World Health Organization
African Region

For the African Governments

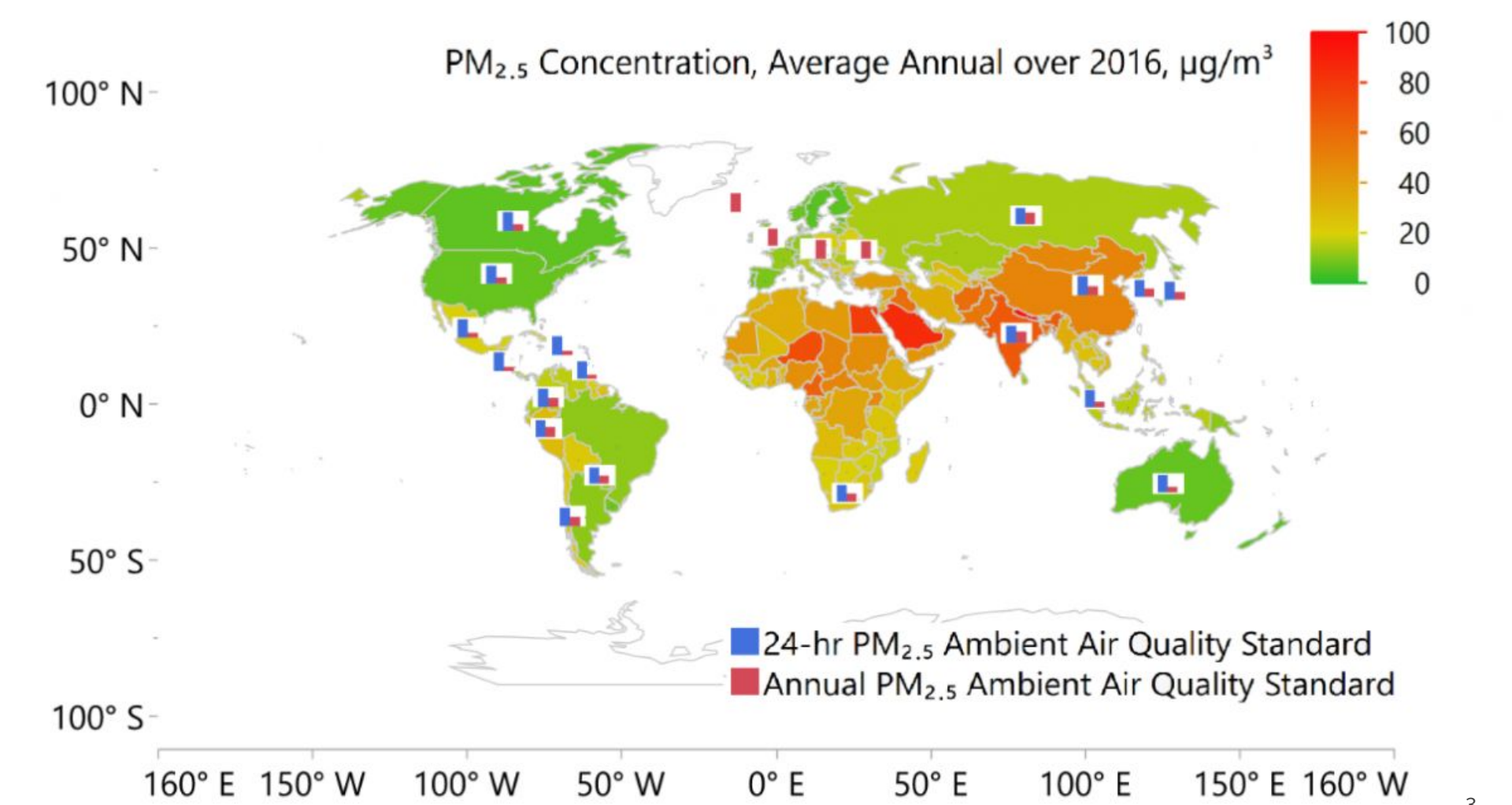
Air Pollution Campaign

Current Situation

→ African countries are dealing with increased air pollution

→ PM2.5: Key indicator of air quality (as less as possible)

