

國立臺灣大學公共衛生學院環境與職業健康科學研究所

碩士論文

Institute of Environmental and Occupational Health Sciences

College of Public Health

National Taiwan University

Master Thesis



應用機器學習方法預測 PM_{2.5}—以大台北地區為例
Application of the Machine Learning Method in PM_{2.5}
Prediction: A Case Study of Taipei Area

黃宇丞

Yu-Cheng Huang

指導教授：林靖愉 博士 郭育良 博士

Advisor: Ching-Yu Lin, Ph.D. Yue-Liang Leon Guo, Ph.D.

中華民國 110 年 8 月

August 2021

國立臺灣大學碩士學位論文
口試委員會審定書

應用機器學習方法預測 PM_{2.5}—以大台北地區為例

Application of the Machine Learning Method in PM_{2.5}

Prediction: A Case Study of Taipei Area

本論文係黃宇丞 君 (R07844012) 在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國 110 年 08 月 16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

鄧育良

(簽名)

(指導教授)

鄧育良

國立臺灣大學碩士學位論文
口試委員會審定書

應用機器學習方法預測 PM_{2.5}—以大台北地區為例
Application of the Machine Learning Method in PM_{2.5}
Prediction: A Case Study of Taipei Area

本論文係黃宇丞 君 (R07844012) 在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國 110 年 08 月 16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

林靖愉 (簽名)

(指導教授)

_____	_____
_____	_____
_____	_____
_____	_____

國立臺灣大學碩士學位論文
口試委員會審定書

應用機器學習方法預測 PM_{2.5}—以大台北地區為例
Application of the Machine Learning Method in PM_{2.5}
Prediction: A Case Study of Taipei Area

本論文係黃宇丞 君 (R07844012) 在國立臺灣大學環境
與職業健康科學研究所完成之碩士學位論文，於民國 110 年
08 月 16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

	(簽名)
(指導教授)	吳章甫
_____	_____
_____	_____
_____	_____
_____	_____

國立臺灣大學碩士學位論文
口試委員會審定書

應用機器學習方法預測 PM_{2.5}—以大台北地區為例

Application of the Machine Learning Method in PM_{2.5}

Prediction: A Case Study of Taipei Area

本論文係黃宇丞 君 (R07844012) 在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國 110 年 08 月 16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

(簽名)

(指導教授)

徐 謹 昇

誌謝



時光飛逝，碩士班兩年的時光也接近尾聲。回首來時路，不禁五味雜陳，一路上幫助我的人太多，此時此刻心中滿是感恩。

本論文能夠順利完成，首先要感謝我的兩位指導教授-郭育良老師及林靖瑜老師，兩位老師對我總是包容，謝謝郭育良老師兩年來的悉心指導，無論是學術上的提攜，論文方向上的建議，老師總是毫無保留的分享。在我的論文方向決定跨領域時，老師也積極地幫我尋找資源，更從開學就讓我參加許多大大小小的會議，以及感謝老師總是放手讓我去修課，讓我可以去探索不同的領域，真的很慶幸能在碩班遇到您作為我的指導老師。也感謝作為共指的林靖瑜老師，老師一直都很溫柔的對待每位學生，口試前也願意花時間聽我演練，也給我許多實用的建議，在此向兩位老師致上我最高的敬意與最誠摯的感謝。

另外，口試期間感謝兩位指導教授以及口試委員 中央研究院資訊科學研究所研究員徐讚昇教授，台灣大學資訊工程學系暨研究所賴飛羆教授，台灣大學環境與職業健康科學研究所吳章甫所長口試期間不吝細心指正，在論文上給予我許多寶貴的建議，讓我的研究更佳完備，在此致上最深的謝意。

碩士班修業期間，也感謝研究室的士群學長及其他學長姐們，在論文遇到問題時願意撥空與我討論並給我建議，並願意提供我相關資料，也謝謝同屆的姮如、宛瑜，謝謝你們在修課及論文上都給了我許多幫助及資訊，因為有你們，讓我在研究生涯中不孤單、且更加順利。

最後謝謝親愛的家人的體諒及支援，讓我在家中可以安心的撰寫論文，謝謝爸爸和媽媽學習上一路以來的支持，讓我在這條路上無後顧之憂，謝謝你們以我為榮，我愛你們！

最後，謹此向所有關心我、幫助過我的人致上謝意，沒有你們就沒有這份論文的產生，謝謝！



中文摘要

背景與目的

PM_{2.5} 細懸浮微粒係指氣動粒徑小於 2.5 微米的粒子，依據不同的成分組成及附著物具有不同的毒性，其孔徑大小足以穿透肺泡至人體血液中，尤其對於長者、幼兒、具有心肺功能疾病者，不論長短期，暴露於細懸浮微粒都具有對危害健康的潛在風險。環保署於 1993 年完成全國空氣品質監測站網的設置，以達到監督空氣品質保障人民健康之目的，而近年來隨著空氣品質受到民眾重視，也越來越多研究嘗試對空氣汙染進行預測。本研究旨在利用不同機器學習模型比較空氣品質預測效力。

材料與方法

本研究針對環保署設立於台北地區以台灣新北市及台北市為主的一般空氣品質測站，蒐集 2018 年至 2019 年間包含 PM_{2.5} 細懸浮微粒等空氣污染物以及其相關的氣象資料，以過去 8 小時之歷史資料推估未來三小時後之 PM_{2.5} 細懸浮微粒濃度，使用 2018 年對模型進行訓練，並用 2019 年的資料進行驗證，以評估模型的效果。模型部分使用傳統線性回歸統計方法作為基準，比較機器學習與深度學習模型對於 PM_{2.5} 細懸浮微粒濃度預測的效力。研究中探討單一模型對於不同測站間、不同模型間的預測效果比較，並考量加入鄰近測站的影響，評估其對不同模型的預測效果提升是否有幫助。

結果

本研究共蒐集納入兩年間 13 測站共 227760 筆逐時資料，23 個變數，各測站的 PM_{2.5} 濃度平均為 15.23 毫克 (標準差為 10.15 毫克)，使用三種模型進行預測，發現以 XGBoost 預測模型效力最高，其次是 LSTM，兩者平均都高於線性回歸模型。在測站方

面，以土城及菜寮站在 R-squared 上表現最好，士林及萬華站表現最差，而加入鄰近測站變項後，比較無納入變項的土城站、士林站及萬華站的預測效果都有所提升，最終的模型對於 2019 年整年的預測達到了 64% 的預測力。



結論

本研究顯示在提前三小時的預測力上 XGBoost 預測模型相較於神經網路及線性回歸具有較佳的預測效果，加入鄰近測站也能提高模型的準確率。

關鍵字：空氣汙染預測、細懸浮微粒、機器學習

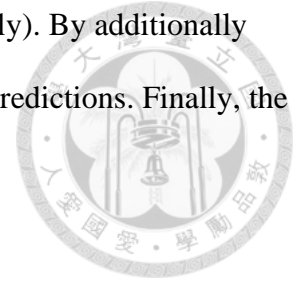
Abstract

Backgrounds: Increasing air pollution has become a grave concern, with researchers finding adverse health effects caused by air pollutants. Among all air pollutants, fine particulate matter (PM_{2.5}) whose aerodynamic diameter is less than 2.5 μ m is of particular concern. Especially for sensitive people, short-term as well as long-term exposure to PM_{2.5} might cause serious hazards. Although the Taiwan Environment Protection Administration has built an air quality monitoring network to monitor the PM_{2.5} concentrations and the government has revised the standards related to pollutants, an accurate and prompt early warning system is urgently needed.

Methods: In this study, we conducted a comprehensive evaluation of several models to predict PM_{2.5} concentrations in the Taipei area. We collected the data of Taipei City and New Taipei City from 2018 to 2019 from the Environmental Protection Administration open data platform, and we applied three kinds of models, i.e., linear regression, machine learning, and deep learning after a series of data preprocessing steps. Depending on the various requirements of models, the dataset can be classified as time-series-oriented and feature-oriented to fit the model. Model performance among stations and various models are compared in our research. We also compared using geographical predictors using nearby stations to see whether they would improve the predictions. The performance of prediction was evaluated using Root Mean Square Error, Mean Absolute Error and R-squared.

Results: In this study, 227760 hourly data from 13 stations were collected, and 23 variables were adopted to train the model. Among all stations, the XGBoost model outperformed the LSTM model followed by the linear regression model. Tucheng and Cailiao station in all the three

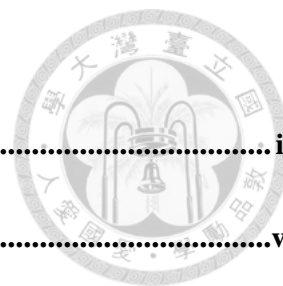
models achieved the best R-squared on average (0.6043, 0.6042 respectively). By additionally considering the influence of nearby stations, most models improved their predictions. Finally, the best models' prediction reached an R-squared value of 0.64.



Conclusion: This study found that the prediction using the 2018-year data in a single station in the Taipei Area can have a performance of 0.64 by using the XGBoost model, which outperformed the LSTM model followed by the linear regression model. Additional features from nearby stations for training are also beneficial to the predictions.

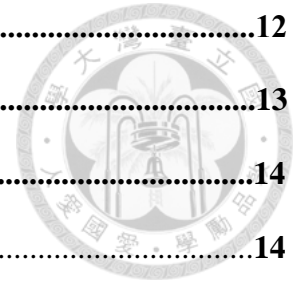
Keywords: Air pollution prediction, Forecasting, PM_{2.5}, Machine learning, Deep learning

Contents



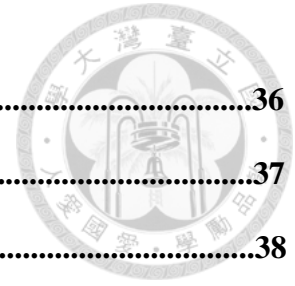
口試委員會審定書.....	i
誌謝.....	v
中文摘要.....	vi
Abstract.....	viii
Contents.....	x
List of Tables.....	xii
List of Figures.....	xiii
Chapter 1 Introduction.....	1
Chapter 2 Materials and methods.....	7
2.1 Data collection.....	7
2.1.1 Database.....	7
2.1.2 Study Area.....	8
2.2 Data Preprocessing	9
2.2.1 Missing value and Data imputation.....	9
2.2.2 Feature engineering.....	9
2.2.3 Data Normalization.....	10
2.3 Algorithms.....	10
2.3.1 Linear Regression using stepwise.....	10
2.3.2 eXtreme Gradient Boosting (XGBoost)	11
2.3.3 Long Short-Term Memory (LSTM)	11

2.4 Experiment Design.....	12
2.5 Forecasting evaluation	13
Chapter 3 Results	14
3.1 Summary statistics	14
3.2 Comparison of different stations	14
3.2.1 Performance of Linear regression using stepwise.....	14
3.2.2 Performance of XGBoost.....	15
3.2.3 Performance of LSTM.....	15
3.3 Comparing of different models.....	15
3.4 Effect of Adding Nearby Station Pollutants Features.....	16
3.4.1 Pearson Correlation.....	16
3.4.2 Effect of Adding Nearby Station Pollutants Features.....	16
3.5 Comparing of using different length of historical data.....	17
Chapter 4 Discussion	18
Chapter 5 Conclusion	21
Reference.....	22



List of Tables

Table 1. Air quality and meteorological data used in this study	36
Table 2. Missing value of 15 pollutants in 2018 and 2019.....	37
Table 3. PM_{2.5} value of 13 stations in 2018 and 2019.....	38
Table 4. PM_{2.5} value of 13 stations in 2018.....	39
Table 5. PM_{2.5} value of 13 stations in 2019.....	40
Table 6. Main results of R², MAE, and RMSE at all stations for PM_{2.5} using Linear regression.....	41
Table 7. Main results of R², MAE, and RMSE at all stations for PM_{2.5} using XGBoost.....	42
Table 8. Main results of R², MAE, and RMSE at all stations for PM_{2.5} using LSTM.....	43
Table 9. Pearson correlation of PM_{2.5} concentrations with 3 hour-lag among 13 stations.....	44
Table 10. Pearson correlation of PM_{2.5} concentrations with 3 hour-lag among 13 stations.....	45
Table 11. Prediction outcomes while training with different length of historical data.....	46



List of Figures

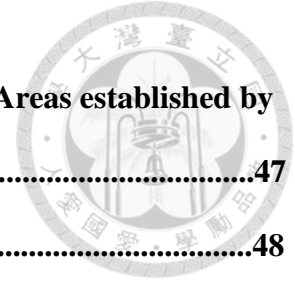


Figure 1. The distribution of air quality monitoring stations in Taipei Areas established by the Taiwan Environmental Protection Administration (EPA).....	47
Figure 2. Linear Regression prediction outcomes in 2019.....	48
Figure 3. XGBoost prediction outcomes in 2019.....	49
Figure 4. LSTM prediction outcomes in 2019.....	50
Figure 5. The comparison of linear regression, XGBoost, and LSTM in prediction performance among all stations when the meteorological and pollutants data are only input.....	51
Figure 6 The comparison of XGBoost prediction performance among six stations when adding pollutants data from nearby stations.....	52
Figure 7 The comparison of XGBoost prediction performance in RMSE and R2 at Tucheng station.....	53
Figure 8 The feature importance of the XGBoost model in Tucheng using 24 hours in 2018 for training.....	54
Figure 9 The feature importance of the XGBoost model in Tucheng using 8 hours in 2014-2018 for training.....	55

Introduction

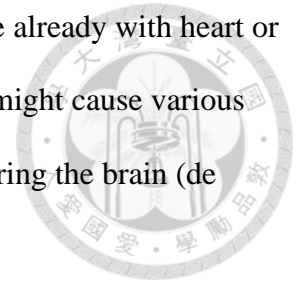
Rapid development of the economy has caused many negative environmental impacts with air pollution being one of them. Air pollution refers to the aggregation of harmful materials in the air. An air pollutant is possibly causing adverse health effects on humans. According to an official report published by World Health Organization in 2008, it mentioned for three-quarters of the world's population, the air pollution concentration values of living environments exceed the WHO's guideline limits; moreover, indoor and outdoor air causes about 7 million premature deaths every year (World Health, 2015).

