Examining Variable Significance in Air Quality Across Cities of Different Urban Form in the USA

AI In urban Planning and Design Project Report – Seoul National University
Pierre-Antoine Lequeu

I- Introduction

Over the last few decades, the escalating concern over air quality has become increasingly widespread. Numerous studies have shed light on a robust correlation between poor air quality and multiple health issues, including a marked increase in the prevalence of conditions such as asthma, various cancers, and cardiovascular diseases. The impact of compromised air quality extends beyond individual health, significantly influencing the day-to-day lives of residents, particularly those dwelling in urban areas. The detrimental effects ripple through communities, affecting outdoor activities and even the seemingly simple act of inhaling fresh air, thereby emphasizing the urgency of implementing sustainable policies.

The tangible connection between air quality and the daily routines of individuals underscores the critical need for comprehensive and effective measures. In addition to its direct impact on public health, air pollution significantly contributes to broader environmental concerns. It plays an active role in the ongoing challenges of global warming and the degradation of ecosystems. As we grapple with the multifaceted consequences of air pollution, there is a pressing need to delve deeper into research aimed at understanding the nuanced causes and implications of poor air quality.

This expanded understanding serves as the foundation for informed decision-making and the development of concrete actions. By unraveling the intricacies of air quality dynamics, we empower ourselves to enact meaningful changes that contribute to a sustainable and healthier future. In essence, the quest for cleaner air is not just a matter of personal well-being; it is a collective responsibility towards the vitality of our planet and the well-being of generations to come.

II- Previous Research

Over the years, improvements in machine learning have significantly improved air quality models. As more and more air quality monitoring stations are built in cities with more and better sensors, as well as satellite imaging and analysis, data for pollutant, meteorological conditions and geographical factors is flowing. This complex interplay between various factors, previously hard to approach with conventional analytics approaches, can now be discerned by machine learning algorithms. These models excel at understanding non-linear relationships and interactions between data from multiple sources.

Air quality analysis is a trending topic, with numerous studies. Mendez and al. [1] did an extensive review of the main breakthroughs in the domain in the last decade. Most common machine learning models were tested on air quality prediction. Deep learning taken aside, Support Vector Regression [2][3], Random Forest regression [4] and linear and multi-linear regressions [5] are the most common algorithms found in the literature. Boosting algorithms, even if less popular than regression, also showed great results on air quality regression tasks [6]. However, deep learning models are the most used models nowadays, as they showed better results than simpler models. Specifically, simple MLP [7] and LSTMs [8] are the most common deep learning algorithms, followed by Convolutional Neural Networks. Many studies also implement hybrid models, using more than one algorithm for the regression task.

In urban planning, previous studies have tried to predict air quality analysis in metropolitan areas such as Taipei using SVM [9] or Seoul using MLP [10]. Most studies focus on high population and high densities places, with little interest to small to medium sized city. Squizzato et al. [11] did focus on the mid-sized Itallian city of Treviso. Liang et al. [12] did an extensive study of the effect of urban form on air quality trend in China, using more than 600 cities of various sizes and socio-economic states, with decade-long air quality data. They showed a very strong correlation between urban form and pm2.5 trends. However, they highlight that the results of this study can only be applied to Chinese cities, as a lot of metrics are influenced by exclusively Chinese politics that led, for example, to *ghost* cities, or at least a high rate of unoccupied dwellings.

This project analyses the significance of different pollutions vectors (such as traffic and power plants) in cities of different urban form and development level in the United States of America and could be considered a proof of concept for further analysis.

III- Project

This part focuses on the technical aspect of the project, the data and the model used.

