First ELECON Workshop
Towards Efficient European and Brazilian Electricity Markets

Sérgio Ramos

scr@isep.ipp.pt

ISEP, Porto, Portugal

25th September 2013

ELECON

ELECTRICITY CONSUMPTION ANALYSIS & ENERGY EFFICIENCY

POLITÉCNICO DO PORTO

gecad

Grupo de Investigação em Engenharia do Conhecimento e Apoio à Decisão

Data Mining Applications in Energy and Power Systems

ELECON
ELECTRICITY CONSUMPTION ANALYSIS & ENERGY EFFICIENCY

Presentation outline

Data preprocessing

Data Mining: Used techniques (clustering and classification)

Case study presentation

- I – Electrical consumers characterization
- II – Data Mining Contributions to Characterize Zonal Prices
- III – Data Mining Based Methodology for Wind Forecasting

ELECON
ELECTRICITY CONSUMPTION ANALYSIS & ENERGY EFFICIENCY

Motivation

- ✓ In real applications data tend to be inconsistent, incomplete and / or wrong

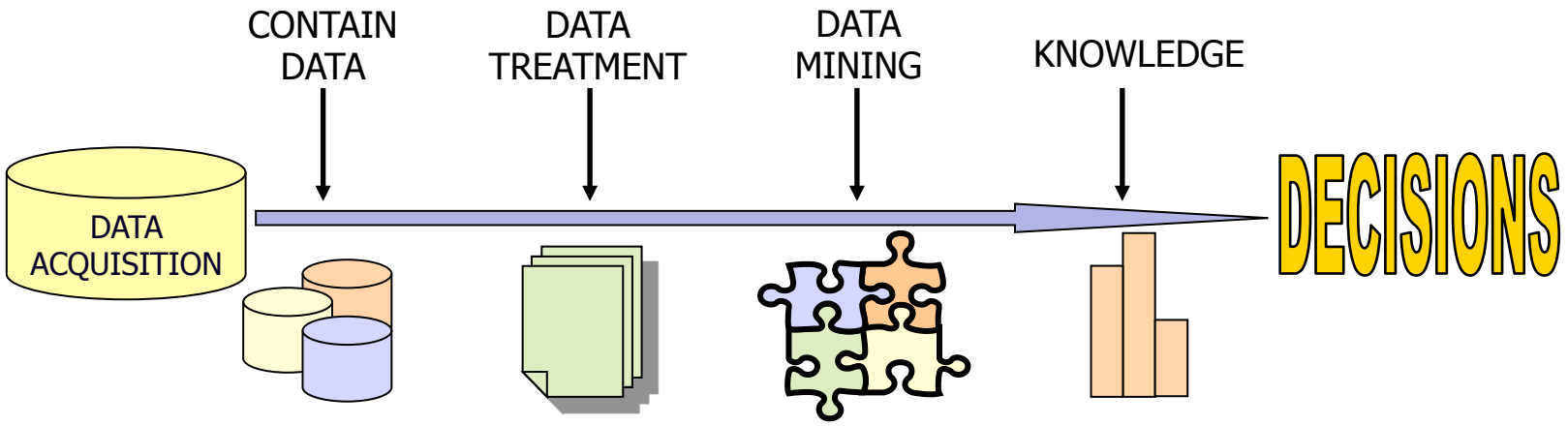
- What happens when data are not correct?

- The knowledge extracted from databases can be reliable?

Obstacles to Knowledge Discovery → Poor data

Introduction

Knowledge Discovery in Databases



ELECON
ELECTRICITY CONSUMPTION ANALYSIS & ENERGY EFFICIENCY

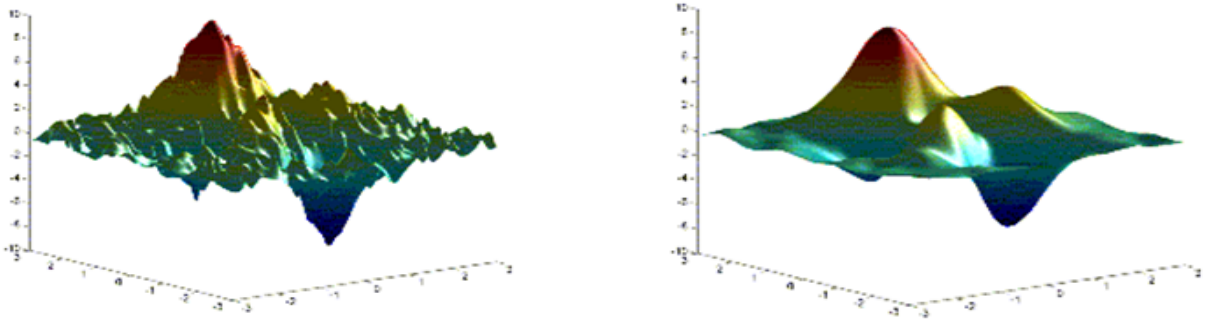
Data preprocessing – Goals

Understand the nature of data

Knowing what useful information is available in a certain data set so that it may be preserved when random subsets of samples are formed

Adapting data according to Data Mining (DM) algorithms

Typically, it is estimated that data preparation takes 70-80% of the whole effort to develop a Data Mining study



Data preprocessing – Techniques

- How can data be preprocessed in order to help improve the quality of data and, consequently, of the mining results?

There are a number of data preprocessing techniques:

- ✓ ***Data cleaning –*** can be applied to remove noise and correct inconsistencies in data
- ✓ ***Data integration –*** merges data from multiple sources into a coherent data store, such as a data warehouse
- ✓ ***Data transformations –*** such as normalization, may be applied. For instance, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements
- ✓ ***Data reduction –*** can reduce data size by aggregating, eliminating redundant features, or clustering

Data Cleaning Phase

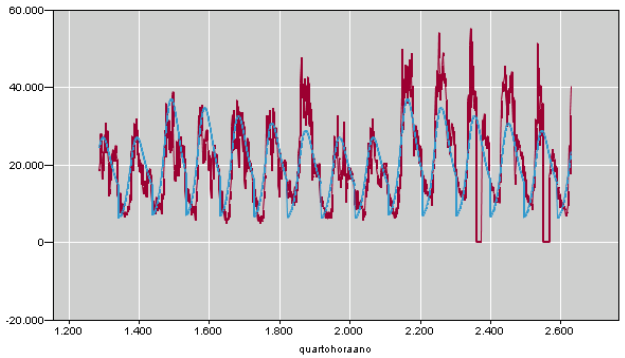
Treatment of missing values

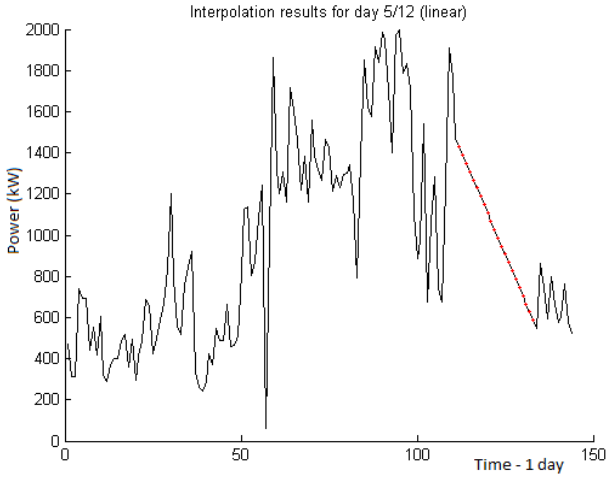
Removing the files without registered values

