Usage of the Kalman filter for Data Cleaning of Sensor Data

Klemen Kenda, Jasna Škrbec, Maja Škrjanc

Jozef Stefan Institute, Artificial Intelligence Lab

Data Cleaning

Detection and correction of the data which is:

- *Corrupt*
- *Inaccurate*
- *Incorrect*
- Incomplete
- Irrelevant
- Duplicated
- Missing

Data cleaning does:

- *Data transformation*
- Elimination of duplicates
- Detection of missing data
- *Error correction*
- Detection of lost information

Architecture

Application Tier

Data Consolidation/Integration Tier

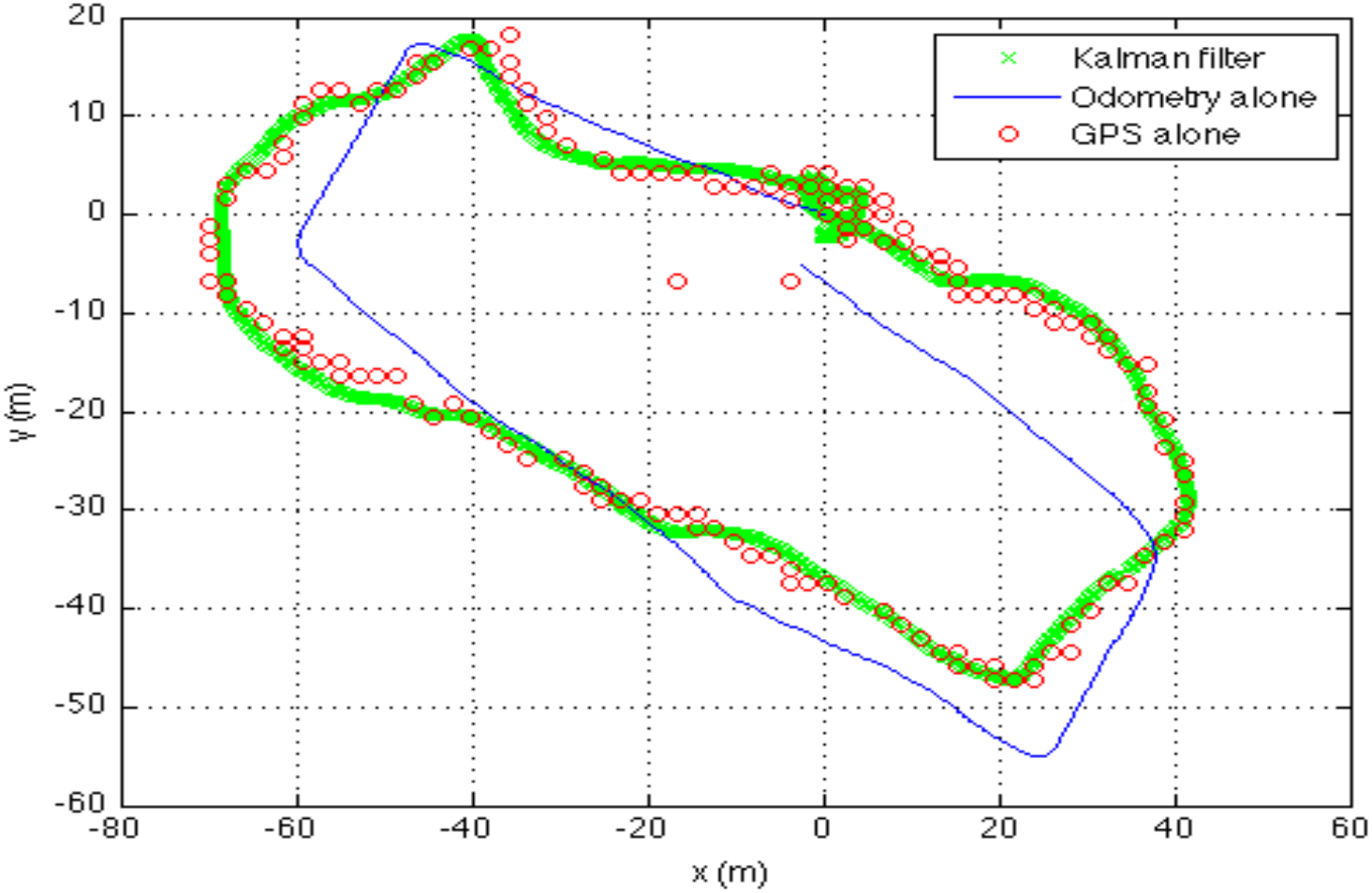
Data Access Tier

Resource Tier

The Kalman Filter (1/2)

- Well known algorythm from 1960
- Rudolf E. Kalman (1930 -)
- Different types: linear, non-linear (EKF) ...
- Usage:
  - Navigation (GPS navigation, electronic compass)
  - Depth measurments
  - Fitting Bezier to noisy/moving point data
  - Tracking objects (missiles, faces, heads, hands)
  - Computer vision (feature tracking/cluster tracking, fusing data from radar, lidar, stereo-cameras)

The Kalman Filter (1/2)

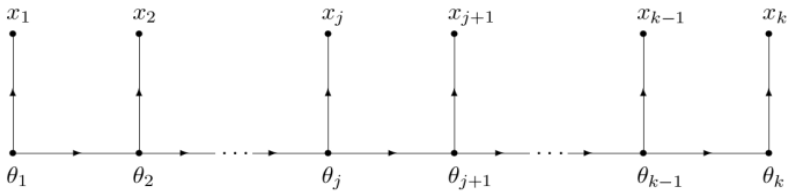


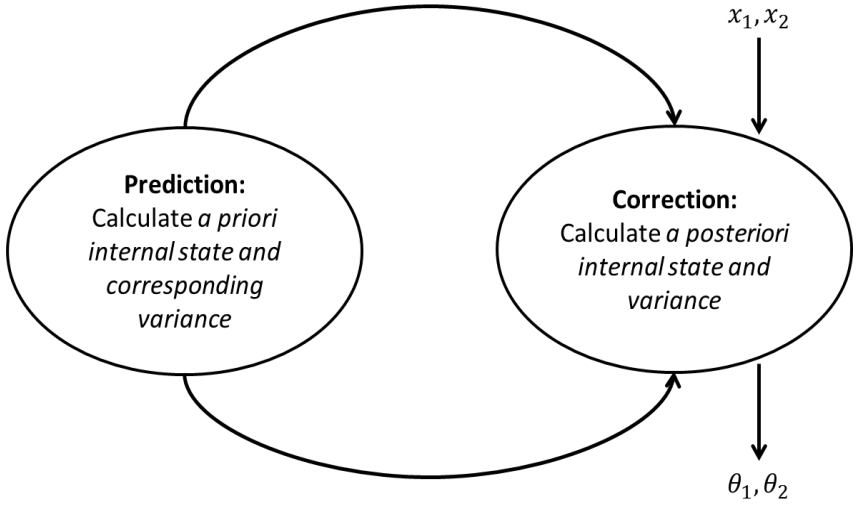
The Kalman Filter (2/2)

- LSE comparison
- Gauss-Markov process
- On-line
- Prediction phase

$$\theta_{k+1}^{-} = \Phi_{\mathrm{k}}\theta_k$$

- Correction phase
- Filter is adaptive





Kalman on Sensor Data

- Properties of Sensor Data
  - On-line
  - High frequency
  - Continuous properties, smooth changes
  - Vague or too complex models
- Second-degree model

