Algorithms for the Spatial Interpolation of Environmental Data

Bachelor Thesis

Simón E. Sánchez Viloria

Leganés, July 2022

Advisor: Harold Molina Bulla PhD.

Outline

1. Introduction

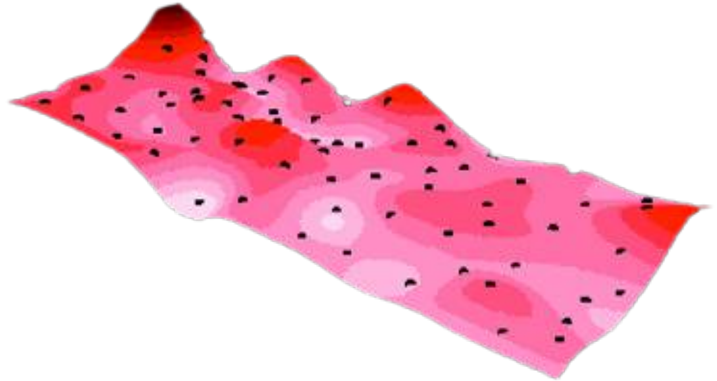
- Introduction
- Methodology
- Results
- Conclusions

Spatial Interpolation

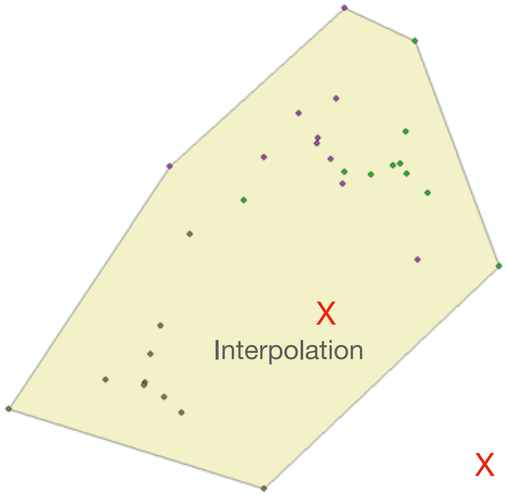
1. Introduction

“Everything is related to everything else, but near things are more related than distant things”

- Waldo Tobler’s "First Law of Geography” (1970)



- Data Imputation
- Surface Models
- Extrapolation



X extrapolation

Related Work

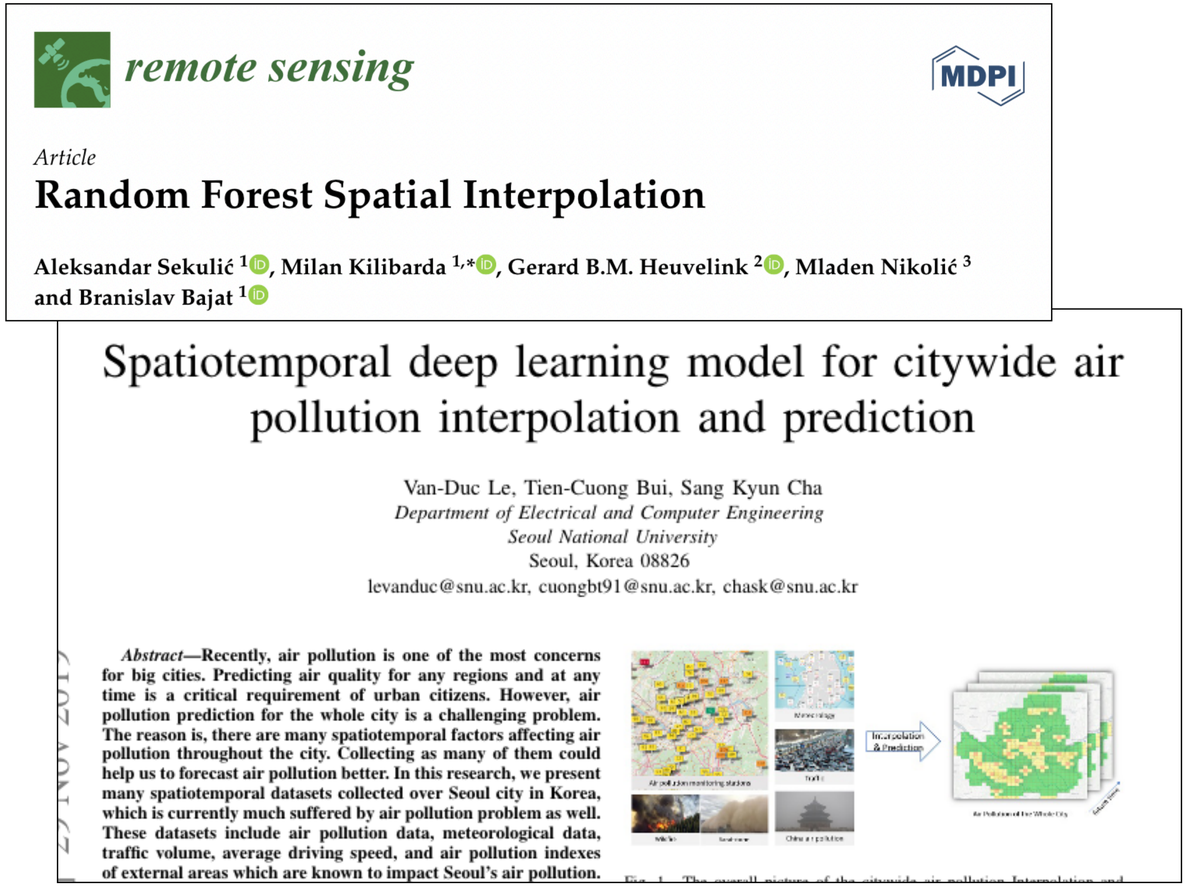
1. Introduction

Many methods have been used in the literature with the specific purpose of spatial interpolation / estimating surface models

Common Approaches:

- Gaussian Process Regression (Kriging)
- Deterministic methods

Kriging in the most popular method for environmental data



remote sensing

MDPI

Article

Random Forest Spatial Interpolation

Aleksandar Sekulić [1], Milan Kilibarda [1,*], Gerard B.M. Heuvelink [2], Mladen Nikolić [3] and Branislav Bajat [1]

Spatiotemporal deep learning model for citywide air pollution interpolation and prediction

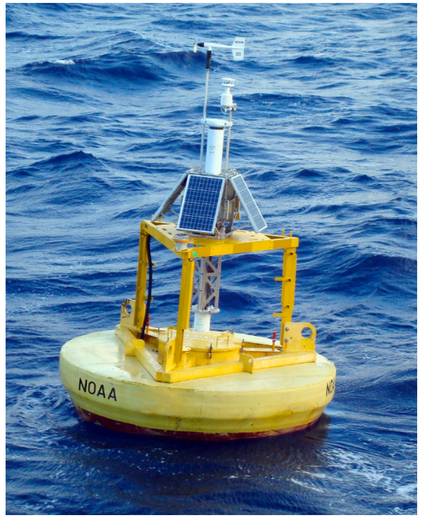
Van-Duc Le, Tien-Cuong Bui, Sang Kyun Cha
Department of Electrical and Computer Engineering
Seoul National University
Seoul, Korea 08826
levanduc@snu.ac.kr, cuongbt91@snu.ac.kr, chask@snu.ac.kr

***Abstract*—Recently, air pollution is one of the most concerns for big cities. Predicting air quality for any regions and at any time is a critical requirement of urban citizens. However, air pollution prediction for the whole city is a challenging problem. The reason is, there are many spatiotemporal factors affecting air pollution throughout the city. Collecting as many of them could help us to forecast air pollution better. In this research, we present many spatiotemporal datasets collected over Seoul city in Korea, which is currently much suffered by air pollution problem as well. These datasets include air pollution data, meteorological data, traffic volume, average driving speed, and air pollution indexes of external areas which are known to impact Seoul's air pollution.**