Fill in missing values manually (if in reduced numbers)

Estimation values

- Average of "k-neighbors" for numeric data
- Linear Regression
- Neural network





Preprocessing Phase - Normalization

Normalization

Min-Max – The Min-max normalization performs a linear transformation of the original input set to a new specific set (typically 0-1)

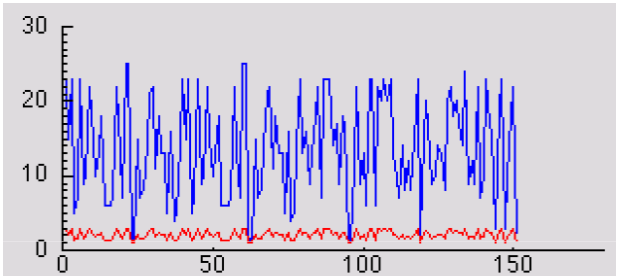
$$y' = \frac{y - \min_1}{\max_1 - \min_1} \times (\max_2 - \min_2) + \min_2$$

The new values min_2, max_2 are defined by the analyst

Preserves exactly all initial data values relations

Does not introduce any changes in the data

The shape of the diagram is maintained



Preprocessing Phase - Normalization

Normalization

Zscore – Transforms data of the input variables, such that:

- The average is 0
- the variance is 1

$$y' = \frac{y - average}{standard \quad deviation}$$

Zscore normalization works well when:

- Do not know the maximum and minimum of the input variables (there are missing values in the sample)

- If has isolated values that dominate the Min-Max normalization

Preprocessing Phase - Normalization

Normalization

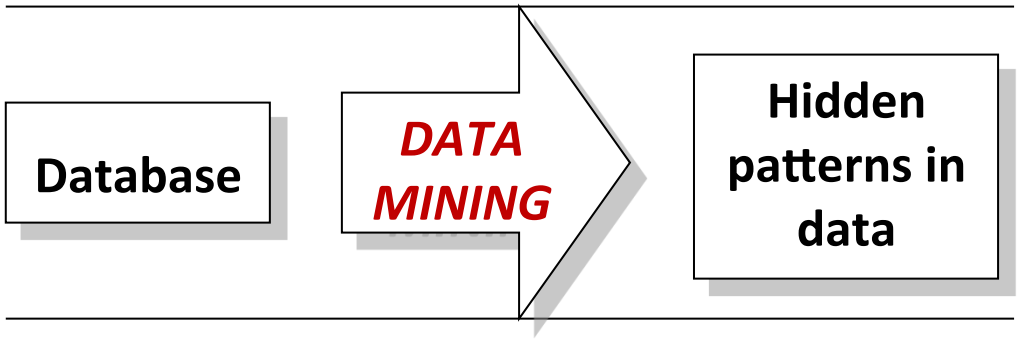
Sigmoidal – Normalizes nonlinear input data into a [-1,1] interval using the sigmoidal function

$$y' = \frac{1 - e^{-\alpha}}{1 + e^{-\alpha}} \qquad with \quad \alpha = \frac{y - average}{standard \quad deviation}$$

The sigmoidal normalization is appropriate when it is intended to include outliers in data set to analyze

Data Mining - Concept

Data Mining (DM) represents the task of finding new knowledge, generally unpredictable, based on a dataset previously collected and properly prepared for this purpose



Data Mining - Concept

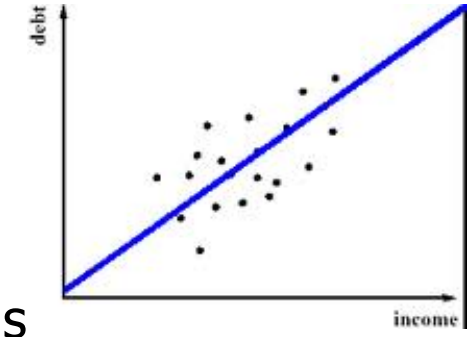
The DM process involves the use of algorithms for determining patterns and relationships in data. The main tasks of DM are:

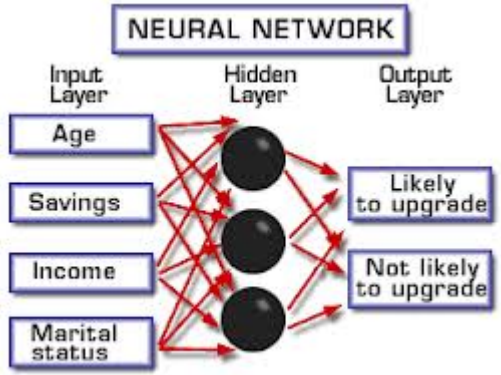
- Classification
- Estimation
- Grouping by affinity or association
- Clustering
- Analysis of deviations

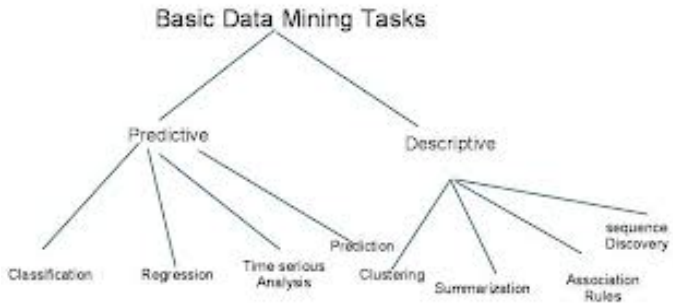
Data Mining - Techniques

The most known DM techniques used are:

- Decision trees
- Regression
- Neural networks
- Genetic algorithms
- Clustering algorithms
- Nearest neighborhood algorithm
- Fuzzy Logic
- Rough Sets



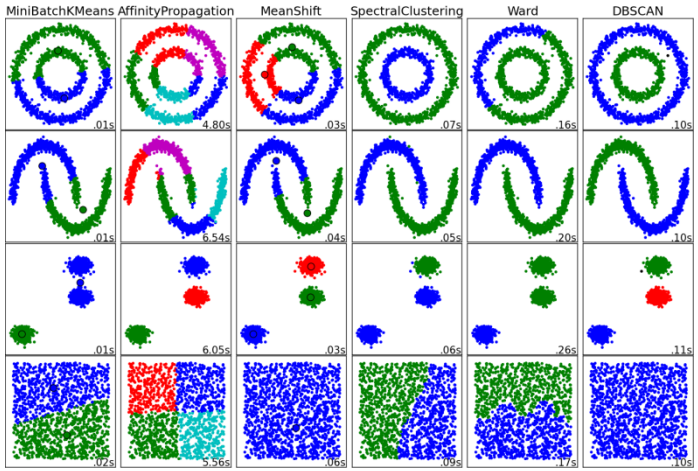




Data Mining - Clustering

Clustering can be defined as the process of partitioning a large database into groups (**or clusters**) based on a concept of similarity or proximity among data

