Aristotle
University
Thessaloniki

Dept. of
Mechanical
Engineering

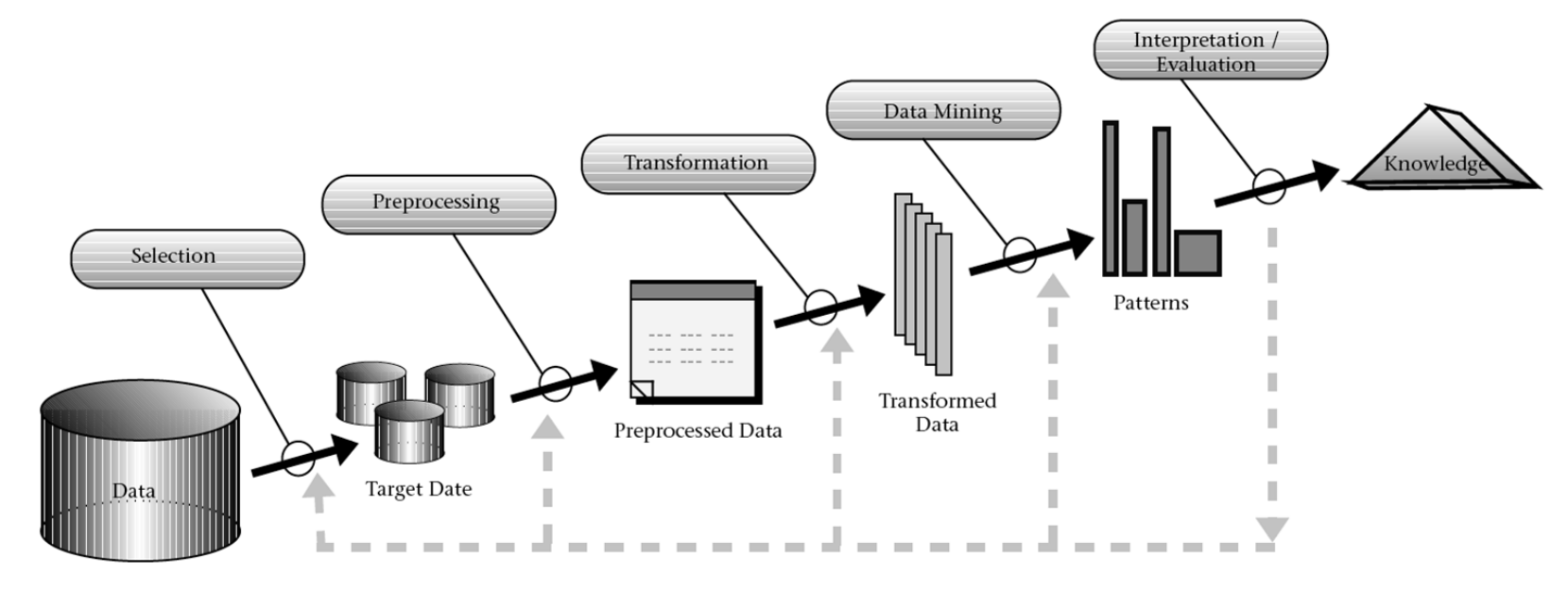
Preprocessing, analyzing and modelling of AQ measurement data

Kostas Karatzas

Informatics Systems and Applications – Environmental Informatics Research Group
Dept. of Mechanical Engineering, Aristotle University, Thessaloniki, Greece
Tel/Fax: +30 2310 994176 kkara@eng.auth.gr

Contents

- Preprocessing
- Analysis
- Modelling



Aristotle
University
Thessaloniki

Dept. of
Mechanical
Engineering

I. Preprocessing

Preprocessing

- The goal is to
  - Identify and remove errors
  - Prepare a reference data set, available for further analysis and modelling

