Modeling airborne indoor and outdoor particulate matter using genetic programming

Rama Rao Karrri,⁎, Mahmoud Mohammadyanb, Mahboobeh Ghoochanid, Reza Ali Mohammadpourie, Yusri Yusupf, Mohd Rafatullahf, Behzad Heibatic, J.N. Sahu,a,b,*

a Petroleum & Chemical Engineering, Universiti Teknologi Brunei, Brunei Darussalam
b University of Stuttgart, Institute of Chemical Technology, Faculty of Chemistry, Stuttgart, Germany
c Health Science Research Center, Department of Occupational Health Engineering, Faculty of Health, Mazandaran University of Medical Sciences, Sari, Iran
d Department of Environmental Health Engineering, Faculty of Public Health, Tehran University of Medical Sciences Tehran, Iran
e Department of Biostatistics, Faculty of Health, Mazandaran University of Medical Sciences, Sari, Iran
f Environmental Technology, School of Industrial Technology, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia

ARTICLE INFO

Keywords:
Air quality
Airborne particles
Particulate matter
Modeling
Genetic programming

ABSTRACT

Airborne particulate matter (PM) is considered to be an essential indicator of outdoor and indoor air quality. In this study, indoor and outdoor PM1, PM2.5, PM10 concentrations were monitored at different locations within the Tehran University campus. It is found that 10% of PM1, PM2.5, and PM10 concentrations were higher than 36.11, 52.48, and 92.13 μg/m3 for indoors respectively. Genetic programming (GP) based methodology is implemented to identify the influence of outdoor PM on the indoor PM and established significant empirical models. The best GP model is identified based on fitness measure and root mean square error. It was observed that the GP based models are perfectly able to mimic the behavioural trends of outdoor particulate matter for PM1, PM2.5, and PM10 concentrations. The model predictions are very similar to the measured values and their variation was less than ± 8%. This analysis confirms the performance of GP based data driven modeling approach to predict the relationship between the outdoor particulate matter and its influence on the indoor particulate matter concentration.

1. Introduction

In recent years, the indoor air quality in residential buildings and workplaces has become an increasing concern as the inhabitants spend more time indoors. Airborne particulate matter (PM) is an essential indicator of indoor and outdoor air quality. Epidemiological studies have documented a significant positive correlation between the daily mean concentration of particles (PM1, PM2.5, PM10) and increased mortality and morbidity attributable to respiratory and cardiovascular diseases (Anderson, Atkinson, Peacock, Marston, & Konstantinou, 2004; Hazarika, Srivastava, & Das, 2017). Major human health issues resulting from PM exposure include aggravation of asthma, bronchitis, and other respiratory problems, leading to increased hospital admissions and decreased life expectancy. Smaller particles have increasingly more severe impacts on human health as compared to larger particles. The international agency for research on cancer recently concluded that exposure to PM in outdoor air is carcinogenic to humans and causes lung cancer (Hamra et al., 2014; Shafaghat, Keyvanfar, Manteghi, & Lamit, 2016). Both short-term and long-term epidemiological studies have reported respiratory and cardiopulmonary effects with increased cardiopulmonary mortality (Gehring et al., 2006; Hoek, Bruckner, Goldbohm, Fischer, & van den Brandt, 2002; Krewski et al., 2004; Le Tertre et al., 2002; McDonnell, Nishino-Ishikawa, Petersen, Chen, & Abbey, 2000; Miller et al., 2007). Ostro (2004) have reported that there exists a significant association between PM10 concentrations and the medical visits by children and elderly persons for lower respiratory and upper respiratory symptoms respectively. Other environmental effects of PM are visibility reduction, acidic precipitation, and the transport of pollutants from industrial regions to remote and pristine areas. Few studies have reported that PM is possibly responsible for global climate change through their direct and indirect role in the earth’s radiation balance (Levy et al., 2013; Zorbas & Skouroupatis, 2016).

