Urban Air Pollution

By Hanna Schaumberger, Fanxing Xi, Karol Palczynski

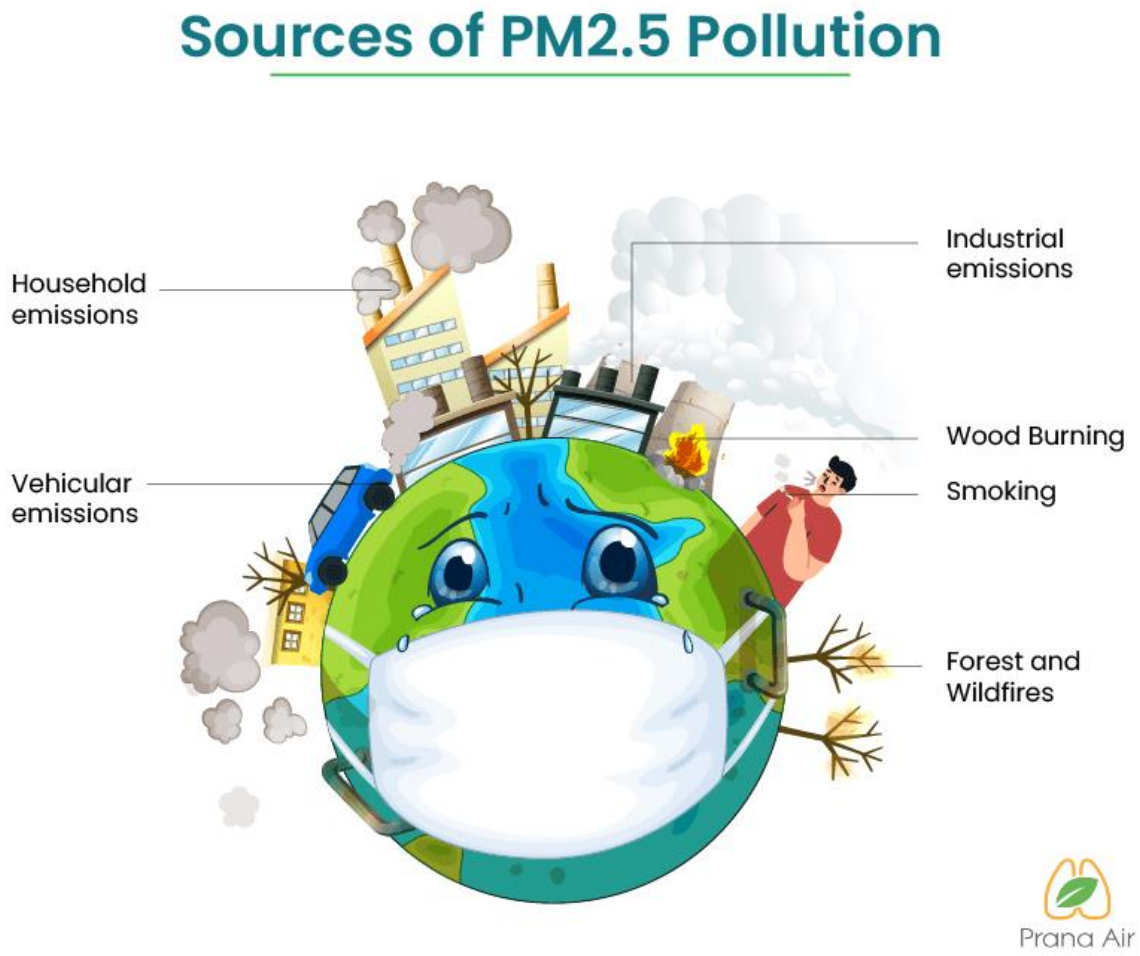
ABOUT US

- We are the scientific advisory board (Task-Force) to fight air pollution around the world
- We are presenting our findings to the politicians from the G12 Summit (without Putin)
- Our Mission is to promote air pollution awareness for citizens and provide a unified and world-wide air quality information

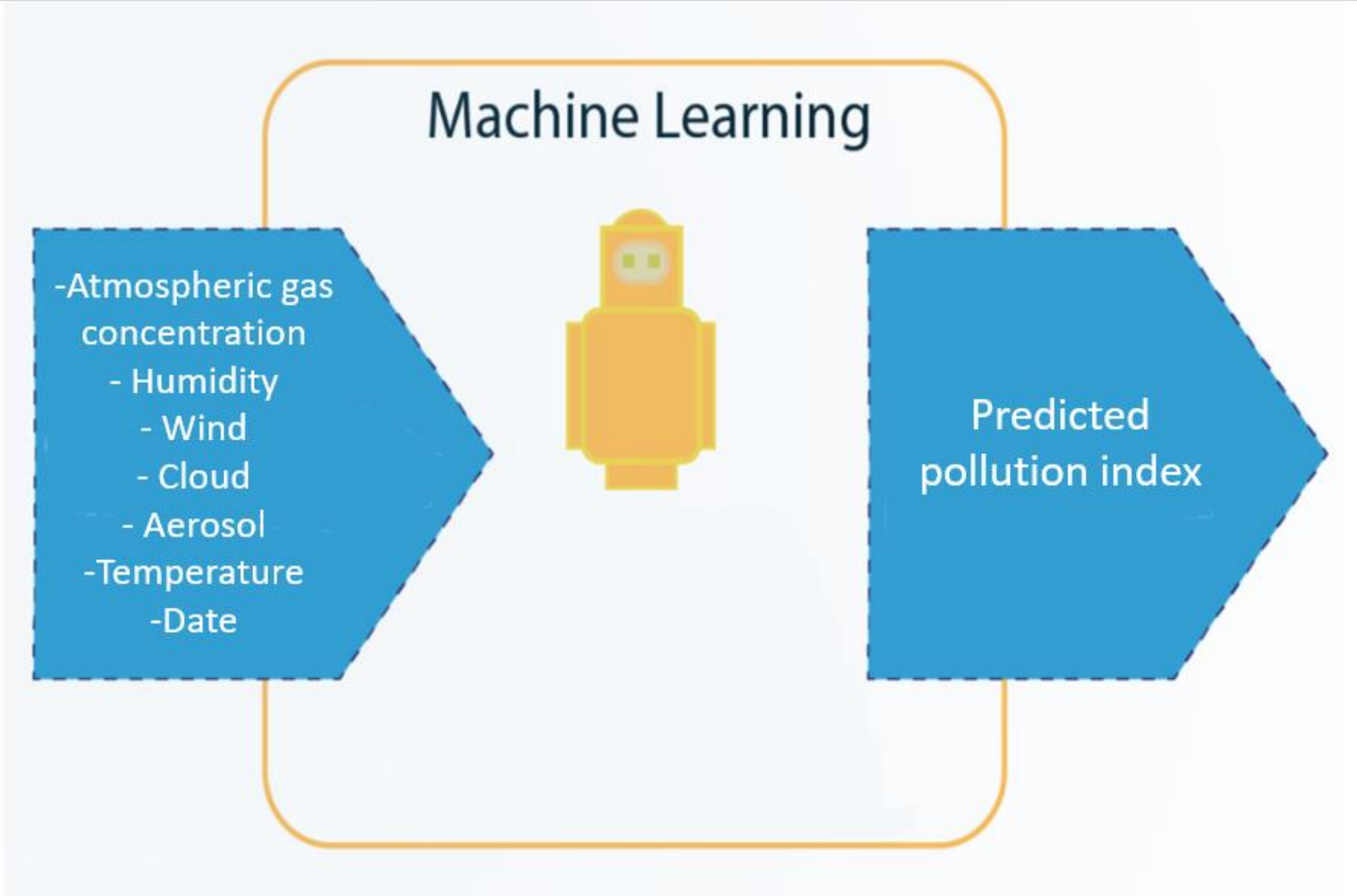
Introduction

- Goal: To predict how air quality (PM2.5 particulate matter concentration) changes in places where we don’t have ground-based sensors for measuring
- Database : We’ve collected weather data and daily observations from the Sentinel 5P satellite tracking various pollutants in the atmosphere. The data covers the last three months, spanning hundreds of cities across the globe.

