Available online at www.sciencedirect.com

ScienceDirect

Building and Environment 43 (2008) 1121–1126

BUILDING AND ENVIRONMENT

www.elsevier.com/locate/buildenv

Application of artificial neural networks to predict prevalence of building-related symptoms in office buildings

Sait C. Sofuoglu*

Environmental Research Center and Department of Chemical Engineering, Izmir Institute of Technology, Gulbahce, Urla, 35430 Izmir, Turkey

Received 2 February 2006; received in revised form 2 February 2007; accepted 2 March 2007

Abstract

Artificial neural networks (ANN) were constructed to predict prevalence of building-related symptoms (BRS) of office building occupants. Six indoor air pollutants and four indoor comfort variables were used as input variables to the networks. A symptom metric was used as the measure of BRS prevalence, and employed as the output variable. Pollutant concentration, comfort variable, and occupant symptom data were obtained from the Building Assessment and Survey Evaluation study conducted by the US Environmental Protection Agency, in which all were measured concurrently. Feed-forward networks that employ back-propagation algorithm with momentum term and variable learning rate were used in ANN modeling. Root mean square error and R^2 value of the simple linear regression between observed and predicted output were used as performance measures. Among the constructed networks, the best prediction performance was observed in a one-hidden-layered network with an R^2 value of 0.56 for the test set. All constructed networks except one showed a better performance than the multiple linear regression analysis.

Keywords: Indoor air quality; Indoor environmental quality; Office buildings; Building-related symptoms; Artificial neural networks

1. Introduction

Symptoms like eye, nose, throat, and skin irritation, difficulty in breathing, headache, fatigue, dizziness experienced by occupants in a building and generally improve away from the building are known as building-related symptoms (BRS), sometimes also called as sick building syndrome. These non-specific symptoms are thought to be the results of job related, personal, psychological [1], and organizational factors [2], in addition to indoor environmental quality (IEQ). Linking IEQ and symptoms of building occupants has been a difficult task. In general, occupant symptoms were studied to correlate with individual environmental factors on individual occupant basis. There have been many studies conducted; some of which reported correlations with BRS: ventilation rate or carbon dioxide [1,3], bacterial endotoxin [4], groups of volatile organic compounds (VOCs) [5], fungal beta-glucans [6], moisture, mold, or moisture related micro-organisms [7–9], higher temperature [10,11]. However, some studies were either inconclusive [12], reported negative relationships [13], or found associations for some pollutants but were inconclusive for others [14–16].

BRS, in part, may be the result of overall IEQ rather than specific individual pollutants or comfort variables. Relationships between overall indoor air/environmental quality and occupant symptom prevalence were investigated on building basis in our previous studies through use of indices that rate indoor air pollution [17] and IEQ [18] in office buildings. Both studies reported high levels of correlations between IAP/IEQ indices and symptom indices; however, the results were based on average values of groups of office buildings.

Human body is a very complex structure; therefore responses of such systems to their environment, the symptoms, are difficult to predict on building, and even on occupant basis. This study aimed to explore the applicability of artificial neural networks (ANN), a method

*Tel.: +90 232 750 6648; fax: +90 232 750 6645.
E-mail addresses: cemilsofuoglu@iyte.edu.tr, saitcemil@iit.edu (S.C. Sofuoglu).

0360-1323/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.buildenv.2007.03.003

inspired by the human brain, to predict prevalence of BRS on building basis using IEQ variables as inputs, and without use of IAP/IEQ indices. In fact, ANN has been proposed to predict atmospheric concentrations of NO_x [19,20], ozone [21,22], benzene [23], SO_2 [24–26], and particulate matter [27,28], but was not used in indoor air quality or predicting occupant symptom prevalence.

The pollutant concentration and occupant symptom data were obtained from the Building Assessment and Survey Evaluation (BASE) study [29] conducted by the US Environmental Protection Agency, EPA, between 1994 and 1998. The BASE study, concurrently, measured a number of indoor air pollutant concentrations in office buildings throughout the United States, and identified occupant symptoms using self-administered questionnaires.

2. Material and methods

2.1. Artificial neural networks

ANN are systems that consists of interconnected neurons that process information. As opposed to the traditional modeling techniques, ANN is a data driven, self-adaptive, black-box method, which learns from examples. Networks can often correctly infer on a population when trained with sufficient data, even if the underlying relationships are unknown or difficult to describe as generally it is in the nonlinear nature of the real-world events. As a result, ANN has found use in many fields including environmental sciences.

