Analysis of the Impact of External Factors on the Air Quality Index: a Machine Learning Approach

https://doi.org/10.31713/MCIT.2024.014

Nataliya Boyko
department of artificial intelligence systems
Lviv Polytechnic National University
Lviv, Ukraine
Nataliya.i.boyko@lpnu.ua

Heorhii Petunin
department of artificial intelligence systems
Lviv Polytechnic National University
Lviv, Ukraine
heorhiipetunin@gmail.com

***Abstract* – In today's world, we often see governments introduce green laws that are actively advertised and promoted to the public. There is active promotion of the green agenda: They create comfortable conditions for electric car manufacturers, provide benefits for the purchase and use of electric cars to citizens, create and develop alternative ways of generating electricity instead of rough heating at thermal power plants, introduce standards for car exhaust emissions that limit manufacturers in production, try to get rid of factories that emit significant amounts of harmful substances into the air by moving them to other countries (Asian countries), leaving only collection workshops or enterprises whose impact is not significant within their country. All of this is done in order to improve air quality in their own country. Although all these changes often seem unnecessary or even harmful to the population and economy, what is the actual result of such actions? This paper presents an analysis of the air composition over the past 50 years, which proves the correctness of the decisions made. It also discusses possible trends in the growth of harmful substances in the air and why they did not materialize.**

***Keywords* – green law; the Air Quality Index; Ecological Footprint; natural resources; the Extreme Gradient Boosted Decision Tree**

I. Introduction

The impact of external factors on air quality is largely the direct responsibility of people. Given the trends of recent decades, people are increasingly thinking about the feasibility of certain laws, projects or measures to improve air quality and reduce emissions of harmful substances into the atmosphere. By finding the extremes of the Air Quality Index (AQI) and comparing them with the years when certain green measures were introduced, we can clearly answer the question of the feasibility and effectiveness of these measures.

In recent years, governments worldwide have increasingly embraced green laws aimed at mitigating environmental degradation, particularly air pollution. Despite skepticism and concerns regarding economic implications, this paper conducts a comprehensive analysis of air composition over the past five decades to evaluate the efficacy of such policies.

The study reveals a discernible improvement in air quality, attributed to a range of legislative measures promoting sustainable practices. These include incentivizing electric vehicle adoption, stringent emissions standards for automobiles, and transitioning to alternative energy sources. Furthermore, the relocation of pollutant-emitting industries to regions with lax regulations has notably reduced domestic air pollution.

By examining historical trends and utilizing empirical data, this paper identifies the effectiveness of green laws in curbing harmful emissions. Contrary to fears of economic strain, the findings demonstrate that prioritizing environmental conservation can yield tangible benefits, including improved public health and enhanced quality of life. Additionally, the analysis offers insights into potential future trajectories, highlighting the sustainability of current policies in maintaining air quality standards.

II. Methods of Solving

Since this work involves time series, well-structured data is essential. No ready-made data frames were found in the public domain, but monitors for individual substances were provided that were well structured by year. After analysing the data from sources [6] and [7], several files were identified that were suitable for converting into a single data set that could be processed by machine learning methods.

Each file had a lot of information about the countries: code, monitor types, etc. Therefore, the data was converted to the form 'Country + Year_1970 + ... + Year_2021'. This resulted in 3 samples of size 30x53. Later, using the obtained data frames, 4 samples were formed, which corresponds to the air quality index.

Since no reliable data on the air quality index for European countries were found, it was decided to generate it manually. Using a data frame of air quality index data from India [5], the XGBoost method was trained to predict the AQI using regression. This model was later used to calculate the AQI for European countries, which resulted in a sample size of 4.

MinMaxScaler doesn't reduce the effect of outliers, but it linearly scales them down into a fixed range, where the largest occurring data point corresponds to the maximum value and the smallest one corresponds to the minimum value (Equation 1).

$$\begin{aligned} X\_std &= (X - X.min(axis = 0)) / (X.max(axis = 0) - X.min(axis = 0)) \\ X\_scaled &= X\_std * (max - min) + min, \end{aligned} \quad (1)$$

where *min*, *max* = feature_range.

This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

In this paper, this was used to bring the numerical data on the same parameters used to calculate the AQI to a single data range; this ensures that the model trained on one data set can be used correctly with new data on other countries.

XGBoost [8] uses Newton-Raphson Method in function space, unlike gradient bousting which works as gradient descent in function space, the loss function uses second order Taylor series to relate to Newton-Raphson method.