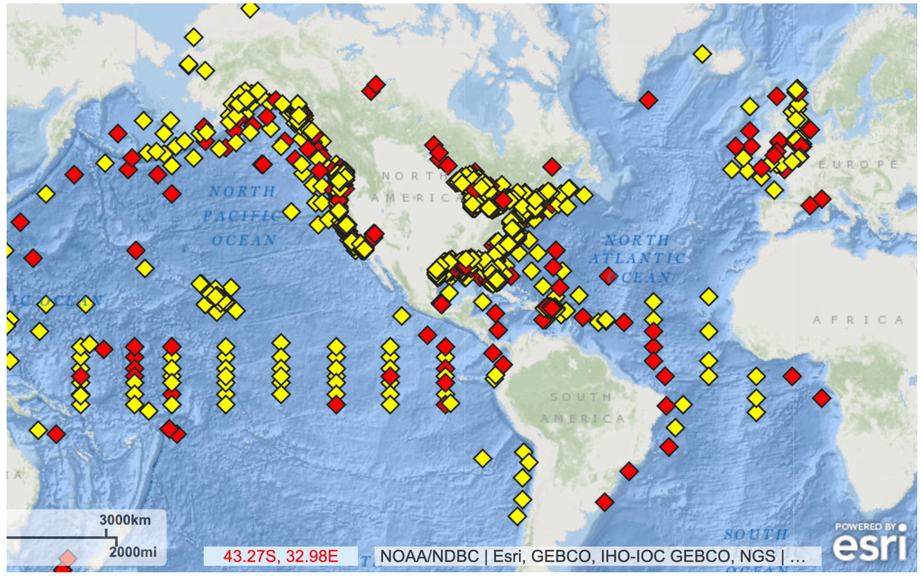
NOAA's NDBC Dataset

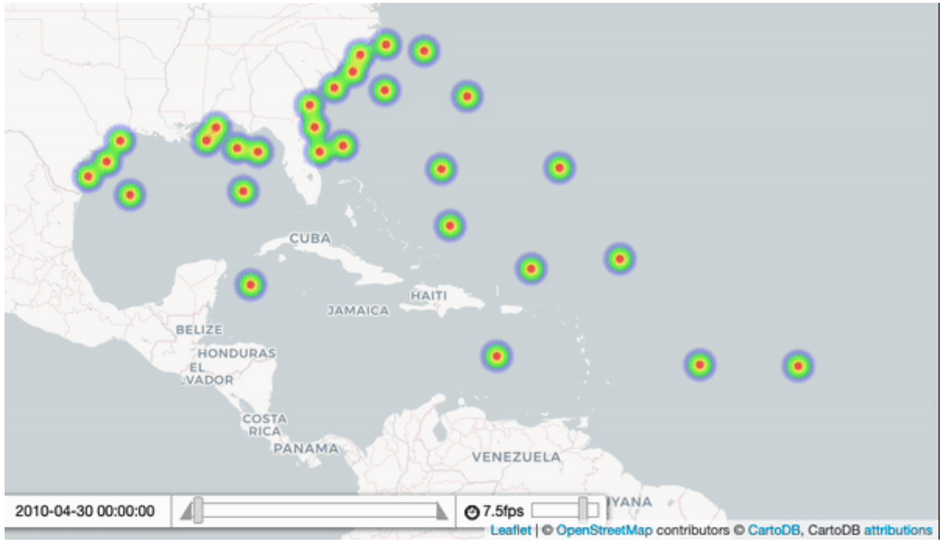
1. Introduction

- Only buoys located near the South-East region of the continental United States, the Gulf of Mexico, and the Caribbean
- Data available from 1978 to 2021
- Most measurements and spatial availability after 2010
- Significant Wave Height as target parameter



One of NOAA's buoys moored in the Gulf of Mexico





Outline

2. Methodology

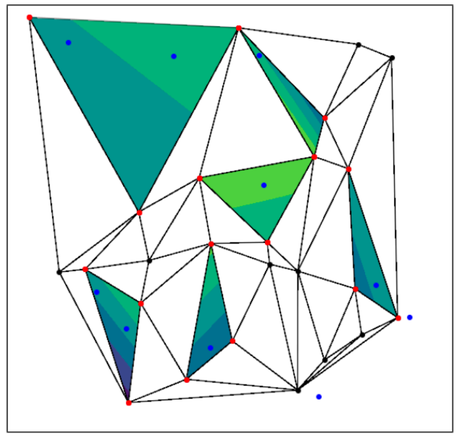
- Introduction
- Methodology
- Results
- Conclusions

Algorithms Studied

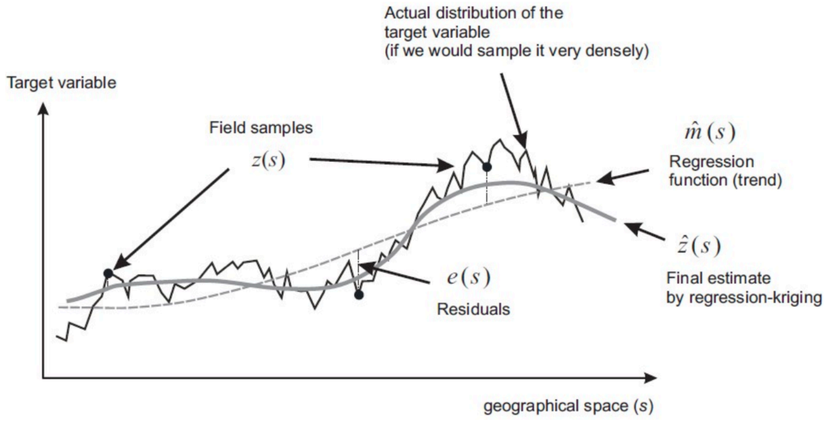
2. Methodology

- **Deterministic Algorithms:** Linear Barycentric, IDW, Radial Basis Functions (Gaussian, Matern, Cubic, …)
- **Statistical Algorithms:** Ordinary and Regression Kriging (Gaussian Process Regression)
- **ML Algorithms:** Tree-Ensemble methods (Random Forest & Gradient Boosting Regression)

In the case of the first two, the estimations are done with variations of the general formula: $\hat{Z}(x_0) = \sum_{i=1}^{n} w_i Z(x_i)$



Linear Barycentric relies on making a tessellation of the convex hull of the area sampled so it can't do extrapolation.



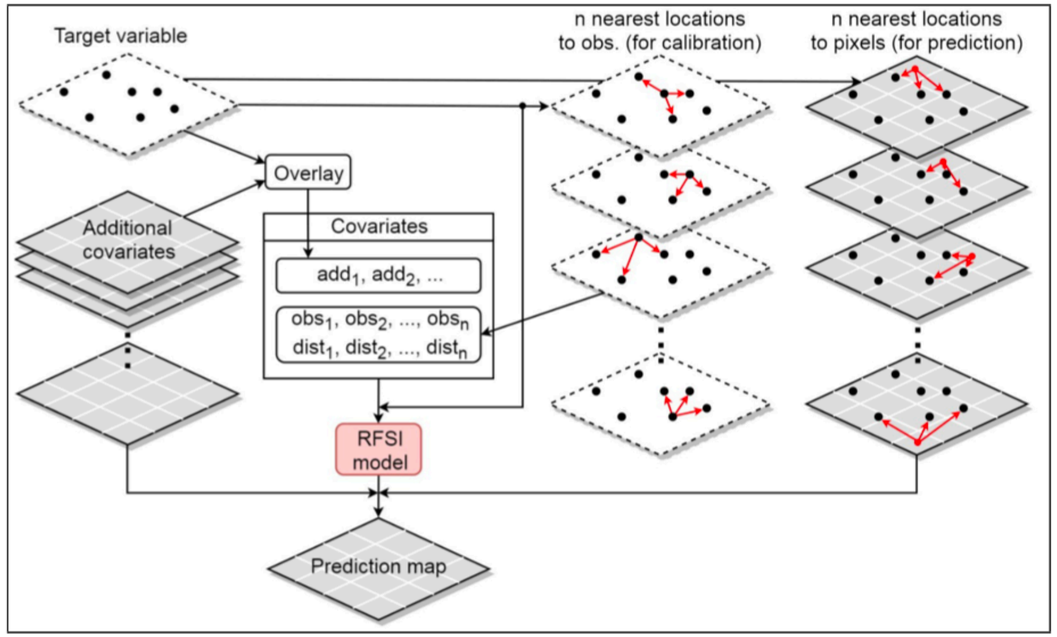
Regression Kriging can be thought as an hybrid approach