Besides the statistic of deaths, plenty of research over the past ten years has provided robust evidence showing that poor air quality was responsible for adverse effects on health (Bai et al., 2018; Ning, Ji, Li, & Sang, 2019; Qiu et al., 2019; Shou et al., 2019; Song et al., 2017).

Particulate matter (PM) below 2.5 μm , which is called $\text{PM}_{2.5}$, is recognized as a major source of mortality among air pollutants. With the growth of the public concern, it has been widely studied around the world and has been found to adversely affect human health, including such problems as cardiovascular, cerebrovascular, and pulmonary diseases (Puett et al., 2009; Stafoggia et al., 2014; C. F. Wu et al., 2016). Moreover, according to existing research (Z. Chen, Wang, Ma, & Zhang, 2013; Dockery et al., 1993; Pope Iii et al., 2002; Sun et al., 2005; Xu, Zhang, Zhang, & Li, 2016; Yu & Stuart, 2017), $\text{PM}_{2.5}$ are found to be strongly correlated with effects of cardiovascular disease.

In conclusion, $\text{PM}_{2.5}$ can penetrate deeply into the lungs when human inhales, causing mainly cardiopulmonary disease but not limited, which includes: chronic bronchitis and nonfatal heart attacks, such as cardiovascular disease (Pope et al., 2004), respiratory symptoms (Dominici et al., 2006), diabetes (Y. Yang et al., 2018) and other adverse influence. Especially, those

vulnerable groups such as children, elders (Simoni et al., 2015), and people already with heart or lung diseases are the most vulnerable. De Prado Bert also observed PM_{2.5} might cause various neurodegenerative diseases by penetrating the blood-brain barrier and entering the brain (de Prado Bert, Mercader, Pujol, Sunyer, & Mortamais, 2018).



However, the short-term effects of PM_{2.5} exposures are relatively less discussed but still of interest to epidemiologists. Some previous epidemiological studies have assessed the effects and observed inconsistent results, indicating that short-term exposure to PM_{2.5} is associated with increased (Chang et al., 2015; Hoffmann et al., 2012; Jacobs et al., 2012; H. Lin et al., 2017; Mar et al., 2005; S. Wu et al., 2013) and decreased changes (Ibald-Mulli et al., 2004; Mirowsky et al., 2015) in blood pressure. For body function, some studies also showed that short-term PM_{2.5} exposures would reduce lung function. For the asthma inhaler user, Williams found that the usage would increase per 1-ug/m³ (Williams, Phaneuf, Barrett, & Su, 2019). Some research also suggests that for every 10-ug/m³ increase there was a reduction in daily peak expiratory flow (Yamazaki et al., 2011).

As a result of health effects, environmental prediction can be beneficial to the protection of human health and welfare from pollution. To monitor and control the possible exposure, many countries have developed their own real-time monitoring network (e.g., <http://www.pm25.in/>) using various stations.

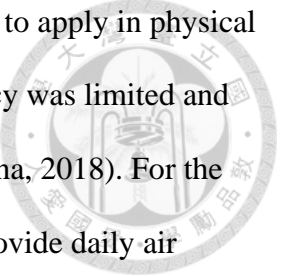
In East Asia, the Taiwan Environmental Protection Administration (EPA) set up 19 weather stations in 1980 in order to observe hourly air pollution data in major cities and to report through the internet in real time. In 1993, the Taiwanese government developed the Taiwan Air Quality Monitoring Network. Until 2021, 78 national air monitoring stations have been established all over Taiwan.

With these stations, the concentration of PM_{2.5} can be checked anytime so that people can decide whether to go out or try to avoid polluted areas. However, in order to know the future air pollution status in advance, we must rely on an air pollution prediction system (T. Liu, Lau, Sandbrink, & Fung, 2018; Y. Wang, Sun, Yang, & Yuan, 2016; Yang, Huang, & Li, 2018).

How to correctly use a single station point to collect history and real-time concentration data and then consider the spatiotemporal correlation among multiple stations hence are important in the air quality prediction field to avoid exposure to hazardous pollutants and protect ourselves from adverse health effects. Moreover, an effective model has a high application value for early warning since it can provide useful information for either guiding government policymaking or vulnerable people's short-term hazard assessment guideline.

PM_{2.5} predictions are challenging because many factors strongly influence PM_{2.5}. Related investigations show that estimation of PM_{2.5} from meteorological measures was carried out by researchers using nonlinear exposure-lag-response models (Z.-Y. Chen et al., 2018).

In recent years, a wealth of research (Cho, Lee, Kwon, & Kim, 2019; Corani, 2005; Delavar et al., 2019; Elangasinghe, Singhal, Dirks, Salmond, & Samarasinghe, 2014; Franceschi, Cobo, & Figueredo, 2018; Xuefei Hu et al., 2014; Maharani & Murfi, 2019; Mingjian, Guocheng, Xuxu, & Zhongyi, 2011; Rybarczyk & Zalakeviciute, 2018; Soh, Chang, & Huang, 2018; J. Wang & Song, 2018; Yi, Zhang, Wang, Li, & Zheng, 2018; Q. Zhou, Jiang, Wang, & Zhou, 2014) has been conducted to predict air pollution. According to different classification aspects, categories of air quality forecasting (AQF) systems might differ from much research (Cheng et al., 2021; Lee et al., 2020; Y. Li, Jiang, She, & Lin, 2018; L. Lin, Chen, Yang, Xu, & Fang, 2020; Ma, Yu, Qu, Xu, & Cao, 2020; Y. Zhou et al., 2019). Generally, physical models and machine learning models are two types of techniques that are used to forecast air quality.



In the 1990s, scientists applied various atmospheric dynamics methods to apply in physical models with complicated equations to calculate great iterations. The accuracy was limited and the importance of new and old data was not able to be identified (Marriboyina, 2018). For the time being, chemical transport models (CTMs) have been widely used to provide daily air quality forecasts (Ghim et al., 2017; Mathur, Yu, Kang, & Schere, 2008; Otte et al., 2005; Žabkar et al., 2015). However, uncertainties in emission inventories and meteorological forecasts as the key parameters of these models might lead to incomplete physical and chemical mechanisms in the CTMs, related to the substantial prediction errors in real values (Cobourn, 2010; Lv, Cobourn, & Bai, 2016).

Apart from physical methods, some traditional $PM_{2.5}$ prediction methods have focused mostly on statistic models (Kiesewetter, Schoepp, Heyes, & Amann, 2015; Lu & Wang, 2005). However, traditional methods are not capable of processing a large amount of multidimensional nonlinear data. Also, the complexity between $PM_{2.5}$ concentration and other climate features makes it more difficult. It is against this backdrop that machine learning models have attracted considerable attention given the numerous benefits they offer. For example, machine learning models provide a new way to analyze air quality in the absence of a physical model (Kurt & Oktay, 2010) by quantifying the underlying complex relationships between air pollutants and potential predictors based on big data sets under various atmospheric conditions (Cobourn, 2010; Hrust, Klaić, Križan, AntoniĆ, & Hercog, 2009). They also provide a promising approach towards dealing with complex nonlinear relationships between various interacting predictors (Zhan et al., 2017). It removes the classical statistical process, which consists of hypothesis distribution, a mathematical model fitting, hypothesis testing and determination of the P-value. Previous studies have indicated that an air quality prediction model is worthy of studying with

the big data and developing machine learning techniques (Huang, Chen, Hwang, Tzeng, & Huang, 2018; Mahajan, Chen, & Tsai, 2018).



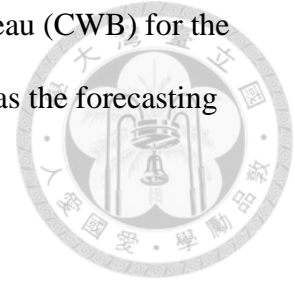
The common machine learning and deep learning models for time series problems include support vector regression (SVR), random forests (RF) (T. Liu et al., 2018), gradient boosting decision tree, multi-layer perceptron (MLP) or called artificial neural networks (ANNs), long short-term memory neural network (LSTM), and so on (Witten, Frank, & Hall, 2011).

Tree-based machine learning methods (e.g., RF) feature tackling linear and nonlinear problems with extra feature importance as a reference of feature values. In another study, RF algorithms have several advantages and have been successfully applied in different countries (X. Hu et al., 2017; Stafoggia et al., 2019; Wei et al., 2019). In particular, another popular machine learning algorithm, the gradient boosting decision tree (GBDT) (Jerome, 2001), namely by iterating multiple trees to make final decisions, is preferred for big data mining due to its interpretability and efficiency; compared to logistic regression, which can only be used for linear regression, all linear or nonlinear problems can be applied to GBDT. It exhibits a greater ability of robustness and generalization to handle complex correlated variables (P. Li, 2012).

In 2016, Chen et al., from the University of Washington, promoted a robust algorithm, named eXtreme Gradient Boosting (XGBoost), based on the GBDT (T. Chen & Guestrin, 2016). Pan (Pan, 2018) has applied the XGBoost algorithm to predict hourly $PM_{2.5}$ concentrations in China. He compared the results with prediction from various models including the random forest, support vector machine, linear regression, and decision tree regression. Among these, XGBoost algorithm demonstrated the best performance in air quality forecasting.

In this paper, the $PM_{2.5}$ forecasting model is proposed using XGBoost, as well as the Long Short-Term Model (LSTM). Air pollution data was collected in 2018-2019 from the Taiwan

Environmental Protection Administration (EPA) and Central Weather Bureau (CWB) for the Taipei area, and was combined into 23 features. We used the data in 2019 as the forecasting testing data.



The contribution of this paper is summarized as below:

- 1) We proposed an efficient small-region prediction model and set up a prediction application to forecast the $PM_{2.5}$ after three hours (3-h $PM_{2.5}$).
- 2) We implemented two well-known $PM_{2.5}$ prediction models, XGBoost and LSTM, for the 3-h $PM_{2.5}$ prediction.
- 3) A comparative analysis was performed for $PM_{2.5}$ prediction in between stations in the Taipei Area.
- 4) We compared our comparative analysis results with other studies in the similar study area.
- 5) We discussed several possible methods to enhance the prediction.

In the following of the paper, Section 2 presents the methods and materials used in the analysis, Section 3 and Section 4 present discussion and results respectively, and conclusions are then drawn in Section 5.

Materials and Method



2.1 Data Collection

2.1.1 Database

The EPA and CWB databases constitute the main sources for air quality forecasting. These systems collect air quality data in Taiwan every hour.

In our study, we adopted 227,760 samples collected from the EPA database from 2018 to 2019. Those data were collected from 13 general air monitoring stations in Taipei area, that is, Taipei City and New Taipei City, including (1) Xizhi, (2) Wanli, (3) Xindian, (4) Tucheng, (5) Banqiao, (6) Xinzhuang, (7) Cailiao, (8) Linkou, (9) Shilin, (10) Zhongshan, (11) Wanhua, (12) Guting, (13) Songshan districts. **Figure 1** shows the distribution of these 13 stations. These 13 stations in Taipei area were considered because they are situated in the most populated area and the financial center in Taiwan.

Additionally, CWB has built an automatic weather station that records weather data including pressure every hour. We extracted the information of station pressure from CWB and combined it with the 16 dimensions of EPA data, as presented in **Table 1**. According to Chuang et al. (2008), high-pressure peripheral circulation, pacific high-pressure systems stretching westerly and weak high-pressure systems are related to terrain blocking and aerosol accumulation (Chuang et al., 2008). Furthermore, features in the prediction model also include time data like the hour of the day, day of the week, weekend or not and year to learn the trend and period of the temporal index. In summary, 23 features are used in our models.

Since May 2014, EPA has annually published linear regression equations for calibration referring to the United States Environmental Protection Agency until Sep 2019. In our study, all

data before Sep 2019 were calibrated by regression and then published publicly. After Sep 2019, it no longer needs to be calibrated since new instruments have passed tests related to regulations.



2.1.2 Study Area

In our study, we selected Taipei City and New Taipei City as our study area. Taipei City and New Taipei City cover the biggest part of the Taipei area, with an area of 2,324 km² and a population of 6.59 million in 2021, which accounts for almost 30% of Taiwan. It is the center of politics, commerce, and culture in Taiwan and people commuting in this area nowadays are forced to face the high-level invasion of PM_{2.5}.

As there is fast-growing concern about air quality issues, Taiwan began to monitor PM_{2.5} concentrations in 2005. Due to the large population of the Taipei area, it also has the densest monitoring stations (19) in Taiwan (76), which are 25% of the total stations with nearly 6% area in Taiwan. The dense monitoring stations tend to provide a more reliable estimation of PM_{2.5} for citizens compared with other counties, which is also the reason we chose the Taipei area in this study.

