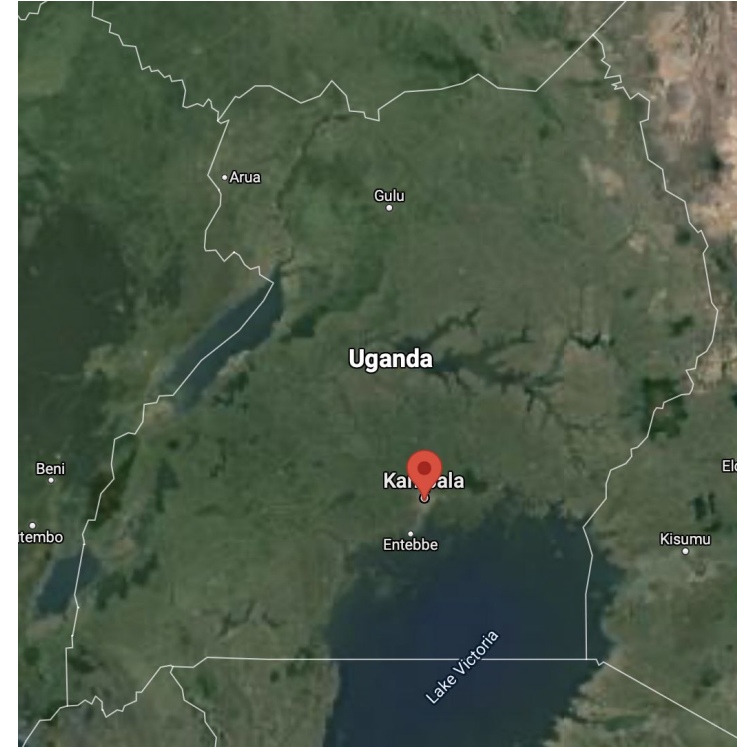


Predicting air quality using meteorological observations in Kampala, Uganda

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

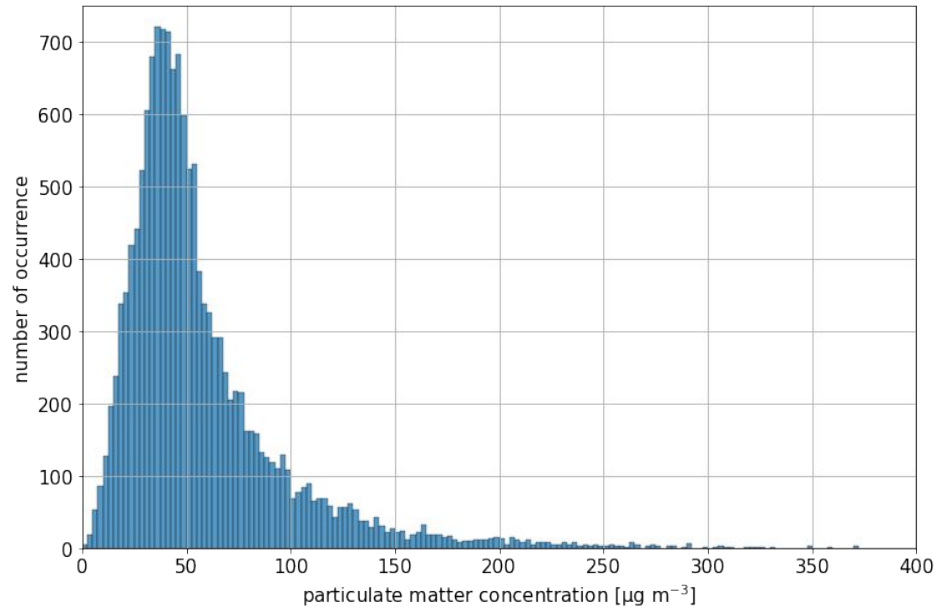
- **Ministry of Health**, which makes the prediction system available to various stakeholders and to the population

