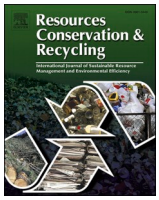
Resources, Conservation & Recycling 184 (2022) 106380

Contents lists available at ScienceDirect

Resources, Conservation & Recycling

journal homepage: www.elsevier.com/locate/resconrec



Forecasting ground-level ozone concentration levels using machine learning

Jianbang Du [a,*], Fengxiang Qiao [b], Pan Lu [a,c], Lei Yu [b]

[a] Upper Great Plains Transportation Institute, North Dakota State University, Fargo, North Dakota, 58105, United States
[b] Innovative Transportation Research Institute, Texas Southern University, Houston, Texas, 77004, United States
[c] Department of Transportation, Logistics, and Finance, North Dakota State University, Fargo, North Dakota, 58105, United States

ARTICLE INFO

Keywords:
Air quality analysis
Ground-level ozone
Forecasting
Machine learning
Frequent pattern mining
Meteorological Measurements

ABSTRACT

Ground-level ozone (GLO) has been widely recognized as a critical air pollutant that has the potential to induce various adverse environmental and health effects. To eliminate its hazardous impacts, the development of an accurate and effective approach to forecast the upcoming pollution concentration levels is in urgent need. Recent studies show that machine learning algorithms have excellent abilities in ground-level ozone concentration forecasting, however many of their forecasting models do not consider the contributing effects on meteorological measurements and traffic situations which were also identified as potential influencing factors to ground-level ozone concentration by some previous research. In the meantime, most of the existing models target short-term forecasting with rough temporal resolutions, such as daily and weekly scales. This paper aims to propose a methodology that can provide long-term GLO concentration forecasting with a high temporal resolution. To achieve this, a frequent pattern mining approach is utilized to analyze the local intercorrelation between GLO concentration and contributing factors such as meteorological parameters and transportation situations. Then, a series of machine learning algorithms were identified to forecast the ground-level ozone concentration levels using traffic and meteorological measurements data. A case study was conducted in the Houston region with 10 years of historical measurements, each of the historical ground-level ozone concentration records is associated with a series of meteorology and traffic situation parameters. Data from 2010 to 2019 was used to select and train the machine learning models, and data from 2020 was used to perform the final validation and evaluation. Results show that the extreme gradient boost (XGBoost) machine learning algorithm provides the most accurate prediction of the hourly ground-level ozone concentration on a yearly scale, which shows the annually forecasting ability and the robustness of the model built. The proposed approach could be applied to other similar regions and other critical air pollutants that are also influenced by transportation and meteorological factors.

1. Introduction

With the rapid growth of the global population and technologies, the consumption of fossil fuels and petrochemicals is raising drastically (Xiao et al., 2018), which is one of the main contributors to air pollution that can induce a series of life-shortening respiratory and lung diseases (Zumla et al., 2015). Furthermore, air pollution plays an important role in adversely affecting ecosystems and flora and fauna species, and triggering other consequences such as greenhouse effects, acid rain, earth fallout, and other environmental damages.

Ground-level ozone (GLO), also known as tropospheric ozone, is one of the most notorious air pollutions identified as a criteria pollutant by the United States Environmental Protection Agency (U.S. EPA). Unlike the primary air pollutants, ground-level ozone as a secondary air pollutant is formed by a series of complex reactions of its precursor chemicals, including NO_x and volatile organic compounds (VOC) in the presence of sunlight (Du et al., 2018; Shao et al., 2009). Only a small amount of ground-level ozone is naturally sourced. While the majority of the ground-level ozone precursor chemicals are sourced from vehicle exhausts, power plant emitters, and oil refineries (Cardelino and Chameides, 1995). Thus, urban areas are more susceptible to ground-level ozone hazards (Bell et al., 2004). And the forecasting of ground-level ozone concentration is extremely beneficial to the public health (Bell et al., 2006). For that purpose, various approaches have been utilized to

* Corresponding author.
E-mail address: jianbang.du@ndsu.edu (J. Du).

https://doi.org/10.1016/j.resconrec.2022.106380
Received 4 February 2022; Received in revised form 21 April 2022; Accepted 25 April 2022
Available online 11 May 2022

forecast and estimate ground-level ozone concentration levels. One of the most widely used algorithms is the neural network (NN) model. The majority of the existing prediction models predict the value of ozone concentration in future time steps based on time series data observed and collected at various stations or locations. In other words, the predictions are historical spatio-temporal-based methods. However, due to the non-linear and complex nature of ground-level ozone concentration, forecasting became a tough task if the prediction model does not account for the GLO formation causes such as traffic and other meteorological conditions. GLO is not emitted directly into the air but is formed by chemical reactions between oxides of nitrogen (NOx) and volatile organic compounds (VOC) under certain meteorological conditions. GLO forms when pollutants are emitted by cars and other sources under the presence of certain meteorological conditions. EPA reported that transportation as a whole is responsible for over 55% of total NO_x emissions and around 10% of VOCs emissions in the U.S. (EPA, 2021). Meteorological conditions could largely influence the ground-level ozone formation (Ding et al., 2004; Gao et al., 2005). For instance, the anticyclones are associated with the large circulation of winds around a high-pressure core that bring clear skies and cooler and drier air, which is highly related to the ground-level ozone concentration (Rodwell and Hoskins, 2001). The intense sunlight and low wind are also directly related to the formation and accumulation of ground-level ozone and its precursors (Ding et al., 2004).

When considering influencing factors in the forecasting models, the results can be improved compared to prediction performance with the models that only use historical spatial and temporal ground-level ozone concentration data. Wen et al. (2019) exhibited a spatiotemporal convolutional long short-term memory (C-LSTME) NN extended model to predict various air pollutions (Wen et al., 2019). The historical air pollution data were utilized in the *k*-nearest neighboring (*k*-NN) model as well as the meteorological data. It was claimed that the model was well performed for different time predictions at different region scales. Maleki et al. (2019) also utilized five meteorological parameters to forecast the ground-level ozone concentration based on the artificial NN algorithm but over various time ranges (Maleki et al., 2019). In the meantime, a case study performed by Elangasinghe et al. (2014) concluded that the forecasting of a ground-level ozone precursor concentration by meteorological measurements and other factors outperformed the linear regression model (Elangasinghe et al., 2014). However, prediction performance considering both traffic emission and meteorological conditions is still underresearched.

Based on different input data types, machine learning models have shown their abilities in air pollution forecasting for a predefined future period. When compared to other forecasting methods, the advantages of machine learning forecasting are significant: the data precession speed is accelerated, automatic forecasting updates, more data capacity, hidden pattern identification of the data, and increased adaptability (Taranenko, 2019). In the research conducted by Shaban et al., 2016, univariate and multivariate machine learning algorithms that include support vector machines (SVM) and M5P model trees were adopted to build one-step and multi-step ahead ground-level ozone concentration forecasting models (Shaban et al., 2016). Results showed that the M5P algorithm yielded more accurate predictions when using different features. A study conducted in China compared different machine learning classification algorithms for 74 cities (Xi et al., 2015), which revealed that the accuracy is positively related to the number of features selected to use, and the combined model is usually better than a unique model. Srivastava et al. (2018) conducted a case study in New Delhi, which implemented and evaluated various classification and regression models to forecast the ground-level ozone concentration (Srivastava et al., 2018). In his research, it is implied that the support vector regression and artificial NN outperformed other algorithms in forecasting.

