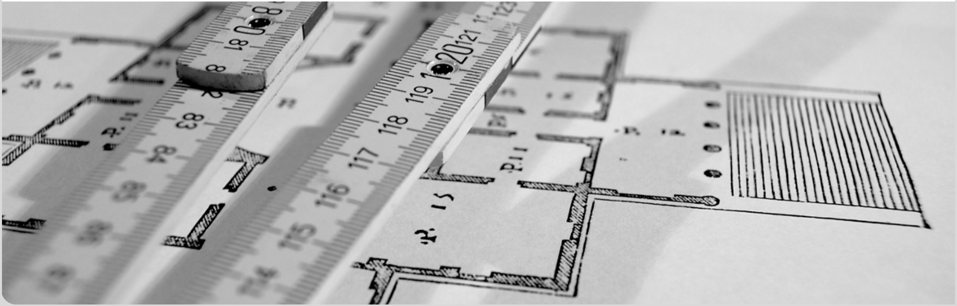
Creating and Evaluating Stochastic Regression Models on the Basis of Heterogeneous Sensor Networks

Bachelors Thesis, Supervisors - Dr. Johannes Riesterer, Dr. Sebastian Lerch

Stanislav Arnaudov | 7. November 2018

TECO - DAS TELECOOPERATION OFFICE



Motivation

Given is a heterogeneous network of sensors – the network contains “good” **and** “bad” sensors.

- good – calibrated and high precision of the measurements
- bad – uncalibrated and low precision of the measurements

Interesting Questions:

- Can we use the bad sensors in order to predict the values of the good ones?
- Can we identify the weak parts of the network?

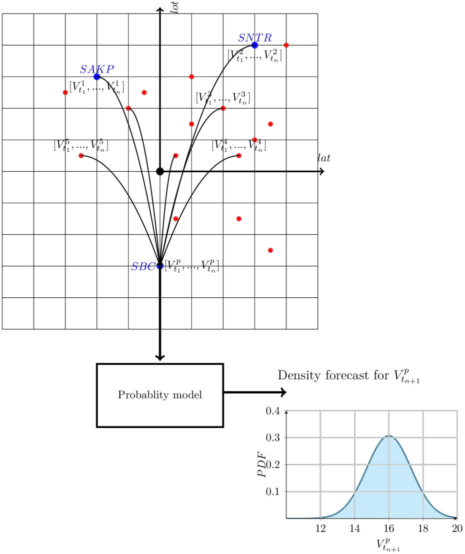
Our approach to the problems:

- Use of stochastic regression models.
- Formal evaluation of the models through the use of proper scoring rules, verification rank histograms and predictive performance checks.
- Analyzing the relevance of each predictor for the final prediction through the use of a feature importance technique.

Data

- Heterogeneous Network of air pollution sensors in Stuttgart.
  - LU-BW (*Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg*) - **3 Sensors** of high quality.
    - Station names: SBC, SNTR, SAKP
  - *luftdaten.info* – Public data from cheap DIY sensors.
- Considered period: the year of 2018.
- Challenges:
  - Noisy data.
  - Holes in the data.
  - Station SAKP provides air pollution values only for the last four months of the considered year.
  - The DIY sensors provide values for each minute of the day, the LUBW sensors - for thirty-minute intervals

General View



Goals and not goals

Goals:

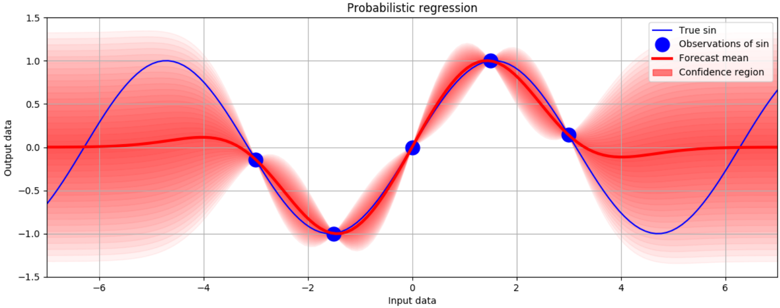
- Investigate how the considered models perform on the data.
- Show that the use of stochastic regression models is a feasible approach to predicting air pollution values in heterogeneous networks of sensors.
- Show that the unreliable parts of the network can be identified through the use of feature importance technique*.

Not goals:

- Build the best possible model for predicting air pollution values.

