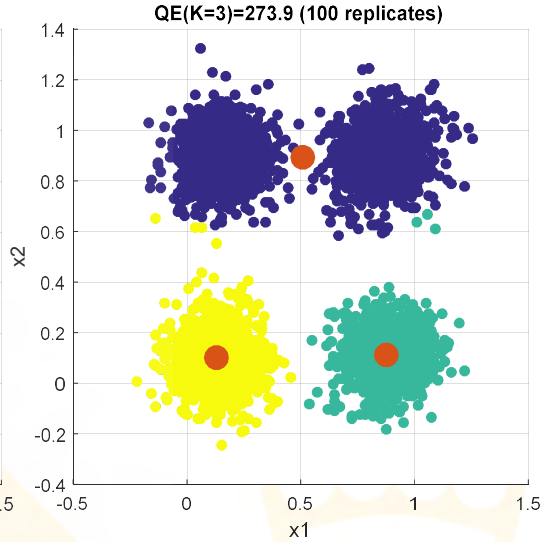
K-means clustering
Illustrative synthetic cases

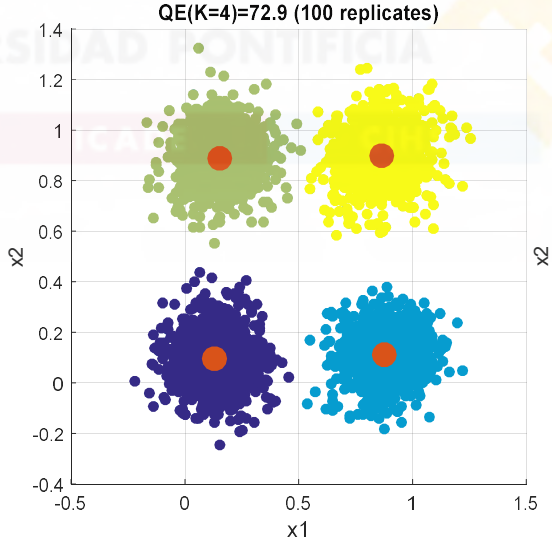
- C3: 4 clusters

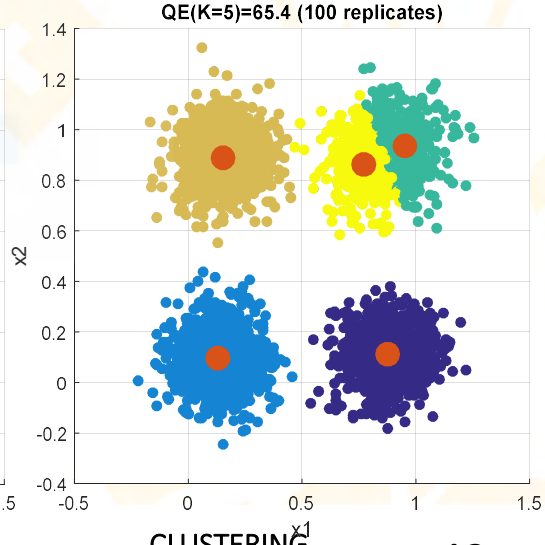








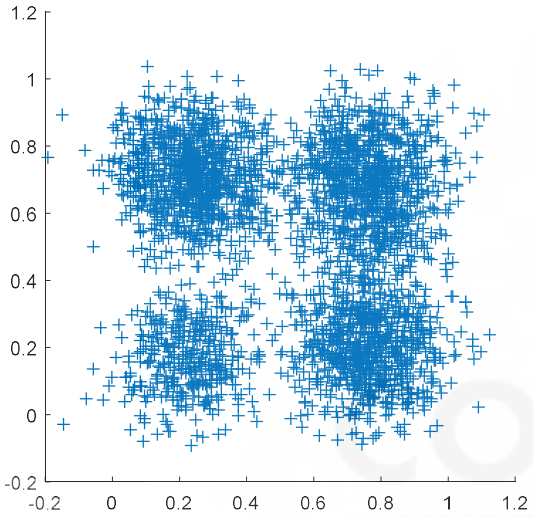


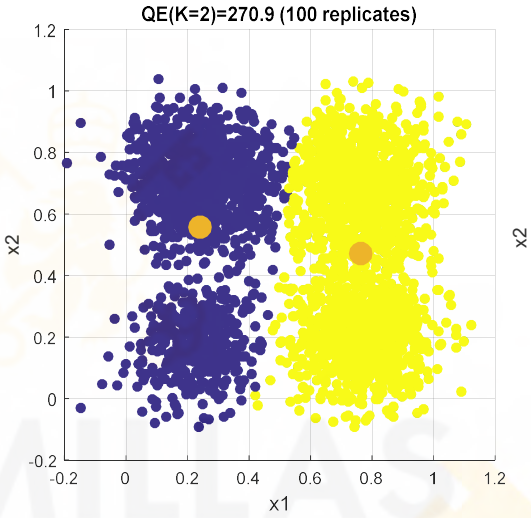


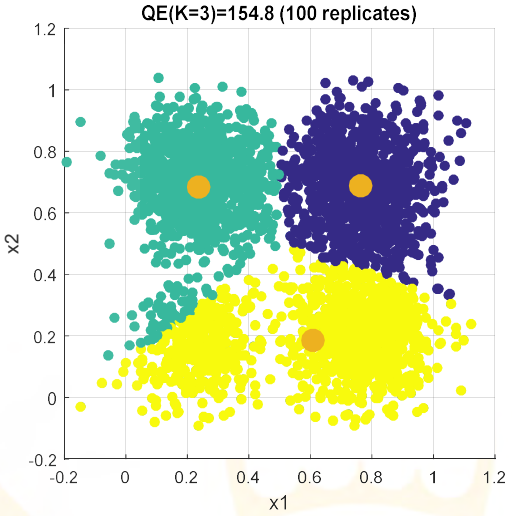
K-means clustering
Illustrative synthetic cases