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

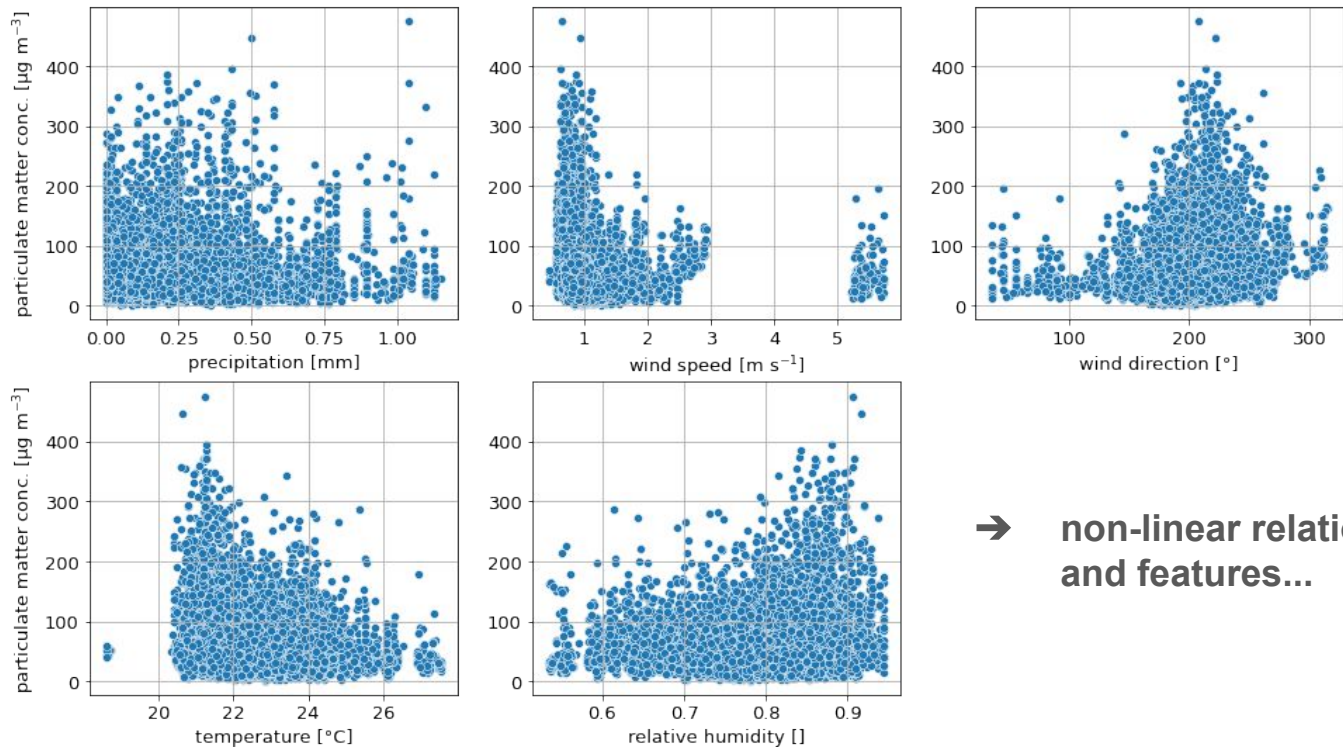
- **PM2.5 concentration [$\mu\text{g m}^{-3}$]** (particles with a diameter smaller than $2.5 \mu\text{m}$)



Health Concern	PM _{2.5} ($\mu\text{g m}^{-3}$)	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**

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

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

- **baseline model**

- > always predict the **mean particulate matter load** of **$58 \mu\text{g m}^{-3}$**

- > score: $\text{RMSE} = 42 \mu\text{g m}^{-3} \sim 72 \% \text{ of the mean particulate matter load}$

Models to predict particulate matter concentration

Model	RMSE [$\mu\text{g m}^{-3}$]	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 **not suitable** due to non-linear target-feature relationships