Algorithms Studied: ML Methods

2. Methodology

- **Tree-Ensemble models:** Random Forest and Gradient Boosted Regression Trees

- **Feature Extraction:** Engineer variables that carry spatial information about the area around the points to be interpolated.

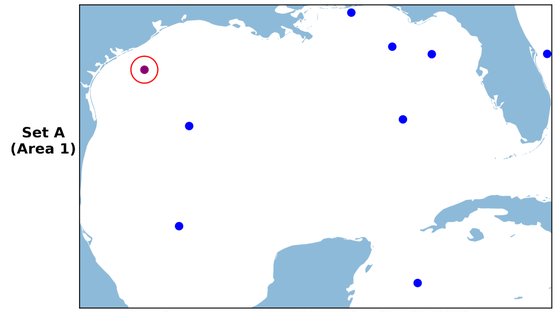


A Framework for extracting spatial information from Sekulic´ et al, 2019

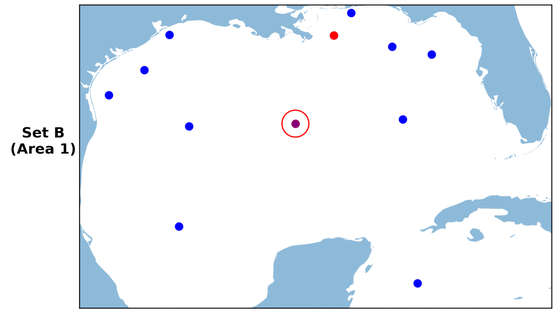
Evaluation

2. Methodology

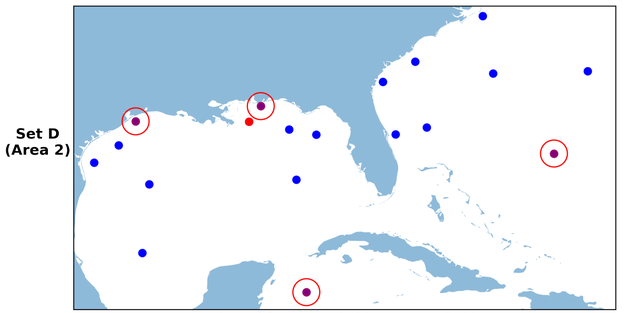
- Three datasets that consist of areas with different spatial configurations
- Multiple test sets per evaluation area
- Validation Metrics: RMSE, MAE and R-Squared



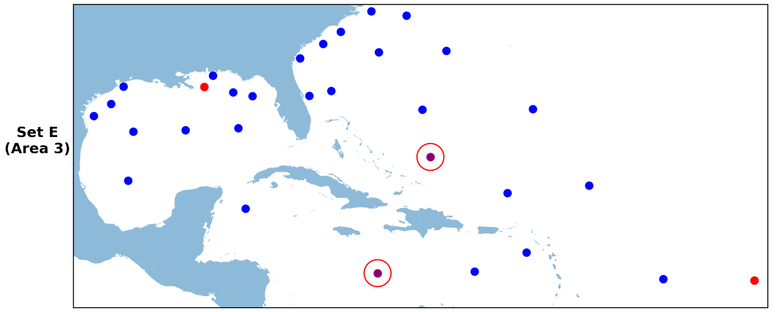
Evaluated from Mar 01, 2021 to Mar 20, 2021



Evaluated from Nov 25, 2016 to Jan 01, 2017



Evaluated from Jun 10, 2015 to Jun 13, 2015



Evaluated from Jul 10, 2021 to Jul 30, 2021

Technical & Experimental Approach

2. Methodology

- **Careful thought was put in the implementation of the experiments**
- **Monitoring done through MLFlow**
- **Pipeline Design and Execution with Kedro**
- **Data Engineering Approach & Deployment Strategy**

