Pollution forecasting using LSTM

Isam Al Jawarneh
College of Computing & Informatics
University of Sharjah
Sharjah, United Arab Emirates
ijawarneh@sharjah.ac.ae

Madyan Omar Bagosher
College of Computing & Informatics
University of Sharjah
Sharjah, United Arab Emirates
U23200049@sharjah.ac.ae

Fatemeh Mohammadi Aghjehmashhad
College of Computing & Informatics
University of Sharjah
Sharjah, United Arab Emirates
U23200047@sharjah.ac.ae

***Abstract*—With the increase in factories and cars, air pollution has become a significant concern. This study presents a predictive analysis of PM pollution (PM2.5) in New York City using Long Short-Term Memory (LSTM) neural networks to forecast 24-hour pollution levels at each data collection location. We computed several statistical metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), for the LSTM model, other machine learning models and networks, such as RNN, MLP, SVR, and Linear Regression. A comparison of the results indicates that the LSTM model outperformed other models in predicting pollution levels.**

***Index Terms*—LSTM, Pollution, Forecasting, Analysis, NYC**

I. Introduction

Air quality describes the level of various pollutants in the air at a specific time and location [1]. These pollutants include gases like Sulphur Dioxide (SO2), Ammonia (NH3), Nitrogen Dioxide (NO2), Carbon Dioxide (CO2), Carbon Monoxide (CO), Ozone (O3), and particulate matter (PM 2.5 and PM 10). Research indicates that short-term exposure to high levels of these pollutants can cause respiratory difficulties, eye irritation, and may affect heart and lung health. Prolonged exposure can result in cancer and harm the body's respiratory, reproductive, neurological, and immune systems. Therefore, forecasting air quality and collecting real-time data on it are crucial fields of study.

This paper explores the use of Long Short-Term Memory (LSTM) neural networks to predict PM pollution levels across different locations in New York City. LSTM networks are a type of deep learning model that are particularly suited for time series forecasting because of their ability to capture long-term dependencies in sequential data. By analyzing historical air quality data collected at various points in the city, we aim to develop a model that can accurately forecast pollution levels 24 hours in advance.

To assess the effectiveness of our approach, we compare the predictive accuracy of the LSTM model against other common machine learning models and neural networks including: Recurrent Neural Networks (RNN), Multi-Layer Perceptrons (MLP), Support Vector Regression (SVR), and Linear Regression. We evaluate these models using several statistical metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Several studies have previously forecasted pollution levels, as noted in references [2] and [3]. However, this study distinguishes itself by using geohash locations to predict pollution levels. Specifically, we utilize data on particulate matter (PM) pollution from previous hours at each location to forecast pollution levels for the upcoming 24 hour.

II. Related Work

A. Pollution level prediction using Machine Learning

The regulation of air pollutant levels is a major concern, necessitating constant surveillance. With the proliferation of urbanization, the imperative to control pollution and identify areas of high pollution escalates. Algorithms, particularly in the realm of machine learning, have revolutionized pollution research, empowering researchers to uncover significant insights into pollution levels across cities and providing them with the tools necessary for thorough investigation. Aditya et al. found that using Logistic Regression and Autoregressive based models performed efficiently to detect the quality of air and predict the level of PM2.5 [4].

B. LSTM method of Pollution forecast in China

China is one of the most populated countries in the world. Thus, it is of utmost importance to ensure the health and safety of the citizens and keep pollution levels to a minimum. In a study conducted by Cheng et al., they used an LSTM neural network to predict pm10 levels in 5 representative cities in China: (Beijing, Taiyuan, Shanghai, Nanjing and Guangzhou). They concluded that their model showed excellent adaptability for various regions in China with different geographical conditions and PM10 characteristics, as they managed to obtain prediction accuracies ranging between (80% - 90%) [5].

III. Methodology

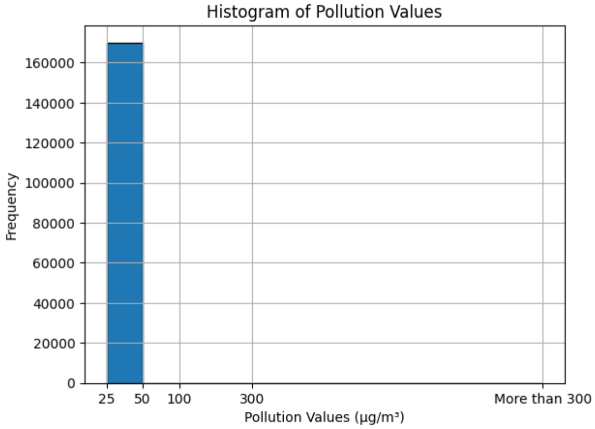
A. Dataset & Data preparation

Our data analysis process began with the acquisition of two distinct datasets: one detailing pollution levels across New York City in CSV format, and the other comprising a GeoJSON file delineating the city's map. Upon initial inspection, the original dataset boasted 31 features, yet to streamline our analysis, we pruned this down significantly, discarding 25 features deemed extraneous to our objectives. From the remaining pool, we carefully selected six pertinent features: Timestamp, Latitude, Longitude, Temperature, Humidity, and PM2.5 (Particulate Matter). However, recognizing the value of geographical context, we opted to enrich our dataset through a geospatial join with the GeoJSON file. This integration augmented our feature count to nine, with the inclusion of essential geographic attributes such as Borough, Borough Code, and Neighborhood. Furthermore, seeking to enhance the dataset's utility for visualization and in-depth analysis, we introduced an additional feature: the calculation of geohashes derived from Latitude and Longitude coordinates. This augmentation not only facilitated spatial analysis but also provided valuable insights into the spatial distribution of pollution levels across the city.