The project can be split into four parts. The first part is the collection of the data needed for the analysis. The second part is the clustering of the cities according to different urban form metrics. The third part focuses on determining the best model for air quality prediction. Finaly, the fourth part analyses variable significance for the model in each cluster and tries to draw conclusions.

a- Dataset

The main dataset used for this project is the DEAP, for *Deciphering Environmental Air Pollution,* dataset, published in 2021 [13]. It consists of daily air quality metrics for the most common pollutants (o3, pm10, pm25, no2, co and so2), meteorological values (temperature, wind speed, wind gust, humidity and pressure) and human activity metrics (distance travelled by cars, population staying at home, population not staying at home and close power plants emissions) for 54 cities of varied sized in the USA.
The original dataset has 71 features (4 for Date, City, County and State and 67 numerical values), but only 20 features are kept in this project. These features are described in Table 1. Moreover, some data cleaning was applied as five cities (Fort Worth TX, Brooklyn NY, Oakland CA, Raleigh NC and Staten Island NY) were removed. They are part of the urban area of bigger cities and are not representative of their environment. For example, Staten Island is expected to have air quality values similar to New-York City, but with very different population and density metrics, thus giving wrong data for our study.

The second dataset was handmade for the project. It has some basic urban form information for the 50 cities of the DEAP dataset: population, urban area population, area, population density, urban area population density and the proportion of water area in the city’s boundaries. More information can be found in Table 2. Some other metrics were looked at, especially representing land-use and city shape as these proved important in previous studies, but they were too hard to collect for all these cities, especially the smaller ones, in the time frame of the project. However, I strongly believe these features would be very interesting in further research.

Feature	Description		Feature	Description
Date	Data of the sample		Humidity_median	Median humidity for the day
City	City of the sample		Dew_median	Median dew for the day
County	County of the city		Wind-speed_median	Median wind speed for the day
State	State of the city		Wind-gust_median	Median wind gust for the day
Population staying at home	Used a measure of domestic emissions.		S02_median	Median SO2 concentration for the day (target feature)
Population not staying at home	Used a measure of away from home emissions.		Pm25_median	Median PM2.5 concentration for the day (target feature)
Mil_miles	Total vehicle travel distance for the sample		O3_median	Median 03 concentration for the day (target feature)
Pp_feat	Calculated feature for the influence of neighboring power plants		PM10_median	Median PM1.0 concentration for the day (target feature)
temperature_median	Median temperature for the day		NO2_median	Median NO2 concentration for the day (target feature)
Pressure_median	Median pressure for the day		CO_median	Median CO concentration for the day (target feature)

Table 1: Features used in the DEAP dataset. The features noted as (target feature) are not used as input but as potential target value.

Feature	Description		Feature	Description
City	City Name		Density	Population density of the city (people/km2)
County	County of the city		Urban_area_density	Population density of the urban area (people/km2)
State	State of the county		Water_area_prop	Water area proportion in the city (%)
Population	Population of the city		Area	Area of the city (km2)
Urban_area_pop	Population of the urban area			

Table 2: Features of the urban form dataset

b- Clustering

The clustering is the process of making groups of cities of similar urban forms. It was made using the K-means algorithm using different sub-group of features of the urban form datasets.

The number of clusters for each sub-group of features was determined “by hand” depending on human evaluation. The sub-group of features and their clustering is as follows: *Population* and *Urban Erea Population* with four clusters (Fig. 1), *Population Density* and *Urban Erea Population Density* with three clusters (Fig. 2), *Proportion of Water Area* with three clusters (Fig. 3) and *Population* and *Density* with four clusters (Fig. 4)