There is a wide variety of clustering algorithms, although there is no single algorithm that can, by itself, discover all sorts of cluster shapes and structures



Data Mining – Clustering algorithms

There are a great range of clustering algorithms, such as:

- K-means algorithm
- Two-step algorithm
- Single-link
- Average-link
- Complete-link
- Ward's-link
- Normalized cut algorithm
- PC-k-means – Pairwise Constrained K-Means
- MPC-K-means – Metric Pairwise Constrained K–Means
- SOM - Self Organizing Features Maps (Kohonen network)
- WEACS – Weighted Evidence Accumulation Clustering using Subsampling

Data Mining – Classification

In classification problems, a set of pre-classified data points are given and the classification algorithm tries **to discover a rule**.

A classification problem is a supervised learning task

The classification model should allow the attribution of a new consumer to a certain cluster, based on the rules generated by the classification model

Case study I –
Electrical consumers characterization

Consumers characterization

Sample 1.022 Medium Voltage Consumers

Utility EDP - Distribuição

Sample description

Data acquisition:
- 1 year
(2010/2011)

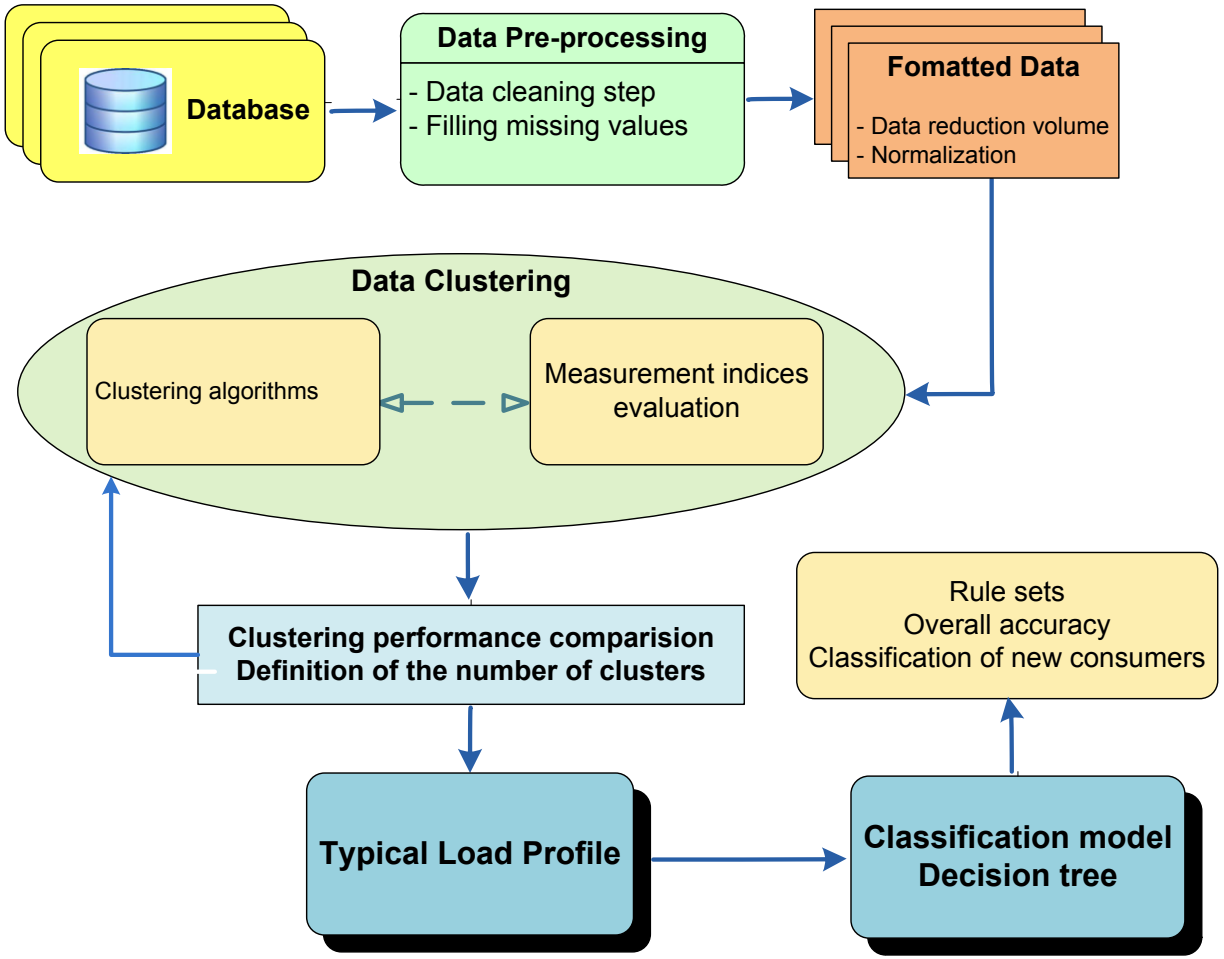
Measurement:
- Active power
- Reactive power
(15-min intervals)

Daily electricity consumer vector:

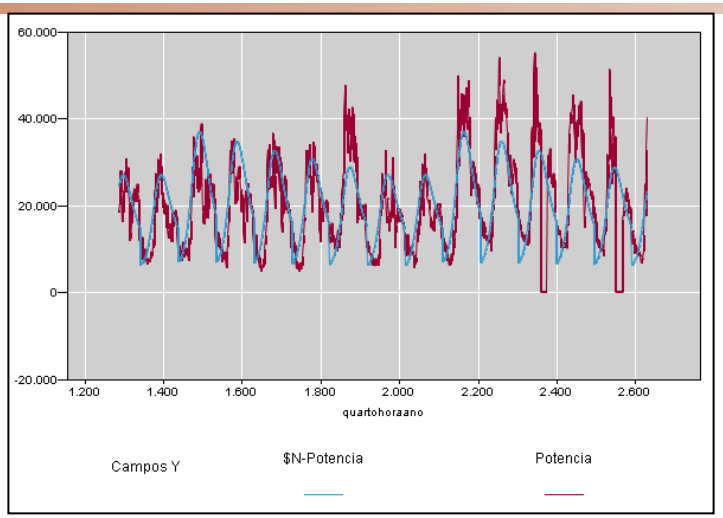
$$\mathbf{l(m) = \{l_1(m), \ldots, l_{96}(m)\} \text{ with } m = \text{number of customers}}$$