What is PM2.5?

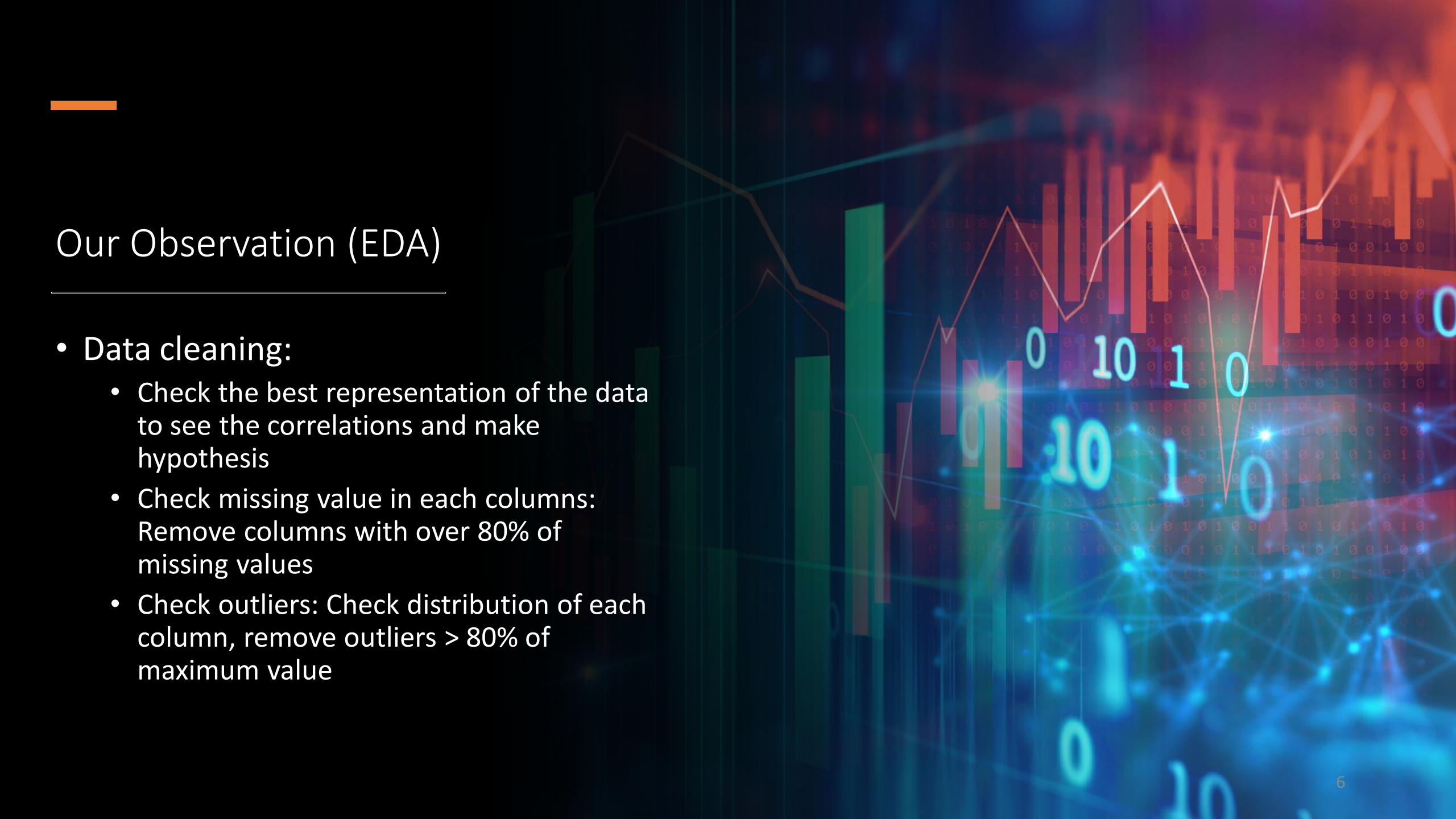

Sources of PM2.5 Pollution
Household emissions
Vehicular emissions
Industrial emissions
Wood Burning
Smoking
Forest and Wildfires
Prana Air

Methodology



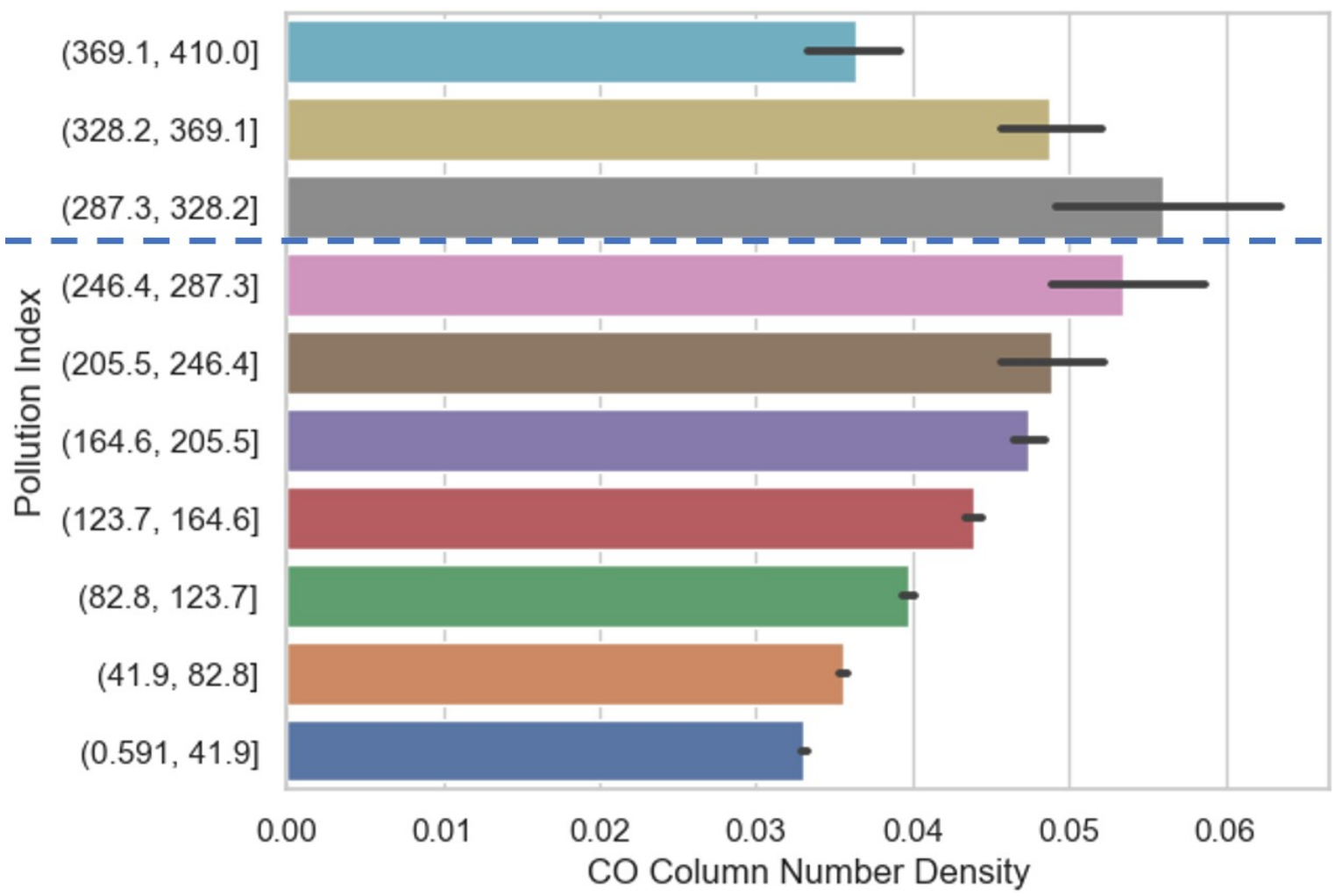
Our Observation (EDA)