Figure 1 shows the location of general stations in Taipei area. Among them, Shilin (SL) station, Guting (GT) station, Songshan (SS) station, Wanhua (WH) station, and Zhongshan (ZS) station are located in Taipei City, while others are in New Taipei City.

Generally, PM_{2.5} pollution is often severe in winter due to the geographical characteristics of the area, which roughly correspond to areas located within the Taipei Basin. Researchers have utilized a weather map to classify the weather patterns for aerosol events in Taipei and found aerosol accumulation often comes with enhanced atmospheric stability and weak winds as a result of geography (Chuang et al., 2008).

2.2 Data preprocessing

The data sets mainly used in our work are from 13 EPA air quality stations from 2018 to 2019. CWB data and time features were matched to the stations and study duration. We use totally 23-dimensional data from the previous eight hours to process the data to train the model. Data preprocessing (Su, Xu, & Tang, 2017) is highly correlated between ozone, PM10, and PM_{2.5}. The data preprocessing steps are as shown in the following:



2.2.1 Missing value and Data imputation

Table 2 shows the missing percentage of the raw data. The main reason for the missing information is measure instrument failure. Additionally, in this study the abnormal values are marked as missing. In order to collect more valid data, we dropped those missing values last for more than three days. And then we used linear interpolation methods to fill the missing.

The linear interpolate method involves using straight linear to construct new data points within the range of a discrete set of known data points. The formula with one data point (x, y) available between (x_1, y_1) and (x_2, y_2) is as following:

$$y = L(x) = y_1 \left(\frac{x - x_2}{x_1 - x_2} \right) + y_2 \left(\frac{x - x_1}{x_2 - x_1} \right)$$

2.2.2 Feature engineering

Some cyclical variables (e.g., wind direction, hour) would be mapped onto a circle under some sine and cosine transforming so that the lowest value for that variable appears right next to the largest value. For hour, the formula is as follows:

$$hour(x - axis) = \sin\left(2 \times \Pi \times \frac{hour}{23}\right)$$

$$hour(y - axis) = \cos\left(2 \times \Pi \times \frac{hour}{23}\right)$$

For wind, we also combined speed and direction to make the calculation reasonable. The following is the formula:

$$Wind(x - axis) = Wind_Speed \times \sin\left(\Pi \times \frac{Wind_Direction}{180}\right)$$

$$Wind(y - axis) = Wind_Speed \times \cos\left(\Pi \times \frac{Wind_Direction}{180}\right)$$

2.2.3 Data Normalization

When the scale of each variable is different, some variables might be dominated by others. In the neural network which uses the gradient descent method, the different scale possibly causes the network iterating for many times before it converges, or leads to its failure to converge. In this work we used log transformation for skewed pollutants data and Min-Max Normalization for other features to rescale the values between 0 and 1. The Min-Max formula is:

$$X_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2.3 Algorithms

2.3.1 Linear Regression using stepwise

Linear regression has been used in many different areas of forecasting and analytic studies (Cortina–Januchs, Quintanilla–Dominguez, Vega–Corona, & Andina, 2015). However, due to its simple structure, some research suggested that they can only predict the general trend or a short term trend (Menon, Bharadwaj, Shetty, Sanu, & Nagendra, 2017). Hence, linear regression would then be treated as a baseline in our study. We feed all input variables into a single model. Besides, we implemented stepwise based on the p-value to boost its prediction performance.

2.3.2 eXtreme Gradient Boosting (XGBoost)

In 2016, Tianqi Chen proposed a robust algorithm named XGBoost which combines software and hardware optimization techniques perfectly, and yield superior results than other methods (T. Chen & Guestrin, 2016). XGBoost represents a highly efficient kind of gradient boosting algorithm, enabling gradient boosting “on steroids” (also known as one of a reason called “Extreme Gradient Boosting”). It offers a parallel boost to the tree and features accurately solving many data science issues accurately. In the structure of the model, it provides several default hyper-parameters which could be tuned or manually adjusted to enhance the model performance. In this study, we conducted grid search methods with three-fold cross validation and repetition of 300 iterations to find the best combination of hyper-parameters which best perform the prediction.

2.3.3 Long Short-Term Memory (LSTM)

The LSTM is an improved version of recurrent neural network (RNN) with the addition of a memory cell able to store information for a long time. As an abbreviation of Long Short-Term Memory, LSTM is commonly used for sequential data processing, such as voice or text processing or other time series problems. LSTM is capable of learning long-term dependencies conquering the weakness in RNN. In each LSTM cell, there are three Sigmoid functions and one Hyperbolic Tan function. For long term problems, LSTM could handle noise, distributed representation, and continuous values (Qiao et al., 2019).

The choice of the optimizer also plays an important role in training. The Adam optimization algorithm, a variant of stochastic gradient descent in deep learning, has recently gained popularity in the fields of computer vision and natural language processing. Thus, in our work we built the LSTM models using the Adam optimizer.

2.4 Experimental Design

The previous studies have primarily focused on long-term forecasting of PM_{2.5} concentrations (X. Li, Peng, Hu, Shao, & Chi, 2016; Nguyen, Starzyk, Goh, & Jachyra, 2012). Given the severity of health impact, reliable and precise PM_{2.5} forecasting is in urgent need. Hence, some studies used time lag for 1-10 hours (Tsai, Zeng, & Chang, 2018). In our work, we used the next three-hour as the limit in order to provide people sufficient time to respond.

Firstly, we used the historical eight hours data to predict the next three-hour data in the same station. Next, we picked the models based on the performance in last stage and considered the spatial influence by nearest stations. Finally, we focused on one station and compared the performance of various combinations of past time periods. Overall, the system workflow was designed to perform model training. The 48 models were conducted in this experiment based on three stages:

Stage One: 39 models, trained using 23 variables in 13 stations for eight hours

Stage Two: six models, trained using 31 variables in six stations including nearby stations' pollutants data for eight hours.

Stage Three: three models, trained with different historical data: eight hours in one year, eight hours in five years, 24 hours in one year.

In stage two, we explored the effect of adding nearby stations pollutant features into our model. We implemented the Pearson correlation analysis of PM_{2.5} for three-hour lags when $t=k$ and $t=k-3$ in-between stations to decide which stations should be added to the variables. Based on the performance of stage one, we chose six models, the best three and the poorest three models, and continued to analyze them.

In stage three, the different training periods were compared. We chose the station with the

best performance in R-squared and compared the models' improvement with different historical data.



2.5 Forecasting evaluation

This paper uses the information of 13 air quality from 2018 to 2019. We use eight hours as the time period to predict the PM_{2.5} concentration value in the next three hour. For prediction, the data of 2018 are used as the training set, and the data of 2019 as the testing set.

In order to quantitatively evaluate the prediction accuracy of the proposed model, we used root-mean-square error (RMSE), mean-absolute error (MAE), and R-squared (R²). RMSE and MAE are commonly used as a measure of the difference between predicted and observed values. The smaller the RMSE and MAE value is, the better the performance of the prediction model there is. However, for the R-squared, which measures the suitability of the model to the sample standard deviation of the predicted value, the larger the value is, the better effect the model has. Also, the actual value is between zero and one. The closer to one, the higher the suitability of the model. Equations are given below:

$$MAE = \frac{1}{N} \sum_{i=1}^N |O_i - P_i|$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (O_i - P_i)^2} = \sqrt{MSE}$$

$$R - squared = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} = 1 - \frac{MSE}{Var(O)}$$

where n is the number of data points, O_i is the observed value (true value), P_i is the prediction value, and \bar{O} is the mean value.

Results



In the Materials and Methods section, we have proposed three prediction models and three experiment stages.

3.1 Summary statistics

The available concentration of $PM_{2.5}$ based on data collected from 13 air quality stations in the Taipei Area from 2018 to 2019 are summarized in **Tables 3-5**. **It can be observed** that the average hourly $PM_{2.5}$ concentrations for all stations range between $12.87 \mu\text{g}/\text{m}^3$ to $15.85 \mu\text{g}/\text{m}^3$, with Linkou having the lowest value and Banqiao the highest. In terms of standard deviation, except for Wanli, all stations had high values over 9. We also found that there were several continuous missing values that spanned more than three days. For instance, Guting in both 2018 and 2019 had periods of three consecutive months of missing hours, so we did not use it in the models. Missing counts can also be inferred from **Table 3-5**.

3.2 Comparison between various stations

Experiments were conducted to determine the predictive performance of the proposed models, with R^2 , RMSE, and MAE serving as performance metrics.

3.2.1 Performance of linear regression using stepwise

We first evaluated the performance of linear regression. The results are presented in **Figure 2 and Table 6**. As a baseline of prediction models, the results of the testing set in 2019 showed that the values of R^2 ranged from 0.37 to 0.56 and the RMSE ranged from $5.27 \mu\text{g}/\text{m}^3$ to $7.73 \mu\text{g}/\text{m}^3$. For R -squared, the best three stations were Tucheng, Cailiao, and Xinzhuang station, while the poorest ones are Shilin, Wanhua, and Songshan. For RMSE, the lowest values fall at Wanli, Cailiao, and Xinzhuang station, while the highest are Shilin, Guting, and Wanhua. For MAE, all stations have similar results as RMSE.

3.2.2 Performance of XGBoost

Through the same grid search process of models with three-fold cross validation and 300 repetitions for each station, the models were finally tuned with the best hyper-parameters among the number of estimators, max depth, learning rate, subsample, and minimum child weight. The results are presented in **Table 7 and Figure 3**.

The results show that the performance of the models in the testing set is the best in R-squared when predicting Tucheng, Cailiao, and Xinchuang, and the poorest in Shilin, Wanhua, and Songshan. In terms of RMSE and MAE, the lowest values occur in Wanli, Cailiao, and Xinzhuang, while the highest ones occur in Shilin, Guting, and Wanhua.

3.2.3 Performance of LSTM

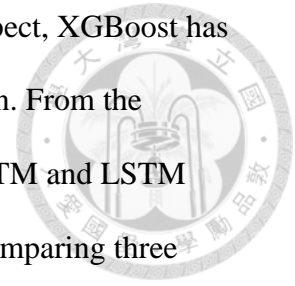
Table 8 and Figure 4 present the results of LSTM.

It can be seen from the tables that the prediction models of LSTM have good R-squared performance in Tucheng, Cailiao, and Xinzhuang, but poor performance in Shilin, Songshan, and Wanhua. For the RMSE, Wanli, Xizhi, and Xinzhuang had the best performance of LSTM in the testing set, while Shilin, Guting, and Wanhua had the poorest ones. For MAE, unlike RMSE in third place, the Cailiao is better than Xinzhuang and Zhongshan is worse than Wanhua. However, their values are quite similar.

3.3 Comparing various models

In our work, approximately 8760 records for each station are used as the testing data set from 1 Jan 2019 to 31 Dec 2019 using the established prediction models such as Linear Regression, XGBoost, and LSTM. **Figure 5** shows the results of the three algorithms in predicting the effectiveness of the PM_{2.5} concentrations after three hours. XGBoost errors on an average 13 stations in the RMSE and MAE aspects are 6.22 and 4.47 respectively, which are

lower than those of the LSTM and Linear Regression. In the R-squared aspect, XGBoost has 0.53 on average, which is higher than those of LSTM and Linear regression. From the perspective of individual stations, XGBoost consistently outperformed LSTM and LSTM consistently outperformed Linear Regression for all stations. That is, by comparing three evaluation methods, XGBoost is better than LSTM and Linear Regression in predicting the PM_{2.5} value after three hours using pollutants and meteorological data for all 13 stations.



3.4 Effect of Adding Nearby Station Pollutants Features

3.4.1 Pearson Correlation

Table 9 illustrates the results of Pearson correlation for three-hour lags when $t=k$ and $t=k-3$ in-between stations. Based on the results of Stage 1, we then chose Tucheng, Cailiao, and Xinzhuang with the best three performances and Shilin, Songshan, and Wanhua with the poorest ones in terms of R-squared scores. Table 3 indicates the Pearson correlation outcomes. It can be seen that Banqiao and Xinzhuang are closest to Tucheng station, which also with the highest correlations. Note Cailiao is further from Tucheng than Xindian but its PM_{2.5} concentrations is more relevant. There are several similar cases in Songshan, Wanhua, Shilin and Cailiao, wherein the most related stations are not always the nearest ones. Only in Xinzhuang station the most relevant stations are consistent with the nearest ones.

3.4.2 Effect of Adding Nearby Station Pollutants Features

Next, we selected the same six stations and three of the most relevant stations for each station for analyzing the effect of adding nearby station pollutants features. The additional variables are presented in **Table 10** and **Figure 6**. The results indicate that the effect may provide some improvement ranging from 0.02 to 0.07 in the R-squared aspect for the prediction

outcome. In the Wanhua station, it provided the most significant enhancement among the three stations. Overall, these stations have an average improvement of 0.03.



3.5 Comparing using various lengths of historical data

To realize the effect of using the various lengths of historical data, we chose Tucheng station and implemented one training with 24 hours based on the same 23 variables using data from 2018 and another with 8 hours using data from 2014 to 2018.

The results showed that prolonging the historical hour to 24 hours did not improve the performance; instead, this resulted in slightly worse results for R-squared, MAE, and RMSE than the original one. However, without changing the length of historical hours data, we additionally collected more observations from 2014 to 2018 which enhanced the performance of all scores (R-squared, MAE, and RMSE). The results are presented in **Table 11** and **Figure 7**.

