




REVIEW ARTICLE

WILEY

Machine learning and statistical models for predicting indoor air quality

Wenjuan Wei¹  | Olivier Ramalho¹ | Laetitia Malingre¹ | Sutharsini Sivanantham¹ | John C. Little²  | Corinne Mandin¹ 

¹Scientific and Technical Center for Building (CSTB), Health and Comfort Department, French Indoor Air Quality Observatory (OQAI), University of Paris-Est, Marne la Vallée Cedex 2, France

²Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia, USA

Correspondence

Wenjuan Wei, Centre Scientifique et Technique du Bâtiment (CSTB), Direction Santé-Confort – Observatoire de la Qualité de l'Air Intérieur (OQAI), 84 Avenue Jean Jaurès, Champs sur Marne, 77447 Marne la Vallée Cedex 2, France.
Email: Wenjuan.WEI@cstb.fr

Funding information

Scientific and Technical Center for Building (CSTB), Grant/Award Number: SEC21; University of La Rochelle

Abstract

Indoor air quality (IAQ), as determined by the concentrations of indoor air pollutants, can be predicted using either physically based mechanistic models or statistical models that are driven by measured data. In comparison with mechanistic models mostly used in unoccupied or scenario-based environments, statistical models have great potential to explore IAQ captured in large measurement campaigns or in real occupied environments. The present study carried out the first literature review of the use of statistical models to predict IAQ. The most commonly used statistical modeling methods were reviewed and their strengths and weaknesses discussed. Thirty-seven publications, in which statistical models were applied to predict IAQ, were identified. These studies were all published in the past decade, indicating the emergence of the awareness and application of machine learning and statistical modeling in the field of IAQ. The concentrations of indoor particulate matter (PM_{2.5} and PM₁₀) were the most frequently studied parameters, followed by carbon dioxide and radon. The most popular statistical models applied to IAQ were artificial neural networks, multiple linear regression, partial least squares, and decision trees.

KEYWORDS

artificial neural networks, data mining, IAQ, partial least squares, particulate matter, regression

1 | INTRODUCTION

Predicting the concentration of indoor air pollutants to characterize indoor air quality (IAQ) has long been an important topic in the domain of indoor air science. Compared to the measurement of indoor concentrations, prediction is non-invasive and can be quick and inexpensive. Predictions can be used to estimate IAQ while buildings are being designed and can also provide long-term evaluation of expected IAQ in some cases.

Indoor air quality is commonly predicted using mechanistic models based on an understanding of the underlying mechanisms governing the fate and transport of indoor air pollutants, including, for example, diffusion of compounds from source materials, convective mass transfer of compounds in air, and sorption of compounds into

or onto sink materials. Mechanistic models have been developed to predict the concentrations of volatile organic compounds (VOCs),¹⁻⁴ semi-volatile organic compounds (SVOCs),⁵⁻⁷ aldehydes,⁸⁻¹⁰ inorganic compounds including radon,¹¹⁻¹⁴ and particulate matter (PM)¹⁵⁻¹⁹ in indoor environments. Mechanistic models require detailed inputs on both indoor sources and sink materials for the target pollutants, building envelope and ventilation conditions, and outdoor concentrations of the target pollutants. The outputs of the models are either dynamic or steady state indoor concentrations of the pollutants. Mechanistic models can be used in unoccupied environments where the above-mentioned information is available and somewhat controlled. They can also be used in occupied environments when the occupancy scenario is well defined. Since mechanistic models require complex input data, they are frequently used

in well-designed case studies. One major advantage of mechanistic models is that they can be used during the design stage of a building to provide insight on possible indoor concentrations before the building is constructed and to help select appropriate low-emission materials for construction. Since mechanistic models require complex input data, it is difficult to retrieve enough information to run mechanistic models when multiple buildings are involved and particularly in real environments when the occupants interact with their indoor environments including indoor sources and window openings. In these cases, statistical models can serve as an alternative approach for IAQ prediction. In indoor air surveys, measured concentration of pollutants can be related to questionnaire data using statistical models that can subsequently be applied to estimate indoor concentrations in new environments. In a specific building, measured concentrations of pollutants can be related to other indoor and/or outdoor parameters using statistical models, so that a pollutant concentration can be predicted from the other parameters. In general, while mechanistic models may appear more trustworthy, statistical models can nevertheless be very useful, especially when the specific mechanisms or their dynamic variation are not well established, and when large data sets exist.

Machine learning and statistical models have been widely used in outdoor environments to predict the concentrations of atmospheric pollutants²⁰⁻²³ and in indoor environments to predict thermal comfort²⁴⁻²⁶ and building energy efficiency.²⁷⁻²⁹ Models that have been commonly used in these studies include various regression models, partial least squares (PLS), decision trees (classification and regression trees), Bayesian hierarchical modeling, generalized boosting models, support vector machine, random forests, generalized linear models, and artificial neural networks (ANN).^{30,31} While these predictive models deal with non-time series data, time series data can be analyzed by time series models based on autoregression, including the autoregressive model and autoregressive integrated moving average model. Time series models are a group of specialized forecasting models that use the historical profile of a parameter to forecast its future values. Time series forecasting is based on continuous or repeated measurements and is frequently used for signal detection and estimation, which has been addressed in several review papers³²⁻³⁴ and applied in different fields including energy^{35,36} and electric vehicles.³⁷

In indoor environments, although some models have been applied to predict IAQ, studies on the depth and breadth of the applications and on the state of the art of statistical predictions of IAQ are lacking. The objectives of this review are therefore to: (a) summarize and compare the common machine learning and statistical modeling methods, (b) discuss how and where the methods can be and have been used in the field of IAQ, mainly focusing on the prediction of the concentrations of indoor air pollutants, and (c) identify opportunities, gaps, and research needs for indoor air science. Due to the difference between predictive models and time series models in terms of data requirements and application purposes, time series models were not included in the review, but are included in the Discussion section.

Practical Implications

- The present work has reviewed the state of the art of IAQ predictions using statistical models.
- Possible ways to improve the application of statistical models to predict IAQ were discussed.

2 | MATERIAL AND METHODS

Reference books³⁸⁻⁴⁵ about statistical methods for prediction have been consulted to describe the main methods according to their common classification. The literature search was limited to the field of IAQ to avoid the large number of applications of statistical models in other fields. Peer-reviewed journal articles, conference papers, and PhD theses were searched using the Google Scholar, Science Direct and Scopus search engines, regardless of the date of publication. The keywords used for the search in either title, abstract, or keywords were "indoor air" AND predict AND concentration AND ("statistical model" OR "big data" OR "data mining" OR "machine learning" OR regression OR "partial least squares" OR "decision tree" OR "regression tree" OR "Bayesian hierarchical modeling" OR boosting OR "support vector machine" OR "random forest" OR "neural network" OR "deep learning").

A hundred and four publications were identified, from which 37 studies addressing machine learning and statistical modeling to predict chemical and particle concentrations in indoor environments were obtained for review. The other 67 studies used statistical analysis to explain but not predict IAQ and were thus discarded from the present work. For each of the 37 studies, the following information was obtained and compared: (a) objective of the study, (b) detailed information on the measurements, including the location, date, and sampling period, (c) pre-analysis of the model, including data transformation and data mining, (d) development of the model, including the inputs, outputs, and structure of the model, and (e) performance metrics of the model, including goodness of fit, mean absolute error, root mean square error, and accuracy.

The retrieved applications of the statistical models used mainly cross-sectional data for prediction although some of them also included the data at the previous sampling time as inputs for model development. Overall, predictive models perform a spatial estimate of a parameter based on knowledge of other parameters and do not deal with time series data.

3 | RESULTS

3.1 | Summary of machine learning and statistical modeling methods

Table 1 briefly describes the primary models that may be applicable to the field of IAQ and summarizes their strengths and weaknesses.

TABLE 1 Strengths and weaknesses of some statistical models and methods

Supervised learning					
Model	Description	Type of response variables	Linearity of the model	Strength	Weakness
Regression models					
Multiple linear regression ^a	A linear regression that describes the relationship between a response variable and several predictive variables	Continuous	Linear	(1) Determines the best predictors of the variable of interest (2) Simplicity of the model to be used for predictions (3) Detects outliers (4) Variables can be selected using stepwise, forward, and backward algorithm	(1) Does not deal with non-linear problems if data is not transformed linearly (2) Requires more observations than variables (3) Multi-collinearity (4) The presence of outliers can seriously bias the regression coefficients
Generalized linear model	A generalization of ordinary linear regression, which allows response variables to have errors that are not normally distributed	Continuous and categorical	Linear ^b	(1) Transformation of non-linear problems to linear problems (2) Variables can be selected using stepwise, forward, and backward algorithms	(1) Requires more observations than variables (2) Multi-collinearity (3) Sensitive to outliers
Regularized regression	A regression that regularizes or shrinks the coefficient estimates toward zero, for example, elastic net, LASSO regression, ridge regression	Continuous	Linear	(1) The number of observations can be lower than the number of variables (2) Prevents overfitting (3) May perform variable selection (4) The multi-collinearity issue is handled	(1) LASSO regression: the number of selected variables is restricted (2) Ridge regression: variables cannot be classified according to their importance in the model.
Partial least squares	A linear regression model that projects the predicted variables and the observable variables to a new space	Continuous and categorical	Linear ^b	(1) The multi-collinearity issue is handled (2) The number of observations can be lower than the number of variables (3) Possibility of having more than one response variable (4) Deals with missing data	(1) The choice of the number of components is subjective: usually with the help of cross-validation and Q^2 indicator (cross-validated R^2).
Principal component regression	A regression based on principal component analysis (PCA)	Continuous	Linear	(1) The multi-collinearity issue is handled (2) The number of observations can be lower than the number of variables	(1) Does not consider the response variable when choosing the principal components

(Continues)

TABLE 1 (Continued)