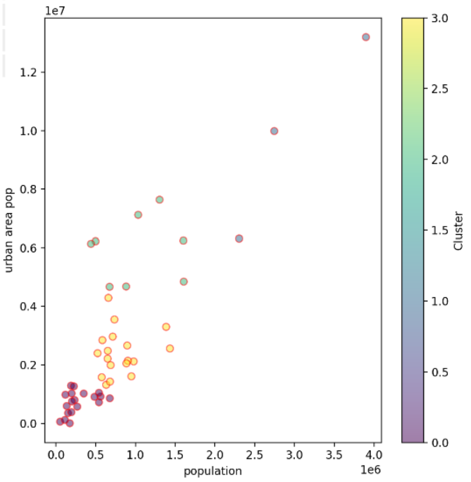


Fig 1: *Population* and *Urban Area Population* clustering

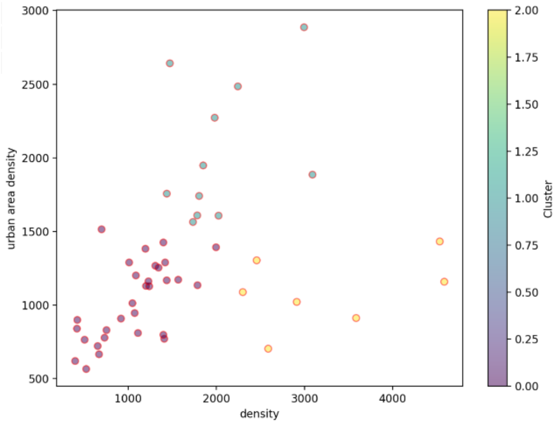


Fig 2: *Density* and *Urban Area Density* clustering

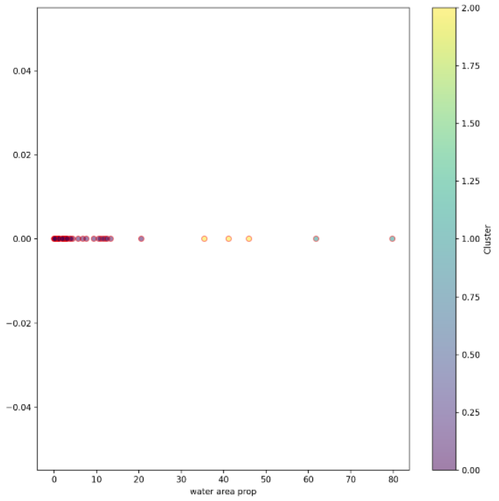


Fig 3: *Proportion of Water Area* clustering

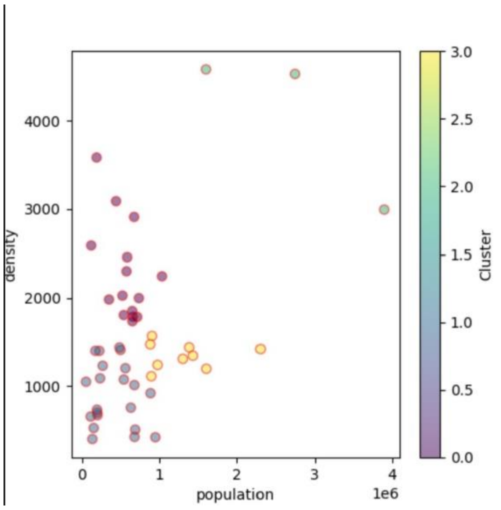


Fig 4: *Population* and *Density* clustering

The choice of the features subset for the different clustering was made so that the clusters are significant but still understandable. A clustering using all features would work but would be very hard to interpret. Therefore, it would be very hard to draw urban planning conclusions from it.

c- Model

In order to avoid different bias for different clusters, a similar model should be used on every one of them. Therefore, the model used should work properly on every cluster. To established which one to use, the one that scores best on the entire dataset is selected. Even though it might not be the optimal one for a specific cluster, it should still be efficient for each one of them.

Two metrics were used to evaluate the models. RMSE (Root Mean Squared Error) is the root of average of the square of the errors and MAPE (Mean Absolute Percentage Error) defines the average of the absolute percentage errors. These metrics were used as they are the two metrics in the DEAP dataset paper to evaluate different models.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{MAPE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{|y_i - \hat{y}_i|}{max(\epsilon, |y_i|)}$$

Three models were tested on the dataset: Gradient Boosting Regressor, Decision Tree Regressor and Bagging Regressor. They were chosen as they showed great results in previous research. Every model hyperparameters is determined by grid search cross-validation for optimal results. There results of the testing are available in Table 3. The train/test split is done as follow: For each city, a 60 days-span is randomly extracted and added to the test set. As each cities have 2 years of data, it makes 9% of the data from testing.