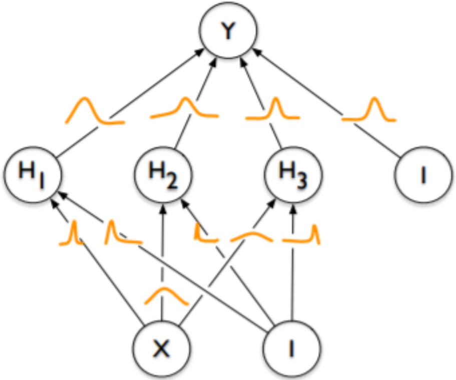
Stochastic Regression Models

- Regression: feature vector $\mapsto$ a real value
- Stochastic Regression: feature vector $\mapsto$ a probability distribution



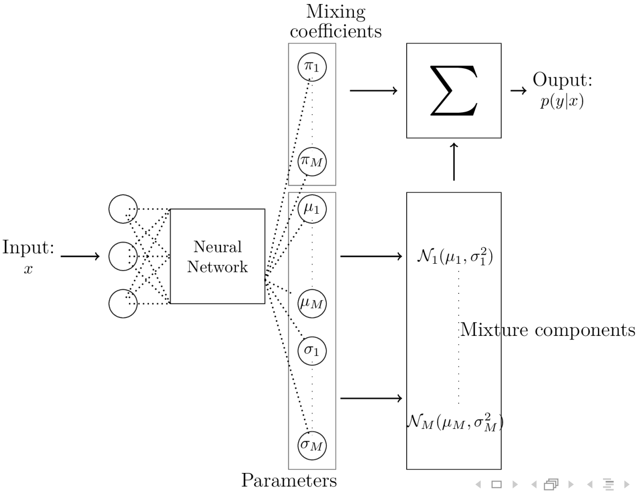
Concrete Models

- Bayesian Neural Networks



Concrete Models

- Mixture Density Networks



Concrete Models

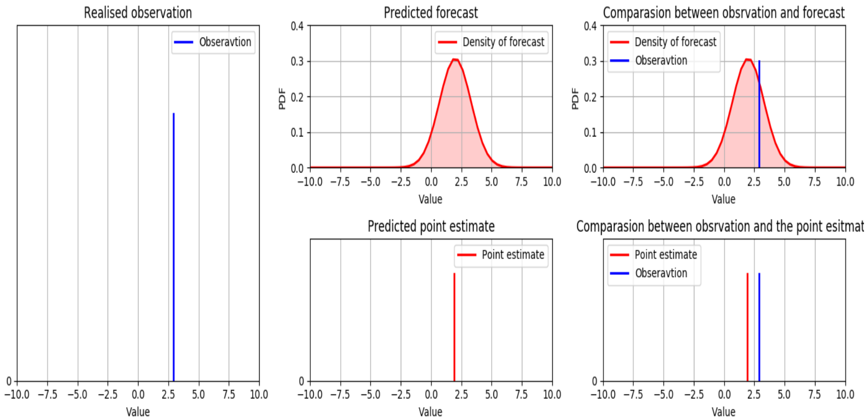
Empirical model

- Past observations are treated as values of a random variable. From that a distribution for a predicted future value is implied.
- We use this model as a baseline.

We hope to achieve better results than the empirical model with the other two models.

Evaluation of probability distributions

- We do not compare point estimate with a realized real value but rather a probability distribution with a real value.

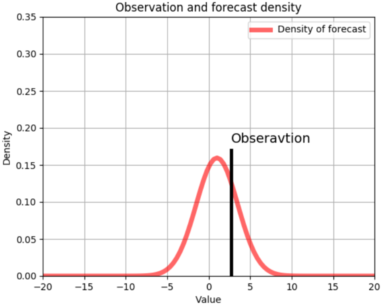


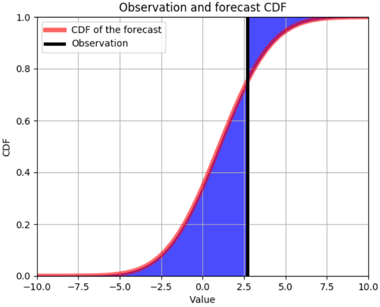
Proper Scoring Rules

Continuous Rank Probability Score

- CRPS compares a distribution with an observation, where both are represented as cumulative distribution functions.

