

# Pattern Selection Problems in Multivariate Time-Series using Equation Discovery

Arne Koopman, Arno Knobbe, Marving Meeng



Universiteit Leiden  
The Netherlands

# InfraWatch



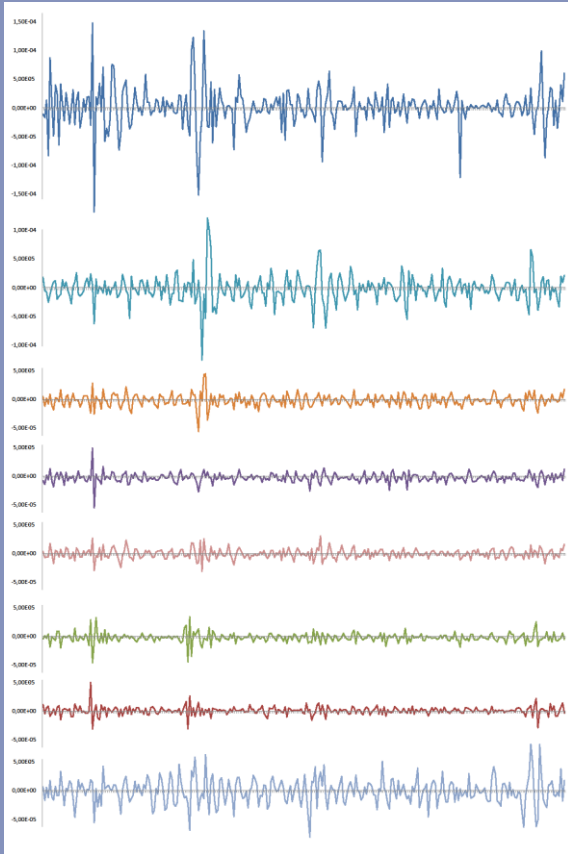
- Data Mining for Infrastructure Asset Management
- Hollandse Brug – a Dutch highway bridge
  - Monitoring of events (i.e. degradation, congestion)

# InfraWatch



- 145 sensors: continuous time-series data
- Various types: geophones, strain sensors, temperature sensors

# (Too) Much Sensor Data?

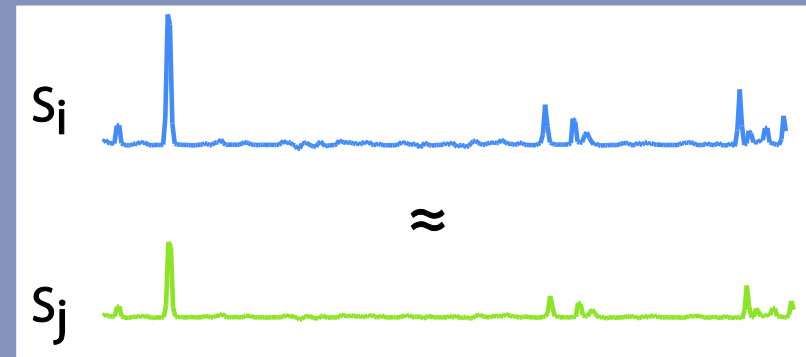


- 145 sensors : 145 continuous streams
- Sampling at 100 Hz : ~4GB /day
- Is all of this useful?
- Or... can we select a few sensors that provide a good view on the whole system?



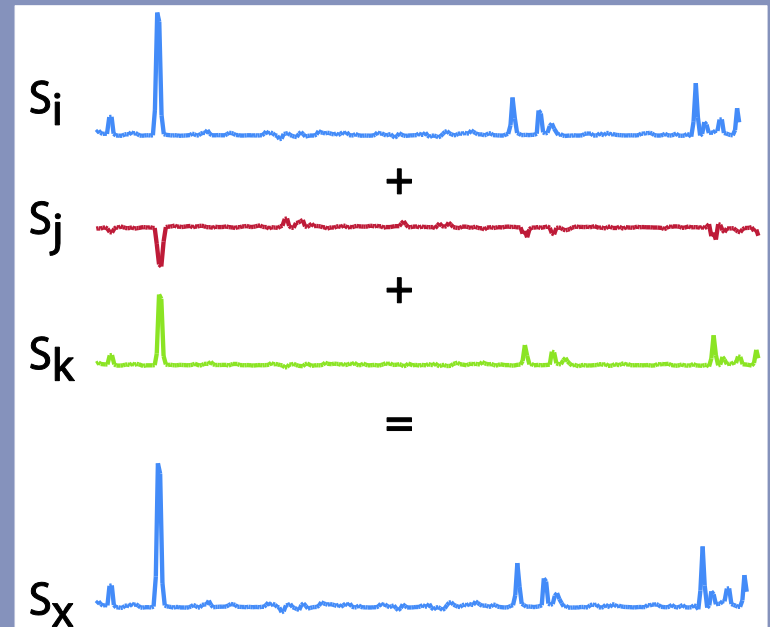
# Relevant Sensors

- Idea: sensors that have similar sensor readings are assumed redundant
- Select a set of non-redundant sensors that provide a overall picture of the complete system



