Data-Driven Modeling of Urban Air Quality Using Traffic Congestion and Wind Speed

Reza Golshan Khavas[1*], Abdolreza Sheikholeslami[2], Maryam Niazi Rad[3]

[1] Assistant Professor, School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran
[2] Associate Professor, School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran
[3] M.Sc. Graduate in Transportation Engineering, School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran

ABSTRACT

Air pollution is a major environmental and public-health concern in urban areas, where road traffic is an important emission source and meteorological conditions influence pollutant dispersion. This study develops a data-driven framework for short-term prediction of carbon monoxide (CO), particulate matter (PM_{10}), and sulfur dioxide (SO_2) using a traffic congestion index and wind speed. Traffic conditions were obtained hourly from the Google Maps traffic layer within a 100-m radius of the Fath Square air-quality monitoring station in southwest Tehran. Data were collected for 30 days during Azar 1401 (November-December 2022), providing 720 hourly observations. Three predictive models, Artificial Neural Networks (ANN), Gated Recurrent Unit (GRU), and Extreme Gradient Boosting (XGBoost), were evaluated under five temporal scenarios representing the original hourly data and two- and three-hour averaging or accumulation of explanatory variables. Model performance was assessed using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). XGBoost provided the best overall performance. With the original hourly data, R^2 values of 0.65, 0.64, and 0.51 were obtained for CO, PM_{10}, and SO_2, respectively. Temporal aggregation did not consistently improve prediction accuracy, indicating that short-term variations in traffic congestion and wind speed contain useful predictive information. The findings demonstrate the potential of combining readily available online traffic information with meteorological observations and machine-learning techniques for short-term urban air-quality prediction and environmental decision support.

KEYWORDS

Air pollution prediction; Traffic congestion index; XGBoost; Machine learning; Wind speed.

* Corresponding Author: Email: rgolshan@iust.ac.ir

1. Introduction

Urban air pollution has intensified with population growth, urbanization, and increasing transportation activity. Traffic congestion can aggravate this problem because vehicles operating at low speeds with frequent stops, acceleration, and deceleration may generate greater emissions and fuel consumption. Previous research has demonstrated substantial public-health impacts associated with traffic congestion [1]. Meteorological conditions are also important because wind affects pollutant dispersion. Studies in Tehran have shown that meteorological variables can contribute to prediction of air-pollutant concentrations using data-driven methods [2].

Machine-learning techniques have increasingly been applied to air-quality forecasting. Shakya et al. [3] used meteorological, vehicular, and emission data in a deep-learning framework to predict $PM_{2.5}$ concentrations. XGBoost has also shown promising performance in air-quality prediction; Wang et al. [4] developed a spatiotemporal XGBoost model for PM2.5, while Lei et al. [5] compared ANN, XGBoost, and other machine-learning methods for hourly air-quality forecasting.

The present study investigates whether a traffic congestion index derived from Google Maps, combined with wind speed, can effectively predict CO, PM_{10}, and SO_2 concentrations. The main contribution is the integration of readily available online traffic-congestion information with local wind conditions for short-term pollutant prediction at the monitoring-station level. In addition, five temporal representations are compared to determine whether temporal aggregation improves prediction relative to the original hourly observations.

2. Methodology

The study focused on the Fath Square air-quality monitoring station in southwest Tehran. Traffic conditions within a 100-m radius were obtained from the Google Maps traffic layer. Four traffic states were coded numerically: green = 1 (free flow), orange = 2 (moderate congestion), red = 3 (heavy congestion), and dark red = 4 (very heavy congestion). The lengths of road segments corresponding to each state were measured and used to calculate a traffic congestion index. Previous research has supported the applicability of online traffic information from Google Maps for traffic analysis [6].

Data were collected hourly for 30 days during Azar 1401 (November-December 2022), yielding 720 observations. CO, PM_{10}, and SO_2 were the target pollutants, while traffic congestion and wind speed were explanatory variables. Five scenarios were examined: (1) original hourly data; (2) two-hour averages; (3) two-hour cumulative traffic congestion; (4) three-hour averages; and (5) three-hour cumulative traffic congestion. ANN, GRU, and XGBoost were evaluated using R^2, MAE, and RMSE.

The temporal scenarios were designed to test whether the immediate traffic condition or a short accumulation of traffic effects was more informative for pollutant prediction. In the averaging scenarios, the explanatory variables were represented by their two- or three-hour averages, whereas in the cumulative scenarios the congestion index was accumulated over the corresponding period. Pollutant observations remained hourly so that the predictive target and the total number of observations were preserved. The same performance criteria were then used for all model-scenario combinations, enabling a direct comparison of the effect of temporal treatment and modeling technique.

