Predicting Historical Air Quality - A Zindi Project by Yasin Ayami and Jonathan Whitaker

Executive Summary

In this project, we present a method for predicting historical air quality (as measured by daily median PM25 concentration) for locations where no ground-based sensors are present, by using weather data and remote sensing data from sources like the Sentinel 5P satellite.

Air quality data is obtained for 555 cities and supplemented by satellite and weather data. This is then used to build a model to predict the air quality for a given date and location. A competition hosted by Zindi was used to crowd-source the creation of the model used, with the winning code forming the basis of our modelling approach.

We use the trained model to create a new dataset of historical air quality predictions for cities across Africa, available at https://github.com/johnowhitaker/air_quality_prediction

1. Introduction

1.1 Background

According to the United Nations, 54% of the world's population resides in urban areas in the year 2014. It is projected that by 2050 this number will increase by 12%. The direct effect of this urban drift has had profound effects on social, economic and ecological systems, causing stresses on the environment and society. The social and economic implications include impacts from human activities such as transport, industrialization, combustion, construction etc., all of which have a direct or indirect bearing on the environment. These pollution sources have led to release of pollutants such as Nitrogen dioxide (NO2), Particulate Matter (PM) and Sulphur dioxide (SO2) into the atmosphere. It is believed that air pollution is influenced by urban dynamics. For instance, in the UK, over 40,000 deaths annually are as a result of air pollution [1,2,3,4].

1.2 The 'Gap in the Map'

Whilst air quality is important because it directly affects the air humans inhale, many cities lack resources to monitor the air quality. For instance, the World Air Quality Index (WAQI) monitors over 10, 000 stations across the world. From these stations only Angola and South

Africa have stations that monitor air quality in the southern region of Africa as can be seen in Figure 1. Sustainable Development Goal number 11 states that cities and human settlements should be inclusive, safe, resilient and sustainable. However, this is not achievable if the status quo remains the same. It is therefore imperative that a mechanism to predict the air quality of cities with no ground-based sensors be derived if this goal is to be realised by 2030.

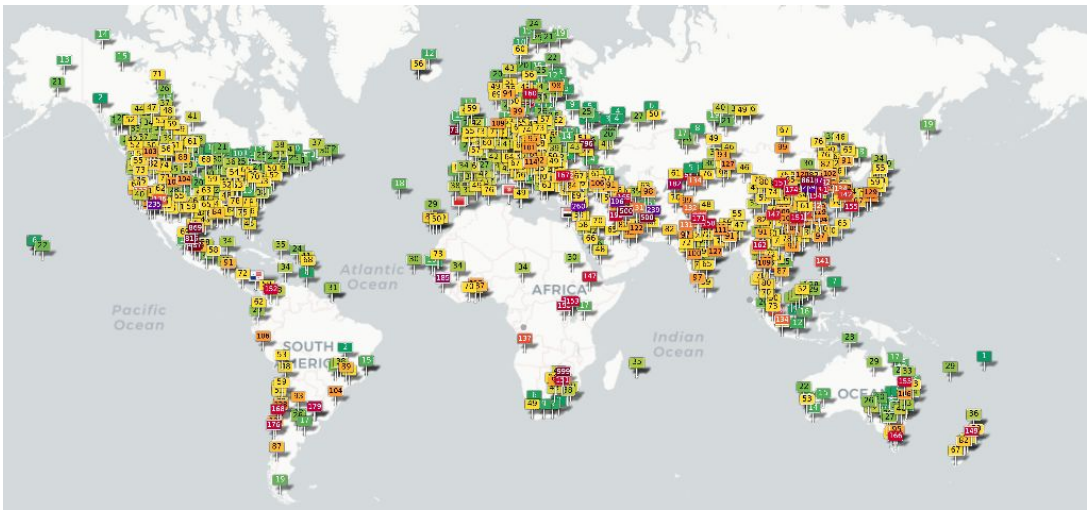

Figure 1: Air Quality sensor locations [5]

Against this background, this project with support from ZINDI and SAEON developed a model that leverages WAQI historical data on air quality to predict the air quality for cities with no ground-based sensors to monitor air quality.

1.3 About the Project

Zindi is a data science competitions platform focused on solving African problems. Participants compete to find the best solutions by creating models and submitting their predictions, which are automatically scored and ranked.

As one of 6 COVID-themed hackathons, air quality data from the start of 2019 was combined with the satellite and other data sources used in this project, and the problem was framed as a prediction challenge. The hackathon was a success, with 240 data scientists enrolled and 134 on the leaderboard. The top solutions were given to Zindi and formed the basis of the modelling component of this project.