Air pollution caused by industrialization is one of the major environmental challenges and different industries contribute to air pollution in different ways (Brusheilwe et al., 2012; Reisen, Bhujel, & Leonard, 2014). Motor vehicle traffic is also an important source of...
particulate pollution in cities of the developing world, where rapid growth, coupled with a lack of adequate transport and land use planning, may result in harmful levels of fine particles (PM$_{2.5}$) in the air (Bereitschaft, 2015; Qiu et al., 2017). The inhabitants in these industrialized cities spend more than 90% of their time in indoor environments carrying various daily activities. Especially students spend approximately 60% of their daytime in their school/college/university campus and hence exposed to closed indoor pollution (An & Yu, 2018). Due to prolonged exposure, they are vulnerable to potential health hazards. Indoor concentrations of air pollutants will vary by time and geographical location. Usually high mobility in the daytime and limited or no movement in the night time is commonly observed inside a regular university campus. Within the university indoors, the main source of airborne particulate matters is due to re-suspension of particles because of mobility of occupants. Any outdoor activities that generate dust or air pollutants might penetrate indoors and remain inside the closed environment. Since students occupy in a closed and confined space over a period of several hours during their lectures/tutorials and laboratories, hence prone to be exposed in long term. Currently, very few studies are reported in open literature about indoor classroom PM concentrations and occupants’ exposure patterns to PM in a university establishment (An & Yu, 2018; Manimaran & Narayana, 2018).

Any studies related to identify correlations between the indoor and outdoor PM will be a great boon for student’s society and accordingly can lead to take necessary primitive measures to minimize the effect. Since the PM data can be available from various measuring devices, efficient data mining techniques can be implemented to understand the behavioural patterns and identify the complex characteristics of the system. In general, various data mining techniques were studied to identify models ranging from algebraic to differential equations system using processed data (Dong et al., 2009; Fernandez-Camacho et al., 2015; Karri, 2011; Karri, Jayakumar, & Sahu, 2017; Jayawardene, 2012; Tzima, Karatzas, Mitkas, & Karathanasis, 2007; Rao, Srinivasan, & Venkateswarlu, 2010; Riga, Tzima, Karatzas, & Mitkas, 2009; Voukantsis et al., 2011). It was found from the literature that genetic programming (GP) based data mining technique is a powerful tool for system identification when little is known about the underlying model structure in the data (Nelles, 2013). The applications on algebraic modeling using the GP approach is extensively available (Hassani, Tjahjowidodo, & Do, 2014; Kandpal, Kalyan, & Samavedham, 2013; Nelles, 2013; Pombeiro, Machado, & Silva, 2017). However, there have been little or no reported applications of genetic programming as a system identification tool to identify the state space model especially in assessing indoor and outdoor air quality.

The main objective of this research study is to identify the influence of outdoor PM on the indoor PM. Indeed, the authors initially tried to identify and establish empirical relation representing outdoor PM as dependent variables. It was found that indoor PM has significant strong correlation with only one outdoor monitored PM concentration. This is due to the fact that the PM concentrations at outdoor locations are far apart, and due to geometrical orientations of the rooms, PM at indoor locations (which are in the vicinity) had similar (strong) correlation with only one outdoor location. This thus led to a simple straightforward linear relation. But it was observed that PM measured at a particular outdoor location is consequently affecting the indoor concentration of many indoor rooms. In this research study, it was observed that the particulate matters entering classrooms through the doors and windows had stronger influence with the outdoor monitoring...
location in its vicinity. So in the established state space models, the influence of outdoor PM on the respective indoor can be estimated based on the significance of the coefficients with respect to the corresponding indoor location. Hence this approach will help us to identify the indoor locations which are strongly influenced by particular outdoor pollution. Therefore, primary objective of this study is to model indoor and outdoor PM concentrations using the genetic programming-based framework and identify the correlation between them.

2. Materials and methods

Airborne particulate matters with size 1, 2.5 and 10 microns respectively (PM$_1$, PM$_{2.5}$ and PM$_{10}$) are measured at specified indoor locations consisting of four classrooms, four corridors, and four faculty offices in the Payambar Azam Campus of the Mazandaran University of Medical Sciences, Iran. Samples are collected at all the monitoring locations using continuous sampling instruments during weekday mornings and afternoons during the month of April until September 2014, wherein this sample period covers both season’s spring and summer. The number of sampling locations monitored in indoors was 12 (labeled as X$_1$ to X$_{12}$) and outdoors were 5 (labeled as Y$_1$ to Y$_5$) respectively as shown in Fig. 1. This figure is rough representation of the monitoring locations and campus area. The orientation is not accurately measured using GIS, hence they are roughly presented and the distance between the room are much far than they are shown here. But this sketch presents a rough indication of monitoring indoor locations and corridor. This study was initially planned only to monitor the PM in class rooms and corridors in their vicinity, but later due to request from few staff members (who reported ill health due to pollution), this study was extended by monitoring their rooms as well. Since the PM need to be measured twice a day, so we chose randomly only four staff members (who reported ill health due to pollution), the study was extended by monitoring their rooms as well. As students spend time in the corridors as well, so we included them as well. All these measurements were taken during the working hours (9 a.m. – 4 pm) from Monday to Friday. We chose four outdoor locations, exterior to the campus and one outdoor location within the campus. All these measurements were done systematically at all locations and in total of 230 samples are collected at each indoor/outdoor location.