This study used feed-forward networks which are successfully employed in environmental studies. In feed-forward networks, the input quantities are fed into input neurons, processed, and then passed onto the next level, hidden layer neurons. In the process, input signal is multiplied by a weight that determines the intensity of the input. The weighted input received from each input neuron is added up by the hidden layer neurons, and associated with a bias before passing the result onto the next level using a transfer function. All bias neurons are connected to all neurons in the next hidden and output layers.

Training is of the most fundamental importance to the ANN in which observed values of the output variable is compared to the network output, and then the error is minimized by adjusting the weights and biases. During training, a neuron receives inputs from a preceding layer, weights each input with a predetermined value, and combines the weighted inputs using

$$\mathrm{net}_j = \sum x_i v_{ij}, \tag{1}$$

where net_j = summation of the weighted input for the jth neuron, x_i = input from the ith neuron to the jth neuron, and v_{ij} = weight from the ith neuron in the preceding layer to the jth neuron in the present layer.

Level of activation is determined by calculating a transfer function value for net_j. In this study, hyperbolic tangent function was employed as activation function. The upper and lower limits of hyperbolic tangent function are $[-1, 1]$; its derivative is continuous [30]. Values of input and output variables were scaled to $[-1, 1]$ range by normalizing the mean and standard deviation of the training set.

Learning is defined as a network's ability to change weights [31]. In this study, the learning of ANN was accomplished by a back-propagation algorithm. Back-propagation is the most commonly used supervised training algorithm in multilayer feed-forward networks. In back-propagation networks, information is processed in the forward direction from the input layer to the hidden layer(s) and then to the output layer. The objective of a back-propagation network is to find the optimal weights which would generate an output vector $\boldsymbol{Y} = (y_1, y_2, \ldots, y_p)$ as close as possible to target values of output vector $\boldsymbol{T} = (t_1, t_2, \ldots, t_p)$ with a selected accuracy, by minimizing a predetermined error function:

$$E = \sum_P \sum_p (y_i - t_i)^2, \tag{2}$$

where y_i = component of an ANN output vector $\boldsymbol{Y}$, t_i = component of a target output vector $\boldsymbol{T}$, $\boldsymbol{p}$ = number of output neurons; and P = number of training patterns.

All the neural network calculations in this study were performed using MATLAB™ Neural Network Toolbox. Over-fitting is a problem for large networks with small data sets, such as the 100 building data set used in this study. Early stopping method was employed to avoid over-fitting. Data set was divided as 60–20–20 buildings as training, validation, and test sets, respectively. A network training function that updates weight and bias values according to gradient descent momentum and an adaptive learning rate is employed with early stopping. Two values (0.90 and 0.75) were tried for momentum constant, MC. Mean square error (MSE) is used as the performance function in gradient descent algorithm.

2.2. The data

Data from the BASE study [29] were used in this work. The BASE study surveyed 100 office buildings from 10 geographical/climatic regions of the United States. Pollutant concentrations and comfort variables were measured, building characteristics and occupant symptoms were determined using questionnaires. The study was conducted according to a standard protocol [32]. Surveyed pollutants included carbon dioxide (CO_2), carbon monoxide (CO), radon, particulate matter (PM_{10} and $PM_{2.5}$), VOCs, airborne bacteria, and fungi in indoor air. Temperature, relative humidity (RH), light, and noise were among the measured comfort variables. CO_2, CO, and comfort variables were measured continuously; PM was collected by inertial impaction onto pre-weighed Teflon air sampling membrane filters using a particle size selection device.

The mass of the collected particulates was determined gravimetrically using a microbalance. VOCs (except aldehydes) were collected in canisters, and analyzed by thermal desorption and gas chromatography/mass spectroscopy. Formaldehyde (HCHO) samples were collected on dinitrophenyl hydrazine cartridges and analyzed by high-performance liquid chromatography. Microbiological contaminants were collected on six-stage Andersen samplers. Occupants completed self-instructed questionnaires that inquired about their symptoms. Persistency and status of the symptoms away from the building were questioned. In this study, persistent and building-related symptoms that were experienced at least 1–3 days per week and that got better away from the building were considered. A symptom index, POPS2, defined as percent of occupants in the sampling area of the office building with two or more persistent symptoms [18] was calculated for BRS. Detailed information regarding questionnaires, sampling, and analytical methods can be found elsewhere [32]. Articles in scientific journals and presentations from *Indoor Air* and *Healthy Buildings* conferences held between 1995 and 2003, describing the study and summarizing preliminary results, are currently listed on the study website [33].