	Model	Description	Type of response variables	Linearity of the model	Strength	Weakness	Reference
Models based on decision trees	Decision tree ^a	A decision support tool that uses a tree-like model of decisions and their possible consequences	Continuous and categorical	Linear ^b	(1) Does not need upstream variable selection (2) Deals with missing data (3) Applicable for continuous and categorical response variable	(1) Requires a large amount of data to be robust (2) Prone to overfitting	44
	Gradient boosting tree	A combination of decision tree algorithms and boosting methods	Continuous and categorical	Linear and non-linear	(1) Improvement of accuracy compared to a single decision tree (2) Reduces overfitting compared to a single decision tree	(1) Requires careful tuning to ensure a good model (2) Cannot be used with parallel processing and thus takes more time to compute	44
	Random forest	An ensemble method that constructs a multitude of decision trees and outputs the mode of the classes or the mean prediction of the individual trees	Continuous and categorical	Linear and non-linear	(1) Improvement of accuracy compared to a single decision tree (2) Reduces overfitting compared to a decision tree (3) Does not need upstream variable selection (4) Good for parallel processing	(1) Not as easy as decision tree to interpret the predictors (2) Multiple parameters to tune (number of features, number of trees, minimum sample leaf size)	44
Artificial neural networks ^a (model)		An interconnected structure of neurons for data classification and prediction using pattern recognition	Continuous and categorical	Linear and non-linear	(1) Deals with missing data (2) Parallel processing (3) Applies to a wide range of problems (eg, image, sound recognition, text, and time series)	(1) Computationally intensive to train, requiring processors with parallel processing power (2) Difficult to tune to ensure that the model learns well (3) Black box (does not explain the behavior of the network)	45
	Support vector machine	A classification to represent the data as points in space so that data of different categories are divided by a clear gap	Categorical	Linear and non-linear	(1) Linear classifier, and can be turned into a non-linear classifier (2) Handles high dimensional data well	(1) Choice of the penalty variable (2) Choice of the kernel	38
Classifiers (model)	Bayes classifier	A family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features	Categorical	Linear and non-linear	(1) Handles high dimensional data well (2) Simple and fast processing (3) Robustness (4) Incrementality	(1) No variable selection (2) No explicit model	38
	k-NN	A data classification approach that estimates how likely a data point is to be a member of one group	Categorical	Linear and non-linear	(1) Simplicity of implementation (2) Incrementality	(1) Slow processing (2) Difficult to tune to ensure that the model learns well (3) Variables should have similar scales (4) Difficult to interpret results	82

(Continues)

TABLE 1 (Continued)

Model	Description	Type of response variables	Linearity of the model	Strength	Weakness	Reference
Unsupervised learning						
Clustering (method)	A method to group a set of objects so that objects in the same group are more similar to each other than to those in other groups, for example, hierarchical clustering, distribution-based clustering, density-based clustering, and k-means clustering.	Not relevant	No prediction data	(1) An optimization process between intraclass variance and interclass variance	(1) Choice of the number of groups (2) Results may vary according to the algorithm that is used and the tuning of the method (3) Variables should have similar scales	41
Factor analysis (method)	A projection of information contained in a multidimensional space on a smaller space to study the similarities/differences between individuals and the relationships between the variables, for example, PCA and MCA	Not relevant	No prediction data	(1) Deals with a large amount of data	(1) If data transformation is not performed, variables should have similar scales	39

^aBasic model of the category, including classification and regression trees.

^bNon-linear problems may be transferred to linear problems.

^cArtificial neural networks can be supervised (eg, feed-forward back-propagation network and cascade correlation) and unsupervised (eg, autoencoder neural network and self-organizing map).

Machine learning algorithms are based on either supervised or unsupervised learning. Supervised learning uses a set of labeled examples as training data and makes predictions for all unknown points.³⁸ Models using supervised learning include regression models (eg, multiple linear regression, generalized linear model, regularized regression, PLS, and PCR), models based on decision trees (eg, gradient boosting tree and random forest), classifiers (eg, Bayes classifier, k-NN, and support vector machine), and some ANNs (eg, feed-forward back-propagation network and cascade correlation). According to the data type of the response variable, these models can be divided into those for continuous variables (eg, pollutant concentration) and categorical variables (eg, air quality indices). Models for continuous variables include multiple linear regression (MLR), regularized regression, PLS, and PCR. Models for categorical variables include Bayes classifier, k-NN, and support vector machine. Models for both types of variables include generalized linear models, decision trees, gradient boosting trees, random forests, and ANNs. These models can also be divided into linear and non-linear models to address linear and non-linear problems. When the response and predictive variables are linearly related or when they are transformed into a linear relationship, linear regression models can be used as suitable prediction models. When the scale of the multiple variables differs greatly, data transformation techniques including normalization, log transformation, and rank transformation can be employed. MLR is the classical regression model due to its simplicity and clear display of the best predictors of the variable of interest, although it does not deal with missing data, requires more observations than variables and risks multi-collinearity problems. These negative aspects of MLR can be well handled when using PLS and PCR, which group individual explanatory variables into components to reduce regression variables. Linear models are simple to develop, easy to use, and are often applied as the first attempt of prediction. When the response and predictive variables are unlikely to be linearly related, other non-linear models can be used more effectively regardless of the data structure.

Unsupervised learning uses unlabeled training data and aims at reducing, summarizing, and synthesizing data.³⁸ Although unsupervised learning cannot provide predictions of unknown data, it helps to understand the structural nature of the data, so that a supervised model can be chosen for the prediction. Unsupervised learning methods include association rule learning, clustering, factor analysis (eg, principal component analysis [PCA] and multiple correspondence analysis [MCA]), and some ANNs (eg, autoencoder neural network and self-organizing map). Among these models, PCA can address continuous variables, MCA can address categorical variables, and clustering can address both types of variables according to the metric used. In unsupervised methods, there is no output to predict, hence no input-output relationship. As such, data linearity is not an issue. However, transformed data will not get the same result as untransformed data and the distance metric used (eg, Euclidean space in PCA and chi-square metric for MCA) will also affect the overall interpretation of the results.

Besides the above-mentioned models, many other statistical models including structural equation models, time series models, duration models, and text mining have been applied in a wide range of fields, including economics, biology, medicine, manufacturing, finance, and marketing domains. These models are not discussed in detail in the present work.

To develop a statistical model to predict IAQ, the overall data set is usually split into three for the training, validation, and testing of the model. Usually about 70% of the data are used for training, while the other 30% are used for model validation (15%) and testing (15%).^{46,47} The most common validation algorithm is the leave-one-out cross-validation.⁴⁸ The model may also be tested using new data sets obtained in other cases than the original data set.⁴⁹ More detailed information on some statistical models that have been applied to the field of IAQ is presented below.

3.1.1 | Artificial neural networks

Artificial neural network is a popular data-driven method based on an interconnected structure of neurons (connected units or nodes).⁵⁰ It learns and predicts on an intuitional basis and is an effective method to solve non-linear problems.⁵¹ It uses a complex combination of weights and functions to convert input variables into predicted variables (outputs) without the need for pre-defined assumptions regarding the relationship among the variables.⁵²

An ANN structure generally consists of an input layer, several hidden layers, and an output layer (Figure 1). Each layer consists of one or more neurons, and each neuron is connected to other neurons of the previous layer through adaptable synaptic weights.⁵¹ During the learning stage, a data set of known inputs and outputs is fed to the model. The weights between the neurons are adjusted by iteration until an optimal solution of the outputs is reached. Then, the ANN model can be used for the prediction of the same outputs at a different time for the same case.

Depending on the structure and technique, ANN can be classified into many types, such as the multilayer perceptron neural network, feed-forward back-propagation neural network, recurrent neural network, and general regression neural network. The most frequently used networks in the literature were feed-forward neural networks. Among these networks, the two popular networks were multilayer perceptron and back-propagation neural networks, which address linearly separable data and non-linearly separable data, respectively.

Most of the studies used traditional training methods (eg, Levenberg-Marquardt and Bayesian regularization) to determine the weights between the neurons. These gradient descent methods may risk reaching a local optimum rather than a global optimum⁵³ not only for ANN, but also for other statistical models.

3.1.2 | Regressions

Regression models include MLR, kernel regression, and PLS, which estimate the relationships among variables. MLR is the basic and

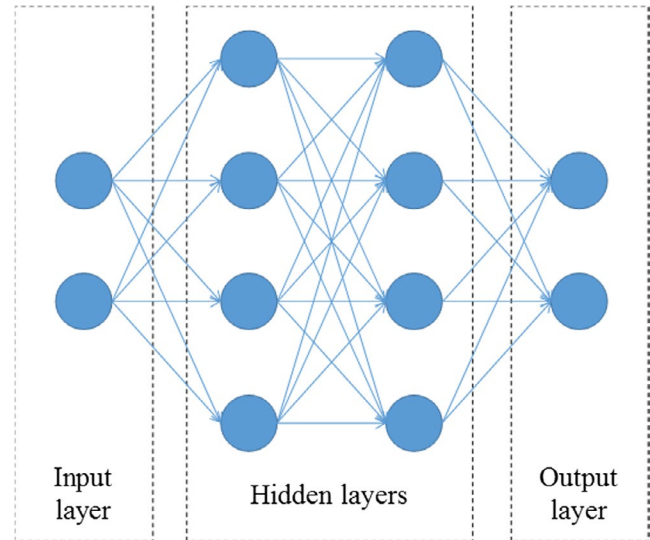


FIGURE 1 Schematic of a four-layer feed-forward artificial neural network

most widely used regression model to address the linear relationship between a response variable (output) and several predictive variables (inputs). An MLR model can be expressed as⁵⁴

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_ix_i + \varepsilon, \quad (1)$$

where y is the output variable, x is the input variable (from x_1 to x_i), a is the regression coefficient (from a_0 to a_i), and ε is the stochastic error associated with the regression.

When using MLR, the input variables should not be correlated with each other (multi-collinearity). A pre-analysis using data mining techniques, such as factor analyses, correlation analysis, and stepwise selection of variables, can identify and remove the multi-collinearity.⁵⁴⁻⁵⁶ Compared to MLR, other regression models including the least absolute shrinkage and selection operator (LASSO) regression³⁸ and stepwise regression⁵⁷ may better explore and select input variables.

Moreover, the PLS model projects input and output variables into a new space to obtain variance-covariance matrices in much the same way as PCA; however, the components are rotated to better explain the output variables.⁵⁸ Compared to MLR, PLS can reduce the dimension of the original variables while maintaining important information.⁵⁹ Cross-validation is used to determine the optimal number of components.