Heterogeneity and quality of data

- Outliers identification
- Missing data handling

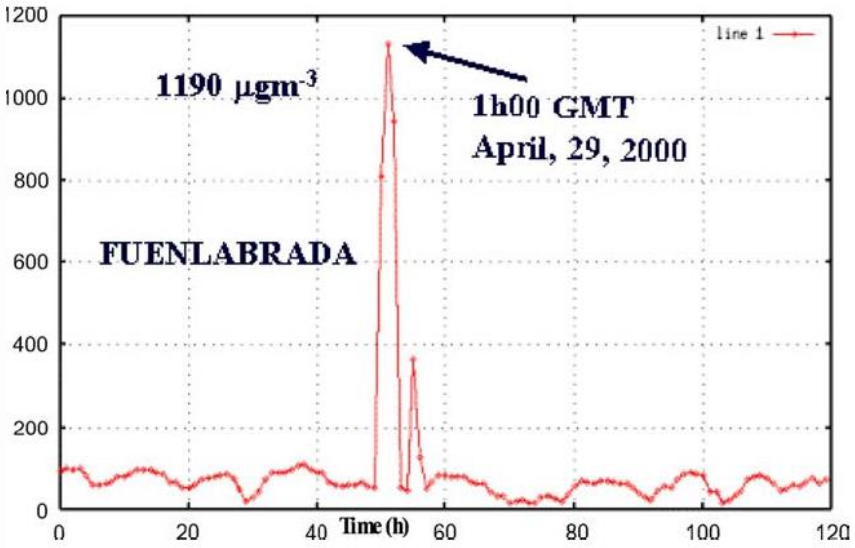
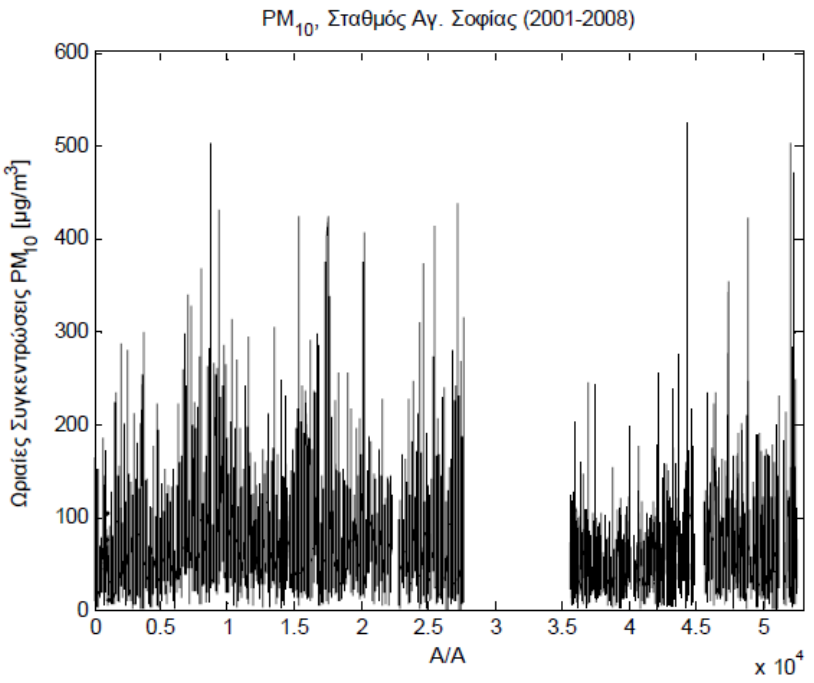
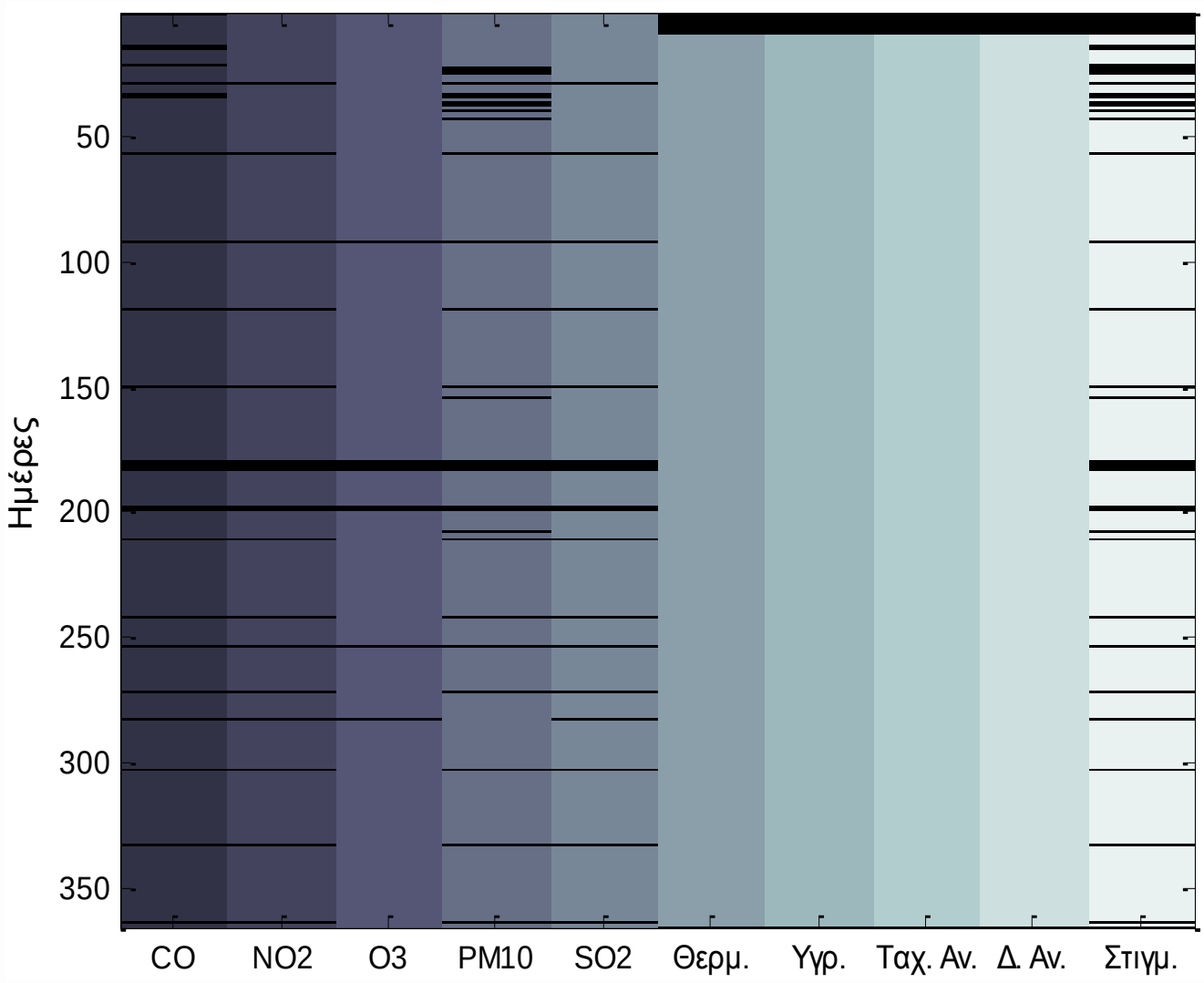