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

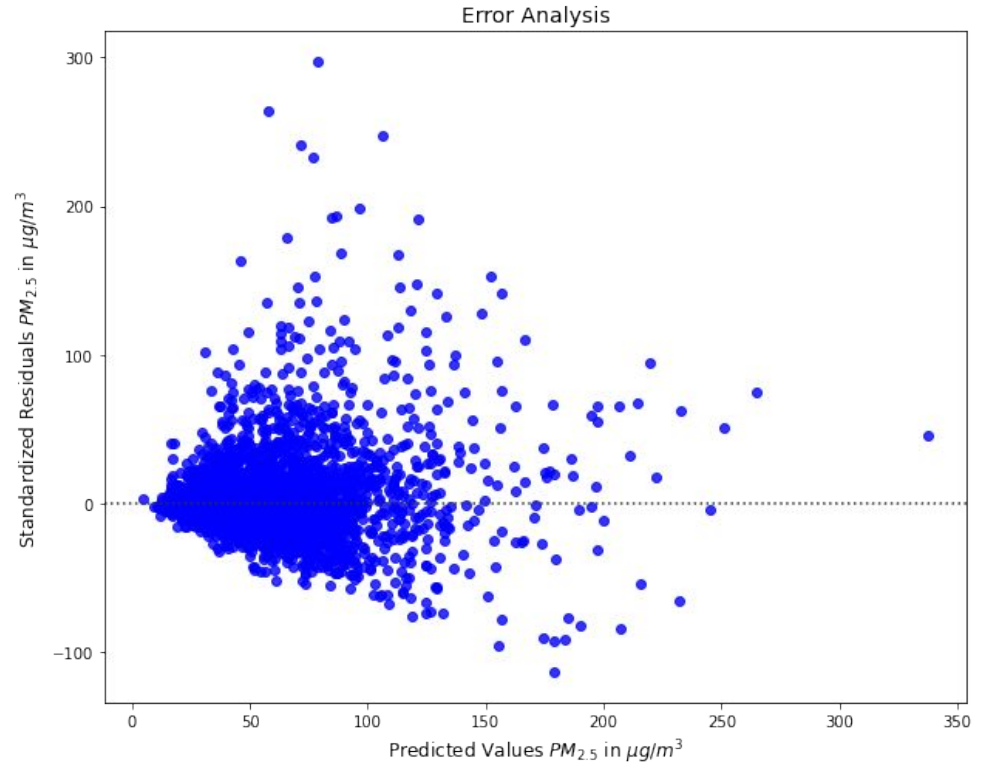
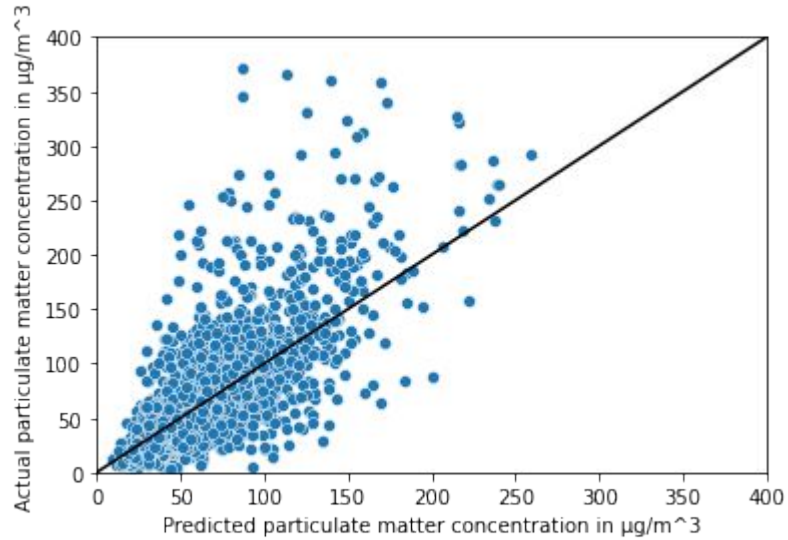
Summary

- **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 $\mu\text{g m}^{-3}$**
 - ~ 48 % of mean particulate matter load
 - ~ deviation of two categories in the upper part of guideline on hazardous levels