Sampling was done on every day, irrespective of the climatic conditions. PM$_1$, PM$_{2.5}$ and PM$_{10}$ were determined by hand-held Environmental Dust Monitor, GRIMM Model EDM 107. This device uses light-scattering technology to count particles using lasers. The lasers measure particle diameters and record them according to their size. Thus, the amounts of PM$_1$, PM$_{2.5}$ and PM$_{10}$ can be measured simultaneously for each sample.

The samples are collected by sucking the ambient air containing particles using an internal calibrated pump with a flow rate of 1.2 L/min. The air sample passes through a measuring cell and then measured by a detector diode laser. Later, this air sample then passes through a polytetrafluoroethylene filter (Dia of 47 mm). The particles deposited on the filter were used for further analysis. This measurement principle is consistent with the guidelines by the USEPA and the EEA (Oliveira, Slezakova, Delerue-Matos, Pereira, & Morais, 2016). This device is connected to data acquisition system which automatically retrieves the respective data in its memory for further analysis.

To study the influence of localised driving parameters (like relative humidity, pressure and temperature) effect on PM at various locations, these parameters are also monitored concurrently. As these airborne particles concentrate or scatter depending on the natural driving factors like temperature, humidity, wind speed and direction, these parameters were also collected from the Mazandaran local meteorological office. The data hence collected from these samples from various locations are extracted from each instrument and analysed using the standard Microsoft® Excel, 2010 for data processing. Further analysis and interpretation of the data are carried out with MATLAB R2014a.

3. Implementation of genetic programming

3.1. GP implementation procedure

Due to the varied parameters’ contribution to PM, we attempted to identify the correlation between PM$_1$, PM$_{2.5}$ and PM$_{10}$ indoors and outdoors respectively. Identification of a relation between the input and output variables, falls under the category of system identification problem. In this framework, input-output behaviour of the unknown process system is approximated using an appropriate model. Appropriate modeling components must be selected to ensure that a model can accurately reproduce the behaviour of the system and estimate the inherent characteristics. To identify these complex non-linear relations, various data mining techniques were applied for different applications (Dong et al., 2009; Fernandez-Camacho et al., 2015; Jayawardene, 2012; Karri, 2011; Rao, Rao, & Venkateswarlu, 2009; Tzima et al., 2007). Genetic programming based data mining technique was identified as a method to approximate the input-output behavior. GP is a powerful tool for system identification when the underlying model structure is unknown and this technique has great potential due to its ability to seek discontinuous complex nonlinear spaces (Babovic & Keijzer, 2002; Hassan et al., 2014; Kandpal et al., 2013; Nelles, 2013; Oliveira et al., 2016; Parasuraman, Elshorbagy, & Carey, 2007).

Genetic algorithm is formulated by representing the solution as individuals in a population that evolves through several generations. The genetic algorithm attempts to find a very good or best solution to the problem by genetically breeding the population of individuals. Numerical parameters that appear in each of the models are estimated in order to minimize the error between the actual output and the model predictions. GP is basically derived from conventional genetic algorithm which imitates the Darwinian principle of reproduction and survival of the fittest and naturally occurring genetic operations such as crossover (recombination) and mutation to solve a suitably posed mathematical problem. It has the advantage of being transparent because it can provide explicit mathematical equations of the model, in contrast to neural network based applications that is more "opaque". It also has the ability to provide a family of models rather than a single model. The latter feature can assist in choosing a physically meaningful model from a population of valid models.

Genetic programming starts with an initial population of randomly generated functions applicable to the problem domain. The important parameters to be initialized are number of runs, number of generations, population size, the probability of genetic operations and specification for termination criterion. The functions will be composed of standard arithmetic operations, programming operations, mathematical functions and/or domain-specific functions. Each individual function in the population is measured in terms of how well it meets the objective which is termed as fitness measure. In system identification, the model serves as a program and is simulated using a simulation engine. The resulting simulated output data for a given input is compared with actual output data. The parameters in the model are optimized based on the root mean square error. The flowchart of an implementation of GP strategy is presented in Fig. 2.

