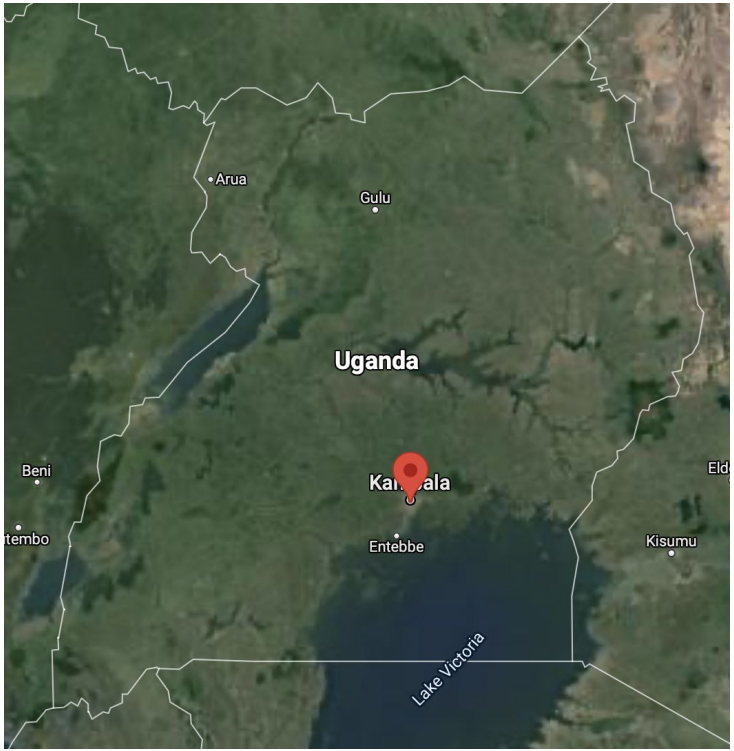
Predicting air quality using meteorological observations in Kampala, Uganda

Konstanze, Paulos, Jerome



Value & Stakeholder

Value

- Predicting **particulate matter concentration** based on **meteorological measurements**
- **Protect community health** by warning population in case of expected high particulate matter load

Stakeholder

- **Ministry of Health**, which makes the prediction system available to various stakeholders and to the population

Dataset

- Dataset result of **AirQo** research initiative of **Makerere University, Kampala, Uganda**

- ca. **15000 weather** and **particulate matter** observation sets at **five locations** in **Kampala**

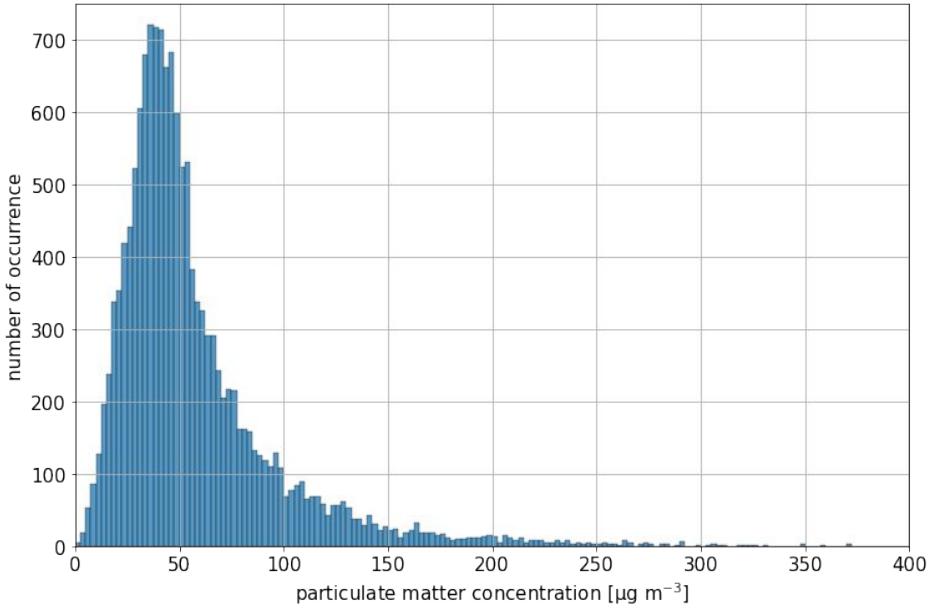
- **per observation set:**

  -> **five days** of **hourly meteorological** measurements (e.g. temperature, precipitation)

