Machine Learning Models for Indoor $PM_{2.5}$ Concentrations in Residential Architecture in Taiwan

Yu Chen, Lin [a], *Yaw Shyan Tsay* [b], *Chien Chen Jung* [c]

[a] Department of Architecture, National Cheng Kung University, Tainan, Taiwan, e74036085@gs.ncku.edu.tw

[b] Department of Architecture, National Cheng Kung University, Tainan, Taiwan, tsayys@mail.ncku.edu.tw

[c] Department of Public Health Assistant, China Medical University, Taichung, Taiwan, ccjung@mail.cmu.edu.tw

Abstract. People typically spend 80-90% of their time indoors. Therefore, establishing prediction models to estimate particulate matter ($PM_{2.5}$) concentration in indoor environments is of great importance, especially in residential households, in order to allow for accurate assessments of exposure in epidemiological studies. However, installing monitoring instruments to collect indoor $PM_{2.5}$ data is both labor and budget-intensive. Therefore, indoor $PM_{2.5}$ concentration prediction models have become critical issues. This study aimed to develop a predictive model for hourly household $PM_{2.5}$ concentration based on the artificial neural network (ANN) method. From January 2019 to April 2020, $PM_{2.5}$ concentration and related parameters (e.g., occupants' behavior information and ventilation settings) were collected in a total of 62 houses and apartments in Tainan, Taiwan (tropical and subtropical region). Overall, 2136 pairs of data and 9 possible variables were used to establish the model. Meteorological data were primarily used to establish the model. Meanwhile, occupants' behavior and building characteristics were generalized as effective opening areas to describe the importance of ventilation in subtropical areas. We performed five-fold cross-validation to assess prediction model performance. The prediction model achieved promising predictive accuracy, with a coefficient of determination (R^2) value of 0.88 and a root mean square error value of 3.35 ($\mu g/m^3$), respectively. Outdoor $PM_{2.5}$ concentrations were the most important predictor variable, followed in descending order by temperature, outdoor carbon dioxide concentration, outdoor relative humidity, and opening effective areas. In summary, we developed a prediction model of hourly indoor $PM_{2.5}$ concentrations and suggest that outdoor meteorological data, building characteristics, and human behavior can be powerful predictors. The results also confirm that the model can be used to predict indoor $PM_{2.5}$ concentrations across seasons.

Keywords: Artificial neural network, Indoor air quality, Subtropical
DOI: https://doi.org/10.34641/clima.2022.247

1. Introduction

Fine particle matter (PM) has been recognized as a key air pollutant that influences occupants' health [1]. Recently, the monitoring methods for $PM_{2.5}$ have primarily been through direct field measurement, which is labor-intensive and costly. Moreover, this method is primarily suitable for outdoor measurements and is difficult to widely use in indoor environments. Since most people spend 85%-90% of their life indoors [2], assessing indoor particle pollution is important for understanding the impact of particle pollution and can avoid exposing occupants to harmful levels of pollution.

In addition to indoor sources like cooking and smoking, outdoor pollution particles are major contributor to indoor concentrations. A previous study observed that outdoor air has more critical influences on indoor air in tropical and subtropical areas than in cold and temperate zones [3]. Using ventilation systems and natural moving winds, outdoor particles can be transported indoors. Furthermore, infiltration from leaks in the building envelopes is another dominant source [4]. Consequently, indoor particle concentration fluctuates according to such mixed factors as climates, seasons, human behaviors, and building characteristics, thus demonstrating the importance of research on local data.

As mentioned, the spatiotemporal distribution of indoor $PM_{2.5}$ concentrations involves complex natural and anthropogenic sources, thus making the