Outline

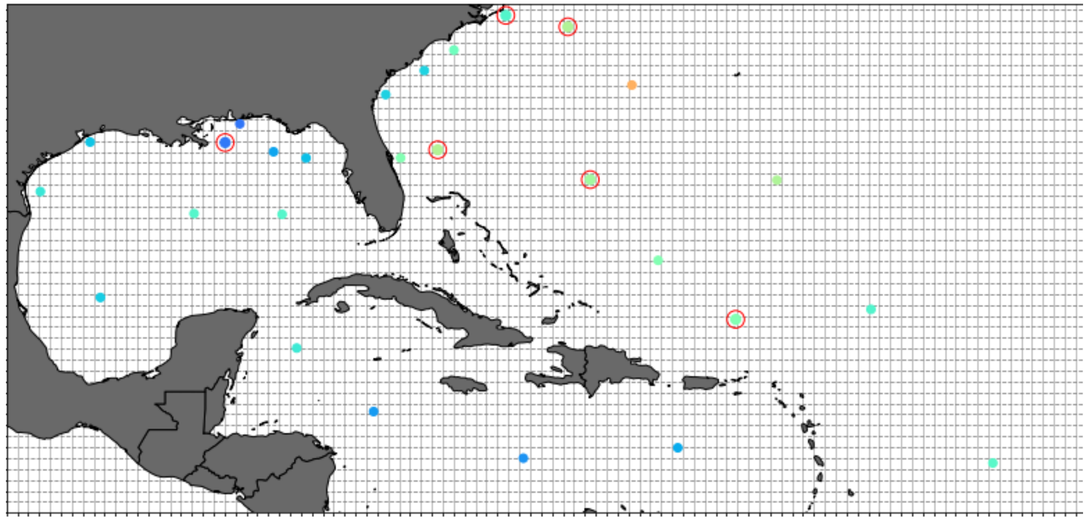
3. Results

- Introduction
- Methodology
- Results
- Conclusions

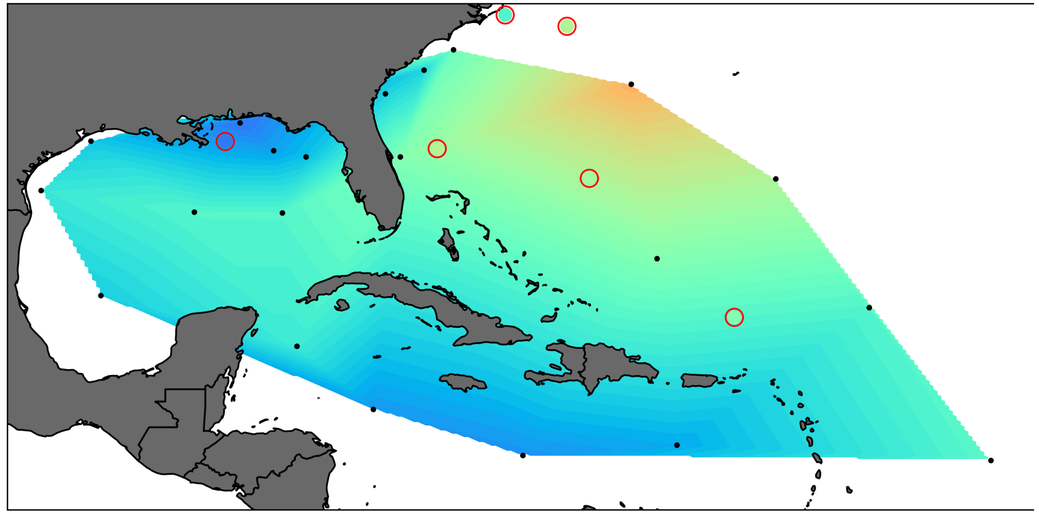
Visual Examples

3. Results

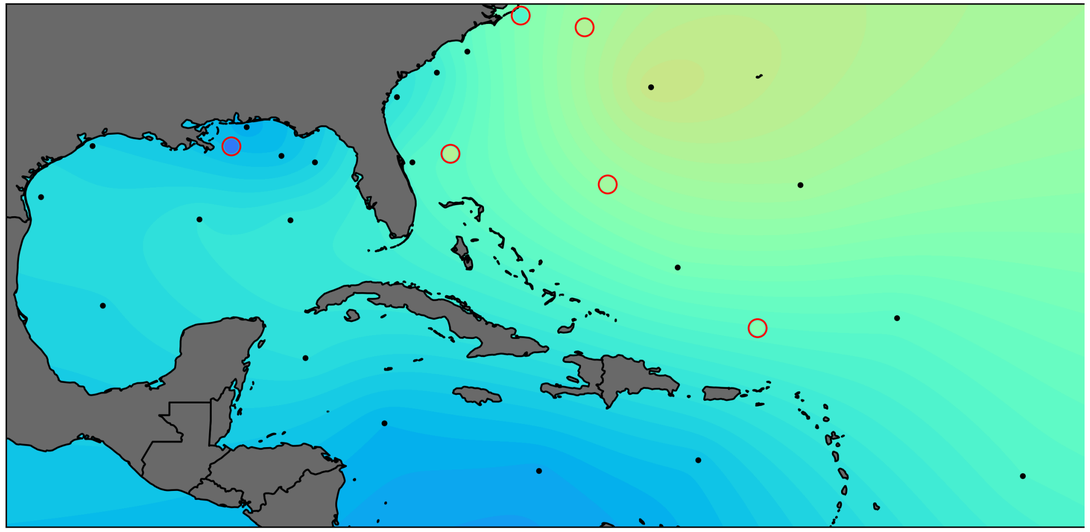
Set F (Area 3)



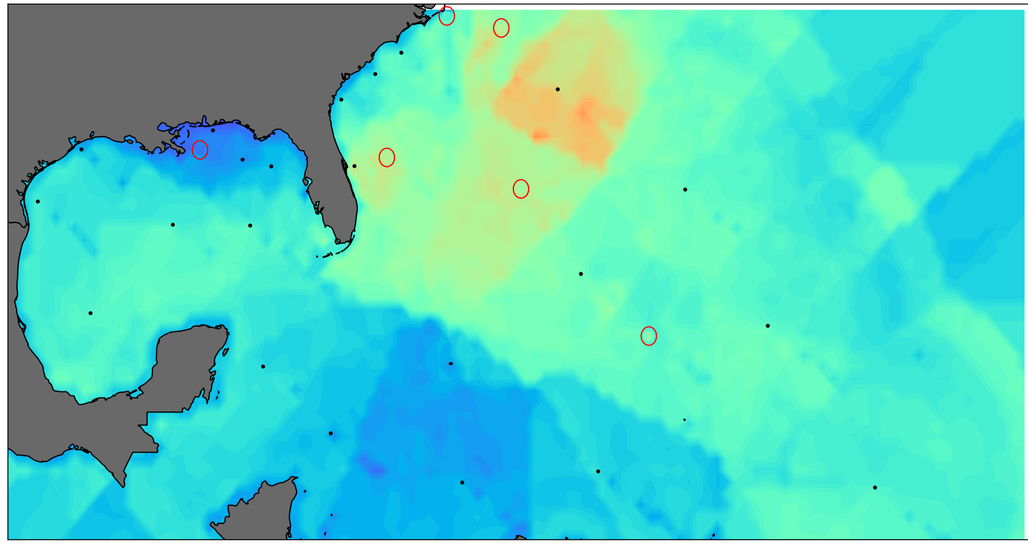
Linear Barycentric Interpolation on Set F (Area 3)

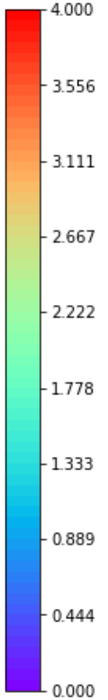


Ordinary Kriging (Spherical) Interpolation on Set F (Area 3)



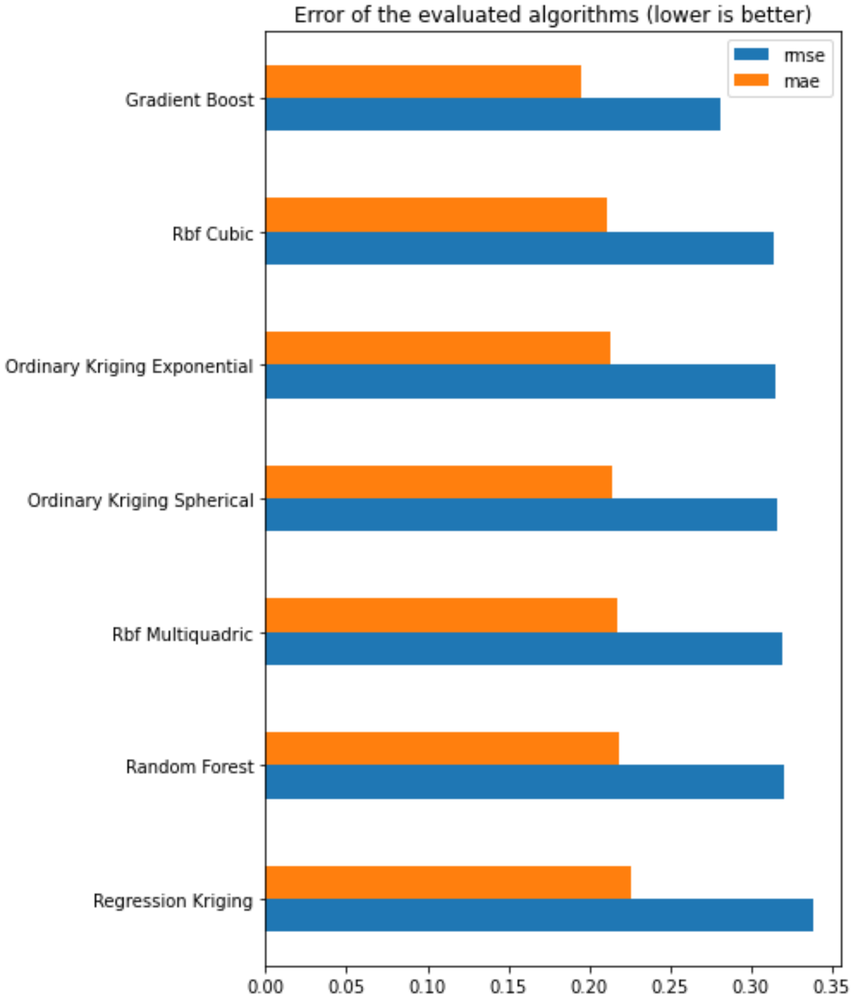
Gradient Boosting Regressor Interpolation on Area 3