  -> one **particulate matter** measurement **24 hours** after last weather observation

Target variable: particulate matter concentration

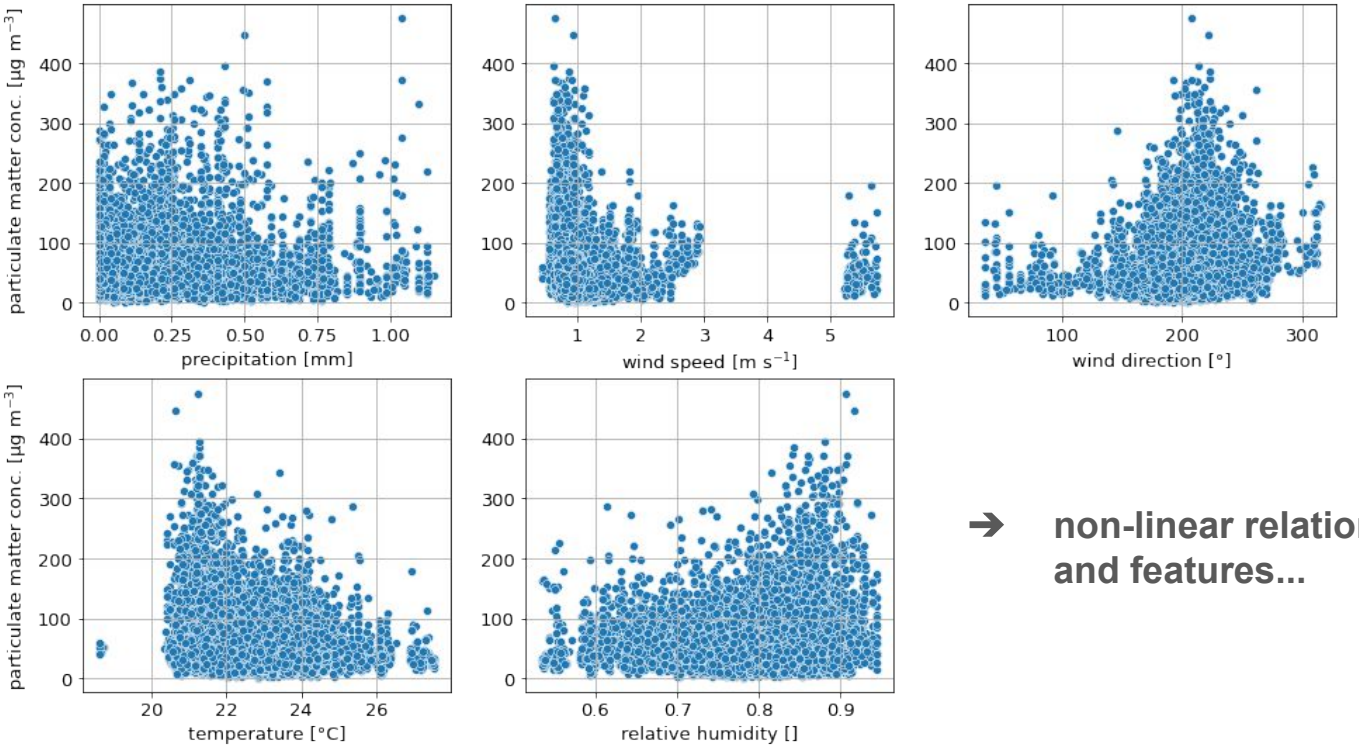
- **PM2.5 concentration [μg m^{-3}]** (particles with a diameter smaller than 2.5 μm)