- Data cleaning:
  - Check the best representation of the data to see the correlations and make hypothesis
  - Check missing value in each columns: Remove columns with over 80% of missing values
  - Check outliers: Check distribution of each column, remove outliers > 80% of maximum value

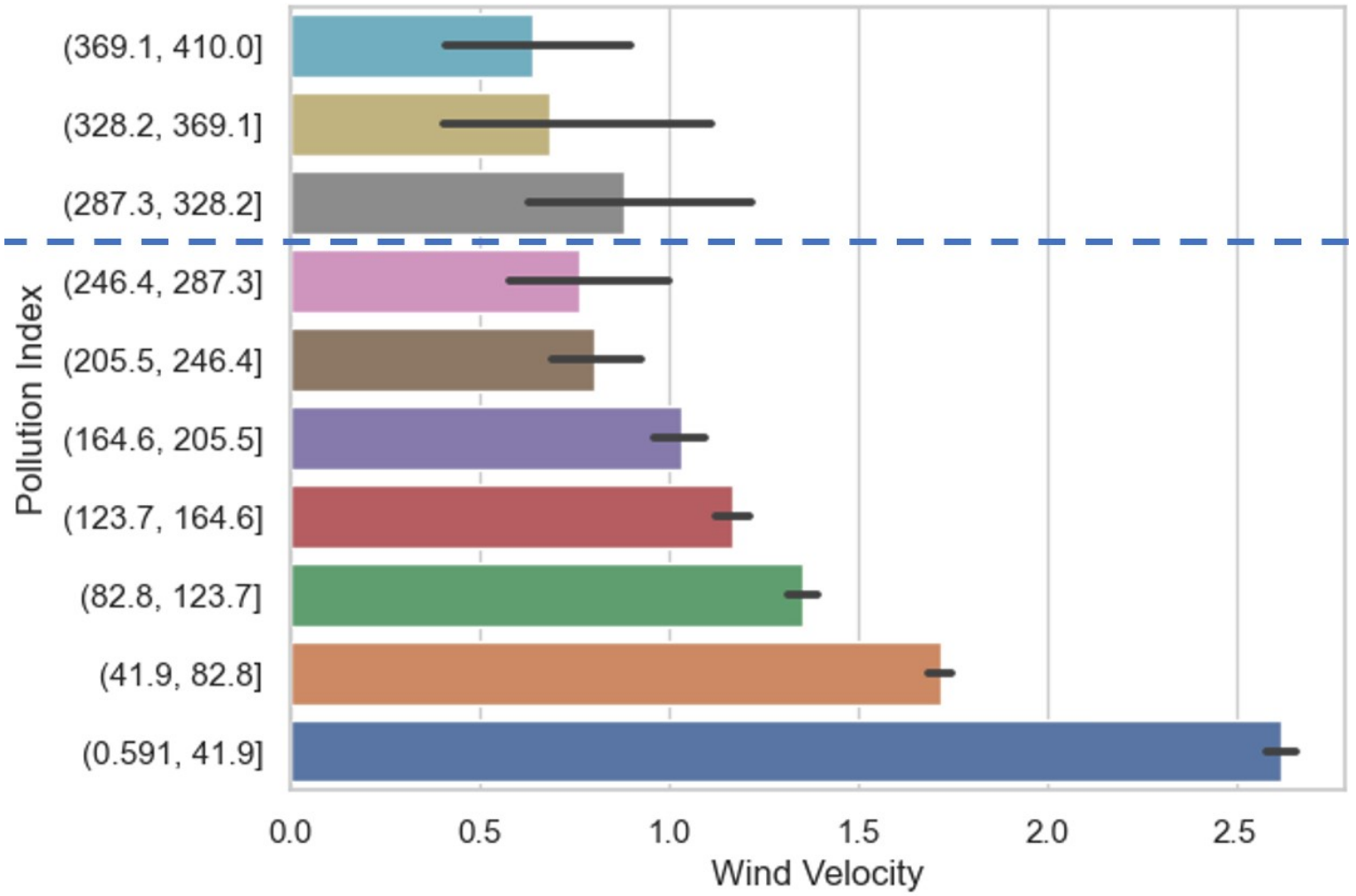


Hypothesis

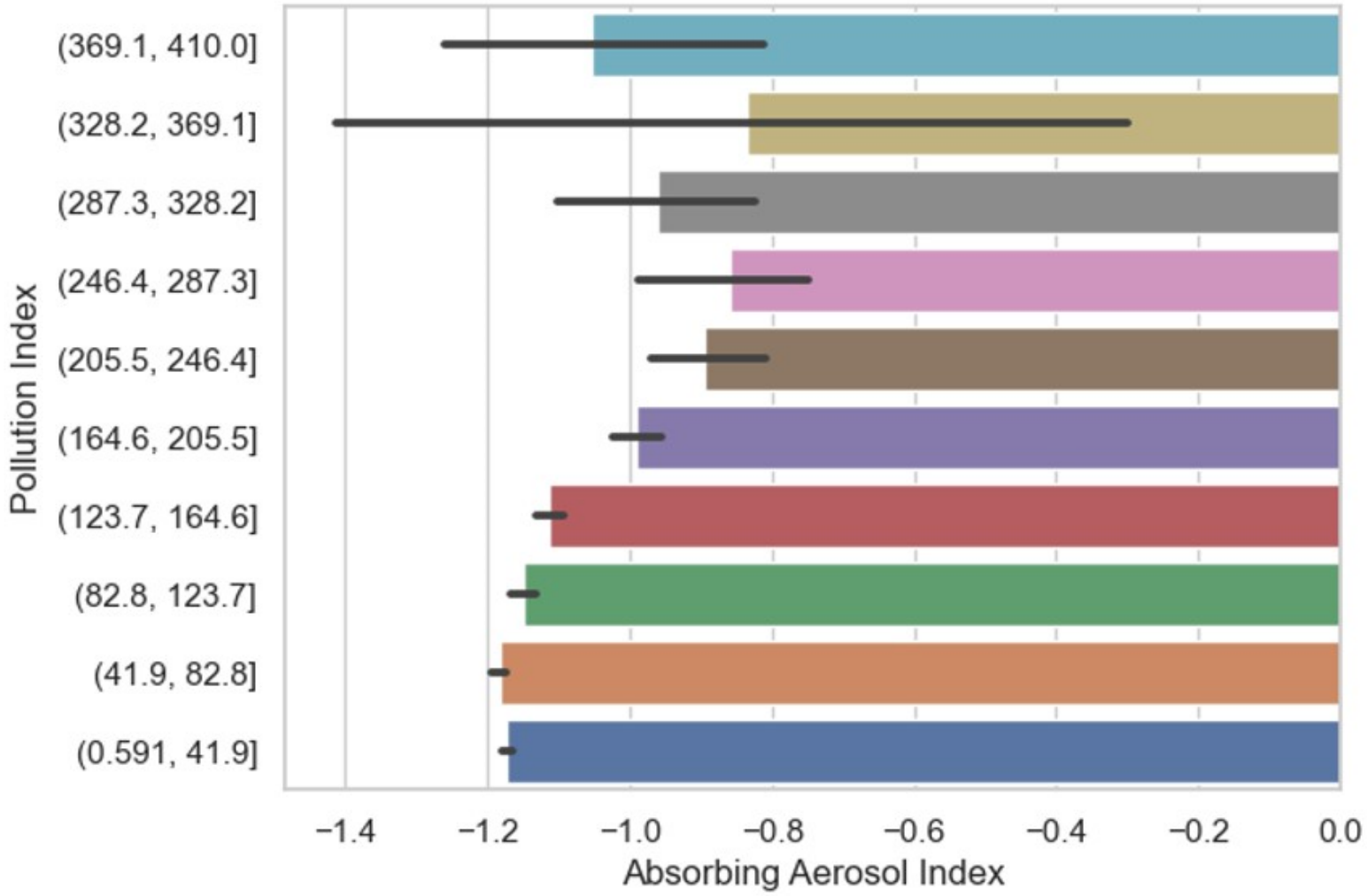
1. **Higher gas column densities, higher target value**
2. **Higher wind velocity, lower target value?**
3. **Aerosol index with target value?**
4. Sensor altitude with target value?
5. Satellite angle correlates with the target value
6. Stratospheric column density no correlation with target value
7. Aerosol index with target value
8. Humidity correlated non-linearly with target value