# Sets of Sensors = Equation

- Sensor  $x$  is described by a sensor set
- Select a set of sensors that have events that coincide : they describe the same events



$$s_x = c_0 + \sum_{s_i \in S} c_i \cdot s_i$$

# Equations

- LaGrange's grammar defines an equation type,
- such as linear:

$$f_x(t) = c_0 + \sum_{s_y \in S} c_y \cdot s_y(t)$$

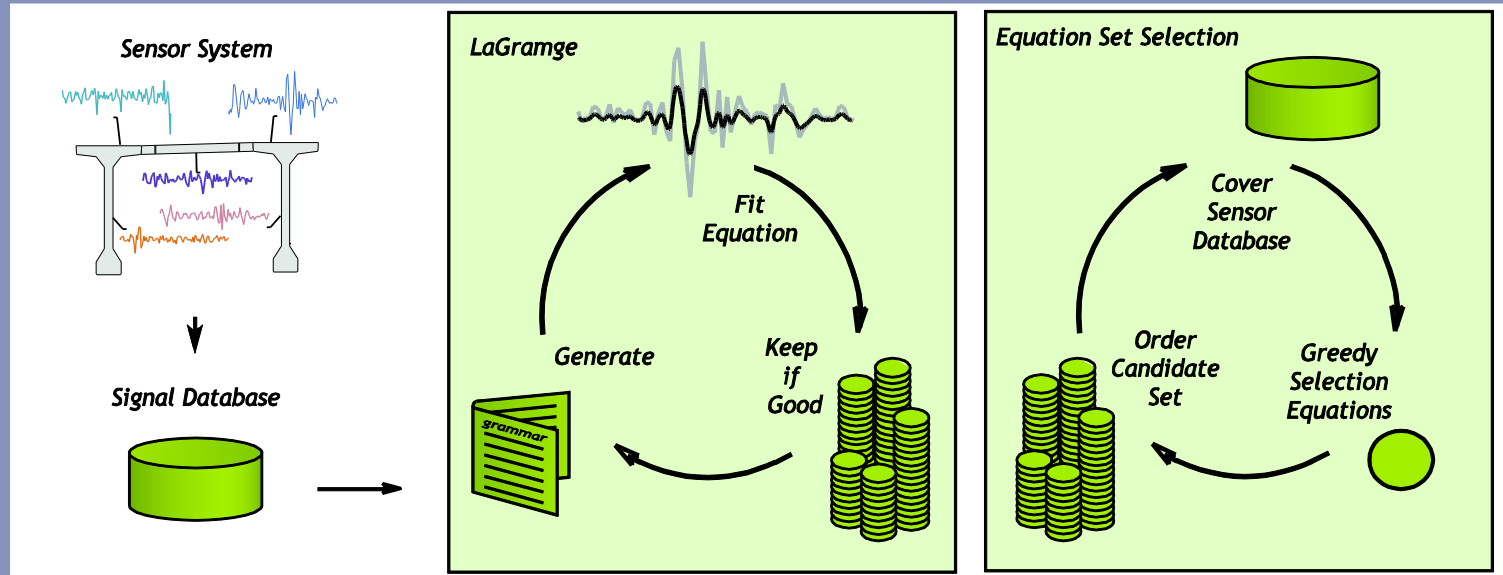
- ..or differential:

$$f_x(t) = c_0 + \sum_{s_y \in S} c_y \cdot \delta s_y(t) / \delta t$$

- ... or, can use expert knowledge to define known relations between signals



# Which Equations?



## LaGrange

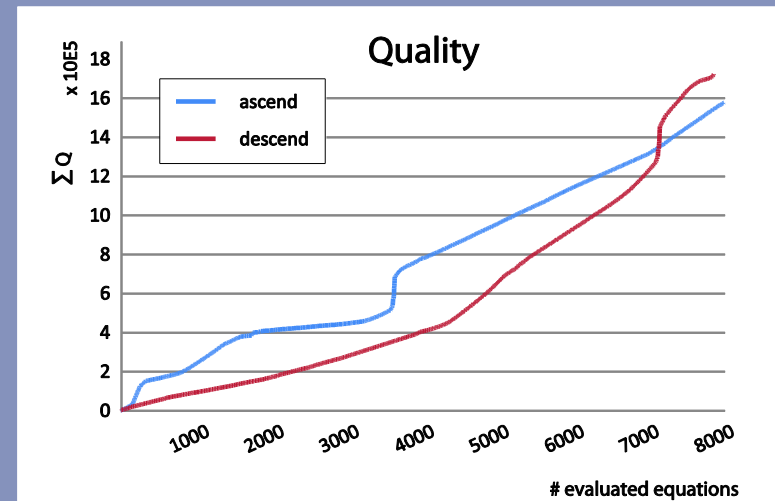
- Fit candidate equations to the data
- Pick all equations that fit the signal well