Health Concern	$PM_{2.5}$ ($\mu g m^{-3}$)	Precautions
Good	0 - 12	None
Moderate	13 - 35	Unusually sensitive people should consider reducing prolonged or heavy exertion
Unhealthy for Sensitive Groups	36 - 55	Sensitive groups should reduce prolonged or heavy exertion
Unhealthy	56 - 150	Everyone should reduce prolonged or heavy exertion, take more breaks during outdoor activities
Very Unhealthy	151 - 250	Everyone should avoid prolonged or heavy exertion, move activities indoors or reschedule
Hazardous	250 +	Everyone should avoid all physical activities outdoors.

Features

- **precipitation**, **wind speed** and **direction**, **temperature**, **relative humidity**
- **average** over **five days** of **weather observations** used



➔ **non-linear relationships between target and features...**

RMSE score & baseline model

- **RMSE (Root Mean Squared Error) score**

  -> can be thought of a kind of mean deviation between the predicted and the observed values

  -> in a way to answer the question: “How far off should we expect our model to be on its next prediction on average?”

- **baseline model**

  -> always predict the test set’s **mean particulate matter load** of **58 µg m^{-3}**

  -> score: RMSE = 42 µg m^{-3} **~ 72 % of the mean particulate matter load**

Models to predict particulate matter concentration

Model	RMSE [µg m^{-3}]	Percentage of the mean particulate matter load [%]
Baseline model	42	~ 72
Multivariate linear regression	41	~ 71
Decision tree	36	~ 62
Random Forest	28	~ 48

→ **linear regression** model here n**ot suitable** due to non-linear target-feature relationships

→ **best prediction** by **Random Forest Regression**

Summary

- ca. **15000 meteorological measurements** used to train and test models to predict **particulate matter load 24 hours ahead**

- features: **precipitation**, **wind speed** and **direction**, **temperature**, **relative humidity**

- **linear regression not suitable due to non-linear target-features relations**

- **best performance** by **Random Forest Regression** with RMSE of **28 µg m^{-3}**
  ➔ ~ 48 % of mean particulate matter load
  ➔ ~ deviation of two categories in the upper part of guideline on hazardous levels