Discussion

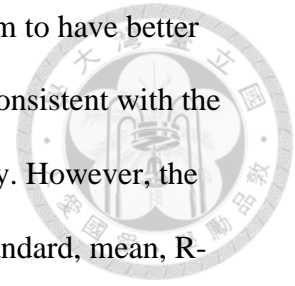


This study collected 227,760 hourly data from EPA and CWB from 2018 to 2019, and we implemented three models for predicting the $PM_{2.5}$ concentration three hours later using historical eight-hour data.

Air quality forecast has aroused attention from governments and scientists for improving the environmental quality of citizens. However, $PM_{2.5}$ predictions are still challenging because spatial and temporal variations strongly influence the formation and transportation of $PM_{2.5}$ (Chu, Huang, & Lin, 2015; Mandal et al., 2020). These limitations of time and space also lead to variations of predictive performance when applying different models in different countries (Deleawe, Kuszniir, Lamb, & Cook, 2010; Zhao, Zhang, Wang, Bai, & Liu, 2010). Even in different areas within the same country, there might exist divergence due to multiple factors (Lee et al., 2020). To make a more meaningful comparison, we compared our predictions with previous studies whose area of interest was also in Taipei. In our study, the XGBoost model we proposed using meteorological and pollutants data for the next three hour resulted in a similar or lower MAE (3.5-5.6) and RMSE (5-8) value compared to previous studies with the same prediction period (Lee et al., 2020; Shih, To, Nguyen, Wu, & You, 2021; Tsai et al., 2018; Y. Zhou et al., 2019) in which LSTM, random forest, and Gradient Boosting Decision were used. However, the outcome differs from evaluation methods, station characteristics, and training and testing periods. Some studies evaluated using R-squared and RMSE (Ho, Chen, & Hwang, 2020; Lee et al., 2020; Y. Li et al., 2018; L. Wang et al., 2020b); some used error rate or NRMSE, which makes the comparison more difficult.

In our study, we found the similarity in stations' performance in different models. Tucheng, Cailiao, and Xinzhuang had the best score in R-squared in XGBoost, LSTM, as well as Linear

regression. In **Figure 7**, it can be seen that stations in New Taipei City seem to have better performance in R-squared and RMSE than those in Taipei City, which is consistent with the prediction in our work that the best three stations are all in New Taipei City. However, the underlying mechanism is unclear. We evaluated the relationship among standard, mean, R-squared, and RMSE and found a significant difference.



In the comparison of Tucheng and Cailiao, we found that from the perspective of RMSE, Cailiao is better than Tucheng but from that of R-squared it is worse than Tucheng, which is because with the same mean squared error, the bigger the standard deviation, the higher the R-squared, according to the formula in the Methods section. However, a higher deviation is not absolute to the high R-squared; for instance, Shilin station whose deviations are similar to Tucheng but lower R-squared.

In terms of enhancement for model prediction, we implemented Pearson correlation in between stations and found that for most stations the most relevant station might not always be the closest ones. This might be because of the monsoon characteristics of Taiwan, which is consistent with the findings of previous studies (Beckerman et al., 2013; Hwa-Lung & Chih-Hsin, 2010). In response, we added the pollutant data of nearby stations and found improvement in R-squared among all models ranging from 0.02 to 0.07, which means this information might represent an important factor to predictions. Furthermore, when we implemented the comparison of different lengths of historical data, we found that the most important features often fell in the past eight hours from the feature importance figures.

To clarify the prediction hour over time, we additionally ran two more models, one using 24 hours historical data in two years and the other using eight hours in five years respectively. We found that longer training time does not improve the prediction and might even be slightly worse

than the original one. In the feature importance information (**Figure 8-9**), the most important features are all within eight hours. However, using five years data with the same historical length provided a better prediction outcome. This implies that the performance might be enhanced by the addition of observations and features related to influence by nearby stations.

Some studies in Taiwan used land-use parameters like traffic, population, satellite data or other potential factors to evaluate the air pollution for daily or monthly prediction, which also showed good results (Kibirige, Yang, Liu, & Chen, 2021; D.-R. Liu, Lee, Huang, & Chiu, 2020; L. Wang et al., 2020a). We plan to study these issues in the future.

Conclusion

This paper demonstrated the PM_{2.5} forecasting model using XGBoost, LSTM, and Linear Regression. The air pollution data were extracted from Taiwan EPA and CWB at the individual general stations from New Taipei City and Taipei City in 2018-2019. We found that the XGBoost model has a higher accuracy (R-squared, RMSE, and MAE) than the two other models, and its performance is similar or better than previous studies. The study also investigated the performance divergence among stations. We found that stations in New Taipei city are more predictable than those in Taipei City. However, the mechanism is still unclear. This work also proposes a feasible direction for improving predictions by adding observations or geographical features.

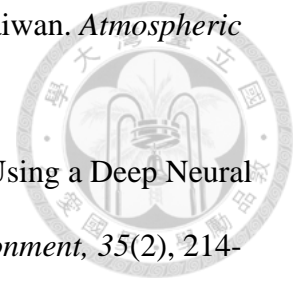
One limitation of this study is that as the population, emissions or traffic data is not present in our datasets, the influence of these factors cannot be predicted. Future work should focus on the methods for enhancing model prediction, particularly by investigating divergence in spatial factors.

References



- Bai, X., Liu, Y., Wang, S., Liu, C., Liu, F., Su, G., . . . Yan, B. (2018). Ultrafine particle libraries for exploring mechanisms of PM_{2.5}-induced toxicity in human cells. *Ecotoxicol Environ Saf*, *157*, 380-387. doi:10.1016/j.ecoenv.2018.03.095
- Beckerman, B. S., Jerrett, M., Serre, M., Martin, R. V., Lee, S. J., van Donkelaar, A., . . . Burnett, R. T. (2013). A hybrid approach to estimating national scale spatiotemporal variability of PM_{2.5} in the contiguous United States. *Environ Sci Technol*, *47*(13), 7233-7241. doi:10.1021/es400039u
- Chang, L. T., Chuang, K. J., Yang, W. T., Wang, V. S., Chuang, H. C., Bao, B. Y., . . . Chang, T. Y. (2015). Short-term exposure to noise, fine particulate matter and nitrogen oxides on ambulatory blood pressure: A repeated-measure study. *Environ Res*, *140*, 634-640. doi:10.1016/j.envres.2015.06.004
- Chen, T., & Guestrin, C. (2016). *XGBoost*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Chen, Z.-Y., Zhang, T.-H., Zhang, R., Zhu, Z.-M., Ou, C.-Q., & Guo, Y. (2018). Estimating PM_{2.5} concentrations based on non-linear exposure-lag-response associations with aerosol optical depth and meteorological measures. *Atmospheric Environment*, *173*, 30-37. doi:10.1016/j.atmosenv.2017.10.055
- Chen, Z., Wang, J.-N., Ma, G.-X., & Zhang, Y.-S. (2013). China tackles the health effects of air pollution. *The Lancet*, *382*(9909), 1959-1960. doi:10.1016/s0140-6736(13)62064-4
- Cheng, F.-Y., Feng, C.-Y., Yang, Z.-M., Hsu, C.-H., Chan, K.-W., Lee, C.-Y., & Chang, S.-C. (2021). Evaluation of real-time PM_{2.5} forecasts with the WRF-CMAQ modeling system

and weather-pattern-dependent bias-adjusted PM2.5 forecasts in Taiwan. *Atmospheric Environment*, 244. doi:10.1016/j.atmosenv.2020.117909



Cho, K., Lee, B.-y., Kwon, M., & Kim, S. (2019). Air Quality Prediction Using a Deep Neural Network Model. *Journal of Korean Society for Atmospheric Environment*, 35(2), 214-225. doi:10.5572/kosae.2019.35.2.214

Chu, H.-J., Huang, B., & Lin, C.-Y. (2015). Modeling the spatio-temporal heterogeneity in the PM10-PM2.5 relationship. *Atmospheric Environment*, 102, 176-182. doi:10.1016/j.atmosenv.2014.11.062

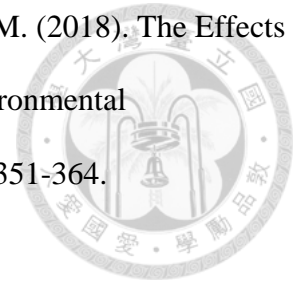
Chuang, M.-T., Chiang, P.-C., Chan, C.-C., Wang, C.-F., Chang, E. E., & Lee, C.-T. (2008). The effects of synoptical weather pattern and complex terrain on the formation of aerosol events in the Greater Taipei area. *Science of The Total Environment*, 399(1), 128-146. doi:<https://doi.org/10.1016/j.scitotenv.2008.01.051>

Cobourn, W. G. (2010). An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. *Atmospheric Environment*, 44(25), 3015-3023. doi:10.1016/j.atmosenv.2010.05.009

Corani, G. (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling*, 185(2-4), 513-529. doi:10.1016/j.ecolmodel.2005.01.008

Cortina-Januchs, M. G., Quintanilla-Dominguez, J., Vega-Corona, A., & Andina, D. (2015). Development of a model for forecasting of PM10 concentrations in Salamanca, Mexico. *Atmospheric Pollution Research*, 6(4), 626-634. doi:<https://doi.org/10.5094/APR.2015.071>

de Prado Bert, P., Mercader, E. M. H., Pujol, J., Sunyer, J., & Mortamais, M. (2018). The Effects of Air Pollution on the Brain: a Review of Studies Interfacing Environmental Epidemiology and Neuroimaging. *Curr Environ Health Rep*, 5(3), 351-364. doi:10.1007/s40572-018-0209-9



Delavar, M., Gholami, A., Shiran, G., Rashidi, Y., Nakhaeizadeh, G., Fedra, K., & Hatefi Afshar, S. (2019). A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran. *ISPRS International Journal of Geo-Information*, 8(2). doi:10.3390/ijgi8020099

Deleawe, S., Kuszniir, J., Lamb, B., & Cook, D. J. (2010). Predicting Air Quality in Smart Environments. *J Ambient Intell Smart Environ*, 2(2), 145-152. doi:10.3233/AIS-2010-0061

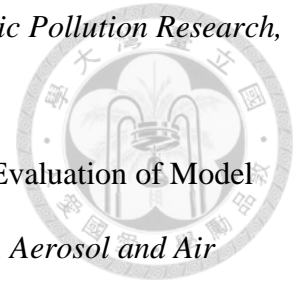
Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., . . . Speizer, F. E. (1993). An Association between Air Pollution and Mortality in Six U.S. Cities. *New England Journal of Medicine*, 329(24), 1753-1759. doi:10.1056/NEJM199312093292401

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., & Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10), 1127-1134. doi:10.1001/jama.295.10.1127

Elangasinghe, M. A., Singhal, N., Dirks, K. N., Salmond, J. A., & Samarasinghe, S. (2014). Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. *Atmospheric Environment*, 94, 106-116. doi:10.1016/j.atmosenv.2014.04.051

Franceschi, F., Cobo, M., & Figueredo, M. (2018). Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks,

Principal Component Analysis, and k-means clustering. *Atmospheric Pollution Research*, 9(5), 912-922. doi:10.1016/j.apr.2018.02.006



Ghim, Y. S., Choi, Y., Kim, S., Bae, C. H., Park, J., & Shin, H. J. (2017). Evaluation of Model Performance for Forecasting Fine Particle Concentrations in Korea. *Aerosol and Air Quality Research*, 17(7), 1856-1864. doi:10.4209/aaqr.2016.10.0446

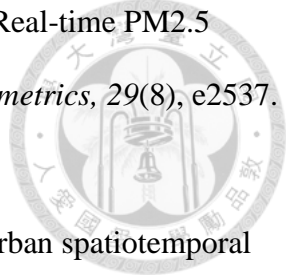
Ho, C. C., Chen, L. J., & Hwang, J. S. (2020). Estimating ground-level PM_{2.5} levels in Taiwan using data from air quality monitoring stations and high coverage of microsensors. *Environ Pollut*, 264, 114810. doi:10.1016/j.envpol.2020.114810

Hoffmann, B., Luttmann-Gibson, H., Cohen, A., Zanobetti, A., de Souza, C., Foley, C., . . . Gold, D. R. (2012). Opposing effects of particle pollution, ozone, and ambient temperature on arterial blood pressure. *Environ Health Perspect*, 120(2), 241-246. doi:10.1289/ehp.1103647

Hrust, L., Klaić, Z. B., Križan, J., Antonić, O., & Hercog, P. (2009). Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. *Atmospheric Environment*, 43(35), 5588-5596. doi:10.1016/j.atmosenv.2009.07.048

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM_{2.5} Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ Sci Technol*, 51(12), 6936-6944. doi:10.1021/acs.est.7b01210

Hu, X., Waller, L. A., Lyapustin, A., Wang, Y., Al-Hamdan, M. Z., Crosson, W. L., . . . Liu, Y. (2014). Estimating ground-level PM_{2.5} concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sensing of Environment*, 140, 220-232. doi:10.1016/j.rse.2013.08.032

- 
- Huang, G., Chen, L. J., Hwang, W. H., Tzeng, S., & Huang, H. C. (2018). Real-time PM2.5 mapping and anomaly detection from AirBoxes in Taiwan. *Environmetrics*, 29(8), e2537. doi:<https://doi.org/10.1002/env.2537>
- Hwa-Lung, Y., & Chih-Hsin, W. (2010). Retrospective prediction of intraurban spatiotemporal distribution of PM2.5 in Taipei. *Atmospheric Environment*, 44(25), 3053-3065. doi:10.1016/j.atmosenv.2010.04.030
- Ibald-Mulli, A., Timonen, K. L., Peters, A., Heinrich, J., Wolke, G., Lanki, T., . . . Pekkanen, J. (2004). Effects of particulate air pollution on blood pressure and heart rate in subjects with cardiovascular disease: a multicenter approach. *Environ Health Perspect*, 112(3), 369-377. doi:10.1289/ehp.6523
- Jacobs, L., Buczynska, A., Walgraeve, C., Delcloo, A., Potgieter-Vermaak, S., Van Grieken, R., . . . Nawrot, T. S. (2012). Acute changes in pulse pressure in relation to constituents of particulate air pollution in elderly persons. *Environ Res*, 117, 60-67. doi:10.1016/j.envres.2012.05.003
- Jerome, H. F. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. doi:10.1214/aos/1013203451
- Kibirige, G., Yang, M.-C., Liu, C.-L., & Chen, M. (2021). *Using Satellite Data on Remote Transportation of Air Pollutants for PM2.5 Prediction in Northern Taiwan*.
- Kiesewetter, G., Schoepp, W., Heyes, C., & Amann, M. (2015). Modelling PM2.5 impact indicators in Europe: Health effects and legal compliance. *Environmental Modelling & Software*, 74, 201-211. doi:10.1016/j.envsoft.2015.02.022

Kurt, A., & Oktay, A. B. (2010). Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks. *Expert Systems with Applications*, 37(12), 7986-7992. doi:10.1016/j.eswa.2010.05.093



Lee, M., Lin, L., Chen, C. Y., Tsao, Y., Yao, T. H., Fei, M. H., & Fang, S. H. (2020). Forecasting Air Quality in Taiwan by Using Machine Learning. *Sci Rep*, 10(1), 4153. doi:10.1038/s41598-020-61151-7

Li, P. (2012). Robust LogitBoost and Adaptive Base Class (ABC) LogitBoost.