3.1.3 | Decision trees

A decision tree uses a tree-like model of decisions and their possible consequences for data classification or regression. A decision tree can be used for classification (classification tree) and prediction (regression tree). A single regression tree may not be able to predict complex problems and is less robust than a random forest regression (RFR). Therefore, RFR as an ensemble of regression trees is the most frequently used decision tree-based model.⁵⁹

Table 2 Summary of IAQ prediction studies using artificial neural networks

Reference	Data	Data transformation	Indoor type	Pre-analysis	Network	Inputs	Outputs	No. of hidden layers
83	BASE study: 100 office buildings from 10 geographic/climatic regions		Office	None	Feed-forward back-propagation network	Indoor TVOC, formaldehyde, CO ₂ , PM _{2.5} , airborne fungi and bacteria, temperature, relative humidity, light, and noise	Building-related symptom index	1
63	Daily average values in S subway station, Seoul, Korea. March 1, 2007-July 13, 2008	No	Subway station	PLS	Recurrent neural network	Indoor PM _{2.5} and PM ₁₀ of the day before, current indoor NO _x	Current daily indoor PM _{2.5}	1
84	BASE study: 100 office buildings from 10 geographic/climatic regions	No	Office	None	Feed-forward back-propagation network	Indoor temperature, relative humidity, air velocity, CO ₂ , TVOC, formaldehyde, PM _{2.5} , airborne fungi and bacteria	Indoor air quality index	2
85	8 apartments from four bedrooms and six living rooms in an apartment building located in Kuopio, Finland. May-October 2011	No	Apartment	None	Multilayer perceptron neural network	Indoor temperature and relative humidity	Indoor CO ₂	NA
86	150 workdays in Beijing campus, China. December 23, 2013-May 9, 2014	No	Office	None	Forward network	Outdoor air quality index, indoor air quality index, outdoor temperature, humidity, atmospheric pressure, and wind speed	Air purification time to reach the goal (indoor PM _{2.5} ≤ 35 µg/m ³ or reaches steady state)	1
46	500 Latin-Hypercube samples of the results predicted by the CONTAM model	No	Dwelling	Sensitivity analysis	Cascade forward network	Indoor temperature, generation rate of internal PM _{2.5} , kitchen window opening area	Indoor PM _{2.5}	2
52	Three buildings measured in 2010-2011	No	Office and shop	None	Feed-forward time-delay neural network	Time of day, barometer level pressure, sea level pressure, outdoor temperature, relative humidity, wind speed, wind direction, Pasquill atmospheric stability class, global solar radiation, outdoor NO ₂ and PM _{2.5}	Indoor NO ₂ and PM _{2.5}	1
47	The measurements were taken in each site for three consecutive days during school hours	No	School	Stepwise regression	Feed-forward back-propagation network	Outdoor and PM _{2.5-10} , indoor CO ₂ and relative humidity	PM _{2.5}	1

Network structure	Training algorithm	Transfer function for the hidden layer	Transfer function for the output layer	Percentage of data for training/ validation/ testing	PA	MSE	RMSE	R^2	NAE	IA
10-10-1	Back-propagation algorithm	Hyperbolic tangent function	NA	60%/ 20%/ 20%			8.4 (testing)			
3-5-1	Steepest gradient method	Sigmoid transfer function	NA	50% training/ 50% validation and testing			13.25 (training), 17.80 (testing)			
9-9-9-1	Back-propagation algorithm	Hyperbolic tangent function	NA	60%/ 20%/ 20%			7.2 (training), 8.8 (testing)	0.77 (training), 0.69 (testing)		
NA	Levenberg-Marquardt algorithm	Hyperbolic tangent function	Linear transfer function	NA			83.14-258.68	0.00-0.39		0.40-0.76
6-16-12	Back-propagation algorithm	NA	NA	NA	0.87					
3-6-6-1	Bayesian regularization training algorithm	NA	NA	70%/ 15%/ 15%		7.7 (testing)				
NA	Levenberg-Marquardt algorithm	Sigmoid transfer function	Linear transfer function	75%/ 10%/ 15%						
4-12-1	Back-propagation algorithm	Hyperbolic tangent sigmoid function (Tansig)	Linear transfer function (Purelin)	70%/ 15%/ 15%	0.9	1.4		0.75	0.18	0.91

(Continues)

TABLE 2 (Continued)

Reference	Data	Data transformation	Indoor type	Pre-analysis	Network	Inputs	Outputs	No. of hidden layers
87	1271 hourly data points, May 8, 2007-June 29, 2007	Normalization	Bus	Vector time series	Back-propagation neural network	Indoor and outdoor temperature and relative humidity, outdoor wind speed, passenger density, ventilation settings, precipitation, light vehicles, heavy vehicles	Indoor CO and CO ₂	1
88	Monthly data	Normalization	Dwelling	None	Self-organizing map	Indoor TVOC, benzene, toluene, and xylene	Indoor air pollution index (IAPvoc)	NA
89	Data measured over seven months at three different times of the day	No	NA	None	Gated recurrent unit network	Past indoor CO ₂ , fine dust, temperature, humidity, light quantity, and TVOC	Current indoor CO ₂ , fine dust, temperature, humidity, light quantity, and TVOC	2
90	Electronic nose	NA	NA	NA	Multilayer perceptron neural network	NA	Current indoor air quality	NA
64	261 measurements	Normalization	Ventilated room	None	Multilayer perceptron neural network	Diameter of the aerosol particle, room inlet air velocity, room geometry, coordinates of distance between the entrance and the exit of an area, density of air, particle stay time in the room	Normalized indoor particle concentration	1
91	249 measured data	No	Building	NA	General regression neural network	Indoor PM _{2.5} , PM ₁₀ , CO ₂ , temperature, and relative humidity	Indoor airborne bacteria	NA
50	Data for every minute, Mons, Belgium. February 11, 2015-February 18, 2015	No	Room	NA	Multilayer perceptron neural network	Indoor temperature and relative humidity	Indoor CO ₂	1
62	Six underground subway stations	Normalization	Subway station	NA	Feed-forward network	Outdoor PM ₁₀ , subway frequency and ventilation rate	Indoor PM ₁₀	1
62	Six underground subway stations	Normalization	Subway station	NA	Feed-forward network	Current outdoor PM ₁₀ , subway frequency, ventilation rate, indoor PM ₁₀ at the previous time	Current indoor PM ₁₀	1
51	Historical data from 5 d measurement consecutively, in which the measurement took 8 h/d	No	Office	NA	Feed-forward network	Indoor CO ₂ historical concentrations from Monday to Friday	Indoor CO ₂ concentrations of Monday and Tuesday of the next week	1

Network structure	Training algorithm	Transfer function for the hidden layer	Transfer function for the output layer	Percentage of data for training/ validation/ testing	PA	MSE	RMSE	R ²	NAE	IA
NA	Back-propagation algorithm	Hyperbolic tangent sigmoid function (Tansig)	Linear transfer function (Purelin)	92 (training)/ 8% (validation)						
NA	NA	NA	NA	NA						
6-1270-1270-6	Adaptive moment estimation optimization algorithm	Sigmoid activation function	NA	NA	0.85					
NA	NA	NA	NA	NA						
12-25-1	Fletcher conjugate gradient algorithm	Log-sigmoid transfer function	Linear function	80%/ 10%/ 10%			0.0003		0.01	
NA	NA	NA	NA	95% (training)/ 5% (testing)			412.7			
2-4-1	Levenberg-Marquardt algorithm	Tansig function	NA	70%/ 15%/ 15%		284.8				
NA	Back-propagation algorithm	Tangent sigmoid	Pure sigmoid	80% (training)/ 20% (validation)			19.08-59.01 (validation)	0.62 (validation)		
NA	Back-propagation algorithm	Tangent sigmoid	Pure sigmoid	80% (training)/ 20% (validation)			20.2-42.82 (validation)	0.7 (validation)		
5-10-2	Levenberg-Marquardt algorithm	NA	NA	NA		1924.3				

(Continues)

TABLE 2 (Continued)

Reference	Data	Data transformation	Indoor type	Pre-analysis	Network	Inputs	Outputs	No. of hidden layers
⁶¹	8761 data in the D-subway station, Seoul, South Korea, January-December 2009	No	Subway station	Kolmogorov-Smirnov test	Recurrent neural networks	Past indoor PM _{2.5}	Current indoor PM _{2.5}	2

Abbreviations: IA, index of agreement; MSE, mean squared error; NA, not available; NAE, normalized absolute error; PA, prediction accuracy; PLS, partial least squares; R^2 , coefficient of determination; RMSE, root mean square error; TVOC, total volatile organic compounds.

3.2 | Applications to indoor air quality

The application of statistical models for predicting the concentrations of indoor air pollutants, for example, CO₂, VOCs, and PM, is much less advanced compared to the applications to predict outdoor air pollutant concentrations. Statistical models can predict IAQ in an existing building based on questionnaires and/or measurements.⁶⁰ The 37 articles about statistical models applied to predict IAQ were published in the past decade, with a median publication date of 2016. Only four papers were published before 2010, while 33 were published after 2010. The predicted variables were indoor concentrations of PM (PM_{2.5} and PM₁₀), carbon dioxide and monoxide (CO₂ and CO), nitrogen oxides (NO_x), radon, airborne culturable bacteria, and IAQ indices consisting of various IAQ parameters. The studies were carried out in dwellings (n = 16), offices (n = 5), schools (n = 3), hospitals (n = 1), transportation infrastructures and vehicles (n = 9), and other types of buildings (n = 3). Twenty-eight studies used cross-sectional data for prediction and nine studies also used data at the previous sampling time as inputs. The models included ANN, regression models (eg, MLR, stepwise regression, PLS, and PCR), models based on decision trees, and PARAFAC. The most frequently used statistical models were ANN (n = 18), regression models (n = 19), and models based on decision trees (n = 4).

