Aristotle
University
Thessaloniki

Dept. of
Mechanical
Engineering

Preprocessing, analyzing and modelling of AQ measurement data: the Aveiro intercomparison

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

Basic aims

- Identify errors
- Harmonize data
- Proceed with analysis
- Modelling
- Services

Aristotle
University
Thessaloniki

Dept. of
Mechanical
Engineering

I. Identify errors

Informatics Systems & Applications Group (ISAG) http://isag.meng.auth.gr

Errors

- Errors in time series measurements
  - Systematic
  - Random
  - Semantic

Errors

- Timestamp drift
  - Attributed to imprecisions in clocks
  - Proposal: correct on the basis of time series "events"
    - Identify increase-decrease patterns and use them as "reference points"
  - Proposal: make use of aggregated values
    - Instead of time series data (every 5 min), calculate and use statistical parameters per hour
      - Mean, Min, Max, Std, Skewness, Kurtosis, Range and Crest factor ($C = \frac{|x|_{\text{peak}}}{x_{\text{rms}}}$)

Errors

- Related to installation location
  - Humidity – condensation influencing instruments
  - Air flow influencing measurements
    - Installation location “asphyxiating” from other instruments
  - Solar radiation (exposure to direct sunlight due to installation)
  - Direct influence of traffic?
  - Any mechanical of electromagnetic vibrations in the area?

- Proposal: “group” and analyze measurements coming from neighboring instruments

Errors

- Related to
  - technology being used
  - sensitivity thresholds per type of technology
  - “Inertia” of measurement devices (that “refuse” to follow a change in the monitored parameter)
- Proposal:
  - “group” and analyze measurements coming from instruments using same-similar technologies
    - Check whether “same” instruments have been installed & used differently
  - Use different time windows for analysis to get rid of periodicities related to technology and not to the phenomenon monitored

Errors

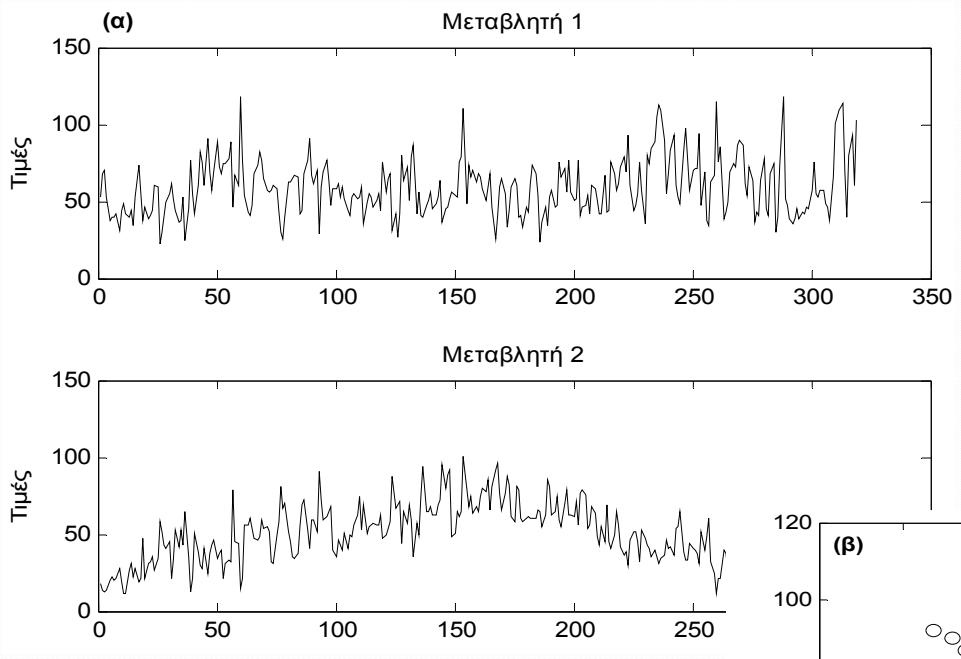
- Semantic:
  - Are we measuring what we think we measure?
    - Cross-sensitivity of gas sensors
      - How to identify and qualify this?
    - Are we correlated with what we think we are related?