→ Increase risk of health problems (asthma, heart disease, low birth weight)



https://mcgill.on.worldcat.org/

About the project

Audience:
- cities and municipalities
- raising awareness on air pollution and potential health risk

Value of Product:
- determination/monitoring of ground level air quality based on satellite data of air pollutants without need of expensive measurement equipment
- suitable for developing African countries

Prediction:
- air quality measure (particle size > 2.5nm) based on satellite data of pollutants

Evaluation Metric: RMSE

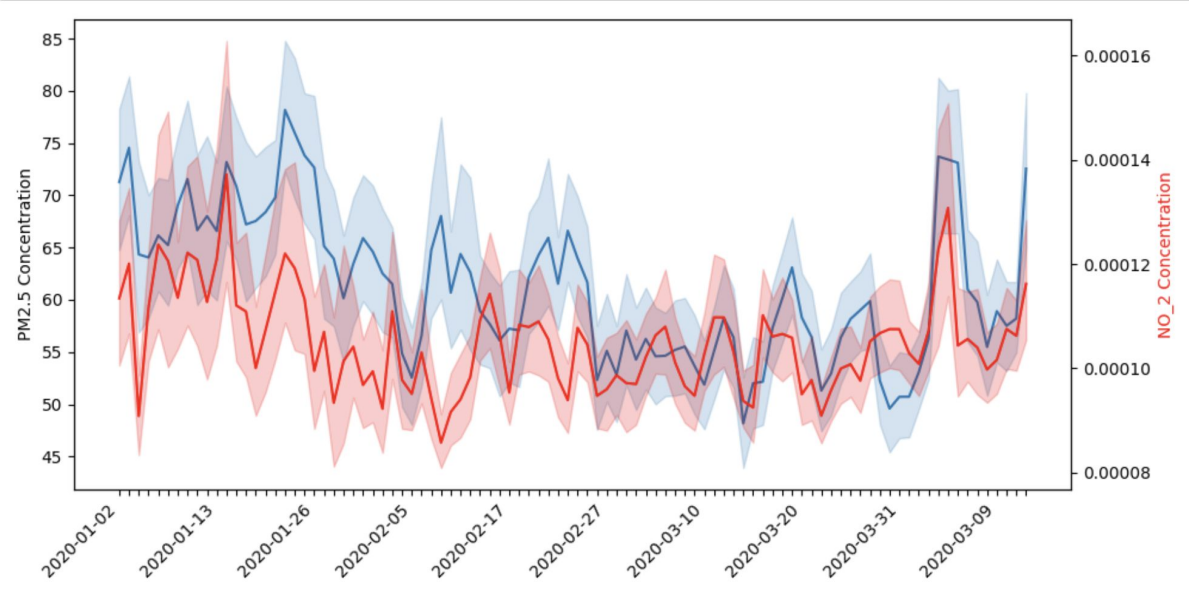
Baseline Model:
- wind magnitude and air pollution have negative linear relationship

Steps

1.	Preprocessing Data	Tobi
2.	Baseline Model	Seda
3.	Predictive Models	Hannes

Understanding the data structure

- Target columns: ground measurements of air quality
- Feature columns: satellite data of air pollutants
- Acquisition time: 3 months
- Location of Acquisition: various cities in Africa



AIM

Predict the blue time series (line) based on multiple other time series

Data Cleaning

- Feature columns **> 80%** missing values: exclude from analysis
- Feature columns **10 - 30 %** missing values: use location specific imputation strategy

- We did not exclude any columns from the model with useless information

Feature Engineering

- log-normal distributions — log-transformation, Min-Max scaler
- normal distributions — Min-Max scaler