Fig. 3. Ozone concentrations in the Fuenlabrada monitoring station (Madrid community) during the ozone episode. The pattern shows data on April 27, 2000–May 1, 2000.

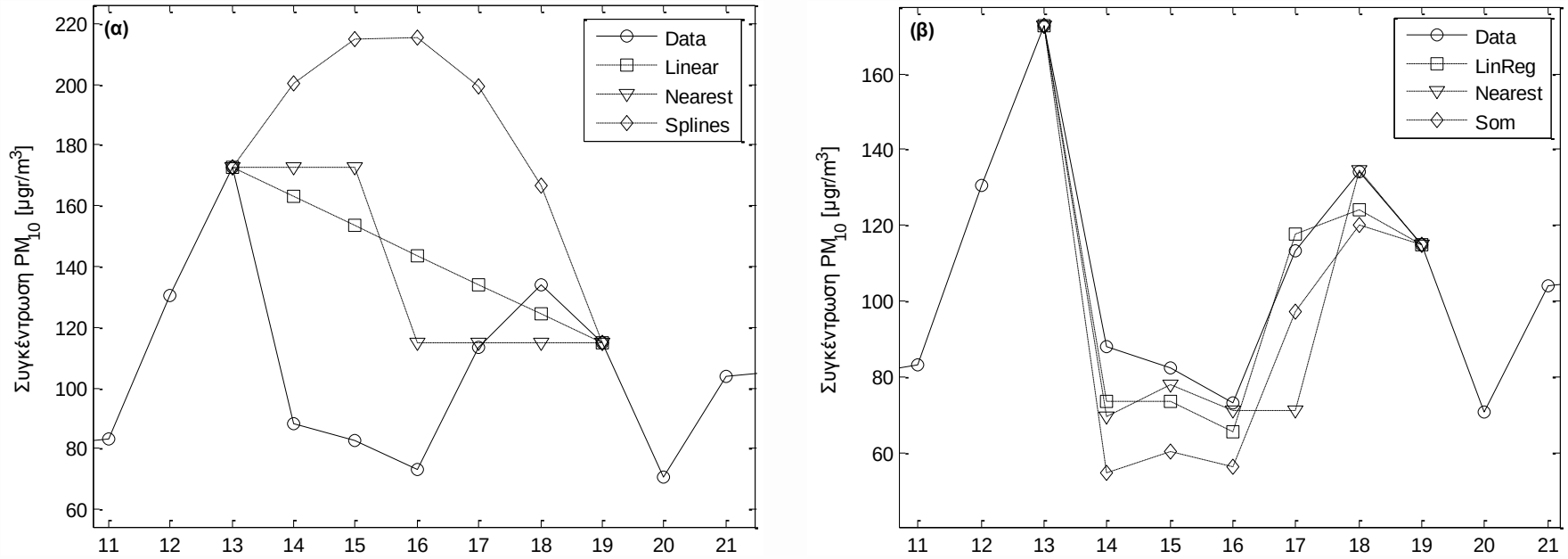


Missing values graphs



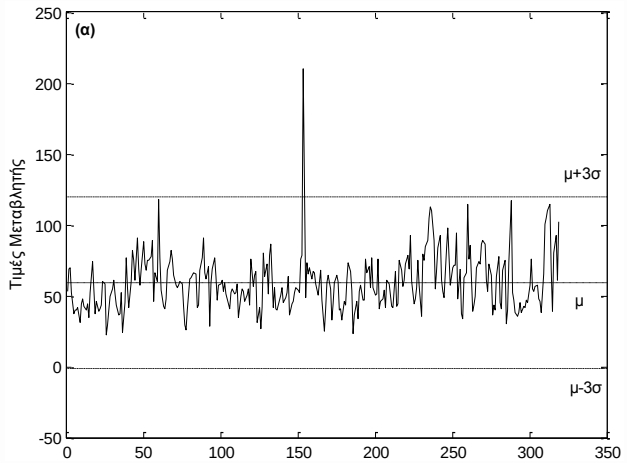
Missing value(s)

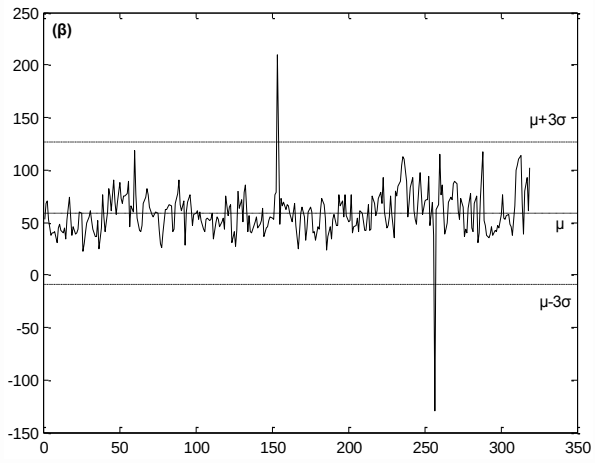
- Missing values handling
  - Removal and replacement of missing values
  - Calculation of missing value(s)



Missing value calculation examples
(a) Interpolation methods and (b) predictive modelling

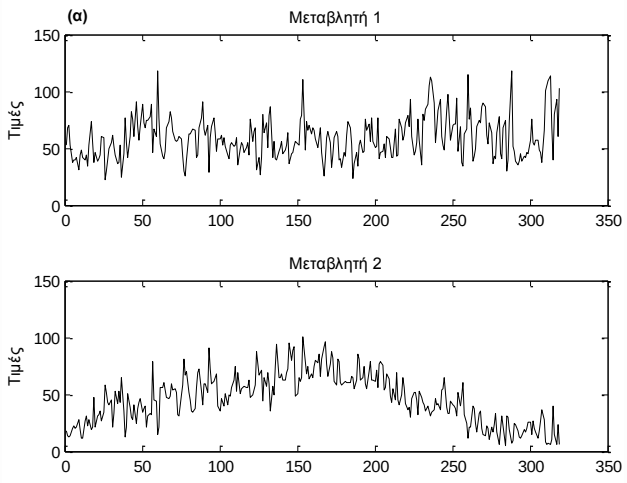
Graphs for identifying outliers

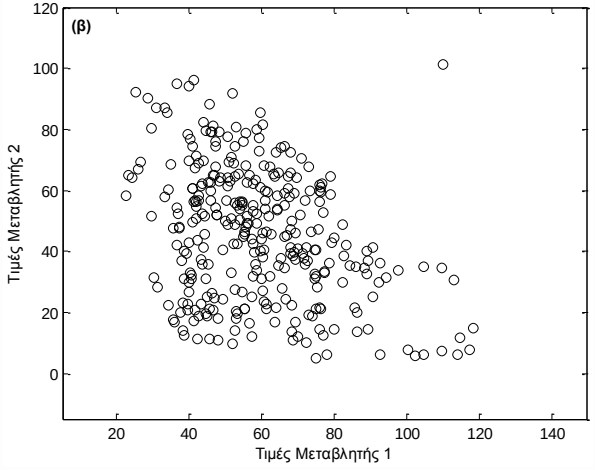




Mean value (μ),

μ-3σ and μ+3σ.





