2009 Fifth International Conference on Natural Computation

Prediction of Indoor Air Quality Using Artificial Neural Networks

Hui Xie, Fei Ma
School of Civil & Environment Engineering
University of Science and Technology Beijing
Beijing, China
e-mail: xiehui20000@sina.com

Qingyuan Bai
Dezhou Yatai Group CO. LTD
Dezhou, Shandong, China

Abstract—**This paper described an application of artificial neural networks (ANNs) to predict the indoor air quality (IAQ). Six indoor air pollutants and three indoor comfort variables were used as input variables to the networks. An occupant symptom metric (PIAQ) was used as the measure of indoor air quality, and employed as the output variable. Pollutant concentration, comfort variable, and PIAQ data were obtained from previous studies. Feed-forward networks that employed back-propagation algorithm with momentum term and variable learning rate were used in ANN modeling. Among constructed networks, the best prediction performance was observed in a two-hidden-layered network with the high correlation coefficient and low root mean square error for the test set. Meanwhile, the constructed networks had a better performance than the multiple linear regression analysis. The results showed that the ANN approach can be applied successfully in predicting indoor air quality.**

Keywords-indoor air quality; artificial neural networks; prediction

I. 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 sick building syndrome (SBS). These non-specific symptoms are thought to be the results of job related, personal, psychological, and organizational factors [1], in addition to indoor air quality (IAQ). Linking IAQ 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 SBS: ventilation rate or carbon dioxide, bacterial endotoxin, groups of volatile organic compounds (VOCs), fungal betaglucans, moisture, mold, or moisture related microorganisms, higher temperature [2–7]. However, some studies were either inconclusive [8], reported negative relationships, or found associations for some pollutants but were inconclusive for others [9,10]. SBS, in part, may be the result of overall IAQ rather than specific individual pollutants or comfort variables. Relationships between overall indoor air quality and occupant symptom prevalence were investigated on building basis in previous studies [11,12], which reported high levels of correlations between indoor air pollution(IAP) indices and symptom indices

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 (ANNs), a method inspired by the human brain, to predict prevalence of SBS on building basis using IAQ variables as inputs, and without use of IAP indices. In fact, ANN has been proposed to predict atmospheric concentrations of NO_x [13], ozone [14], benzene [15], SO_2 [16,17], and particulate matter [18], but was not used in indoor air quality or predicting occupant symptom prevalence.

II. ARTIFICIAL NEURAL NETWORKS

ANN is one of the major branches of artificial intelligence, consisting of massively interconnected nonlinear memoryless processing elements known as neurons or nodes. 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 multilayer feed-forward neural networks which have been successfully employed in some environmental studies. The network usually consists of an input layer, some hidden layers and an output layer. In its simple form, each single neuron is connected to other neurons of a previous layer through adaptable synaptic weights. Knowledge is usually stored as a set of connection weights (presumably corresponding to synapse-efficacy in biological neural systems). Training is the process of modifying the connection weights, in some orderly fashion, using a suitable learning method. The network uses a learning mode, in which an input is presented to the network along with the desired output and the weights are adjusted so that the network attempts to produce the desired output. The weights, after training, contain meaningful information whereas before training they are random and have no meaning. Fig. 1 illustrates how information is processed through a single node.

978-0-7695-3736-8/09 $25.00 © 2009 IEEE
DOI 10.1109/ICNC.2009.502

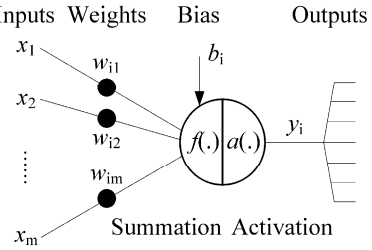

Figure 1. Information processing in an artificial neuron.

The output of the neuron is given by:

$$y(t+1) = a\left(\sum_{j=1}^{m} w_{ij} x_j(t) - b_i\right) \quad \text{and}$$
$$f_i = \text{net}_i = \sum_{j=1}^{m} w_{ij} x_j - b_i \tag{1}$$

where, $x = (x_1, x_2, \ldots, x_m)$ represent the m input applied to the neuron, w_{ij} represent the weights for input x_i, b_i is a bias value, $a(.)$ is activation function.

Learning is defined as a network's ability to change weights [19]. 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. An ANN with a back-propagation algorithm learns by changing the connection weights, and these changes are stored as knowledge.

Training of a network is basically a process of arriving at an optimum weight space of the network. The descent down the error surface is made using the following rule:

$$\Delta w_n = -n \frac{\partial E}{\partial w_n} \tag{2}$$

where n is the learning rate, w_n is the weight of the connection between the ith neuron of the input layer and the jth neuron of the hidden layer. The update of weight for the $(n + 1)$th pattern is given as:

$$w_n(n+1) = w_n(n) + \Delta w_n \tag{3}$$

The error E is the mean squared error and determined by the following relation:

$$E = \sum_{k=1}^{m} (O_k(n) - O'_k(n))^2 \tag{4}$$

where $O_k(n)$ is the output determined by the network for the nth pattern and $O'_k(n)$ is the corresponding output given in the training data set.

