Gaussian Processes for Predicting Air Quality in Kampala

Clara Stoddart, Lauren Shrack, Richard Sserunjogi, Usman Abdul-Ganiy, Engineer Bainomugisha, Deo Okure, Ruth Misener, Jose Pablo Folch, Ruby Sedgwick

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023

Air quality

- Recognised as the **number one environmental threat** to global health by the World Health Organisation
- Estimated to cause **7 million** deaths annually
- **PM2.5** recognised as a particularly dangerous pollutant

AirQo

Research group in Kampala aiming to empower communities across Africa with information about the quality of the air they breathe

Developed a network of low-cost air quality monitors across Uganda, Kenya, Senegal and Cameroon.

Aim of the project

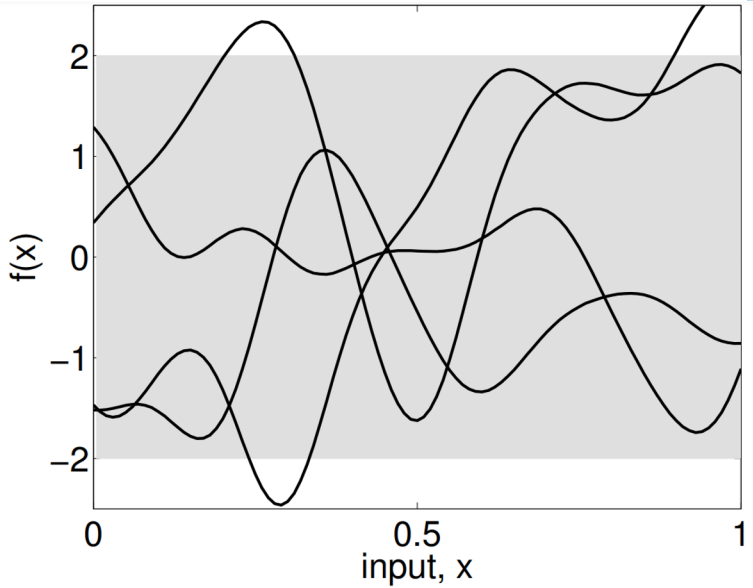
- Use machine learning models to predict PM2.5 levels in places where there are no sensors
- Forecast PM2.5 levels

Background

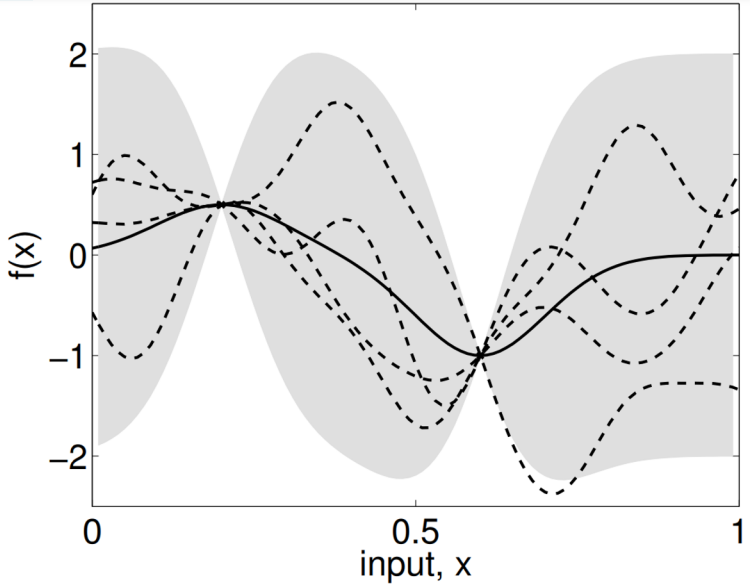
Gaussian Processes

A probabilistic approach to machine learning, giving us highly flexible, data-efficient models, which provide uncertainty estimates