However, given the various studies that utilized the machine learning algorithms to forecast the GLO concentration level for a future period, the majority of them tried to make only short-term forecasting with rough temporal resolutions to achieve a higher prediction accuracy (Ma et al., 2020). For example, Zhan et al. (2018) and Di et al. (2017) used the random frost and neural network models to make the daily 8-hours mean prediction of GLO (Di et al., 2017; Zhan et al., 2018). Wang et al. (2015) and (2016) made two weeks mean GLO concentration forecasting by the land-use regression (LUR) models (Wang et al., 2015, 2016). Beelen et al., (2009); Wolf et al., (2017), and De Hoogh et al. (2018) also utilized the LUP models to make the annual average forecasting (Beelen et al., 2009; De Hoogh et al., 2018; Wolf et al., 2017). This study aims to overcome the short-term and rough temporal resolution forecasting limitations and provide a methodology to build a robust model with the annual hourly prediction ability and comparable accuracy. To achieve this goal, firstly, a frequent data mining analysis will be performed to reveal the local influencing trends. Furthermore, seven machine learning algorithms will be built, compared, and analyzed based on their performance. After parameters tuning and selection, a best-performed model will be selected to perform the real-world forecasting to be finally evaluated.

2. Methodology

This paper involves using historical air pollution concentration levels, meteorological measurements, and traffic status data for temporal and frequent pattern mining analysis and machine learning forecasting. The flowchart of this research is shown in Fig. 1.

As Fig. 1 indicates, data collection and processing are information sources of this research, while all calculations and analyses are based on the data collected. Three parts of data should be collected: air pollution concentration data, meteorological measurements data, and traffic activity data. All three parts of the data are preprocessed and transformed into datasets with the same format for further analyses. The statistical and temporal analyses are based on a series of statistical tests such as the correlation test and temporal characteristics analysis. The datasets are then integrated into a uniform dataset where each air pollution record is associated with a series of meteorological and traffic attributes. The frequent pattern mining is performed on the binned integrated dataset by frequent pattern (FP)-growth and Apriori algorithms, which yield the frequent pattern rules with corresponding support values. The machine learning forecasting process is conducted on the split integrated dataset. The proper machine learning models are identified through a model selection process, while the employed models are trained for air pollution forecasting. The forecasting results are validated and evaluated with part of the real data from the dataset.

2.1. Frequent pattern mining

In this research, the object of pattern discovery is to find the inherent regularities in ground-level ozone concentration level records, in which the influencing factors that include solar, temperature, pressure, etc. are considered as *items*, and one or more items are considered as the *itemset* (*s*). A *k-itemset* is represented as $X = (x_1, x_2, \ldots, x_k)$, and the absolute *support* of *X* is the frequency of the occurrences of itemset *X*. The relative *support s* is the percentage of transactions that contain *X*, which is also the probability an air pollutant concentration level record contains *X*. An *itemset X* is frequent if the *support* of *X* is no less than a minimum support (*minsup*) threshold (σ).

The *support, confidence*, and interestingness measurement index LIFT (*l*) can be calculated using Eqs. (1)-3 (Lin et al., 2015).

$$s(C,\ D) = s(C \cup D) = \frac{n(C \cup D)}{n(T)} \tag{1}$$

$$c(C,\ D) = \frac{s(C \cup D)}{s(C)} \tag{2}$$

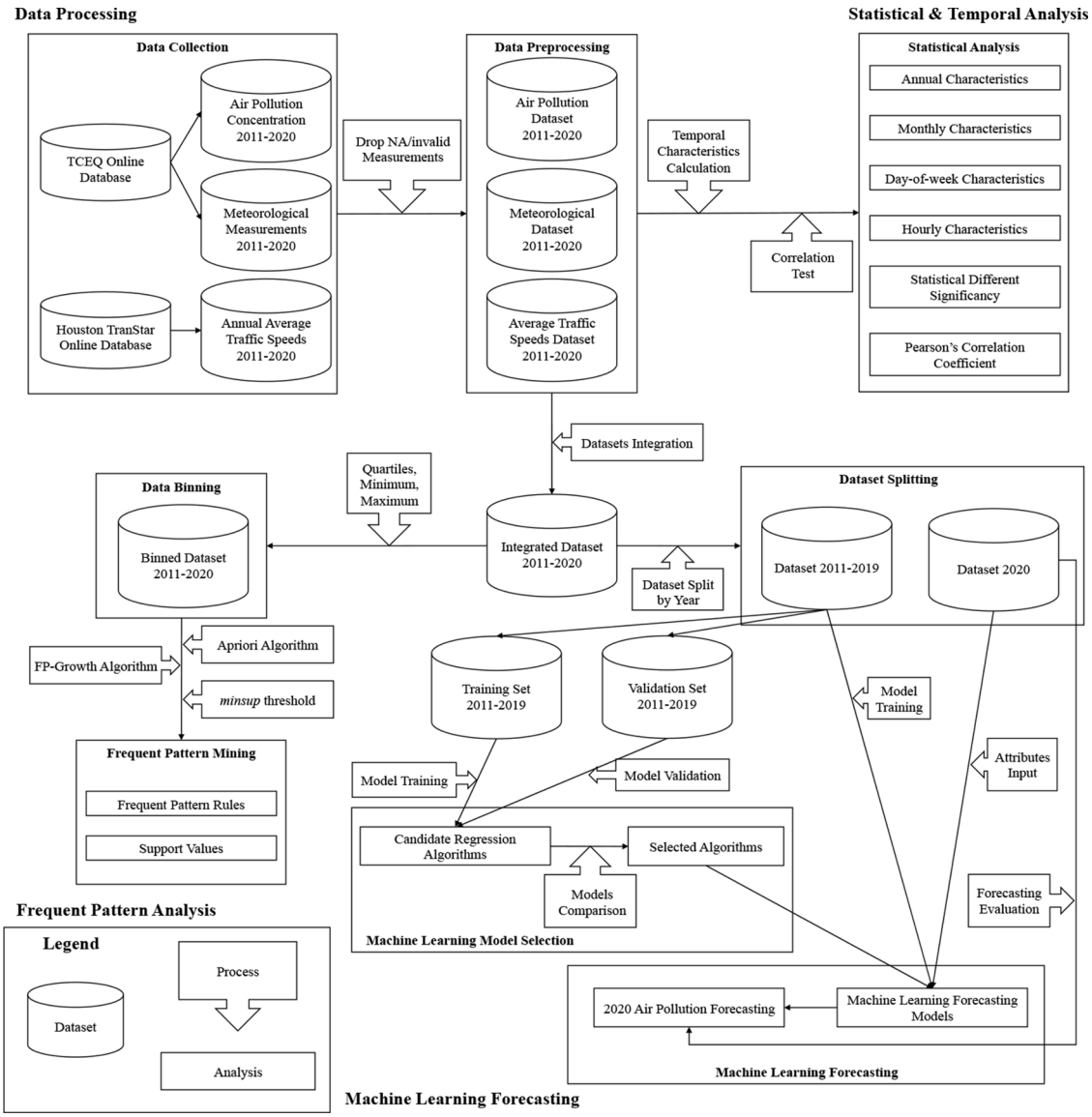


Fig. 1. Flow Chart of this Research.

$$l(C,\ D) = \frac{c(C \cup D)}{s(D)} = \frac{s(C \cup D)}{s(C) * s(D)} \tag{3}$$

where,

$s(C, D)$: the *support* for air pollutant concentration C and influencing factor measurement D occurring together, ranging (0, 1),
$n(C, D)$: the number of events when C and D occur together,
$n(T)$: the number of total events,
$c(C, D)$: the *confidence* for event D to occur when event C occurs, ranging (0, 1),
$l(C, D)$: the interestingness measurement LIFT (ranging [0, ∞]) for event D to occur when event C occurs, which tells how C and D are correlated.

In this research, The *support* $s(C, D)$ can provide the scale of an air pollutant concentration record occurring on a set of influencing items. *Confidence* $c(C, D)$ is the likelihood of an *item* occurring if another *item* happened. The LIFT illustrates the increase in an air pollutant concentration record when another item happened. Two typical and well-developed frequent pattern mining algorithms used in this research are the Apriori algorithm and the frequent pattern (FP)-growth algorithm (Aggarwal et al., 2014). The Apriori algorithm scans all possible *itemsets* and conducts all calculations while the FP-growth algorithm does not consider all possible *itemsets*. Based on the inherent mathematical theory of these algorithms, the mining results of the Apriori and FP-growth algorithms are the same with different calculation efficiencies depending on the *minsup* values (Xin et al., 2005). Thus, both algorithms are utilized in the analysis.