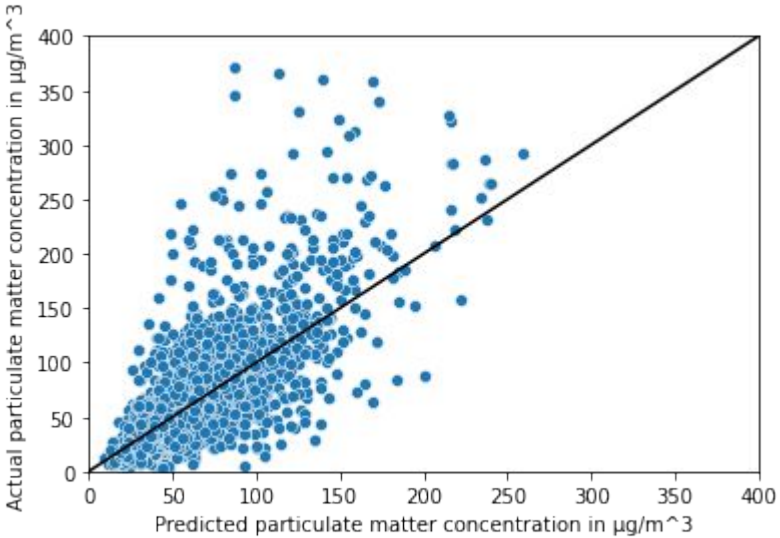
Outlook

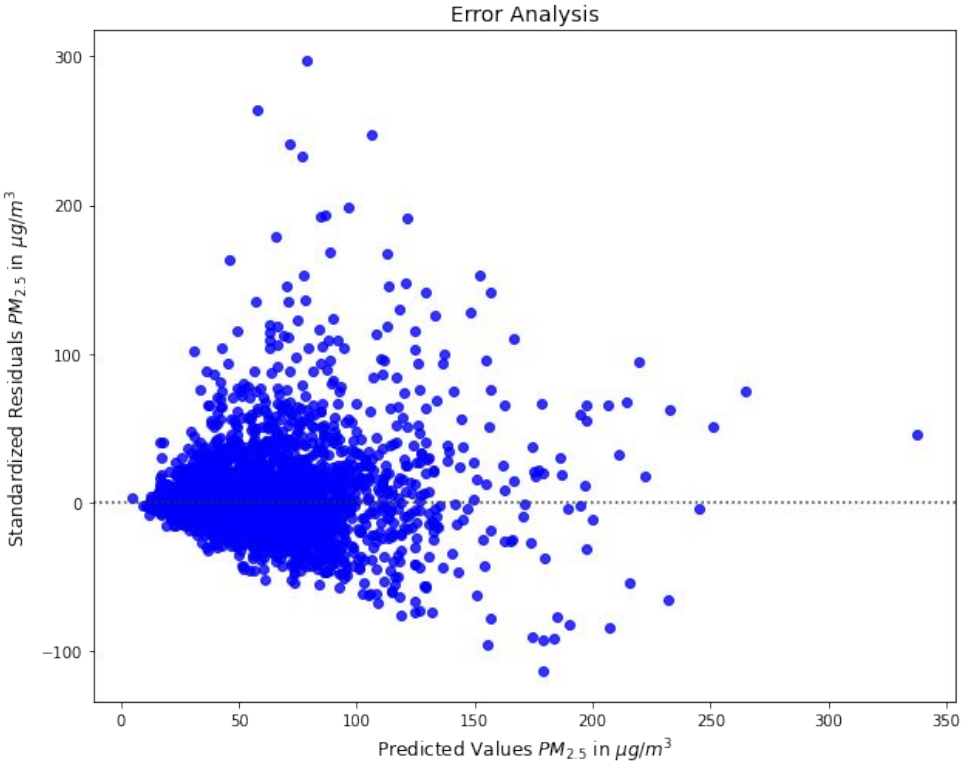
- result of error analysis: **model underestimates large particulate matter concentrations**
  - ➔ separate model formulation for extreme and thus the most dangerous particulate matter concentrations required?

- **model analysis** for the **five locations separately** and more sophisticated analysis for **different averaging periods**