Gaussian Processes

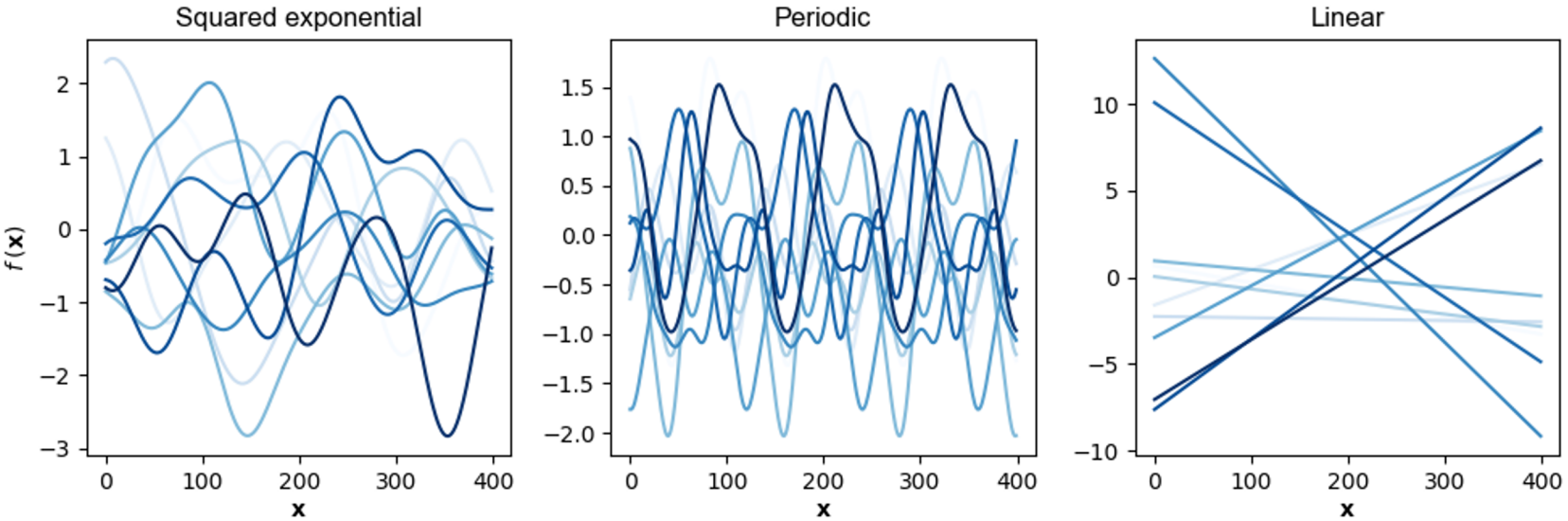


Function samples from our prior distribution



Function samples from our posterior after observing two data points

Example Kernels



The functions we can learn are define by the kernels

Sparse Gaussian Processes

- Gaussian processes require the inversion of NxN covariance matrices, where N is the number of data points
- Hence have time complexity of $O(N^3)$
- Sparse GPs instead base all computations on a subset of the most representative M data points, called inducing points