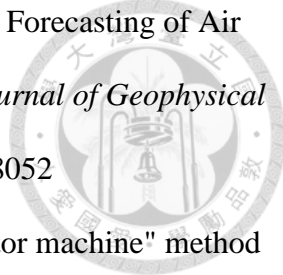
Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environ Sci Pollut Res Int*, 23(22), 22408-22417. doi:10.1007/s11356-016-7812-9

Li, Y., Jiang, P., She, Q., & Lin, G. (2018). Research on air pollutant concentration prediction method based on self-adaptive neuro-fuzzy weighted extreme learning machine. *Environ Pollut*, 241, 1115-1127. doi:10.1016/j.envpol.2018.05.072

Lin, H., Guo, Y., Zheng, Y., Di, Q., Liu, T., Xiao, J., . . . Wu, F. (2017). Long-Term Effects of Ambient PM_{2.5} on Hypertension and Blood Pressure and Attributable Risk Among Older Chinese Adults. *Hypertension*, 69(5), 806-812. doi:10.1161/HYPERTENSIONAHA.116.08839

Lin, L., Chen, C.-Y., Yang, H.-Y., Xu, Z., & Fang, S.-H. (2020). Dynamic System Approach for Improved PM_{2.5} Prediction in Taiwan. *IEEE Access*, 8, 210910-210921. doi:10.1109/access.2020.3038853

Liu, D.-R., Lee, S.-J., Huang, Y., & Chiu, C.-J. (2020). Air pollution forecasting based on attention-based LSTM neural network and ensemble learning. *Expert Systems*, 37(3), e12511. doi:<https://doi.org/10.1111/exsy.12511>

- 
- Liu, T., Lau, A. K. H., Sandbrink, K., & Fung, J. C. H. (2018). Time Series Forecasting of Air Quality Based On Regional Numerical Modeling in Hong Kong. *Journal of Geophysical Research: Atmospheres*, *123*(8), 4175-4196. doi:10.1002/2017jd028052
- Lu, W. Z., & Wang, W. J. (2005). Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. *Chemosphere*, *59*(5), 693-701. doi:10.1016/j.chemosphere.2004.10.032
- Lv, B., Cobourn, W. G., & Bai, Y. (2016). Development of nonlinear empirical models to forecast daily PM_{2.5} and ozone levels in three large Chinese cities. *Atmospheric Environment*, *147*, 209-223. doi:10.1016/j.atmosenv.2016.10.003
- Ma, J., Yu, Z., Qu, Y., Xu, J., & Cao, Y. (2020). Application of the XGBoost Machine Learning Method in PM_{2.5} Prediction: A Case Study of Shanghai. *Aerosol and Air Quality Research*, *20*(1), 128-138. doi:10.4209/aaqr.2019.08.0408
- Mahajan, S., Chen, L.-J., & Tsai, T.-C. (2018). Short-Term PM_{2.5} Forecasting Using Exponential Smoothing Method: A Comparative Analysis. *Sensors*, *18*(10). doi:10.3390/s18103223
- Maharani, D., & Murfi, H. (2019). Deep Neural Network For Structured Data - A Case Study Of Mortality Rate Prediction Caused By Air Quality. *Journal of Physics: Conference Series*, *1192*. doi:10.1088/1742-6596/1192/1/012010
- Mandal, S., Madhipatla, K. K., Guttikunda, S., Kloog, I., Prabhakaran, D., Schwartz, J. D., & GeoHealth Hub India, T. (2020). Ensemble averaging based assessment of spatiotemporal variations in ambient PM_{2.5} concentrations over Delhi, India, during 2010-2016. *Atmos Environ (1994)*, *224*. doi:10.1016/j.atmosenv.2020.117309

Mar, T. F., Koenig, J. Q., Jansen, K., Sullivan, J., Kaufman, J., Trenga, C. A., . . . Neas, L.

(2005). Fine particulate air pollution and cardiorespiratory effects in the elderly.

Epidemiology, 16(5), 681-687. doi:10.1097/01.ede.0000173037.83211.d6

Marriloyina, V. (2018). *A survey on Air Quality forecasting Techniques*.

Mathur, R., Yu, S., Kang, D., & Schere, K. L. (2008). Assessment of the wintertime performance

of developmental particulate matter forecasts with the Eta-Community Multiscale Air

Quality modeling system. *Journal of Geophysical Research*, 113(D2).

doi:10.1029/2007jd008580

Menon, S. P., Bharadwaj, R., Shetty, P., Sanu, P., & Nagendra, S. (2017, 15-16 Dec. 2017).

Prediction of temperature using linear regression. Paper presented at the 2017

International Conference on Electrical, Electronics, Communication, Computer, and

Optimization Techniques (ICEECCOT).

Mingjian, F., Guocheng, Z., Xuxu, Z., & Zhongyi, Y. (2011). *Study on Air Fine Particles*

Pollution Prediction of Main Traffic Route Using Artificial Neural Network. Paper

presented at the 2011 International Conference on Computer Distributed Control and

Intelligent Environmental Monitoring.

Mirowsky, J. E., Peltier, R. E., Lippmann, M., Thurston, G., Chen, L. C., Neas, L., . . . Gordon,

T. (2015). Repeated measures of inflammation, blood pressure, and heart rate variability

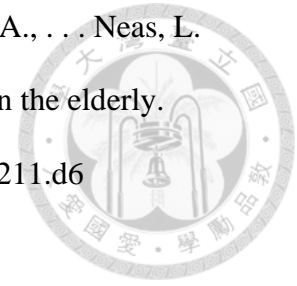
associated with traffic exposures in healthy adults. *Environ Health*, 14, 66.

doi:10.1186/s12940-015-0049-0

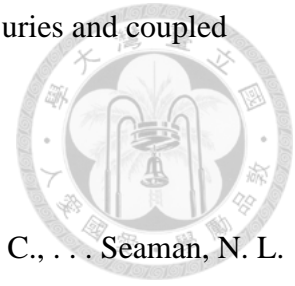
Nguyen, V. A., Starzyk, J. A., Goh, W. B., & Jachyra, D. (2012). Neural network structure for

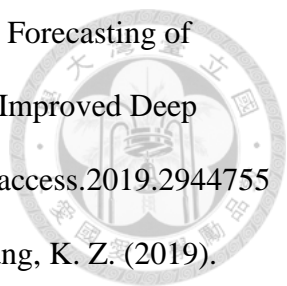
spatio-temporal long-term memory. *IEEE Trans Neural Netw Learn Syst*, 23(6), 971-983.

doi:10.1109/TNNLS.2012.2191419

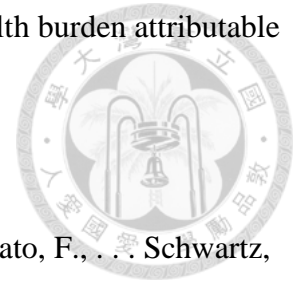



- Ning, X., Ji, X., Li, G., & Sang, N. (2019). Ambient PM_{2.5} causes lung injuries and coupled energy metabolic disorder. *Ecotoxicol Environ Saf*, *170*, 620-626.
doi:10.1016/j.ecoenv.2018.12.028
- Otte, T. L., Pouliot, G., Pleim, J. E., Young, J. O., Schere, K. L., Wong, D. C., . . . Seaman, N. L. (2005). Linking the Eta Model with the Community Multiscale Air Quality (CMAQ) Modeling System to Build a National Air Quality Forecasting System. *Weather and Forecasting*, *20*(3), 367-384. doi:10.1175/WAF855.1
- Pan, B. (2018). Application of XGBoost algorithm in hourly PM_{2.5} concentration prediction. *IOP Conference Series: Earth and Environmental Science*, *113*. doi:10.1088/1755-1315/113/1/012127
- Pope, C. A., 3rd, Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, E. E., Krewski, D., & Godleski, J. J. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. *Circulation*, *109*(1), 71-77. doi:10.1161/01.CIR.0000108927.80044.7F
- Pope Iii, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA*, *287*(9), 1132-1141. doi:10.1001/jama.287.9.1132
- Puett, R. C., Hart, J. E., Yanosky, J. D., Paciorek, C., Schwartz, J., Suh, H., . . . Laden, F. (2009). Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Environ Health Perspect*, *117*(11), 1697-1701.
doi:10.1289/ehp.0900572

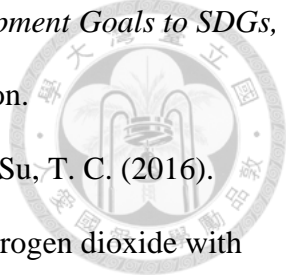


- 
- Qiao, W., Tian, W., Tian, Y., Yang, Q., Wang, Y., & Zhang, J. (2019). The Forecasting of PM2.5 Using a Hybrid Model Based on Wavelet Transform and an Improved Deep Learning Algorithm. *IEEE Access*, 7, 142814-142825. doi:10.1109/access.2019.2944755
- Qiu, Y. N., Wang, G. H., Zhou, F., Hao, J. J., Tian, L., Guan, L. F., . . . Zhang, K. Z. (2019). PM2.5 induces liver fibrosis via triggering ROS-mediated mitophagy. *Ecotoxicol Environ Saf*, 167, 178-187. doi:10.1016/j.ecoenv.2018.08.050
- Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Applied Sciences*, 8(12). doi:10.3390/app8122570
- Shih, D. H., To, T. H., Nguyen, L. S. P., Wu, T. W., & You, W. T. (2021). Design of a Spark Big Data Framework for PM2.5 Air Pollution Forecasting. *Int J Environ Res Public Health*, 18(13). doi:10.3390/ijerph18137087
- Shou, Y., Huang, Y., Zhu, X., Liu, C., Hu, Y., & Wang, H. (2019). A review of the possible associations between ambient PM2.5 exposures and the development of Alzheimer's disease. *Ecotoxicology and Environmental Safety*, 174, 344-352. doi:10.1016/j.ecoenv.2019.02.086
- Simoni, M., Baldacci, S., Maio, S., Cerrai, S., Sarno, G., & Viegi, G. J. J. o. T. D. (2015). Adverse effects of outdoor pollution in the elderly. *2015*, 7(1), 34-45.
- Soh, P.-W., Chang, J.-W., & Huang, J.-W. (2018). Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations. *IEEE Access*, 6, 38186-38199. doi:10.1109/access.2018.2849820

- Song, C., He, J., Wu, L., Jin, T., Chen, X., Li, R., . . . Mao, H. (2017). Health burden attributable to ambient PM_{2.5} in China. *Environ Pollut*, 223, 575-586.
doi:10.1016/j.envpol.2017.01.060
- Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., De' Donato, F., . . . Schwartz, J. (2019). Estimation of daily PM₁₀ and PM_{2.5} concentrations in Italy, 2013-2015, using a spatiotemporal land-use random-forest model. *Environ Int*, 124, 170-179.
doi:10.1016/j.envint.2019.01.016
- Stafoggia, M., Cesaroni, G., Peters, A., Andersen, Z. J., Badaloni, C., Beelen, R., . . . Forastiere, F. (2014). Long-term exposure to ambient air pollution and incidence of cerebrovascular events: results from 11 European cohorts within the ESCAPE project. *Environ Health Perspect*, 122(9), 919-925. doi:10.1289/ehp.1307301
- Su, F., Xu, Y., & Tang, X. (2017, 25-27 Oct. 2017). *Short-and mid-term load forecasting using machine learning models*. Paper presented at the 2017 China International Electrical and Energy Conference (CIEEC).
- Sun, Q., Wang, A., Jin, X., Natanzon, A., Duquaine, D., Brook, R. D., . . . Rajagopalan, S. (2005). Long-term Air Pollution Exposure and Acceleration of Atherosclerosis and Vascular Inflammation in an Animal Model. *JAMA*, 294(23), 3003-3010.
doi:10.1001/jama.294.23.3003
- Tsai, Y.-T., Zeng, Y.-R., & Chang, Y.-S. (2018). *Air Pollution Forecasting Using RNN with LSTM*. Paper presented at the 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech).