Case study I – *Electrical consumers characterization*

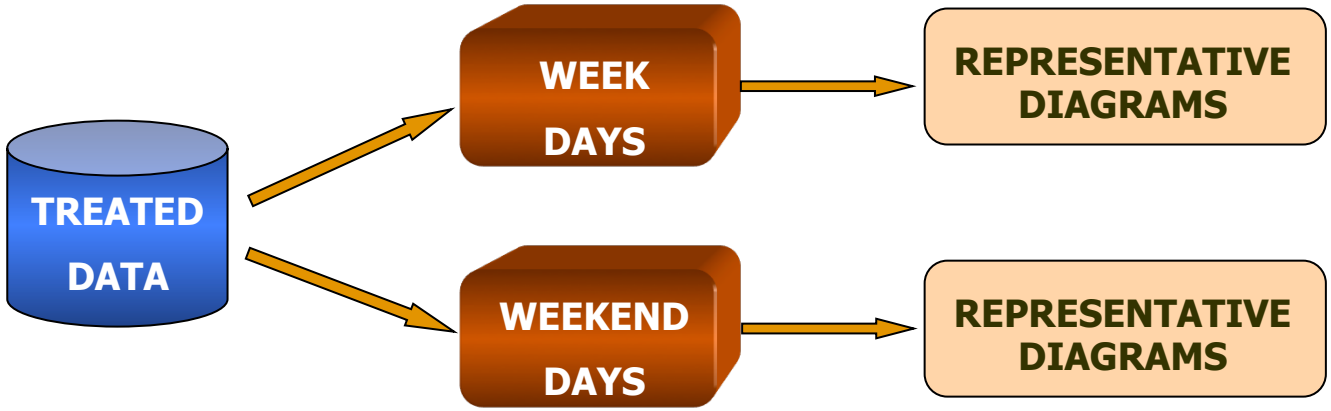
Proposed methodology architecture



Case study I – *Electrical consumers characterization*



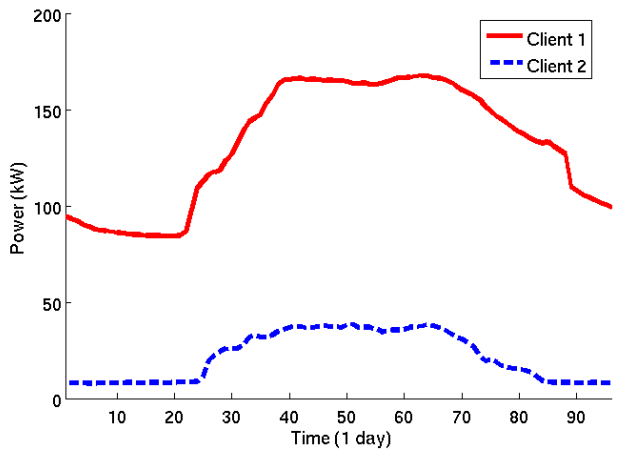
- ✓ **Preprocessing phase**
  - Verify all consumers records
  - **Identify missing values**
  - To estimate missing values of measures a *multi layer perceptron* – MLP – artificial neural net was used
  - **Data reduction – Distinguish working and weekend days**

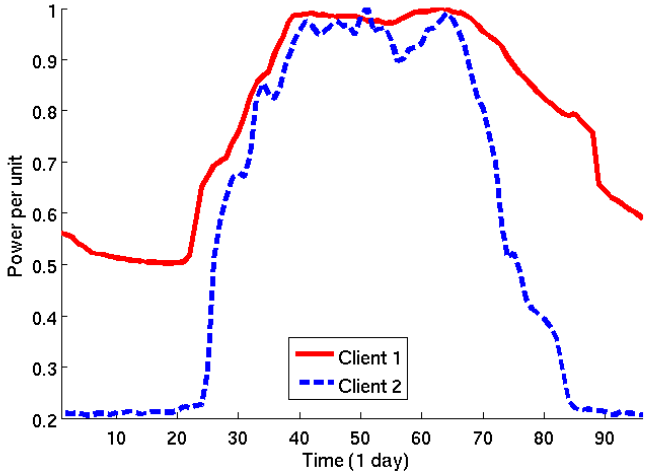


Case study I – *Electrical consumers characterization*

Preprocessing Data

- Verify all records belonging to customer's files
- Power consumption normalization – [0-1] (to be compared among them)





Case study I – *Electrical consumers characterization*

Clustering Process:

To assemble the representative load diagrams in clusters 4 clustering algorithms have been used

- K-means algorithm – (KM)
- Normalized Cut algorithm – (NC)
- Pairwise Constrained K-Means (PC KM)
- Metric Pairwise Constrained K-Means (MPC KM)

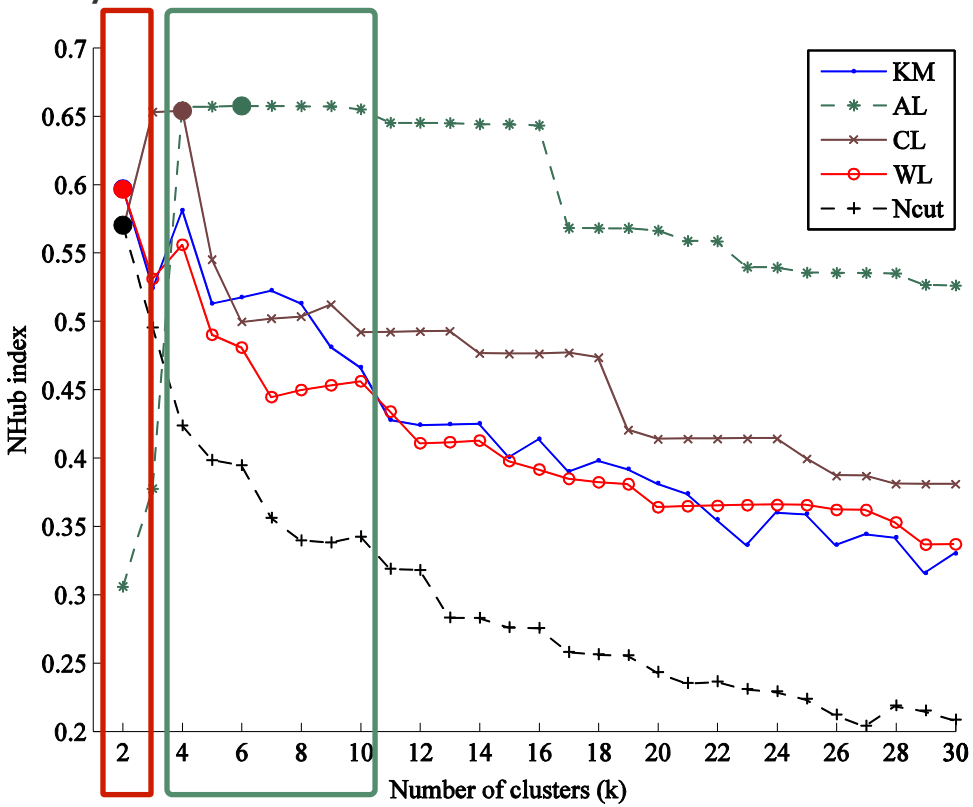
ELECTRICITY CONSUMPTION ANALYSIS & ENERGY EFFICIENCY

Case study I – *Electrical consumers characterization*

Clustering Process validation:

The choice of the clustering algorithm and the numbers of clusters were based on the performance of 8 validity indices

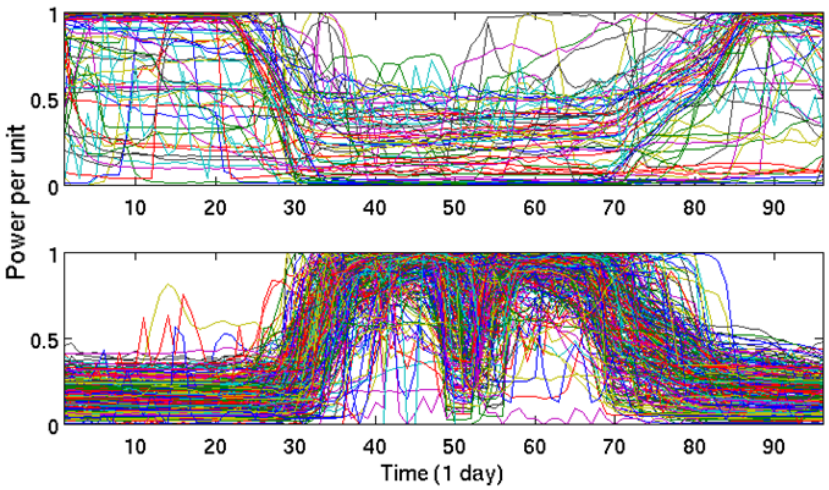
- Normalized Hubert Statistic – (NH)
- Dunn index – (D)
- Davies-Bouldin index – (DB)
- SD validity index – (SD)
- Silhouette statistic – (S)
- Index I – (I)
- XB cluster validity index – (XB)

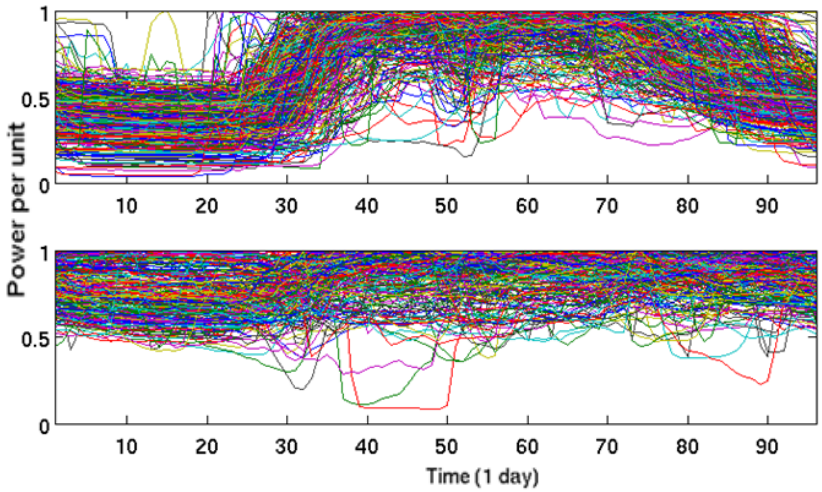


Case study I – *Electrical consumers characterization*

Cluster Validity Results:

Cluster Range		SD	PS	DB	XB	S	I	D	NH
2-20	Cluster algorithm	KM	NC	KM	**KM**	**KM**	**KM**	MPC	PC
	Number of clusters	4	2	2	**3**	**3**	**3**	18	3
4-17	Cluster algorithm	**KM**	**KM**	**KM**	PC	**KM**	**KM**	MPC	PC
	Number of clusters	**4**	**4**	**4**	5	**4**	**4**	15	5
5-14	Cluster algorithm	KM	MPC	**KM**	PC	KM	**KM**	PC	PC
	Number of clusters	8	6	**6**	5	7	**6**	12	5

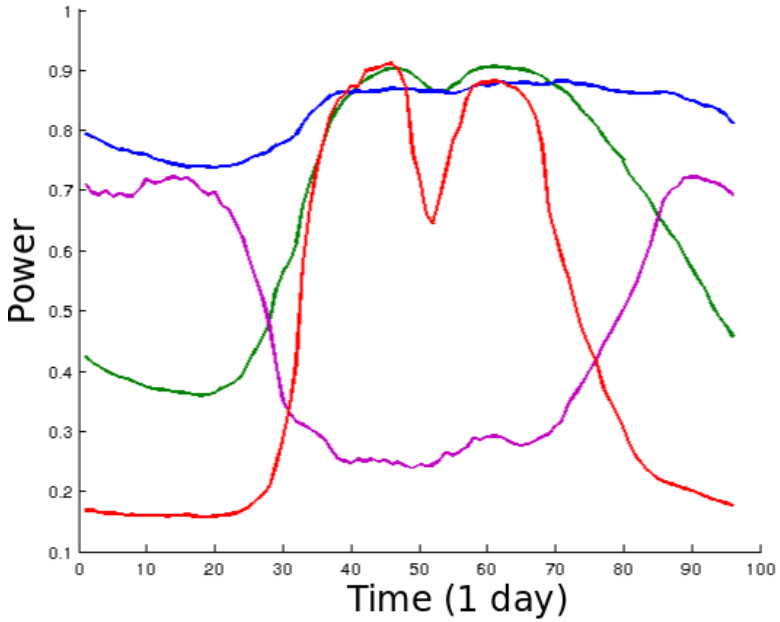




Case study I – *Electrical consumers characterization*

Clustering Results – Case I

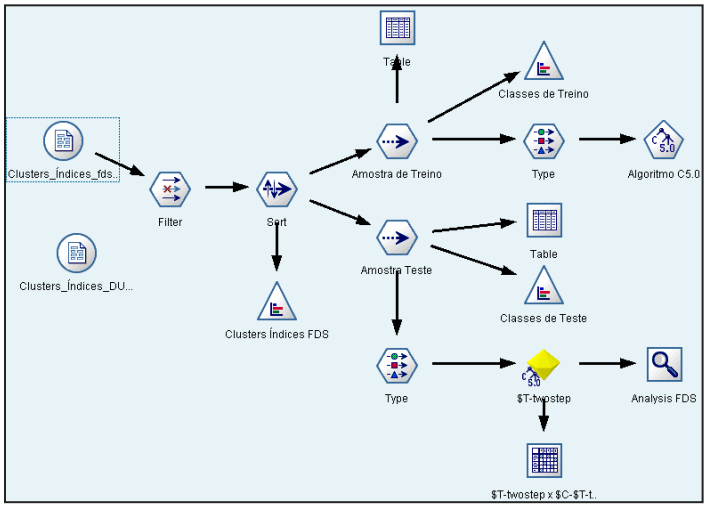
- K-means algorithm
- 4 clusters (k=4)
- Representative load diagram – working days



Case study I – *Electrical consumers characterization*

Consumers classification:

- To build a classification model, that applied to new unclassified records, will allow to foresee the class to which it belongs
- In the future it will allow to attribute to each new consumer the consumption profile that best represents it



Case study I – *Electrical consumers characterization*

- ✓ **C5.0 Algorithm**
  - **Decision Tree:**
  - Application simplicity
  - Result in tree form
  - Generation of rules

$$f1 = \frac{P_{av,day}}{P_{\max,day}} \qquad f2 = \frac{P_{\min,day}}{P_{\max,day}}$$

$$f3 = \frac{P_{\min,day}}{P_{av,day}} \qquad f4 = \frac{1}{3}\frac{P_{av,night}}{P_{av,day}}$$

$$f5 = \frac{1}{8}\frac{P_{av,lunch}}{P_{av,day}}$$

- ✓ **Derive from the daily load diagrams**
  - **Give information about:**
    - The daily load curve shape
    - The consumption pattern of each consumer

Case study I – *Electrical consumers characterization*

✓ Rule set for the week days classification model:

if $f_1 \leq 0.57$ and $f_3 \leq 0.21$	then cluster 7
if $f_1 \leq 0.57$ and $f_3 \leq 0.21$ and $f_1 > 0.24$ and $f_4 \leq 0.10$	then cluster 6
if $f_1 \leq 0.57$ and $f_3 \leq 0.21$ and $f_1 > 0.24$ and $f_4 > 0.10$	then cluster 4
if $f_1 \leq 0.57$ and $f_3 \leq 0.21$ and $f_1 > 0.44$ and $f_4 > 0.10$	then cluster 7
if $f_1 \leq 0.57$ and $f_3 > 0.21$ and $f_5 \leq 0.61$	then cluster 5
if $f_1 \leq 0.57$ and $f_3 > 0.21$ and $f_5 > 0.61$	then cluster 4
if $f_1 \leq 0.57$ and $f_1 > 0.45$ and $f_4 \leq 0.20$	then cluster 3
if $f_1 \leq 0.57$ and $f_1 > 0.45$ and $f_4 > 0.20$	then cluster 5
if $f_1 > 0.57$ and $f_1 \leq 0.71$ and $f_4 \leq 0.23$	then cluster 8
if $f_1 > 0.57$ and $f_1 \leq 0.71$ and $f_4 > 0.23$	then cluster 2
if $f_1 > 0.57$ and $f_4 \leq 0.21$	then cluster 2
if $f_1 > 0.57$ and $f_4 > 0.21$	then cluster 1

✓ Overall accuracy:

89%

Case study II – *Data Mining Contributions to Characterize Zonal Prices*

Locational Marginal Prices (LMP) database in 3.891 electrical buses from the California ISO (CAISO) have been used to identify economical zones

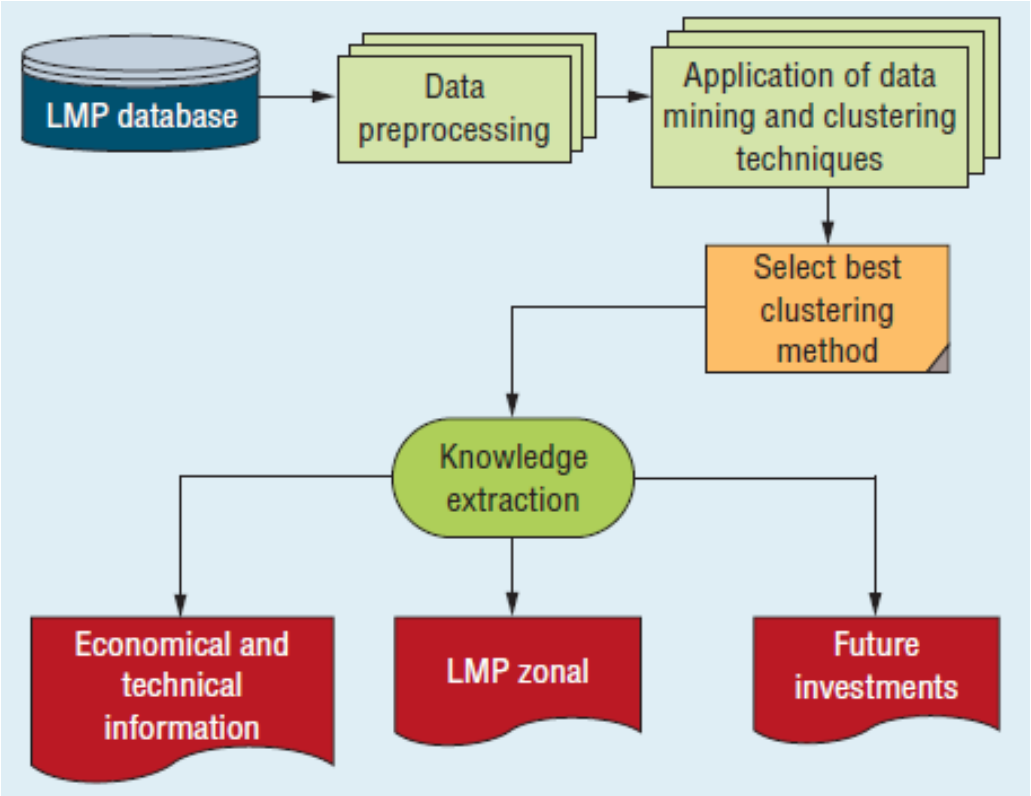
- LMP consist of three components as follows:

$$LMP_i = LMP^{energy} + LMP_i^{loss} + LMP_i^{cong}$$

- where
  - LMP_i - locational marginal price at bus *i* ($/MWh)
  - LMP^{energy} - marginal energy price of system ($/MWh)
  - LMP_i^{loss} - marginal loss price at bus *i* ($/MWh)
  - LMP_i^{cong} - marginal congestion price at bus *i* ($/MWh)

Case study II – *Data Mining Contributions to Characterize Zonal Prices*

Proposed methodology architecture



Based on the proposed data mining methodology, the knowledge extracted from the LMP historical databases allows the ISO to formulate relevant decisions concerning network planning and investment in electricity generation.