Sparse Gaussian Processes

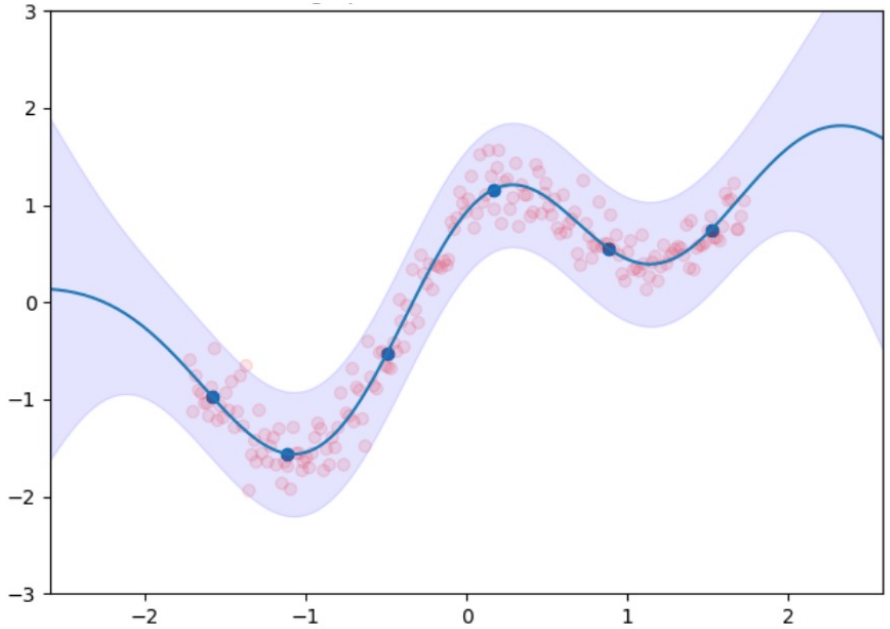


image taken from: https://ludwigwinkler.github.io

Sparse Gaussian Process Methods

Spatio-Temporal Variational GPs:

- Adaptation tailored for spatio-temporal tasks
- Reduces temporal scaling from cubic to linear

[2] Oliver Hamelijnck, William J. Wilkinson, Niki A. Loppi, Arno Solin, and Theodoros Damoulas. Spatio-temporal variational Gaussian Processes