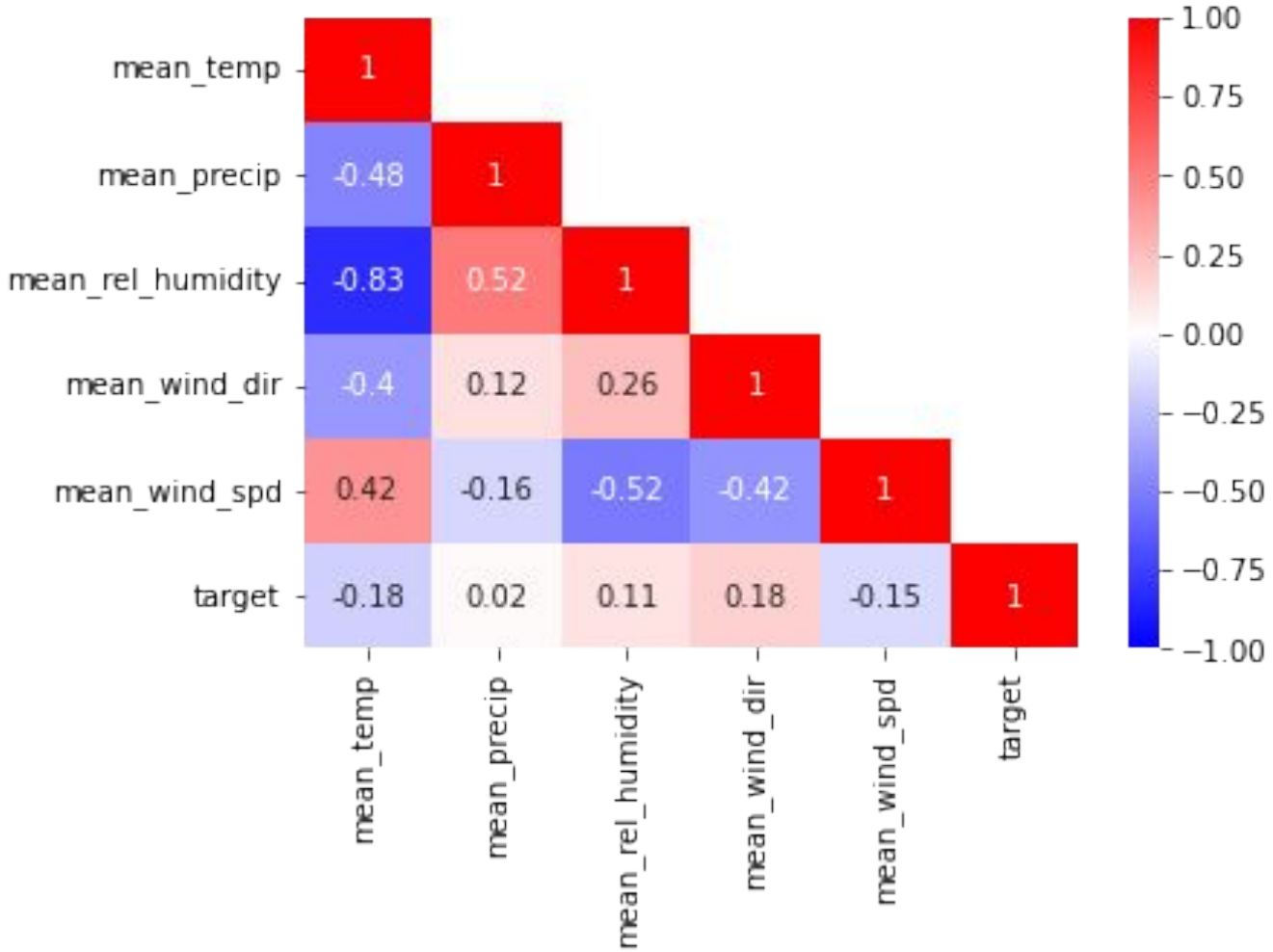
- **better averaging of wind direction**