3.2. Modeling approach

Data driven modeling approach which is being a conventional data mining technique to establish a model based on the widespread data (Babovic & Keijzer, 2002; Gaileli et al., 2014; Orouji, Bozorg Haddad, Pahlavani, & Mariño, 2013; Pandey, Pan, Das, Leahy, & Kwapisinski, 2015; Parasuraman et al., 2007; Ramos, Almeida, Simões, & Pereira, 2017; Solomatine & Ostfeld, 2008) is explored in this research study. Incidentally, genetic programming based data driven modeling was found to be an effective tool to establish a multi-nonlinear model (Hassani et al., 2014; Kandpal et al., 2013; Nelles, 2013; Rao et al., 2009). In this regard, genetic programming is implemented to identify
the state space model to determine the relationship between the measured outdoors particulate matters versus measured indoors. The empirical state space model was developed using the major contributing variables (locations) and other significant variables (excluding the correlated ones). Initially, the PM\(_1\) measured at outdoors (\(x_1\) – \(x_5\)) are identified as a function of corresponding PM measured indoors at all locations. State space model is identified by implementing genetic programming for each outdoor location as written in Eq. (1).

\[
x_{1-5} = f(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}, y_{12})
\]  

(1)

The algorithm is implemented for training, and the model is established with a various number of chromosomes for genetic mutation. This algorithm is implemented using the genetic modeling system toolbox by Tun and Lakshminarayanan (Oliveira et al., 2016), and the procedure and methodology of implementation are shown in Fig. 2. The GP model is evaluated by training on the given set of sample PM data and significance of selected GP model is tested on the testing data. Here the number of generations is chosen as 30, and population size is 128. A maximum number of identical chromosomes are assigned to be 3. The probability of mutation, crossover, reproduction, permutation and adaptation are chosen to be 2, 95, 5, 5, 10 and 0 respectively. Well-known statistical measures such as Akaike Information Criterion (AIC) and Rissanen’s Minimum Description Length (MDL) are employed as fitness measure to discriminate or choose the best-optimized model. The number of trials and maximum optimization time are fixed at 800 and 1200 s respectively.To train/develop a model using nonlinear regression framework and check the statistical significance of model parameters for the combined model structure, the 80% of the PM data is used. The remaining 20% of the data is used to validate (testing) the efficacy of the identified model.

4. Results and discussion

The descriptive statistics of the indoor and outdoor sampling locations for PM\(_1\), PM\(_{2.5}\) and PM\(_{10}\) concentrations respectively are shown in Table 1. The mean indoor PM\(_1\) concentration (18.70 μg/m\(^3\)) was much higher than the outdoor mean PM\(_1\) concentration (16.46 μg/m\(^3\)). Similarly, it was also observed that the mean indoor PM\(_{10}\) concentration (60.22 μg/m\(^3\)) was much higher than the corresponding outdoor mean PM\(_{10}\) concentration (53.42 μg/m\(^3\)). These observations are definitely alarming, as the students in university spend more time in the class (indoors) and hence exposed to more particulate matter, which will severely affect their health. The indoor concentration was higher than outdoor can be attributed to the fact that the mobility of dust particles increases due to the constant movement of students. Further, the classrooms are cleaned once a week, which means more dust particles accumulate in the classroom floors. The scattering of particulate matters will be much more if indoors fans are used instead of air conditioners. The mean outdoor PM\(_{2.5}\) is higher than the annual standard mean PM\(_{2.5}\) (12.0 μg/m\(^3\)) recommended by USEPA (2013). The mean indoor and outdoor PM\(_{10}\) are also higher than the annual standard mean PM\(_{10}\) (40.0 μg/m\(^3\)) endorsed by EEA (Guerreiro, de Leeuw, & Poltescu, 2013).

The cumulative frequency analysis, reported in Table 1 presents that 50% of indoor PM\(_1\), PM\(_{2.5}\) and PM\(_{10}\) concentrations were 16.75, 26.24 and 62.67 μg/m\(^3\) respectively. It is also observed that 10% of PM\(_1\), PM\(_{2.5}\) and PM\(_{10}\) concentrations were higher than 36.11, 52.48 and 92.13 μg/m\(^3\) for indoors respectively and 27.54, 39.94 and 83.83 μg/m\(^3\) for outdoors respectively. From the standard deviation values of each parameter, indoor PM concentrations were more varied than outdoor PM concentrations. The 90th percentile results also indicate that indoor PM concentrations were generally higher than outdoor concentrations.