2-D outlier detection

Harmonized data overview

Descriptive statistics. The main goal is to calculate a basic set of statistical measures describing the measurements:

- Basic analysis & central tendency measures
  - Mean value, median and most frequent values
- Variation or dispersion measures
  - Standard deviation
- “Shape” measures
  - Skewness, Kurtosis

Descriptive statistics

Measure	Formulae	Comment
Mean Value	$\frac{1}{N}\sum_{i=1}^{N} x_i$	Sensitive to outliers
Median Value	$x_{N/2}$, αν N περιττός $(x_{N/2} + x_{N/2+1})/2$, αν N άρτιος όπου x οι τιμές της μεταβλητής ταξινομημένες	Not influenced by outliers
Trimmed Mean	$\frac{1}{N}\sum_{i=1}^{N} x_i$ Εξαιρώντας συγκεκριμένο ποσοστό των ακραίων τιμών	Useful for measurements following the normal distribution. Robust to outliers
Mode	The most frequent value among observations	Useful in the presence of outliers
Geometric Mean	$\left(\prod_{i=1}^{N} x_i\right)^{1/N}$	Useful when measurements follow logarithmic or asymmetric distributions. Sensitive to outliers
Harmonic Mean	$\frac{N}{\sum_{i=1}^{N} x_i}$	Useful when measurements follow logarithmic or asymmetric distributions. Sensitive to outliers

Descriptive statistics

Measure	Formulae	Comment
Standard Deviation	$\sigma = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{\frac{1}{2}}$ or $\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2}$	Useful for measurements following the normal distribution. Sensitive to outliers
Variance	$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ or $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$	Useful for measurements following the normal distribution. Sensitive to outliers
Mean Absolute Deviation	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \bar{x}\rvert$	Useful for measurements following the normal distribution. Less sensitive to outliers in comp to std
Interquantile Range	50% of the sorted values of x	Less representative if values follow the normal distribution. Robust to outliers
Range	$max(x_i) - min(x_i)$	Very sensitive to outliers
$\bar{x}$: arithmetic mean		

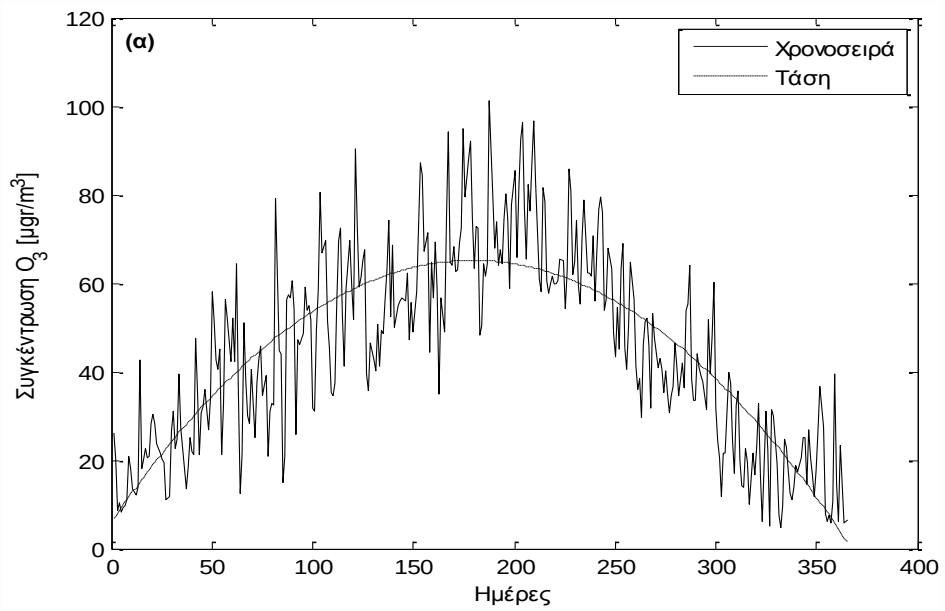
Visual inspection

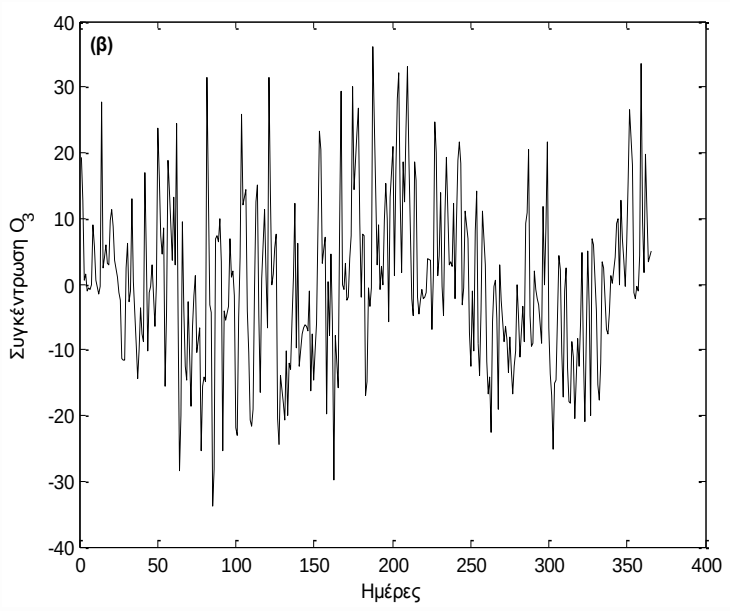
- Basic time series graphs
  - Per parameter
  - Per group of parameters (AQ and meteo groups)
  - One parameter, all institutes, versus reference measurements