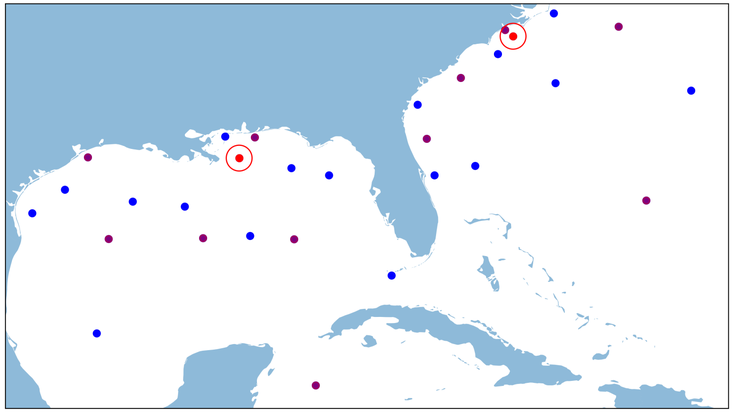
Comparison of the Algorithms

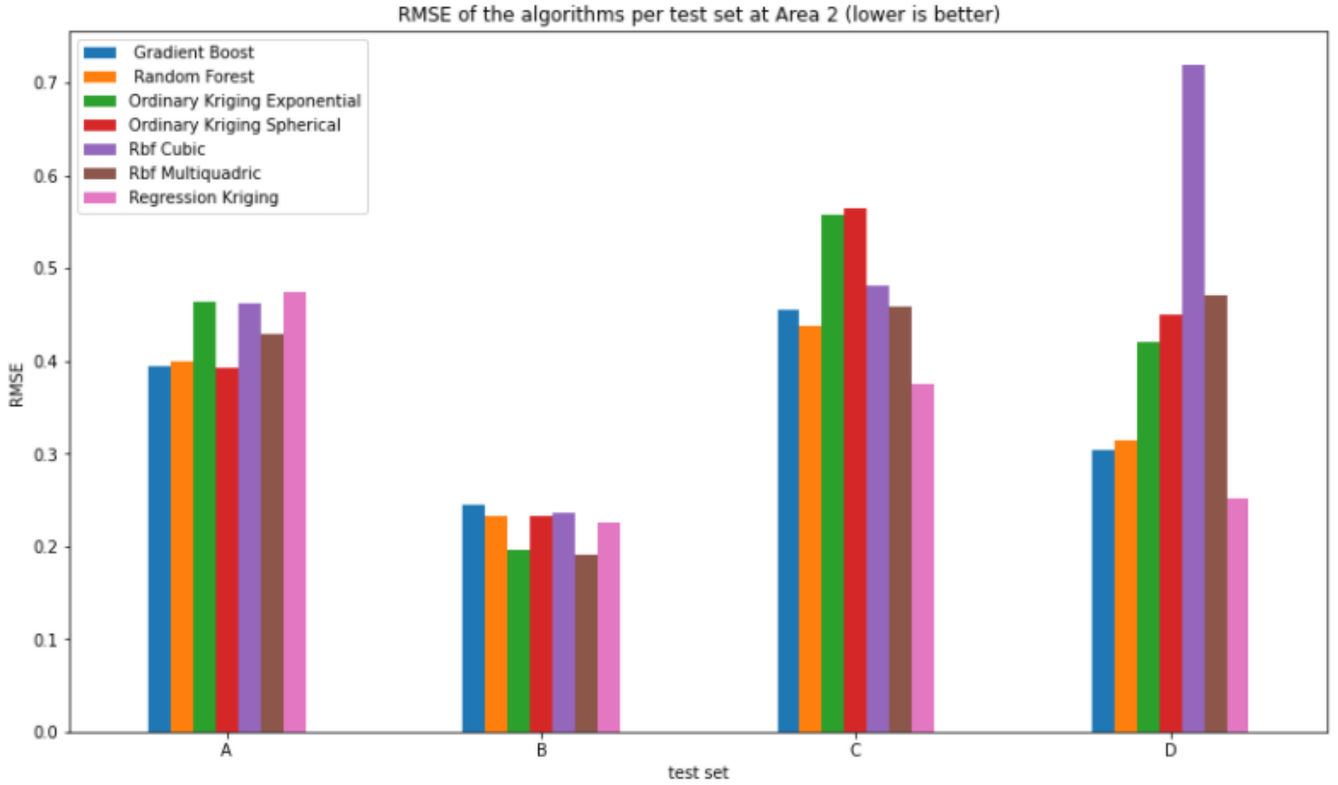
3. Results



Evaluation on Area 2

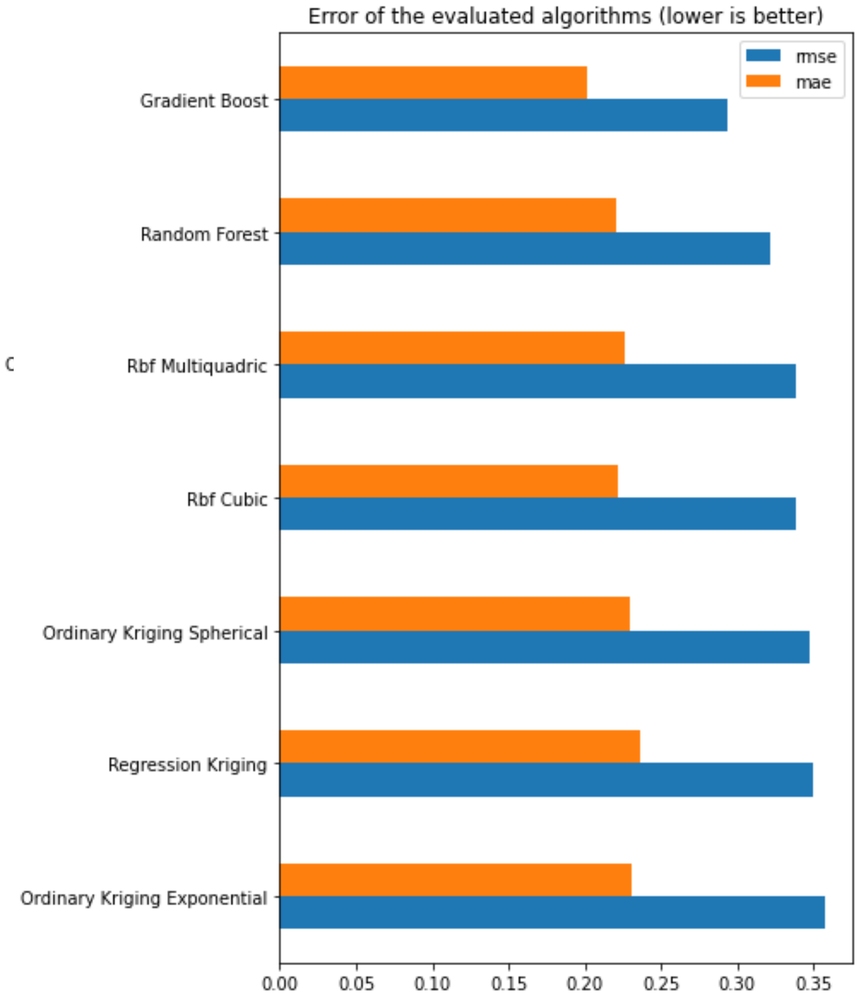
- Set a requires some extrapolation
- Sets b and c only interpolation
- Set d is entirely extrapolation