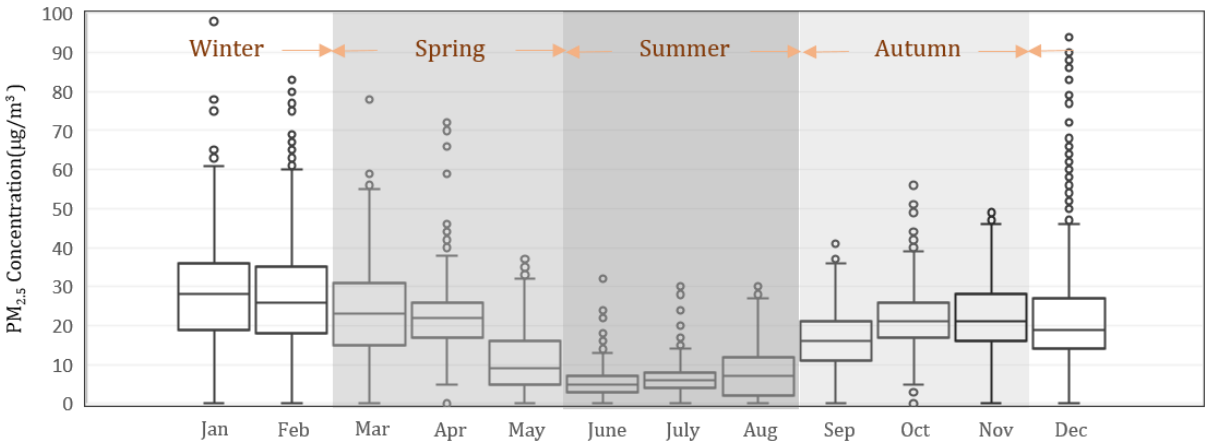

Fig. 1 - Historical plots of the outdoor PM2.5 concentrations at Tainan station (120'12'E; 22'59'N).

control and accurate prediction of $PM_{2.5}$ a challenging task.

In recent years, with the development of machine learning, such algorithms as random forest [5], gradient boosting regression [6], and artificial neural networks [7] have been used to overcome the above drawbacks of linear models with multiple parameters.

The purpose of this study is to explore the features of pollution sources and establish a prediction model for indoor concentration based on the long-term datasets collected from residential buildings in southern Taiwan. The results of this study can provide occupants with valuable information, including air quality control strategies and their risks to human health.

2. Dataset analysis

2.1 wide area data

The target city was Tainan, which is the sixth-largest city in Taiwan. Tainan is located between tropical and subtropical regions, and the annual temperature and relative humidity are 24.3 ℃ and 77.2%, respectively. **Fig. 1** shows the historical distribution of outdoor $PM_{2.5}$ concentrations from 2018 to 2020. Measurement $PM_{2.5}$ data were downloaded from the Tainan air quality monitoring network, which is an officially organized open-source website.

Fig. 1 illustrates the monthly temporal variation of the outdoor $PM_{2.5}$ concentrations. This figure indicates that the $PM_{2.5}$ concentration reached a peak during winter, dropped gradually through spring and summer, and started to increase again during autumn. The reason being that starting from September, the prevailing wind (northeast monsoon) affects the island, thereby easily dispersing and diluting air pollutants, as well as bringing pollutants from nearby China. Furthermore, the subtropical high pressure over the Pacific Ocean in the summer causes high convection of air mass, which leads to the vertical dispersion of pollutants [8].

2.2 survey objects

From Oct 2019 to Jan 2020, hourly data were collected from 62 houses in Tainan using real-time devices (TSI Q-TRAK and TSI DUST-TRAK) over a day (**Fig. 3**). Occupants' behavior and building characteristics were also recorded through questionnaires filled out by family members. Two out of three of the surveyed objects were located in the city center, and the rest of them sat in the nearby suburban An-nan district. The processes and specifications are shown in **Fig. 3**.