The winning solution was shared on GitHub but is not on its own very useful as it was created within the framework of the hackathon. The goal of this project is to build out a complete solution, re-creating and documenting every phase from data sourcing to modelling to application. The final product is a set of historical predictions for cities across South Africa that can hopefully be used by researchers and policymakers going forward.

2. Methods

2.1 Data Sources

Historical Data on air quality was obtained from the World Air Quality Index (WAQI) website from the years 2015 to 2020. This site has historical data for cities across the globe with over 10,000 stations recording different air quality metrics. The data queried from WAQI contained the following fields:

- Date: The day the data was collected
- Country: The country in which the station is stationed.
- City: The city in which the station is stationed.
- Specie: The various types of pollutants available which included co, pm10, o3, so2, no2, pm25, psi, uvi, neph, aqi, mepaqi, pol, temperature, humidity, pressure, wd, wind-speed, d, pm1, wind-gust, precipitation, dew, wind speed, wind gust.
- Count: number of readings of a specie on a specific day
- Min: the minimum reading of a specie on a specific day
- Median: the median reading of a specie on a specific day
- Max: the maximum reading of a specie on a specific day
- Variance: the variance of a specie on a specific day

The latitude and longitude was further added to the datasets for each city. However, for this project, only data consisting of readings for particulate matter (PM25) as a pollutant aggregated by the city was used. More so, the count, min , max and variance columns were excluded from the dataset.

As inputs to the model, various data layers in Google Earth Engine were queried for each location covered by the training data, and then again for the new locations where the predictions were to be made. The inputs included static variables (one measurement used to cover the whole period) and several time-series consisting of daily measurements which were used for the

- Nighttime Light Intensity - A measure that correlates with economic intensity. [https://developers.google.com/earth-engine/datasets/catalog/NOAA_DMSP-OLS_CALIBRATED_LIGHTS_V4]
- Population Density - measured as mean, maximum and minimum density within 5km of the location being considered. [https://developers.google.com/earth-engine/datasets/catalog/CIESIN_GPWv411_GPW_Population_Density]
- Weather Data (daily) was obtained from the Global Forecasting System (GFS) dataset, [link]
- Sentinel 5p Data from the following collections: 'L3_NO2', 'L3_O3', 'L3_CO', 'L3_HCHO', 'L3_AER_AI', 'L3_SO2', 'L3_CH4', 'L3_CLOUD'. More information can be found at https://developers.google.com/earth-engine/datasets/catalog/sentinel-5p

For identifying new locations where we could make predictions, several datasets were considered. We ultimately settled on:

- A list of major cities in Africa derived from https://simplemaps.com/data/world-cities - 1466 locations in total
- A list of cities in SA from the same source (73 total)
- A new dataset obtained by identifying population clusters within South Africa, based on the GPWv411 dataset (Gridded Population of the World, Version 4, https://doi.org/10.7927/H49C6VHW.). Identifying clusters with a density greater than 1000 people/km^2 over more than 2 km^2 resulted in 496 additional locations across South Africa. Earth Engine Script