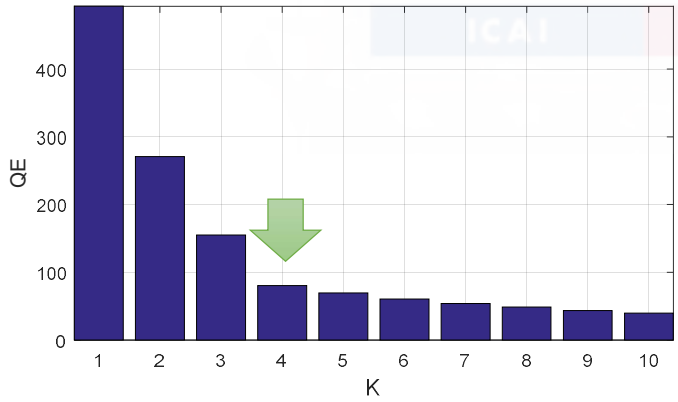
Maybe 3 clusters make no sense to you

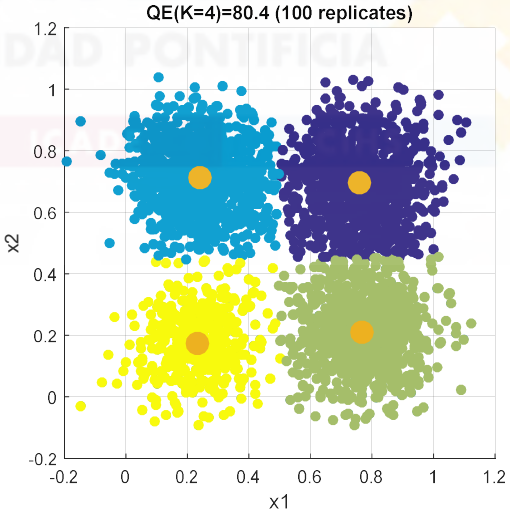
- C4: 3 or 4 clusters?