2.2. Machine learning forecasting

Seven most representative machine learning algorithms are considered to develop the forecasting models: polynomial regression (PR), multilayer perceptron (MLP), XGBoost (XGB), SVM, random forest (RF), linear regression (LR), and (k-NN) algorithms. There are basically two functions of machine learning models: classification and regression. Some algorithms are for both. Since the input data are continuous with float data types rather than discrete integer ones, regression algorithms are employed in this research. The historical data from 2011 to 2019 were utilized to train the selected models, while the real air pollution measurements in the year 2020 were used to evaluate the prediction accuracies. The datasets can be separated into two subsets: the historical dataset that contains the data from 2011 to 2019 and the target dataset

that contains the data in 2020. The historical dataset is used to train and select the model, while the target dataset is used for forecasting. In addition, the historical dataset is further divided into the training dataset validation datasets that both include attributes and classes. The candidate machine learning algorithms are trained by the training dataset, and the accuracy of the developed models was evaluated by the validation dataset. If the accuracy is not acceptable, then the parameters of the models shall be adjusted or the algorithm would be rejected. If the accuracy is accepted, then the target dataset will be fed to the forecasting model, and the prediction results will be provided. After the model is trained, the forecasting inputs are the corresponding hour's meteorological measurements and traffic situation, and the output is the ground-level ozone concentration.

Several evaluation matrices can be used to evaluate the model's prediction performance. For instance, accuracy, recall, and precision can be used for the classification models, while the mean squared error (MSE) and root mean squared error (RMSE) can be used for regression model evaluation. The objective of this research is to build a forecasting model that can be used for different regions and different air pollutants. To present the relative error percentage that can be used for inter-comparison for different air pollutants instead of the absolute error value, the normalized root means squared error (NRMSE) is commonly used on regression model evaluation as shown in Eq. (4), which can be compared between models and datasets with different scales (Nash and Sutcliffe, 1970; Ris et al., 1999; Willmott et al., 1985).

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{4}$$

The differences between the maximum value y_{max} and the minimum are the range for normalization with the outliers excluded. Moreover, the prediction vs. actual (PVA) plot, which is a scatter plot showing great data visualization (Piñeiro et al., 2008), is also utilized to characterize the regression prediction performance. As an ideal case, the model predictions should exactly meet the measured real lines. Thus, the line $y = x$ in the PVA plot is typically displayed along with the fitted trendline from the predicted points, while the angle between these two lines illustrates how the predictions deviated from the actual measurements. Furthermore, in this study, the prediction lines in the PVA plot were moved parallelly so they will cross the real measured lines at the origin point. This process will only increase the visual and comparison clarity without influencing the results as the angles between the two lines were not changed. In addition, to better compare the results from this research to previous models, the widely use coefficient of determination (R^2) scores were calculated to show how the regression model fits the real observed data.

3. Case study

The Houston-Galveston-Brazoria (HGB) area is a nonattainment area (NA) for ozone in the United States (Flynn, 2018). Thus, the case study was performed targeting the HGB area, which uses 10 years' (2011 - 2020) historical ground-level ozone concentration records and relevant meteorological measurements and transportation activity data. Based on the inherent relationships between the air pollution concentration and influencing factors, a pool of supervised machine learning models was employed, compared, and evaluated, while the best performing model was selected to predict the air pollution concentration level for year 2020 in terms of the model's prediction capability.

3.1. Data collection & preprocessing

The air pollution and meteorology data were collected from the Texas Commission on Environmental Quality (TCEQ, 2021), which is currently operating more than 200 air monitoring stations that are serving over 25 million population statewide in Texas, including industrial and large population regions. Other than ground-level ozone, the meteorological parameters are also collected, including solar radiation, temperature, pressure, precipitation, relative humidity, resultant wind speed, and resultant wind direction. Considering raw data issues in most monitoring sites, such as invalid/null data, discontinued measurements, inactivated sites, and unrepresentative site locations, the air monitoring site 403 (TCEQ site name: Clinton C403/C304/AH113) was selected to represent the Houston urban area. The site is located at 9525½ Clinton Drive near Interstate Highway IH-610 and the Houston ship channel.

The hourly ground-level ozone concentration is collected in parts per billion (ppb), which is the average reading for every five minutes. The solar radiation in Langleys per minute (Langleys/min) is measured by the total electromagnetic radiation emitted by the sun at the monitoring site. The temperature in Fahrenheit (°F) degrees is measured near the site. Precipitation in inches is the rainfall to the ground in liquid or solid form. The relative humidity is the percentage of the moisture in the air ranging from 0% (no humidity) to 100% (fully saturated air). The wind speed in miles per hour (mph) and direction is a single vector measured by converting the five-minute wind speeds and directions. The wind direction is measured to the nearest degree on a 360-degree compass, where 0° (360°) means the north, while 180° the south. The barometric pressure was recorded in millibars.

Traffic activity data from the years 2011 to 2020 were collected from the Houston Transportation Management Center (TMC) TranStar (2021). The data were recorded in 15-minute intervals from 5:00 to 19:00 during a full years' worth of weekdays. Traffic activities on IH-610 East Loop, which is the nearest to the monitoring site 403, were collected, including the annual average speeds along the southbound from Wayside Drive to Broadway Street (4.9 miles) and along the northbound from SH-225 to Gellhorn Drive (4.9 miles). Ten-year historical data were collected, while there were fifty-six 15-minute time interval records each day for each direction, and the annual on-road speed data were calculated from the average of the daily data. On the Houston TranStar database, it is showing that out of that time range, the on-road speeds are not impacted by the traffic volume. Thus, one of the assumptions of this study is that the on-road speeds out of this range are free-flow speeds. All collected raw data, especially for those from TCEQ, contain various invalid measurements to be preprocessed, such as NA (not enough data), AQI (data rejected), and LIM (lost data). Data containing invalid records are eliminated from the study. A total of 14,124 invalid records were removed at the data preprocessing stage, while the remaining 73,236 valid records were prepared for the following analysis.

The original data are in continuous data types. However, frequent pattern analysis is better performed on discrete data types such as inters and labels. In this research, the parameters other than precipitation are categorized and binned based on each of their minimum, maximum, and quartile values. Precipitation is classified into two bins due to most of its values are not discrete enough, which are bin zero for no precipitation and bin one for having precipitation. Table 1 shows the example of a binned record for one influencing factor record.

As shown in Table 1, each air pollutant concentration level record contains a series of binary information on all related factors. The full binning information is shown in Appendix A attached to this paper.

3.2. Temporal analysis

The annual, monthly, daily, and hourly temporal analysis for ground-level ozone concentration levels is performed for the entire 10-year period. The analysis results are shown in Fig. 2.

In Fig. 2, the yearly, monthly, and daily analysis results are shown in the box-whisker plots in Figs. 2 (a)-(c), respectively, where the x-axis shows the time scale while the y-axis shows the measurements of ground-level ozone concentration. The boxes show the interquartile ranges (IQR) with the upper and lower edges of the boxes showing the upper and lower quartiles (Q3 and Q1). The upper and lower short bars

Table 1
Example of a binned solar record.