3. Results and discussion

3.1. Descriptive statistics

Descriptive statistics for pollutant and comfort variables and the symptom index are presented in Table 1. Pollutant concentrations show large variations with right skewed distributions. Concentrations of two of the six pollutants are distributed log-normally, while concentrations of three pollutants are best described with gamma distribution, and Weibull is the best fitting distribution for one pollutant. Comfort variable distributions are more symmetrical compared to pollutant concentration distributions, and POPS2 values are normally distributed.

3.2. ANN modeling

In this study, one or two-hidden-layered networks were employed. Hyperbolic tangent function was used as the transfer function. The database was divided into three sections for early stopping. Data from 60% of the buildings were used in training the networks, 20% were designated as the validation set, and the remaining 20% of the buildings were employed in testing. Performance of the networks was determined by two measures: coefficient of determination (R^2) for the regression between observed and modeled values of the output variable, and root mean square error (RMSE) about the modeled values. Six pollutants ($CO_2, PM_{2.5}$, HCHO, total VOCs, bacteria, fungi) and four comfort variables (temperature, RH, light and noise levels) were considered as input variables. All buildings had close mean noise levels, i.e., variation in mean noise level among buildings was small. Since the contribution of mean noise would have been low in predicting BRS due to small variation, the range of measured noise in each building was determined, and was considered as the noise input variable values. CO was not selected as an input variable because a large portion of the data were below detection limit; PM_{10} was not included because it is correlated with $PM_{2.5}$, and $PM_{2.5}$ penetrates more into the respiratory system. POPS2 was used as the output variable.

Structures of constructed networks and their performance levels are shown in Table 2. R^2 values ranged from 0.14 for the 20-neuron two-hidden-layered network (Model-8) to 0.56 for the 10-neuron one-hidden-layered network (Model-1). Structure of the best performing network is presented in Fig. 1. Training and testing performance of the network is shown in Fig. 2. All RMSE values were close, taking values around 8% except for one network (10.6%). Increasing the number of neurons in the hidden layers from 10 to 20, and reducing the value of the MC from 0.90 to 0.75 decreased the level of performance in terms of R^2 values.

Considering only the better performing networks, those with testing R of >0.60 (Models-1,3, and 5), all three models were consistent in having a problem predicting low

Table 1
Descriptive statistics of input and output variables

Statistic	TVOC ($\mu g\,m^{-3}$)	HCHO ($\mu g\,m^{-3}$)	CO_2 (ppm)	$PM_{2.5}$ ($\mu g\,m^{-3}$)	Fungi ($cfu\,m^{-3}$)	Bacteria ($cfu\,m^{-3}$)	Temp. (°C)	RH (%)	Light (lux)	Noise (dB)	POPS2 (%)
Mean	2168	15.7	573	8.0	62.9	46.1	20.7	44.9	446	44	57.0
Median	1613	14.8	528	6.9	42.3	39.1	21.2	46.0	425	43.2	57.5
Standard deviation	1664	8.3	127	3.8	68.3	26.1	2.0	15.8	154	9.1	10.5
Minimum	159	3.4	381	2.4	3.5	0.0	15.7	13.2	183	25.7	25
Maximum	9820	43.6	983	24.7	333.3	134.3	25.5	74.4	953	71.6	81.5
Distribution	LN	*W*	*G*	LN	LN	*G*	*W*	*N*	*L*	LN	*N*
Parameters	$\mu = 2178$	$L = 2.68$	$L = 355$	$\mu = 7.98$	$\mu = 66.8$	$L = 0$	$L = 11.2$	$\mu = 44.9$	$\mu = 435$	$\mu = 44$	$\mu = 57$
	$\sigma = 1793$	Sc = 14.46	Sc = 78	$\sigma = 3.87$	$\sigma = 94.2$	Sc = 19.1	Sc = 10.3	$\sigma = 15.8$	Sc = 85	$\sigma = 9$	$\sigma = 11$
	—	Sh = 1.59	Sh = 2.81	—	—	Sh = 5.39		—	—	—	—

G: gamma, *L*: logistic, LN: lognormal, *N*: normal, *W*: Weibull, Sc: scale, Sh: shape.