A general view of the non-regularised XGBoost algorithm:

- Input: training set $\{(x_1, y_i)\}_{i=1}^{N}$,
- differentiable loss function $L(y, F(x))$,
- number of weak learners M,
- learning rate α.

Algorithm:

1. Initialise the model with a constant value (Equation 2): $\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta)$
2. For $m = 1$ to M:
3. Calculate the "gradients" and "hessians" (Equation 3):

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$
$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}. \quad (3)$$

4. Fit a basic/weak learner using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$, by solving the following optimisation problem (Equation 4):

$$\hat{\phi}_m = \arg\min_{\phi \in \Phi} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2. \quad (4)$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x).$$

5. Model update (Equation 5):

$$\hat{f}_m(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x). \quad (5)$$

6. Result (Equation 6):

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x). \quad (6)$$

III. Experiments

Let's review the results that provide valuable visual representations of changes in substance concentrations over time, serving as important tools for analyzing air quality trends and evaluating the effectiveness of environmental policies and measures. Further analysis of these graphs, in conjunction with the estimated AQI data, can yield deeper insights into the dynamics of air pollution and the impacts of regulatory measures.

Figures 1, 2, 3 show graphs of changes in the content of these substances in the air each year in the range from 1970 to 2021.

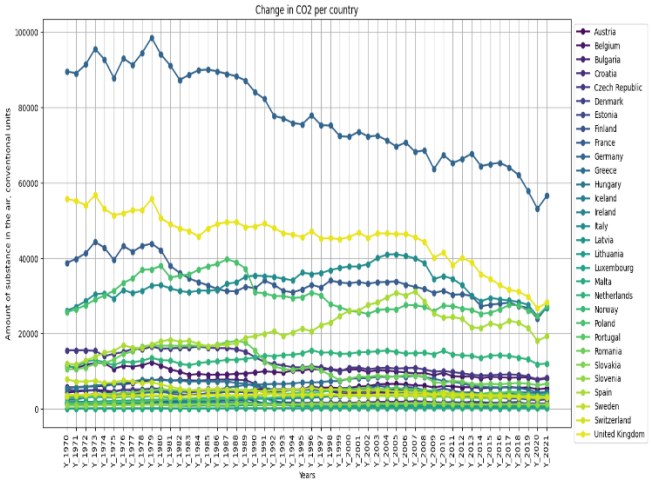

Figure 1. Change in CO_2 per country

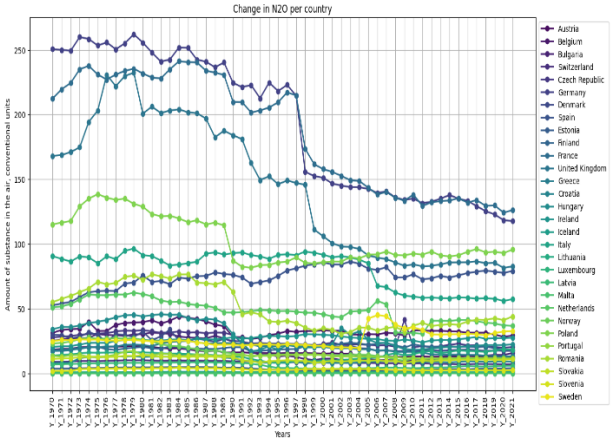

Figure 2. Change in N2O per country

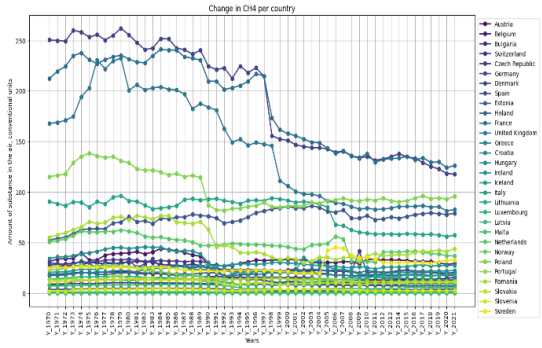

Figure 3. Change in CH4 per country

IV. Results

At this stage, all the objectives were achieved, except for the construction of a sample with events that could affect the change in the air quality index. In general, some information about large countries has been collected, but it has not yet been structured and brought to the form of a sample.