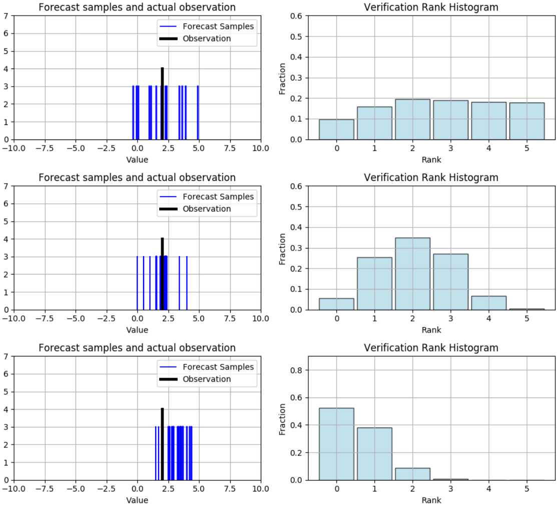
$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}\{y \leq x\})^2 dx$$





Verification Rank Histograms

- Does the observation behave like a random sample from the forecast distribution?



Feature importance

- Assess the **relative** importance of all features used by a model.
  - Important: Importance is relative only to the model itself.
- Random shuffling of each predictor/feature in the test set one at a time and observing the increase/decrease in mean the CRPS value compared to the unpermuted features.

Training

- Approach
  Based on the data and the examined types of models there are several orthogonal considerations to make when it comes to training:
  - Target LUBW-Station – SAKP, SBC, SNTR
  - Integration period – one day, twelve hours, one hour
  - Use of the other two LUBW station – yes or no
  - Predicted air pollution value – PM10 or PM2.5
  - Used model - BNN or MDN (two types of MDNs)

  We wanted to investigate every possible combination.
- Challenges:
  - BNNs are very hard to train.
  - A lot of possible ways to train several different models.

Results

Results presented here: representative sample of all results.

Results

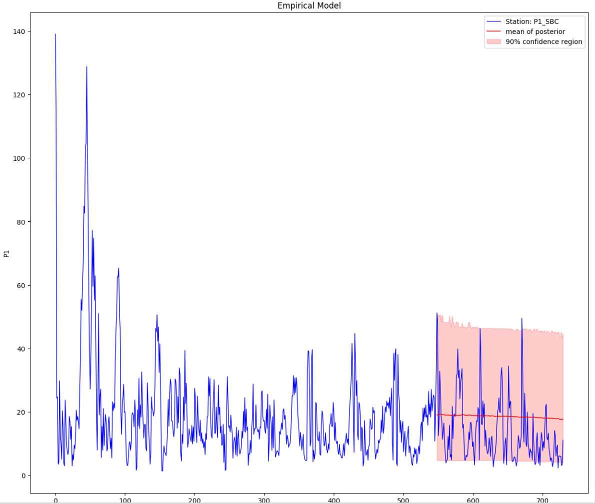
Results presented here: representative sample of all results.

- Plots showing how exactly the built models predict the values of their train and test sets.

Results

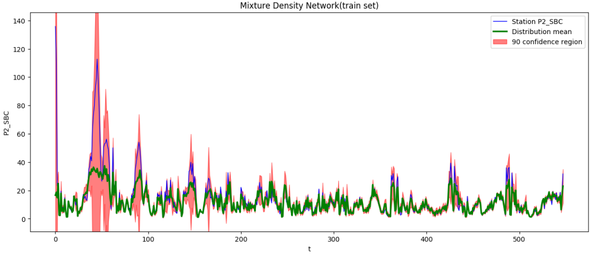
Karlsruhe Institute of Technology

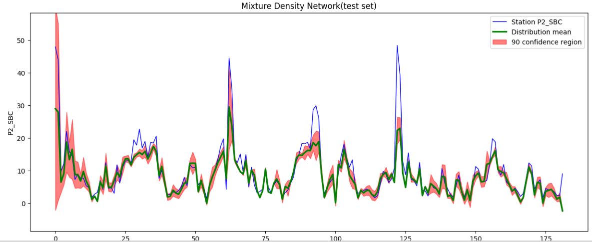
Empirical Model, twelve hour averaged data