Table 2
Results of the network structure optimization

Network number	Network structure	Momentum constant	Training R	Testing R	Testing RMSE
1	10-10-1	0.90	0.78	0.75	8.4
2	10-10-1	0.75	0.61	0.43	8.0
3	10-20-1	0.90	0.71	0.68	8.9
4	10-20-1	0.75	0.57	0.49	7.8
5	10-10-10-1	0.90	0.88	0.62	10.6
6	10-10-10-1	0.75	0.68	0.54	7.5
7	10-20-20-1	0.90	0.62	0.51	7.7
8	10-20-20-1	0.75	0.44	0.36	7.6

R: correlation coefficient, RMSE: root mean square error.

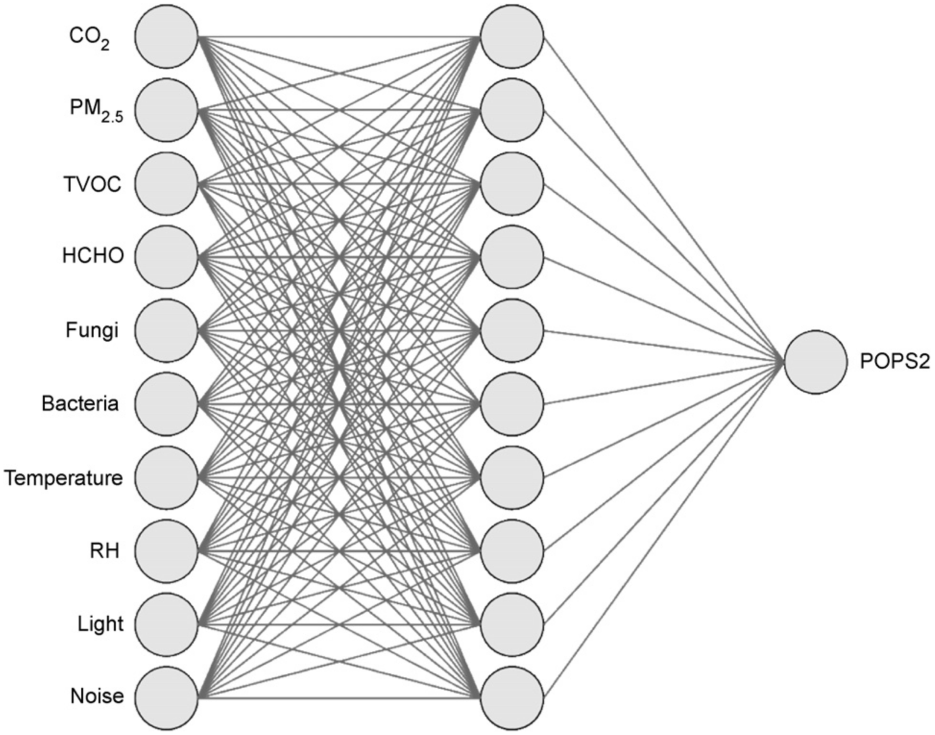

Fig. 1. Structure of the best performing network.

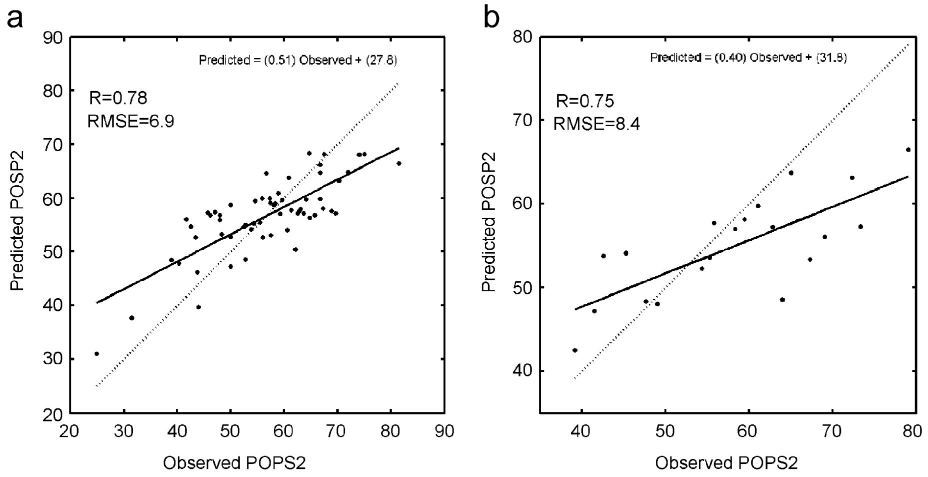

Fig. 2. Observed vs. predicted POPS2 by Model-1 for (a) training and (b) testing sets.

and high ends of the POPS2 range. Although Model-5 produced the highest RMSE value, about 1.25 times compared to the other two models, it performed better than the other models at both ends of the POPS2 range.