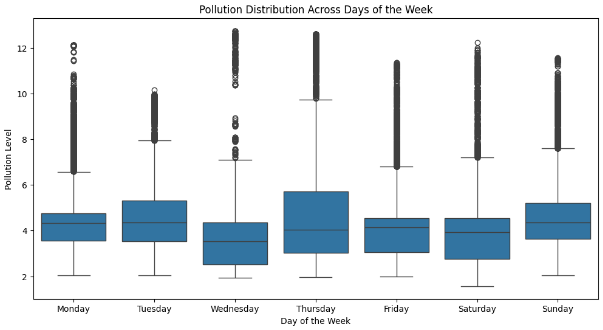
B. Data Analysis and Visualization

Various components can be considered when determining air quality. In our data, we focus on "pollution" and the geographic locations in New York City. The maximum pollution value and the minimum value are 12.74 and 1.57, respectively. The following picture is a histogram of pollution values which help us to understand the overall distribution of pollution levels.

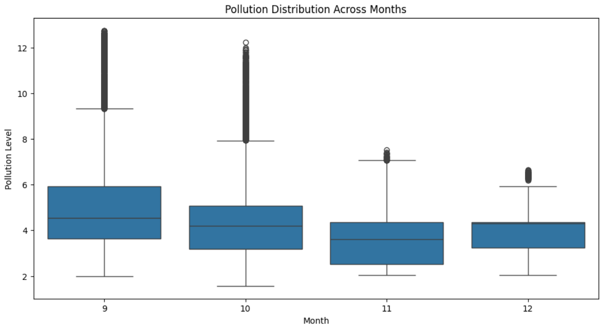
The histogram below shows a left-skewed (or negatively skewed) distribution, indicating that the majority of the pollution values are low, with fewer instances of very high pollution levels. The highest bar is in the first interval, suggesting that most pollution measurements are at the lower end of the scale, possibly within clean to moderately polluted air quality ranges. The skewness of the data's distribution suggests that median value would be more representative of typical pollution levels than mean value.



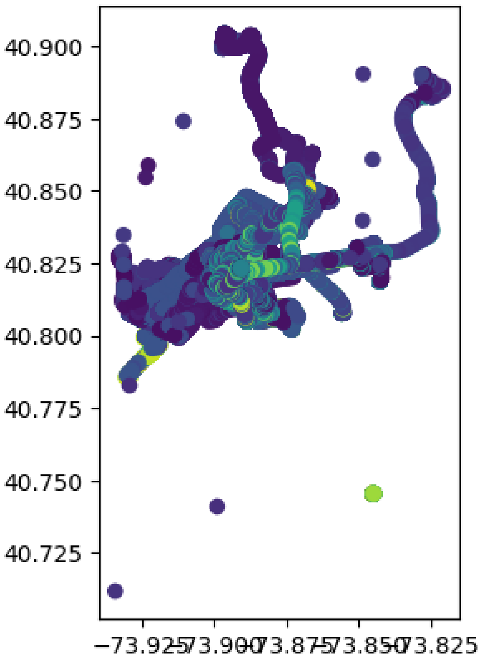
Also, we can see the pollution distribution across days of the week in the following picture. It shows that the median pollution level for all days are almost similar . These appear fairly consistent across the days, suggesting a similar central tendency for pollution regardless of the day of the week. The IQRs seem relatively consistent, though some days like Thursday show a slightly larger IQR, indicating more variability.



Additionally, we can see the distribution of pollution levels across four months —September (9), October (10), November (11), and December (12) in the below picture. It is worth noting that the median appears relatively stable across the months, with slight variations. Also, the height of the boxes, which represent IQR, suggests that September and October have a broader range of pollution levels compared to November and December. September shows the most variability with the widest box and whiskers extending further, indicating more sporadic pollution levels. In contrast, December shows the least variability with a smaller IQR and shorter whiskers.



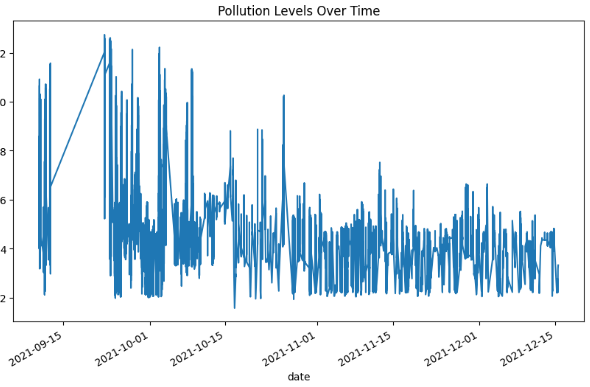
Also, the following map can help us to visualize the spatial distribution of pollution levels:



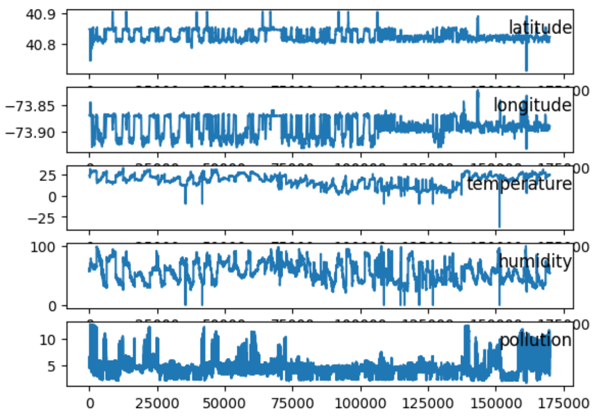
We can see areas with high and low pollution concentration. In fact, the concentration patterns suggest that pollution is not evenly distributed across the city but is rather concentrated in specific areas like urban areas, industrial areas and etc. More precisely, there are specific areas where pollution levels are particularly high. These are indicated by the regions with warmer colors on the density scale. Such visual cues often point to urban centers, busy roads, industrial areas, or other zones with high human activity that could contribute to pollution. Also, there are areas with cooler colors, suggest places where pollution is less concentrated. These might correspond to parks, bodies of water, less developed, or more residential areas with fewer pollution sources.