1. Higher gas column densities, higher target value



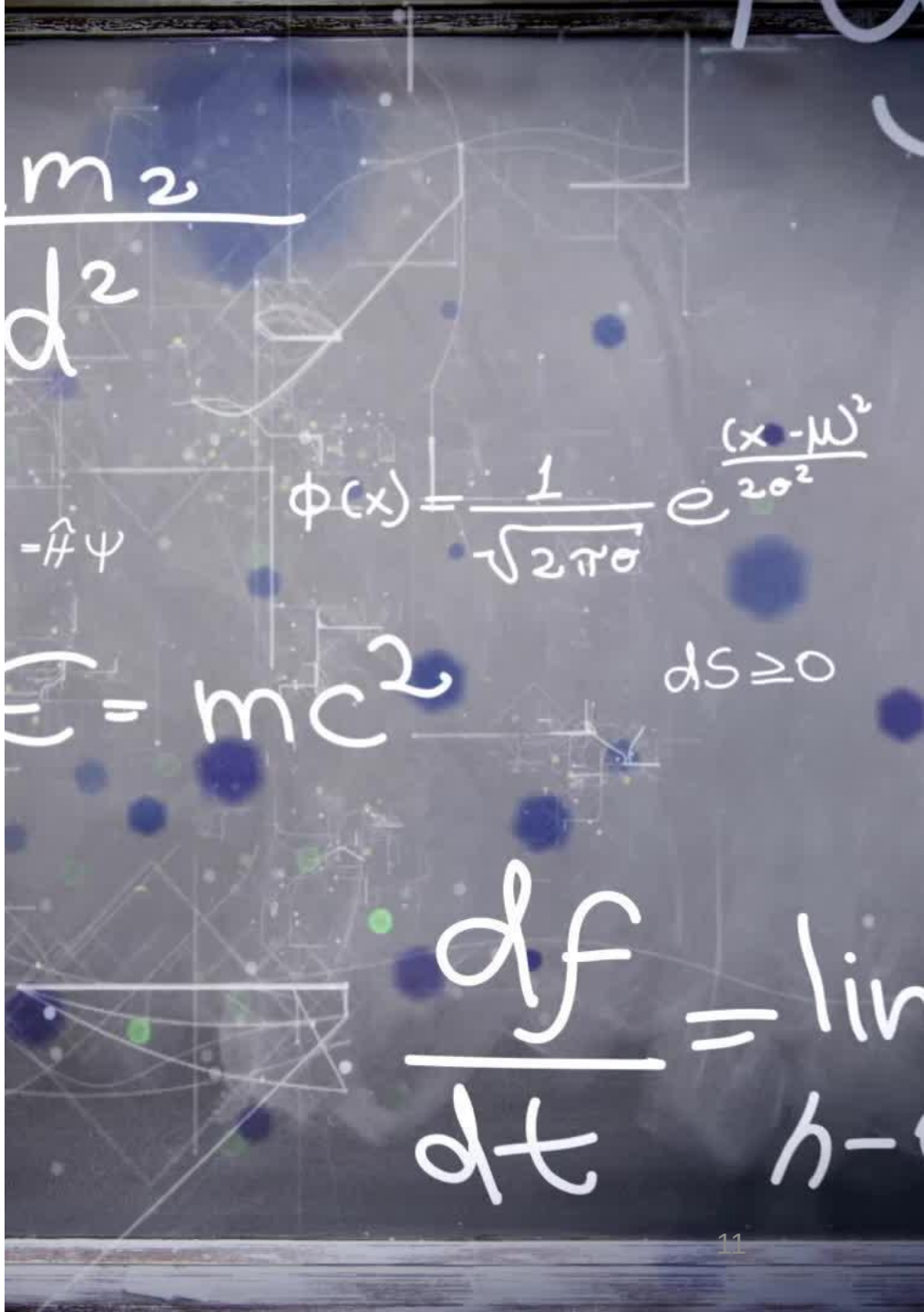
2. Higher wind velocity, lower target value?



3. The higher Aerosol index the higher the target value?

Machine learning Model

1. Impute and Scale
2. Select models: Different regression models were tested; Random forest, LGBM regressor, XGB regressor selected based on their performances
3. Stacking regressor: combining selected models with linear regression
4. Grid Search for best parameters: no improvement within limited time