Ground-level Ozone				Solar			
Ground-level Ozone_bin_1	Ground-level Ozone_bin_2	Ground-level Ozone_bin_3	Ground-level Ozone_bin_4	Solar_bin_1	Solar_bin_2	Solar_bin_3	Solar_bin_4
0	0	1	0	0	0	0	1

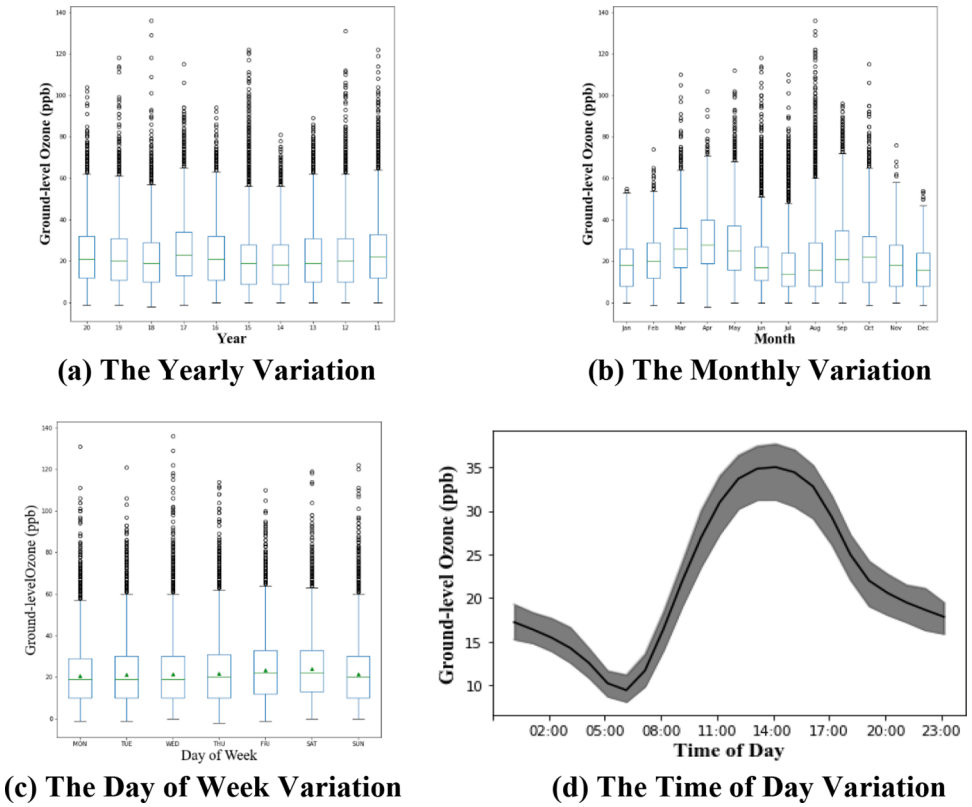


Fig. 2. Temporal Analysis of Ground-Level Ozone Concentration.

are the maximum (Q3+1.5* IQR) and minimum (Q1–1.5* IQR), while the short green bars in the boxes are the median, with the black spots being the outliers. For the time of day concentration plot in Fig. 2 (d), the gray areas show the variances of the average values of each category, and the black solid lines show the mean values of the variances. As shown in Fig. 2 (a) for yearly ground-level ozone concentration, there is no decreasing or increasing trend over the 10 years. For all years, the minimum concentrations were around 0 ppb, the maximums concentration was around 60 ppb, and the outliers were below 140 ppb. The year 2017 had the highest concentration range, while the years 2014, 2015, and 2018 had the lowest concentration ranges. In the monthly plot in Fig. 2 (b), the ground-level ozone concentration shows two significant peaks each year with the highest concentration at around 70 ppb for both April and September, and the highest concentration in July and December at around 50 ppb is the two lower valleys. In summary, one can find that seasonal patterns in ozone concentration are observed at the monitoring site: the ground-level ozone concentrations were higher in the spring and fall seasons and lower in the summer and winter seasons. From the day of week plot in Fig. 2 (c), no clear pattern is observed. The ground-level ozone concentration level is slightly higher on Fridays and Saturdays with the average concentration of 22 ppb compared with the other weekdays. From the time of day plot in Fig. 2 (d), the ground-level ozone concentration during a day increases from 7:00, which is the lowest point of a day, to its peak at around 14:00, then starts to decrease ranging from 18 to 30 ppb.

3.3. Correlation analysis

To determine the relationships among the parameters that are considered in this research, the inter-correlation values among all parameters are listed in Table 2.

Table 2 shows the matrix of Pearson *r* values between each pair of variables. The ground-level ozone had relatively weaker negative linear correlations with pressure, precipitation, wind direction, relatively strong negative correlations with relative humidity, and average traffic speed, together with relatively strong positive correlations with solar radiation, temperature, and resultant wind speed. The correlations among the influencing factors are listed below.

- The solar radiation had relatively weaker negative correlations with pressure and average traffic speed, and weaker positive correlations with precipitation and wind direction, a relatively strong negative correlation with relative humidity, and relatively strong negative correlations with temperature and resultant wind speed.
- The outdoor temperature had relatively weaker negative correlations with precipitation and average traffic speed, relatively strong negative correlations with pressure and relative humidity, and relatively stronger positive correlations with resultant wind speed and wind direction.
- Pressure had a relatively weaker negative correlation with precipitation, relatively strong negative correlations with relative humidity, resultant wind speed, and wind direction, and a relatively weak correlation with average traffic speed.

Table 2
Inter-correlations among parameters.

	Ozone	Solar	Temperature	Pressure	Precipitation	Humidity	Wind speed	Wind direction	Traffic situation
Ozone	1	0.497	0.248	-0.065	-0.009	-0.501	0.103	-0.038	-0.121
Solar		1	0.318	-0.007	0.012	-0.449	0.113	0.059	-0.067
Temperature			1	-0.539	-0.012	-0.145	0.215	0.148	-0.069
Pressure				1	0.020	-0.135	-0.160	-0.106	0.007
Precipitation					1	0.095	-0.023	-0.003	-0.013
Humidity						1	-0.106	-0.062	0.128
Wind speed							1	0.187	-0.078
Wind direction								1	0.023
Traffic situation									1