The data was covered from 2021-09-10 at 12:29:09 to 2021-12-15 at 14:35:55 which makes 96 days and 02:06:46.

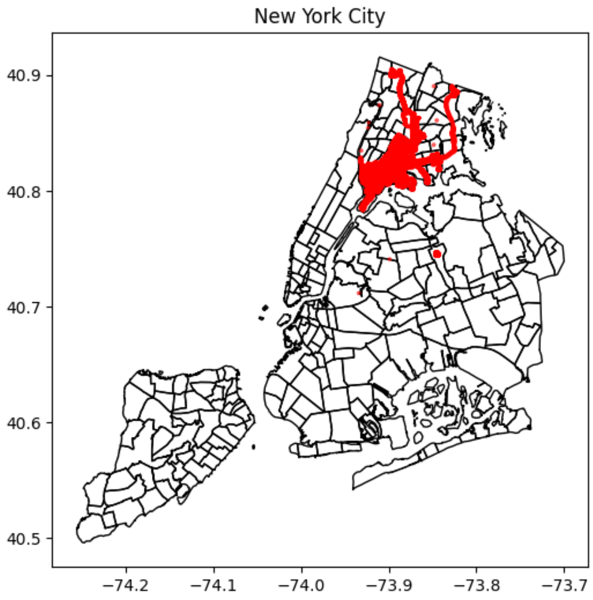
For investigating temporal trend of pollution, we plot the fluctuation of pollution levels over time in the following figure:



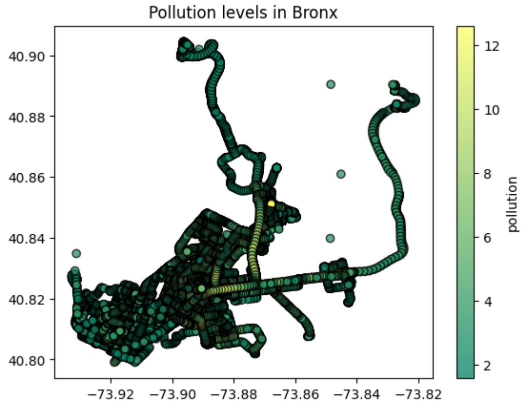
The plot shows a significant variation in pollution levels over time. There are visible peaks and troughs indicating that on some days, pollution levels were much higher than on others. There appears to be a downward trend in pollution levels as time progresses. Initially, there are more instances of higher peaks, which seem to become less frequent towards the end of the time series. The sharp spikes, especially noticeable at the beginning of the time series, suggest transient events that caused pollution levels to shoot up. These could be due to specific incidents, like industrial releases, traffic congestion, or other episodic events. The continuity of the data indicates consistent monitoring without apparent gaps in data collection, which is good for analysis. We also provide the graph for features in our dataset such as: Latitude, Longitude, temperature, humidity and pollution as follow:

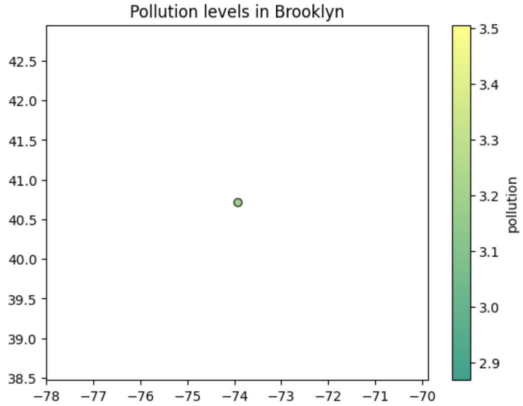


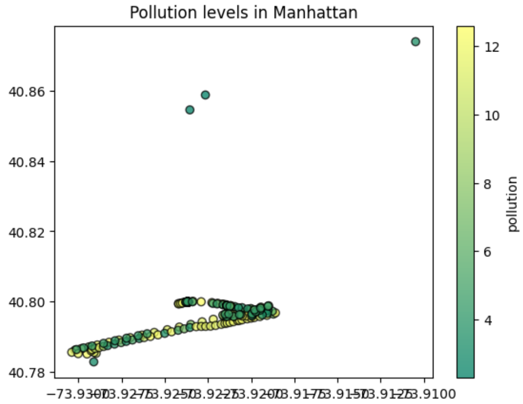
First, by creating the scatter plot on top of a map of New York City, we managed to visualize pollution data. In fact, the following map provides spatial context for understanding the distribution of pollution within New York City, where each point on the map represents a pollution data point, and the size of the point reflects the pollution level.

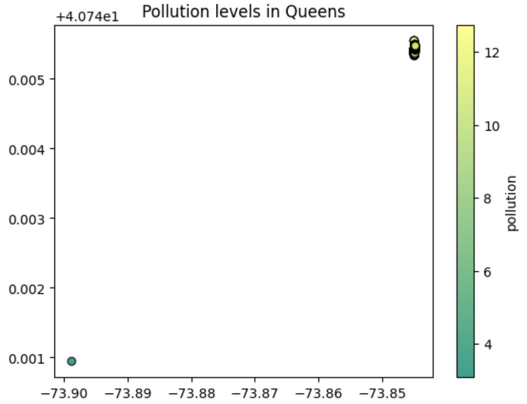


From this map, we can see where pollution is concentrated and how it spreads across different areas within the city. Additionally, the following scatter plots help us to visualize pollution levels in each area.