3. Discussion and Results

XGBoost consistently provided better overall predictive performance than ANN and GRU. For the original hourly scenario, XGBoost achieved R^2 values of 0.65, 0.64, and 0.51 for CO, PM_{10}, and SO_2, respectively. The corresponding R^2 values were 0.21, 0.16, and 0.12 for ANN and 0.16, 0.25, and 0.27 for GRU. Table 1 summarizes the XGBoost results across the five temporal scenarios.

Table 1. R^2 values of XGBoost under different temporal scenarios

Scenario	CO	PM_{10}	SO_2
Hourly	0.65	0.64	0.51
2-h avg.	0.60	0.58	0.52
2-h cum.	0.62	0.58	0.52
3-h avg.	0.55	0.61	0.52
3-h cum.	0.58	0.61	0.47

As shown in Table 1, temporal aggregation did not consistently improve performance. The hourly scenario produced the highest R^2 for CO and PM_{10}, while the maximum R^2 for SO_2 increased only marginally from 0.51 to 0.52 in Scenarios 2-4. For the hourly scenario, XGBoost produced MAE values of 0.78, 15.15, and 3.45 and RMSE values of 1.11, 21.09, and 4.87 for CO, PM_{10}, and SO_2, respectively. The larger absolute errors for PM_{10} can partly be attributed to its wider concentration range.

Statistical analysis showed that wind speed and traffic congestion were significant variables, with P-values of 0.000108 and 0.000009, respectively. The strong performance of XGBoost is consistent with previous applications to air-quality prediction [4,5]. Deep-learning time-series forecasting has also shown useful performance for air quality [7], while the XGBoost framework provides a scalable basis for nonlinear boosted-tree prediction [8]. Nevertheless, differences among pollutants indicate that traffic congestion and wind speed alone cannot fully explain pollutant variations, particularly for SO_2.

A closer comparison of the five temporal scenarios confirms that the effect of aggregation depends on the pollutant. For CO, the hourly XGBoost model achieved $R^2 = 0.65$, whereas the two-hour average, two-hour cumulative, three-hour average, and three-hour cumulative scenarios produced R^2 values of 0.60, 0.62, 0.55, and 0.58, respectively. For PM_{10}, the corresponding values were 0.64, 0.58, 0.58, 0.61, and 0.61. Thus, none of the aggregated scenarios exceeded the hourly model for these two pollutants. For SO_2, aggregation produced only a very small improvement: R^2 increased from 0.51 for hourly data to 0.52 in the two-hour average, two-hour cumulative, and three-hour average scenarios, while it decreased to 0.47 for the three-hour cumulative scenario. Considering all pollutants together, the hourly representation therefore provided the most consistent performance.

The error indicators lead to the same general conclusion. In the hourly scenario, XGBoost yielded RMSE values of 1.11 for CO, 21.09 for PM_{10}, and 4.87 for SO_2, with MAE values of 0.78, 15.15, and 3.45, respectively. PM_{10} had the largest absolute errors, which should be interpreted in relation to its substantially wider observed concentration range. The comparatively lower predictability of SO_2 also suggests that its short-term concentration may be affected by sources and atmospheric processes that are not adequately represented by local traffic congestion and wind speed alone. Consequently, the proposed model should be viewed as a parsimonious traffic-and-weather-based predictor rather than a complete representation of all mechanisms controlling urban pollutant concentrations.

The comparison among algorithms is also noteworthy. ANN and GRU did not benefit sufficiently from the available one-month dataset to match the performance of XGBoost. In contrast, the tree-boosting structure of XGBoost was able to represent nonlinear interactions between the explanatory variables without requiring a long temporal sequence. This result is compatible with previous air-quality applications of XGBoost [4,5]. From an operational perspective, the finding is important because the required traffic input can be obtained from an online traffic service rather than from an extensive fixed traffic-counting network. When combined with local wind observations, this provides a relatively simple framework for near-term estimation of pollutant levels around an urban monitoring station.

The scenario-level results also provide insight into the temporal behavior of the input variables. The decrease in CO performance from $R^2 = 0.65$ for hourly data to 0.55 for the three-hour average suggests that smoothing the traffic and wind series removes short-lived variations that are relevant to local CO concentrations. PM_{10} shows a similar, although less pronounced, pattern: the hourly value of 0.64 remains higher than the 0.58-0.61 range obtained after aggregation. SO_2 behaves differently, with a marginal increase from 0.51 to 0.52 in several aggregated scenarios. Because this improvement is very small and is accompanied by deterioration in the three-hour cumulative case, there is no consistent evidence that accumulation improves the overall forecasting framework. The results therefore support retaining the finest available hourly resolution for short-term operational prediction.