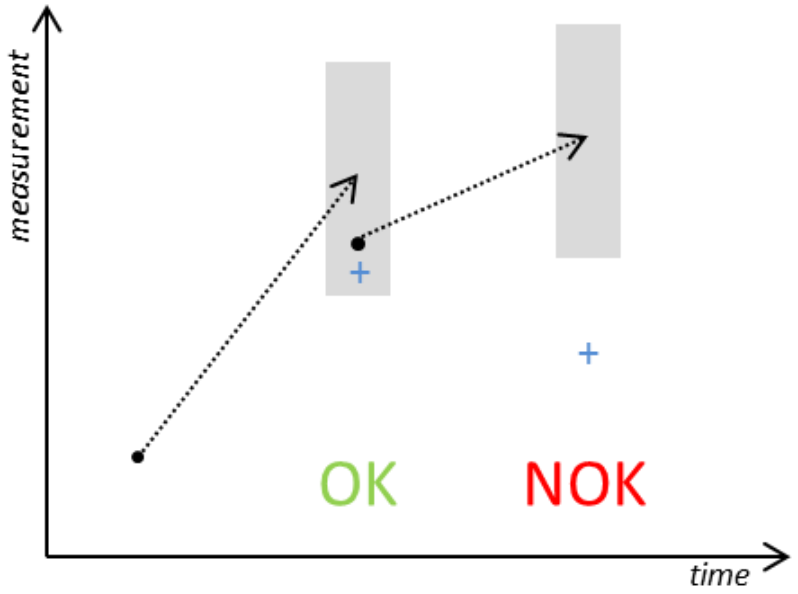
$$\bar{\theta}_{k+1} = \Phi_k \theta_k$$

$$\theta_k = (A, dA/dt, d^2A/dt^2)$$

$$\Phi_k = \begin{pmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{pmatrix}$$

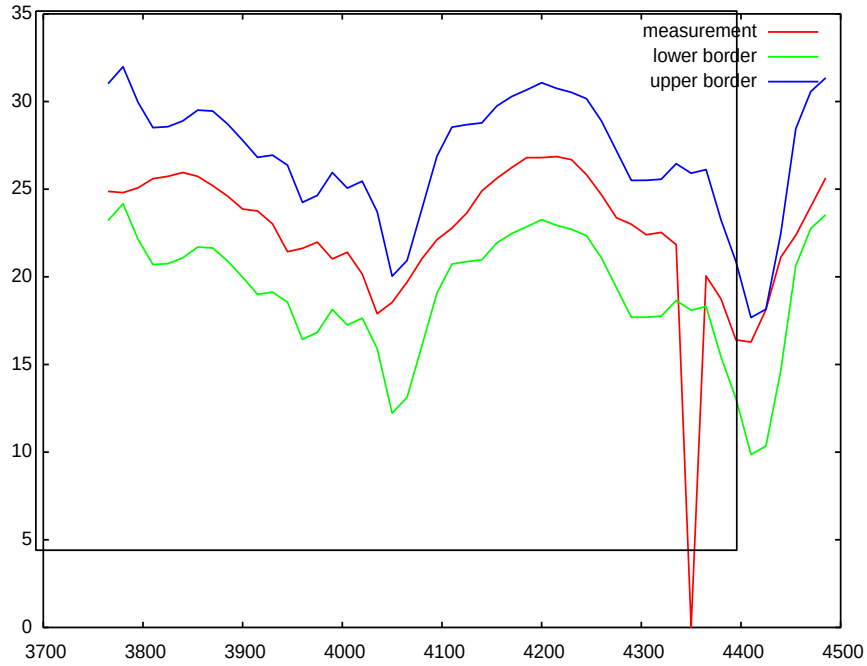
Outlier Detection

- Using the model nature of the prediction phase
- Determining a gap (can be in units of variance)
- Semi-supervised method
- Outlier or missing value replaced with prediction value