Results

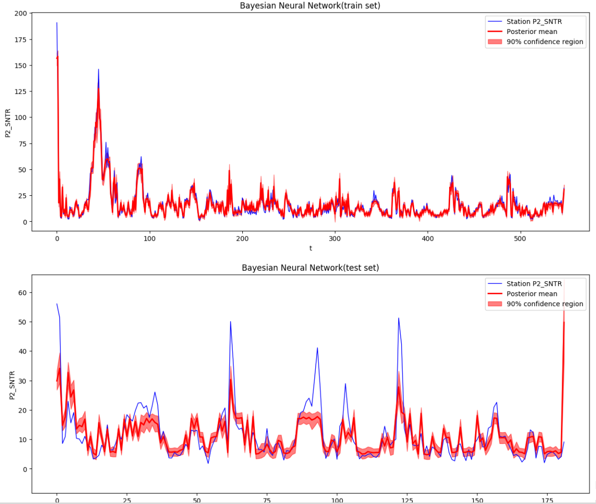
MDN, twelve hour averaged data





Results

BNN, twelve hour averaged data



Results

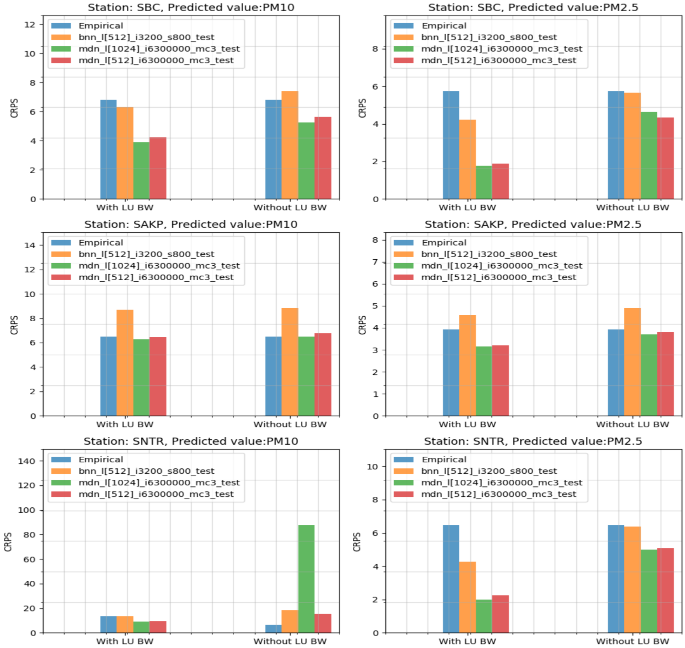
Results presented here: representative sample of all results.

- Plots of the models modeling their respective test and train sets.

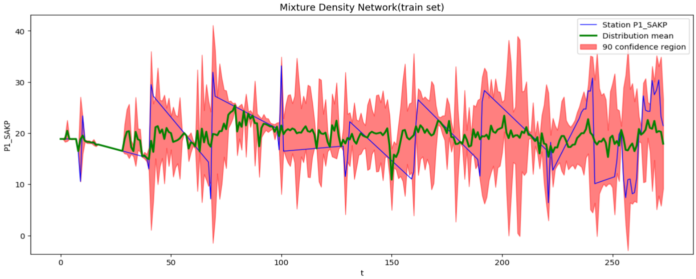
Results

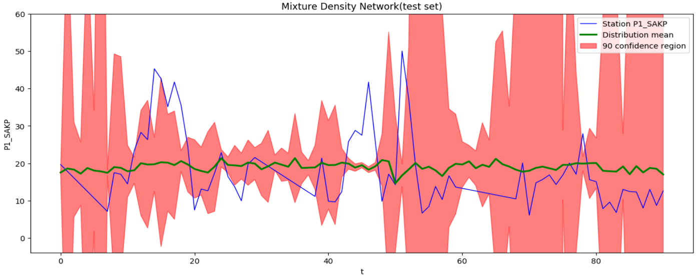
Results presented here: representative sample of all results.

- Plots of the models modeling their respective test and train sets.
- Plots comparing the built models with respect their CRPS score across LUBW Station, use of LUBW data and predicted air pollution value.


Station: SBC, Predicted value:PM10
Station: SBC, Predicted value:PM2.5
Station: SAKP, Predicted value:PM10
Station: SAKP, Predicted value:PM2.5
Station: SNTR, Predicted value:PM10
Station: SNTR, Predicted value:PM2.5
Empirical
bnn_l[512]_i3200_s800_test
mdn_l[1024]_i6300000_mc3_test
mdn_l[512]_i6300000_mc3_test
CRPS
With LU BW
Without LU BW

Results


Mixture Density Network(train set)
Station P1_SAKP
Distribution mean
90 confidence region
P1_SAKP
t


Mixture Density Network(test set)
Station P1_SAKP
Distribution mean
90 confidence region
P1_SAKP
t

Results

Results presented here: representative sample of all results.

- Plots of the models modeling their respective test and train sets.
- Plots comparing the built models with respect their CRPS score across LUBW Station, use of LUBW data and predicted air pollution value.