- 
- Wang, J., & Song, G. (2018). A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction. *Neurocomputing*, 314, 198-206. doi:10.1016/j.neucom.2018.06.049
- Wang, L., Bi, J., Meng, X., Geng, G., Huang, K., Li, J., . . . Liu, Y. (2020a). Satellite-based assessment of the long-term efficacy of PM2.5 pollution control policies across the Taiwan Strait. *Remote Sensing of Environment*, 251, 112067.
doi:<https://doi.org/10.1016/j.rse.2020.112067>
- Wang, L., Bi, J., Meng, X., Geng, G., Huang, K., Li, J., . . . Liu, Y. (2020b). Satellite-based assessment of the long-term efficacy of PM2.5 pollution control policies across the Taiwan Strait. *Remote Sensing of Environment*, 251. doi:10.1016/j.rse.2020.112067
- Wang, Y., Sun, M., Yang, X., & Yuan, X. (2016). Public awareness and willingness to pay for tackling smog pollution in China: a case study. *Journal of Cleaner Production*, 112, 1627-1634. doi:10.1016/j.jclepro.2015.04.135
- Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., & Cribb, M. (2019). Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. *Remote Sensing of Environment*, 231. doi:10.1016/j.rse.2019.111221
- Williams, A. M., Phaneuf, D. J., Barrett, M. A., & Su, J. G. (2019). Short-term impact of PM2.5 on contemporaneous asthma medication use: Behavior and the value of pollution reductions. *Proc Natl Acad Sci U S A*, 116(12), 5246-5253.
doi:10.1073/pnas.1805647115
- Witten, I. H., Frank, E., & Hall, M. A. (2011). Chapter 4 - Algorithms: The Basic Methods. In I. H. Witten, E. Frank, & M. A. Hall (Eds.), *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)* (pp. 85-145). Boston: Morgan Kaufmann.

- 
- World Health, O. (2015). *Health in 2015: from MDGs, Millennium Development Goals to SDGs, Sustainable Development Goals*. Geneva: World Health Organization.
- Wu, C. F., Shen, F. H., Li, Y. R., Tsao, T. M., Tsai, M. J., Chen, C. C., . . . Su, T. C. (2016). Association of short-term exposure to fine particulate matter and nitrogen dioxide with acute cardiovascular effects. *Sci Total Environ*, 569-570, 300-305.
doi:10.1016/j.scitotenv.2016.06.084
- Wu, S., Deng, F., Huang, J., Wang, H., Shima, M., Wang, X., . . . Guo, X. (2013). Blood Pressure Changes and Chemical Constituents of Particulate Air Pollution: Results from the Healthy Volunteer Natural Relocation (HVNR) Study. *Environmental Health Perspectives*, 121(1), 66-72. doi:10.1289/ehp.1104812
- Xu, L., Zhang, Z., Zhang, Q., & Li, P. (2016). Mycotoxin Determination in Foods Using Advanced Sensors Based on Antibodies or Aptamers. *Toxins (Basel)*, 8(8).
doi:10.3390/toxins8080239
- Yamazaki, S., Shima, M., Ando, M., Nitta, H., Watanabe, H., & Nishimuta, T. (2011). Effect of hourly concentration of particulate matter on peak expiratory flow in hospitalized children: a panel study. *Environ Health*, 10, 15. doi:10.1186/1476-069X-10-15
- Yang, G., Huang, J., & Li, X. (2018). Mining sequential patterns of PM_{2.5} pollution in three zones in China. *Journal of Cleaner Production*, 170, 388-398.
doi:10.1016/j.jclepro.2017.09.162
- Yang, Y., Guo, Y., Qian, Z. M., Ruan, Z., Zheng, Y., Woodward, A., . . . Lin, H. (2018). Ambient fine particulate pollution associated with diabetes mellitus among the elderly aged 50 years and older in China. *Environ Pollut*, 243(Pt B), 815-823.
doi:10.1016/j.envpol.2018.09.056


- 
- Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. (2018). *Deep Distributed Fusion Network for Air Quality Prediction*. Paper presented at the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.
- Yu, H., & Stuart, A. L. (2017). Impacts of compact growth and electric vehicles on future air quality and urban exposures may be mixed. *Sci Total Environ*, 576, 148-158.
doi:10.1016/j.scitotenv.2016.10.079
- Žabkar, R., Honzak, L., Skok, G., Forkel, R., Rakovec, J., Ceglar, A., & Žagar, N. (2015). Evaluation of the high resolution WRF-Chem (v3.4.1) air quality forecast and its comparison with statistical ozone predictions. *Geoscientific Model Development*, 8(7), 2119-2137. doi:10.5194/gmd-8-2119-2015
- Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., . . . Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM_{2.5} concentrations across China using a spatially explicit machine learning algorithm. *Atmospheric Environment*, 155, 129-139.
doi:10.1016/j.atmosenv.2017.02.023
- Zhao, H., Zhang, J., Wang, K., Bai, Z., & Liu, A. (2010). A GA-ANN model for air quality predicting. *ICS 2010 - International Computer Symposium*.
doi:10.1109/COMPSYM.2010.5685425
- Zhou, Q., Jiang, H., Wang, J., & Zhou, J. (2014). A hybrid model for PM_{2.5} forecasting based on ensemble empirical mode decomposition and a general regression neural network. *Sci Total Environ*, 496, 264-274. doi:10.1016/j.scitotenv.2014.07.051
- Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., Wang, Y. S., & Kang, C. C. (2019). Multi-output support vector machine for regional multi-step-ahead PM_{2.5} forecasting. *Sci Total Environ*, 651(Pt 1), 230-240. doi:10.1016/j.scitotenv.2018.09.111

Table 1. Air quality and meteorological data used in this study

#	Item	Source	Units
Input parameters			
1	SO ₂	EPA	ppb
2	CO	EPA	ppm
3	O ₃	EPA	ppb
4	PM ₁₀	EPA	µg/m ³
5	PM _{2.5}	EPA	µg/m ³
6	NO _x	EPA	ppb
7	NO	EPA	ppb
8	NO ₂	EPA	ppb
9	Ambient Temperature (AMB_TEMP)	EPA	°C
10	RAINFALL	EPA	mm
11	Relative humidity (RH)	EPA	%
12	WIND_SPEED (WS, instantaneous value)	EPA	m/sec
13	WIND_DIREC (WD, instantaneous value)	EPA	degress
14	WS_HR (hourly average)	EPA	m/sec
15	WD_HR (hourly average)	EPA	degress
16	Station Pressure (StnPres)	CWB	hPa
Output parameters			
17	Next three hour PM _{2.5} concentration	EPA	µg/m ³

Table 2. Missing value of 15 pollutants in 2018 and 2019

pollutants	2018			2019		
	missing	total	percentages (%)	missing	total	percentages (%)
station	0	113880	0	0	113880	0
AMB_TEMP	584	113296	0.005	1337	112543	1.188
CO	1198	112682	0.011	2071	111809	1.852
NO	2339	111541	0.021	3926	109954	3.571
NO2	2339	111541	0.021	3642	110238	3.304
NOx	2339	111541	0.021	3637	110243	3.299
O3	1292	112588	0.011	2464	111416	2.212
PM10	2113	111767	0.019	3536	110344	3.205
PM2.5	5133	108747	0.047	2475	111405	2.222
RAINFALL	866	113014	0.008	1394	112486	1.239
RH	602	113278	0.005	1217	112663	1.080
SO2	1858	112022	0.017	3208	110672	2.899
wind_x	616	113264	0.005	1172	112708	1.040
wind_y	616	113264	0.005	1172	112708	1.040
wind_x_HR	625	113255	0.006	1076	112804	0.954
wind_y_HR	625	113255	0.006	1076	112804	0.954

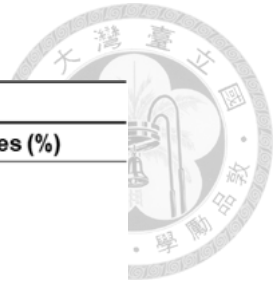



Table 3. PM_{2.5} value of 13 stations in 2018 and 2019




Abbr.	Station	Count	Type	Mean ($\mu\text{g}/\text{m}^3$)	Maximum ($\mu\text{g}/\text{m}^4$)	Standard deviation ($\mu\text{g}/\text{m}^5$)
ZS	Zhongshan	17520	Normal	15.15	83.00	9.74
GT	Guting	15131	Normal	14.10	95.00	9.40
TC	Tucheng	17520	Normal	15.81	100.00	10.59
SL	Shilin	17520	Normal	14.06	109.00	10.55
XD	Xindian	17520	Normal	13.47	88.00	9.45
XZ	Xinzhuang	17520	Normal	14.21	77.00	9.48
SS	Songshan	17520	Normal	14.64	89.00	9.33
BQ	Banqiao	17520	Normal	15.85	85.00	10.18
LK	Linkou	16314	Normal	12.87	72.00	9.88
XZ	Xizhi	17373	Normal	13.84	103.00	9.69
CL	Cailiao	17520	Normal	14.55	78.00	9.44
WH	Wanhua	17520	Normal	14.16	85.00	10.03
WL	Wanli	17179	Normal	13.93	90.00	7.65

Table 4. PM_{2.5} value of 13 stations in 2018



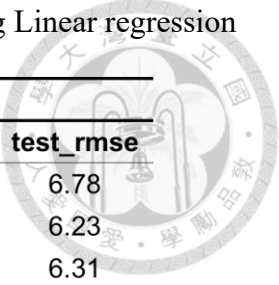
Abbr.	station	count	Type	Mean ($\mu\text{g}/\text{m}^3$)	Maximum ($\mu\text{g}/\text{m}^4$)	Standard deviation ($\mu\text{g}/\text{m}^5$)
ZS	Zhongshan	8760	Normal	14.43	83.00	9.17
GT	Guting	8555	Normal	14.14	95.00	10.24
TC	Tucheng	8760	Normal	14.52	80.00	10.19
SL	Shilin	8760	Normal	12.81	71.00	9.71
XD	Xindian	8760	Normal	12.01	75.00	8.70
XZ	Xinzhuang	8760	Normal	13.70	75.00	9.00
SS	Songshan	8760	Normal	14.43	89.00	8.92
BQ	Banqiao	8760	Normal	14.85	77.00	9.36
LK	Linkou	8760	Normal	11.30	72.00	8.94
XZ	Xizhi	8613	Normal	13.63	82.00	8.52
CL	Cailiao	8760	Normal	13.55	78.00	9.14
WH	Wanhua	8760	Normal	13.21	85.00	9.26
WL	Wanli	8419	Normal	12.94	90.00	7.44

Table 5. PM_{2.5} value of 13 stations in 2019



Abbr.	station	count	Type	Mean ($\mu\text{g}/\text{m}^3$)	Maximum ($\mu\text{g}/\text{m}^4$)	Standard deviation ($\mu\text{g}/\text{m}^5$)
ZS	Zhongshan	8760	Normal	15.88	73.00	10.23
GT	Guting	6576	Normal	14.04	63.00	8.18
TC	Tucheng	8760	Normal	17.09	100.00	10.83
SL	Shilin	8760	Normal	15.31	109.00	11.20
XD	Xindian	8760	Normal	14.93	88.00	9.94
XZ	Xinzhuang	8760	Normal	14.72	77.00	9.91
SS	Songshan	8760	Normal	14.85	84.00	9.72
BQ	Banqiao	8760	Normal	16.85	85.00	10.86
LK	Linkou	7554	Normal	14.68	69.00	10.58
XZ	Xizhi	8760	Normal	14.04	103.00	10.72
CL	Cailliao	8760	Normal	15.55	77.00	9.63
WH	Wanhua	8760	Normal	15.11	76.00	10.67
WL	Wanli	8760	Normal	14.89	58.00	7.73

Table 6. Main results of R2, MAE, and RMSE at all stations for PM_{2.5} using Linear regression



		2018			2019		
	variables	train_r2	train_mae	train_rmse	test_r2	test_mae	test_rmse
Tucheng	25	0.62	4.84	6.68	0.56	4.86	6.78
Cailiao	28	0.62	4.27	5.96	0.53	4.55	6.23
Xinzhuang	30	0.63	4.38	6.07	0.51	4.56	6.31
Wanli	23	0.58	3.67	5.03	0.50	3.79	5.27
Banqiao	23	0.62	4.84	6.70	0.49	4.79	6.66
Linkou	27	0.61	4.66	6.62	0.47	4.78	6.52
Guting	24	0.47	4.49	5.95	0.46	5.06	7.51
Xizhi	28	0.68	4.34	6.03	0.46	4.61	6.26
Xindian	29	0.57	4.80	6.53	0.45	4.69	6.43
Zhongshan	31	0.49	5.52	7.33	0.43	5.14	6.94
Songshan	35	0.53	5.12	6.70	0.42	4.99	6.82
Wanhua	27	0.49	5.73	7.64	0.41	5.25	7.09
Shilin	26	0.44	6.34	8.40	0.37	5.89	7.73

Table 7. Main results of R2, MAE, and RMSE at all stations for PM_{2.5} using XGBoost