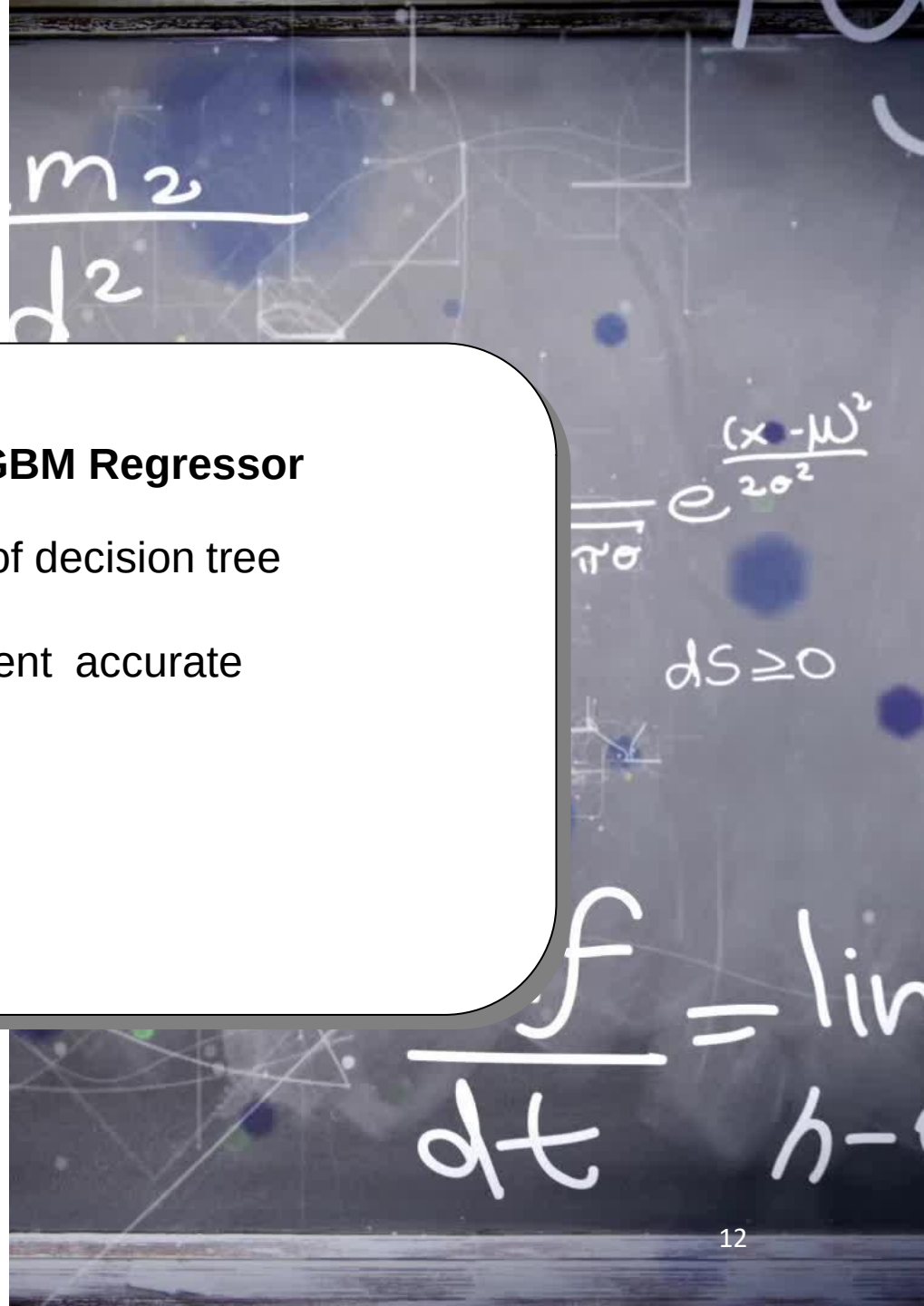


Machine learning Model

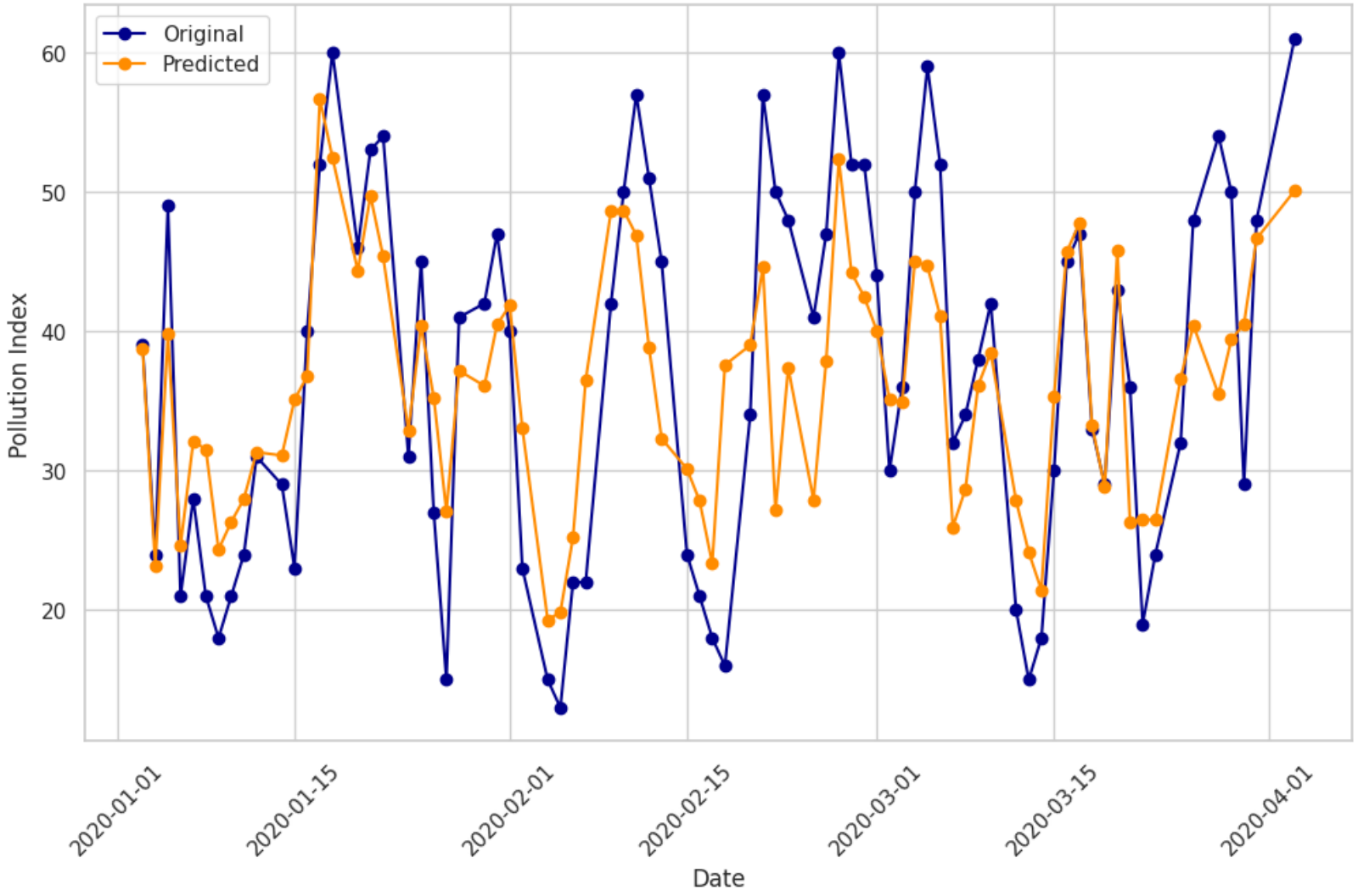
1. Impute and Scale
2. Select models: Different regression models were tested; Random forest, **LGBM regressor**, XGB regressor selected based on their performances
3. Stacking regressor: combining selected models with linear regression
4. Grid Search for best parameters: no improvement within limited time

Light GBM Regressor

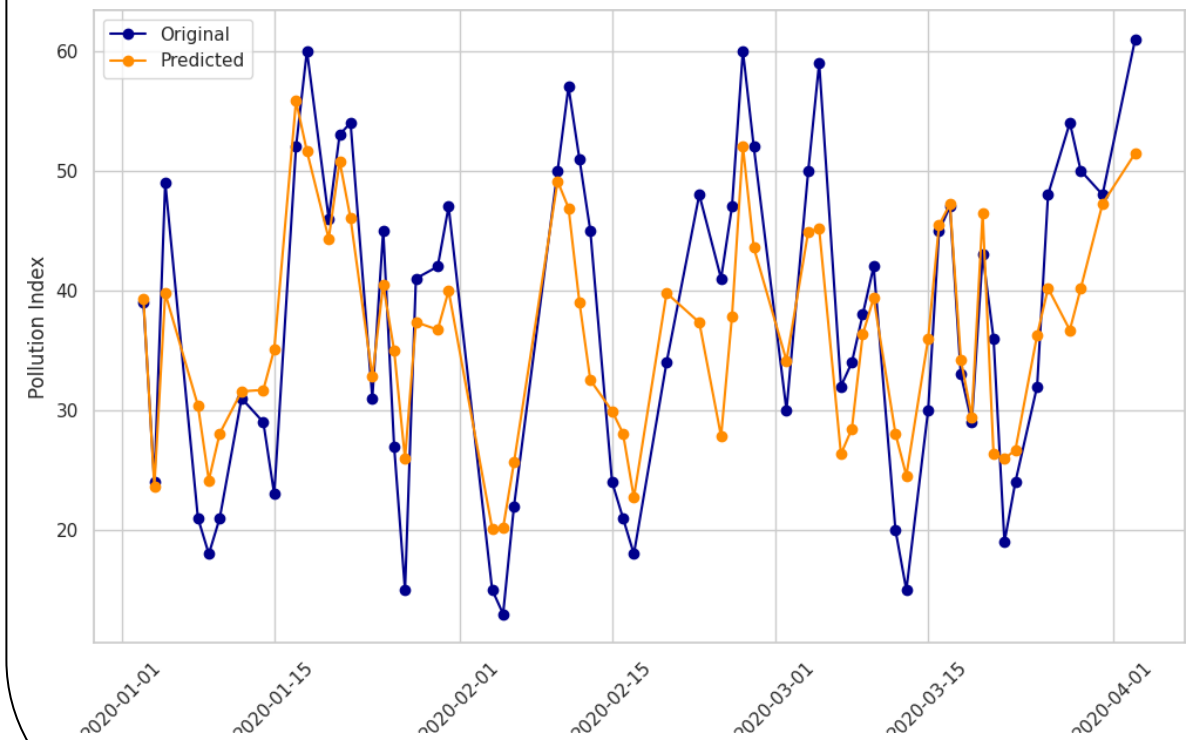
- New type of decision tree
- Fast, efficient accurate
- Parallel