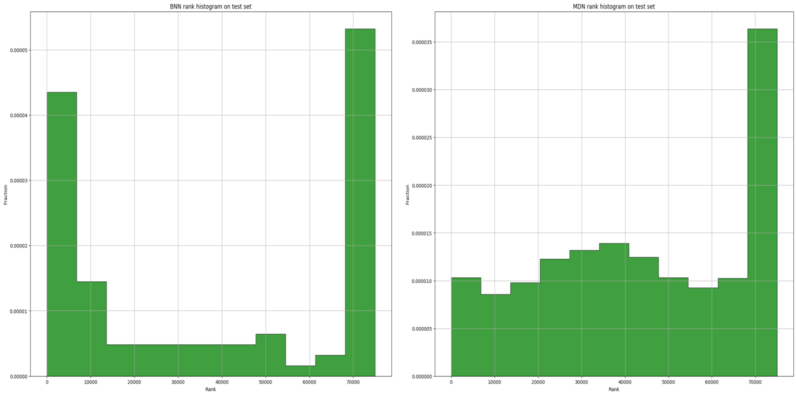
Results

Results presented here: representative sample of all results.

- Plots of the models modeling their respective test and train sets.
- Plots comparing the built models with respect their CRPS score across LUBW Station, use of LUBW data and predicted air pollution value.
- Two rank histograms reflecting the results of all built models.

Results

BNN (one day average) and MDN (one hour average), Rank Histograms



Results

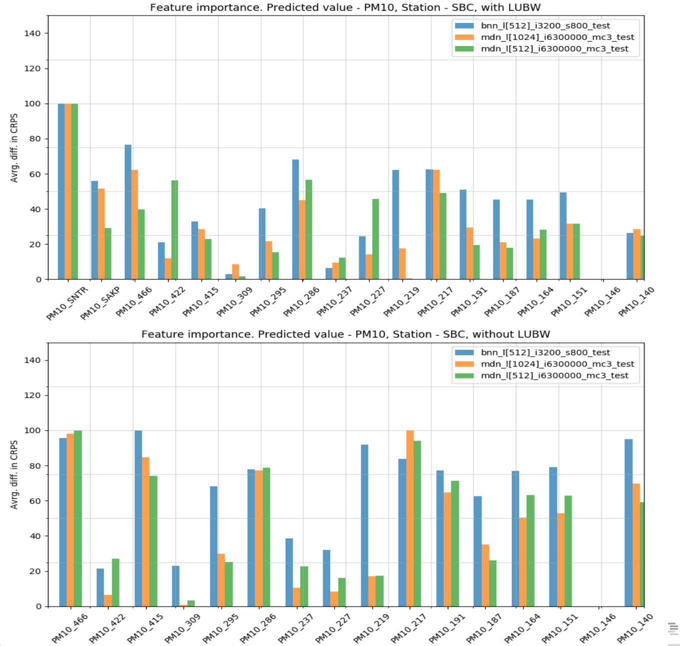
Results presented here: representative sample of all results.

- Plots of the models modeling their respective test and train sets.
- Plots comparing the built models with respect their CRPS score across LUBW Station, use of LUBW data and predicted air pollution value.
- Two rank histograms reflecting the results of all built models.

Results

Results presented here: representative sample of all results.

- Plots of the models modeling their respective test and train sets.
- Plots comparing the built models with respect their CRPS score across LUBW Station, use of LUBW data and predicted air pollution value.
- Two rank histograms reflecting the results of all built models.
- Feature importance data.



Feature importance. Predicted value - PM10, Station - SBC, with LUBW
bnn_l[512]_i3200_s800_test
mdn_l[1024]_i6300000_mc3_test
mdn_l[512]_i6300000_mc3_test
Avrg. diff. in CRPS
PM10_SNTR
PM10_SAKP
PM10_466
PM10_422
PM10_415
PM10_309
PM10_295
PM10_286
PM10_237
PM10_227
PM10_219
PM10_217
PM10_191
PM10_187
PM10_164
PM10_151
PM10_146
PM10_140
Feature importance. Predicted value - PM10, Station - SBC, without LUBW
bnn_l[512]_i3200_s800_test
mdn_l[1024]_i6300000_mc3_test
mdn_l[512]_i6300000_mc3_test
Avrg. diff. in CRPS
PM10_466
PM10_422
PM10_415
PM10_309
PM10_295
PM10_286
PM10_237
PM10_227
PM10_219
PM10_217
PM10_191
PM10_187
PM10_164
PM10_151
PM10_146
PM10_140

Conclusion

The results show:

- Stochastic regression models are a viable approach for predicting values in the investigated sensor network.
- Few sensors have consistently low feature importance for the built models.
- There is clearly room for improvement of the models.

Thank you for your attention.

Questions?