The significance tests reinforce the role of the two explanatory variables but should be interpreted together with the predictive results. The reported P-values indicate statistically meaningful relationships between pollutant concentrations and both wind speed and the congestion index within the study dataset. Wind can alter the local dilution and transport of pollutants, whereas the congestion index provides an indirect representation of the intensity and operating condition of nearby traffic. However, statistical significance does not imply that these variables explain all changes in pollutant concentration. The moderate R^2 values, especially for SO_2, indicate that other factors such as background pollution, non-traffic emission sources, atmospheric stability, temperature, humidity, and conditions outside the 100-m study radius may also contribute to the observed concentrations.

The relative performance of the algorithms should also be considered in relation to the size and structure of the dataset. The analysis was based on 720 hourly observations collected during one month. Sequence-based models such as GRU can benefit from substantially longer time series when temporal dependencies are strong and repeated seasonal patterns are available. In contrast, XGBoost can efficiently partition a limited tabular dataset and represent nonlinear interactions without requiring a long training sequence. This characteristic likely contributed to its stronger performance in the present application. The finding is compatible with previous air-quality studies reporting useful XGBoost performance [4,5], while deep-learning studies have shown that recurrent approaches become particularly effective when larger multivariate time series are available [3,7]. Thus, the present comparison should be understood as evidence for this dataset and application rather than as a universal ranking of the algorithms.

From a practical perspective, the framework is attractive because traffic input can be obtained from an online mapping service rather than from a dense network of fixed detectors. This can reduce data-collection requirements and support short-term assessment around monitoring stations where conventional traffic counts are unavailable. Nevertheless, the model was developed for one station and one month; spatial and seasonal transferability has therefore not been established. Future work should test multiple stations and seasons and incorporate additional meteorological and traffic variables to evaluate whether the method can support reliable real-time warning or traffic-management strategies.

4. Conclusions

This study investigated short-term urban air-pollution prediction using a Google Maps-derived traffic congestion index and wind speed. XGBoost achieved the best overall performance among ANN, GRU, and XGBoost. Using the original hourly observations, R^2 values were 0.65 for CO, 0.64 for PM_{10}, and 0.51 for SO_2. Temporal aggregation generally did not improve prediction accuracy, suggesting that short-term variations in traffic and wind contain useful information. Both explanatory variables were statistically significant. The proposed approach demonstrates the potential of integrating readily available online traffic information with meteorological observations for short-term urban air-quality prediction. Future studies should examine longer periods, additional monitoring stations, and further traffic and meteorological variables.

5. References

[1] Levy, J.I., Buonocore, J.J., and Von Stackelberg, K., 2010. "Evaluation of the public health impacts of traffic congestion: a health risk assessment". Environmental Health, 9, pp. 1–12.

[2] Farhadi, R., Hadavifar, M., Moeen al-Dini, M., and Amin Toosi, M., 2019. "Prediction of air pollutant concentrations in Tehran based on meteorological variables using artificial neural networks and linear regression in hot and cold seasons". Journal of Natural Environment, 73(1), pp. 115–127. (in Persian)

[3] Shakya, D., Deshpande, V., Goyal, M.K., and Agarwal, M., 2023. "PM2.5 air pollution prediction through deep learning using meteorological, vehicular, and emission data: A case study of New Delhi, India". Journal of Cleaner Production, 427, pp. 139278.

[4] Wang, Z., Wu, X., and Wu, Y., 2023. "A spatiotemporal XGBoost model for PM2.5 concentration prediction and its application in Shanghai". Heliyon, 9(12), pp. e22569.

[5] Lei, T.M., Ng, S.C., and Siu, S.W., 2023. "Application of ANN, XGBoost, and other ML methods to forecast air quality in Macau". Sustainability, 15(6), pp. 5341.

[6] Golshan Khavas, R., Shariat Mohaymany, A., and Rahimi, Z., 2022. "Evaluation of travel time data from Google Maps and Neshan services". In 13th National Congress on Civil Engineering, Isfahan, Iran. (in Persian)

[7] Freeman, B.S., Taylor, G., Gharabaghi, B., and Thé, J., 2018. "Forecasting air quality time series using deep learning". Journal of the Air & Waste Management Association, 68(8), pp. 866–886.

[8] Chen, T., and Guestrin, C., 2016. "XGBoost: A scalable tree boosting system". In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.