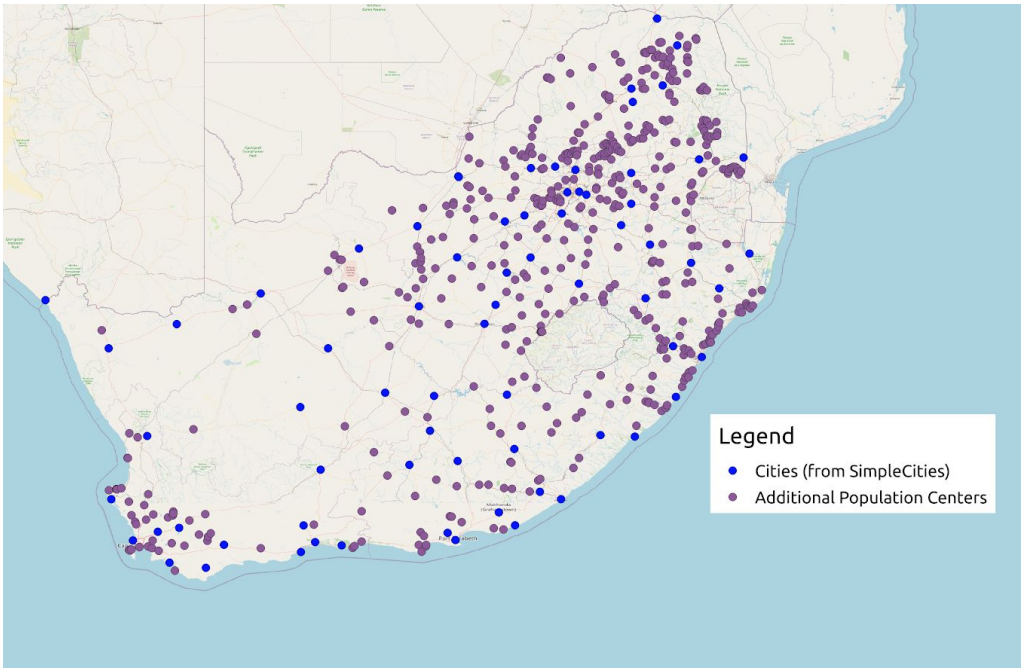

Figure 2 - Locations for predictions within South Africa

2.2 Zindi Competition

A version of this problem was presented on Zindi in the form of a Hackathon, where participants were presented with the satellite and weather data, along with the target variable (median PM25 concentration) for a subset of the cities in the dataset. The challenge was to create a model that could predict the air quality measurements for the cities in the test set.

The winning code was submitted to Zindi for review and was used as inspiration for the modelling stages of this project.

2.3 Feature Engineering and Modelling

The winning submission provided a good starting point for the modelling and feature engineering process. However, the nature of the competition framework meant that a lot of tweaking was required to transform the winning code into a single documented process. Specifically:

- The winning solution trained an ensemble of different models, combining the predictions together to achieve a small improvement in the final score. We replaced this ensemble with a single model for simplicity.
- Some features created during the competition, such as measurement frequency for a city, cannot be replicated for new locations, and so they had to be dropped.

The final proposed solution is documented in the 'Modelling' notebook in the GitHub Repository. To summarize the process:

2.3.1 Feature Engineering

- The training data is read in and combined with the time-series data sampled from GEE in the 'Data Prep' notebook.
- A few unneeded columns are dropped.
- For columns with only a small number of missing values, the missing data is replaced with the mean reading for that location.
- For variables with a high correlation to the target, several lagged versions of those variables are created.
- The date column is transformed into several additional features.

2.3.2 Modelling and evaluation

- The data is split by city in k-fold cross-validation. For each fold, a catboost model is trained with the parameters used by the winning solution to the Zindi competition, and the predictions for the test set are stored and used for evaluation.
- The metric chosen is RMSE, although several other metrics are implemented (see next section).

2.4 Sharing and Analysing predictions

We make the model predictions available CSV files - one for South African cities, a second for the additional population centers and a third for major cities across Africa. In addition, a simple dashboard is created where users can view the air quality timeseries - http://www.datasciencecastnet.com/airq/ and the data is also being made available through our partner SAEON.

3. Results

When predicting daily median PM25 values for unseen locations, our model had an RMSE of 27 ppm. Smoothing the predictions and the targets with a 10-day rolling average gave a much lower RMSE of 23.7.