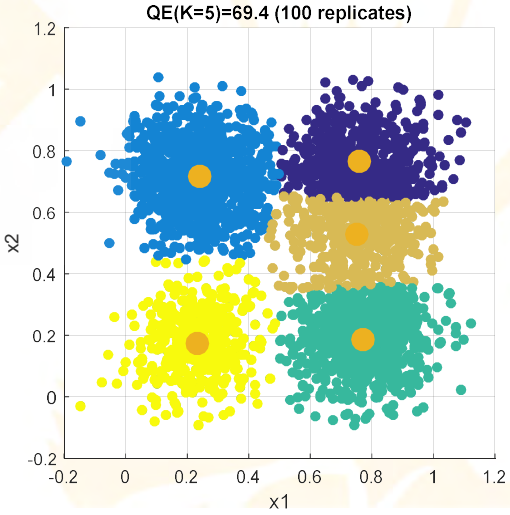






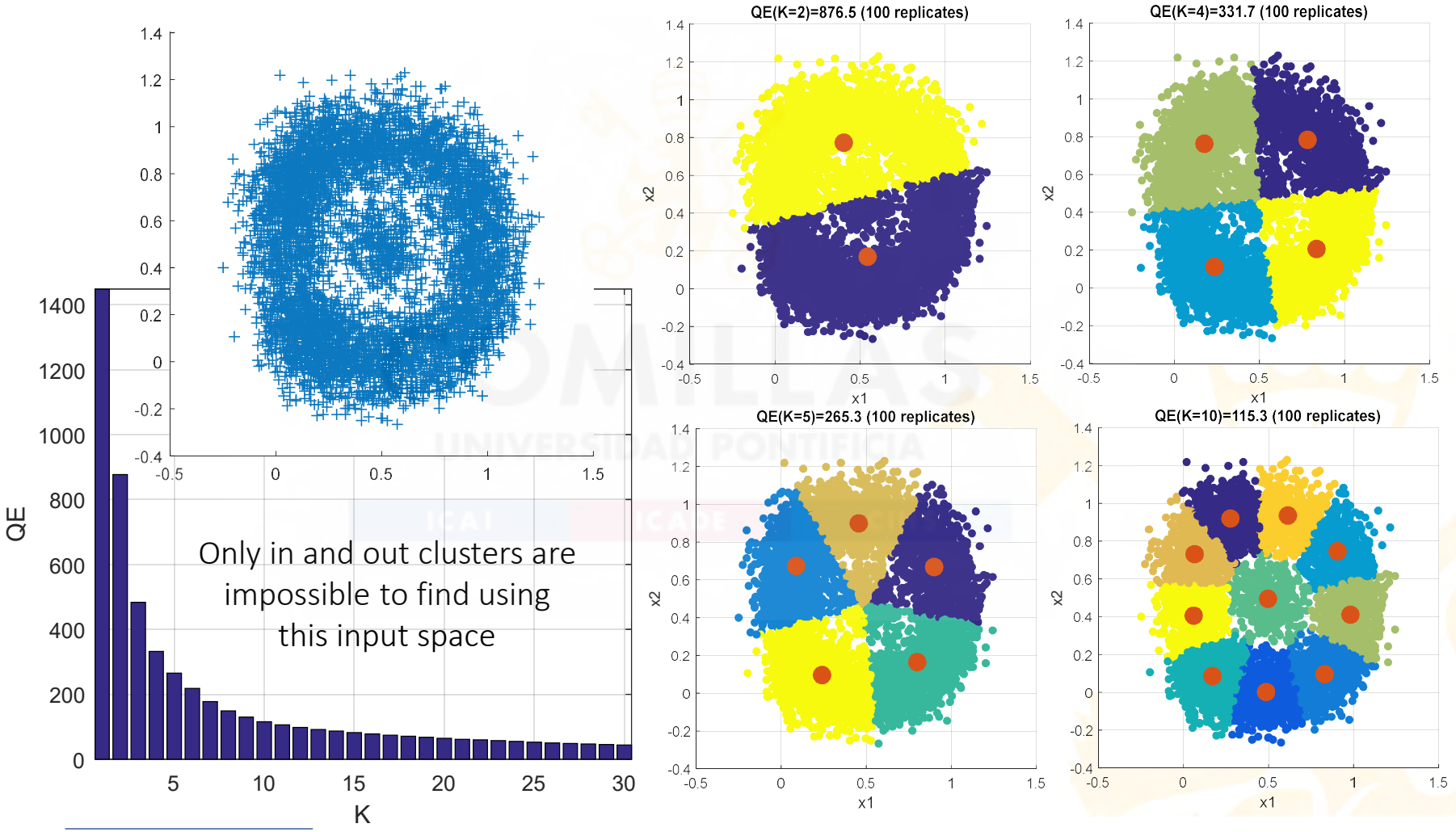






K-means clustering
Illustrative synthetic cases

- C5: 2 clusters (in and out)