However, the XGBoost machine learning method has now been implemented, with the model trained on alternative data (a sample with data on the content of certain harmful substances in the air and a reliable indicator of the air quality index). This approach allowed the model to be trained on real data, which made it possible to calculate the air quality index for each time series in European countries with an accuracy of 98%. We also processed a significant amount of data, which allowed us to collect, normalise and make available for analysis data on the air content of 3 harmful substances over the past 52 years.

The data was structured and illustrated with the help of graphs; this allows for a quick assessment of the extremes of both individual substances and the air quality index itself.

Given the results, this study can be used as a basis for the following research:

- It is possible to collect additional data for the time period corresponding to the one chosen in the study and conduct a more accurate analysis of the impact of certain substances on the air quality index.
- New approaches to solving the regression problem and predicting the growth or decline of the predicted value in the future can be tested on these samples.
- Since the impact of human activities on air quality change has not been previously assessed using machine learning methods, this work provides an impetus for the development of this topic.

This study also confirmed the opinion of other authors about the effectiveness of the machine learning method called XGBoost. Taking into account the positive experience of other authors, most of whom in modern studies note the high efficiency and accuracy of the same algorithm, it was used as a prototype in this study. The fact of its effectiveness was confirmed, as we managed to achieve high accuracy rates on both training and validation data. We were also pleased with the ease of use and additional customization of the algorithm architecture: there are many libraries that allow calculating accuracy metrics for the regressor or classifier based on the chosen algorithm; it is possible to combine it with other popular algorithms; the training time for prediction is also satisfactory.

V. Conclusion

The data on the content of harmful substances in the air was collected. To be able to work with them using machine learning methods, all data was brought to a single standard and format. In addition, a new unique sample was created, containing air quality indices for each country in the time period from 1970 to 2021. This indicator is calculated with an accuracy of 97% or an absolute deviation of 11 conventional units in the range of values from 0 to 489.

A way to estimate the air quality index was proposed using a rather interesting approach: knowing the exact level of the air quality index in Indian cities, and having information on the content of all the substances considered in the paper, the XGBoost Regressor model was trained, which took into account only the parameters used in the analysis of European countries (although there were others in the sample) and showed an accuracy of 97% in calculating the AQI. Later, the data for European countries were brought to a similar numerical range as for Indian cities, which allowed the same model to be used to calculate the AQI for the selected countries. This approach provided a new sample of data, which was so lacking for the study.

Having evaluated the graphs obtained during the tests, we can say that the model estimated the AQI well, as for most countries the same dynamics can be clearly observed in the graphs of substance content and air quality level in pairs.

References

[1] O. Savvateeva, D. Sokolova, M. Semernya, Assessment of the Urban Air Environment Based on Bioindication Studies, in: IOP Conference Series Earth and Environmental Science, Volume 688(1), 2021, P. 012022. https://doi.org/10.1088/1755-1315/688/1/012022.

[2] S. Subramaniam, N. Raju, A. Ganesan, N. Rajavel, M. Chenniappan, C. Prakash, A. Pramanik, A.K. Basak, S. Dixit, Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review, in: Sustainability, Volume 14, 2022, 9951. https://doi.org/10.3390/su14169951.

[3] B. Li, C. Liu, Q. Hu, M. Sun, C. Zhang, Y. Zhu, T. Liu, Y. Guo, G.R. Carmichael, M. Gao, A Deep Learning Approach to Increase the Value of Satellite Data for PM2.5 Monitoring in China, in: Remote Sens., Volume 15, 2023, 3724. https://doi.org/10.3390/rs15153724.

[4] Bellinger, Colin, et al., Mapping historical air pollution levels in an urban environment using convolutional neural networks and ground-based imaging, in: Environmental Science & Technology, Volume 52.10, 2018, P. 6060-6067.

[5] https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india

[6] D. Carslaw, K. Ropkins, Openair—An R package for air quality data analysis, in: Environmental Modelling & Software, Volume 27, 2012, P. 52-61. https://doi.org/10.1016/j.envsoft.2011.09.008

[7] J. Boonphun, Ch. Kaisornsawad, P. Wongchaisuwat, Machine learning techniques for predicting air pollution: A review, in: Environment International, Volume 130, 2019, 104910. https://doi.org/10.1051/e3sconf/201912003004

[8] W. Wei, O. Ramalho, L. Malingre, Machine learning models for predicting indoor air quality parameters: A review, in: Sustainable Cities and Society, in: Indoor Air, Volume 52, 2020, 101864. https://doi.org/10.1111/ina.12580