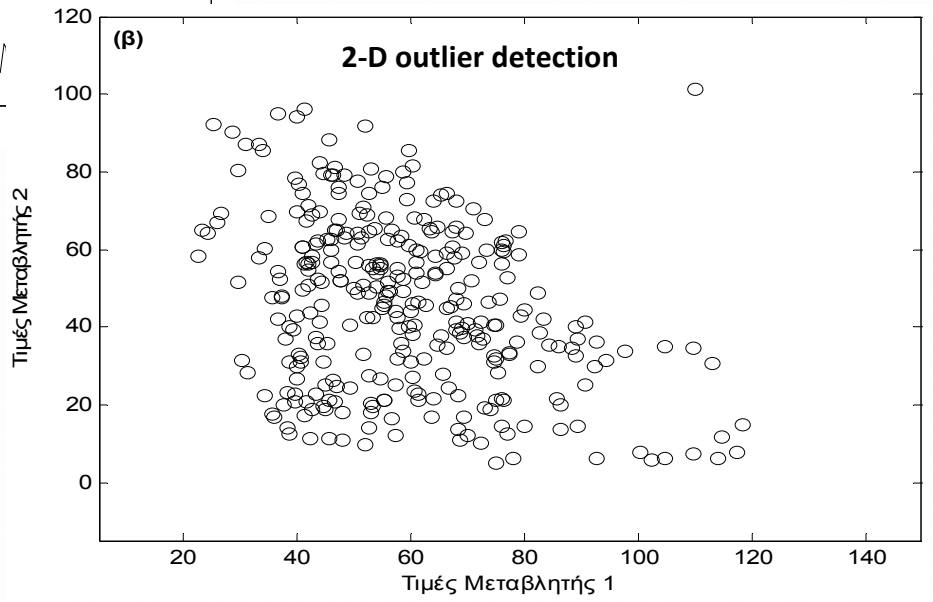
II. Harmonization

Harmonization

- Same parameters, same time stamp
- Same aggregated values
- Outliers: agree on a way to identify and treat
  - They might represent instrument sensitivity and should not be “banned”

Harmonization: identifying outliers





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

Harmonization: Visual inspection

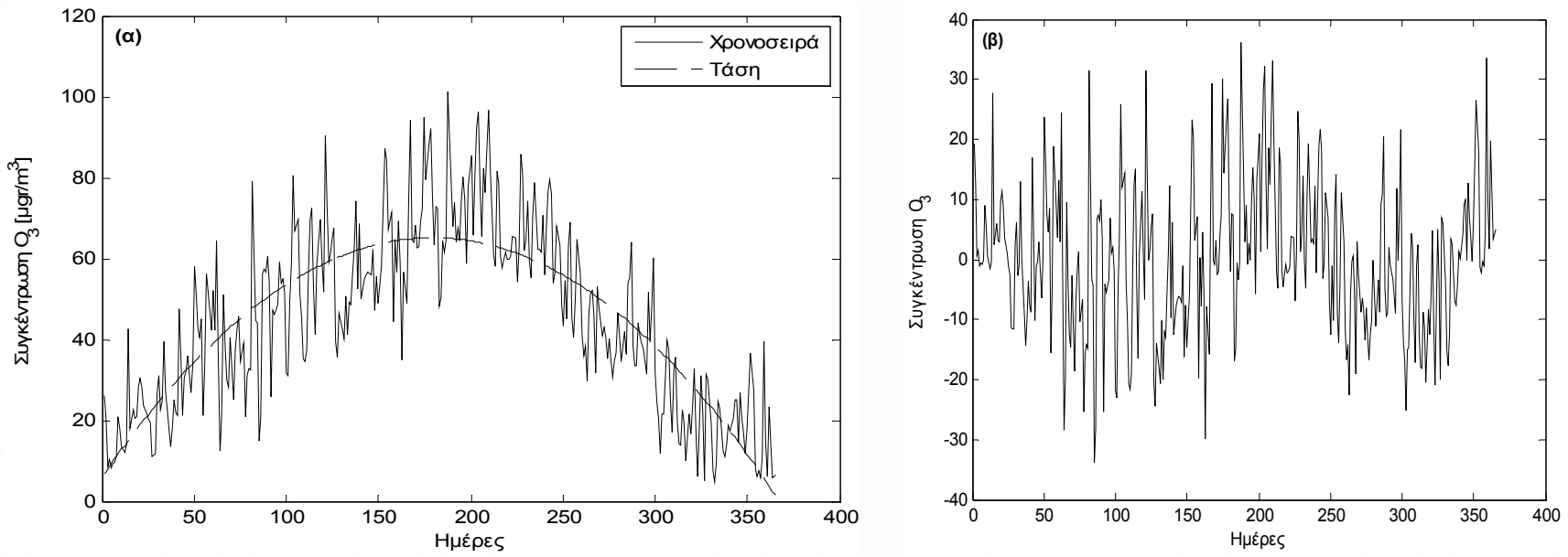
- Basic time series graphs
  - Per parameter
  - Per group of parameters (AQ and meteo groups)
  - One parameter, all institutes, versus reference measurements