4.1. Genetic programming based on data-driven modeling of PM\(_1\) concentrations

Based on the standard statistical measures like best fitness measure and root mean square error (RMSE), the following empirical state space models were identified from the training dataset is presented in Eqs. (2)-(6) obtained for a chromosome number of 1683. In these empirical models, only a few variables are shown as a function, and henceforth the other missing variables can be attributed to the fact that they are either non-correlated or least significant in predicting the behaviour of the outdoor variable. These models determined by testing on the training data set and its efficacy of these models for \(y_1\) and \(y_2\) are shown in Fig. 3. The performance of the respective models for \(y_3\), \(y_4\) and \(y_5\) are very similar and hence theirs plots are not presented here due to space constraint. These empirical models depicts those indoor locations which are strongly influenced by the outdoor locations. In these models, the least significant (less influenced) indoor location presented very low coefficient and hence neglected from the model expressions. As expected, the indoor locations which are in the vicinity of considered outdoor locations, shown to be strongly influenced.

\[
y_1 = 0.2479 \times x_1 + (0.5551 \times x_3 + 1.864 \times \left(\frac{x_1}{x_2}\right) + 2.32 \times 10^{-1} \times x_5)
\]  

(2)

\[
y_2 = 0.172 \times x_1 - 0.01926 \times (x_2)^2 + x_7 + 0.7492 \times \frac{x_2}{x_4} 
\]  

(3)

\[
y_3 = -1.632((-0.4316 \times x_9) + (0.4824 \times 10^{-1} \times x_2 \times x_5 - 0.1195 \times x_4)) 
\]  

(4)

\[
y_4 = 0.2611 \times x_0 + ((5.701 + 0.667 \times 10^{-11} \times (x_7 - x_8)) + 0.4059 \times x_9) 
\]  

(5)

\[
y_5 = 8.437 + 0.01434 \times (x_5 \times x_0) - 0.4615 \times 10^{-11} \times (x_7 - 2.610 \times x_0) \times x_6 
\]  

(6)

To validate the performance of identified empirical models, these models are implemented to predict the values for the testing data set and these predicted values from each corresponding model are compared against the measured values and the following time series trends were plotted for PM\(_1\) and the five outdoor locations (\(y_1\)-\(y_5\)), shown in Fig. 4. From these profiles, it can be observed that the GP based models are perfectly able to mimic the behavioural trends of outdoor particulate matter (PM\(_1\)). These models may slightly overestimate or underestimate on few occasions, but the overall trend and highest peaks are very much matching. The PM\(_1\) is below 40 μg/m\(^3\) (Hamra et al., 2014) at most events, but on a particular day (11th May 2014), the particulate...
concentration is very high, and this behaviour is observed in all the locations both indoor as well as outdoor. This behaviour can be attributed to the fact that drilling of ground for masonry work was carried on this day within the university campus. Owing to the fact that more dust scattered during this work, the highest particulate matter concentration is observed at the locations as well as similar behaviour was also found in particulate matter PM2.5 & PM10. Presence of PM1 above 15 μg/m³ in the atmosphere, can pose severe health issues. These smaller particles can penetrate more deeply into the human body and aggravate asthma, bronchitis, and other respiratory problems, and further leading to cardiovascular symptoms.

To further observe the variation of predicted model values against the measured values, and quantify the percentage of over-estimation/under-estimation, the scatter plot being best tool to quantify the error deviation and hence presented in Fig. 5 for the four locations Y1 - Y4. To differentiate the difference between the predicted and measured values, a 45° (y = x) line is also plotted in each sub-plot. All the points lying on this 45° line shows that the measured and predicted values are very much similar. The points above the 45° line depict the values of over-estimation and points values below the 45° line depicts the benefits of underestimation. It is observed in all the locations; the points are very close to 45° line, which itself indicates that the model predictions are very similar to the measured values and their variation was less than ± 8%. This marginal variation can be attributed to internal generation of PM due to student mobility and external influence of other factors around the campus. This analysis confirms the performance of quality of GP based data-driven modeling procedure to predict the relationship between the outdoor particulate matter and its influence on the indoor particulate matter concentration.

4.2. Genetic programming based data driven modeling of PM2.5 concentrations

As explained in the earlier section for identifying the state space model for PM1, in the similar way, an analysis is performed to identify PM2.5 measured at outdoors (Y1–Y5) as a function of corresponding PM measured at all indoors locations. The state space model is determined by implementing genetic programming for each outdoor location as written in Eq (1). The algorithm is implemented for training data set (80% of the measured PM2.5 data), and the models are established with various numbers of chromosomes for genetic mutation, and the best model is identified by varying the number of chromosomes and based on the standard statistical measures like best fitness measure and RMSE. The following empirical models were determined from the training dataset for PM2.5 measured at outdoors (X1–X5) as a function of corresponding PM measured at all indoors locations are shown in Eqs. (7)–(11) obtained for a chromosome number of 2248. The functional relationship of outdoor locations PM with respect to indoor locations represented in these empirical state space models are very similar to relations obtained for PM1.