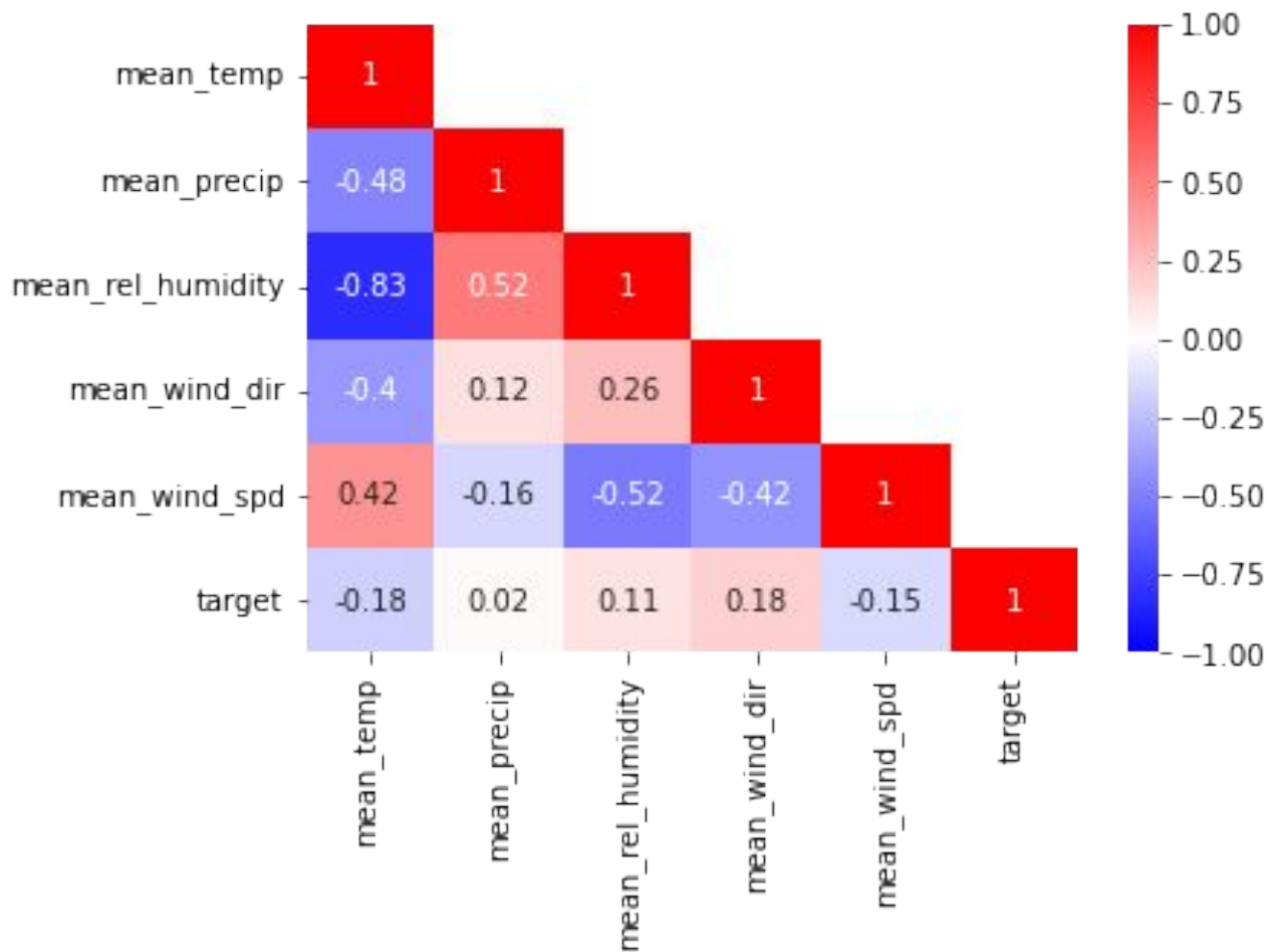
Outlook

- 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**