Selected Baseline Model

- Poisson Regression

Assumption

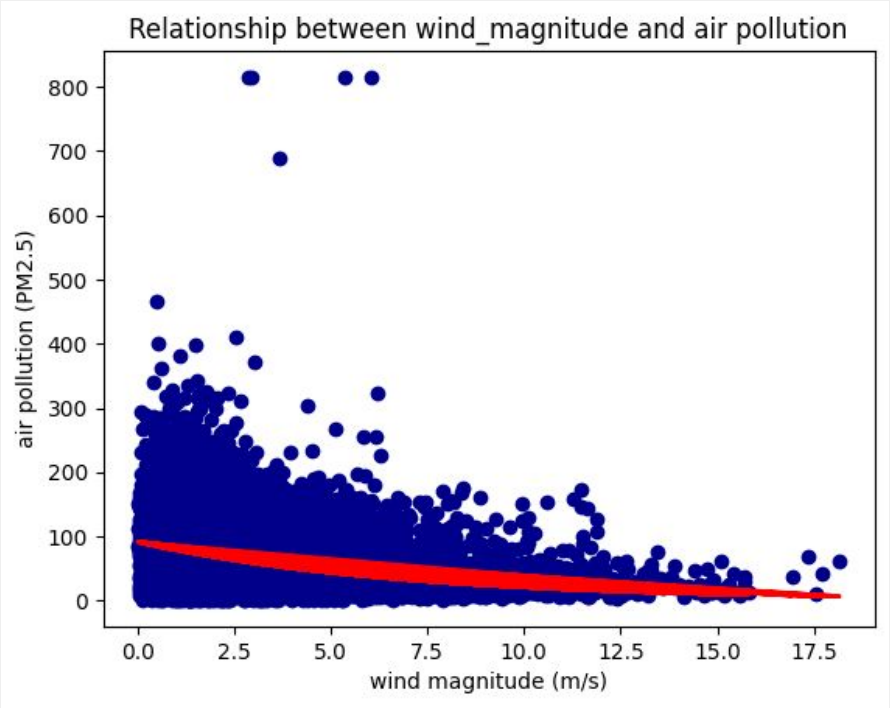
- Wind blows dust/pollution particles away
- Less wind magnitude → higher PM2.5 concentration

Evaluation Metric

- **RMSE (Train) = 43.81**
- **RMSE (Test) = 42.68**

Why RMSE?

- easy to interpret
- prediction accuracy
- performance comparison of different models