The literature, in general, investigated the relationship between indoor air quality and symptom prevalence on the individual occupant basis, and used logistic regression as the method of approach. There are a few comparable studies in the literature that approached the issue on the building basis, two of which are our previous index studies: the IAPI [17] and the IEI [18]. However, the relationships were based on average values of groups of buildings. Results of this study show that by employing ANN modeling, occupant symptom prevalence on building basis can be predicted using IEQ variables as inputs, and without grouping the buildings.

Multiple linear regression analyses were conducted to compare the results of ANN modeling. POPS2 was considered as the dependent variable; the 10 IEQ variables were considered as the independent variables. The R^2 value for the model was 0.14. The ANOVA test showed that the model was significant only at $p = 0.16$. Among the 10 independent variables, only formaldehyde ($p = 0.03$), temperature ($p = 0.08$), and bacteria ($p = 0.09$) were significant at significance level of 0.10. In addition, p-value for light was 0.14 for the *t*-statistic. The *t*-statistic *p*-values for the remaining independent variables were >0.40. Residuals showed no patterns for any of the independent variables.

The results obtained from the ANN modeling are encouraging compared to the constructed multiple linear regression model ($R^2 = 0.14$). R^2 values for the regression between observed and modeled output variable ranged between 0.26 and 0.56 for the networks trained with an MC of 0.90. The best performance ($R^2 = 0.56$) was obtained by a model that had one hidden layer with 10 neurons. The level of performance supports the claim that a part of the variation is explained by factors other than indoor environmental variables, such as job related, personal, and psychological variables. Large number of data sets is preferred in artificial intelligence methods to capture the variations in data; therefore, it would be reasonable to expect a better performance at the individual occupant level.

In conclusion, ANN modeling was applied to predict an index of occupant symptom prevalence using levels of indoor air pollutant concentrations and comfort variables as input variables at building level for the US office buildings. Back-propagation feed-forward networks that use hyperbolic tangent function as the transfer function are shown to predict occupant symptom prevalence with $R^2 > 0.50$.

References

[1] Mendell MJ. Non-specific symptoms in office workers: a review and summary of the epidemiologic literature. Indoor Air 1993;3:227–36.

[2] Arnold K. Sick building syndrome solutions. Professional Safety 2001;46:43–4.

[3] Seppanen O, Fisk WJ, Mendell MJ. Association of ventilation rates and CO_2 concentrations with health and other responses in commercial and institutional buildings. Indoor Air 1999;9:226–52.

[4] Teeuw KB, Vandenbroucke-Grauls CM, Verhoef J. Airborne gram-negative bacteria and endotoxin in sick building syndrome. A study in Dutch governmental office buildings. Archives of Internal Medicine 1994;154:2339–45.

[5] Ten Brinke J, Selvin S, Hodgson AT, Fisk WJ, Mendell MJ, Koshland CP, et al. Development of new volatile organic compound (VOC) exposure metrics and their relationship to 'sick building syndrome' symptoms. Indoor Air 1998;8:140–52.

[6] Rylander R, Lin RH. (1,3)-beta-*D*-glucan—relationship to indoor air-related symptoms, allergy and asthma. Toxicology 2000; 152:47–52.

[7] Bornehag CG, Blomquist G, Gyntelberg F, Jarholm B, Malmberg P, Nordvall L, et al. Dampness in buildings and health: Nordic interdisciplinary review of the scientific evidence on associations between exposure to 'dampness' in buildings and health effects (NORDDAMP). Indoor Air 2001;11:72–86.

[8] Haverinen U, Husman T, Vahteristo M, Koskinen O, Moschandreas D, Nevalainen A, et al. Comparison of two-level and three-level classifications of moisture-damaged dwellings in relation to health effects. Indoor Air 2001;11:192–9.

[9] Mendell MJ, Fisk WJ, Petersen MR, Hines CJ, Dong M, Faulkner D, et al. Indoor particles and symptoms among office workers: results from a double-blind cross-over study. Epidemiology 2002;13: 296–304.