  We can thus identify common behavior between sensors and use this in the next step to group sensors and proceed with further analysis
- Dispersion plots

Time series analysis

- De-trenting

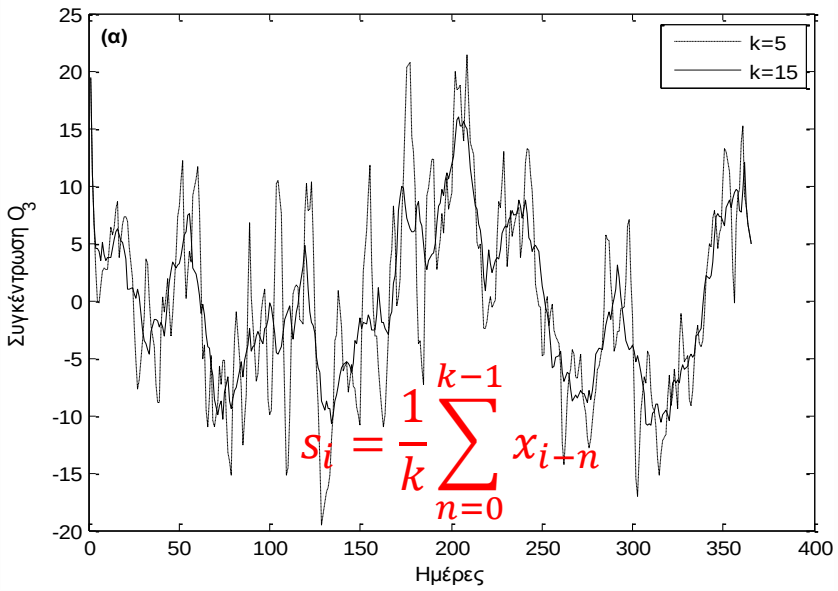


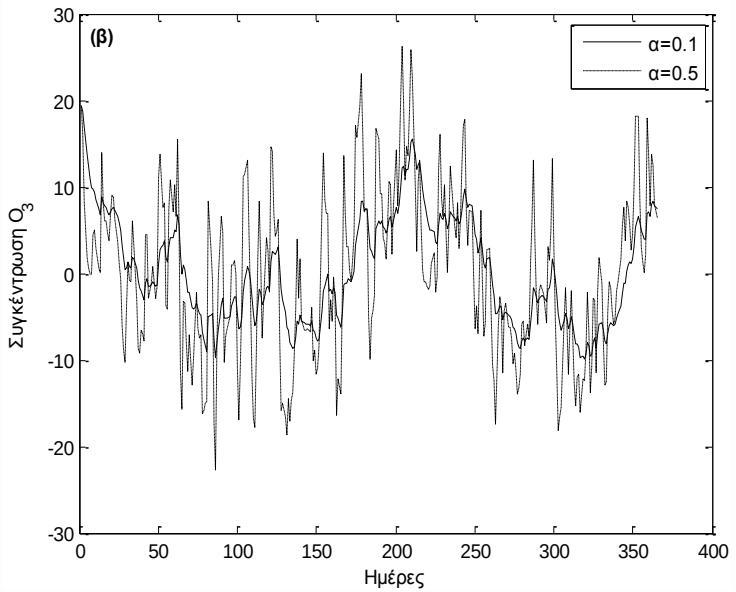


Identification and removal of trend (2nd degr. polyonym) from O3 time series: (α) before (β) after

Periodicity identification

- Identify and isolate periodicities





Smoothed O3 time series (after de-trenting) (α) running mean (k=5, k=15) and (β) exponential smoothing (α=0.1, α=0.5).