\[
y_1^* = (x_1 + 59.47 * x_1)/(3.395 * x_1 + x_5) + 0.4859 * x_9 \quad (7)
\]

\[
y_2^* = x_4 - 0.4134E-02 * x_2 - (-0.1470 * x_4 + x_9) + 0.326 * x_4 \quad (8)
\]

\[
y_3^* = 0.866 * x_12 + x_9/(-0.743 * x_11 + 0.4075 * x_12 + 1.52 * x_4) \quad (9)
\]

\[
y_4^* = 0.7158E-02 * (x_3 * x_9) + 0.782 * x_5 * 0.1041 * x_7 \quad (10)
\]

\[
y_5^* = 0.28 * x_3 + 0.635 * x_10/x_1 + 0.782 * x_7 * 0.1041 * x_8 \quad (11)
\]

To further validate the performance of these predicted empirical models, as done for PM1, these models are implemented to predict the values for the testing data set and these predicted values from each
model are compared against the corresponding measured values for PM$_{2.5}$, the following time series plots at the five outdoor locations (Y1–Y4) are shown in Fig. 6. From these profiles, it can be observed that the GP based models are also able to perfectly mimic the behavioural trends of outdoor particulate matter, PM$_{2.5}$. As observed in this figure, on one particular occasion, the PM$_{2.5}$ is much above the mean value from all the remaining occasions. This event has occurred on 11$^{th}$ May 2014, wherein as mentioned earlier that this highest peak of particulate matter concentration is due to the fact that drilling of ground for masonry work was carried on this day. The variations in the PM concentration are observed at all the outdoor locations. It was found that there is factory within the close proximity of university campus that produces MDF sheets. According to residents within the university campus, they observed that frequent night activity of plants produces lot of dust, which forms a haze type situation in the early morning. These unsettled pollution causes irritation in eyes for morning walkers and haze hinders the visibility. The outdoor PM$_{2.5}$ is much higher than the permissible annual standard mean PM$_{2.5}$ (12.0 $\mu$g/m$^3$) per day recommended by US environment protection agency (USEPA). Similarly, the outdoor PM$_{2.5}$ is higher than the annual standard mean PM$_{2.5}$ (25.0 $\mu$g/m$^3$) recommended by the EEA.

To further quantify the variation of predicted empirical models output against the corresponding measured values for PM$_{2.5}$, the scatter plots for the four locations Y1, Y2, Y3 and Y5 are presented in Fig. 7. These scatter plot reveals the quality of predicted models outputs and quantifies the overall percentage of over-estimation/under-estimation. It is observed that many points lie on this 45° line, which indicated that for certain occasions, the predicted values are very close to the measured PM values. On other few instances, the predicted values are spread firmly on either side of the 45° line and are within the variation of less than ± 10%. This spread of predicted values close to the $y = x$ line depicts that the model predicted values are not too far from the measured values. This marginal variation can be attributed to accountability of external influence of natural driving parameters and other localised factors. Even though the model able to pick the highest variation (peak) behaviour on a particular day, but their values marginally underestimated, which can also be seen in all the scatter plots presented in Fig. 7. These scatter plots thus confirms the quality of performance of GP based data-driven modeling technique to predict the relationship between the outdoor particulate matter, PM$_{2.5}$ and its influence on the indoor particulate matter concentration.
Fig. 5. Scatter plots showing the variation of measured against predicted at locations Y1–Y4, particulate matter, PM$_{1.0}$.

Fig. 6. Plots depicting the performance of predicted against the corresponding measured PM$_{2.5}$ at locations Y1–Y4.
4.3. Genetic programming based data driven modeling of PM$_{10}$

To further examine the quality of performance of GP to predict the relationship between the outdoor particulate matter, PM$_{10}$, a similar exercise is done to identify the model by training & validating the models. The algorithm is implemented for training, and thus the model is established with various numbers of chromosomes for genetic

Fig. 7. Scatter plot showing the variation of measured against predicted at locations Y$_1$, Y$_2$, Y$_3$ & Y$_5$ particulate matter, PM$_{2.5}$.

Fig. 8. Time series plots depicting the performance of predicted against the corresponding measured PM$_{10}$ at locations Y$_1$–Y$_4$. 