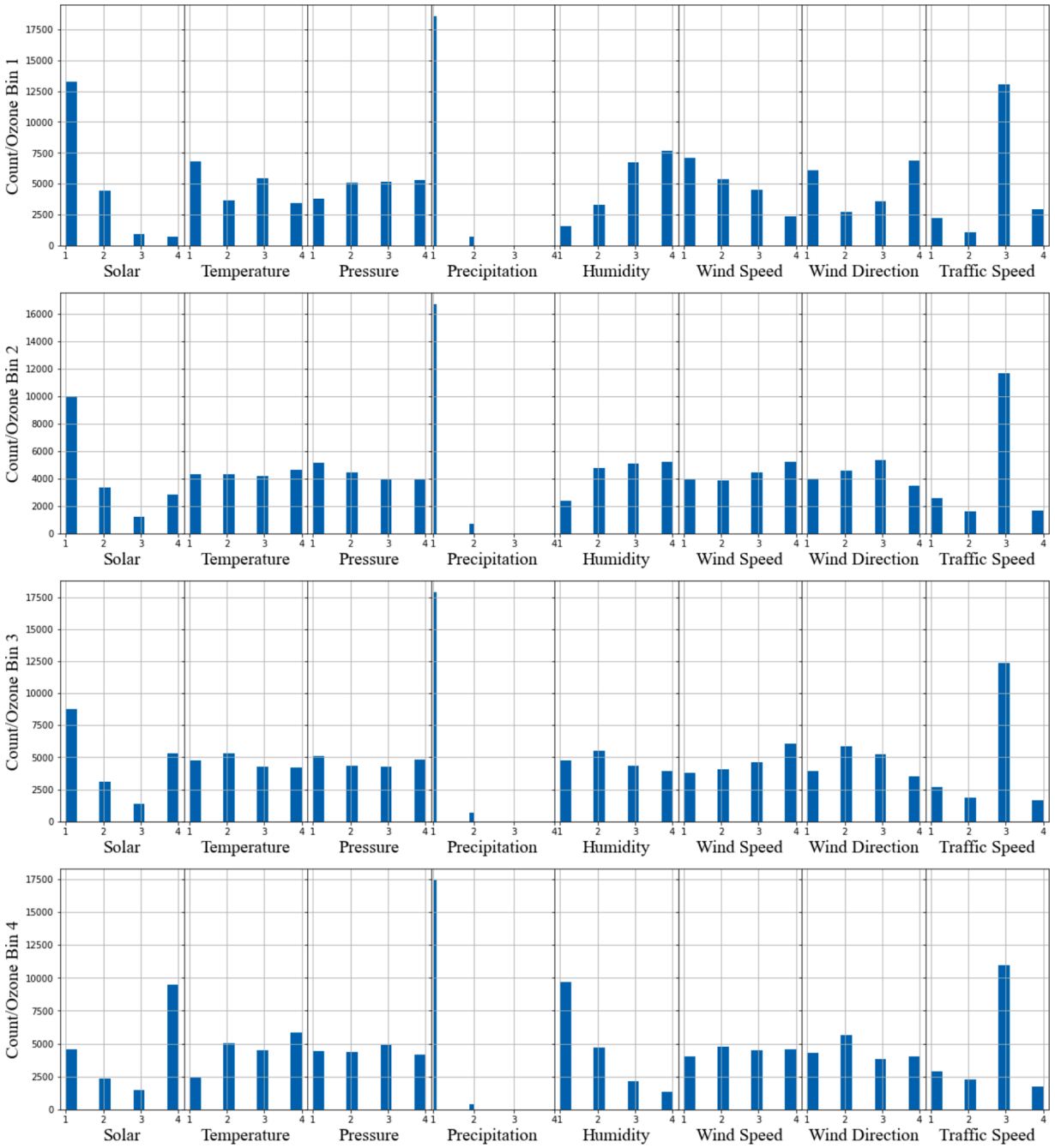


Fig. 3. Count Distributions of the Influencing Factors of Ground-Level Ozone.

- Precipitation had relatively weaker negative correlations with resultant wind speed, wind direction, and average traffic speed, and a relatively weak positive correlation with relative humidity.
- Relative humidity had a weaker negative correlation with wind direction, a relatively strong negative correlation with resultant wind speed, and a relatively strong positive correlation with average traffic speed.
- The resultant wind speed had a relatively weaker negative correlation with average traffic speed and a relatively strong positive correlation with wind direction.
- The wind direction had a relatively weaker position correlation with average traffic speed.

3.4. Frequent pattern mining

Fig. 3 shows the bin distribution of each influencing factor of ground-level ozone concentration. Some significant trends can be found by comparing each column of the figure vertically. For ground-level ozone concentration level categories from bin 1 to bin 4, the number of solar radiation decreased for bin 1 while decresed for bin 4; Counts of temperature in bin 1 decreased while the number of temperature in bin 4 increased; the number of humidity level in bin 1 increased and in bin 4 decreased significantly, and the resultant wind speed in bin 1 decreased and in bin 4 increased. These results are consistent with the previous correlation analysis.

Appendix B shows the frequent pattern mining rules with the highest support for ground-level ozone. Based on Appendix B, when the solar radiation level in [0, 0.01] Langleys/min, the temperature level in (-∞, 62.5] F, the pressure level in (-∞, 1013.1] millibars, the precipitation level in no precipitation, the relative humidity level in (83.8, 100] percentage, the resultant wind speed level in (0, 3.4] mph, the wind direction in [0, 70] degree compass, and the traffic speed level in (55, 60] mph, the ground-level ozone concentration level tends to be in (0, 11] ppb with the support of 0.265%.

When the solar radiation level in [0, 0.01] Langleys/min, the temperature level in (81.4, +∞) F, the pressure level in (1013.1, 1015.9) millibars, the precipitation level in no precipitation, the relative humidity level in (57.9, 73.8] percentage, the resultant wind speed level in (8.3, +∞) mph, the wind direction in (165, 222] degree compass, and the traffic speed level in (55, 60] mph, the ground-level ozone concentration level tends to be in (11, 20] ppb with the support of 0.24%.

When the solar radiation level in [0, 0.01] Langleys/min, the temperature level in (73.8, 81.4] F, the pressure level in (-∞, 1013.1] millibars, the precipitation level in no precipitation, the relative humidity level in (73.8, 83.8] percentage, the resultant wind speed level in (8.3, +∞) mph, the wind direction in (165, 222] degree compass, and the traffic speed level in (55, 60] mph, the ground-level ozone concentration level tends to be in (20, 31] ppb with the support of 0.213%.

When the solar radiation level in (0.414, +∞) Langleys/min, the temperature level in (81.4, +∞) F, the pressure level in (-∞, 1013.1] millibars, the precipitation level in no precipitation, the relative humidity level in (0, 57.9] percentage, the resultant wind speed level in (8.3, +∞) mph, the wind direction in (165, 222] degree compass, and the traffic speed level in (55, 60] mph, the ground-level ozone concentration level tends to be in (31, +∞) ppb with the support of 0.132%.

3.5. Machine learning forecasting

The machine learning forecasting analysis of this study includes two parts: the first is the model selection and parameters tuning, which involves the data from 2011 to 2019 by the 10-fold cross-validation and constructing the best performing model. The second part is to use the developed and selected model to perform real-world forecasting that serves to provide a forecast example and further evaluation with the data from the year 2020 which serves as a test set to truly evaluate model prediction power. Table 3 shows the NRMSEs, R^2 scores, density functions, and the PVA plots for all candidate algorithms, which provide the information that can be mutually compared that is previously introduced in Section 2.

As shown in Table 3, the second column shows the NRMSE values, the third column shows the R^2 scores, while the fourth and fifth columns demonstrate the distribution plots and PVA plots for all candidate algorithms. In the distribution plots, the model training set is shown in black lines, the model validation set is shown in red lines, and the model prediction set is shown in the blue line. Since the model training set was randomly selected as 80% of the entire database and the model validation set was the remaining 20%, the distribution of the validation should be close to that of the training set, which means the black line and the red line should be very close to each other.

As an interpretation, the blue line and the red line should be very close for good algorithms with accurate predictions. However, the distribution line plot only shows the differences in distribution among relevant sets, which cannot show the level of errors like the NRMSE. Thus, even though the gap between the blue line and the red line is relatively large, the NRMSE value could still be relatively small. The PVA plot compares the model prediction set with the actual measurements, in which the red line is a 45-degree line that shows the 100% accuracy match, while the blue line is the trend line of the prediction scatters. The more accurate the model prediction is, the smaller the prediction error angle would be between the red line and the blue line.