Normalization & useful transformations

- Variance normalization (all values between 0 and 1)

$$x' = \frac{x - \mu_x}{\sigma_x}$$

- Logarithmic (for big differences)

$$x' = \ln(x - x_{min} - 1)$$

- Trigonometric transformation (cyclic nature)

$$x' = 1 + tan\left(x + \frac{\pi}{4}\right)$$

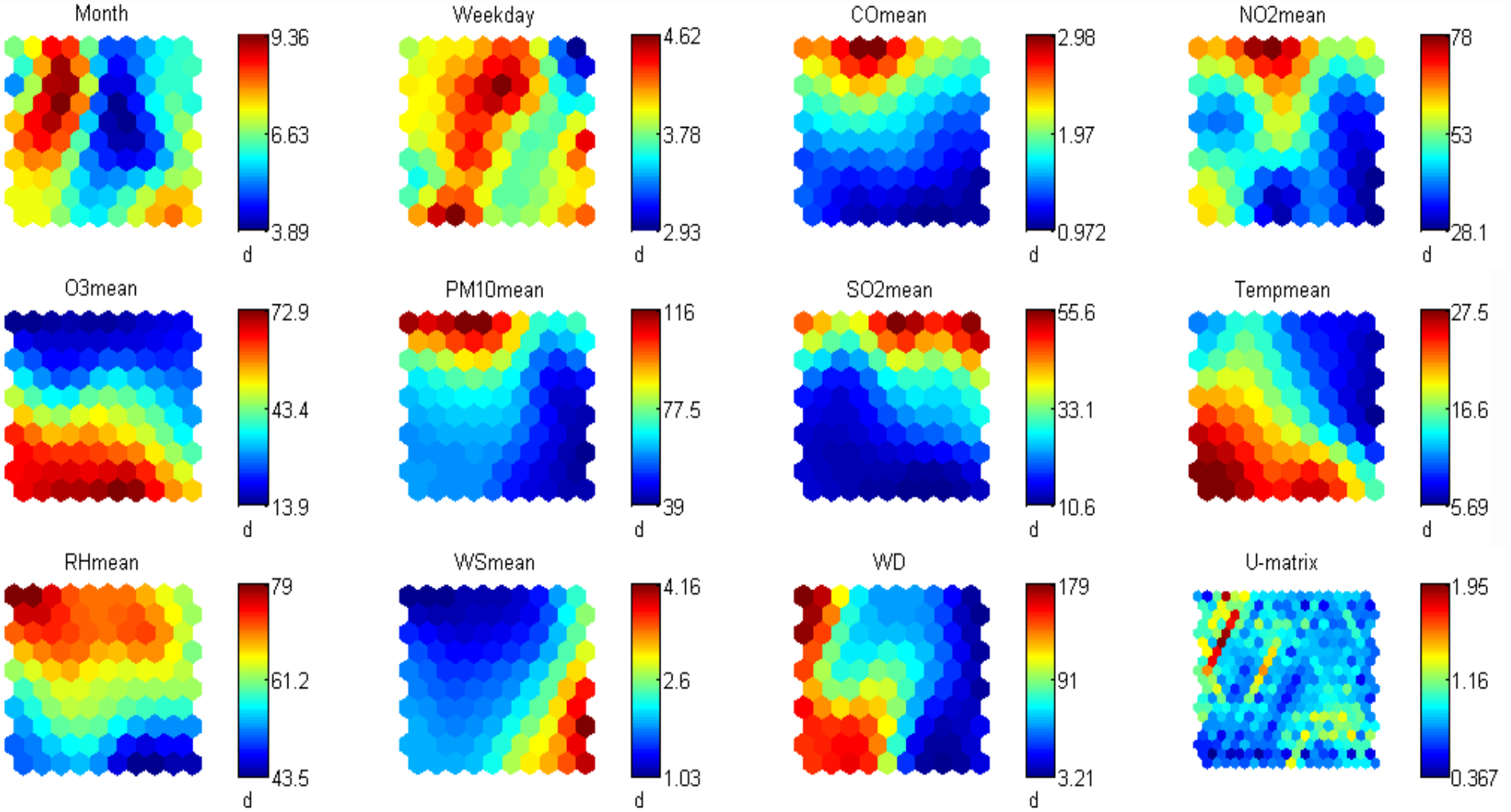
Aristotle
University
Thessaloniki

Dept. of
Mechanical
Engineering

II. Analysis

AQ data analysis

- The goal is to identify the “most” important parameters and their basic “relationships”
  - Covariance matrix
  - Correlation coefficient matrix
  - Information gain criterion
  - PCA
  - SOM
  - K-means clustering