3.2.1 | Artificial neural networks

Artificial neural network is the most popular method for the prediction of IAQ, having been applied in 18 studies. A summary of IAQ prediction studies using ANN is provided in Table 2. The ANN models were developed for different sites (including dwellings, offices, schools, and subway stations) across the world to address several IAQ parameters using various techniques. When using cross-sectional data, the inputs and outputs are different variables at the same time. The input variables can be either continuous time-dependent variables, or continuous or categorical time-independent variables, such as the surface area of a dwelling and the nature of a building material. The output variables that were studied include PM_{2.5}, PM₁₀, CO, CO₂, NO₂, airborne bacteria, total VOCs (TVOC), and some IAQ indices consisting of many

IAQ parameters. The R^2 values for the available ANN models in the field of IAQ regardless of inputs and outputs vary between 0.65 and 0.79 for feed-forward back-propagation neural network, and between 0.39 and 0.62 for multilayer perceptron neural network. However, it remains unclear whether feed-forward back-propagation neural network has a higher performance than multilayer perceptron neural network for three reasons. First, the models were not developed based on the same data set. The R^2 value may be influenced by the structure of the data. Second, it was unclear in most of the available studies whether the R^2 values were calculated for the training set or validation set. Third, the R^2 value also depends on the number of observations. When using some data at the previous sampling time in addition to the current cross-sectional data, the inputs and outputs can be the same variables at different lag times (eg, the inputs include the variables at the previous sampling time, and the outputs are the variables at the current time). Several recurrent neural networks were studied to process the inputs for a time sequence. Loy-Benitez et al⁶¹ compared three types of recurrent neural networks (ie, standard recurrent neural networks, long short-term memory, and gated recurrent unit) to predict PM_{2.5} concentration in a subway station in Seoul. The gated recurrent unit model had the highest performance (R^2 : 0.65). Although these studies used data at the previous sampling time, the models did not train on the historical profiles of the variables and the objective remained to predict rather than forecast IAQ. ANN usually provided acceptable estimates of IAQ with only one hidden layer (R^2 : 0.62-0.79), while the number of neurons in the hidden layer depended largely on the number of inputs.⁴⁷

The PM concentration in indoor air was the most frequently studied parameter of the ANN applications indoors. The models have shown generally good performances (R^2 : 0.62-0.79, normalized absolute error: 0.01-0.19, index of agreement: 0.89-0.95). The inputs for the models are shown in Figure 2. Most of the prediction models used regional environmental variables including the outdoor PM concentration, temperature, and wind speed, as inputs. These outdoor variables can provide good performance explaining indoor air PM of outdoor origin. In a study of PM_{2.5} and PM_{2.5-10} concentrations in twelve naturally ventilated schools located in Palestine, the R^2 value was found to be between 0.65 and 0.79,⁴⁷

Network structure	Training algorithm	Transfer function for the hidden layer	Transfer function for the output layer	Percentage of data for training/ validation/ testing	PA	MSE	RMSE	R^2	NAE	IA
1-50-100-1	NA	NA	NA	70% (training)/ 30% (testing)			20.9	0.65		

indicating the robustness of the method when the main PM source is the outdoor environment. For other indoor environments, such as dwellings, where indoor sources are not negligible, models are scarce. Das et al proposed a few indoor variables, such as the generation rate of internal $PM_{2.5}$, to address the indoor source.⁴⁶ It should be noted that providing this kind of input is not easy and is critical to the prediction model. Therefore, further studies are needed to study the inputs and the performance of the models when strong indoor sources are present. When data from the previous sampling time were included in addition to the current data, models in the literature tended to employ indoor variables rather than outdoor variables as inputs. For example, a model using outdoor and indoor PM_{10} concentrations (1-minute average), train frequency, and ventilation rate at the previous sampling time as inputs performed well (R^2 : 0.7), predicting the current indoor PM_{10} concentration in a subway station.⁶² In addition to the indoor PM concentrations at the previous sampling time, some studies also used indoor temperature and NO_x concentration at the current time as inputs to predict indoor PM concentrations at the current time.⁶³ In these studies, data from the previous sampling time were simply used as normal input data rather than time series data. Sixteen parameters were only used once as inputs among all the studies for the model development. The literature did not disclose the reasons for the choice, and there was no evidence indicating that these parameters could better predict IAQ. They may simply be used as an attempt.

Since statistical models are data-driven, the most crucial issue for applying ANN in the field of IAQ and achieving a robust prediction is the selection of input variables. To search for patterns in the data and identify important variables, several studies performed pre-analyses before the modeling development using data mining techniques including stepwise regression, principal component analysis, and PLS.^{46,63} The pre-analysis can filter the variables that are not important for the output and reduce the complexity of the model. However, these analyses usually assume a linear relationship between inputs and outputs. For some studies, the value of input variables differed by several orders of magnitude, and the data were normalized in various ways to reflect the relative significance of the variables.⁶⁴

3.2.2 | Regressions

A summary of IAQ prediction studies using regression models is shown in Table 3. Some regression models, including MLR and stepwise regression models, have been applied to indoor PM and NO_2 in three types of buildings: schools, dwellings, and subway stations.^{47,55,65-67} Other regression models, including kernel regression and Bayesian spatial quantile regression, have been developed to predict indoor radon concentration at large scales in Switzerland and Italy.^{65,68} For a given environment, the model performance depends largely on the selection

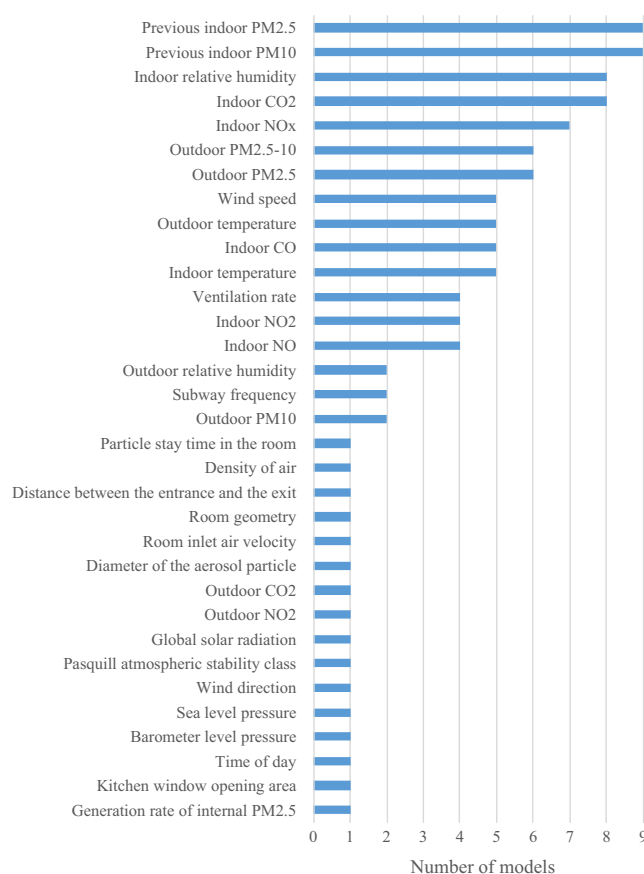


FIGURE 2 Inputs for the study of particulate matter concentrations in indoor air using artificial neural networks

Table 3 Summary of IAQ prediction studies using regression models

Reference	Data	Data transformation	Indoor type	Pre-analysis	Model
⁶³	Daily average values are measured in a subway station, Seoul, Korea. March 1, 2007-July 13, 2008	No	Subway station	PLS	MLR
⁶³	Daily average values are measured in a subway station, Seoul, Korea. March 1, 2007-July 13, 2008	No	Subway station	None	MLR
⁴⁷	Three consecutive days during school hours	Normalization	School	Stepwise regression	MLR
⁴⁷	Three consecutive days during school hours	Normalization	School	Stepwise regression	MLR
⁴⁷	Three consecutive days during school hours	Normalization	School	Stepwise regression	MLR
⁴⁷	Three consecutive days during school hours	Normalization	School	Stepwise regression	MLR
⁴⁷	Three consecutive days during school hours	Normalization	School	Stepwise regression	MLR
⁴⁷	Three consecutive days during school hours	Normalization	School	Stepwise regression	MLR
⁹²	RIOPA study. Summer 1999-spring 2001 spring. 48-h sampling in 374 non-smoking homes in Houston (TX), Los Angeles County (CA), and Elizabeth (NJ).	No	Dwelling	Stepwise regression	MLR
⁶⁹	One school situated in the high-traffic center of the city, one school located at the periphery of the city with mild traffic, and a rural school surrounded by undisturbed residential housing with gardens, fields and meadows about 20 km south-west of Prague.	No	School	None	MLR
⁵⁴	Twelve naturally ventilated schools in Gaza Strip (Palestine). October 2011-May 2012 (academic year)	Normalization	School	Stepwise regression	MLR
⁶⁰	Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study: 445 observations from 323 homes. January 2014-December 2015.	No	Dwelling	Variance inflation factor	MLR
⁶⁶	71 daily samples, Y-subway station online number three of Seoul Metro. October 2007-April 2008.	No	Subway station	None	MLR
⁶⁶	71 daily samples, Y-subway station online number three of Seoul Metro. October 2007-April 2008.	No	Subway station	FID	MLR
⁵⁵	Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study: 540 women, January 2014-December 2015.	No	Dwelling	Correlation analysis	MLR