	2018			2019		
	train_r2	train_mae	train_rmse	test_r2	test_mae	test_rmse
Tucheng	0.74	4.09	5.55	0.62	4.42	6.26
Cailiao	0.74	3.58	4.91	0.61	4.01	5.73
Xinzhuang	0.72	3.80	5.23	0.58	4.05	5.80
Banqiao	0.83	3.39	4.50	0.57	4.29	6.12
Linkou	0.72	3.92	5.60	0.56	4.28	5.94
Wanli	0.67	3.22	4.44	0.55	3.54	5.01
Xizhi	0.80	3.49	4.85	0.54	4.17	5.74
Zhongshan	0.70	4.33	5.64	0.51	4.75	6.43
Xindian	0.71	4.00	5.38	0.50	4.41	6.15
Guting	0.65	3.67	4.86	0.50	4.98	7.25
Songshan	0.63	4.57	5.94	0.48	4.67	6.42
Wanhua	0.57	5.28	6.98	0.48	4.91	6.68
Shilin	0.60	5.46	7.11	0.42	5.59	7.37

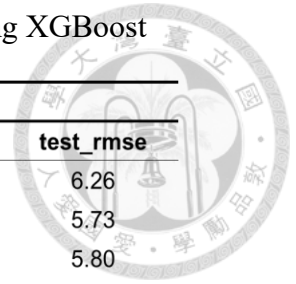


Table 8. Main results of R2, MAE, and RMSE at all stations for PM_{2.5} using LSTM

	2018			2019		
	train_r2	train_mae	tain_rmse	test_r2	test_mae	test_rmse
Tucheng	0.71	4.20	5.81	0.60	4.55	6.48
Cailiao	0.72	3.68	5.10	0.55	4.28	6.11
Xinzhuang	0.70	3.86	5.39	0.55	4.29	6.04
Wanli	0.64	3.33	4.62	0.53	3.53	5.09
Banqiao	0.71	4.19	5.86	0.53	4.48	6.41
Xizhi	0.78	3.59	5.07	0.52	4.20	5.87
Linkou	0.69	4.09	5.92	0.50	4.48	6.30
Xindian	0.63	4.36	6.01	0.49	4.46	6.23
Zhongshan	0.57	5.03	6.70	0.47	4.92	6.65
Guting	0.55	4.13	5.52	0.47	5.04	7.48
Wanhua	0.55	5.28	7.13	0.46	4.91	6.79
Songshan	0.59	4.76	6.26	0.45	4.81	6.63
Shilin	0.50	5.92	7.92	0.40	5.66	7.55

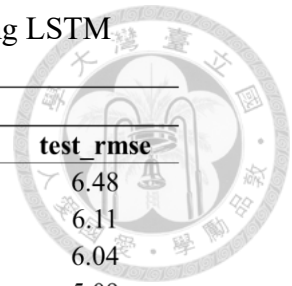


Table 9. Pearson correlation of PM_{2.5} concentrations with 3 hour-lag among 13 stations

**2010-2019 three hour lag Pearson correlation of PM_{2.5} concentrations among Taipei area 13 stations
for each station when t=k (Predicting Objective)**

	Linkou	Xinzhuang	Tucheng	Banqiao	Cailiao	Wanhua	Shilin	Zhonshang	Guting	Xindian	Songshuan	Xizhi	Wanli	
Other station when t=k-3	0.67	0.64	0.67	0.66	0.65	0.71	0.69	0.66	0.68	0.74	0.72	0.75	Wanli	Wanli
	0.68	0.72	0.72	0.72	0.72	0.75	Shilin	0.74	0.71	0.73	0.74	0.70	0.64	Shilin
	Linkou	0.78	0.77	0.76	0.77	0.74	0.74	0.70	0.68	0.72	0.71	0.73	0.67	Linkou
	0.75	0.80	0.79	0.79	Cailiao	0.78	0.76	0.77	0.73	0.73	0.76	0.75	0.63	Cailiao
	0.69	0.71	0.74	0.74	0.72	0.75	0.71	0.73	0.72	0.76	0.76	Xizhi	0.70	Xizhi
	0.63	0.72	0.71	0.72	0.71	0.76	0.72	Zhonshang	0.73	0.71	0.77	0.69	0.59	Zhonshang
	0.63	0.70	0.71	0.71	0.70	0.76	0.71	0.78	0.73	0.74	Songshuan	0.72	0.63	Songshuan
	0.67	0.72	0.73	0.74	0.73	Wanhua	0.73	0.77	0.73	0.74	0.76	0.71	0.62	Wanhua
	0.75	Xinzhuang	0.81	0.82	0.80	0.78	0.76	0.77	0.72	0.73	0.76	0.73	0.62	Xinzhuang
	0.62	0.68	0.70	0.70	0.68	0.74	0.68	0.73	Guting	0.72	0.73	0.68	0.59	Guting
	0.73	0.82	0.83	Banqiao	0.79	0.79	0.75	0.77	0.76	0.76	0.76	0.75	0.63	Banqiao
	0.73	0.80	Tucheng	0.82	0.77	0.77	0.74	0.75	0.73	0.76	0.75	0.74	0.63	Tucheng
	0.66	0.70	0.74	0.73	0.70	0.76	0.71	0.73	0.74	Xindian	0.75	0.73	0.66	Xindian

Table 10. Pearson correlation of PM_{2.5} concentrations with 3 hour-lag among 13 stations

	2019 (original)			2019 (nearby)		
	r2	mae	rmse	r2	mae	rmse
Tucheng	0.62	4.42	6.26	0.64	4.48	6.16
Cailiao	0.61	4.01	5.73	0.63	3.90	5.46
Xinzhuang	0.58	4.05	5.80	0.61	4.10	5.69
Wanhua	0.48	4.91	6.68	0.55	4.58	6.22
Songshan	0.48	4.67	6.42	0.51	4.58	6.24
Shilin	0.42	5.59	7.37	0.45	5.42	7.22
avg	0.53	4.61	6.38	0.56	4.51	6.17

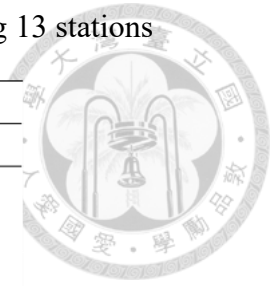


Table 11. Prediction outcomes while training with different length of historical data

	2019		
	test_r2	test_mae	test_rmse
1 years 24hr	0.621	4.434	6.277
5 years 8 hr	0.634	4.337	6.170
1 years 8hr	0.622	4.425	6.265



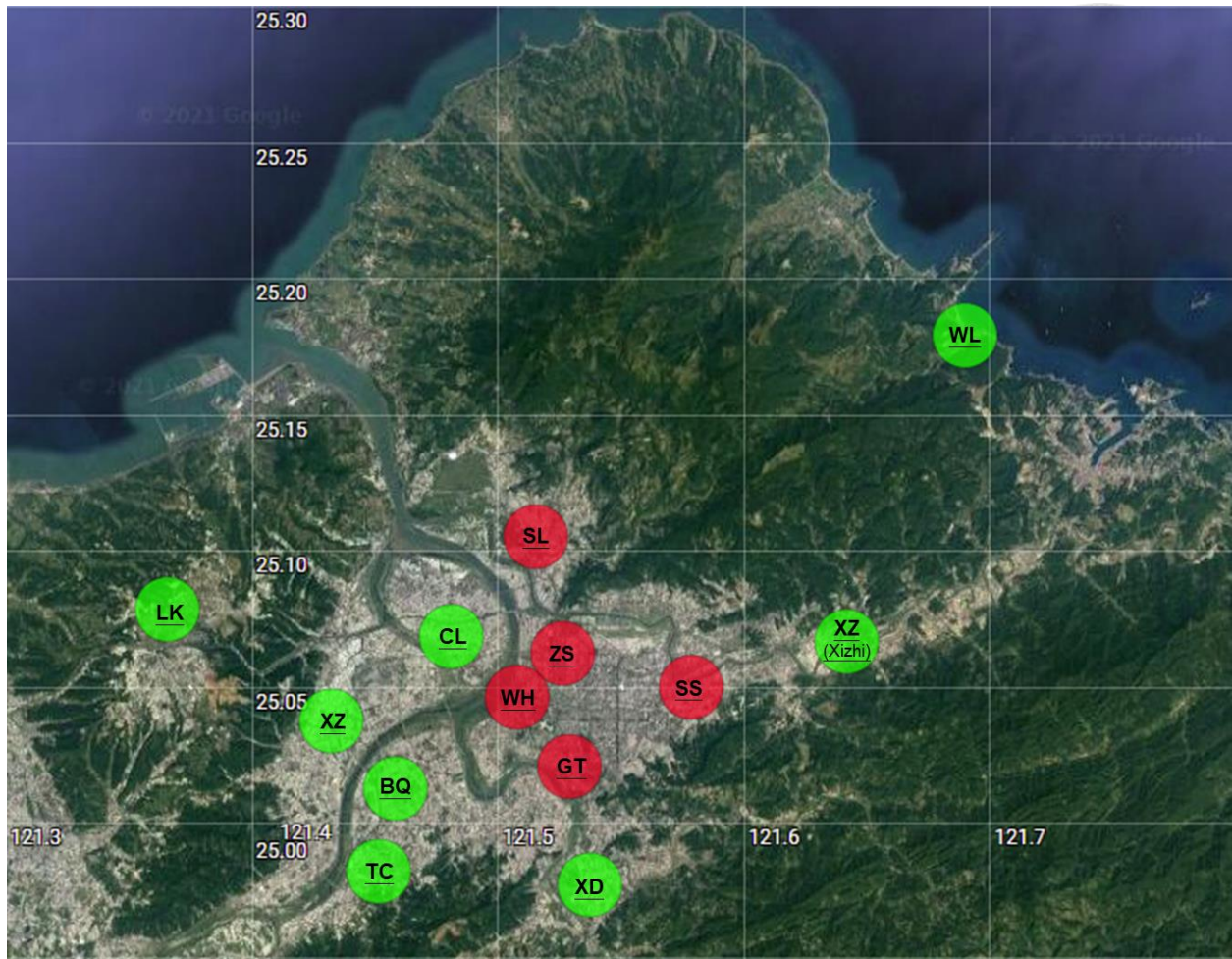


Figure 1. The distribution of air quality monitoring stations in Taipei Areas established by the Taiwan Environmental Protection Administration (EPA)

