# Advanced data analysis and machine learning

## INTEL SENSORS

MARTIN KOSTELNIC, MARIANNE JAKONEN, AHMED ANSARI

Data preprocessing:

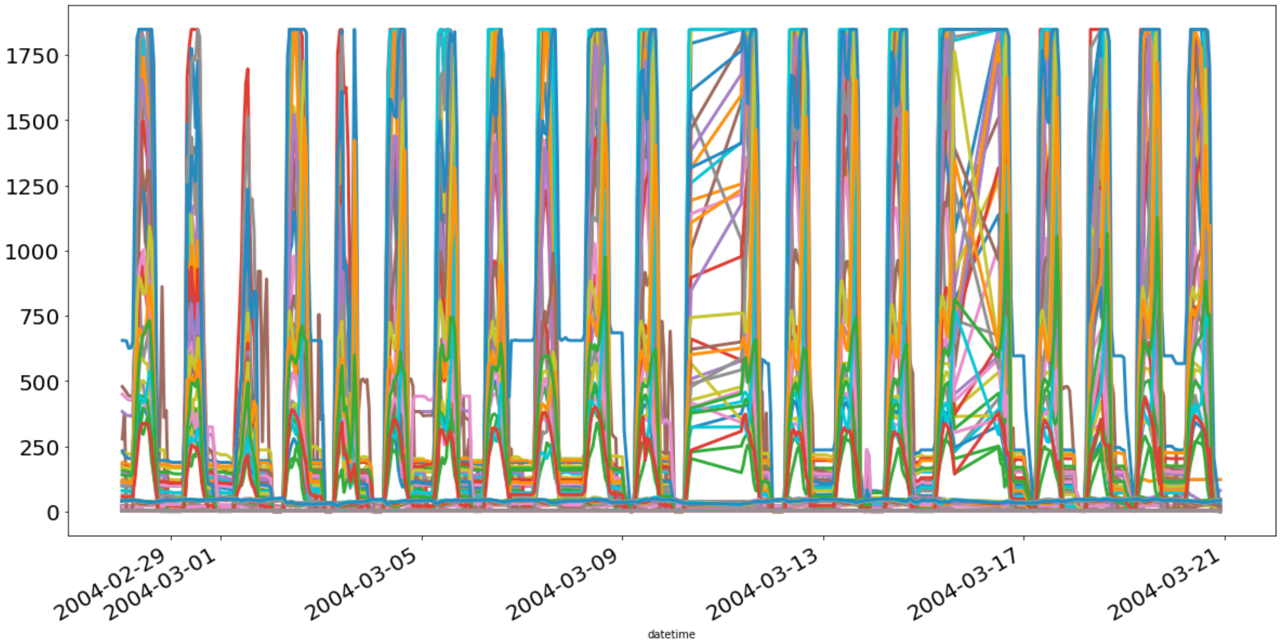
At first some basic filtering was performed on dataset based on theoretical minimum and maximum values of different variables to remove outliers. For example theoretical maximum for humidity measurements is 100% so all values above this were removed.

Data was then rearranged by sensor position id (”mote_id”) and measured variable (temperature, humidity, light, voltage). Dataframe was sorted based on datetime variable.

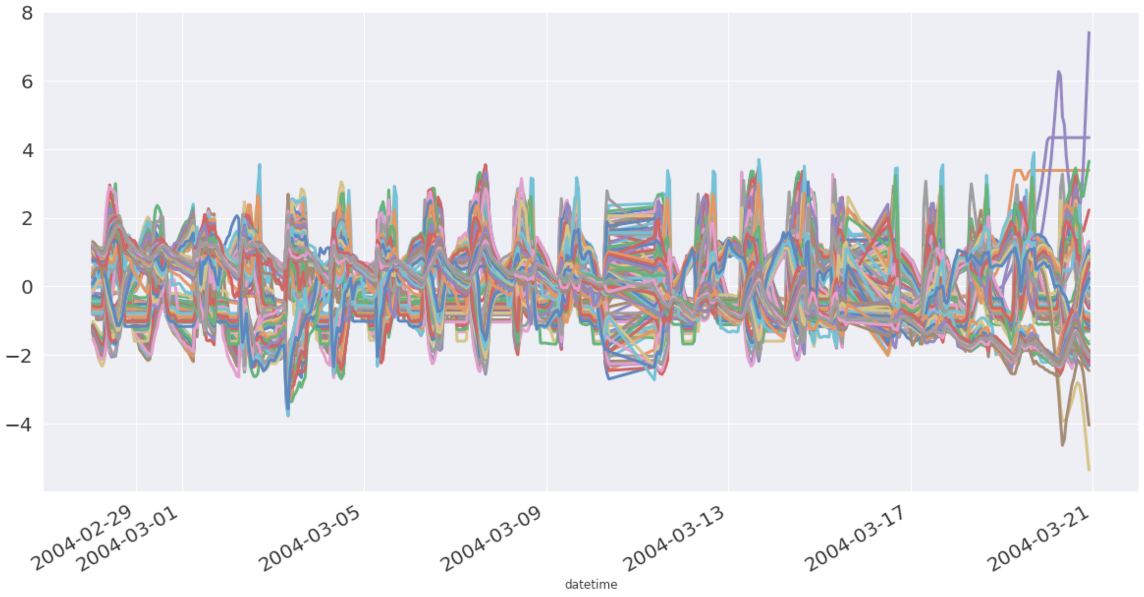
Original dataset has 2313682 rows and 58 columns (variables), measurement interval is 31 seconds. To reduce the size of Dataset it was resampled using 2H sampling interval. Missing values were removed from dataset as follows:

- Rows where all values were missing were dropped
- Rows where more then 80% of data was missing were dropped
- Colums where more than 10 values were missing were dropped

The dataset was reduced to 461 rows and the amount of variables was increaced to 158 as a result of rearranging the dataset by sensor positions. Remaining missing values were interpolated using "linear" method. After removing missing values, the data was plotted. As seen from the figure, the different variables are in different scales.



Data was scaled and centered by using standardization, where the mean is reduced from observations and then divided by standard deviation. The result of standardization is seen in the figure below.


8
6
4
2
0
−2
−4
2004-02-29
2004-03-01
2004-03-05
2004-03-09
2004-03-13
2004-03-17
2004-03-21
datetime