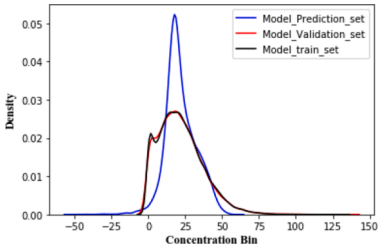
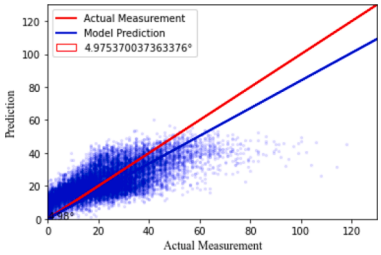
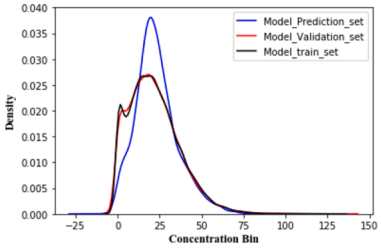
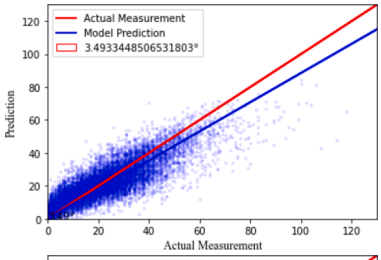
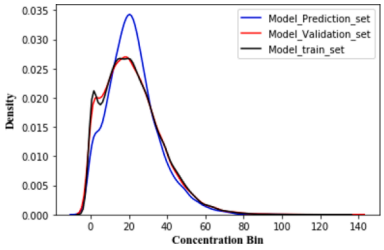
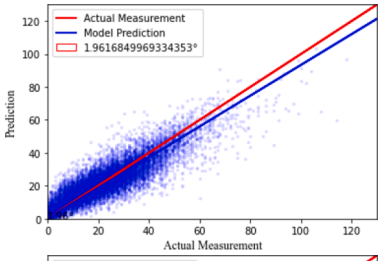
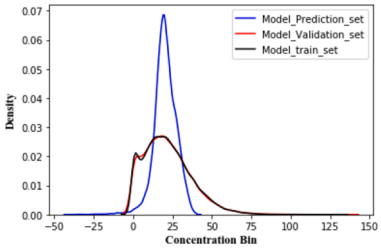
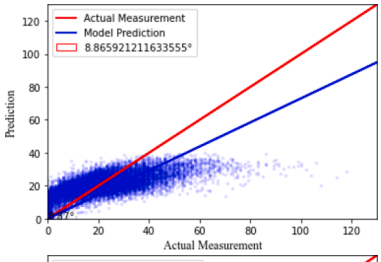
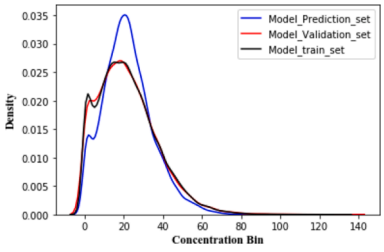
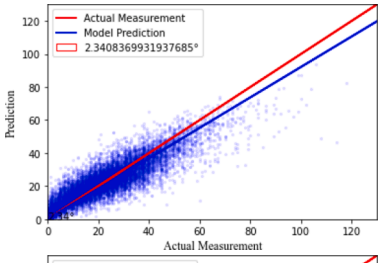
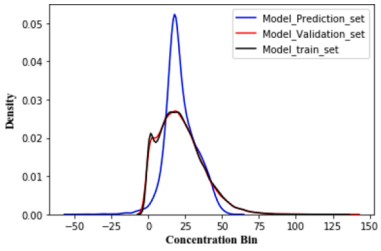
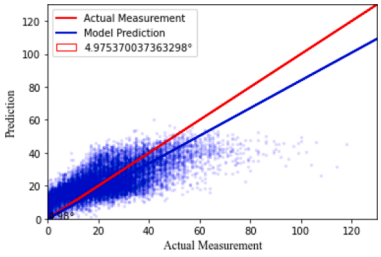
From the comparison, all candidate models yielded smaller NRMSE values with all being lower than 7%, and the prediction error angles were all smaller than 9° The blue trend lines in the PVA plots were all below the red 45-degree lines, which means the concentrations predicted tend to be lower than the real values. Among the models, the *k*-NN and SVM models showed relatively poorer fits for the dataset with the NRMSE value of 6.74% and 6.72%, and the prediction error angles of 3.44° and 8.87°, respectively. The PR, RF, and LR models were more accurate than the *k*-NN and SVM models; meantime, the XGB model prediction of ground-level ozone concentration had the lowest NRMSE of 5.07% and the highest R^2 score of 0.785.

The prediction error angle for the XGB model is 1.96° in the PVA plot, which means the prediction was very close to the real values, and this model will fit the prediction needs for ground-level ozone. Other models also have acceptable prediction error angles, especially the 2.34° angle for the RF model. However, that is still more underperforming than the XGB model. Thus, the XGB model will be chosen to perform the forecasting in the following step.

The data from the year 2020 are then used to make the prediction validation and verification. Based on the analysis and comparison from Table 4, the NRMSE value of the XGB model is 6.50% and the R^2 score is 0.776, which are poorer than those in the model selection process. However, the model selection process utilized the randomly selected validation dataset rather than all of 2020′s data in the forecasting. Thus, a higher NRMSE and lower R^2 score values could be expected. By the distribution plots, the counts of the prediction values, which range from around 0 to 20 ppb, are less than those of the actual measurements; and the counts of the prediction values, which range from around 20 to 40 ppb, are more than those of the actual measurements. From the PVA plots, the prediction values of both models tend to be lower than the real data. In addition, the comparison plots are shown in Fig. 4.

From the comparison plots shown in Fig. 4, two ground-level ozone concentration seasonal peaks can be predicted by the XGB model. The predictions tend to be smoother than the real data, which means the prediction values are with less variance than the real measurements. The concentration trend in terms of increasing and decreasing could be perfectly predicted by both models. For 0 to 700, 3000 to 4500, and after 5800 h, the predictions tended to be higher than the real values. For the other hours, the predictions were lower than the real values. Especially for the fall season concentration peak, the real values were significantly higher than the prediction. The error lines were smoothly around zero before the 3000 h point, which means the predictions for the first half-

Table 3
The NRMSE, distribution plots, and PVA plots for all candidate algorithms.

Algorithm	NRMSE	R2	Distribution Plot and PVA plot of Ground-level Ozone	
PR	5.68%	0.757		
MLP	5.32%	0.707		
XGB	5.07%	0.785		
SVM	6.72%	0.414		
RF	5.24%	0.781		
LR	6.58%	0.562		

(*continued on next page*)

Table 3 (*continued*)

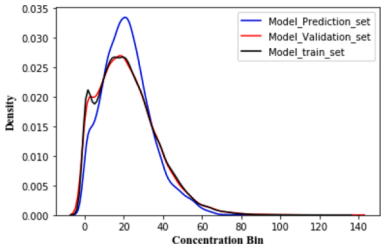
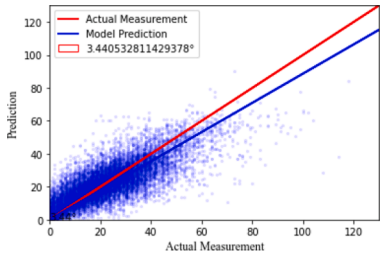
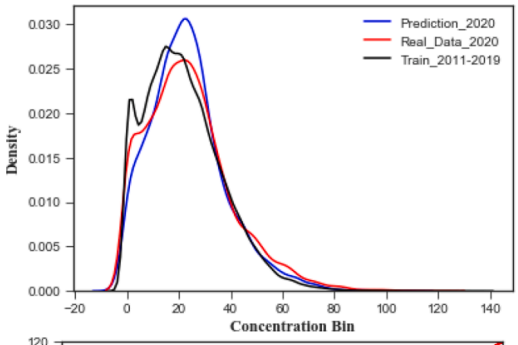
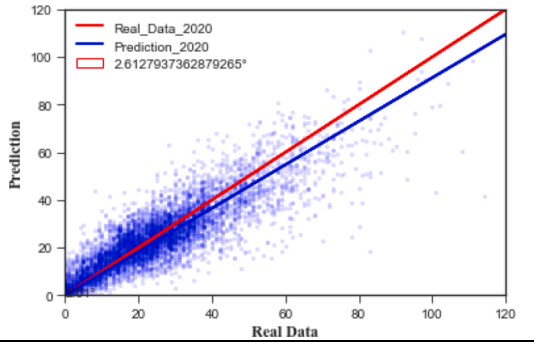
Algorithm	NRMSE	R2	Distribution Plot and PVA plot of Ground-level Ozone
KNN	6.74%	0.5	 

Table 4
The NRMES of forecasting ground-level ozone concentration by different models.

Model	XGB
NRMSE	6.50%
R2	0.776
Distribution Plot	
PVA Plot	

year were more accurate and stable. From this result, the annual hourly predictions of ground-level ozone based on meteorological measurements and average traffic speeds are made with higher accuracies.