Inputs	Outputs	Percentage of data for training/ validation/testing	R ²	PA	RMSE	NAE	IA
Indoor PM _{2.5} and PM ₁₀ of the day before, current indoor NO _x	Current indoor PM _{2.5} and PM ₁₀	50% training/ 50% validation and testing			12.7 (training), 18.06 (testing)		
Indoor PM _{2.5} and PM ₁₀ of the day before, current indoor NO, NO ₂ , NO _x , CO, CO ₂ , temperature, and humidity	Current indoor PM _{2.5} and PM ₁₀	50% training/ 50% validation and testing			12.22 (training), 20.74 (testing)		
Outdoor PM _{2.5} , indoor relative humidity	Indoor PM _{2.5}	70%/ 15%/ 15%	0.58	0.77	1.7	0.33	0.86
Outdoor PM _{2.5} and PM _{2.5-10} , ventilation rate, wind speed, indoor temperature	Indoor PM _{2.5}	70%/ 15%/ 15%	0.69	0.85	1.3	0.38	0.87
Outdoor PM _{2.5} , PM _{2.5-10} and CO ₂	Indoor PM _{2.5}	70%/ 15%/ 15%	0.7	0.85	1.4	0.23	0.88
Outdoor PM _{2.5} and PM _{2.5-10} , indoor relative humidity	Indoor PM _{2.5-10}	70%/ 15%/ 15%	0.44	0.66	2.1	0.26	0.75
Outdoor PM _{2.5-10} and temperature, indoor relative humidity	Indoor PM _{2.5-10}	70%/ 15%/ 15%	0.56	0.75	1.4	0.38	0.83
Outdoor PM _{2.5} , PM _{2.5-10} , and relative humidity	Indoor PM _{2.5-10}	70%/ 15%/ 15%	0.57	0.75	2.1	0.29	0.83
Outdoor PM _{2.5} , indoor tobacco, sweeping incense, sander chainsaw, cooking, diesel vehicle parked	Indoor PM _{2.5}	NA	0.35				
Outdoor PM _{2.5-10} , number of person-hours per day	Indoor PM _{2.5-10}	NA	0.71				
Outdoor PM _{2.5} and PM ₁₀ , indoor temperature, ventilation rate	Indoor PM _{2.5} and PM ₁₀	70% training/ 30% validation	0.84				
87 potential predictor variables (3 for outdoor PM _{2.5} , three for intervention status, three from housing assessment data, 15 from questionnaire data, five from meteorological data, and 58 from geographic data)	Indoor PM _{2.5}	NA	0.525 (training), 0.505 (validation)		0.48 (training), 0.49 (validation)		0.82 (training), 0.80 (validation)
Past indoor PM ₁₀ and PM _{2.5} , current outdoor PM ₁₀ , temperature, humidity, wind speed, the number of passengers in the subway	Current indoor PM _{2.5} and PM ₁₀	63% training/ 37% testing			5.41 (training), 71.16 (testing)		
Past indoor PM ₁₀ , current outdoor PM ₁₀ , the number of passengers	Current indoor PM _{2.5} and PM ₁₀	63% training/ 37% testing			22.34 (training), 47.77 (testing)		
Season, outdoor PM _{2.5} , number of air cleaners deployed, passenger density	Indoor PM _{2.5}	NA	0.524 (training), 0.502 (validation)		0.48 (training), 0.48 (validation)		0.82 (training), 0.8 (validation)

(Continues)

TABLE 3 (Continued)

Reference	Data	Data transformation	Indoor type	Pre-analysis	Model
56	A real-time 24-h IAQ monitoring was conducted at each of the 10 household between March and July 2014 in India	No	Dwelling	Stepwise regression	MLR
65	238769 indoor radon concentration measurements carried out in 148458 houses in Switzerland	NA	Dwelling	NA	Kernel regression
49	CO ₂ measurements are constantly transmitted through internet	Normalization	NA	NA	Kernel regression
93	PM _{2.5} sampled in a subway station in Seoul, January 4-26, 2010, with a sampling time interval of 1 h	NA	Subway station	NA	Least squares support vector regression
68	Annual average radon concentrations measured in almost 2400 buildings, mainly homes, mostly at ground level in Abruzzo since early 90s	NA	All types	Stepwise analysis	Bayesian spatial quantile regression
67	105 homes in the Durban metropolitan area, South Africa	Log transformation	Dwelling	None	Stepwise regression
67	82 homes in the Durban metropolitan area, South Africa	Log transformation	Dwelling	None	Stepwise regression
94	PM _{2.5} concentrations in both indoor and outdoor air were simultaneously measured using online particulate counters in 13 households in Beijing, China	Log transformation	Dwelling	None	Exponential regression
94	PM _{2.5} concentrations in both indoor and outdoor air were simultaneously measured using online particulate counters in 13 households in Beijing, China	Log transformation	Dwelling	None	Exponential regression
48	Airborne PM _{2.5} was collected over 48 h in 23 residential environments (single family and apartment buildings) in the Cincinnati-Kentucky-Indiana tristate region from September 2015 through August 2017	No	Dwelling	None	Linear regression
95	Seoul subway station. March 2008-February 2009.	No	Subway station	None	PLS
58	Platform and waiting room were measured at three min intervals. November 21-25, 2011	No	Subway station	None	PLS

Inputs	Outputs	Percentage of data for training/ validation/testing	R ²	PA	RMSE	NAE	IA
Indoor PM ₁₀ , PM _{2.5} , CO ₂ , temperature, relative humidity	Indoor PM ₁	NA	0.81-0.98				0.92-0.99
Coordinates, detector type, floor level and inhabitation, year of construction, building type, foundation type, altitude, outdoor temperature, lithology	Indoor radon concentration	NA	0.78				
Indoor temperature, humidity, light, the time of measurement	Indoor CO ₂	70% training/ 30% validation/ new data set testing	0.95 (training), 0.88 (validation), 0.79 (testing)				
Indoor NO, NO ₂ , NO _x , PM ₁₀ , CO, CO ₂ , temperature and relative humidity	Indoor PM _{2.5}	82% training/ 18% testing	0.81		6.32		
Building characteristics	Indoor radon concentration	NA					
Type of housing structure, total number of rooms in the household, type of primary cooking fuel, season, number of household smokers	Indoor PM ₁₀	NA	0.33 (validation)		0.54 (validation)		
Distance from the roadway, type of housing structure, type of primary cooking fuel, use of secondary cooking fuel, burning of incense in the house, season	Indoor NO ₂	NA	0.28 (validation)		0.45 (validation)		
Outdoor PM _{2.5} with 10 min of lag time	Indoor PM _{2.5}	NA	0.867				
Outdoor PM _{2.5} with 80 min of lag time	Indoor PM _{2.5}	NA	0.861				
Outdoor black carbon in PM _{2.5} , air infiltration, presence of HVAC filter, window opening, candles	Indoor black carbon in PM _{2.5}	NA	0.76 (validation)				
Current indoor NO ₂ , temperature and relative humidity, previous day data of indoor PM _{2.5} and PM ₁₀	Current indoor PM _{2.5} and PM ₁₀	NA			14.71		
Outdoor PM ₁₀ , waiting room PM ₁₀ , fan speed of the ventilation system, subway schedule, and time	Indoor PM ₁₀	69% training, 31% testing			7.87		

(Continues)

TABLE 3 (Continued)

Reference	Data	Data transformation	Indoor type	Pre-analysis	Model
⁶⁶	71 daily samples, Y-subway station online number three of Seoul Metro. October 2007-April 2008.	No	Subway station	FID	PLS

Abbreviations: FID, Fisher's linear discriminant; IA, index of agreement; MLR, multiple linear regression; NA, not available; NAE, normalized absolute error; PA, prediction accuracy; PLS, partial least squares; R^2 , coefficient of determination; RMSE, root mean square error.

of inputs. For example, the R^2 value of an MLR model for the prediction of $PM_{2.5}$ in a school for three consecutive days during school hours was 0.58 using outdoor $PM_{2.5}$ concentration and indoor relative humidity as inputs.⁴⁷ The R^2 value increased to 0.69 when the inputs were the concentrations of outdoor $PM_{2.5}$ and $PM_{2.5-10}$, ventilation rate, wind speed, and indoor temperature. This is because ventilation rate can strongly affect indoor/outdoor transport of PM. A prediction study of indoor $PM_{2.5}$ concentration compared MLR, LASSO regression, and stepwise regression for the same data set.⁵⁵ The LASSO and stepwise regressions had better performance, indicated by R^2 and root mean square error (RMSE), than MLR during the training period, due to better variable selection procedures. However, the three regression models showed similar performance during the validation period.

Unlike the ANN model, for which the weights of input variables are hidden, those for an MLR model can be quantified and expressed as regression coefficients in Equation 1. Outdoor PM concentrations are the main variables that affect the indoor PM concentrations for the schools and subway stations. The indoor PM concentrations were all positively correlated with the outdoor PM concentrations except for one school situated in a rural area far away from the traffic,⁶⁹ where the variation in indoor PM concentrations may be caused by student activities. Following the outdoor PM concentrations, the person-hours per day was also an important variable for indoor PM concentrations for schools because student activity is an important source of indoor PM.⁷⁰

Since MLR is the most widely used regression model, some studies used MLR as the basic model and compared other prediction models with MLR for the same data set. For the prediction of indoor PM concentrations, two studies have developed both MLR and ANN models.^{47,63} A comparison of the RMSE between the MLR and ANN models for the same studies is shown in Figure 3. The RMSE values for MLR during the training period are generally higher than those for ANN, probably because the relationship between the inputs and outputs tends to be non-linear. The RMSE values for both MLR and ANN increase substantially and are similar during the validation and testing periods.

A summary of IAQ prediction studies using PLS is shown in Table 3. The three studies were all carried out in Seoul subway stations and aimed to predict $PM_{2.5}$ and PM_{10} concentrations in

the air. The models relied on outdoor PM concentrations and indoor PM concentrations at the previous sampling time. The input variables were similar to those for the ANN models. For the prediction of PM concentrations in a subway station, both MLR and PLS models were used.⁶⁶ A comparison of the RMSE between the MLR and PLS models shows that the RMSE values for the PLS model are higher than those for the MLR for training, but lower for testing (Figure 3). The RMSE values for training and testing are similar, indicating the robustness of the PLS models for the studied cases.

3.2.3 | Decision trees

A summary of IAQ prediction studies using models based on decision trees is shown in Table 4. The four studies were carried out in dwellings and an emergency room and aim to predict $PM_{2.5}$, radon, and virus in the air. The models relied on outdoor PM concentrations for $PM_{2.5}$ predictions,⁵⁵ lithological units for radon predictions,⁵⁹ and fine dust particle for virus predictions.⁷¹ The R^2 values are 0.74-0.94 for training and 0.33-0.49 for validation. The decrease of R^2 values for validation indicates that for the studied cases, the well-trained model may need improvements for further predictions. Similar gaps between the training and validation sets are also seen for other performance metrics (eg, RMSE and index of agreement). For two predictions of $PM_{2.5}$ concentrations in dwellings, RFR models were compared with MLR,^{55,60} which showed that the RMSE values for the RFR models were less than those for the MLR for training. For validation, the RMSE values for the RFR models largely increase (Figure 3), indicating that the RFR models may be less robust than the MLR models for the studied cases.