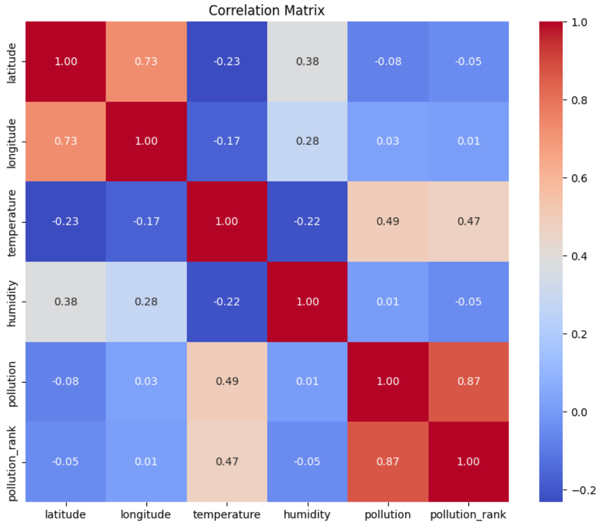




We recall that a correlation matrix is a table that shows the correlation coefficients between variables in a dataset. Each cell in the table represents the correlation coefficient between two variables. The correlation coefficient measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship
- -1 indicates a perfect negative linear relationship
- 0 indicates no linear relationship

We calculated the correlation matrix for our data and it looks like the "Temperature" is the variable that has the highest correlation with the response variable "Pollution", at the par with 0.49. This suggests that higher temperatures might be associated with higher levels of pollution. 'Temperature' and 'humidity' show a moderate negative correlation of -0.22, which could indicate that higher temperatures often correspond to lower humidity levels. Several pairs of variables, like 'humidity' and 'pollution rank' (-0.05), show little to no correlation, indicating no linear relationship between these variables.



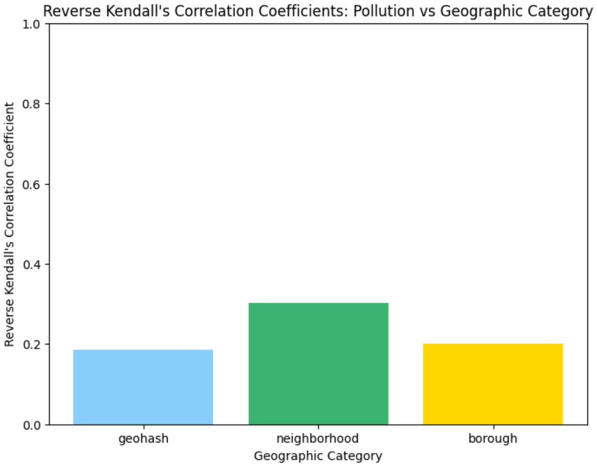
IV. Feature Selection

In this section, we want to investigate any association between pollution (also pollution rank) with geohash variable using several methods:

A. Reverse Kendall's correlation coefficient

We calculated the average rank of pollution values per geohash category as well as the size of each geohash category (i.e., the number of observations in each category).Since both are ordinal data, it is suitable to use Reverse Kendall's correlation coefficient in order to find the association between the average rank of pollution values (ordinal data) and the size of each geohash category (ordinal data). Also, we repeated this strategy by using the neighborhood variable and borough variable instead of geohash variable, and we got the following results:

- Using "geohash" variable: 0.18607703406282353
- Using "neighborhood" variable: 0.3031491751402424
- Using "borough" variable: 0.19999999999999998



One can see that the association between the average rank of pollution values per neighborhood category and the size of each neighborhood is positive and is higher than the other. Larger neighborhoods, which likely correlate with higher population density or more intense urban activities, tend to have higher pollution levels. This suggests that as the activity and population in a neighborhood increase, so does the pollution, potentially due to factors like traffic, industrial emissions, and commercial activities. This would indicate that larger neighborhood areas tend to have higher pollution levels on average.

Also, it is notable that the size of geohash areas and the average pollution rank shows the weakest association, suggesting that geohashes do not align as closely with factors that influence pollution levels. Since geohashes are arbitrary geographic divisions, they might not correspond well with natural or urban boundaries that impact pollution, such as roads, industrial zones, or natural barriers.

Also, we repeated the above algorithm with different precision=[3, 4, 5]. We got the following results:

- Using "geohash" variable: 1.0
- Using "neighborhood" variable: -0.10734744162189888
- Using "borough" variable: 0.39999999999999997

We got the same result for each precision in [3,4,5], and this can suggest that additional precision beyond a certain point doesn't significantly alter the ordering of the geohash (respectively, neighborhood and borough) categories concerning pollution ranks. As we can see the association between the average rank of pollution values per geohash category and the size of each geohash is positive and perfect. This implies that within each truncated geohash (precision reduced to 3, 4, 5), there is a perfect alignment between the size of the geohash category and the average pollution rank. This makes sense, since the geohash precision is reduced, each category might encompass a smaller number of data points or a more uniform area. This creates a scenario where the more observations within a geohash, the higher the rank of pollution observed, simply due to uniform distribution of few varied data points.

On the other hand, the correlation coefficient of -0.107 between the average rank of pollution values per neighborhood category and the size of each neighborhood indicates a very slight negative correlation, suggesting that as the number of observations in neighborhoods increases, the average pollution rank slightly decreases. This might imply that larger neighborhoods (or those with more data points) do not necessarily have higher pollution levels, possibly due to effective pollution control, larger green spaces, or less dense industrial activity.