3.6. Discussions

The results from the temporal analysis are consistent with previous studies. For instance, in the United States, Logan found a peak value of the ground-level ozone during spring and the minimum value in the summer (Logan, 1985). Lal et al. spotted a higher level of ground-level ozone during fall, which the reason was claimed as the higher transportation activities (Lal et al., 2000). In the meantime, the hourly ground-level ozone is found to be higher around the daytime (Pudasainee et al., 2006), which is similar to the results of this research. The correlation analysis between ground-level ozone concentration and the influencing factors showed some significant trends. The relationship between ground-level ozone and pressure and heavy rains that Emanuel found in the western Pacific area was not observed in this research due to the different study locations (Emanuel, 2003). However, the strong correlations between ground-level ozone level and solar radiation and humidity were proved by this research.

The proposed frequent pattern mining analysis provides similar results as the correlation analysis, which could be used in future studies directly as the intercorrelations between the influencing factors and ground-level ozone concentration levels vary by location. One of the advantages of this study's forecasting model is that it doesn't overly rely on the air pollution time series data input. It utilized the corresponding influencing factors' measurements to predict the pollution concentration. The forecasting of ground-level ozone When compared with the traditional statistical forecasting method (Du, 2018), the models proposed in this research are more accurate, practical, and efficient, especially for special events. Furthermore, the long-term forecasting ability of the models built in this research is more advanced when compared with the majority of the previous studies. For instance, Zhan and Di's models predicted the daily maximum 8-hours mean with R^2 scores of 0.69 and 0.76 (Di et al., 2017; Zhan et al., 2018); Wang's LUR model predicted the two weeks average concentration with R^2 scores of around 0.65 (Wang et al., 2015); Beelen and De Hoogh's LUR models predicted the annual average concentration with R^2 scores of 0.70 and 0.63 (Beelen et al., 2009; De Hoogh et al., 2018; Wolf et al., 2017). From the results of this research (R^2 = 0.776), the model built can make the annual forecasting ability of hourly data and yielded a comparable accuracy, which also shows the robustness of this model.

4. Conclusions

This research collected 10 years of ground-level ozone historical concentration data from 2011 to 2020 along with meteorological measurements and traffic situations and has proposed a practical data mining-based forecasting approach. Based on the literature review, the types, sources, potential hazards, and influencing factors of ground-level ozone were identified, which helped to identify the independent variables for ground-level ozone.

A case study was conducted in Houston as one of the ozone non-attainment areas. The temporal analysis was performed by the yearly, monthly, day of week, and time of day pollution concentrations. As shown in the intercorrelation and the Pearson *r* matrixes, the ground-level ozone is relatively more correlated with the relative humidity, average traffic speed, solar radiation, temperature, and resultant wind speed. A series of candidate machine learning algorithms were employed to forecast the ground-level ozone using nine years of data from 2011 to 2019, while the XGB algorithm outperformed other algorithms with the lowest NRMSE and error angles in the PVA plots and the highest R2 score. The year 2020 meteorological measurements and traffic situation data were used to forecast the ground-level ozone forecasting for the further validation of the XGB model. As a result, the NRMSE reached 6.50% with R^2 of 0.776 and the error angle was 2.61°, which is with higher credibility when compared with existing models from the previous studies. The forecasting accuracy parameters are calculated based on an annual hourly concentration prediction, which shows the robustness and long-term high temporal resolution forecasting ability of the model built.

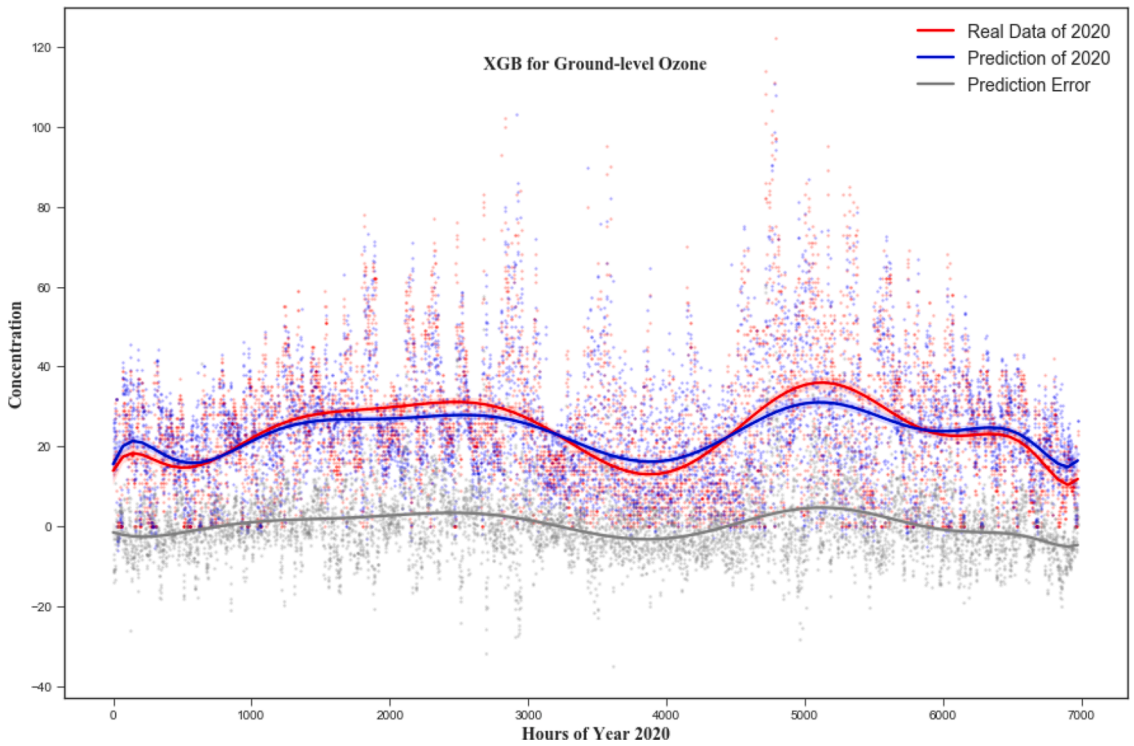


Fig. 4. The Comparison Plots of the XGB Model.

However, there are some limitations to this research. There is an assumption that the on-road average speeds out of the 5:00 to 19:00 time range are free-flow speeds. It is recommended to access more detailed traffic activity databases that might further improve the forecasting accuracy. Other socioeconomic factors such as economic activities could be considered during the modeling stage. The micro-scale forecasting accuracy could be even improved by identifying more machine learning and deep learning algorithms as the candidate modeling pool.

CRediT authorship contribution statement

Jianbang Du: Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Fengxiang Qiao:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Pan Lu:** Writing – original draft, Writing – review & editing. **Lei Yu:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Appendix A. Bins and labels for parameters

Measurements	Bins	Labels
Ground-level Ozone (ppb)	(0, 11]	1
	(11, 20]	2
	(20, 31]	3
	(31, +∞)	4
Solar (Langleys/min)	[0, 0.01]	1
	(0.01, 0.25]	2
	(0.25, 0.414]	3
	(0.414, +∞)	4
Temperature (F)	(-∞, 62.5]	1
	(62.5, 73.8]	2
	(73.8, 81.4]	3
	(81.4, +∞)	4
Pressure (millibars)	(-∞, 1013.1]	1

(*continued on next page*)

(*continued*)

Measurements	Bins	Labels
	(1013.1, 1015.9]	2
	(1015.9, 1019.6]	3
	(1019.6, +∞)	4
Precipitation (inches)	[0]	0
	(0, +∞]	1
Relative Humidity (%)	(0, 57.9]	1
	(57.9, 73.8]	2
	(73.8, 83.8]	3
	(83.8, 100]	4
Resultant Wind (mph)	(0, 3.4]	1
	(3.4, 5.6]	2
	(5.6, 8.3]	3
	(8.3, +∞)	4
Wind Direction (Degree Compass)	[0, 70]	1
	(70, 165]	2
	(165, 222]	3
	(222, 360)	4
Traffic Speed (mph)	(0, 50]	1
	(50, 55]	2
	(55, 60]	3
	(60, +∞)	4

Appendix B. Frequent pattern mining rules with the highest support for ground-level ozone

Parameters	Bins			
Ground-level Ozone Level	**1**	**2**	**3**	**4**
Solar	1	1	1	4
Temperature	1	4	3	4
Pressure	4	2	1	1
Precipitation	0	0	0	0
Humidity	4	2	3	1
Wind Speed	1	4	4	4
Wind Direction	1	3	3	3
Traffic Speed	3	3	3	3
Support	0.00265	0.0024	0.00213	0.00132

References

Aggarwal 2021. Houston TranStar: Yearly Speed Averages. http://traffic.houstontranstar.org/hist/histmain.aspx.