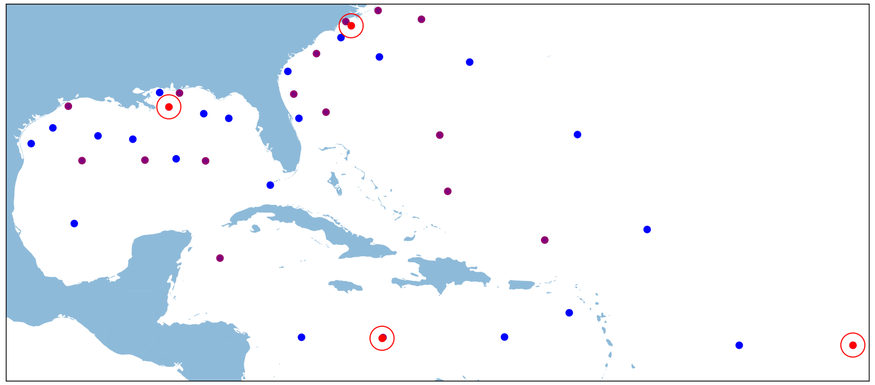
Comparison of the Algorithms

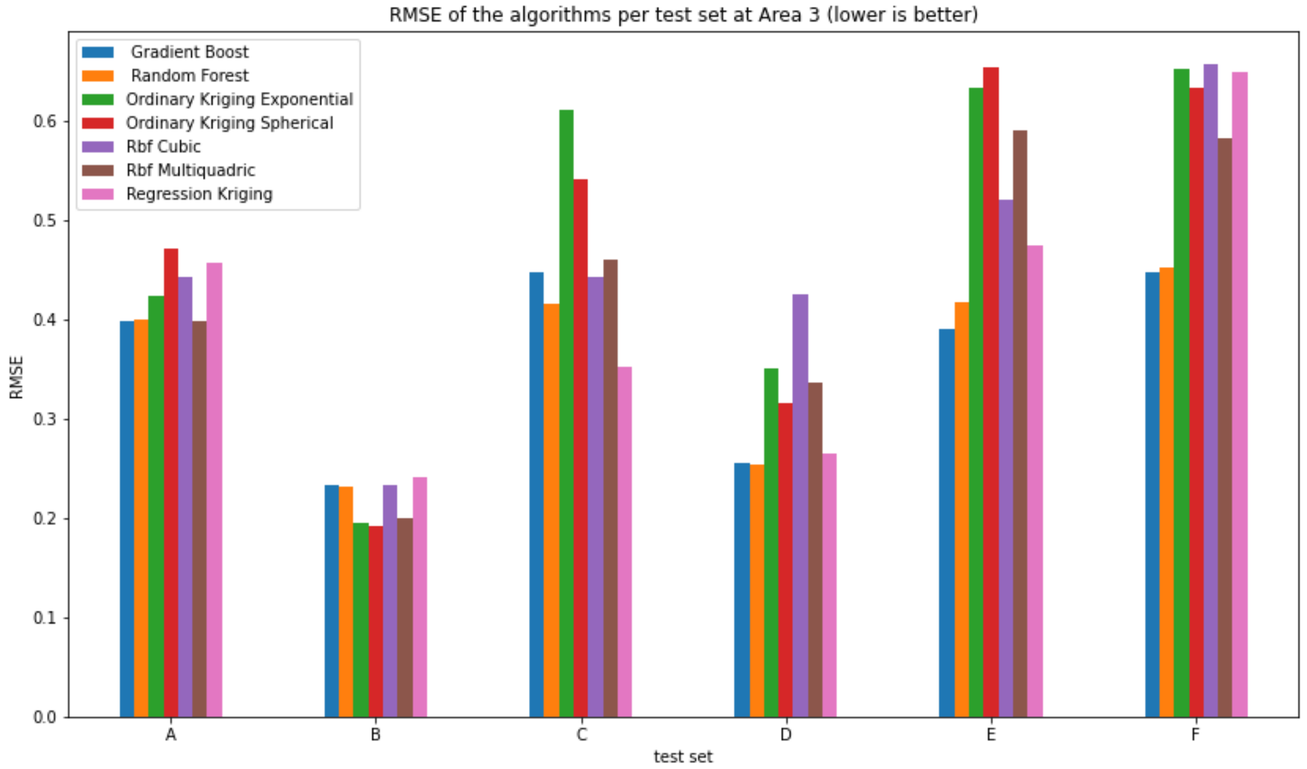
3. Results



Evaluation on Area 3

- Sets a, d, e, f require extrapolation
- Sets b and c only need interpolation





Outline

4. Conclusions

- Introduction
- Methodology
- Results
- Conclusions

Conclusions

4. Conclusions

- While geo-statistical methods such as Kriging dominate the space of Spatial Interpolation in the environmental sciences, there are cases where ML can prove to be a more accurate and/or feasible alternative.

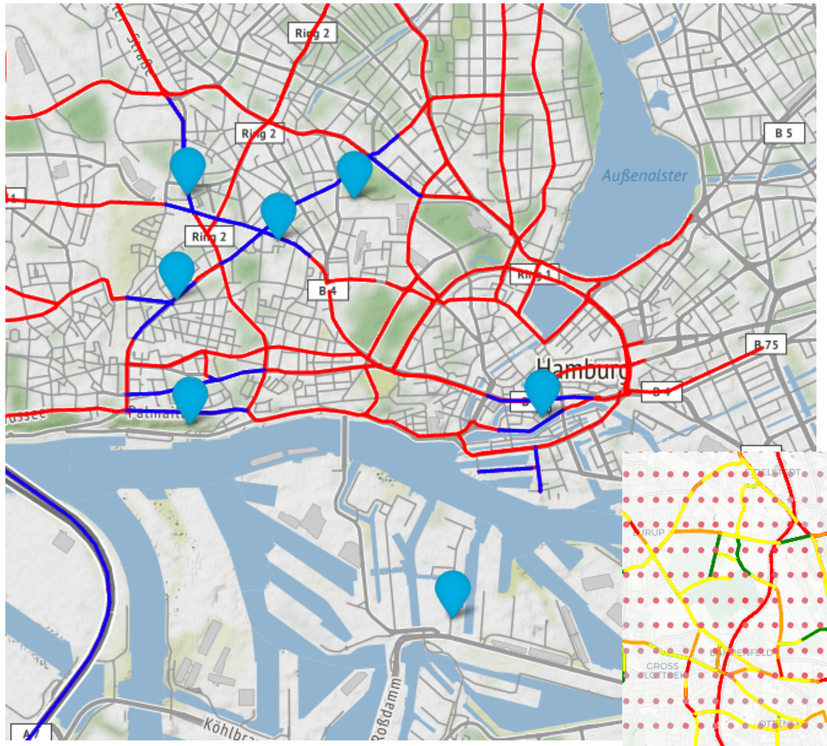
- We can use techniques such as RK to take advantage of the accuracy of ML methods to extrapolate with the capability to estimate uncertainty that makes Kriging so popular

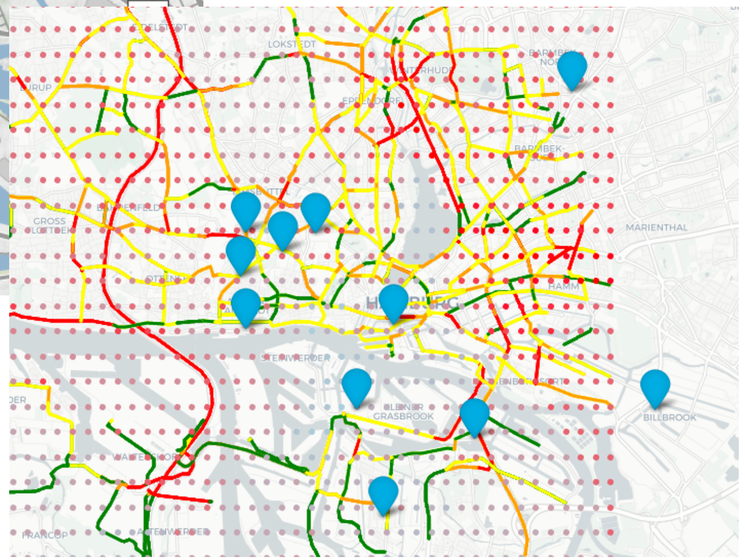
- The techniques and methodology used in this work can be easily implemented with other similar datasets to tackle this problem in real-world scenarios…