Model	Gradient Boosting	Bagging	Decision Tree
RMSE	**9.80**	10.54	14.08
MAPE	**0.27**	0.29	0.40

Table 3: Results in RMSE and MAPE for three different models

The chosen model the gradient boosting regressor which scored best in both RMSE and MAPE. While the results are not great, they match the results of the DEAP study. It is probable that the DEAP dataset lacks some explanatory features of pm2.5 concentration, and do not allow for better models by itself. The hyperparameters are tuned as follow: learning_rate = 0.1, max_iter = 300, max_leaf_node = 40 and min_sample_leaf = 5. Results of the model compared to reality on real data over two months is displayed in figure 5. Overall, it seems that the model is able to understand the overall trend of the pm2.5 concentration but has a hard time getting the quick fluctuations.

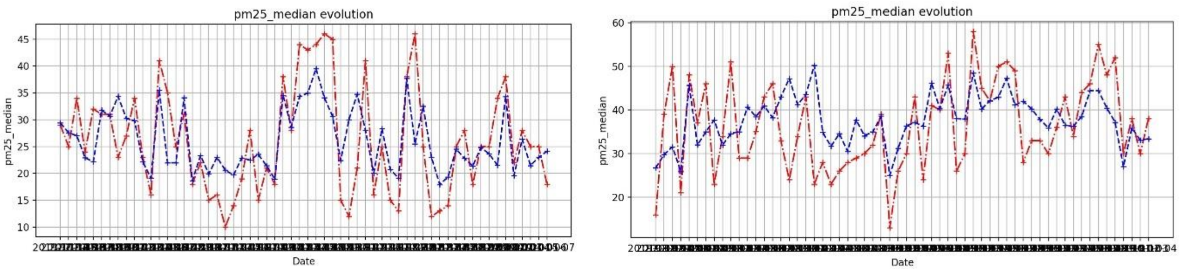


Fig 5: predicted pm2.5 concentration (blue) and real pm2.5 concentration (red) over two months for two different cities.

d- Feature importance

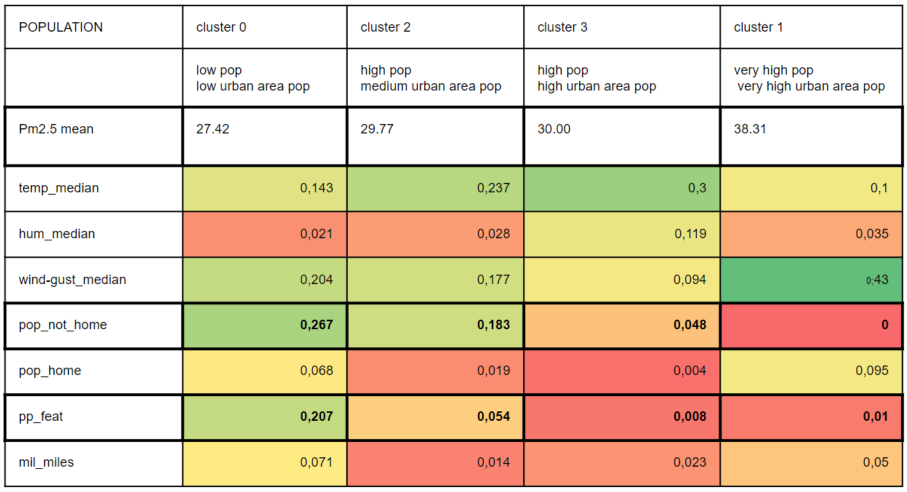
Now that the model is defined, it is trained on a regression task on every cluster. The goal is to compare the features importance in different urban form to understand the underlying link between urban form and air quality.

The feature importance is evaluated using the permutation importance. This algorithm evaluates the importance of a feature by permuting the values of said feature between different line and evaluate how much it modifies the output. The more it modifies it, the more this feature is important.

For every one of the four clustering groups, a table with the pm2.5 mean and the feature importance for each cluster is given. From that, trends are highlighted.