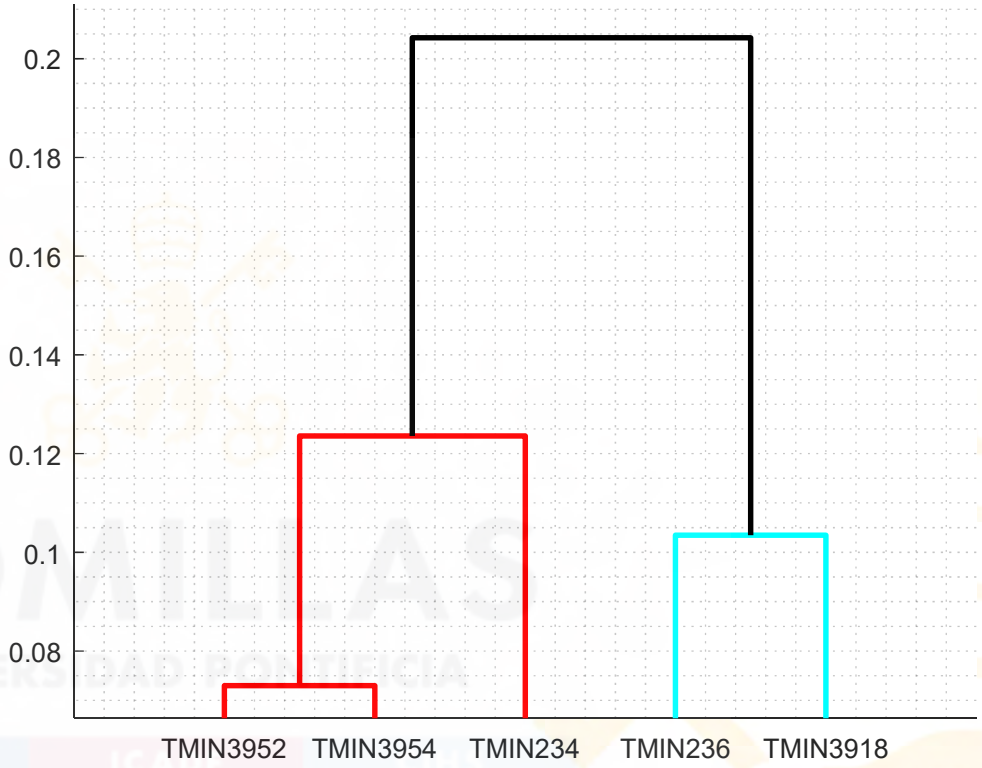


5

Quiz

Quiz
Question 1

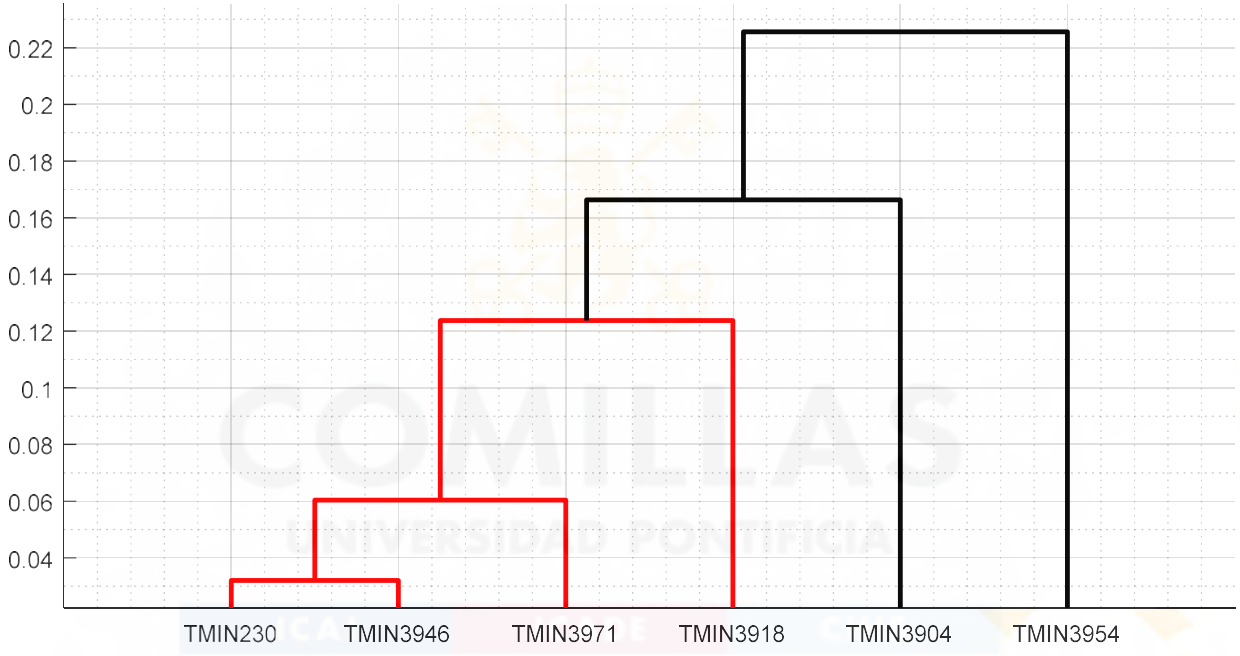
- Según el dendrograma:



A. Si se elige como umbral de disimilitud 0.1 entonces se forman 4 clusters.
B. Si se elige como umbral de disimilitud 0.1 entonces se forman 3 clusters.
C. Si se elige como umbral de disimilitud 0.1 entonces se forman 2 clusters.

Quiz
Question 2

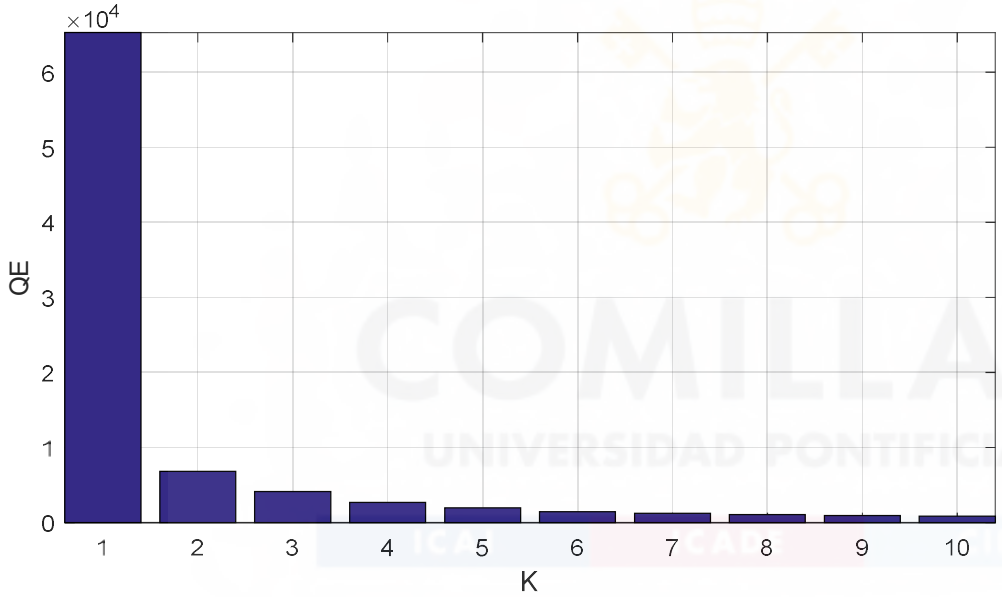
- Según el dendrograma:

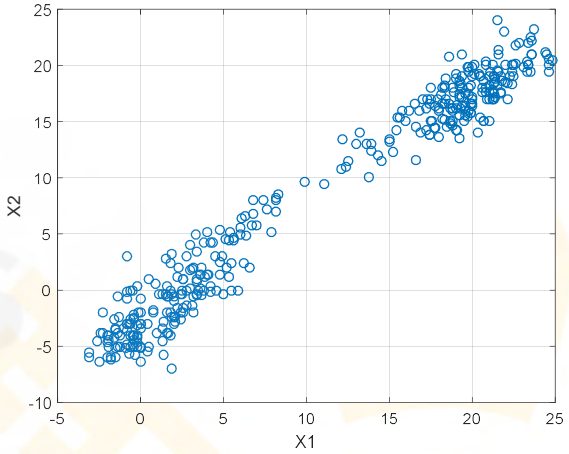


A. Los clústeres obtenidos son: ('TMIN3904'), ('TMIN3918') y ('TMIN230', 'TMIN3946', 'TMIN3954', 'TMIN3971').
B. Los clústeres obtenidos son: ('TMIN230'), ('TMIN3954') y ('TMIN3904', 'TMIN3918', 'TMIN3946', 'TMIN3971').
C. Los clústeres obtenidos son: ('TMIN3904'), ('TMIN3954') y ('TMIN230', 'TMIN3918', 'TMIN3946', 'TMIN3971').

Quiz
Question 3

- Se ha utilizado K-means para realizar una agrupación de las observaciones en un espacio de dos variables. A partir de la representación de la evolución del error de cuantización con el nº de clústeres, se puede afirmar que:

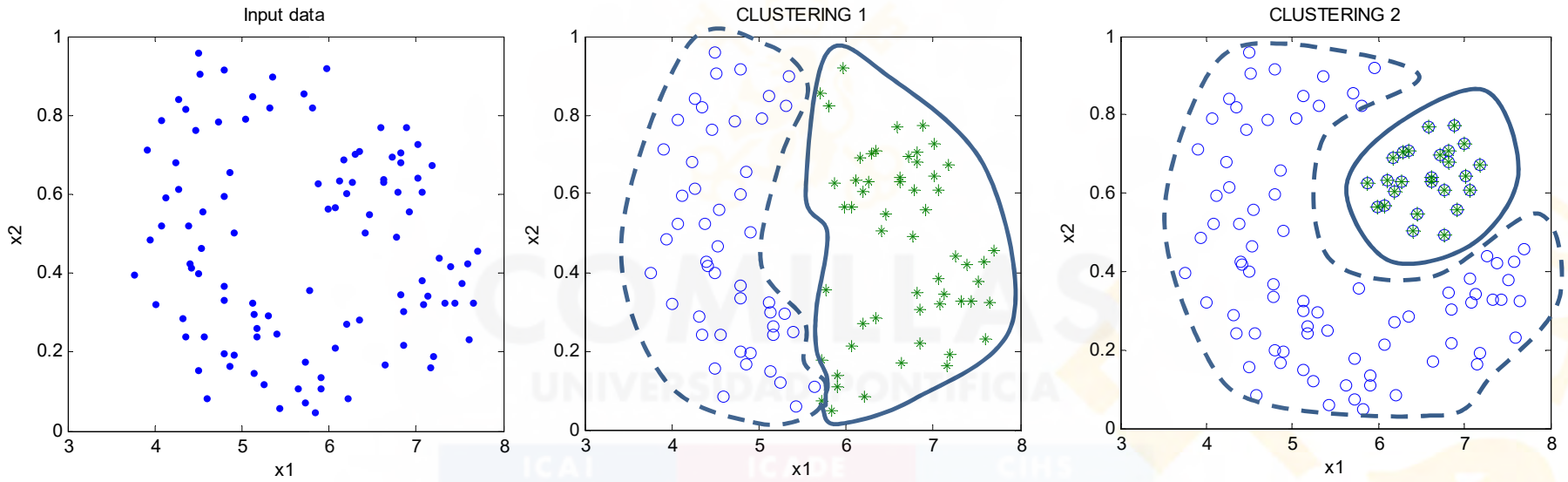




A. Parece razonable elegir 4 clústeres.
B. Parece razonable elegir 2 clústeres.
C. Parece razonable elegir 3 clústeres.

Quiz
Question 4

- Se tiene un conjunto de datos con dos variables x1 y x2, representado en la primera figura ("Input data"). Si se ejecuta el algoritmo K-means para obtener dos grupos de observaciones utilizando la distancia euclídea, entonces



A. Los dos grupos obtenidos están representados en la figura "CLUSTERING 2".
B. Los dos grupos obtenidos están representados en la figura "CLUSTERING 1".
C. Debido a la correlación entre x1 y x2, no es posible ejecutar K-means.