[10] Jaakkola JJ, Heinonen OP. Sick building syndrome, sensation of dryness and thermal comfort in relation to room temperature in an office building: need for individual control of temperature. Environment International 1989;15:163–8.

[11] Mendell MJ, Naco GM, Wilcox TG, Sieber WK. Building-related risk factors and work-related lower respiratory symptoms in 80 office buildings. In: Levin H, editor. Indoor air '02: Proceedings of the ninth international conference on indoor air quality and climate, Santa Cruz, CA. Indoor Air 2002; 1:103–8.

[12] Skov P, Valbjorn O, DICS Group. The Danish town hall study—A one-year follow-up: indoor air '90: In: Walkinshaw DS, editior. Indoor air '90: Proceedings of fifth International Conference on indoor air quality and climate, vol. 1, Toronto, 1990. p. 787–91.

[13] Sundell J, Andersson B, Andersson K, Lindwall T. Volatile organic compounds in ventilating air in buildings at different sampling points in the building and their relationship with the prevalence of occupant symptoms. Indoor Air 1993;3:82–93.

[14] Armstrong CW, Sheretz PC, Llewellyn GC. Sick building syndrome traced to excessive total suspended particulates (TSP). In: IAQ '89 the human equation: Human health and comfort, ASHRAE, Atlanta, GA, 1989.

[15] Hodgson MJ, Collopy P. Symptoms and the microenvironment in the sick building syndrome: a pilot study. In: IAQ '89 the human equation: Human health and comfort, ASHRAE, Atlanta, GA, 1990.

[16] Hodgson MJ, Frohlinger J, Permar E, Tidwell C, Traven ND, Olenchock SA, et al. Symptoms and microenvironmental measures in nonproblem buildings. Journal of Occupational Medicine 1991;33: 527–33.

[17] Sofuoglu SC, Moschandreas DJ. The link between symptoms of office building occupants and in-office air pollution: the indoor air pollution index. Indoor Air 2003;13:332–43.

[18] Moschandreas DJ, Sofuoglu SC. The indoor environmental index and its relationship with symptoms of office building occupants. Journal of the Air and Waste Management Association 2004;54: 1440–51.

[19] Gardner MW, Dorling SR. Neural network modeling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 1999;33:709–19.

[20] Perez P, Trier A. Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Chile. Atmospheric Environment 2001;35:1783–9.

[21] Gardner MW, Dorling SR. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmospheric Environment 2000;34:21–34.

[22] Wang W, Lu WZ, Wang XK, Leung AYT. Prediction of maximum daily ozone level using combined neural network and statistical characteristics. Environment International 2003;29:555–62.

[23] Viotti P, Liuti G, Di Genova P. Atmospheric urban pollution: applications of an artificial neural network, ANN, to the city of Perugia. Ecological Modelling 2002;148:27–46.

[24] Chelani AB, Rao CVC, Phadke KM, Hasan MZ. Prediction of sulphur dioxide concentration using artificial neural networks. Environmental Modeling and Software 2002;17:161–8.

[25] Lu WZ, Fan HY, Lo SM. Application of revolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong. Neurocomputing 2003;51:387–400.

[26] Sofuoglu SC, Tayfur G, Birgili S, Sofuoglu A. Forecasting ambient air SO_2 concentrations using artificial neural networks. Energy Sources Part B 2006;1:127–36.

[27] Perez P, Trier A, Reyes J. Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment 2000;34:1189–96.

[28] Perez P, Reyes J. Prediction of maximum of 24-h average of PM_{10} concentrations 30 h in advance in Santiago, Chile. Atmospheric Environment 2002;34:4555–61.

[29] Burton LE, Baker B, Hanson G, Girman JG, Womble SE, McCarthy JF. Baseline information on 100 randomly selected office buildings in the United States (BASE): gross building characteristics. In: Proceedings of healthy buildings, Vol. 1, 2000. p. 151–6.

[30] Fu L. Neural networks in computer intelligence. New York: McGraw-Hill; 1994.

[31] Engel A. Complexity of learning in artificial neural networks. Theoretical Computer Science 2001;265:285–306.

[32] US EPA. A standardized EPA protocol for characterizing indoor air in large office buildings. Washington, DC: US Environmental Protection Agency; 1994.

[33] US EPA. ⟨http://www.epa.gov/iaq/base/publications.html⟩, 2006.