3.2.4 | Other models

Besides the four main types of models, other statistical models were also applied in IAQ studies including PCR⁵⁴ and PARAFAC.⁷² The PCR model consisting of a PCA followed by a regression was used to predict indoor PM concentrations in schools. The PARAFAC model is a generalization of PCA to higher order arrays, which was used to predict indoor concentrations of PM, NO_x , CO, and CO_2 . The input variables for the PCR and PARAFAC models were consistent with those for the PLS models.

Inputs	Outputs	Percentage of data for training/validation/testing	R^2	PA	RMSE	NAE	IA
Past indoor PM_{10} and $PM_{2.5}$, current outdoor PM_{10} , temperature, humidity, wind speed, the number of passengers in the subway	Current indoor $PM_{2.5}$ and PM_{10}	63% training/ 37% testing			17.23 (training), 21.53 (testing)		

4 | DISCUSSION

4.1 | Time series models

The previously described models represent some statistical models for IAQ prediction. Although some of these applications also used data at the previous sampling time as inputs, they did not perform time series analysis and did not provide forecasting of a parameter based on its historical profile.

Time series models based on autoregression are specialized to perform time series analysis. The most commonly used linear models include the autoregressive model, autoregressive moving average model, and autoregressive integrated moving average model. The most commonly used non-linear models include autoregressive conditional heteroskedasticity model and generalized autoregressive conditional heteroskedasticity model.⁷³ Since the forecasting of a time series model is based on the historical profile of the target IAQ parameter, it requires continuous sensing technologies and online measurements. Time series models have been applied to many cases of outdoor environment^{74,75} and building energy studies.³⁵ In indoor environments, time series models have also been applied to forecast temperature,⁷⁶ relative humidity,^{76,77} CO_2 concentration,⁷⁶⁻⁷⁸ and CO concentration.⁷⁸ The four parameters can be measured continuously using low-cost online sensors.

Formaldehyde concentration was forecasted using a non-linear time series hybrid model based on spectral band decomposition coupled with a threshold autoregressive model.⁷⁹ The hybrid model was used to forecast indoor formaldehyde concentrations. The study was carried out in an open-plan office occupied by 6-8 people where formaldehyde was measured continuously. The only input for the model was the historical formaldehyde concentration in the office. The model analyzed the fluctuation pattern of the historical formaldehyde concentration and provided a forecast without any additional input variable. The historical formaldehyde concentration was first decomposed using fast Fourier transform (FFT) into several FFT components filtered by selecting specific cutoff frequencies. Two non-linear time series models, that is, a threshold autoregressive (TAR) model and a Chaos dynamics-based model, were employed after the FFT for the forecasting. Since the model completely depends on the historical concentration at the site, a large amount of historical data is needed for the training of

the model to capture all the fluctuating features of the formaldehyde concentration and achieve an acceptable forecasting performance. The hybrid model was based on 20 000 observations (1 minute intervals) during 14 continuous days for a forecast horizon of 12 hours ahead.

4.2 | Perspectives on statistical models and IAQ prediction

Compared to mechanistic IAQ models, where the nature of the relationship between inputs and outputs is specified in terms of underlying mechanisms, statistical models seek an optimal relationship between inputs and outputs to describe the observed data. Mechanistic models have advantages in understanding the physical and chemical phenomena of target pollutants in the indoor environment while statistical models are developed based on data sets consisting of between dozens and thousands of different indoor environments. Due to the nature of data mining, machine learning, and statistics, statistical models can efficiently handle studies carried out in multiple indoor environments. So far, there have been no studies that compare the predictions of mechanistic and statistical models in an indoor environment. Considering the state of the art of the two types of models, a possible comparison may be carried out in an unoccupied room where detailed information on the room and its interior materials is obtained, and indoor and outdoor environmental parameters are monitored for the models. Moreover, in an occupied environment, occupant activities can affect environmental parameters and pollutant concentrations at specific times and for varying durations. To date, this interaction between human and environment is not easy to predict using mechanistic models, while statistical models can take into account this interaction by integrating occupant-specific variables among other building- or environment-related parameters as predictors of IAQ. Therefore, a combination of mechanistic and statistical models may be developed to predict IAQ in an occupied building, where the mechanistic model addresses the physical and chemical phenomena, and the statistical model addresses the human behavior.

The number of applications of machine learning and statistics in the field of IAQ is to date relatively low compared to their applications in outdoor environment and energy efficiency. Nevertheless, the use of ANN,⁴⁷ MLR,⁵⁶ and PLS⁵⁸ models for

TABLE 4 Summary of IAQ prediction studies using models based on decision trees

Reference	Data	Data transformation	Indoor type	Pre-analysis	Model
71	NA	NA	Emergency room	NA	Decision trees
60	Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study: 445 observations from 323 homes. January 2014-December 2015.	No	Dwelling	None	RFR
59	72460 indoor radon measurements in 63076 buildings in Switzerland.	NA	NA	k-medoids clustering	RFR
55	Ulaanbaatar Gestation and Air Pollution Research (UGAAR) study: 540 women. January 2014-December 2015.	No	Dwelling	Correlation analysis	RFR

Abbreviations: IA, index of agreement; MSE, mean squared error; R^2 , coefficient of determination; RFR, random forest regression; RMSE, root mean square error.

the prediction of indoor PM concentrations showed acceptable performances in some cases, provided adequate input variables were selected to explain the variations in the response variables.⁴⁷ Statistical models for other IAQ parameters, including VOCs and aldehydes, remain scarce. No model has been developed for indoor SVOC prediction.

Addressing the following aspects may improve the performance of the statistical models to predict IAQ. First, many statistical methods have been developed and may be applied in the field of IAQ. Each of them has positive and negative aspects. To apply these methods for IAQ prediction, the first challenge encountered is the selection of the appropriate methods for the target case. Statistical models can be generally classified into two categories, linear and non-linear models, addressing linear and non-linear problems. Non-linear problems may be transformed into linear problems after suitable data transformations. Besides the inter-model difference, a single model type may include several learning algorithms, each with strengths and weaknesses, leading to difficulties for selection of the most suitable method. A common solution is to make different assumptions and test several methods. An alternative approach is to train several different models and combine the prediction results into a more robust overall estimate. For future studies, the parsimony principle should be used with the best model always being the simplest possible and using the least required inputs. The performance of a statistical model is usually evaluated by comparing the measured and predicted results using some performance metrics, including prediction accuracy, RMSE, coefficient of determination, mean absolute percent error, BIC (Bayesian information criterion), AIC (Akaike information criterion), and normalized absolute error. The

choice and interpretation of the performance metrics were rather subjective in the literature obtained for the present review. There was no consensus on the performance metrics that should be used or the minimum criteria required to evaluate the performance of a model. It can also depend on the application and objective of the prediction. Moreover, these metrics do not always show the same results regarding the performance of the models, which leads to difficulties in comparison.

Second, because there is no mechanistic relationship between inputs and outputs, the capacity of a model to explain the variance in the outputs depends on selecting appropriate input variables. Current studies frequently select common indoor variables (eg, CO₂ and PM concentrations) and local environmental variables (eg, wind speed, outdoor temperature, and outdoor relative humidity). Then these variables are analyzed and ranked using data mining techniques. Inputs including indoor volume and surface area, pollutant emission rate from indoor sources, and infiltration coefficients for outdoor/indoor pollutant transfer have rarely been considered in the existing applications. There is therefore a need to improve the availability of more specific data by developing databases on emission rates, deposition velocities from various indoor sources and pollutants, as well as outdoor/indoor transfer coefficients.

Third, because statistical models are trained based on existing data, they are applicable for use in buildings where conditions are similar to those from which the data for training were obtained.⁵² To build a statistical model for a group of environments, the training set must be representative of the type of these environments.

In summary, statistical methods are practically useful when all possible relationships of the variables that could affect the target

Inputs	Outputs	MSE	RMSE	R^2	IA
Find dust particle number	Types of virus in the particles				
87 potential predictor variables (3 for outdoor $PM_{2.5}$, three for intervention status, three from housing assessment data, 15 from questionnaire data, five from meteorological data, and 58 from geographic data)	Indoor $PM_{2.5}$	0.16 (training), 0.27 (validation)	0.4 (training), 0.52 (validation)	0.744 (training), 0.478 (validation)	0.86 (training), 0.73 (validation)
Building type, foundation type, year of construction, detector type, coordinates, temperature, altitude, clustered lithological units	Indoor radon			0.33	
Season, outdoor $PM_{2.5}$, number of air cleaners deployed, passenger density	Indoor $PM_{2.5}$		0.22 (training), 0.51 (validation)	0.938 (training), 0.489 (validation)	0.96 (training), 0.78 (validation)

pollutant are unknown. Most of the time, this happens when the activities of the occupants can affect the concentration of pollutants, which is almost always the case in real occupied environments. No mechanistic model can determine the existing relationships between occupant activities and the indoor concentration of pollutants unless assumptions are made and pre-designed activity schedules are provided, that is defining a scenario for the occupants. But because statistical models rely on existing data measurements, they cannot be applied to predict indoor air quality in buildings that have not been constructed if a representative database does not exist. In that case, only mechanistic models can prove useful. Whenever IAQ measurements within a large portion of the building stock are achieved, statistical predicting models might also be used for this objective.

Mechanistic models remain a reference to identify the underlying mechanisms and estimate their relative contribution, particularly in model environments and unoccupied situations. They still rely on the available knowledge and databases for specific parameters that remain scarce considering the large number of materials and products used in indoor environments. They can also help to identify key variables of interest to consider in statistical models. Future research will need to combine both statistical and mechanistic models in the hope of predicting IAQ more efficiently in real occupied environments, and “theory-guided data science”⁸⁰ may be useful for this purpose.