4.3. Genetic programming based data driven modeling of PM$_{10}$

To further examine the quality of performance of GP to predict the relationship between the outdoor particulate matter, PM$_{10}$, a similar exercise is done to identify the model by training & validating the models. The algorithm is implemented for training, and thus the model is established with various numbers of chromosomes for genetic
mutation, and the best model is identified based on the fitness measure and RMSE. The following empirical models as shown in Eqs. (12)–(16), were identified for PM10 measured at outdoors (Y1–Y5) as a function of corresponding PM measured at all indoor locations for a chromosome number of 1278. In this empirical models, only whose coefficients had strong significance are represented and other are neglected from the expression. The functional relationship of outdoor locations PM with respect to indoor locations represented in these empirical state space models are very similar to relations obtained for PM1 and PM2.5. Hence it confirms that variation in PM1, PM2.5, & PM10 are very much similar and strongly influence those indoor locations which are in the vicinity of outdoor locations.

\[
\hat{y}_1 = 0.678 * (x_1 - 0.3838 * (x_1 - x_0) + 13.29 * (x_0/x_1) - 0.061 * x_0) \\
\hat{y}_2 = 3.690 * x_2 - 4.668 * (-0.0486 * x_4 + 0.7034 * x_3 - (x_4 + x_3)/(x_7 * x_12)) \\
\hat{y}_3 = 0.861 * x_9 + 40.73 * ((3 * x_4 - 0.08197 * x_1)/(x_5 * x_12)) \\
\hat{y}_4 = \frac{122}{0.069 * x_7 + 0.134 * x_9} - 0.014 + 0.013 * (x_5 x_7 - 10.32 * x_3 + 0.453 * x_10) \\
\hat{y}_5 = 0.3521 * x_4 + 0.65 * x_6 + 0.0352 * x_7 - 0.4 * x_90/x_7
\]

As performed for PM1 & PM2.5, these models are implemented to predict the values for the testing data set and these predicted values from each model are compared against the corresponding measured for PM10, the following time series plots as shown in Fig. 8 are plotted at the four outdoor locations (Y1–Y4). The particulate concentration is spread across from 20 μg/m³ to 140 μg/m³ with wide fluctuations on every day. The fluctuations in the particulate matter concentration can be attributed to the fact that dust released from the nearby MDF sheet producing plant has higher suspended particles. Even though the outdoor PM10 concentrations did not exceed the respective USEPA standards of 150 μg/m³, but due to the suspension of higher diameter airborne particles in the atmosphere causes aggravation of asthma, bronchitis, and other respiratory problems. From the profiles shown in Fig. 8, it can be observed that these predicted models are also able to perfectly mimic the behavioural trends of outdoor particulate matter at all the locations. Even though the quality of performance of not as good as observed for predicting PM1 & PM2.5, nevertheless the overall trend, peaks and troughs are well matched. These results thus confirm that the GP based models are perfectly able to mimic the behavioural trends of outdoor particulate matter for PM1, PM2.5 and PM10.

The variation of predicted empirical models output against the corresponding measured values for PM2.5 is quantified by plotting the corresponding scatter plots for the four locations Y1, Y2, Y3 and Y4. These profiles are presented in Fig. 9. As expected, very few values fall on y = x line, but all other points are evenly spread on either side of 45˚ line. This even spread of points depicts that the predicted model outputs are not too far from the measured values and hence the quality of model outputs are very much in-line with the measured values. The variation of model predictions as compared to the measured values are within the admissible range of ± 10%. Hence these profiles, also confirm that the GP based models can perfectly mimic the behavioural trends of outdoor particulate matter, PM10.

In this research study, the PM at indoor locations are taken as independent variables for building relationship models, owing to the fact that the PM at indoor and outdoor are very much inter-dependent. This approach was initially investigated other way round, where the outdoor locations are taken as independent variables and the indoor PM concentrations are predicted, but this exercise has resulted in no significant relationship models. This is due to the fact that the PM concentrations at outdoor locations are far apart, and due to geometrical orientations of the rooms, PM at indoor locations (which are in the vicinity) had similar (strong) correlation with the same outdoor location. So all the rooms which are facing towards a particular outdoor monitoring location had similar functional relationship with different weightages. Hence this approach is not so effective to estimate the inherent
characteristics of indoor PM. In this regard, the relationship models are identified; wherein the PM concentrations at different indoor locations which are being influenced by a particular outdoor PM concentration. From the empirical state space model shown in Eq. (2) for PM$_1$, shows that PM measured at $y_1$, outdoor location is influencing the indoor locations $x_1$, $x_2$, $x_3$, and $x_4$ with different weightages. This indicates that the variation in indoor PM$_1$ concentration at $x_1$ can be described from the existing outdoor PM$_1$ concentration $y_1$, and feature variance can be completely explained due to strong correlation (weightage). Whereas, the feature variance at $x_2$ can be partially described by outdoor PM$_1$ concentration $y_1$ and remaining variation can be due to internal high mobility/classroom activities. Similar trends are also observed for PM$_{2.5}$ as well as PM$_{10}$, thus inferring that PM at $y_1$, is influencing the indoor locations $x_1$, $x_2$, $x_3$, and $x_4$. Parallel analogies can be interpreted from the equations represented for other outdoor locations and their influence on the respective indoor PM concentration for PM$_1$, PM$_{2.5}$ as well as PM$_{10}$. The uncertainty that are not able to captured by the non-linear models, are due to other driving factors like rainfall, wind, and so on. Overall, the outcomes derived from the model predictions and measured values facilitate the knowledge to characterize the uncertainty and estimate the influence of natural driving parameters. These results also confirm that the GP based models are perfectly able to mimic the behavioural trends of outdoor particulate matter for PM$_1$, PM$_{2.5}$ as well as PM$_{10}$.