Quiz
Answers

- Q1-A
- Q2-C
- Q3-B
- Q4-B

6

Real examples

Real cases
Clustering of Twitter data

2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)

Clustering and Sentiment Analysis on Twitter Data

Shreya Ahuja[1]
Department of CSE
Amity University, Noida
India
shreyaahuja96@gmail.com

Gaurav Dubey[2]
Department of CSE
Amity University, Noida
India
gdubey@amity.edu

***Abstract*- Twitter is a social media platform is a great place where people from all parts of the world can make their opinions heard. Twitter produces around 500 million of tweets daily which amounts to about 8TB of data. The data generated in twitter can be very useful if analyzed as we can extract important information via opinion mining. Opinions about any news or launch of a product or a certain kind of trend can be observed well in twitter data. The main aim of sentiment analysis (or opinion mining) is to discover emotion, opinion, subjectivity and attitude from a natural text. In twitter sentiment analysis, we categorize tweets into positive and negative sentiment.**

Clustering is a protean procedure in which identically resembled objects are grouped together and form a pack or cluster. We conducted a study and found out that the use of clustering can quickly and efficiently distinguish tweets on the basis of their sentiment scores and can find weekly and strongly positive or

Clustering is a job in which we assign certain groups or classes to certain objects such that the objects within the same group or class are more similar than those in the other distinguished group or class [11].

In sentiment analysis, various things are considered as a group of things, example, sentiment scores, polarity, subjectivity, objectivity etc. I use unsupervised learning such as clustering to group such things with one another.

A. Sentiment Analysis

Identifying the mood or opinion of a person's view written in natural language is known as sentiment analysis. The positive or negative polarity is assigned after identification of the opinion [12]. There are many techniques which are applied

Real cases
Clustering of Twitter data

E.Cluster Analysis

K-means, Fuzzy C-means, Hierarchical clustering and Mixture of Gaussians are commonly used clustering techniques. For this research, as the data is very large and Euclidean distance needs to be computed in less time to make the analysis more efficient, K-means algorithm technique is used [17]. The chief objective is to assign k number of centroids which is for each cluster.

1. Choose K number of clusters randomly, far from each other.
2. Calculate the distance between each data point and the centroid, i.e. cluster center.
3. The data point is assigned to the cluster with whose centroid it has the minimum distance.
4. Once all data points have been assigned to the clusters, the centroids are recalculated using data from results obtained in previous step with the help of the formula:

$$C_i = (1/n) * \sum_{j=1}^{n} x_i^{(j)}$$

5. Repeat steps 2-4 until no new k centroids are formed.

To choose the appropriate value of k for k means clustering, elbow method and silhouette methods are used which are graphical methods [20]. Silhouette method requires a lot of

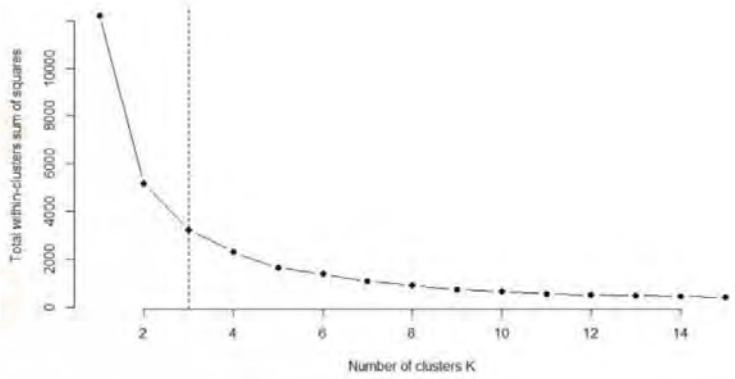
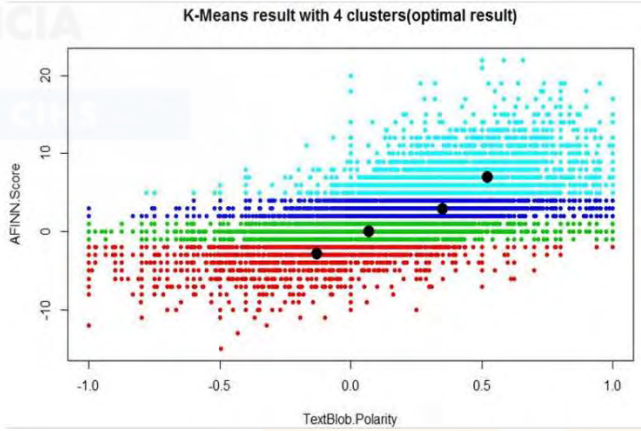


Figure 3: Graph showing elbow method to find 'K'.



Real cases
Clustering of Twitter data

Cluster 'blue'- It includes the tweets with subjectivity greater than 0.4 and going up to 1.0 and sentiment scores from -1.0 to +0.3, i.e. from extremely negative to very slightly positive. Hence this cluster covers highly opinioned, negative tweets and is the most widespread.

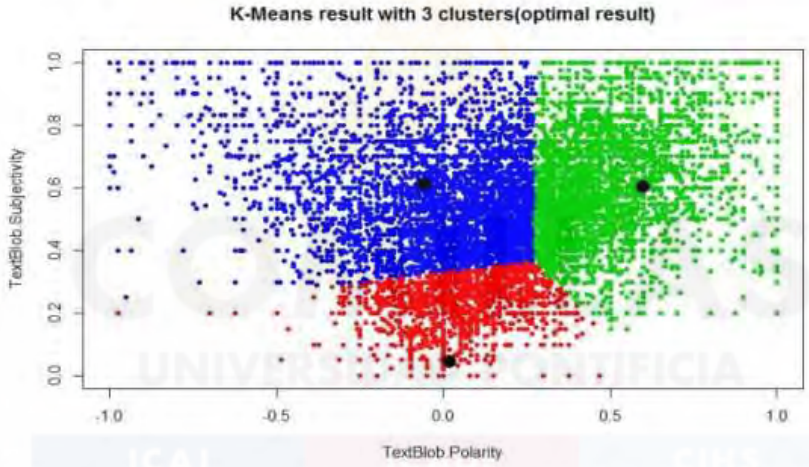


Figure 7: Resulting clusters for the plot b/w Polarity and Subjectivity scores using K-means clustering.