Selected advanced model: Ensemble Methods

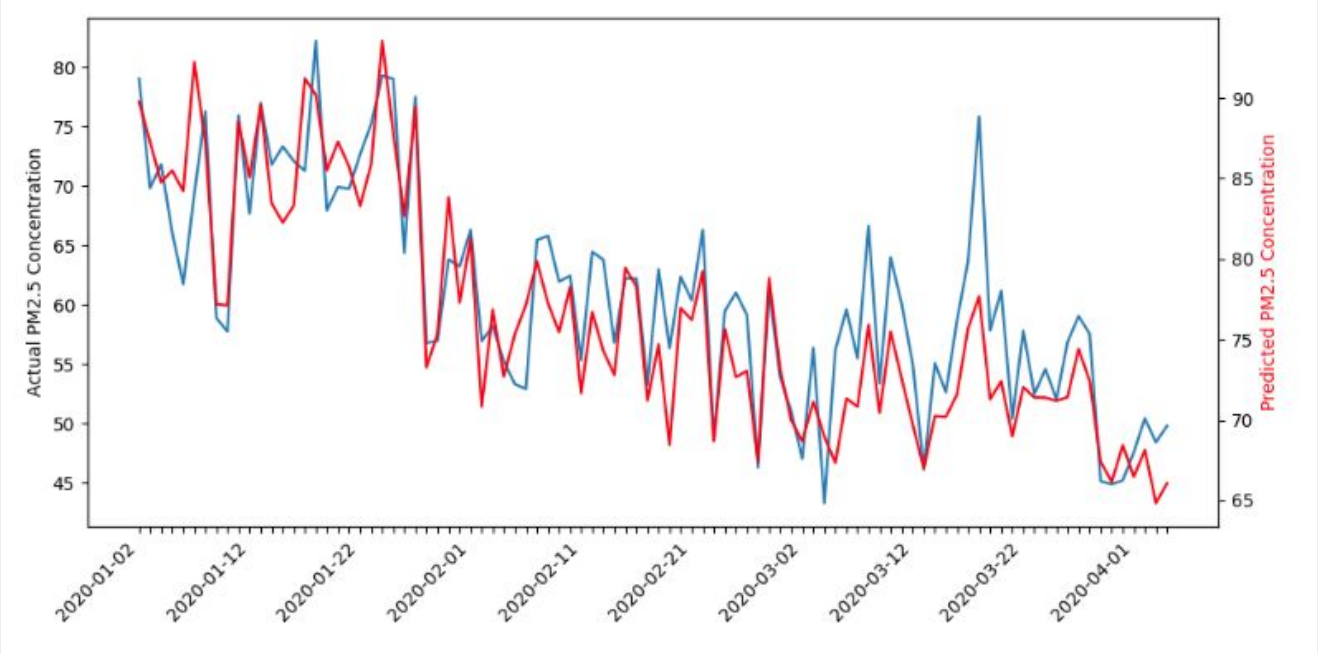
- XGBoost
- Random Forest

Error

- RMSE = 32.65

Features

n = 70



Conclusions

Most important features:

- no2 sensor azimuth angle
- so2 column density

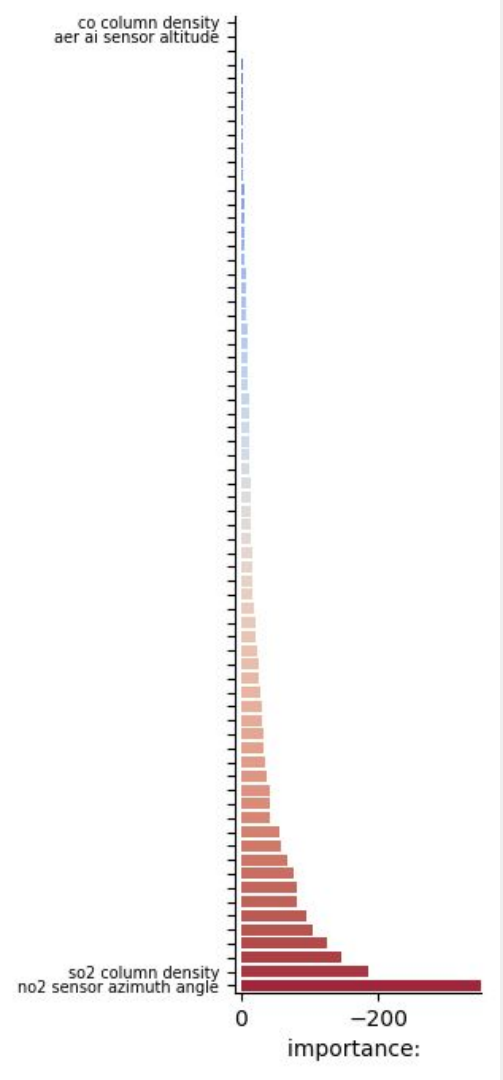
Least important features:

- aer ai sensor altitude
- co column density

=> possibility to provide recommendation which pollutants are critical for air quality

Our predictions have improved compared to baseline model

Features with sensor data should be imputed with different strategy



Thank you for your attention

Seda
Tobias
Hannes
Priscila

References

- https://zindi.africa/competitions/zindiweekendz-learning-urban-air-pollution-challenge
- Foto: https://mcgill.on.worldcat.org/