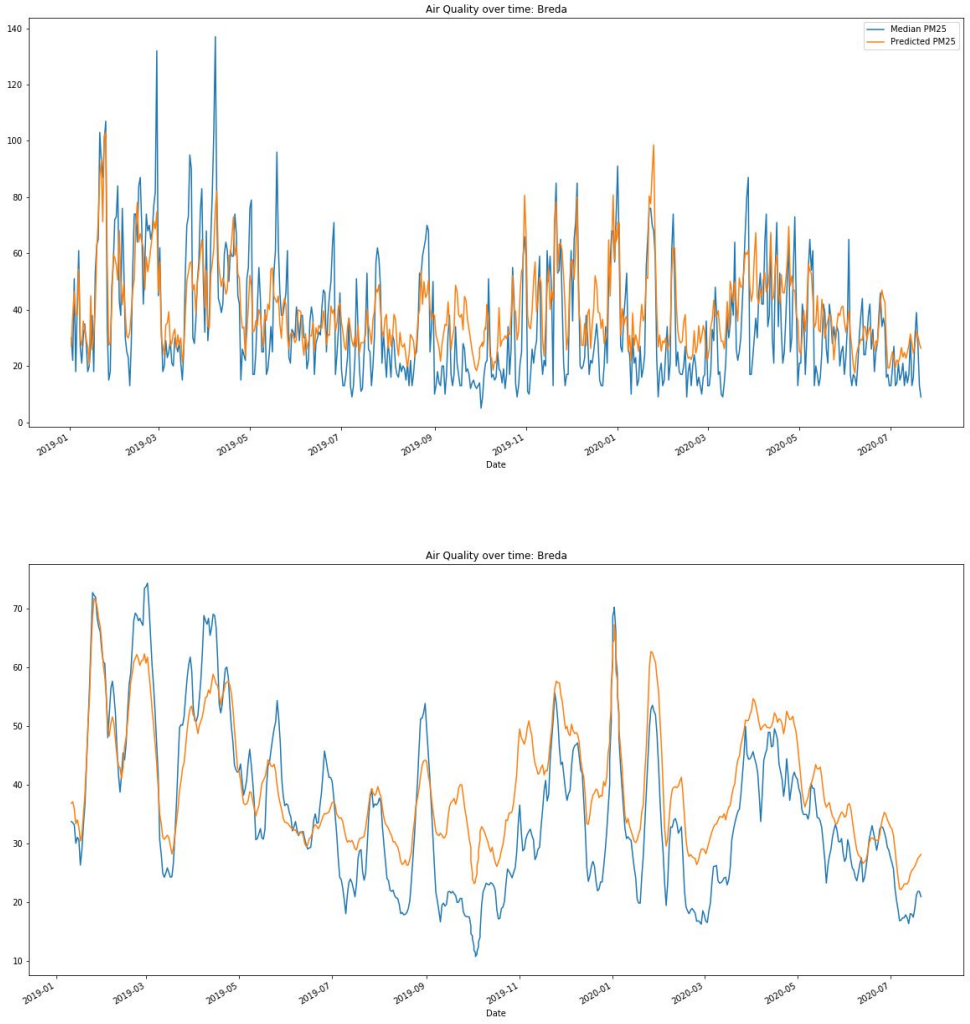

Figure X: Predictions vs observed values for Breda, daily (top) and smoothed (bottom)

To add context to these errors, we group the predicted PM25 concentrations into the ranges used by [organisations] to assign risk: 0-50 for healthy, 50-100 for moderate-risk and so on.

The model assigns the correct air quality class ~70% of the time, and the adjacent class (off by one) a further 27% of the time. In other words, it is off by more than one division less than 3% of the time.

Table 1: Additional Metrics

Metric	**Score**
RMSE	27.07
RMSE (Smoothed Predictions - moving window size 10)	21.9
MAE	19.2
MAE (Smoothed Predictions)	15.74
R^2	0.615
R^2 (Smoothed Predictions)	0.633

4. Discussion

Based on the results above, we believe that the model is accurate enough to be useful. However, it does have some limitations:

- Sensor readings are aggregated across a city. This can miss situations where there may be concentrated areas of low air quality. Notable examples of this can be found in several South African cities.
- The training data consists of mainly larger towns/cities, so the model performance may be lower when considering smaller population centers.
- These models are able to capture large-scale trends in outdoor air quality, but still do not address the issue of **indoor** air quality - a major issue especially in places where cooking methods such as wood stoves are still commonplace[6].

5. Conclusions and Recommendations

This method allows us to predict historical air quality for new locations with enough accuracy to draw meaningful conclusions about the air quality even for places without any sensor data. Despite its limitations, we hope that this serves as a step in the right direction, filling in the ‘gap in the map’ which inspired this project.

6. References

1. Chen, Q., Wang, W., Wu, F., De, S., Wang, R., Zhang, B. and Huang, X., 2019. A survey on an emerging area: Deep learning for smart city data. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *3*(5), pp.392-410.

2. Deng, Z., Weng, D., Chen, J., Liu, R., Wang, Z., Bao, J., Zheng, Y. and Wu, Y., 2019. Airvis: Visual analytics of air pollution propagation. *IEEE transactions on visualization and computer graphics*, *26*(1), pp.800-810.

3. Yi, X., Zhang, J., Wang, Z., Li, T. and Zheng, Y., 2018, July. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 965-973).

4. Zhou, Y., De, S., Ewa, G., Perera, C. and Moessner, K., 2018. Data-driven air quality characterization for urban environments: A case study. *IEEE Access*, *6*, pp.77996-78006.
5. World Air Quality Index, 2020
6. Jafta, N., Barregard, L., Jeena, P.M. and Naidoo, R.N., 2017. Indoor air quality of low and middle income urban households in Durban, South Africa. Environmental research, 156, pp.47-56.