Cluster 'green'- It includes the tweets with subjectivity greater than 0.2 and going up to 1.0 and sentiment scores from +0.4 to +1.0, i.e. this cluster as a whole contains only positive tweets with varying subjectivity.

Real cases
Clustering of European temperatures

Article

Time Series Decomposition of the Daily Outdoor Air Temperature in Europe for Long-Term Energy Forecasting in the Context of Climate Change

Santiago Moreno-Carbonell, Eugenio F. Sánchez-Úbeda * and Antonio Muñoz

Institute for Research in Technology (IIT), ICAI School of Engineering, Comillas Pontifical University, 28015 Madrid, Spain; santiago.moreno@iit.comillas.edu (S.M.-C.); antonio.munoz@iit.comillas.edu (A.M.)
* Correspondence: eugenio.sanchez@iit.comillas.edu (E.F.S.-Ú.)

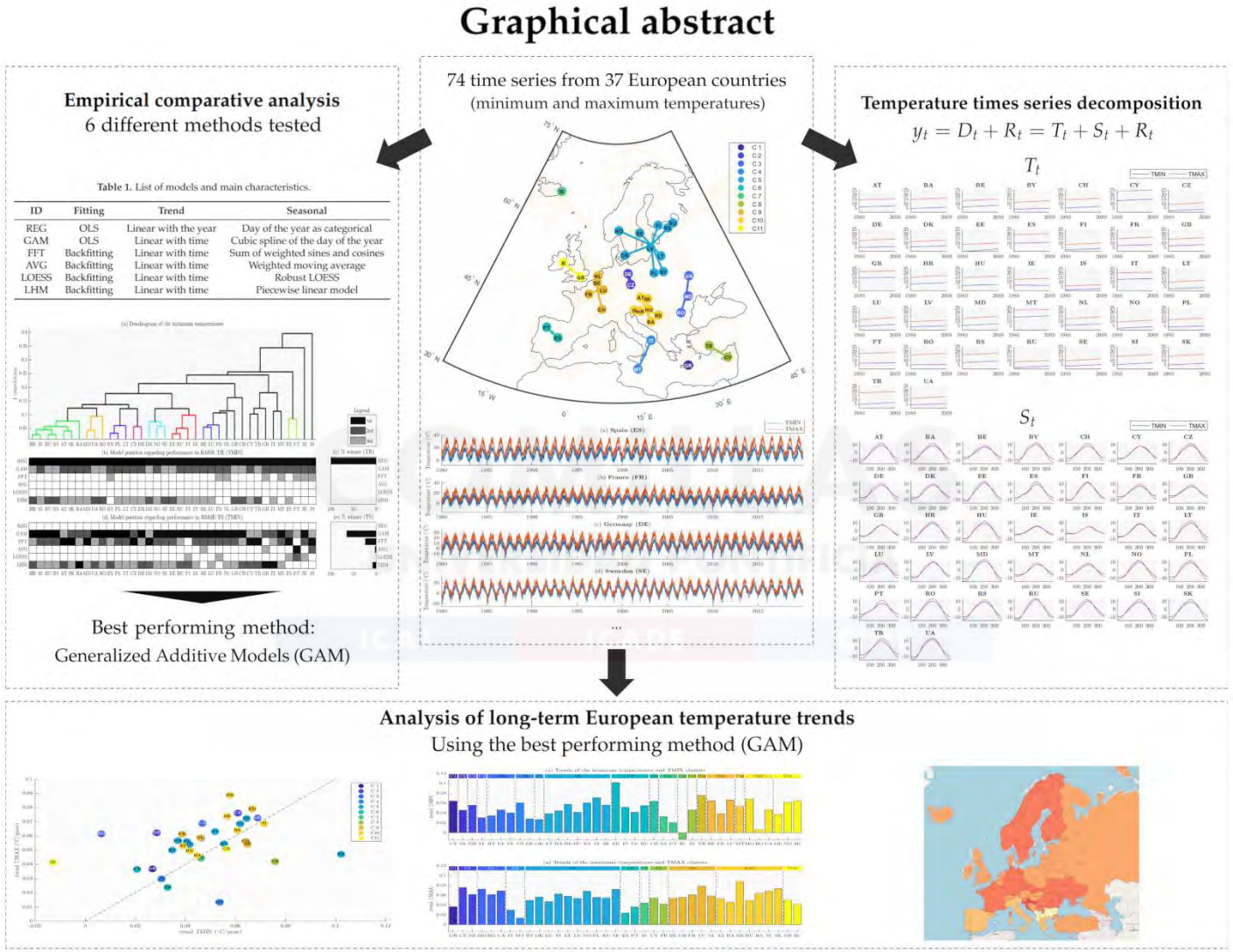
Received: 22 February 2020; Accepted: 24 March 2020; Published: 29 March 2020

Abstract: Temperature is widely known as one of the most important drivers to forecast electricity and gas variables, such as the load. Because of that reason, temperature forecasting is and has been for years of great interest for energy forecasters and several approaches and methods have been published. However, these methods usually do not consider temperature trend, which causes important error increases when dealing with medium- or long-term estimations. This paper presents several temperature forecasting methods based on time series decomposition and analyzes their results and the trends of 37 different European countries, proving their annual average temperature increase and their different behaviors regarding trend and seasonal components.