5 | CONCLUSIONS

Studies on IAQ based on large data sets from indoor environments worldwide have led to research on statistical methods for the analysis

of observed data. The present study carried out a review of the state of the art of statistical models for IAQ prediction and identified strengths and weaknesses of current applications. Several statistical models

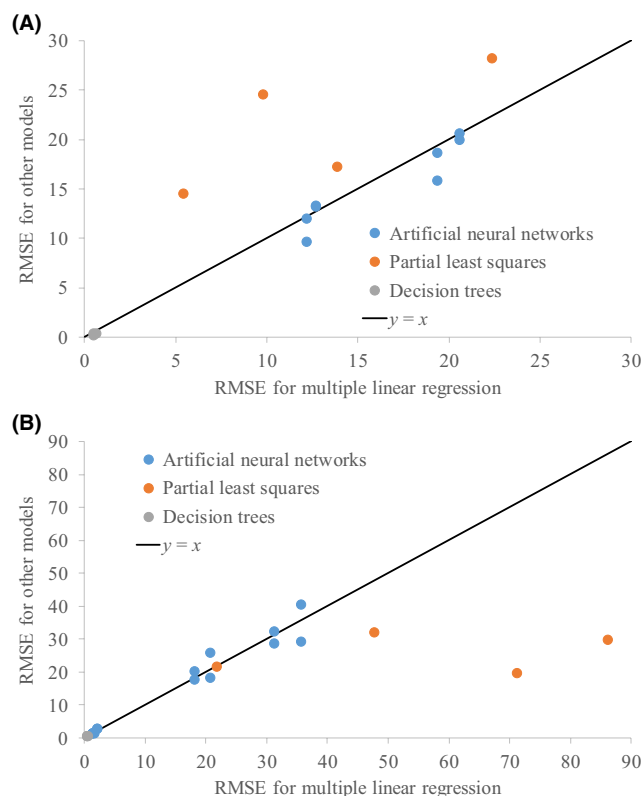


FIGURE 3 Comparison of the root mean square error (RMSE) between the statistical models for predicting the airborne particle concentration ($\mu\text{g}/\text{m}^3$). (A) training, (B) validation and testing

including ANN, linear regression models, and decision tree-based models have been developed in the past decade, focusing on predicting the concentrations of airborne particles, CO, CO₂, NO₂, and radon in indoor environments. Predictions of indoor aldehydes, VOCs, and SVOCs using statistical models remain scarce. Studies of these chemical compounds can be envisaged with the help of low-cost sensors in the future. Moreover, to develop a statistical model for a data set, it is advisable to test and compare different models including ANN, PLS, RFR, and MLR, before choosing the most suitable model for the specific case. Robust predictions arise often by averaging the outputs of different models. Although studies of IAQ in the literature have not yet used this approach, it should be done in future.

ACKNOWLEDGEMENTS

The work received research funding from the Scientific and Technical Center for Building (CSTB) program on Health and Comfort in buildings (Grant SEC21). John Little thanks CSTB and the University of La Rochelle for the grant received to support his contribution to this work.

ORCID

Wenjuan Wei  <https://orcid.org/0000-0002-8232-8186>

John C. Little  <https://orcid.org/0000-0003-2965-9557>

Corinne Mandin  <https://orcid.org/0000-0001-8462-8812>

REFERENCES

- Liu Z, Ye W, Little JC. Predicting emissions of volatile and semivolatile organic compounds from building materials: a review. *Build Environ*. 2013;64:7-25.
- Little JC, Hodgson AT, Gadgil AJ. Modeling emissions of volatile organic compounds from new carpets. *Atmos Environ*. 1994;28:227-234.
- Mendez M, Blond N, Blondeau P, Schoemaeker C, Hauglustaine DA. Assessment of the impact of oxidation processes on indoor air pollution using the new time-resolved INCA-Indoor model. *Atmos Environ*. 2015;122:521-530.
- Carslaw N. A new detailed chemical model for indoor air pollution. *Atmos Environ*. 2007;41:1164-1179.
- Liu C, Zhang Y, Benning JL, Little JC. The effect of ventilation on indoor exposure to semivolatile organic compounds. *Indoor Air*. 2015;25:285-296.
- Xu Y, Hubal E, Clausen PA, Little JC. Predicting residential exposure to phthalate plasticizer emitted from vinyl flooring: a mechanistic analysis. *Environ Sci Technol*. 2009;43:2374-2380.
- Wei W, Ramalho O, Mandin C. A long-term dynamic model for predicting the concentration of semivolatile organic compounds in indoor environments: application to phthalates. *Build Environ*. 2019;148:11-19.
- Wei W, Howard-Reed C, Persily A, Zhang Y. Standard formaldehyde source for chamber testing of material emissions: model development, experimental evaluation, and impacts of environmental factors. *Environ Sci Technol*. 2013;47:7848-7854.
- Bourdin D, Mocho P, Desauziers V, Plaisance H. Formaldehyde emission behavior of building materials: on-site measurements and modeling approach to predict indoor air pollution. *J Hazard Mater*. 2014;280:164-173.
- Panagopoulos IK. A CFD simulation study of VOC and formaldehyde indoor air pollution dispersion in an apartment as part of an indoor pollution management plan. *Aerosol Air Qual Res*. 2011;11:758-762.
- Dimitroulopoulou C, Ashmore MR, Byrne MA, Kinnorsley RP. Modelling of indoor exposure to nitrogen dioxide in the UK. *Atmos Environ*. 2001;35:269-279.
- Van Hooff T, Blocken B. CFD evaluation of natural ventilation of indoor environments by the concentration decay method: CO₂ gas dispersion from a semi-enclosed stadium. *Build Environ*. 2013;61:1-17.
- Jelle BP. Development of a model for radon concentration in indoor air. *Sci Total Environ*. 2012;416:343-350.
- Kumar A, Chauhan RP, Joshi M, Sahoo BK. Modeling of indoor radon concentration from radon exhalation rates of building materials and validation through measurements. *J Environ Radioact*. 2014;127:50-55.
- Chen F, Yu S, Lai A. Modeling particle distribution and deposition in indoor environments with a new drift-flux model. *Atmos Environ*. 2006;40:357-367.
- Schneider T, Alstrup Jensen K, Clausen PA, et al. Prediction of indoor concentration of 0.5-4 µm particles of outdoor origin in an uninhabited apartment. *Atmos Environ*. 2004;38:6349-6359.
- Goyal R, Khare M. Indoor air quality modeling for PM₁₀, PM_{2.5}, and PM_{1.0} in naturally ventilated classrooms of an urban Indian school building. *Environ Monit Assess*. 2011;176:501-516.
- Tran DT, Alleman LY, Coddeville P, Galloo J-C. Indoor particle dynamics in schools: determination of air exchange rate, size-resolved particle deposition rate and penetration factor in real-life conditions. *Indoor Built Environ*. 2017;26:1335-1350.
- Hussein T, Kulmala M. Indoor aerosol modeling: basic principles and practical applications. *Water, Air, Soil Pollut Focus*. 2008;8:23-34.
- Ausati S, Amanollahi J. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting PM_{2.5}. *Atmos Environ*. 2016;142:465-474.
- Niu M, Wang Y, Sun S, Li Y. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM_{2.5} concentration forecasting. *Atmos Environ*. 2016;134:168-180.
- Sousa S, Martins F, Alvimferraz M, Pereira M. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environ Model Softw*. 2007;22:97-103.
- Strebel K, Espinosa G, Giralt F, et al. Modeling airborne benzene in space and time with self-organizing maps and Bayesian techniques. *Environ Model Softw*. 2013;41:151-162.
- Thomas B, Soleimani-Mohseni M. Artificial neural network models for indoor temperature prediction: investigations in two buildings. *Neural Comput Appl*. 2006;16:81-89.
- Patil SL, Tantau HJ, Salokhe VM. Modelling of tropical greenhouse temperature by auto regressive and neural network models. *Biosyst Eng*. 2008;99:423-431.
- He F, Ma C. Modeling greenhouse air humidity by means of artificial neural network and principal component analysis. *Comput Electron Agric*. 2010;71:S19-S23.
- Tsanas A, Xifara A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy Build*. 2012;49:560-567.
- Edwards RE, New J, Parker LE. Predicting future hourly residential electrical consumption: a machine learning case study. *Energy Build*. 2012;49:591-603.
- Chou J-S, Ngo N-T. Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. *Appl Energy*. 2016;177:751-770.