Aristotle
University
Thessaloniki

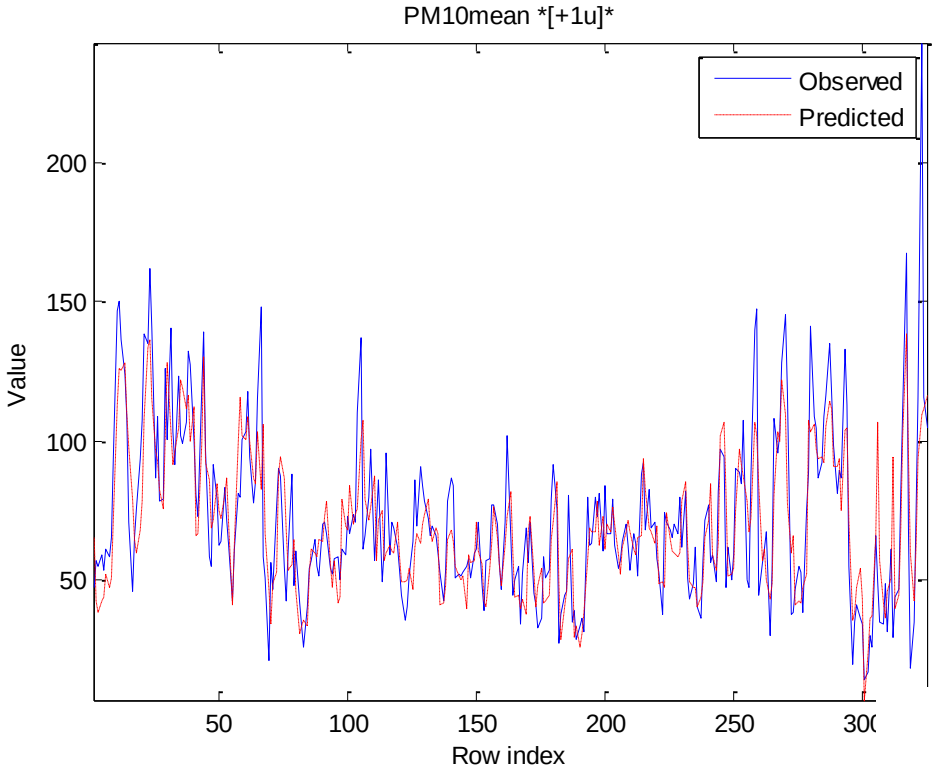
Dept. of
Mechanical
Engineering

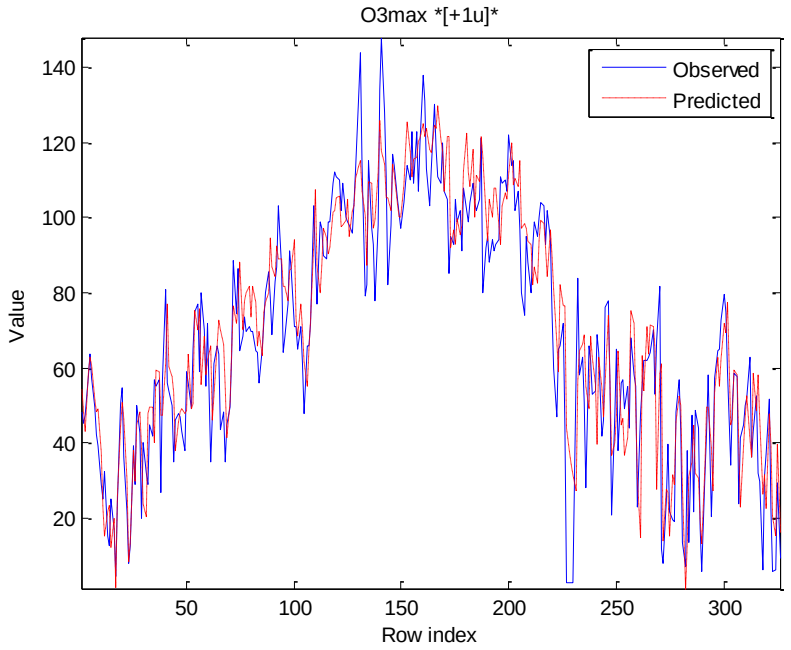
III. Modelling

Modelling

- Data – oriented modelling. The goal is
  - Behavior reproduction (descriptive modelling)
  - Forecasting (predictive modelling)

- Algorithms
  - Linear regression (just for reference)
  - Decision trees
  - ANNs
  - SVMs


PM10mean *[+1u]*
Observed
Predicted
Value
Row index
50
100
150
200
250
300


O3max *[+1u]*
Observed
Predicted
Value
Row index
20
40
60
80
100
120
140
50
100
150
200
250
300

And the final goal should be

- Development of new, innovative, personalized, georeferenced, quality of life related, everyday activity associated….
- Services!!!

The facts (1/2)

- We already have services providing information on:
  - Current air quality (monitoring stations)
  - AQ forecasts

The facts (2/2)

- What we would like to receive
  - Everything else! , i.e.
    - How will the quality of my life, in the place where I live will develop tomorrow
    - What do others like myself have reported and are expected to experience tomorrow
      - Biking like I do
      - Commuting like I do
      - Having everyday habits similar to mine
    - Which are the places and times of the day that I can use to move around and feel better if possible.

Collaboration is the only way

Consider joined design, informatics and environmental workshops with hands-on sessions for service designers

Thank you for your attention!