B. Spearman's Rank Correlation Coefficient

This measures the strength and direction of association between the ranks of 'geohash' (respectively, neighborhood and borough) and 'pollution'. It is suitable when the relationship is monotonic but not necessarily linear. We remind that for calculating the Spearman's Rank Correlation Coefficient, one can use the following formula:

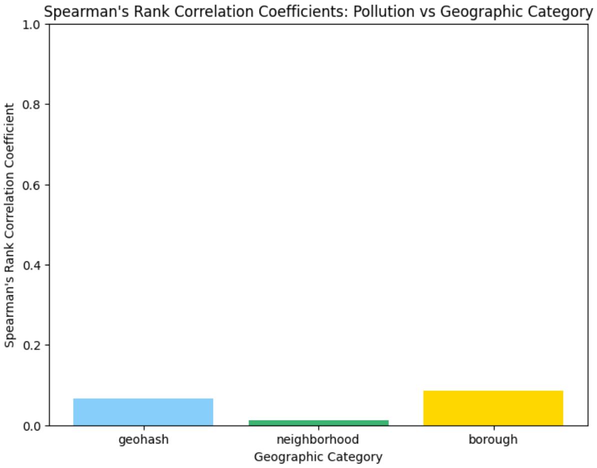
$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where ρ is Spearman's rank correlation coefficient. $\sum d^2$ is the sum of the squared rank differences, and n is the number of observations. This formula assesses how well the relationship between two variables can be described using a monotonic function. If $\rho = 1$, there is a perfect positive monotonic relationship, and if $\rho = -1$, there is a perfect negative monotonic relationship. A ρ close to 0 suggests no monotonic relationship. First, we calculated the Spearman's Rank Correlation Coefficient between 'geohash' (respectively, neighborhood and borough) and 'pollution' without applying precision, and we got the following:

- Geohash vs Pollution: 0.06654169631677545
- Neighborhood vs. Pollution: 0.011637772263997271
- Borough vs. Pollution: 0.08611017903128607

P-values:

- Geohash vs Pollution: 4.454497640901313e-166
- Neighborhood vs. Pollution: 1.5984415142310416e-06
- Borough vs. Pollution: 4.068933168365492e-277



From the result, one can see that overall there is generally a weak association between the ranks of geographic categorizations (geohash, neighborhood, and borough) and pollution. For example, the correlation coefficient of -0.066 indicates a very weak negative monotonic relationship between geohash and pollution. This suggests that increases in geohash ranks are weakly associated with decreases in pollution ranks, but the relationship is not strong or may not be significant. The extremely small p-value (approximately 4.454×10^{-16}) suggests that the result is statistically significant, even though the correlation itself is weak.

Second, we calculated the Spearman's Rank Correlation Coefficient between the ranks of 'geohash' and 'pollution' with different precision=[3, 4, 5]. The results are as follow:

- Precision 3: -0.06654169631677545
- P-value: 4.454497640901313e-166
- Precision 4: -0.03478746127459427
- P-value: 1.1042075713576446e-46
- Precision 5: 0.005006042655342761
- P-value: 0.03901315711545015

From the results, we can conclude that as the precision of the geohash increases from 3 to 5, the magnitude of the correlation coefficient decreases, moving closer to zero. This suggests that as geohash precision increases—thereby defining smaller and more localized geographic areas—the association between geohash and pollution levels becomes weaker.

C. ANOVA (Analysis of Variance)

ANOVA, or Analysis of Variance, is a statistical method used to analyze the differences among means of three or more groups. It assesses whether the means of different groups are statistically significantly different from each other. ANOVA tests the null hypothesis that all group means are equal against the alternative hypothesis that at least one group mean is different.

We need the F-statistic and p-value for the ANOVA test. We recall that the F-statistic measures the ratio of the variance between groups to the variance within groups. And the p-value indicates the probability of obtaining the observed results (or more extreme results) if the null hypothesis is true. The decision rule is that if the p-value is less than a chosen significance level (e.g., 0.05), we can reject the null hypothesis and conclude that there are statistically significant differences in pollution levels across different geohash (respectively, neighborhood and borough) categories. Otherwise, we fail to reject the null hypothesis, suggesting no significant differences in pollution levels between the groups.

First, we show the ANOVA test results for 'geohash' (respectively, neighborhood and borough) and 'pollution' without applying precision:

- Neighborhood - F-statistic: 194.2967984990712, P-value: 0.0
- Borough - F-statistic: 1571.4452929273557, P-value: 0.0
- Geohash - F-statistic: 3520.2944091142294, P-value: 0.0

These results provide a clear indication that pollution levels vary significantly across different geographic scales (geohash, borough, and neighborhood).

More precisely, the F-statistic for geohash is the highest (3520.2944091142294), indicating that the differences in mean pollution levels among different geohashes are the most pronounced among the three categorizations. This suggests a strong spatial variability in pollution distribution at the geohash level.

On the other hand, the F-statistic for borough (1571.4452929273557) is substantial but lower than that for geohash, indicating significant but less pronounced differences among boroughs compared to geohashes.

It is noticeable that neighborhoods, while still showing significant differences with an F-statistic of 194.2967984990712, exhibit the least variability among the three. This suggests that pollution levels within neighborhoods are more homogeneous compared to the larger categorizations of boroughs and geohashes.

These facts are supported by the p-value as well. A P-value of 0.0 is highly statistically significant. This result provides strong evidence to reject the Null hypothesis, indicating that there are significant differences in pollution levels between the different geographic categorizations.

Second, we show the ANOVA test results for 'geohash' and 'pollution' with applying precision=[3, 4, 5]:

- Precision Level '3': F-statistic = 3520.2944091142294, P-value = 0.0
- Precision Level '4': F-statistic = 1762.629585813516, P-value = 0.0
- Precision Level '5': F-statistic = 412.3721028769785, P-value = 0.0

One can see that as the precision level of geohash increases, the F-statistic decreases. This indicates that as the geohash codes become more fine-grained (moving from precision 3 to 5), the differences in mean pollution levels between geohash categories become less pronounced. For example at Precision 5, the F-statistic is significantly lower, indicating that the more localized the geohash, the less variance there is among the categories in terms of pollution levels.

The P-value of 0.0 across all precision levels indicates that the differences observed in mean pollution levels are statistically significant allowing us to confidently reject the null hypothesis which states there are no differences.

D. Kruskall-Wallis test

The Kruskal-Wallis test is a non-parametric method used to determine if there are statistically significant differences between the medians of two or more independent groups. It is especially useful when the assumptions required for ANOVA test are not met, such as when the data are not normally distributed or when the groups have unequal variances.