30. Bellinger C, Mohamed Jabbar MS, Zaiane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*. 2017;17:907.
31. Zhao H, Magoulès F. A review on the prediction of building energy consumption. *Renew Sustain Energy Rev*. 2012;16:3586-3592.
32. Cheng C, Sa-Ngasoongsong A, Beyca O, et al. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. *IIE Trans*. 2015;47:1053-1071.
33. Fu T. A review on time series data mining. *Eng Appl Artif Intell*. 2011;24:164-181.
34. Singh P. A brief review of modeling approaches based on fuzzy time series. *Int J Mach Learn Cybern*. 2017;8:397-420.
35. Deb C, Zhang F, Yang J, Lee SE, Shah KW. A review on time series forecasting techniques for building energy consumption. *Renew Sustain Energy Rev*. 2017;74:902-924.
36. Wang Q, Li S, Li R. Forecasting energy demand in China and India: using single-linear, hybrid-linear, and non-linear time series forecast techniques. *Energy*. 2018;161:821-831.
37. Majidpour M, Qiu C, Chu P, Gadh R, Pota HR. Fast prediction for sparse time series: demand forecast of EV charging stations for cell phone applications. *IEEE Trans Ind Informatics*. 2015;11:242-250.
38. Mohri M, Rostamizadeh A, Talwalkar A. *Foundation of Machine Learning*. (Dietterich T, ed.). Cambridge, MA: MIT Press; 2012.
39. Lebart L, Morineau A, Piron M. *Statistique Exploratoire Multidimensionnelle*. Paris: Dunod; 1995.
40. Esposito Vinzi V, Chin WW, Henseler J, Wang H. *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Heidelberg, Dordrecht, London, New York: Springer; 2010.
41. Aggarwal CC, Reddy CK. *Data Clustering: Algorithms and Applications*. London: Chapman and Hall; 2014.
42. McCullagh P, Nelder JA. *Generalized Linear Models*, 2nd edn. London, New York: Chapman and Hall; 1983.
43. Scott MA, Simonoff JS, Marx BD. *The SAGE Handbook of Multilevel Modeling* (Seaman J, ed.). London, Thousand Oaks, New Delhi, Singapore: SAGE; 2013.
44. Rokach L, Maimon O. *Data Mining with Decision Trees: Theory and Applications*, 2nd edn. Singapore: World Scientific Publishing Co., Pte. Ltd; 2015.
45. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach* (Pompli M, ed.). Upper Saddle River, NJ: Prentice-Hall, Inc; 1995.
46. Das P, Shrubsole C, Jones B, et al. Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling. *Build Environ*. 2014;78:171-182.
47. Elbayoumi M, Ramli NA, Fitri Md Yusof NF. Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM_{2.5}-10 and PM_{2.5} concentrations in naturally ventilated schools. *Atmos Pollut Res*. 2015;6:1013-1023.
48. Isiugo K, Jandarov R, Cox J, et al. Predicting indoor concentrations of black carbon in residential environments. *Atmos Environ*. 2019;201:223-230.
49. Carlos G, Valeria F, Guillermo V. Use of non-industrial environmental sensors and machine learning techniques in telemetry for indoor air pollution. *ARNP J Eng Appl Sci*. 2018;13:2702-2712.
50. Khazaei B, Shiehbeigi A, Haji Molla Ali Kani AR. Modeling indoor air carbon dioxide concentration using artificial neural network. *Int J Environ Sci Technol*. 2019;16:729-736.
51. Putra J, Safrilah MI. The prediction of indoor air quality in office room using artificial neural network. In: AIP Conference Proceedings 1977;020040. <http://aip.scitation.org/doi/abs/10.1063/1.5042896>
52. Challoner A, Pilla F, Gill L. Prediction of indoor air exposure from outdoor air quality using an artificial neural network model for inner city commercial buildings. *Int J Environ Res Public Health*. 2015;12:15233-15253.
53. Zhang T, You X. Improvement of the training and normalization method of artificial neural network in the prediction of indoor environment. *Proc Eng*. 2015;121:1245-1251.
54. Elbayoumi M, Ramli NA, Md Yusof N, Yahaya A, Al Madhoun W, UI-Saufie AZ. Multivariate methods for indoor PM₁₀ and PM_{2.5} modelling in naturally ventilated schools buildings. *Atmos Environ*. 2014;94:11-21.
55. Yuchi W, Gombojav E, Boldbaatar B, et al. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environ Pollut*. 2019;245:746-753.
56. Sharma D, Jain S. Impact of intervention of biomass cookstove technologies and kitchen characteristics on indoor air quality and human exposure in rural settings of India. *Environ Int*. 2019;123:240-255.
57. Rawlings JO, Pantula SG, Dickey DA. *Applied Regression Analysis: A Research Tool*, 2nd edn. New York, Berlin, Heidelberg: Springer Science & Business Media; 2001.
58. Lee S, Kim MJ, Kim JT, Yoo CK. In search for modeling predictive control of indoor air quality and ventilation energy demand in subway station. *Energy Build*. 2015;98:56-65.
59. Kropat G, Bochud F, Jaboyedoff M, Laedermann J, Murith C, Palacios M. Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. *J Environ Radioact*. 2015;147:51-62.
60. Yuchi W. Modelling fine particulate matter concentrations inside the homes of pregnant women in Ulaanbaatar, Mongolia; 2017.
61. Loy-benitez J, Vilela P, Li Q, Yoo C. Ecotoxicology and environmental safety sequential prediction of quantitative health risk assessment for the fine particulate matter in an underground facility using deep recurrent neural networks. *Ecotoxicol Environ Saf*. 2019;169:316-324.
62. Park S, Kim M, Kim M, et al. Predicting PM₁₀ concentration in Seoul metropolitan subway stations using artificial neural network (ANN). *J Hazard Mater*. 2018;341:75-82.
63. Kim M, Kim Y, Sung S, Yoo C. Data-driven prediction model of indoor air quality by the preprocessed recurrent neural networks. In: ICCAS-SICE 2009 – ICROS-SICE International Joint Conference 2009, Proceedings; 2009.
64. Gheziel A, Hanini S, Mohamedi B, Ararem A. Particle dispersion modeling in ventilated room using artificial neural network. *Nucl Sci Tech*. 2017;28:5.
65. Kropat G, Bochud F, Jaboyedoff M, et al. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: an application to Switzerland. *Sci Total Environ*. 2015;505:137-148.
66. Lim J, Kim Y, Oh T, et al. Analysis and prediction of indoor air pollutants in a subway station using a new key variable selection method. *Korean J Chem Eng*. 2012;29:994-1003.
67. Jafta N, Barregard L, Jeena PM, Naidoo RN. Indoor air quality of low and middle income urban households in Durban, South Africa. *Environ Res*. 2017;156:47-56.
68. Sarra A, Fontanella L, Valentini P, Palmeri S. Quantile regression and Bayesian cluster detection to identify radon prone areas. *J Environ Radioact*. 2016;164:354-364.
69. Braniš M, Šafránek J. Characterization of coarse particulate matter in school gyms. *Environ Res*. 2011;111:485-491.
70. Braniš M, Řezáčová P, Domasová M. The effect of outdoor air and indoor human activity on mass concentrations of PM₁₀, PM_{2.5}, and PM₁ in a classroom. *Environ Res*. 2005;99:143-149.
71. Choi M-L, Lim MJ, Kwon Y-M, Chung D-K. A study on the prediction method of emergency room (ER) pollution level based on deep learning using scattering sensor. *J Eng Appl Sci*. 2017;12:2560-2564.
72. Kang O, Liu H, Kim M, Kim JT, Wasewar KL, Yoo C. Periodic local multi-way analysis and monitoring of indoor air quality in a

- subway system considering the weekly effect. *Indoor Built Environ*. 2013;22:77-93.
73. Montgomery DC, Jennings CL, Kulahci M. *Introduction to Time Series Analysis and Forecasting*. New York, NY: John Wiley & Sons; 2008.
 74. Mirzavand M, Ghazavi R. A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resour Manag*. 2015;29:1315-1328.
 75. Solazzo E, Galmarini S. Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation. *Atmos Environ*. 2015;112:234-245.
 76. Yu T-C, Lin C-C. An intelligent wireless sensing and control system to improve indoor air quality: monitoring, prediction, and preaction. *Int J Distrib Sens Networks*. 2015;11:140978.
 77. Han Z, Gao RX, Fan Z. Occupancy and indoor environment quality sensing for smart buildings. In: 2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings. 2012:882-887. <http://ieeexplore.ieee.org/document/6229557/>
 78. Kadiyala A, Kumar A. Multivariate time series based back propagation neural network modeling of air quality inside a public transportation bus using available software. *Environ Prog Sustain Energy*. 2015;34:1259-1266.
 79. Ouaret R, Ionescu A, Petrehus V, Candau Y, Ramalho O. Spectral band decomposition combined with nonlinear models: application to indoor formaldehyde concentration forecasting. *Stoch Environ Res Risk Assess*. 2018;32:985-997.
 80. Karpatne A, Atluri G, Faghmous JH, et al. Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans Knowl Data Eng*. 2017;29:2318-2331.
 81. Jolliffe IT. A note on the use of principal components in regression. *Appl Stat*. 1982;31:300-303.
 82. Fix E, Hodges J. Discriminatory analysis, nonparametric discrimination: consistency properties; 1951.
 83. Sofuoglu SC. Application of artificial neural networks to predict prevalence of building-related symptoms in office buildings. *Build Environ*. 2008;43:1121-1126.
 84. Xie H, Ma F, Bai Q. Prediction of indoor air quality using artificial neural networks. In: 2009 Fifth International Conference on Natural Computation. 2009:414-418. <http://ieeexplore.ieee.org/document/5365009/>
 85. Skön J, Johansson M, Raatikainen M, Leiviskä K, Kolehmainen M. Modelling indoor air carbon dioxide (CO₂) concentration using neural network. *World Acad Sci Eng Technol Int Sci Index*. 2012;6:737-741.
 86. Chen X, Zheng Y, Chen Y, et al. Indoor air quality monitoring system for smart buildings. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct; 2014.
 87. Kadiyala A, Kumar A. Vector-time-series-based back propagation neural network modeling of air quality inside a public transportation bus using available software. *Environ Prog Sustain Energy*. 2016;35:7-13.
 88. Mentese S, Tasdibi D, Orak E. Estimation of sources and factors affecting indoor VOC levels using basic numerical methods. *AIMS Environ Sci*. 2016;3:827-841.
 89. Ahn J, Shin D, Kim K, Yang J. Indoor air quality analysis using deep learning with sensor data. *Sensors*. 2017;17:2476.
 90. Cociorva S, Iftene A. Indoor air quality evaluation in intelligent building. *Energy Procedia*. 2017;112:261-268.
 91. Liu Z, Li H, Cao G. Quick estimation model for the concentration of indoor airborne culturable bacteria: an application of machine learning. *Int J Environ Res Public Health*. 2017;14:857.
 92. Meng QY, Spector D, Colome S, Turpin B. Determinants of indoor and personal exposure to PM_{2.5} of indoor and outdoor origin during the RIOPA study. *Atmos Environ*. 2009;43:5750-5758.
 93. Liu H, Yoo C. A robust localized soft sensor for particulate matter modeling in Seoul metro systems. *J Hazard Mater*. 2016;305:209-218.
 94. Qi M, Zhu X, Du W, et al. Exposure and health impact evaluation based on simultaneous measurement of indoor and ambient PM_{2.5} in Haidian, Beijing. *Environ Pollut*. 2017;220:704-712.
 95. Kim M, Sankararao B, Kang O, Kim J, Yoo C. Monitoring and prediction of indoor air quality (IAQ) in subway or metro systems using season dependent models. *Energy Build*. 2012;46:48-55.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Wei W, Ramalho O, Malingre L, Sivanantham S, Little JC, Mandin C. Machine learning and statistical models for predicting indoor air quality. *Indoor Air*. 2019;29:704-726. <https://doi.org/10.1111/ina.12580>

APPENDIX

Data mining	A process to discover patterns in large data sets
Forecasting	A temporal estimate of a parameter based on prior knowledge of the parameter in a given space
Machine learning	The application of algorithms using computer systems to build models based on existing data for the prediction of unknowns
Mechanistic model	A physically based mathematical model
Prediction	A spatial estimate of a parameter based on knowledge of other parameters in a given space
Statistical model	A mathematical model based on statistical assumptions