Population clustering

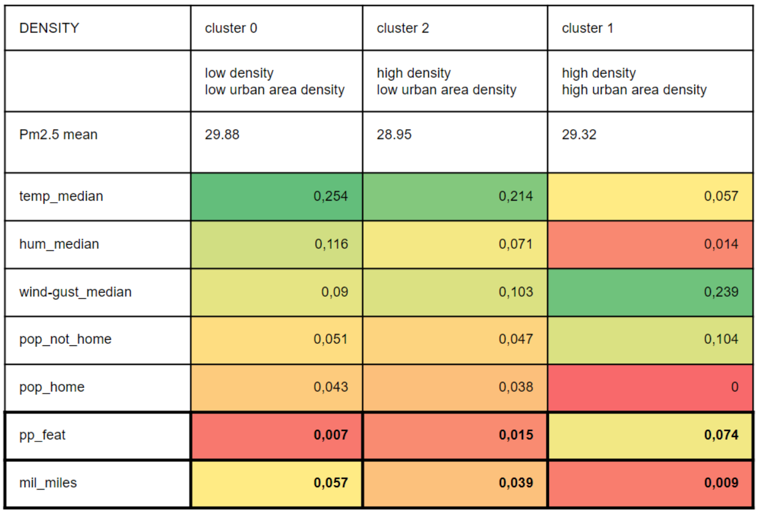
A first trend for the population clustering (Table 4) is the pm2.5 concentration mean. The higher the population, the higher the concentration. This corroborates previous studies that highlighted that bigger cities have worse air quality. Moreover, two trends seem to appear. First, the *population not staying at home* feature is more important the lower the population is. This underlines that daily population mobility is more important in small city. This could be due to the lack of proximity between habitations and workplace or stores. Similarly, it seems that small cities are more affected by powerplants that bigger one. Indeed, one explanation could be that these plants are built in the countryside to not affect big cities too much, therefore being closer to smaller cities.



POPULATION	cluster 0	cluster 2	cluster 3	cluster 1
	low pop low urban area pop	high pop medium urban area pop	high pop high urban area pop	very high pop very high urban area pop
Pm2.5 mean	27.42	29.77	30.00	38.31
temp_median	0,143	0,237	0,3	0,1
hum_median	0,021	0,028	0,119	0,035
wind-gust_median	0,204	0,177	0,094	0;43
pop_not_home	**0,267**	**0,183**	**0,048**	**0**
pop_home	0,068	0,019	0,004	0,095
pp_feat	**0,207**	**0,054**	**0,008**	**0,01**
mil_miles	0,071	0,014	0,023	0,05

Table 4: feature importance for population clustering

Density clustering



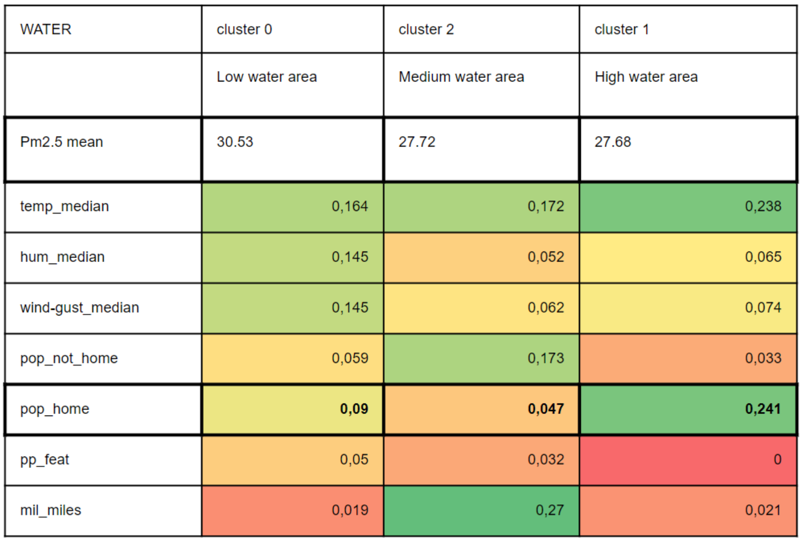
DENSITY	cluster 0	cluster 2	cluster 1
	low density low urban area density	high density low urban area density	high density high urban area density
Pm2.5 mean	29.88	28.95	29.32
temp_median	0,254	0,214	0,057
hum_median	0,116	0,071	0,014
wind-gust_median	0,09	0,103	0,239
pop_not_home	0,051	0,047	0,104
pop_home	0,043	0,038	0
pp_feat	**0,007**	**0,015**	**0,074**
mil_miles	**0,057**	**0,039**	**0,009**