First, we show the Kruskal-Wallis test results for 'geohash' (respectively, neighborhood and borough) and 'pollution' without applying precision:

- Neighborhood: Statistic = 3194.7088275131123, P-value = 0.0
- Borough: Statistic = 1281.6427998977276, P-value = 3.178456778603967e-276
- Geohash: Statistic = 752.7255492814534, P-value = 1.025109961523071e-165

In all cases, the extremely small P-values indicate that the results are statistically significant, allowing us to reject the null hypothesis that the medians of pollution levels across different categories are equal. This suggests that there are significant differences in pollution levels between the groups defined by neighborhood, borough, and geohash.

In the case of comparison, neighborhood has the highest statistic among the three categories which indicates the strongest degree of differences among the groups within this categorization. This suggests that neighborhoods vary greatly in terms of pollution levels.

On the other hand, geohash has the lowest statistic. It suggests that while there are still significant differences in pollution levels across different geohashes, these differences are less pronounced than those observed at the neighborhood and borough levels.

Second, we show the Kruskal-Wallis test results for 'geohash' and 'pollution' with applying precision=[3, 4, 5]:

- Precision Level '3': Statistic = 752.7255492814534, P-value = 1.025109961523071e-165
- Precision Level '4': Statistic = 767.5797478376305, P-value = 2.0997884702223404e-167
- Precision Level '5': Statistic = 1549.7445498734908, P-value = 0.0

It is clear the p-values are very small in all three precision levels. This means we can reject the null hypothesis that states there are no differences in the median pollution levels across different geohash categories at each respective precision level.

Interestingly, the Kruskal-Wallis statistic significantly increases from precision level 4 to 5. While the statistic initially increases slightly from level 3 to 4, the jump to level 5 is much more noticeable. The increase in the Kruskal-Wallis statistic with finer geohash precision suggests that as we zoom in further (higher precision), the differences in pollution levels across smaller geographical areas become more noticeable.

E. Kendall Tau's Correlation Coefficient

This measures the strength and direction of association between the ranks of 'geohash' and 'pollution'. It is suitable for ordinal data or when the relationship is not necessarily linear.

Kendall's Tau is a rank-based correlation coefficient that measures the ordinal association between two variables. We calculated the Kendall's Tau correlation coefficient and the associated p-value. The correlation coefficient ranges from -1 to 1. The p-value indicates the significance of the correlation coefficient. If the p-value is less than a chosen significance level (e.g., 0.05), we can reject the null hypothesis and conclude that there is a statistically significant association between 'geohash' and 'pollution'. Otherwise, we fail to reject the null hypothesis, suggesting no significant association. We got the following results by applying precision=[3, 4, 5]:

KENDALL TAU'S CC RESULTS FOR NEIGHBORHOOD:

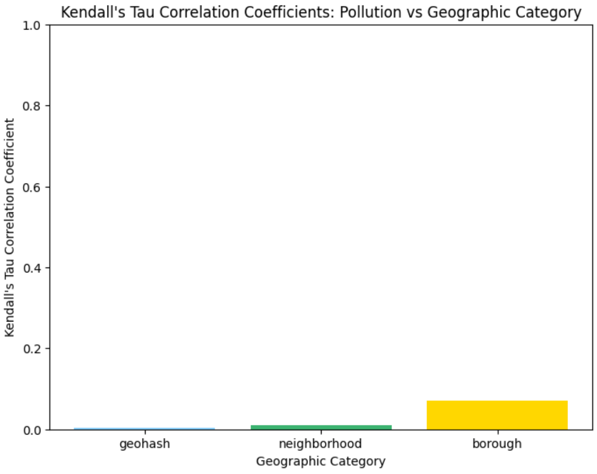
- For neighborhood: Correlation = 0.009259718338660744, P-value = 1.3397951600462843e-07

KENDALL TAU'S CC RESULTS FOR BOROUGH:

- For borough: Correlation = 0.07034127731459229, P-value = 4.592161998189867e-276

KENDALL TAU'S CC RESULTS FOR GEOHASH:

- For geohash Precision Level '3': Correlation = -0.054383698700505266, P-value = 1.0251099615631986e-165
- For geohash Precision Level '4': Correlation = -0.028405622420667093, P-value = 1.2769262199553581e-46
- For geohash Precision Level '5': Correlation = -0.003214327504560284, P-value = 0.08134248790613795



One can see that for two categories neighborhood and geohash the correlation is very close to zero but positive, suggesting a very weak direct relationship between geohash rank (res. neighborhood rank) and pollution levels. The association between borough and pollution is higher than the others, however still is a small value.

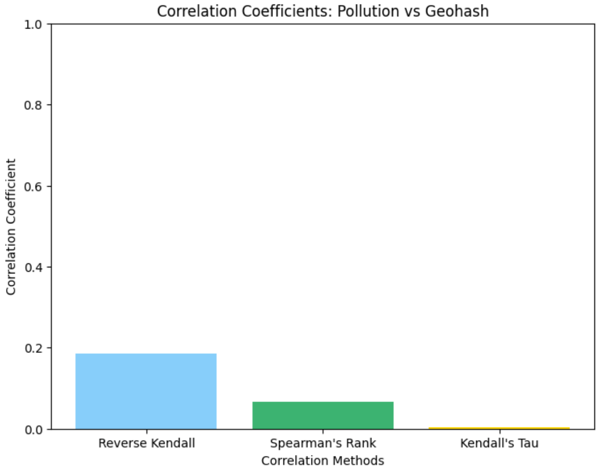
In the case of geohash, we can say that as precision increases from Level 3 to Level 5, the correlation decreases from a weak negative to almost negligible negative. This trend suggests that the inverse relationship between geohash rank and pollution weakens significantly as the precision increases. At precision level 5, we see that p-value=0.08 and is not less than the significance level= 0.05, and so we do not reject null hypothesis. This means that there is no significant association between 'geohash' and 'pollution' at this level.