Case study II – *Data Mining Contributions to Characterize Zonal Prices*

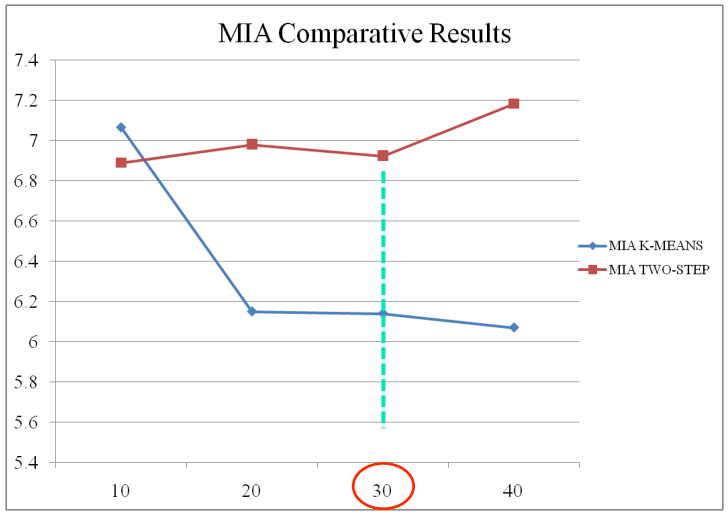
Clustering Process validation:

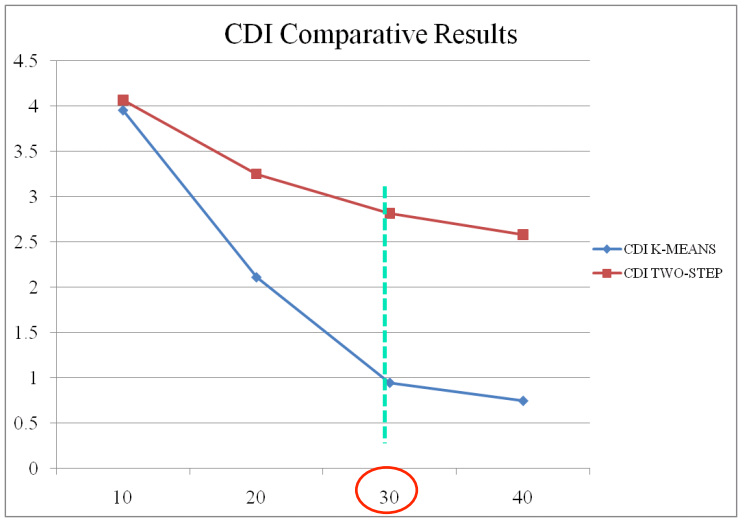
- MIA – Mean Index Adequacy

$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})}$$

- CDI - Clustering Dispersion Indicator

$$CDI = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{2.n^{(k)}}\sum_{n=1}^{n^{(k)}} d^2(l^{(m)}, C^{(k)})\right]}}{\sqrt{\frac{1}{2K}\sum_{k=1}^{K} d^2(r^{(k)}, R)}}$$





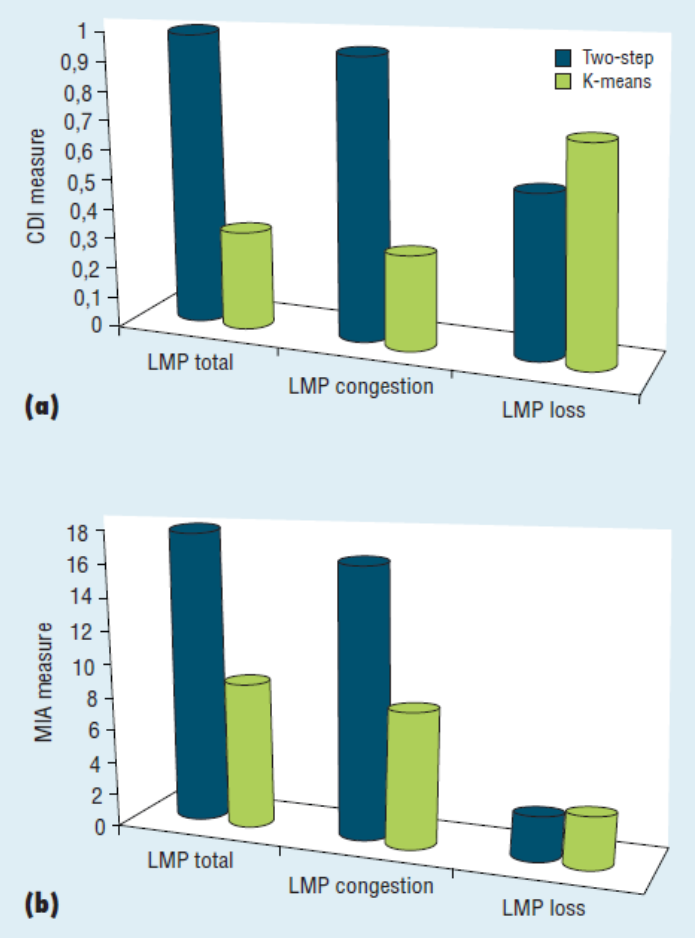
Case study II – *Data Mining Contributions to Characterize Zonal Prices*

Clustering performance comparison between:

- K-means algorithm
- Two-step algorithm

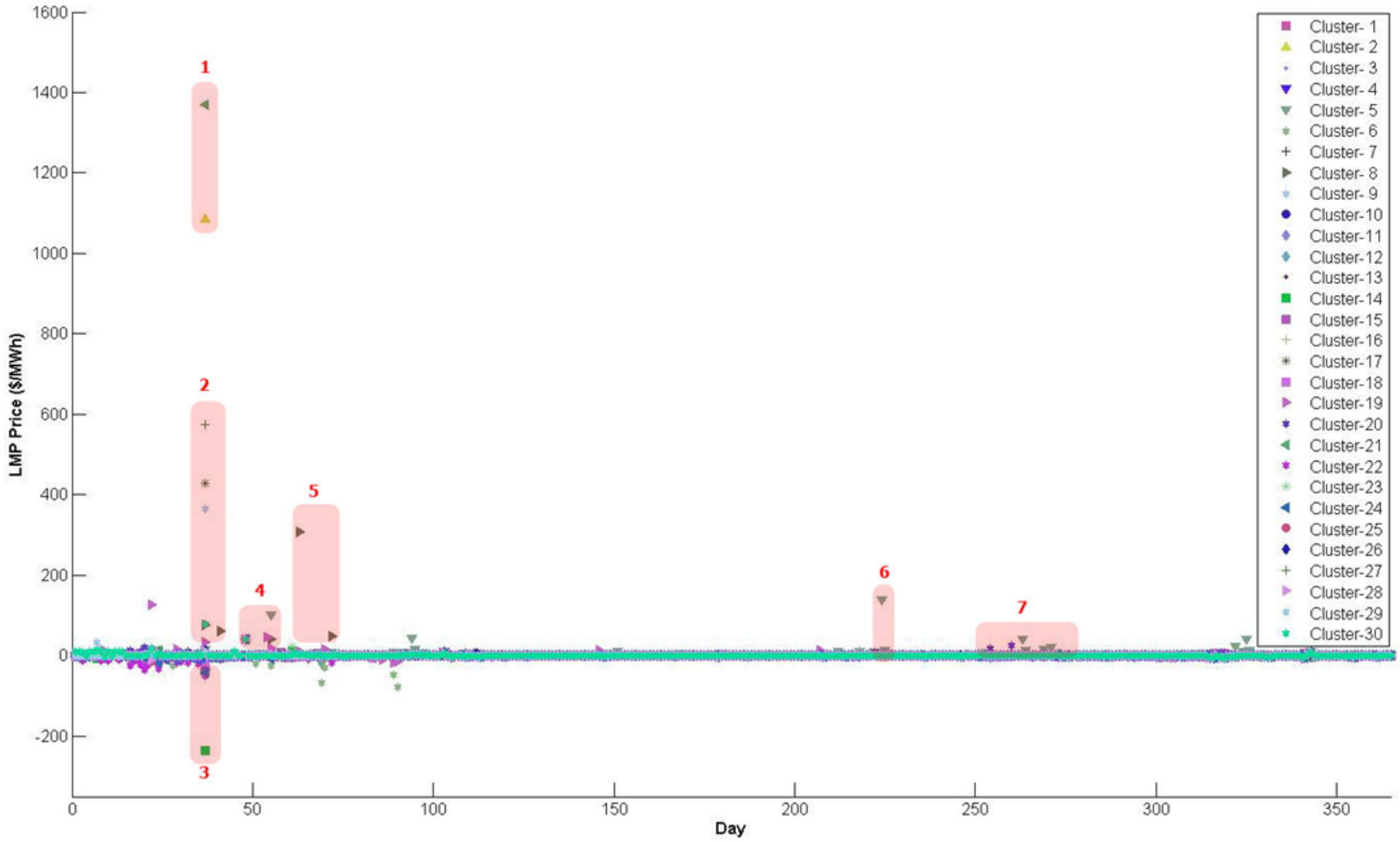
$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})}$$