Table 5: feature importance for density clustering

It is harder to interpret this density clustering (Table 5). First of all, it seems that pm2.5 concentration mean is not correlated with the density, which do not follow other studies assessment.

Some trends seem to appear, but they are not as clear as with the population clustering. It seems that high density cities are more prone to power plant pollution and less prone to car pollution. While the latter makes sense as we could expect higher density means closer workplace and stores, the former is more difficult to interpret.

Water area proportion clustering

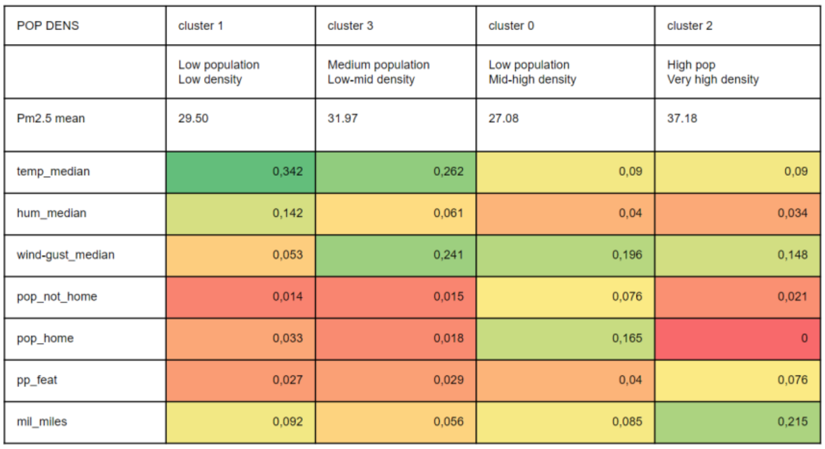


WATER	cluster 0	cluster 2	cluster 1
	Low water area	Medium water area	High water area
Pm2.5 mean	30.53	27.72	27.68
temp_median	0,164	0,172	0,238
hum_median	0,145	0,052	0,065
wind-gust_median	0,145	0,062	0,074
pop_not_home	0,059	0,173	0,033
pop_home	**0,09**	**0,047**	**0,241**
pp_feat	0,05	0,032	0
mil_miles	0,019	0,27	0,021

Table 6: feature importance for water area clustering

Previous studies showed that water has a positive impact and lowers pm2.5 concentration. Table 6 gets the same result, having a have pm2.5 concentration mean in cities with low water area proportion (mostly 0 to 15%). No trend is really highlightable, except from the *population at home* metric., that seem to be very important in cities with a high water area proportion.

Population and density clustering



POP DENS	cluster 1	cluster 3	cluster 0	cluster 2
	Low population Low density	Medium population Low-mid density	Low population Mid-high density	High pop Very high density
Pm2.5 mean	29.50	31.97	27.08	37.18
temp_median	0,342	0,262	0,09	0,09
hum_median	0,142	0,061	0,04	0,034
wind-gust_median	0,053	0,241	0,196	0,148
pop_not_home	0,014	0,015	0,076	0,021
pop_home	0,033	0,018	0,165	0
pp_feat	0,027	0,029	0,04	0,076
mil_miles	0,092	0,056	0,085	0,215

Table 7: feature importance for population & density clustering

This last clustering group, using population and density (table 7), was expected to give meaningful results but is actually really difficult to interpret. Apart from the already well-known results that metropolitan areas (high population and density) have a worse air quality, no other conclusion can be drawn from this clustering.