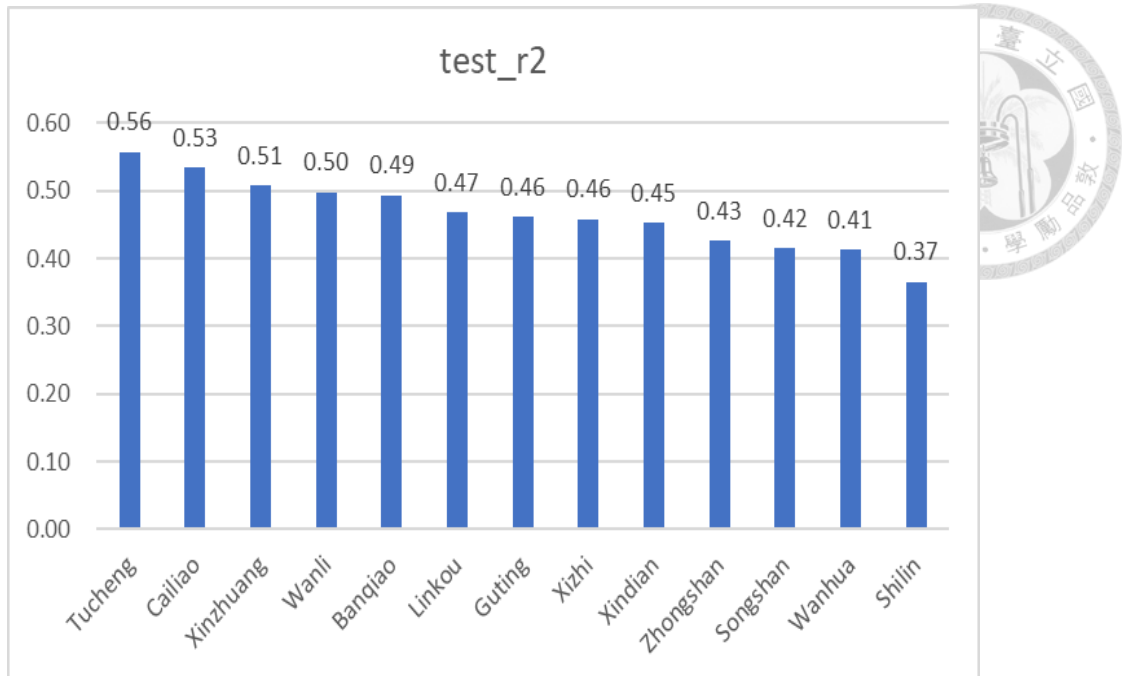


Figure 2. Linear Regression prediction outcomes in 2019

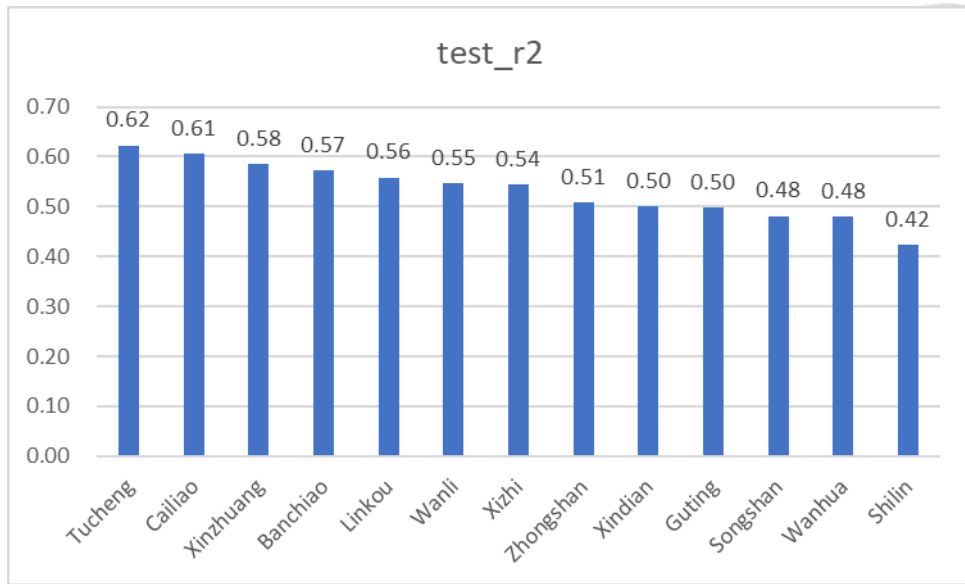


Figure 3. XGBoost prediction outcomes in 2019

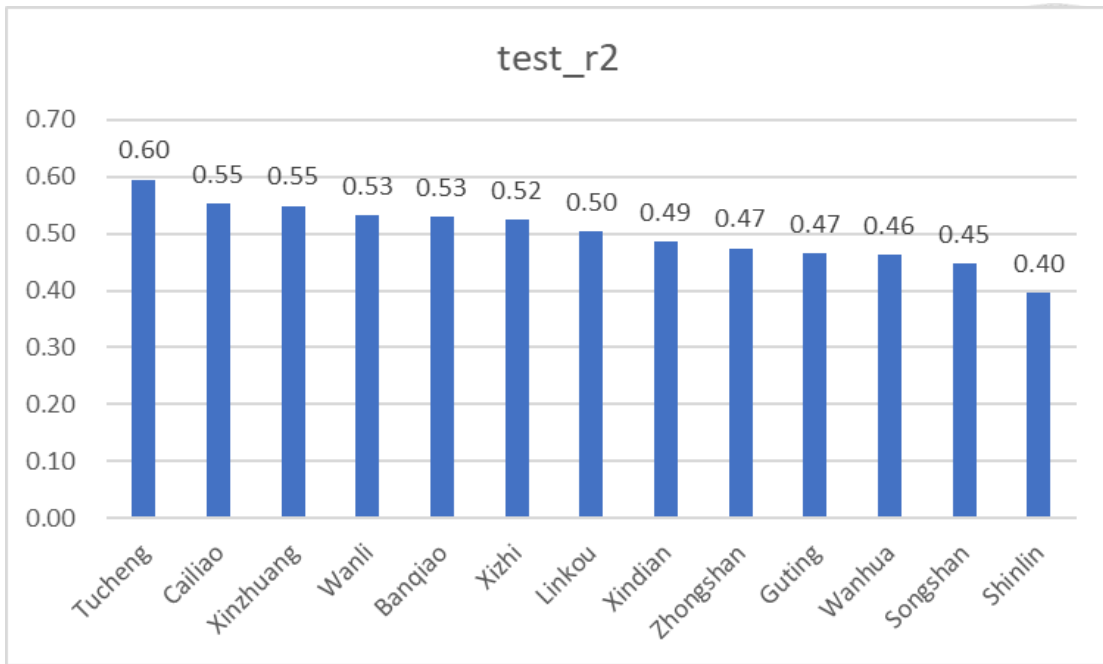


Figure 4. LSTM prediction outcomes in 2019

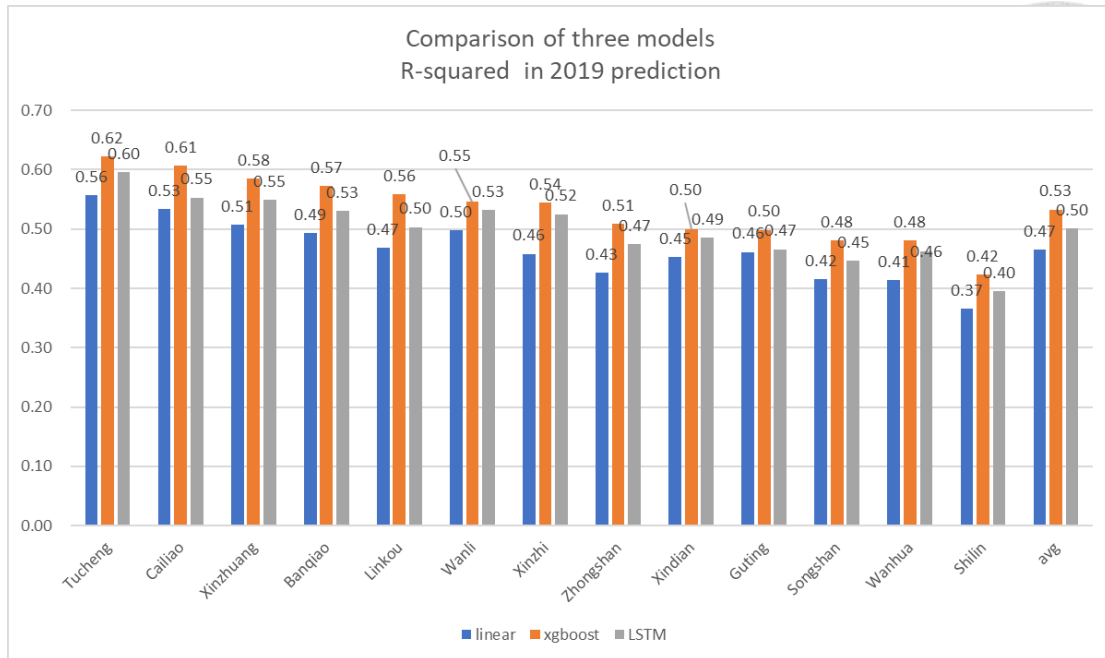


Figure 5. The comparison of linear regression, XGBoost, and LSTM in prediction performance among all stations when the meteorological and pollutants data are only input.

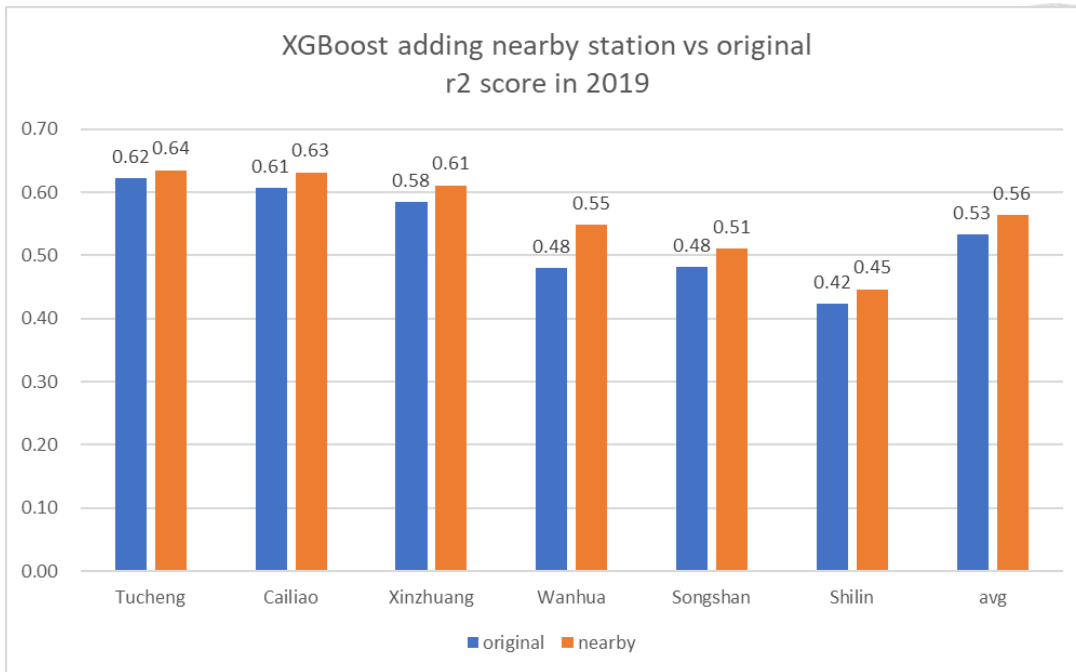


Figure 6. The comparison of XGBoost prediction performance among six stations when adding pollutants data from nearby stations.

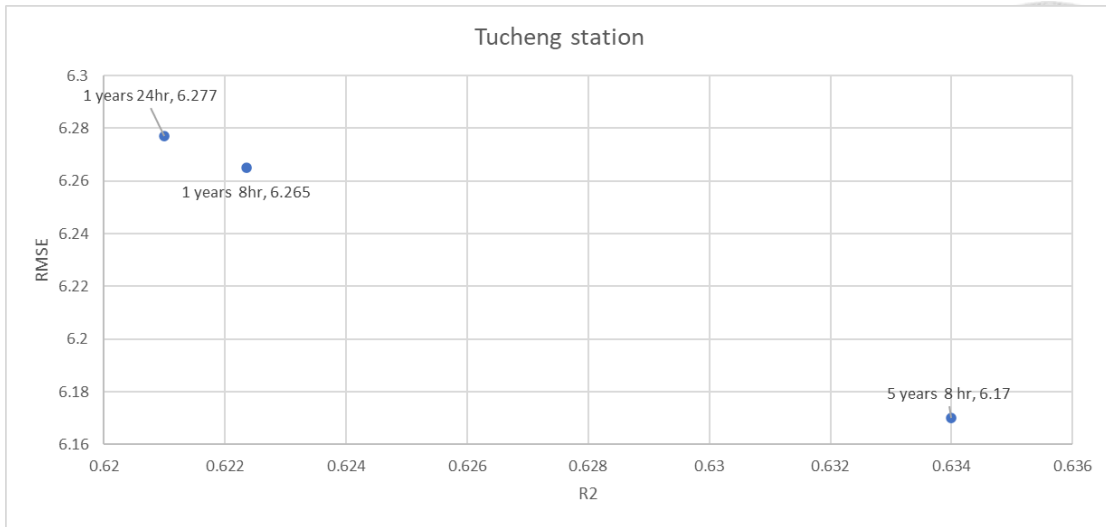


Figure 7. The comparison of XGBoost prediction performance in RMSE and R2 at Tucheng station.

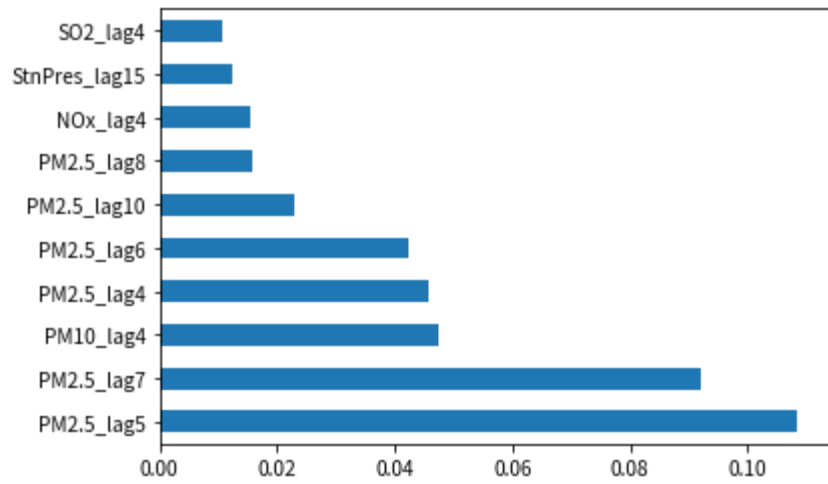


Figure 8. The feature importance of the XGBoost model in Tucheng using 24 hours in 2018 for training.

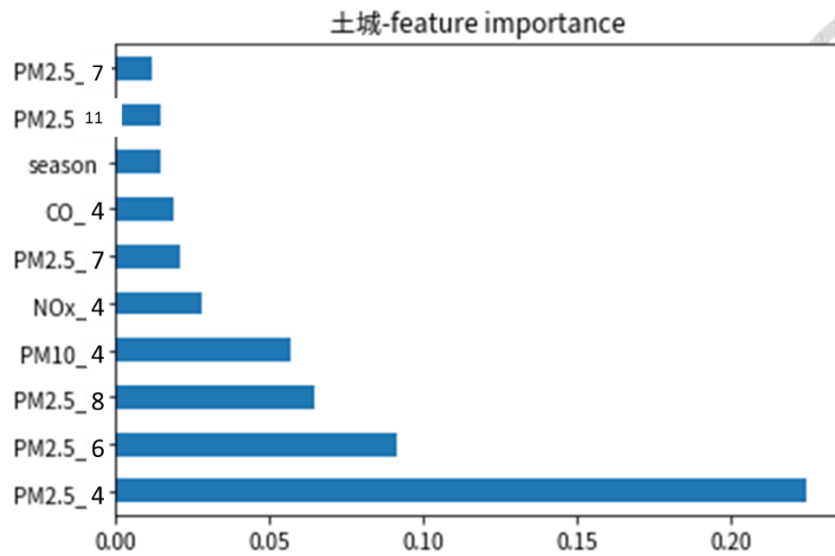


Figure 9. The feature importance of the XGBoost model in Tucheng using eight hours in 2014-2018 for training.