Our model

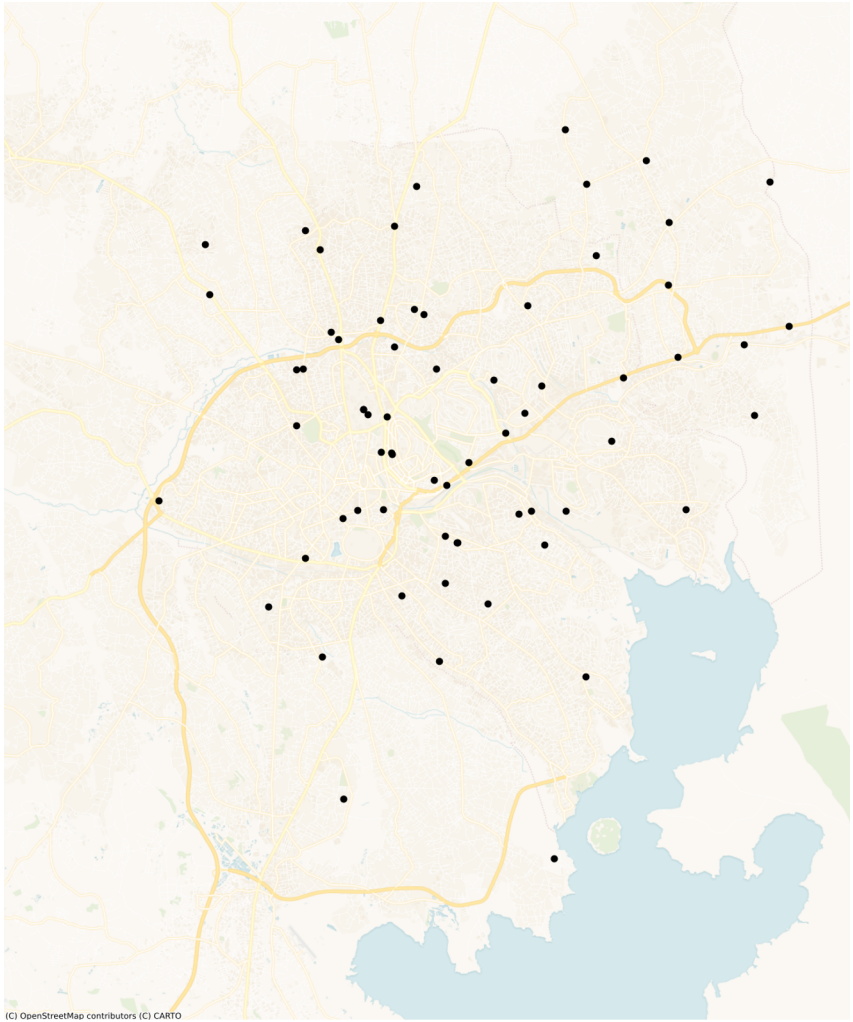
Contributions

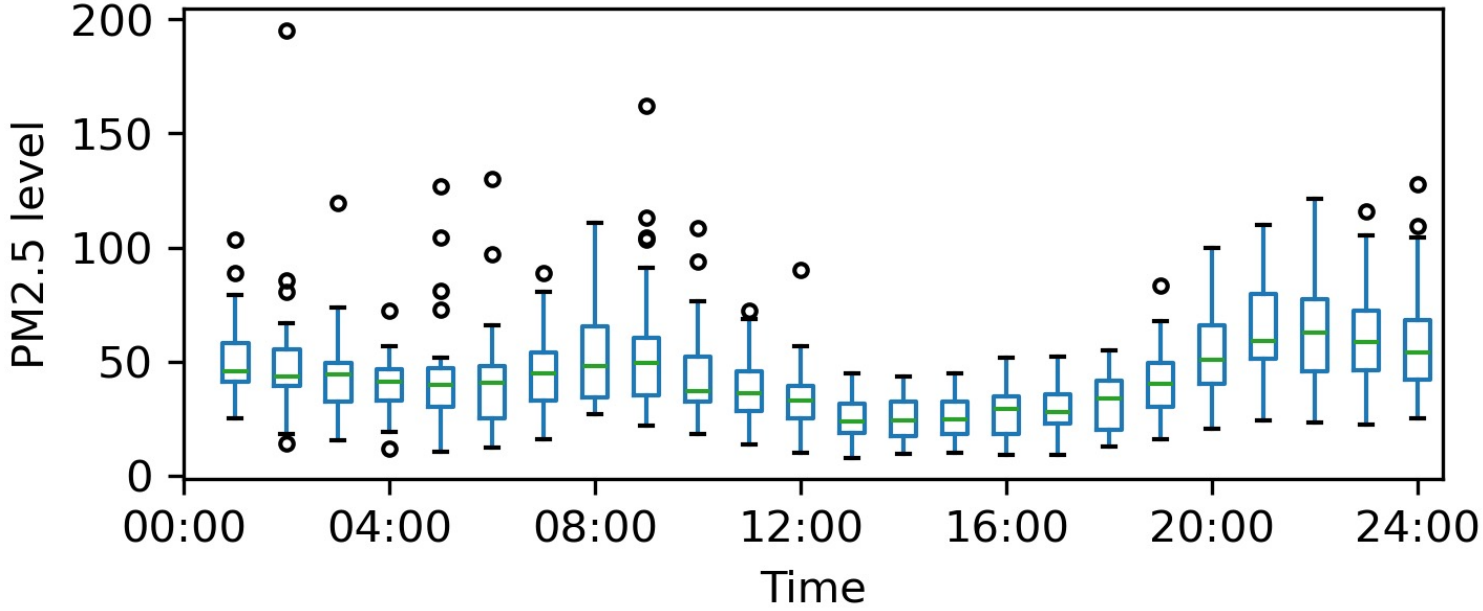
- Developed and tested Gaussian Process model for both forecasting and nowcasting PM2.5 levels in the air. Comparing the effect of:

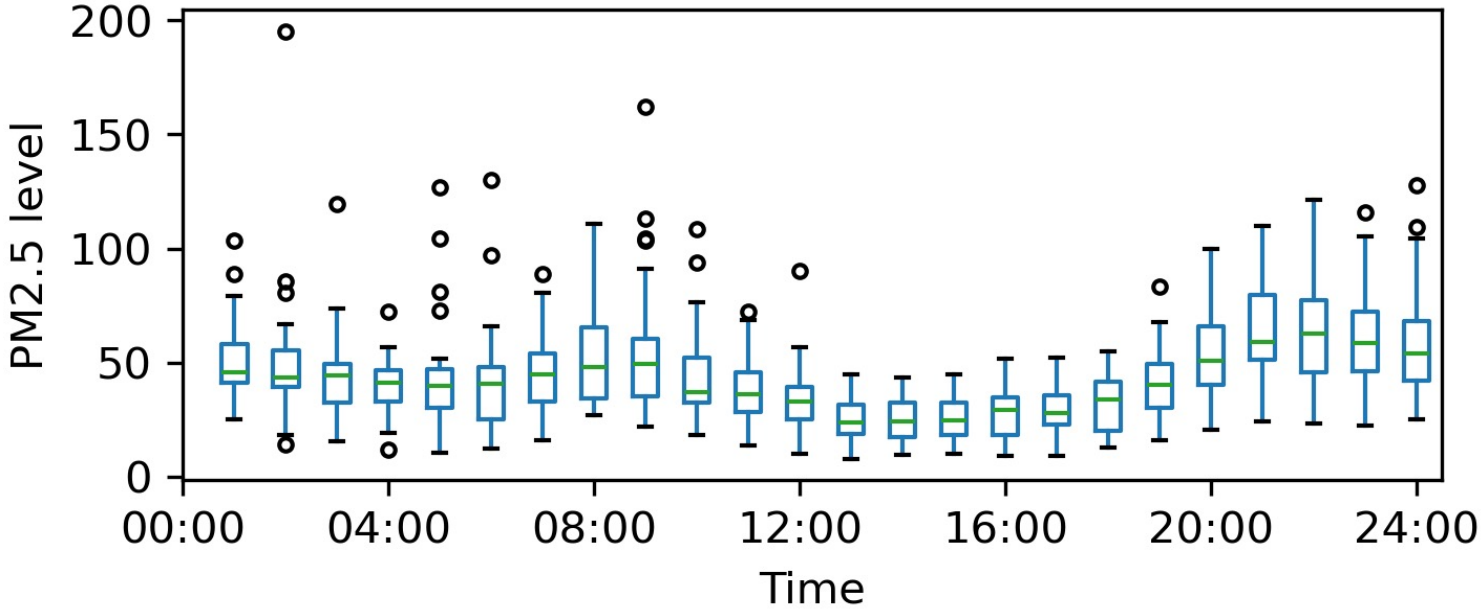
1. Using a periodic and non-periodic kernels
2. Removing outliers from the data
3. Incorporation of meteorological information
4. Sparse approximation methods