IV- Implications in Urban Planning

Before diving into urban planning itself, it is interesting to highlight the importance of the weather on pm2.5 concentration. The three first features in the tables (4, 5, 6, 7) that represent temperature, humidity, and wind gust, almost always have a very high feature importance: up to 0.34 for temperature, meaning that the output value changes by 34% when changing the temperature value. While this should not influence urban planning decisions as it is not (yet) controllable, I believe it is a fact that is important to keep in mind when tackling air quality problems.

On an urban planning and design point of view, the study indicates that smaller cities, with lower populations, are more vulnerable to emission from human movement. Therefore, planners should make an effort to limit daily commuting distances (such as going to work or groceries shopping) within these cities to limit the impact of such emissions. Similarly, a larger population is associated with worsened air quality, pushing planners to promote a more dispersed distribution of the population, and limit the exode to metropolitan areas. Moreover, the introduction of more water areas emerges as a potential solution, as it has been found to effectively reduce pm2.5 concentrations.

However, addressing pm2.5 concentrations, and to a larger extent air quality, is difficult and drawing conclusive insight from the project is almost impossible, given the multi-faceted nature of the analysis. Changes in one aspect of urban planning may affect other variables and yield counterproductive results.

Consequently, urban planners need to take a nuanced and comprehensive approach on their work to foster a healthier and more sustainable urban environments.

V- Limitation and further improvements

a- Limitations

This project showed some limitations that were not solved. The most important is about the feature importance evaluation. This evaluation is extremely dependent on the train-test split made for training. Therefore, launching the algorithm again can give very different values of feature importance. While the very clear trends tend to stay the same, this behavior is not acceptable to make usable and serious conclusions on the problematic. Possible improvements are explored in the improvements part.

b- Improvements

Multiple further improvements for the project were pointed up, either while working on the project, or given during the project presentation.

First, the data used for the clustering is very limited. Other metrics would definitely be interesting to take into account, such as taking into account land-use mix or city shape. [12] introduced multiple variables that had a strong correlation to air quality and using them would help drawing better conclusions. As stated earlier in this report, the time limits of the project did not allow for an extensive data collection process, but itwould be encouraged in further research.

Secondly, it was brought up that the subsets of features used in the clustering process could be different. For example, the correlation matrix showed a very strong correlation between the population and the urban area population. Therefore, this clustering might not be appropriate. Moreover, another clustering using area, density and the proportion of water area might be interesting.

On the technical and modelling choices, two different ways of handling the problem were given.

When evaluating the variable significance for each cluster, the model is retrained on the sub-dataset of cities every time. This is useless, as the model could be trained once, and the variable significance evaluated on each sub-dataset. This would solve some stability problems, due to the differences in size of different cluster. The biggest clusters contain around 30 cities, while the smallest ones only contain 3 cities. These small clusters have very limited training size, and the model is not as good when trained on them.

Additionally, the way train/test split is handled might not be optimal. Even if it the data is not considered as a time-series in this study, it is likely that the real data distribution has some time-series feature. Consequently, removing two random months might affect the model because of data leakage: Information used in the training process could be found in the testing process. To avoid this problem, a training using 3-fold or 5-fold cross-validation might allow for a more resilient model.

Finally, recent studies have shown the impressive results of LSTMs on time-series analysis, including in air quality analysis. Using this type of model might results in better performances and in a more reliable way of estimating pm2.5 concentration. A better model would also mean a better and more realistic evaluation of feature importance, thus giving better information for urban planners.

VI- Conclusion

To conclude, this project has some interesting insights on urban form and air quality. It reinforces the findings of previous studies that emphasize the pivotal role of urban form in the management of air quality. Notably, areas with higher population densities consistently exhibit poorer air quality, aligning with established trends in urban research. However, divergences from other studies become apparent in certain aspects, particularly in the observation that density has a low impact on PM2.5 concentration. These nuances prompt a cautious interpretation of the study's conclusions. The significance of each feature is shown to be highly dependent on the specific train/test set of each cluster, making it challenging to discern a clear and overarching trend. Consequently, each drawn conclusion warrants further in-depth analysis to determine its validity and practical applicability.