Appendix

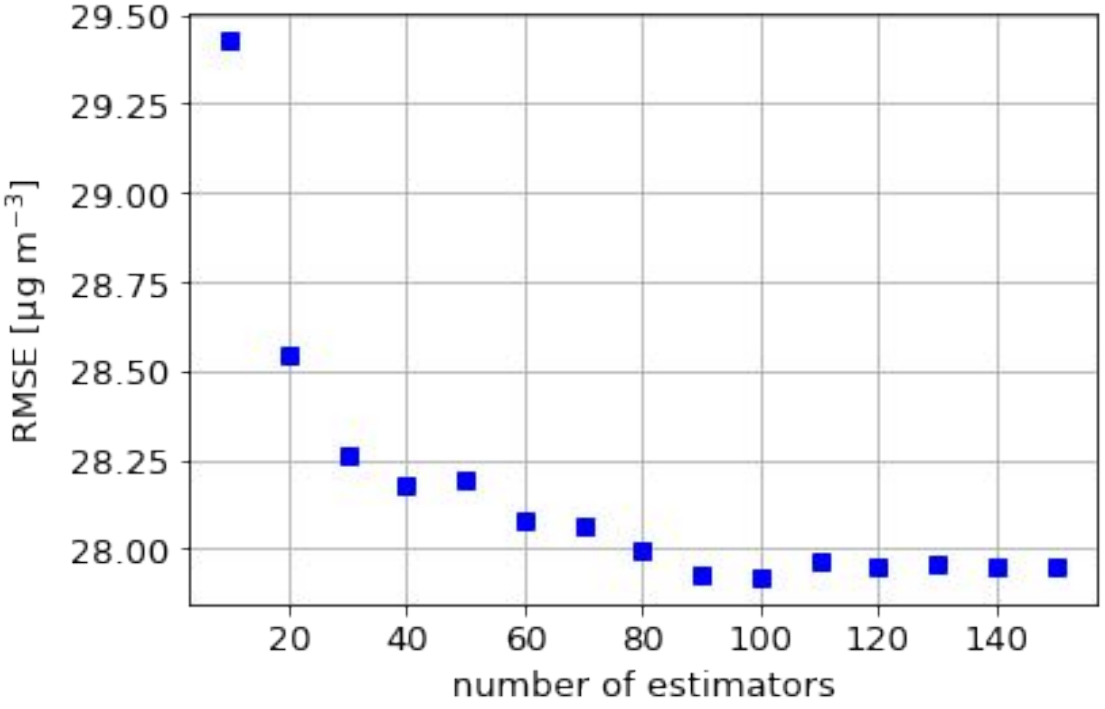

Actual particulate matter concentration in µg/m^3
Predicted particulate matter concentration in µg/m^3
0
50
100
150
200
250
300
350
400


Error Analysis
Standardized Residuals $PM_{2.5}$ in $\mu g/m^3$
Predicted Values $PM_{2.5}$ in $\mu g/m^3$
300
200
100
0
−100
0
50
100
150
200
250
300
350

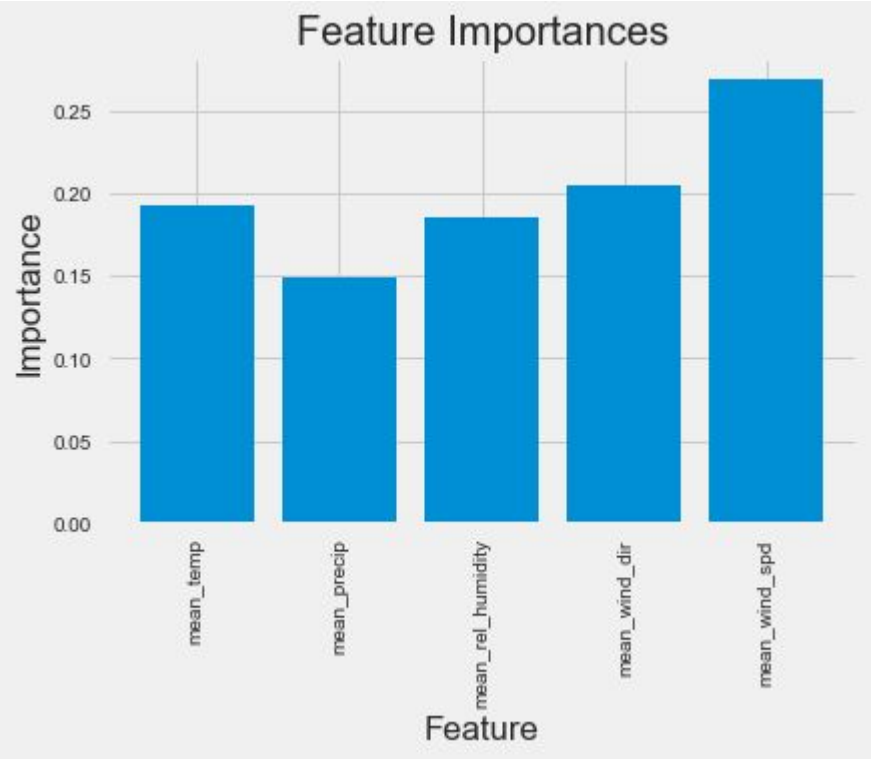
Appendix


mean_temp
mean_precip
mean_rel_humidity
mean_wind_dir
mean_wind_spd
target
1
-0.48 1
-0.83 0.52 1
-0.4 0.12 0.26 1
0.42 -0.16 -0.52 -0.42 1
-0.18 0.02 0.11 0.18 -0.15 1
mean_temp
mean_precip
mean_rel_humidity
mean_wind_dir
mean_wind_spd
target
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