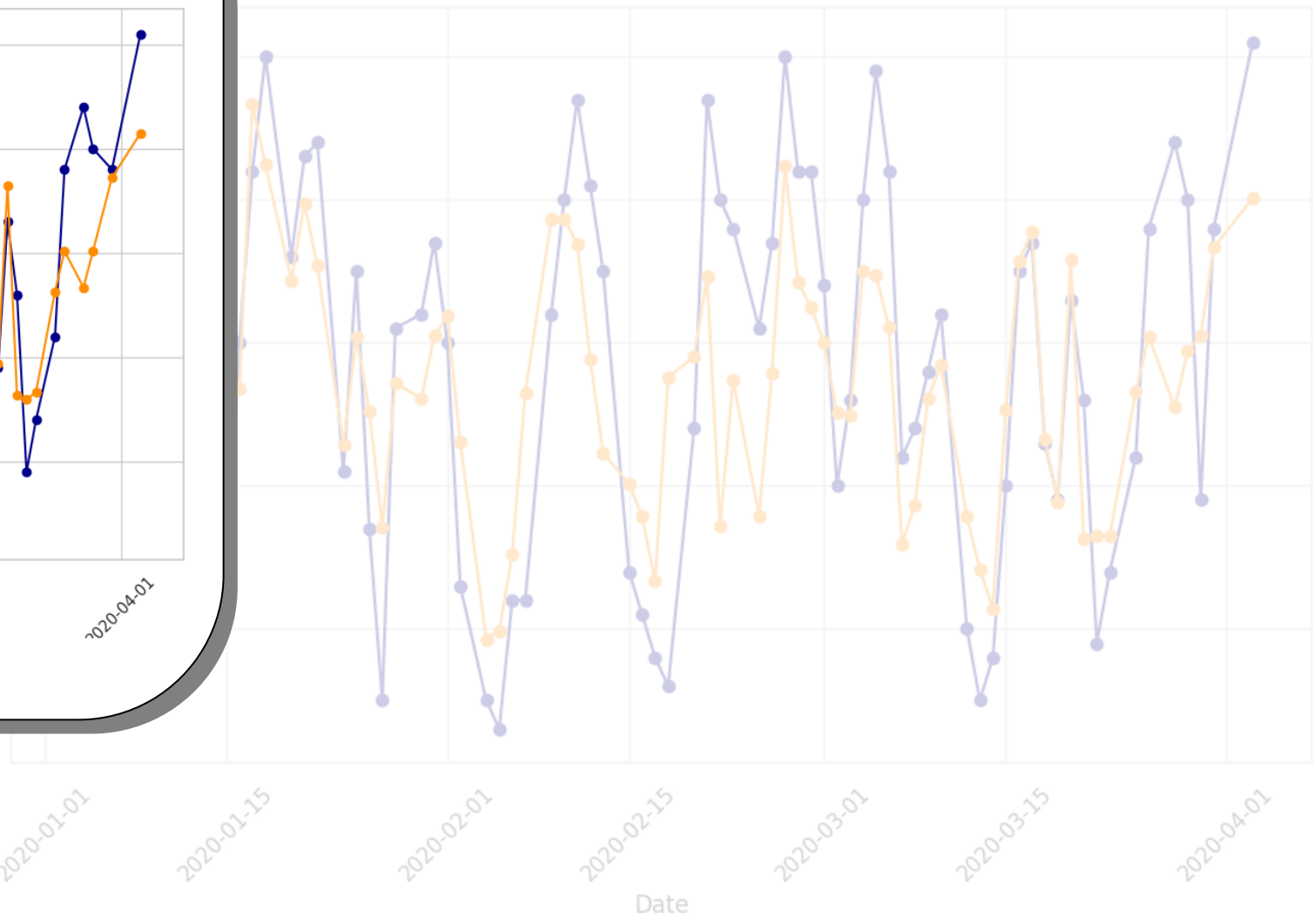


Result 1: Pollution over time in one Place-ID



Training data set

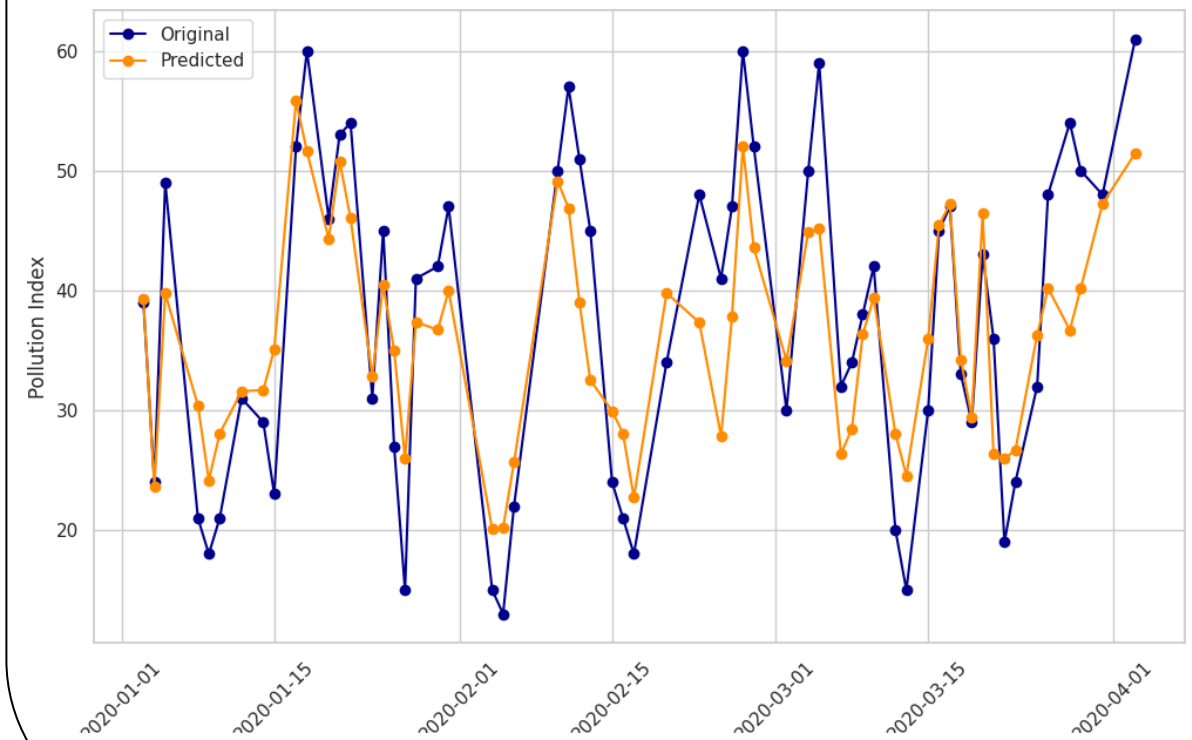




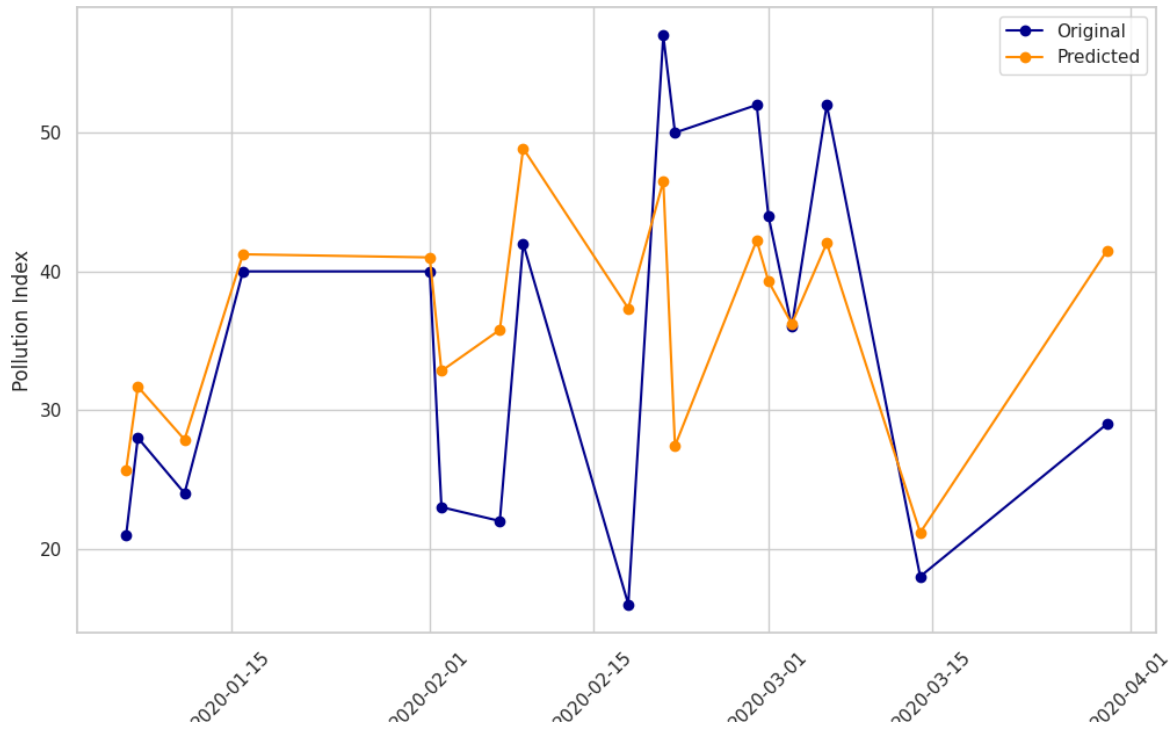
R2-Score: 0.91

MSE: 212

Training data set

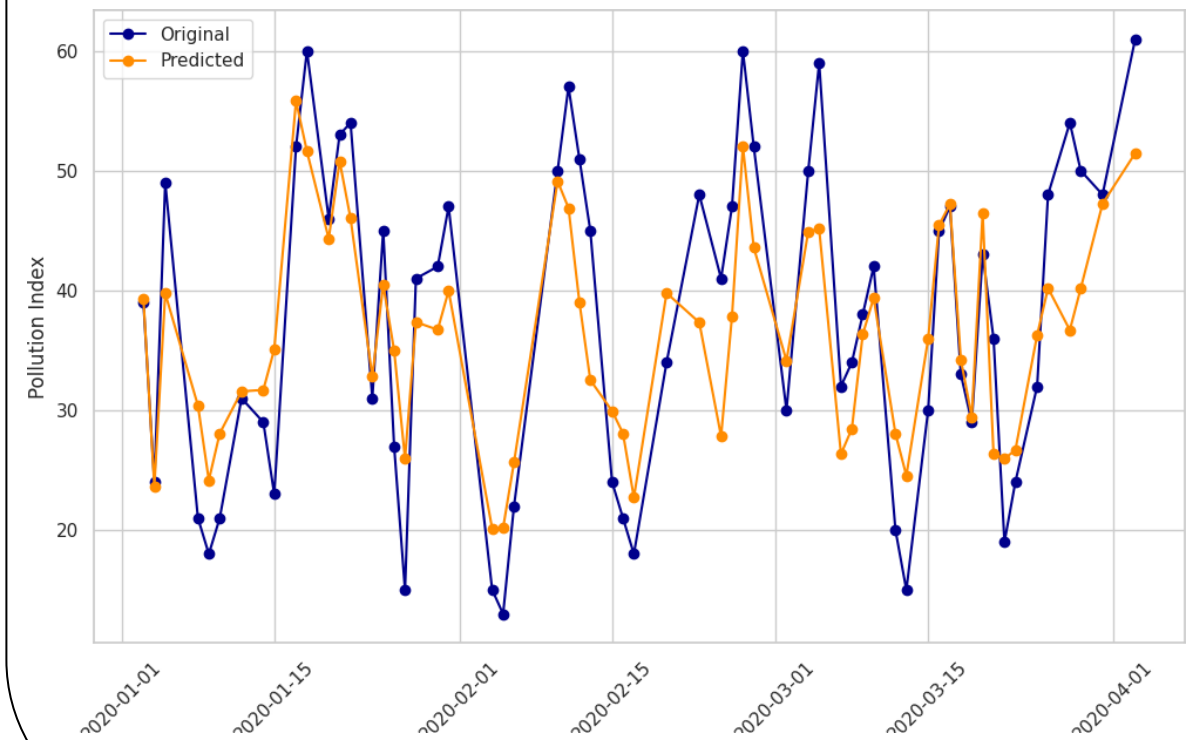


Test data set