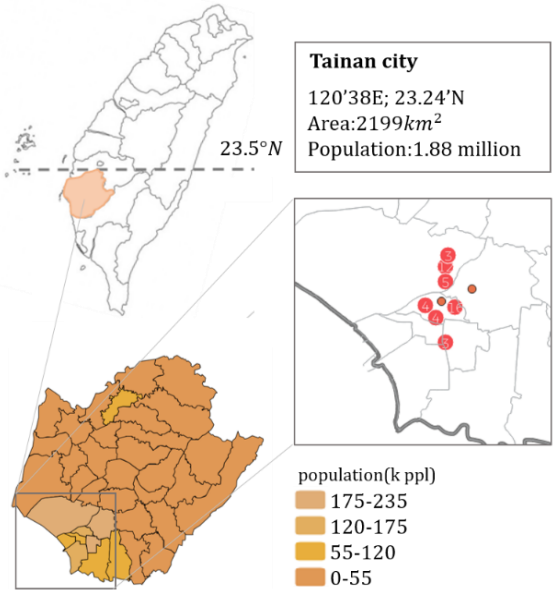

Fig. 2 - Location of the survey object

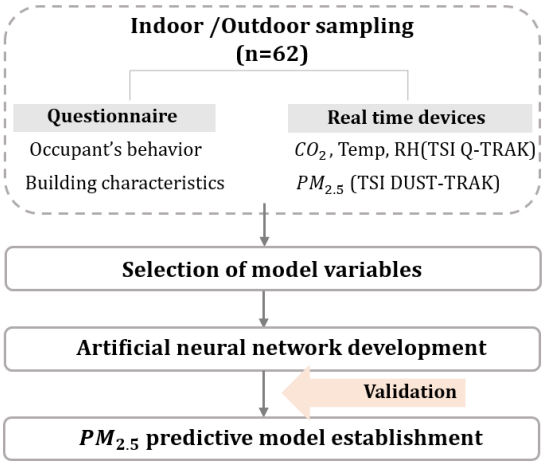

Fig. 3 - Research workflow

Tab. 1 - Summary of building characteristics and human activities of surveyed households

Variables	Description	Results(n,%)	Variables	Description	Results(%)
Building characteristics (n=62 houses)			human activities (t=2135 hrs)		
building types	Townhouse Apartment	42(67.7%) 20(32.3%)	electric devices(computer, printer)	YES	9%
Building age	<20 years 20-40 years >40 years	29(46.8%) 30(48.4%) 3(4.8%)	cooking(oven, microwave)	YES	0.7%
Building floor level	1-5 floors 6-10 floors >10 floors	62(83.9%) 6(9.7%) 4(6.5%)	intense burning	YES	10%
Number of openings	1 2 3	54(87.1%) 6(9.7%) 2(3.2%)	smoking	YES	0%
Near main road(<8m)	YES	12(19.4%)	Airpurifier	YES	15%

2.3 building characteristics and human activities

The vast majority (67.7%) of the objects were two to five-story houses, which is the most common residential type in southern Taiwan. Most buildings were 20-40 years old and were located away from the main road. All of the measured spaces were rooms with more than one opening.

As presented in **Tab. 1,** 9%, 0.7%, 0% and 10% of the time occupants indulged in using the computer, cooking, smoking and intense burning. Furthermore, air purifiers were turned on for more than 15% of the observed time.

2.4 correlation analysis of indoor and outdoor $PM_{2.5}$ concentration

I/O ratio simply describes the relationship between indoor and outdoor particle concentrations according to equation (1):

$$\text{I/O ratio} = \frac{C_{in}}{C_{out}} \quad (1)$$

Where C_{in} and C_{out} are the indoor and outdoor particle concentrations (μg/m^3), respectively. The summary of the measured objects shows a mean value of 0.8 and varies within the range of 0.14-2.37, as shown in **Tab. 2**. The low I/O ratio is strongly related to the presence of few indoor sources, window opening behavior, and the use of filtration systems, such as 0.71 for $PM_{2.5}$ with few indoor sources as also found by Clayton et al. [9].

2.5 data analysis

Among the observed objects, indoor pollution sources were seldom recorded, which indicates that meteorological parameters and building characteristics were the main factors to influence indoor $PM_{2.5}$ concentrations. Previous studies also pointed out that outdoor $PM_{2.5}$ was a major variable in $PM_{2.5}$ prediction models [10]. In particular, opening windows is a universal behavior for adapting to Taiwan's hot and humid climate [11]. Therefore, outdoor $PM_{2.5}$ concentration plays an important role in predicting indoor $PM_{2.5}$ concentrations.

3. Machine learning models

3.1 artificial neural networks

Artificial neural network (ANN) has been proven to be a powerful tool to deal with nonlinear and noisy environmental data when traditional models have failed to predict [12].