$$CDI = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{2.n^{(k)}}\sum_{n=1}^{n^{(k)}} d^2(l^{(m)}, C^{(k)})\right]}}{\sqrt{\frac{1}{2K}\sum_{k=1}^{K} d^2(r^{(k)}, R)}}$$



Case study II – *Data Mining Contributions to Characterize Zonal Prices*

The figure reports the general tendency of the LMP-Cong value during the year for all clusters



Case study III – Data Mining Based Methodology for Wind Forecasting

A Data-mining-based methodology have been used to helps the wind forecasting

- Considering a real database – 3 years (2008, 2009, 2010)
  (National Wind Technology – National Renewable Energy Laboratory, EUA)
- Use of an artificial neuronal network (ANN-*MLP*)
- Evaluation of the Mean Absolute Error (MAE) / Mean Absolute Percentage Error (MAPE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - y_i\right|$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right|$$

Where:

x_i - actual value

y_i - prediction value

n - Number of forecasted periods

- Evaluation of the estimated accuracy value

Case study III – Data Mining Based Methodology for Wind Forecasting

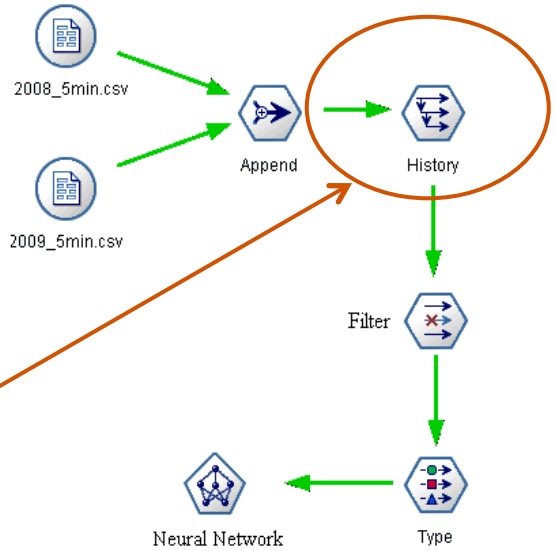
Starting from a historical database concerning 3 years:

- Wind Speed
- Wind direction
- Temperature
- Recorded cadence of 1 minute – almost 5 millions records to be analyzed

Simulation of the ANN including:

- Several input attributes configurations
- Specification of the **span** and **offset** parameters

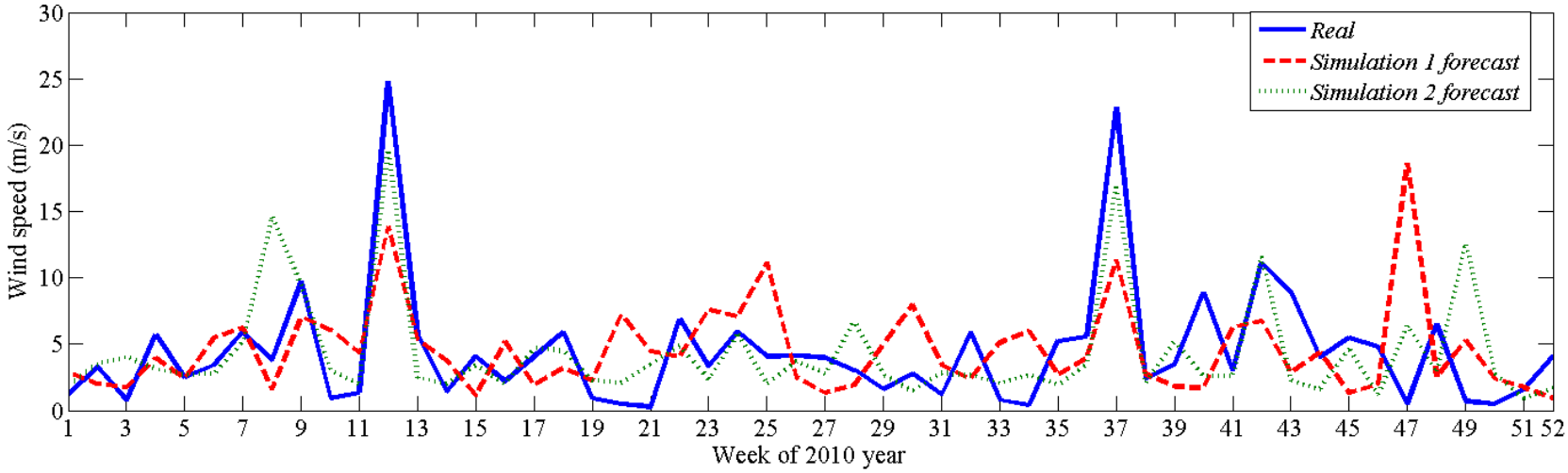
Where:
k – span parameter (5 min.)
n – offset parameter (5 min.)



Case study III – Data Mining Based Methodology for Wind Forecasting

Estimated accuracy of the ANN simulation model obtained:

Simulation	Description		Offset (*k*) (5 min.)	Span (*n*) (5 min.)	MAE	Estimated Accuracy
1	24 hours of the previous day	Wind speed	288	288	1,977	92,46%
2	10 hours of the previous day		288	120	2,066	92,39%
3	previous 24 hours		1	288	0,840	97,11%
4	previous 10 hours		1	120	0,823	97,17%
5	**previous 5 hours**		**1**	**60**	**0,818**	**97,61%**
6	previous 5 hours including wind direction		1	60	0,843	97,18%
7	previous 5 hours including wind direction and temperature		1	60	0,846	97,18%



First ELECON Workshop
Towards Efficient European and Brazilian Electricity Markets

Thank you…

ISEP, Porto, Portugal

25th September 2013