  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

III. Analysis

Time series analysis

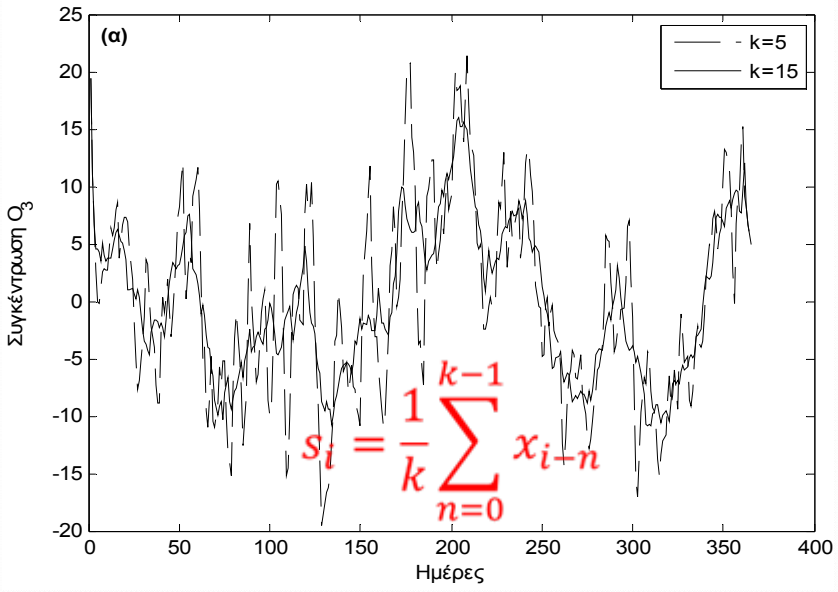
- De-trenting

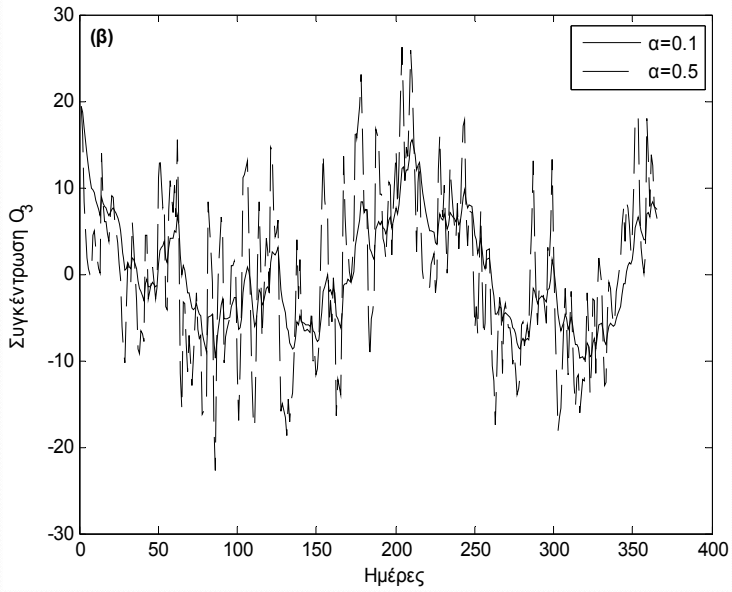


Identification and removal of trend (2nd degr. polyonym) from O3 time series: (α) before (β) after

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)