PoC of Hamburg Air Quality Data with Breeze Technologies DE

4. Conclusions

- Estimate spatial NO2 values on the Hamburg area using air quality, traffic, and weather parameters
- Regression Kriging with Gradient Boost Regression was used.
- The addition of two more data sources meant more complex pipelines
- Deployment was done using Azure Machine Learning Services





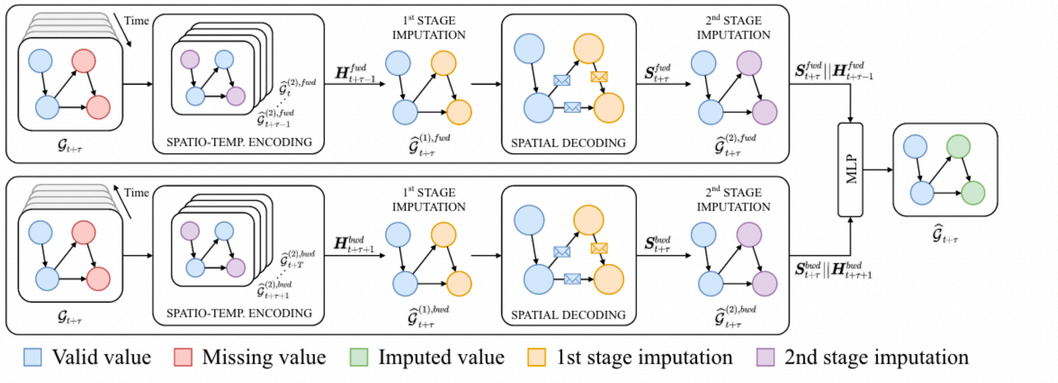
Future Work

4. Conclusions

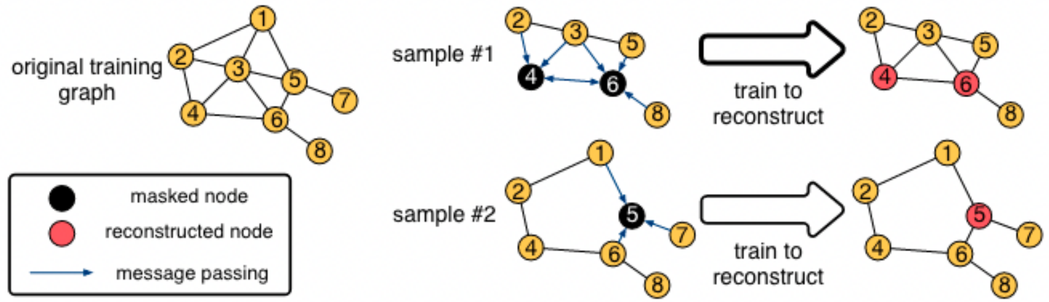
- **Deep Learning Approaches:**
  - Graph Neural Networks
  - ConvLSTM
  - Generative Methods

- **Evaluate other areas, datasets**

- **Live Deployment**



Frameworks for training a Spatial Interpolation GNN sourced from Wu et al, 2020 (top) and Cini et al, 2021 (bottom)