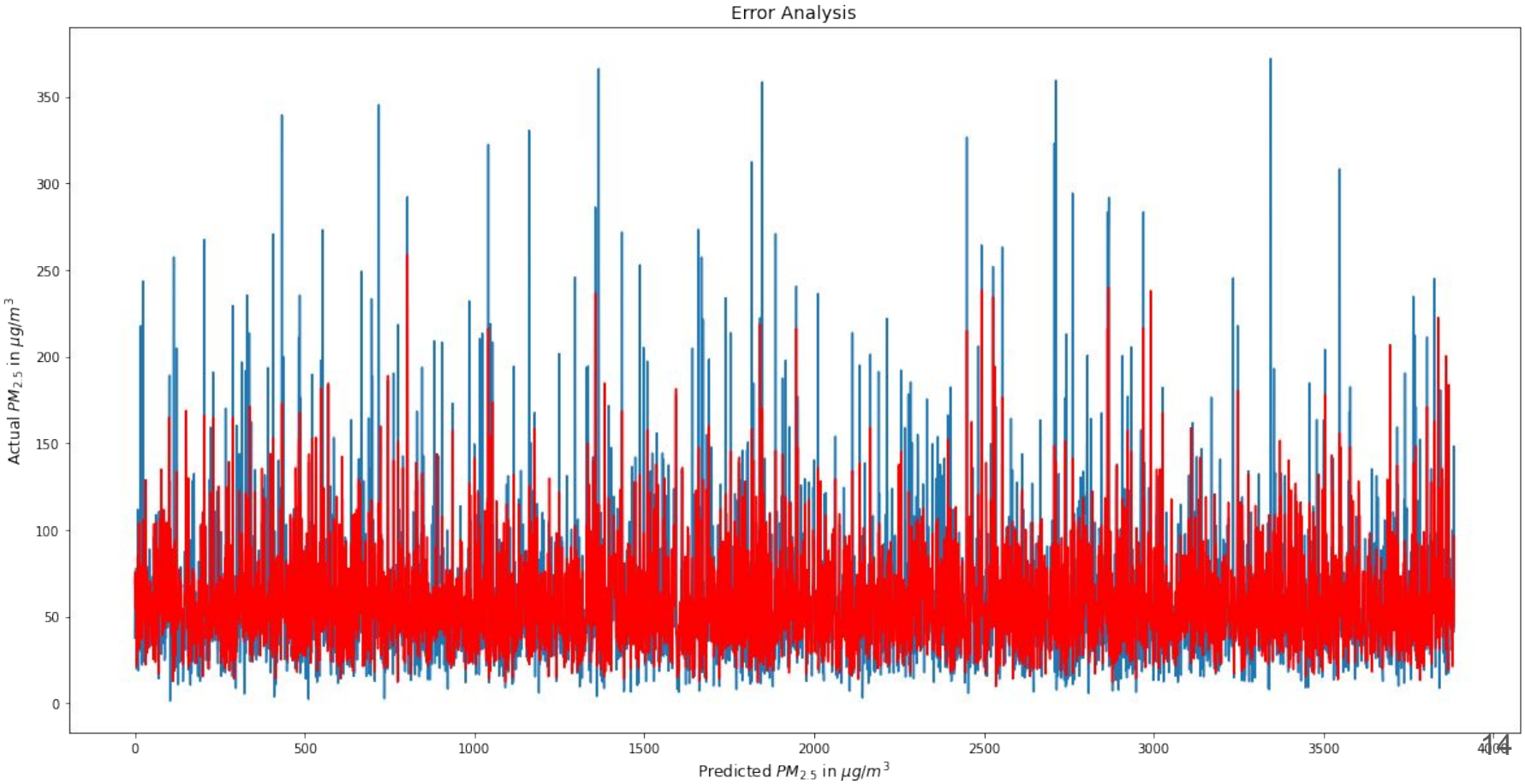
Appendix


29.50
29.25
29.00
28.75
28.50
28.25
28.00
RMSE [µg m^{-3}]
20
40
60
80
100
120
140
number of estimators