Aggarwal, C.C., Bhuiyan, M.A., Al Hasan, M., 2014. Frequent Pattern Mining Algorithms: A Survey, Frequent Pattern Mining. Springer, Cham, pp. 19–64.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., Briggs, D.J., 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Sci. Total Environ. 407 (6), 1852–1867.

Bell, M.L., McDermott, A., Zeger, S.L., Samet, J.M., Dominici, F., 2004. Ozone and short-term mortality in 95 US urban communities, 1987-2000. JAMA 292 (19), 2372–2378.

Bell, M.L., Peng, R.D., Dominici, F., 2006. The exposure–response curve for ozone and risk of mortality and the adequacy of current ozone regulations. Environ. Health Perspect. 114 (4), 532–536.

Cardelino, C., Chameides, W., 1995. An observation-based model for analyzing ozone precursor relationships in the urban atmosphere. J. Air Waste Manag. Assoc. 45 (3), 161–180.

De Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., 2018. Spatial PM2. 5, NO2, O3 and BC models for Western Europe–evaluation of spatiotemporal stability. Environ. Int. 120, 81–92.

Di, Q., Rowland, S., Koutrakis, P., Schwartz, J., 2017. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. J. Air Waste Manag. Assoc. 67 (1), 39–52.

Ding, A., Wang, T., Zhao, M., Wang, T., Li, Z., 2004. Simulation of sea-land breezes and a discussion of their implications on the transport of air pollution during a multi-day ozone episode in the Pearl River Delta of China. Atmos. Environ. 38 (39), 6737–6750.

Du, J., 2018. Temporal Characteristics of Particulate Matter 2.5 Concentration and Their Correlations with Weather Condition and Traffic Volume. Texas Southern University.

Du, J., Li, Q., Qiao, F., Yu, L., 2018. Estimation of vehicle emission on mainline freeway under isolated and integrated ramp metering strategies. Environ. Eng. Manag. J. 17 (5), 1237–1248.

Elangasinghe, M.A., Singhal, N., Dirks, K.N., Salmond, J.A., 2014. Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmos. Pollut. Res. 5 (4), 696–708.

Emanuel, K., 2003. Tropical cyclones. Annu. Rev. Earth Planet Sci. 31 (1), 75–104.

EPA, 2021. Smog, Soot, and Other Air Pollution from Transportation, in: EPA, U. (Ed.).

Flynn, J., 2018. An Investigation of Background Ozone and Particulate Matter Levels in the Houston/Galveston/Brazoria Metropolitan Area, in: TCEQ (Ed.).

Gao, J., Wang, T., Ding, A., Liu, C., 2005. Observational study of ozone and carbon monoxide at the summit of mount Tai (1534m asl) in central-eastern China. Atmos. Environ. 39 (26), 4779–4791.

Lal, S., Naja, M., Subbaraya, B., 2000. Seasonal variations in surface ozone and its precursors over an urban site in India. Atmos. Environ. 34 (17), 2713–2724.

Lin, L., Wang, Q., Sadek, A.W., 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Res. Part C: Emerging Technol. 55, 444–459.

Logan, J.A., 1985. Tropospheric ozone: seasonal behavior, trends, and anthropogenic influence. J. Geophysical Res.: Atmos. 90 (D6), 10463–10482.

Ma, R., Ban, J., Wang, Q., Li, T., 2020. Statistical spatial-temporal modeling of ambient ozone exposure for environmental epidemiology studies: a review. Sci. Total Environ. 701, 134463.

Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Birgani, Y.T., Rahmati, M., 2019. Air pollution prediction by using an artificial neural network model. Clean Technol. Environ. Policy 21 (6), 1341–1352.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. (Amst) 10 (3), 282–290.

Piñeiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M., 2008. How to evaluate models: observed vs. predicted or predicted vs. observed? Ecol. Modell. 216 (3–4), 316–322.

Pudasainee, D., Sapkota, B., Shrestha, M.L., Kaga, A., Kondo, A., Inoue, Y., 2006. Ground level ozone concentrations and its association with NOx and meteorological parameters in Kathmandu valley, Nepal. Atmos. Environ. 40 (40), 8081–8087.

Ris, R., Holthuijsen, L., Booij, N., 1999. A third-generation wave model for coastal regions: 2. Verification. J. Geophysical Res.: Oceans 104 (C4), 7667–7681.

Rodwell, M.J., Hoskins, B.J., 2001. Subtropical anticyclones and summer monsoons. J. Clim. 14 (15), 3192–3211.

Shaban, K.B., Kadri, A., Rezk, E., 2016. Urban air pollution monitoring system with forecasting models. IEEE Sens. J. 16 (8), 2598–2606.
Shao, M., Zhang, Y., Zeng, L., Tang, X., Zhang, J., Zhong, L., Wang, B., 2009. Ground-level ozone in the Pearl River Delta and the roles of VOC and NOx in its production. J. Environ. Manag. 90 (1), 512–518.
Srivastava, C., Singh, S., Singh, A.P., 2018. Estimation of air pollution in delhi using machine learning techniques, 2018 International Conference on Computing, Power and Communication Technologies (GUCON). pp. 304–309.
Taranenko, L., 2019. How to apply machine learning to demand forecasting. https://mobidev.biz/blog/machine-learning-methods-demand-forecasting-retail.
TCEQ, 2021. Texas commission on environmental quality. https://www.tceq.texas.gov/agency/air_main.html.
Wang, M., Keller, J.P., Adar, S.D., Kim, S.-Y., Larson, T.V., Olives, C., Sampson, P.D., Sheppard, L., Szpiro, A.A., Vedal, S., 2015. Development of long-term spatiotemporal models for ambient ozone in six metropolitan regions of the United States: the MESA Air study. Atmos. Environ. 123, 79–87.
Wang, M., Sampson, P.D., Hu, J., Kleeman, M., Keller, J.P., Olives, C., Szpiro, A.A., Vedal, S., Kaufman, J.D., 2016. Combining land-use regression and chemical transport modeling in a spatiotemporal geostatistical model for ozone and PM2. 5. Environ. Sci. Technol. 50 (10), 5111–5118.
Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., Chi, T., 2019. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Sci. Total Environ. 654, 1091–1099.
Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophysical Res.: Oceans 90 (C5), 8995–9005.
Wolf, K., Cyrys, J., Harcinícková, T., Gu, J., Kusch, T., Hampel, R., Schneider, A., Peters, A., 2017. Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. Sci. Total Environ. 579, 1531–1540.
Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y., Jin, D., 2015. A comprehensive evaluation of air pollution prediction improvement by a machine learning method, 2015 IEEE International Conference On Service Operations And Logistics, And Informatics (SOLI). pp. 176–181.
Xiao, Q., Chang, H.H., Geng, G., Liu, Y., 2018. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. Environ. Sci. Technol. 52 (22), 13260–13269.
Xin, D., Han, J., Yan, X., Cheng, H., 2005. Mining compressed frequent-pattern sets, Proceedings of the 31st International Conference on Very large Data Bases. pp. 709–720.
Zhan, Y., Luo, Y., Deng, X., Grieneisen, M.L., Zhang, M., Di, B., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollution 233, 464–473.
Zumla, A., George, A., Sharma, V., Herbert, R.H.N., Oxley, A., Oliver, M., 2015. The WHO 2014 global tuberculosis report—further to go. The Lancet Global Health 3 (1), e10–e12.