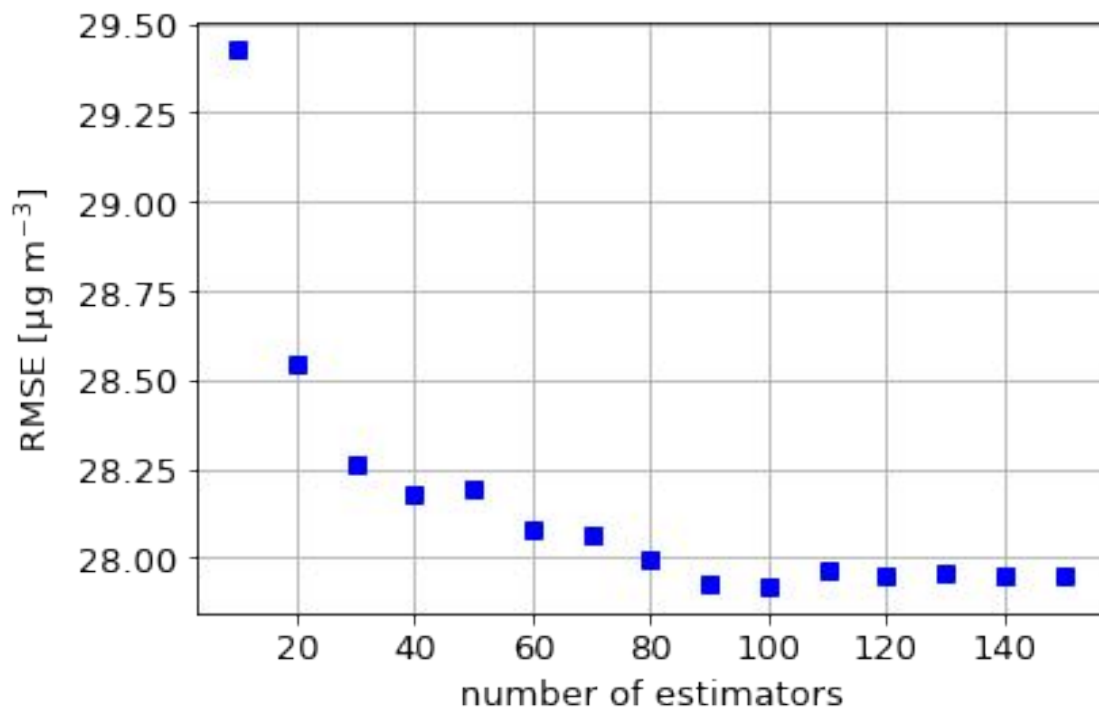
Appendix



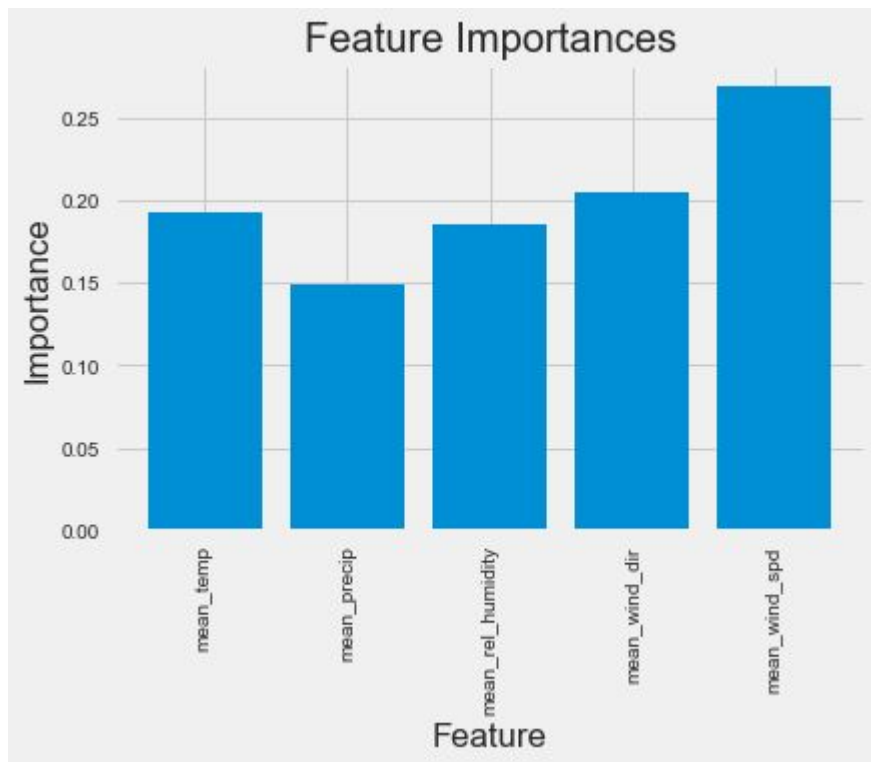