Dataset

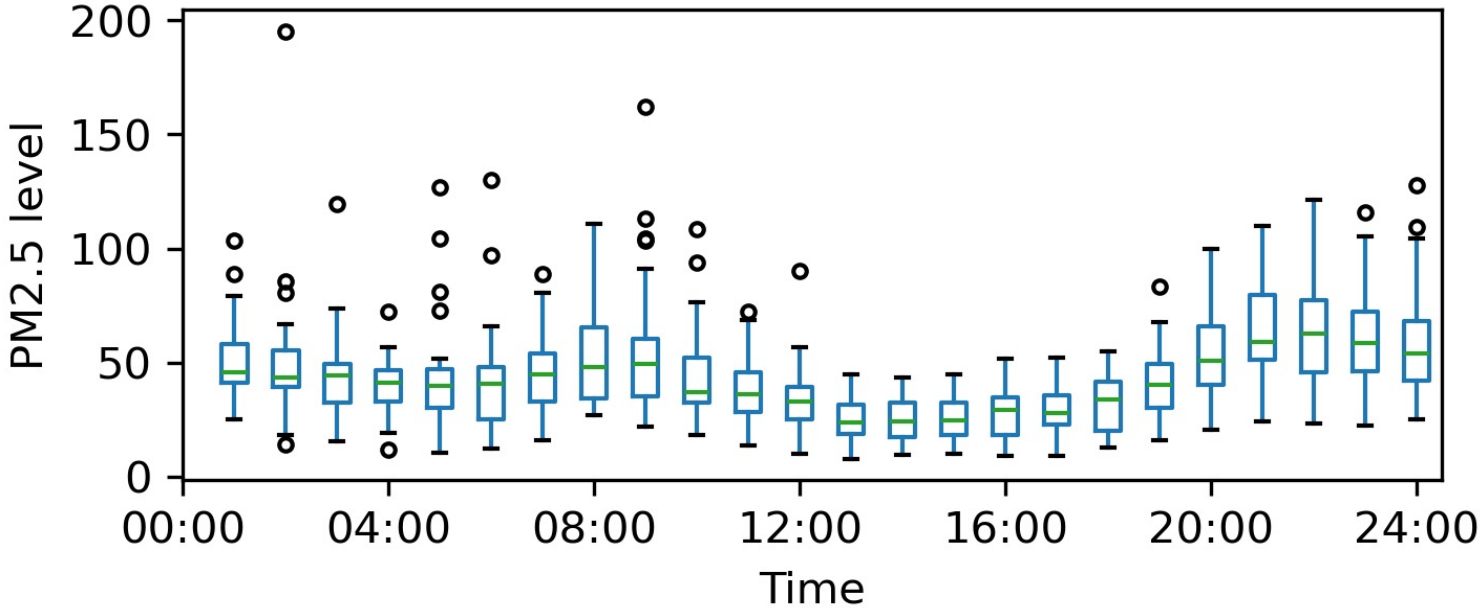
- Provided by AirQo
- We chose to use data from the month of November 2021
- Large number of null or missing entries which were removed
- Total of 66 sensor locations