A typical ANN model contains interconnected input, hidden, and output layers. The information processed by the hidden layer is assigned weights and is then adjusted by an activation function that determines how they transmit data to the next layer. The output layer predicts a value and compares it with the ground truth to assess the error according to the loss function (generally by minimizing the error).

3.2 selection of variables

Outdoor $PM_{2.5}$ concentrations were matched to indoor $PM_{2.5}$ concentration and meteorological data (temperature, CO_2, and humidity) on an hourly basis. Since building types and stories generally do not differ, the information was excluded. Furthermore, household appliances were only activated a small fraction of time, so those variables were also discarded in the training datasets.

In contrast, window opening behavior was of great importance, as described in the previous section. Its influence because of natural ventilation was considered using the "opening effective area" parameter, which was defined as the product of area of the main openings and the occupants' behavior(True(1) or False(0)). Considering the

Tab. 2 - Descriptive statistics of the variables

n=1316	Symbol	Mean±STD	Min	Max	Median
Meteorological data					
Outdoor $PM_{2.5}$ concentration(μg/m^3)	C_{out}_$PM_{2.5}$	21.3±14.0	0.9	115.2	17.5
Outdoor CO_2 concentration(ppm)	C_{out}_CO_2	385.6±65.9	253..9	766.5	377.5
Outdoor temperature(℃)	$Temp_{out}$	26.0±5.1	15.1	44.0	26.0
Outdoor relative humidity(%)	RH_{out}	66.5±7.5	26.8	91.1	66.6
Indoor $PM_{2.5}$ concentration(μg/m^3)	C_{in}_$PM_{2.5}$	15.3±9.7	0.8	78.5	13.1
Indoor CO_2 concentration(ppm)	C_{in}_CO_2	490.5±189.3	308.0	1967.8	430
Indoor temperature(℃)	$Temp_{in}$	26.4±3.1	19.0	33.8	26.4
Indoor relative humidity(%)	RH_{in}	65.6±6.7	43.7	82.5	66.4
IO Ratio		0.8±1.2	0.14	2.37	0.7
Building characteristics					
Opening effective area(m^3)	A	0.9±1.7	0	13.1	0
Space volume	V	71.0±38.8	14.9	213.3	62.6

uncontrolled flow of air through cracks and leaks in the building envelope, the space volume was also included as a building information variable.

The statistics for all variables are described in **Tab. 2**. The measured indoor $PM_{2.5}$ concentrations were 15.3±9.7 μg/m^3 at 62 houses, and the outdoor $PM_{2.5}$ concentrations ranged between 0.9 and 115.2 μg/m^3, with a mean value of about 21.3 μg/m^3.

3.3 variables correlation

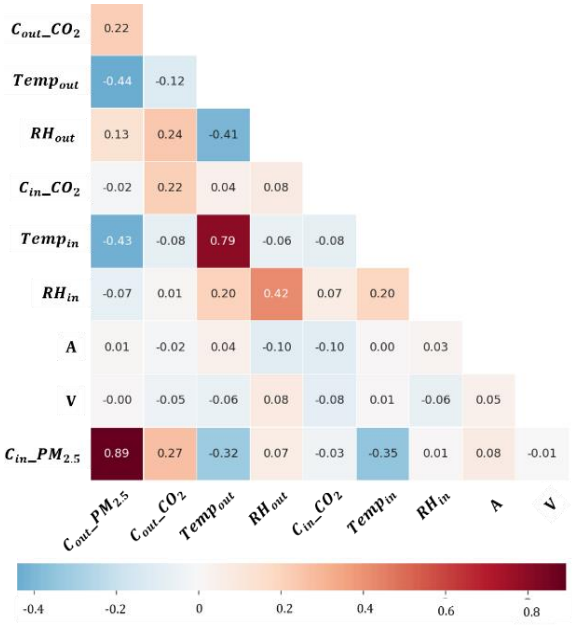

Fig. 4 - Variables correlation matrix

A correlation matrix (**Fig. 4**) illustrates the correlation between the indoor $PM_{2.5}$ and other variables using Pearson correlation coefficients. Of the nine outdoor predictor variables (including outdoor $PM_{2.5}$ concentration, CO_2 concentration, temperature, humidity, effective opening area, and space volume), all predictors were found to be correlated to indoor $PM_{2.5}$. The largest correlation coefficient was 0.89 between the indoor and outdoor $PM_{2.5}$ concentrations. The indoor $PM_{2.5}$ concentration showed a positive correlation with outdoor CO_2 concentrations (0.29), outdoor humidity (0.09), and effective opening area (0.08). Furthermore, a negative correlation was observed with the temperature factors, (-0.33) and (-0.35), respectively.