As a proof of concept, this study serves as a foundation for future investigations into the intricate links between urban form and air quality. The suggestion to explore land-use mix metrics emerges as a potential avenue for gaining deeper insights into the complex relationships at play. Moreover, incorporating state-of-the-art models like LSTM (Long Short-Term Memory) is proposed as a means to enhance the stability and realism of feature importances, thereby advancing the precision of air quality management models in urban planning. In light of these considerations, this study encourages a nuanced and evolving approach to understanding and improving the connections between urban structure and air quality dynamics.

The introduction and conclusion of this project report were enhanced using ChatGPT 3.5.

[1] Zhang, W., Wu, Y., & Calautit, J. K. (2022). A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment. *Renewable and Sustainable Energy Reviews*, *167*, 112704.

[2] Sotomayor-Olmedo A, Aceves-Fernandez MA, Gorrostieta-Hurtado E, Pedraza-Ortega JC, Vargas-Soto JE, Ramos-Arreguin JM, Villaseñor-Carillo U (2011) Evaluating trends of airborne contaminants by using support vector regression techniques. In: CONIELECOMP 2011, 21st International Conference on Electrical Communications and Computers, pp 137–141. https://doi.org/10.1109/CONIELECOMP.2011.5749350

[3] García Nieto PJ, Combarro EF, Del Coz Díaz JJ, Montañés E (2013) A svm-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): a case study. Appl Math Comput 219(17):8923–8937. https://doi.org/10.1016/j.amc.2013.03.018

[4] Zhang C, Yuan D (2015) Fast fne-grained air quality index level prediction using random forest algorithm on cluster computing of spark. In: 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), pp 929–934 . https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.177

[5] Barthwal A, Acharya D (2018) An internet of things system for sensing, analysis forecasting urban air quality. In: 2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), pp 1–6 . https://doi.org/10.1109/CONECCT.2018.8482397

[6] Madan, T., Sagar, S., & Virmani, D. (2020, December). Air quality prediction using machine learning algorithms–a review. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 140-145). IEEE.

[7] Agarwal S, Sharma S, Suresh R, Rahman MH, Vranckx S, Maiheu B, Blyth L, Janssen S, Gargava P, Shukla VK, Batra S (2020) Air quality forecasting using artifcial neural networks with real time dynamic error correction in highly polluted regions. Sci Total Environ 735:139454. https://doi.org/10.1016/j.scitotenv.2020.139454

[8] Schürholz D, Nurgazy M, Zaslavsky A, Jayaraman PP, Kubler S, Mitra K, Saguna S (2019) Myaqi: Contextaware outdoor air pollution monitoring system. In: Proceedings of the 9th International Conference on the Internet of Things. IoT 2019, pp 1–8. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3365871.3365884D

[9] Liu, C. C., Lin, T. C., Yuan, K. Y., & Chiueh, P. T. (2022). Spatio-temporal prediction and factor identification of urban air quality using support vector machine. Urban Climate, 41, 101055.

[10] Lee, H., Lee, J., Oh, S., Park, S., & Mayer, H. (2023). Air pollution assessment in Seoul, South Korea, using an updated daily air quality index. Atmospheric Pollution Research, 14(4), 101728.

[11] Squizzato, S., Cazzaro, M., Innocente, E., Visin, F., Hopke, P. K., & Rampazzo, G. (2017). Urban air quality in a mid-size city—PM2. 5 composition, sources and identification of impact areas: From local to long range contributions. Atmospheric Research, 186, 51-62.

[12] Liang, L., & Gong, P. (2020). Urban and air pollution: a multi-city study of long-term effects of urban landscape patterns on air quality trends. Scientific reports, 10(1), 18618.

[13] Bhattacharyya, M., Nag, S., & Ghosh, U. (2021). Deciphering Environmental Air Pollution with Large Scale City Data. arXiv preprint arXiv:2109.04572.