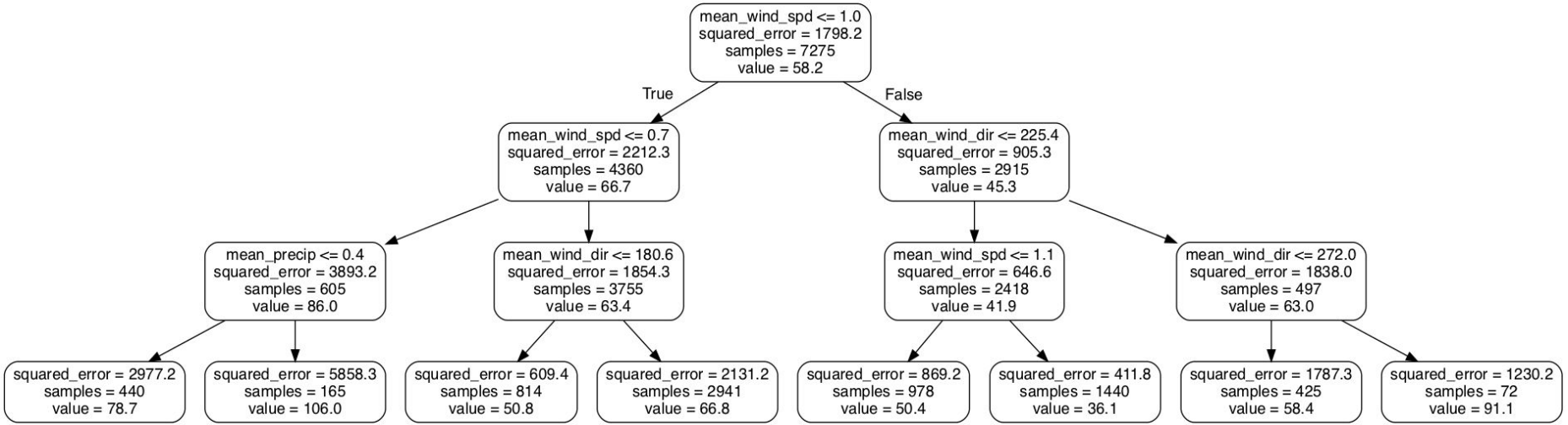
Appendix


Feature Importances
Importance
0.25
0.20
0.15
0.10
0.05
0.00
mean_temp
mean_precip
mean_rel_humidity
mean_wind_dir
mean_wind_spd
Feature

Appendix


Error Analysis
Actual $PM_{2.5}$ in $\mu g/m^3$
Predicted $PM_{2.5}$ in $\mu g/m^3$
0
50
100
150
200
250
300
350
0
500
1000
1500
2000
2500
3000
3500
4000

Appendix


mean_wind_spd <= 1.0
squared_error = 1798.2
samples = 7275
value = 58.2
True
False
mean_wind_spd <= 0.7
squared_error = 2212.3
samples = 4360
value = 66.7
mean_wind_dir <= 225.4
squared_error = 905.3
samples = 2915
value = 45.3
mean_precip <= 0.4
squared_error = 3893.2
samples = 605
value = 86.0
mean_wind_dir <= 180.6
squared_error = 1854.3
samples = 3755
value = 63.4
mean_wind_spd <= 1.1
squared_error = 646.6
samples = 2418
value = 41.9
mean_wind_dir <= 272.0
squared_error = 1838.0
samples = 497
value = 63.0
squared_error = 2977.2
samples = 440
value = 78.7
squared_error = 5858.3
samples = 165
value = 106.0
squared_error = 609.4
samples = 814
value = 50.8
squared_error = 2131.2
samples = 2941
value = 66.8
squared_error = 869.2
samples = 978
value = 50.4
squared_error = 411.8
samples = 1440
value = 36.1
squared_error = 1787.3
samples = 425
value = 58.4
squared_error = 1230.2
samples = 72
value = 91.1