3.4 model development

This research adopted the Google Colaboratory platform with python language to generate the ANN model for prediction. First, the collected data was proven to be normalized before training to achieve better performance [13]. The missing pairs of data were then dropped to create a robust model. The total training sets were consequently composed of 1316 pairs of data and were then divided into training, validation, and testing datasets with a ratio of 6:2:2. The hyperparameters for the ANN training algorithms are shown in **Tab. 3**.

In total, nine variables served as potential input parameters, and the objective output was the indoor $PM_{2.5}$ concentrations. The structure of the proposed basic model was set up using nine neurons in the input layer for the nine input parameters and single neurons in the output layer for the output variables. The model has three hidden layers, and each layer has 30 neurons, which we tested and verified to be the best performing network.

Tab. 3 - Summary of the ANN model settings

Hyperparameter	Model
Neurons	[30,30,30]
Optimizer(learning rate)	Adam(0.001)
Batch size	20

4. Results

Prediction model performance was assessed using the coefficient of determination (R^2) and root mean square error (RMSE). We used five-fold cross-validation to confirm model reliability. The R^2 of the cross validation ranged between 0.85-0.90, with an average of 0.88.

Fig. 5 displays a plot of the test set target values against the measured values from the ANN-based regression model implemented in this work. The RMSE was about 3 (μg /m^3) in all validation cases, indicating significant prediction performance; since the measured indoor $PM_{2.5}$ ranged between 0.8 to 78.5 μg /m^3, the error was relatively small.

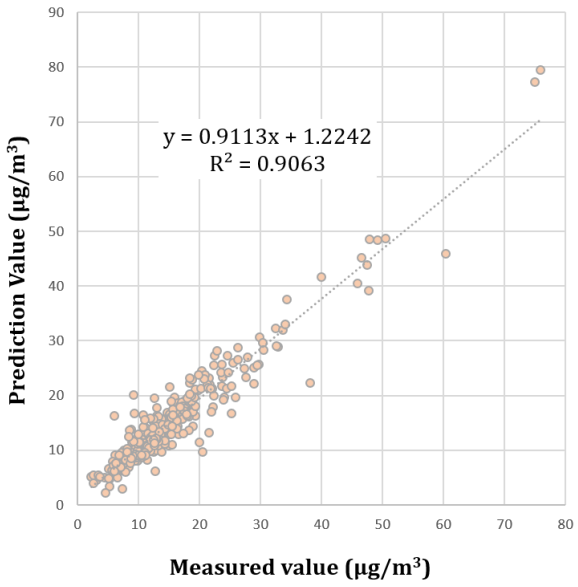

Fig. 5 - Relationship between predicted and measured hourly average indoor $PM_{2.5}$ concentration

5. Conclusions and discussions

In this study, we established a prediction model for hourly average indoor $PM_{2.5}$ concentrations conducting the ANN method. The collected data revealed that outdoor $PM_{2.5}$ was the key predictor, which was consistent with the findings of previous studies [3][5][10]. Furthermore, ventilation and building characteristics alongside human activities were comprehensively included by the effective opening area as a highly relative variable in the model. The results indicated that controlling the window opening behavior is beneficial for reducing indoor $PM_{2.5}$ concentrations. This prediction model can also be adopted in future research for automated ventilation systems in smart buildings.

In this study, about half of the buildings were aged 20-40 years, and over 80% of them were classified as two- to five-story townhouses. According to the official statistical data, they represent the typical building type and age distributions of the area [14]. These data reflected that the prediction model can be applied throughout Taiwan.

Studies in the past have indicated that indoor occasion human activities, especially smoking and incense stick burning, can be influential factors in $PM_{2.5}$ concentrations [15]. However, the above behavior was seldom recorded in our field measurements, which minimized their impact on indoor pollution. For further research, object selections and detailed records could be included and tested as variables to enhance the prediction model.