4.4. Analysis of natural driving parameters on particulate matter concentrations variations

To identify the influence of natural driving parameters like wind, relative humidity, rainfall, solar hours and temperature fluctuations in airborne PM at various locations, a correlation analysis is performed using Matlab. PM$_1$ at the outdoor locations are compared against the above mentioned natural driving parameters and its correlation to each parameter is shown in Fig. 10. This profile presents the overall view of the influence (correlation) of each natural driving parameter against the PM. The parameters like temperature, relative humidity, solar hours, wind direction and wind speed, all display a significant correlation with the PM$_1$ variation, except the precipitation. It was also observed that evaporation has a strong negative correlation with PM$_1$. This profile is for PM$_1$ at all the outdoor locations. Similar profiles are obtained for PM$_{2.5}$ & PM$_{10}$, but due to space limitations, all these profiles are not presented here. However, similar observations are found in these profiles as well. The strong positive and negative correlations of major natural driving parameters is very significant indicating the variations in the PM at various locations. As expected, precipitation is found to be insignificant owing to the fact that the PM measured are indoor locations. Therefore, these studies thus confirms that the natural driving parameters influencing the scattering of particulate matter in both indoor and outdoor locations. These studies and outcomes cater resilient environments for smarter cities which further enhance the framework of monitoring and improving air quality in built environment and cities.
5. Conclusions

Since students usually spend approximately 60% of their daytime in the university and are regarded as particularly vulnerable to potential health hazards due to longer exposure to closed indoor pollution. In this research study, indoor and outdoor PM$_1$, PM$_{2.5}$, PM$_{10}$ concentrations were monitored at 12 indoors and 5 outdoors locations, spread across the university campus. It was found that the mean indoor PM$_{10}$ concentrations (18.70, 60.22$\mu g/m^3$) are much higher than corresponding outdoor concentration (16.46, 53.42$\mu g/m^3$). The mean indoor and outdoor PM$_{10}$ found to be higher than the annual standard mean endorsed by European and US environment agencies. Genetic programming based data driven modeling was implemented to identify the state space model and these models are perfectly able to mimic the behavioural trends of outdoor PM$_1$, PM$_{2.5}$ and PM$_{10}$ concentrations.

The model predictions are very close to the measured values and their variation are within the permissible limits of ± 10%. Even though the PM at indoor locations are taken as independent variables for building relationship models, but the PM at indoor and outdoor are very much inter-dependent. The outcomes derived from the model predictions and measured values facilitate the knowledge to characterize the uncertainty and estimate the influence of natural driving parameters. It was observed that the dispersion of PM at each indoor location is due to external particulate matter as well as due to internal generation of PM (mobility of students and teaching activities). So for a range of 50 m – 200 m, the PM in indoor location may not vary quite significantly, whereas, the PM in indoor location in the same range, will vary significantly depending on the activities held at each indoor location and orientation of room. Even though monitoring at indoor is expensive and tedious, due to concern of student health, it’s a good practice to sample regularly at each indoor location which has to be done on a regular periods. Therefore this research study outlines the prominence of healthy built environment and air quality management for development of sustainable cities.

Conflict of interest

Authors have no conflict of interest.

References

Babovic, V., & Keijzer, M. (2002). Rainfall runooffluence of natural driving parameters. It


programming based predictive models for municipal solid waste gasification in a fluidized bed gasifier. Biosresource Technology, 179, 524–533.

Pombeiro, H., Machado, M. J., & Silva, C. (2017). Dynamic programming and genetic algo-

rithms to control an HVAC system: Maximising thermal comfort and minimizing cost with seasonal weather conditions. Sustainable Cities and Society, 34, 228–238.