## Selection

- Pick equation set that models the system well

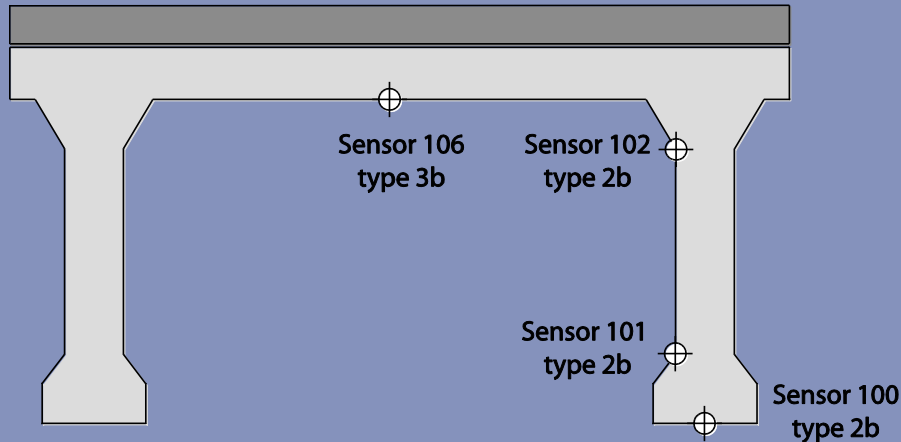
# Selecting on Quality

- Signals in similar range, therefore: do not boost signals too much
- Select equations with scalars  $c$  close to 1
- 2 greedy search strategies: ascending and descending size ordered candidate equations



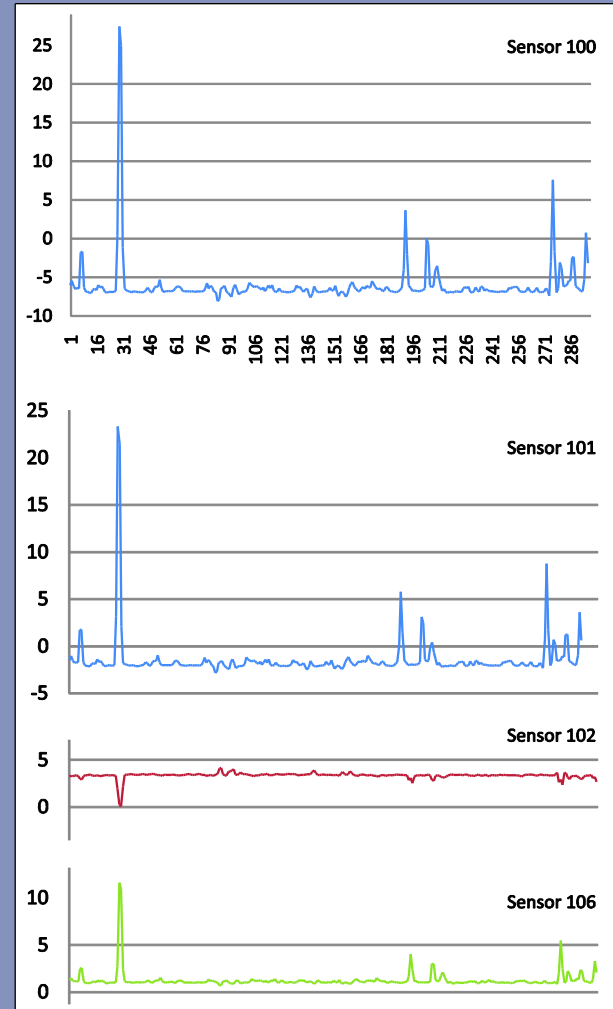
$$f = c_0 + \sum_{s_i \in S} c_i \cdot s_i$$
$$Q(f) = \frac{1}{\sum_{c \in f} |\log |c||}$$

# Compact Equations - Example



$$s_{100} = 1.12 \cdot s_{101} + 0.3 \cdot s_{102} + 0.2 \cdot s_{106}$$

- Sensor 100 is explained by sensors that are close by, and have signals that are correlated



# Final Remarks

## Equation Discovery

- LaGrange suitable to bridge the gap between continuous data and pattern discovery
- Equation sets can be used as a compact description of a continuous system

## InfraWatch

- Visit our website: [www.infrawatch.com](http://www.infrawatch.com)