This Taiwan-based prediction model differs from those of other climate regions, where infiltration can be the primary pathway entering the building. Occupants who embrace natural ventilation should be warned about outdoor air quality levels. The established ANN model using outdoor environmental data and simple building factors has advantages over the classic multiple linear regression model with higher R^2 and lower RMSE [10], which can be used for rapid pollution regulation.

6.References

[1] Huang, Y. L., Chen, H. W., Han, B. C., Liu, C. W., Chuang, H. C., Lin, L. Y., & Chuang, K. J. Personal exposure to household particulate matter, household activities and heart rate variability among housewives. PloS one, 9(3), e89969.

[2] N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, A.M. Tsang, P. Switzer, J.V. Behar, S.C. Hern, W.H. Engelmann. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants, J. Expo. Sci. Environ. Epidemiol., 11 (3) (2001), p. 231

[3] Chien-Cheng Jung, Nai-Yun Hsu, Huey-Jen Su. Temporal and spatial variations in IAQ and its association with building characteristics and human activities in tropical and subtropical areas, Building and Environment, Volume 163,2019,106249.

[4] Chun Chen, Bin Zhao. Review of relationship between indoor and outdoor particles: I/O ratio, infiltration factor and penetration factor, Atmospheric Environment, Volume 45, Issue 2,2011,Pages 275-288.

[5] XuXu, D., Liu, Z., Li, Y., Li, N., Chartier, R., Chang, J., Wang, Q., Wu, Y., & Li, N.C. Estimating hourly average indoor PM2.5 using the random forest approach in two megacities, China. Building and Environment.

[6] Doreswamy, Harishkumar K S, Yogesh KM, Ibrahim Gad. Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models, Procedia Computer Science, Volume 171, 2020, Pages 2057-2066,

[7] Gholamreza GoudarziK. Hopke, Mohsen YazdaniPhilip. Forecasting PM2.5 concentration using artificial neural network and its health effects in Ahvaz, Iran, chemosphere.2021.131285.

[8] Lee, M., Lin, L., Chen, CY. et al. Forecasting Air Quality in Taiwan by Using Machine Learning. Sci Rep 10, 4153

[9] C A ClaytonL Perritt, E D Pellizzari, K W Thomas, R W Whitmore, L A Wallace, H Ozkaynak, J D SpenglerR. Particle Total Exposure Assessment Methodology (PTEAM) study: distributions of aerosol and elemental concentrations in personal, indoor, and outdoor air samples in a southern California community. J Expo Anal Environ Epidemiol. 1993 Apr-Jun;3(2): 227-250

[10] Jung, C.-C.; Lin, W.-Y.; Hsu, N.-Y.; Wu, C.-D.; Chang, H.-T.; Su, H.-J. Development of Hourly Indoor PM2.5 Concentration Prediction Model: The Role of Outdoor Air, Ventilation, Building Characteristic, and Human Activity. Int. J. Environ. Res. Public Health 2020, 17, 5906.

[11] Lu, H.-Y.; Lin, S.-L.; Mwangi, J.K.; Wang, L.-C.; Lin, H.-Y. Characteristics and source apportionment of atmospheric PM2.5 at a coastal city in southern Taiwan. Aerosol Air Qual. Res. 2016, 16, 1022–1034.

[12] Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., ... & Cawley, G. Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment, 37(32), 4539-4550.

[13] JoJun-Mo. (2019). Effectiveness of Normalization Pre-Processing of Big Data. Journal of the KIECS: 547-552.

[14] Interior of the Ministry. (2018). Statistical Report of Ministry of the Interior. Taipei, Taiwan.

[15] Gurung, G.; Bradley, J.; Delgado-Saborit, J.M. Effects of shisha smoking on carbon monoxide and PM2.5 concentrations in the indoor and outdoor microenvironment of shisha premises. Sci. Total Environ. 2016, 548, 340–346

Data Statement

The datasets generated during and/or analysed during the current study are not available because the findings of this study are available on request from the corresponding author, but the authors will make every reasonable effort to publish them in near future.