F. Comparison Between Correlation coefficient obtained from different methods

In the last section, we calculated several correlation coefficients with different methods. In this chapter, we want to compare them thorough geographic category. We compare the results obtained from the methods such as Reverse Kendall, Spearman's Rank, Kendall's Tau for geohash, neighborhood and borough.

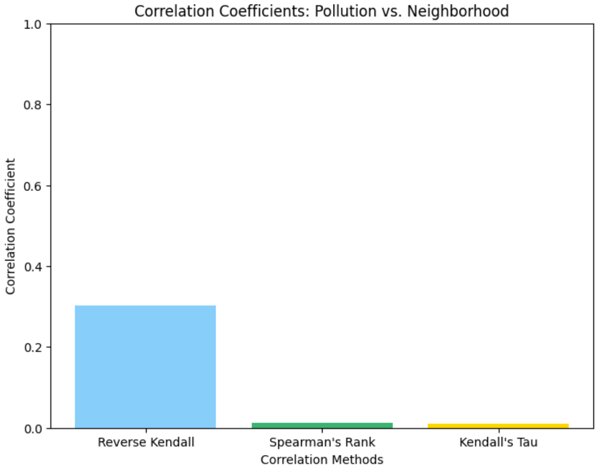
1. Pollution vs. geohash

As we can see from the below bar chart, the correlation coefficient that gained from Reverse Kendall is higher than the two other methods in the geohash geographical category. This implies the higher association between the level of pollution and geohash variable with respect to the other methods. The Spearman's Rank method also shows a moderate association between the level of pollution and geohash variable and Kendall's Tau methods shows almost zero correlation between the level of pollution and geohash.



1. Pollution vs. neighborhood

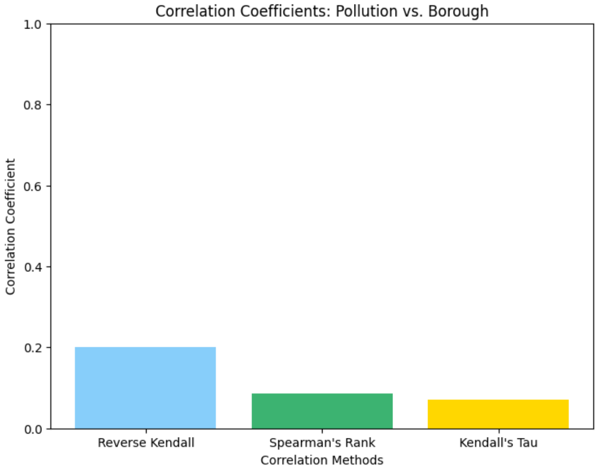
The below bar chart shows the correlation coefficient that obtained from Reverse Kendall is higher than the two other methods in the neighborhood geographical category. This implies the higher association between the level of pollution and neighborhood variable with respect to the other methods. Two other methods show almost zero correlation between the level of pollution and the neighborhood geographical category.



3. Pollution vs. Borough

One can see from the below bar chart that the correlation coefficient that gained from Reverse Kendall is higher than the two other methods in the borough geographical category.

This implies the higher association between the level of pollution and borough variable with respect to the other methods. The two other methods almost show the same correlation coefficient between the level of pollution and borough.



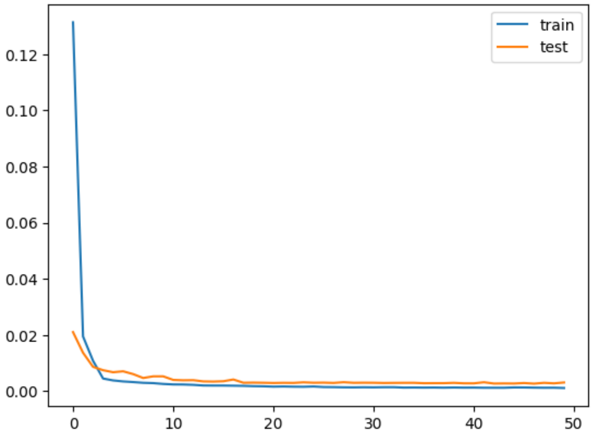
G. Training the proposed model

Recall that our dataset contained 10 features, we only utilized 5 features to train our model: Latitude, Longitude, Temperature, Humidity and Pollution (Target Variable).

The data that we worked on is known as fine-grained low cost air quality (AQ) data. This data is geo-referenced. We used this data and framed a forecasting problem where, given the pollution (PM, particulate matters) for prior hours, we forecast the pollution at the next 24 hour. The data is a time-series data that we have for several locations, actually street-by-street level locations (granular piecimel precision). We reframed the problem so that it is at each location (represented by a geohash), given the pollution (PM, particulate matters) for prior hours, we forecast the pollution at the next 24 hour.

We split our data into train and test sets. We built a model specifically skilled at forecasting. After training it on one year of data, we checked its ability to predict pollution for the next four years. This way, we made sure the model wasn't just memorizing; it was truly learning and adapting to make accurate predictions over a longer period.

The following graph shows the training loss and validation loss over the course of 50 epochs from fitting a Long Short-Term Memory (LSTM) neural network.