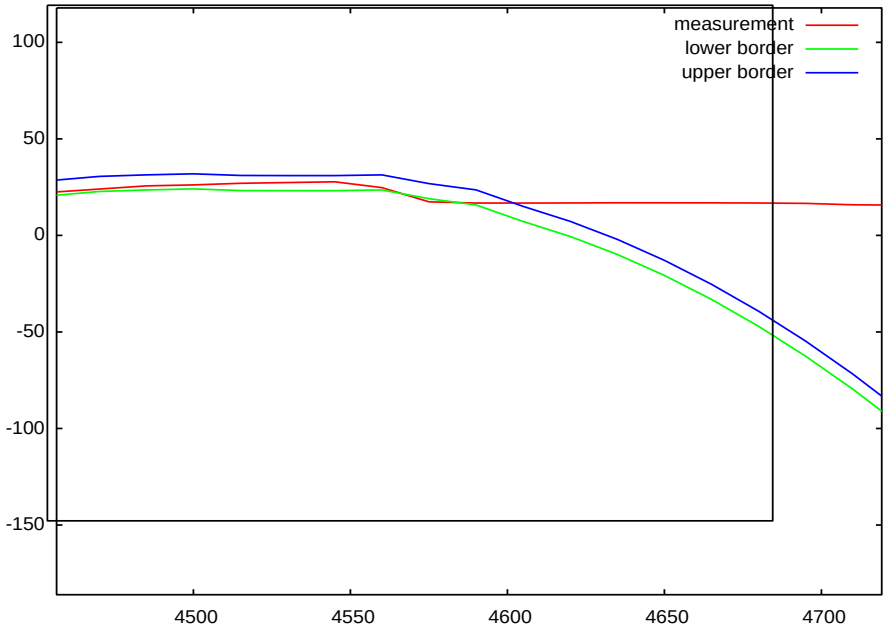


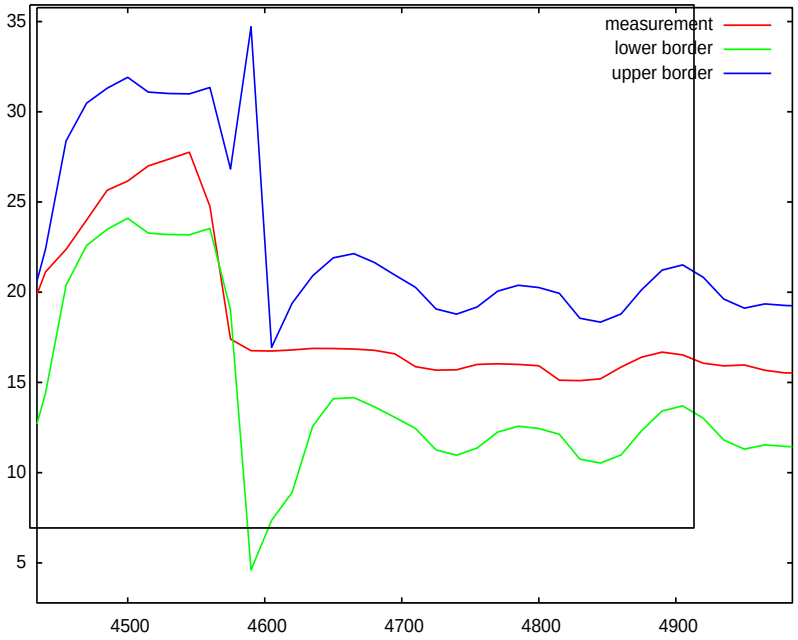
Results

- Temperature data from the NRG4Cast project (July/August 2013)
- Gap determined at 5σ
- Choice of model parameters (!)



Instability of the algorythm





Conclusions and Future Work

- Methodology for Data Cleaning
- Architecture
- Tested basic linear Kalman models on Environmental dataset
- Suggestion on solving the instability problem
- Complex initialization of the filter (gradient descent and similar methods)
- Testing the improved filter developed within JSI

Questions?

This work was supported by the Slovenian Research Agency, by the Ministry of Education, Science and Sport within the Competence Center Open Communications Platform and the ICT Programme of the EC under PlanetData (ICT-NoE-257641), ENVISION (IST-2009-249120) and NRG4Cast (ICT-EeB- 600074).