(C) OpenStreetMap contributors (C) CARTO

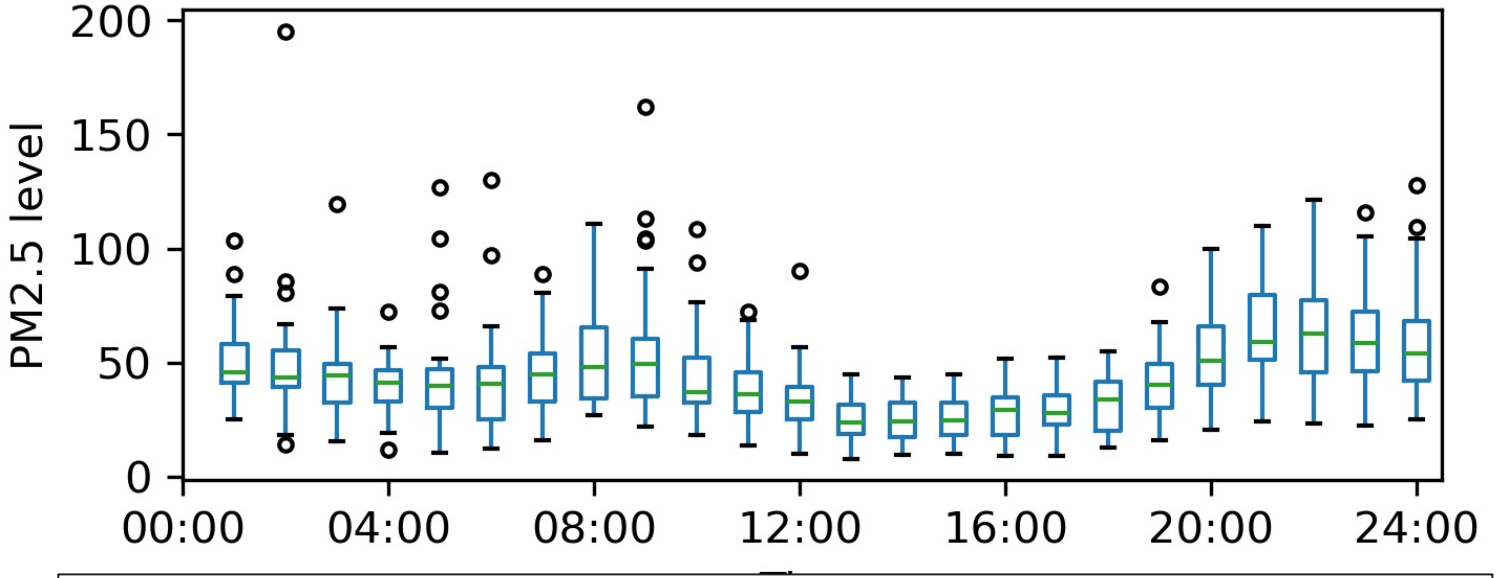

200
150
100
50
0
PM2.5 level
00:00
04:00
08:00
12:00
16:00
20:00
24:00
Time



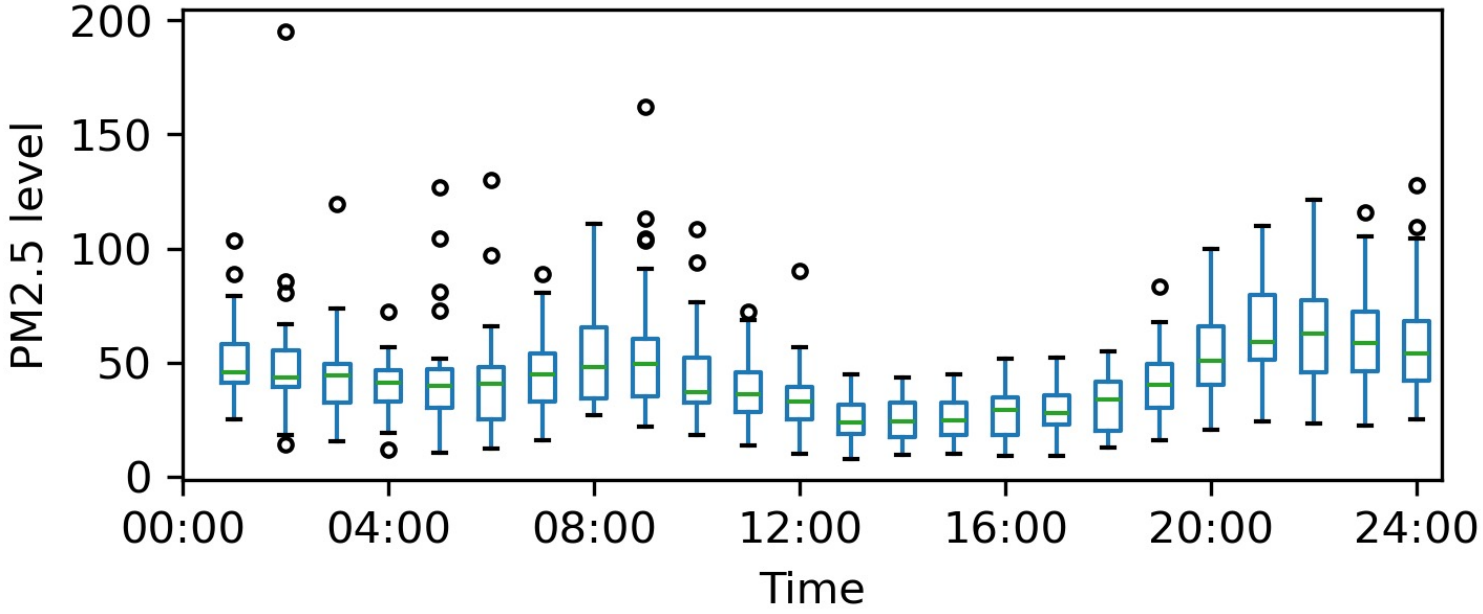
1. We can observe a **clear periodic pattern** related to commute times



2. There is so much data that **sparse approximations** are needed to train on the full monthly data-set



3. There are a lot of **outliers in the data**, and it is not obvious we can predict them. It could be the case they are caused by rare events which do not correlate with the data and may bias the model.