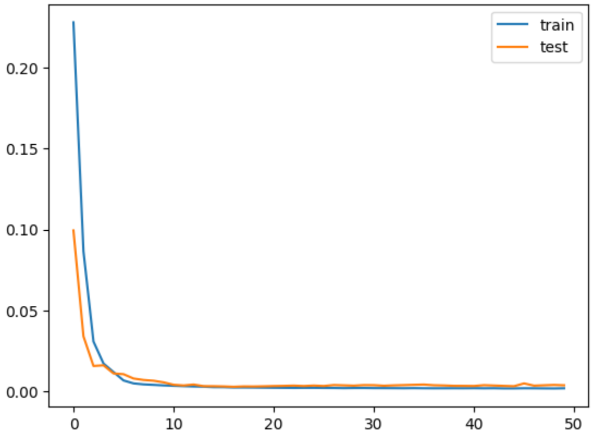
The graph shows that both training and validation loss show a sharp decrease in the initial few epochs, which indicates that the model is learning from the data. After the initial sharp decrease, both lines flatten out, suggesting that the model has reached a point where making further improvements is harder. This means that the model has started to converge on a solution. The final values of the loss functions (training and validation) are very close to each other and close to zero. The close proximity of the training and validation loss suggests that the model is generalizing well and thus should perform similarly on new, unseen data.

H. Prediction using LSTM

By tailoring our model to each specific location identified by its geohash, we've created a series of forecasts that predict the levels of particulate matter (PM) pollution for the next 24 hours. Each location's data, shaped by its unique environmental factors and patterns, helps the model to "learn" from the past hours of pollution data and make these hourly predictions. We forecast the pollution at the next 24 hours.

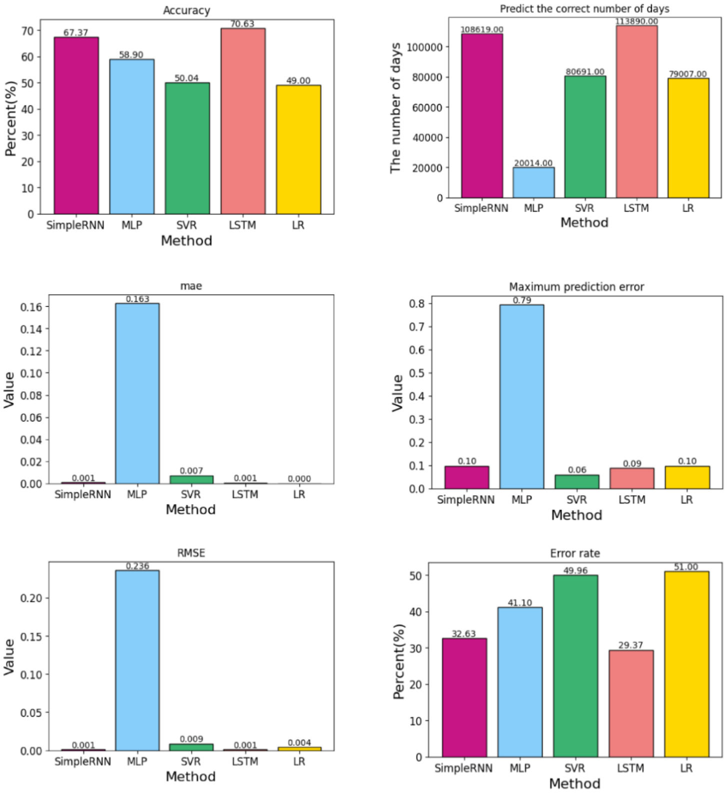
I. Other models used

We also trained several other models and networks such as: Support vector regression model, Linear regression model, Multi Layer perceptron and Recurrent Neural network. Our RNN achieved results that are almost on par with the proposed model. The RNN's training graph is displayed below:



V. Results

During the evaluation process, we calculated various metrics such as the mean absolute error (MAE) which is a measure that tells us the average magnitude of the errors in our forecasts. Our intention was to prove that the our proposed LSTM model is the most efficient in forecasting compared to the other models used. Below are some of the metrics calculated for each model:



The MLP neural network, while adept at handling complex relationships within data, falls short in time series forecasting tasks. Both machine learning models exhibited poor performance due to their inadequacy in dealing with complex relationships and sequential data. Although the RNN model's results were comparable to LSTM, it still lacked effectiveness, likely because it struggled to capture long-term dependencies during training. Our proposed LSTM model achieved the highest accuracy of 70.63% and the lowest RMSE score of 0.001. The model comprised 4 memory cells and underwent training for 50 epochs with a batch size of 32. Below is a table summarizing the evaluation of each model:

TABLE I
MODEL EVALUATION

Model	MAE	MSE	RMSE	Max Prediction Error	Accuracy
LSTM	0.00060	0.00000	0.001	0.088	70.63
MLP	0.16276	0.05552	0.236	0.793	58.90
RNN	0.00069	0.00000	0.001	0.097	67.37
SVR	0.00722	0.00008	0.009	0.060	50.04
LR	0.00011	0.00000	0.004	0.097	49.00

VI. CONCLUSION

We calculated several correlation , and we saw that most of the time Reverse Kendall's correlation coefficient shows the higher correlation between the level of pollution and geographic categories. Also, we calculated several statistical metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) for our proposed LSTM model and other models and networks such as: RNN, MLP, SVR, and Linear Regression. A comparison between the obtained values indicates that the LSTM model was better at predicting the pollution.

REFERENCES

[1] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, "Real-time air quality forecasting, part i: History, techniques, and current status," *Atmospheric Environment*, vol. 60, pp. 632–655, 2012.

[2] J. Brownlee, "Multivariate time series forecasting with lstms in keras," Oct 2020. [Online]. Available: https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/

[3] J. Wang, J. Li, X. Wang, J. Wang, and M. Huang, "Air quality prediction using ct-lstm," *Neural Computing and Applications*, vol. 33, pp. 4779–4792, 2021.

[4] C. Aditya, C. R. Deshmukh, D. Nayana, and P. G. Vidyavastu, "Detection and prediction of air pollution using machine learning models," *International journal of engineering trends and technology (IJETT)*, vol. 59, no. 4, pp. 204–207, 2018.

[5] Y. Chen, S. Cui, P. Chen, Q. Yuan, P. Kang, and L. Zhu, "An lstm-based neural network method of particulate pollution forecast in china," *Environmental Research Letters*, vol. 16, no. 4, p. 044006, 2021.