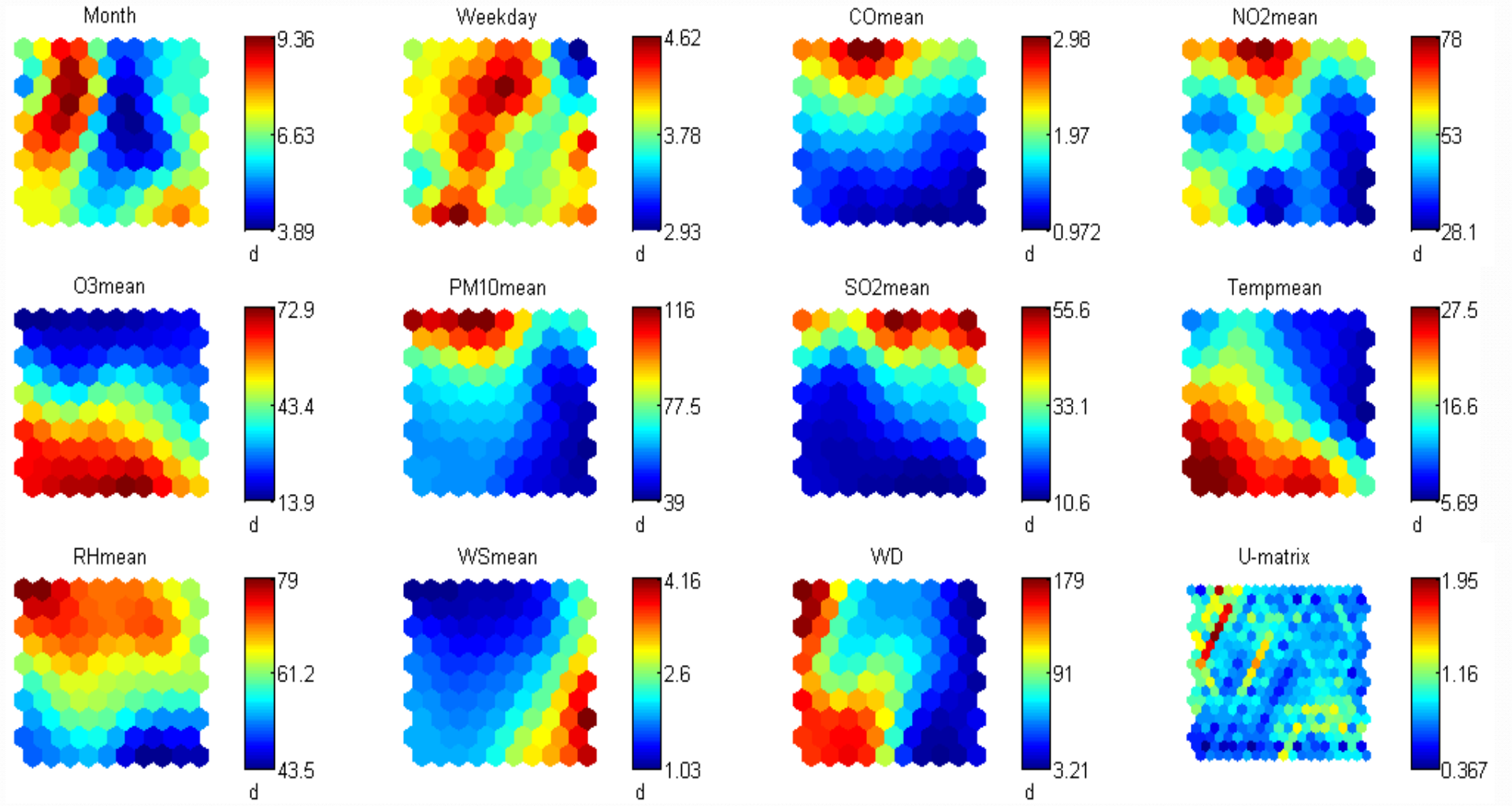
$$x' = \ln(x - x_{min} - 1)$$

- Trigonometric transformation (cyclic nature) for wind direction

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

AQ data analysis

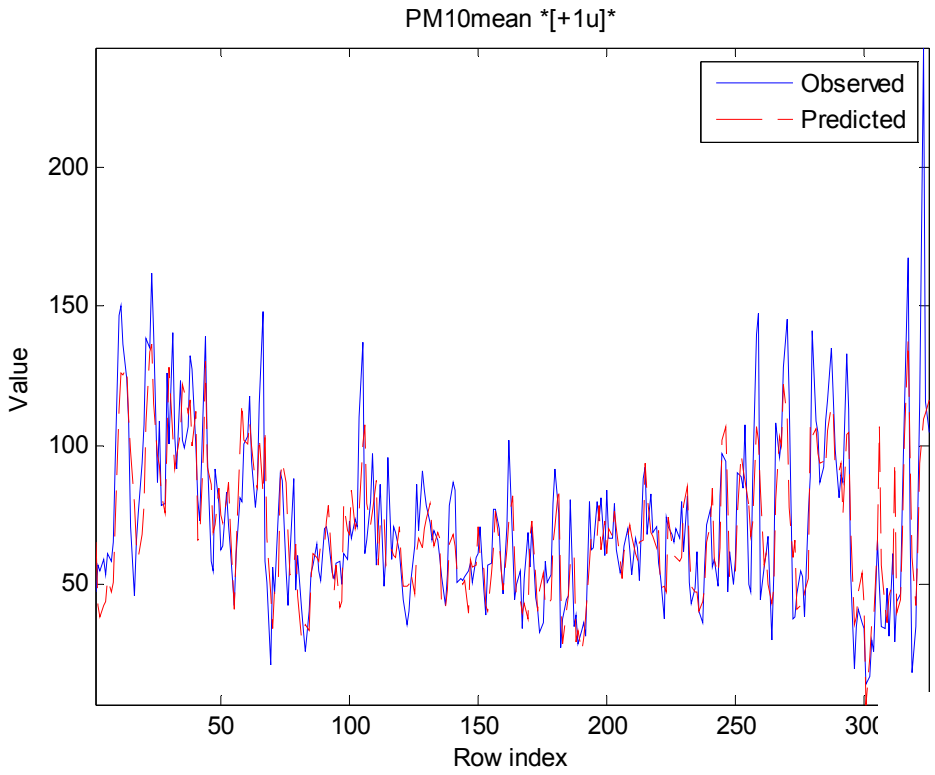
- The goal is to identify the “most” important parameters and their basic “relationships”
  - Covariance matrix
  - Correlation coefficient matrix
  - Information gain criterion
  - PCA
  - SOM
  - K-means clustering

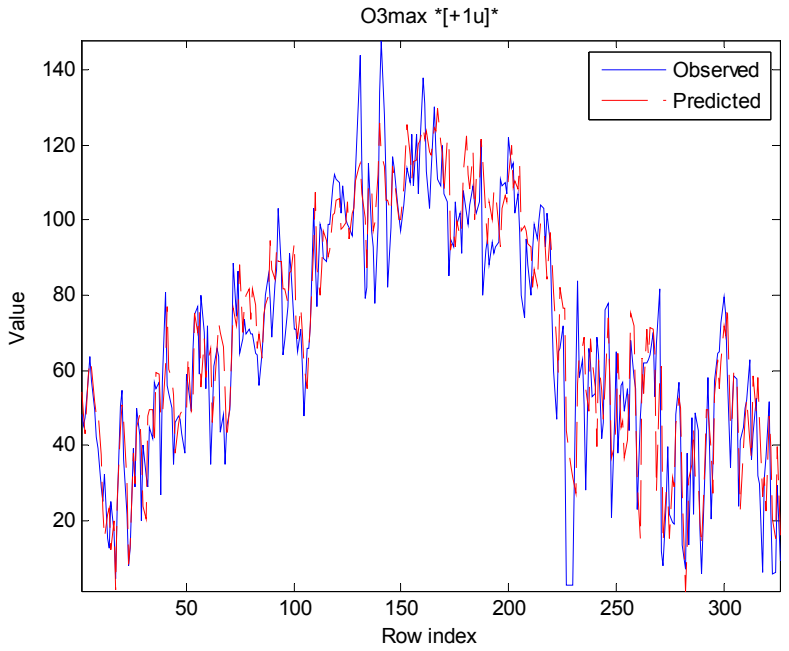

Month
9.36
6.63
3.89
d
Weekday
4.62
3.78
2.93
d
COmean
2.98
1.97
0.972
d
NO2mean
78
53
28.1
d
O3mean
72.9
43.4
13.9
d
PM10mean
116
77.5
39
d
SO2mean
55.6
33.1
10.6
d
Tempmean
27.5
16.6
5.69
d
RHmean
79
61.2
43.5
d
WSmean
4.16
2.6
1.03
d
WD
179
91
3.21
d
U-matrix
1.95
1.16
0.367

IV. 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





V. Services

Services

- What if the Aveiro dataset represents the (near) future of "*sensors everywhere*".
- How do we deal with quality issues?
- Can we demonstrate added value?

Thank you for your attention!