4. We can incorporate **extra predictors**, for example meteorological information.

Testing the model

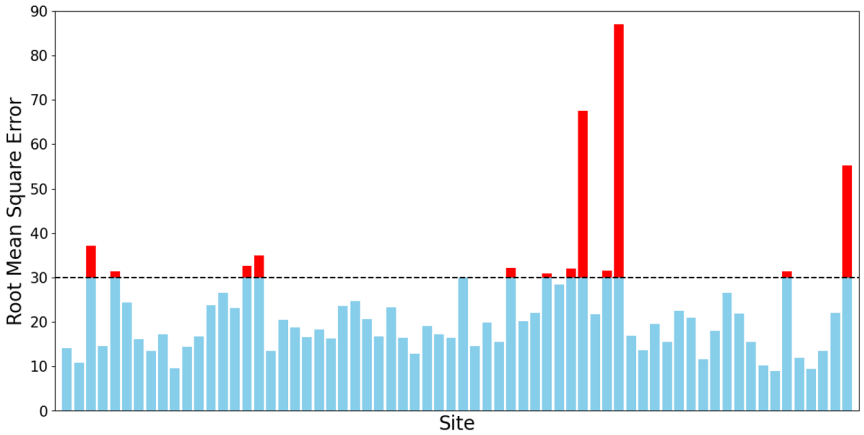
- We use Root Mean Squared Error (RMSE) as performance metric
- Nowcasting: Leave-one-out cross validation on all the sensor locations
- Forecasting: Predicting pollution levels on the last day of the month

Conclusion

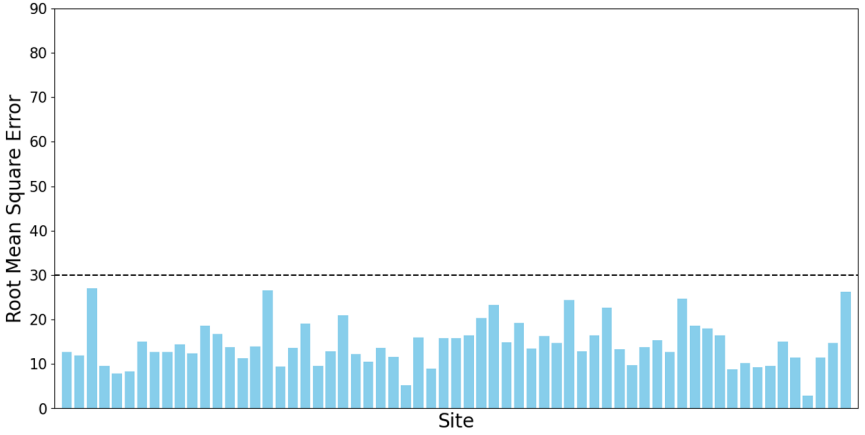
Conclusions

- Overall, our models proved to be better at forecasting than nowcasting. This is probably due to the need to have a better spatial kernel which can capture non-stationarities in the data, and the need for more sensors. However, the nowcasting models still have good predictions and could be used in practice.

Sparse Approximations: **Nowcasting vs Forecasting**



ST-SVGP model



SVGP model

Conclusions

- For **forecasting**, a periodic sparse Gaussian Processes proved to be a very strong yet simple model. Removing outliers seemed to make a big difference in the predictive performance as they are rare events tend to bias the model.

Conclusions

- Incorporating additional inputs, such as meteorological information cannot be done naively. It did not lead to any noticeable improvements; however, it also did not negatively affect the model either.

Conclusions

- Incorporating additional inputs, such as meteorological information cannot be done naively. It did not lead to any noticeable improvements; however, it also did not negatively affect the model either.

Future work

- ST-SVGP model
  - Periodic kernel
- Predicting multiple pollutants
- Decision making using uncertainty estimates from the model