Keywords: temperature forecasting; time series; decomposition methods; generalized additive models; cross-validation; climate change

https://www.mdpi.com/1996-1073/13/7/1569

Real cases
Clustering of European temperatures



Real cases
Clustering of European temperatures

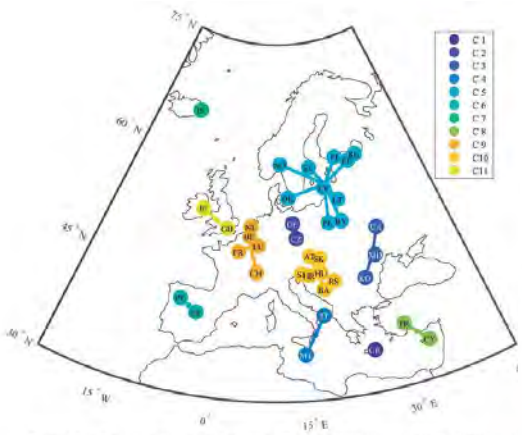


Figure 8. Location of the reference weather stations. The coloured clusters correspond to those formed using the dendrogram of Figure 7 (*Top*), based on the maximum temperature.

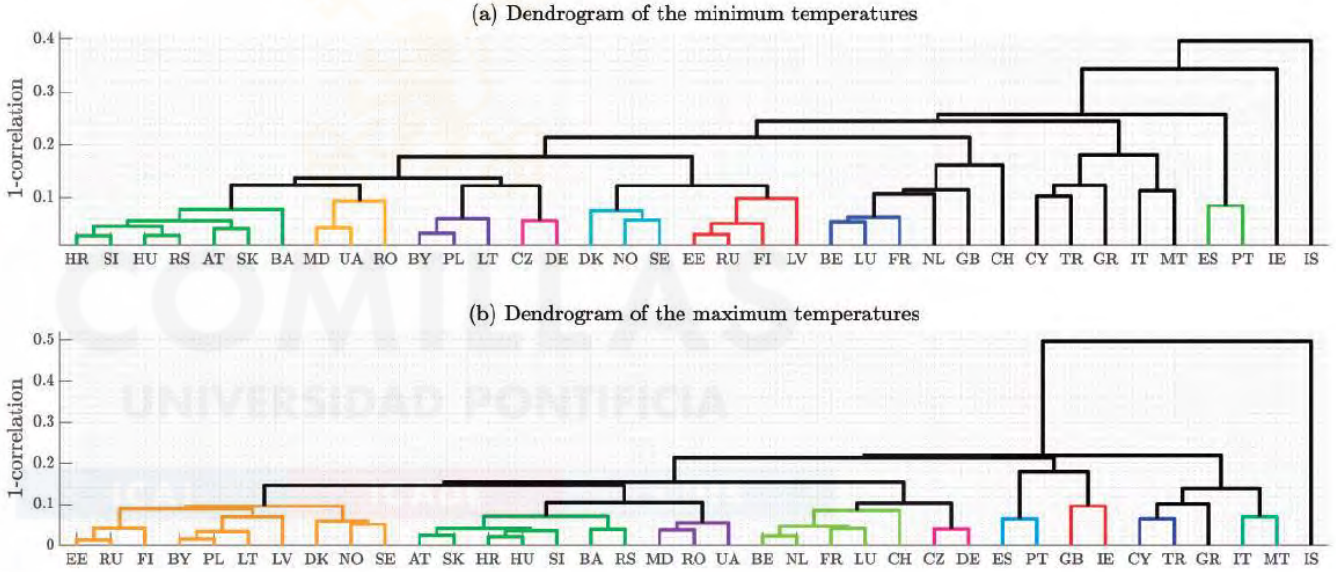


Figure 7. Dendrograms of the 37 European weather stations, based on the (**a**) minimum and (**b**) maximum temperatures. The coloured clusters correspond to those formed using a correlation threshold of 0.9.

A

Dictionary

Clustering
Dictionary

- **Bidding curve** – curva de oferta
- **Clustering analysis** – análisis de conglomerados
- **Quantization error** – error de cuantización
- **Dendrogram** – dendrograma
- **Dissimilarity distance** – distancia de disimilitud, disparidad
- **k-means clustering** – agrupación de k-medias
- **Linkage method** – Método de enlace, vinculación
- **Non-overlapping clusters** – conglomerados no solapados
- **Pattern** – patrón, perfil
- **Hierarchical clustering**- agrupación jerárquica
- **Similarity distance** –distancia de similitud (semejanza)

Thank you for your attention

Eugenio Sánchez Úbeda