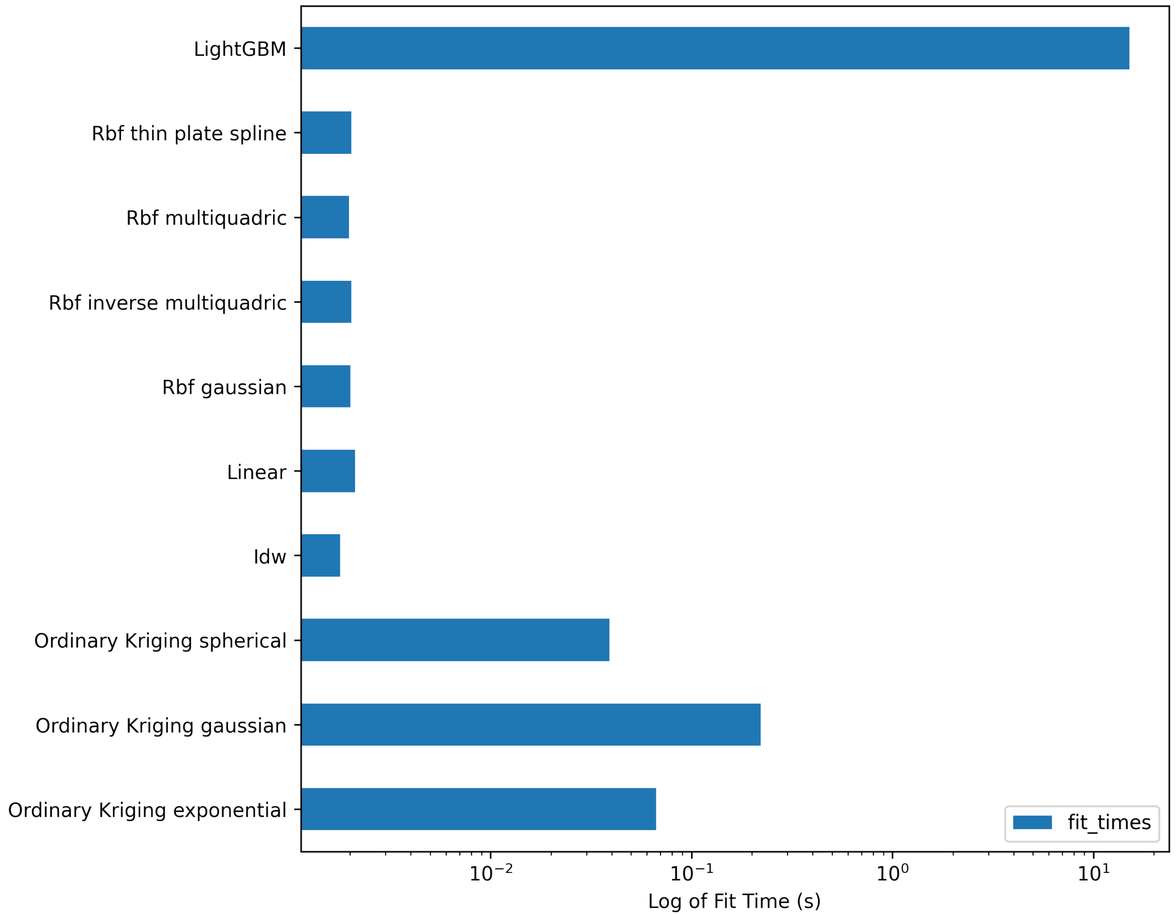


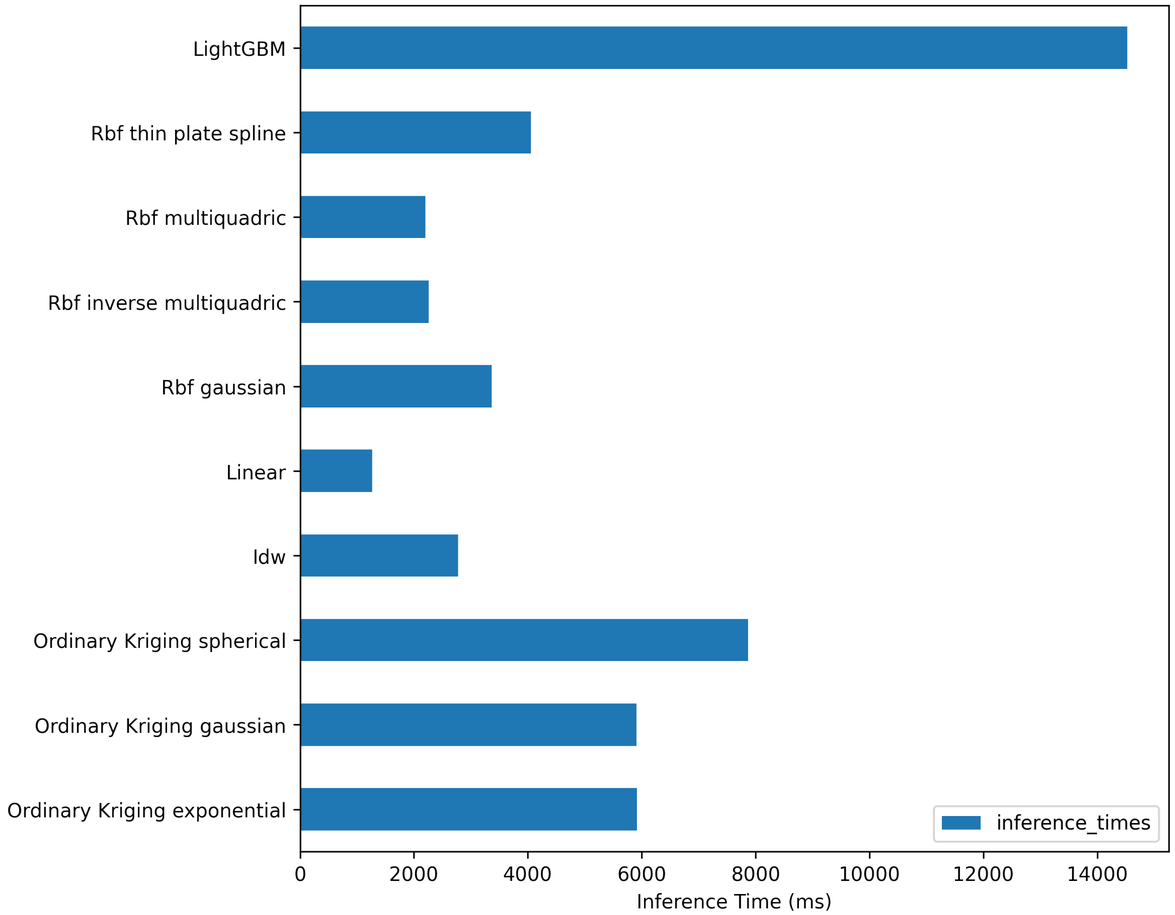
(a) Training process of IGNNK

Questions / Feedback

Thanks for listening!

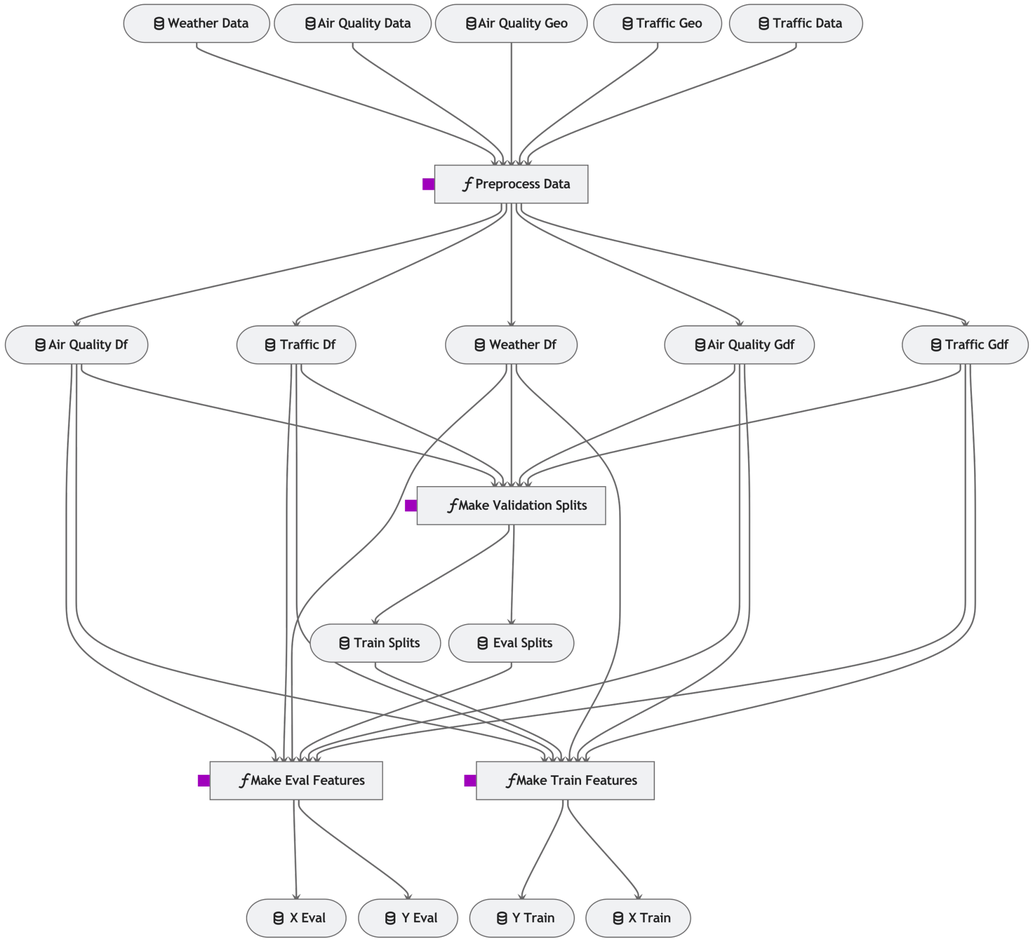
Appendix: Fit and Inference times



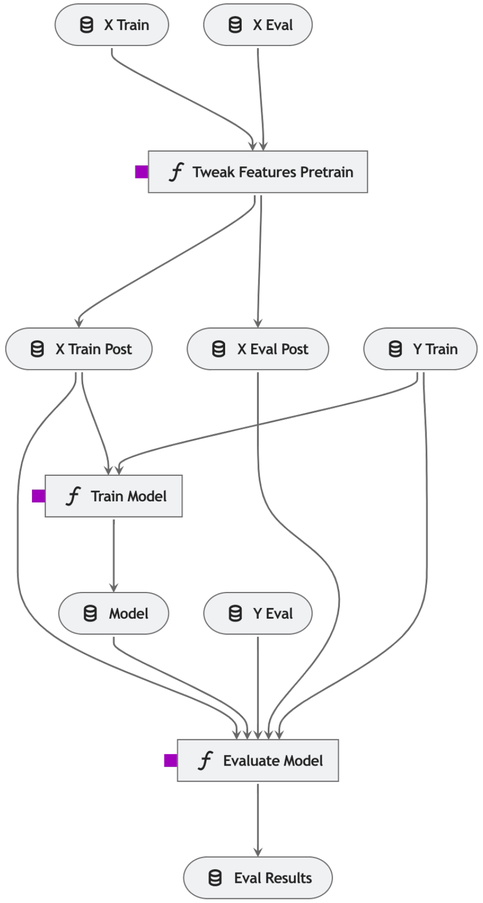


*For Regression Kriging just add the times of LightGBM (Gradient Boost) plus Ordinary Kriging

Appendix: Pipelines and Workflows



Feature extraction pipeline for the Hamburg Air Quality dataset



Pipeline to train and evaluate an ML model

Appendix: ML Feature Importances