Appendix



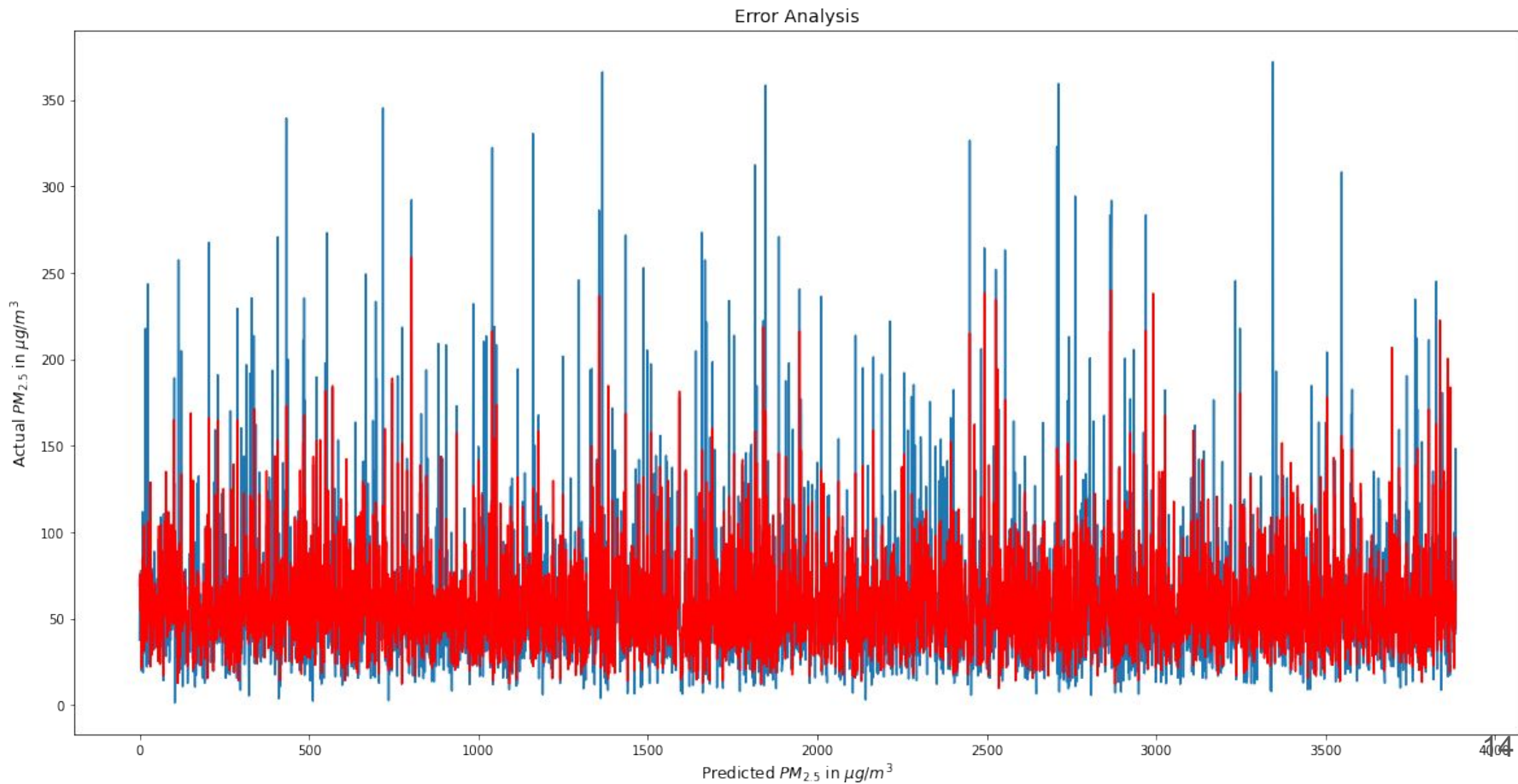
Appendix



Appendix



Appendix



Appendix