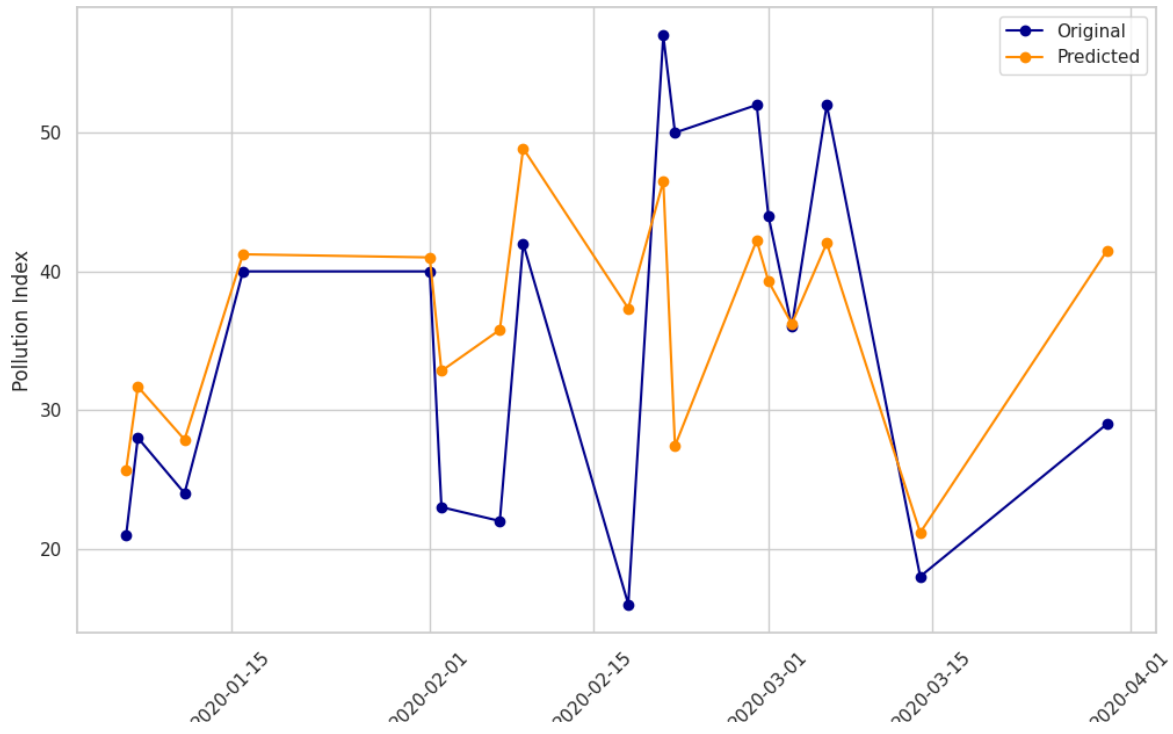


R2-Score:	0.91	0.68
MSE:	212	682

Training data set

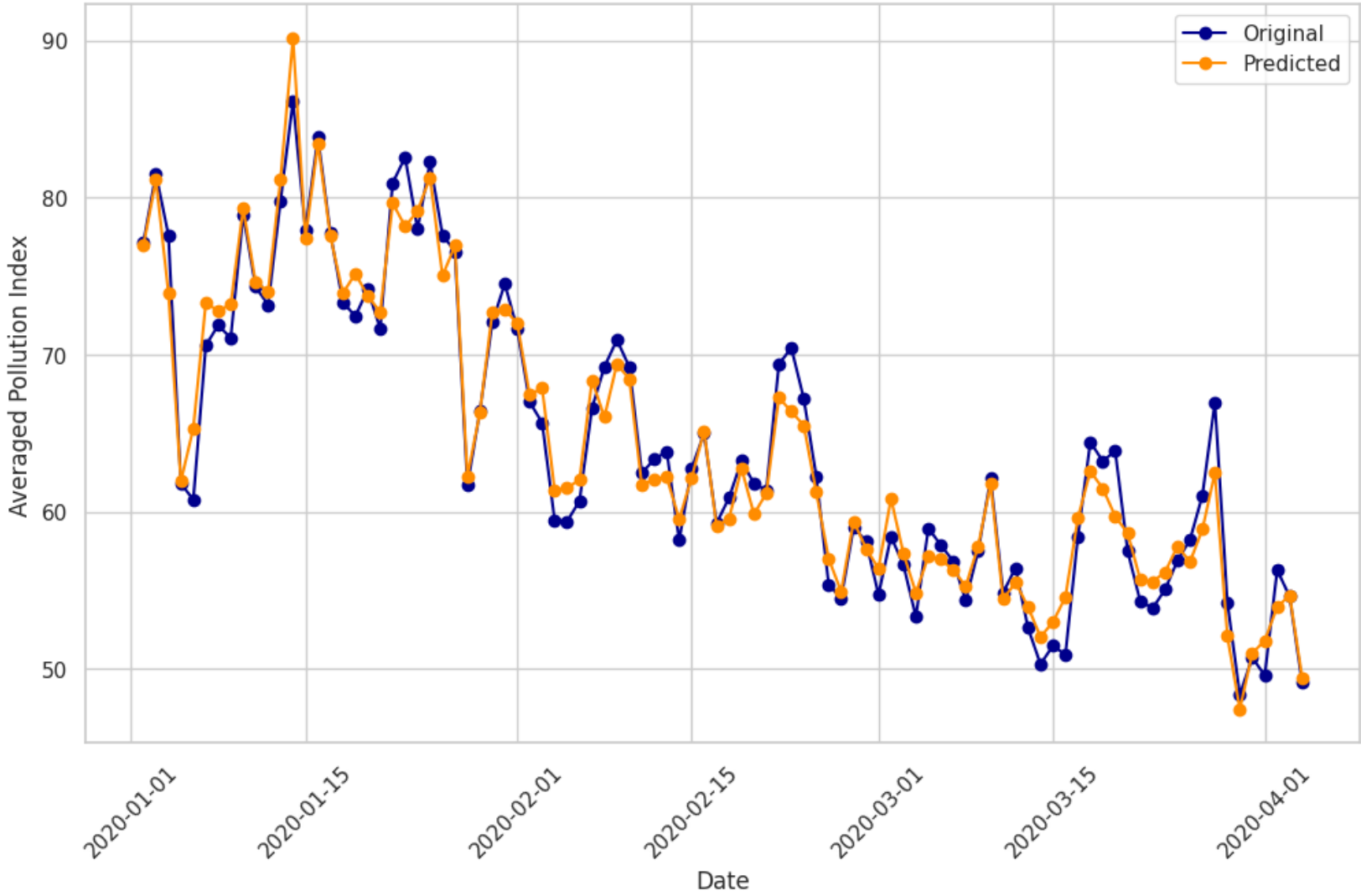


Test data set

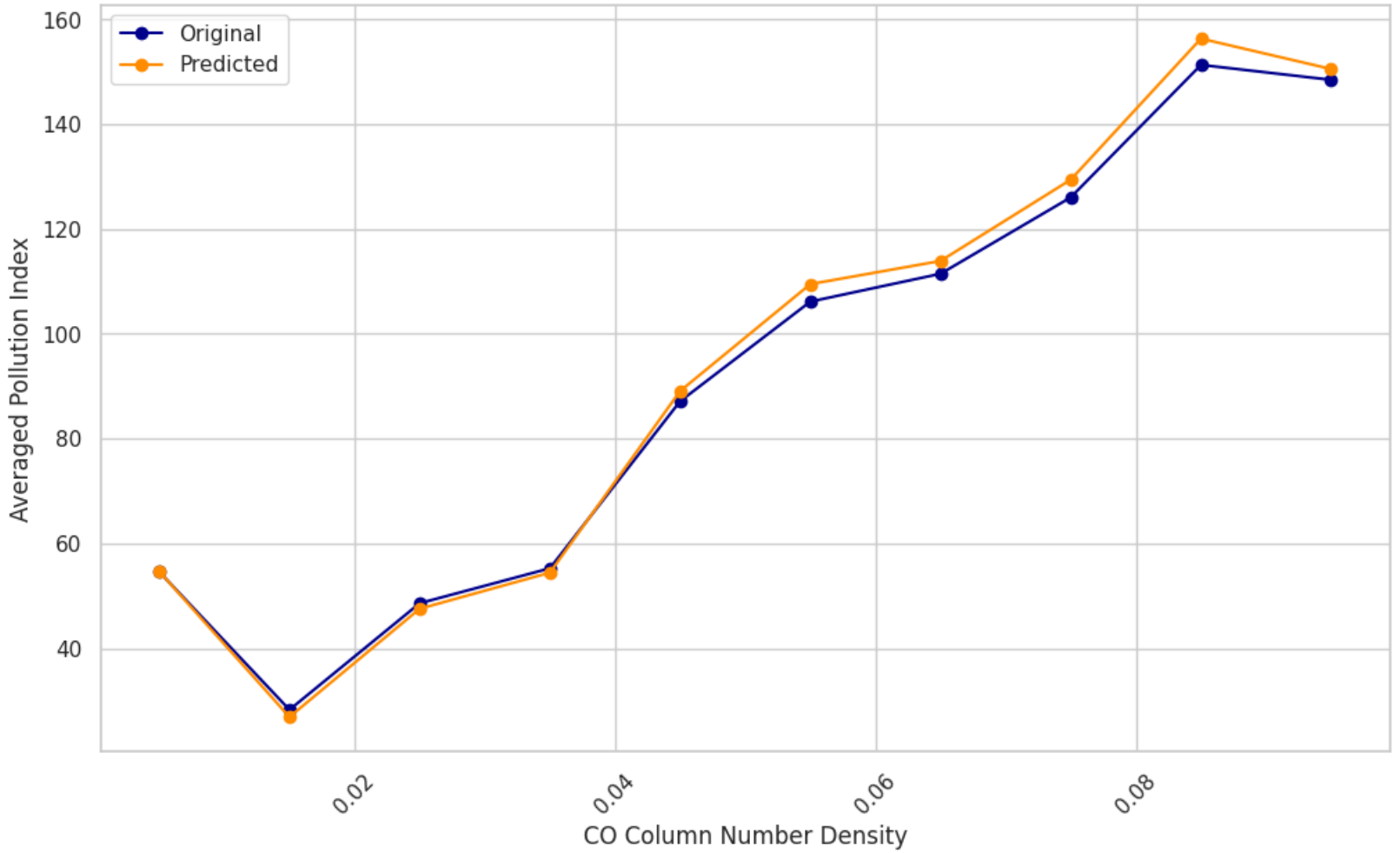


R2-Score:	0.91		0.68
MSE:	212		682
		R2-Score:	0.34
Baseline model:			Lin.Reg. with column density features
		MSE:	1416

Result 2: Pollution over time averaged over all Place-IDs



Result 3: Bin-averaged Pollution vs. Co



Baseline model:	R2-Score:	0.34
	MSE:	1416
Main model:	R2-Score:	0.68
	MSE:	682

Conclusion

Predict pollution index basing on daily weather and Sentinel 5P satellite data without the need of ground-based sensors

Thank you for your attention