The weight change rule is a development of the perception learning rule. Weights are changed by an amount proportional to the error at that unit times, the output of the unit feeding into the weight. The output unit error is used to alter weights on the output units. Then, the error at the hidden nodes is calculated (by back-propagating the error at the output units through the weights), and the weights on the hidden nodes altered using these values. For each data pair to be learned a forward pass and backward pass is performed. This is repeated over once again until the error is at a low enough (or is given up). The input and the hidden layers consists of linear processing units as neurons, whereas, the output layer consists of non-linear processing units as the neurons. In this study, hyperbolic tangent function was employed as activation function and defined as:

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{5}$$

where z is the weighted sum of the inputs for a processing unit. Thus, the outputs are determined for each epoch, the error (E) calculated and the weights updated till a user specified error goal or epoch goal is reached successfully.

III. DATABASE

In order to train, validate and test the ANN prediction model, Data from the BASE study [20] were used in this work. The BASE study surveyed 100 office buildings from 10 geographical/climatic regions. 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 [21]. Surveyed pollutants included carbon dioxide (CO_2), particulate matter ($PM_{2.5}$), VOCs, airborne bacteria, and fungi in indoor air. Air temperature (T_{air}), relative humidity (RH), and velocity (V_{air}) were among the measured comfort variables. CO_2 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 liquid chromatography. Microbiological contaminants were collected by 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, PIAQ, defined as percent of occupants in the sampling area of the office building with two or more persistent symptoms [12] was calculated for SBS. Detailed information regarding questionnaires, sampling, and analytical methods can be found elsewhere [21].

Descriptive statistics for pollutants and comfort variables and the symptom index (PIAQ) were listed in Table 1. Pollutant concentrations show large variations with right skewed distributions. Concentrations of three pollutants are distributed log-normally, while concentrations of two 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 PIAQ values are normally distributed.

TABLE I. STATISTICS OF INPUT AND OUTPUT VARIABLES

Statistic	T_{air} (°C)	*RH* (%)	V_{air} (m/s)	CO_2 (ppm)	VOC (μg/m^3)	HCHO (μg/m^3)	$PM_{2.5}$ (μg/m^3)	Fungi (cfu/m^3)	Bacteria (cfu/m^3)	PIAQ (%)
Mean	20.7	44.9	0.26	573	2168	15.7	8.0	62.9	46.1	55.0
Median	21.2	46.0	0.29	528	1613	14.8	6.9	42.3	39.1	56.5
Standard deviation	2.0	15.8	1.05	127	1664	8.3	3.8	68.3	26.1	11.5
Min.	15.7	13.2	0.09	381	159	3.4	2.4	3.5	0.0	22.0
Max.	25.5	74.4	0.67	983	9820	43.6	24.7	333.3	134.3	87.5
Distribution	*W*	*N*	*N*	*G*	*LN*	*W*	*LN*	*LN*	*G*	*N*
Parameters	*L*=11.2 Sc=10.3	μ=44.9 σ=15.8 –	μ=0.27 σ=1.21 –	*L*=355 Sc=78 Sh=2.81	μ=2178 σ=1793 –	*L*=2.68 Sc=14.46 Sh=1.59	μ=7.98 σ=3.87 –	μ=66.8 σ=94.2 –	*L*=0 Sc=19.1 Sh=5.39	μ=55 σ=12 –

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

IV. ANN MODEL FOR IAQ PREDICTION

As net architecture authors used a 3-layer perceptron model (see Fig. 2). The first input layer contains the input variables of the net. Here, there were nine neurons (including six pollutants: CO_2, $PM_{2.5}$, HCHO, TVOCs, bacteria, and fungi and three comfort variables: T_{air}, *RH* and V_{air}) in the input layer. The number of hidden layers and value of neurons in each hidden layer are the parameters to be chosen in the perceptron model. Therefore, one/two hidden layers and different values of neurons were chosen to optimize the ANN performance. The last layer is the output layer, which consists of the target of the forecasting model. Here, PIAQ was used as the output variable. Hyperbolic tangent function was used as the transfer function. The database was divided into three sections for early stopping. 60% of Data were used in training the networks, 20% were designated as the validation set, and the remaining 20% were employed in testing. Performance of the networks was evaluated by two criteria: 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.

The root mean square error is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - p_i)^2} \tag{6}$$

where *a* and *p* sets are the actual and predicted output sets, respectively, *N* is the number of the points in the data set.

The coefficient of determination is determined from

$$\mathrm{R}^2 = 1 - \left(\frac{\sum_{i=1}^{N}(a_i - p_i)^2}{\sum_{i=1}^{N} p_i^2} \right) \tag{7}$$

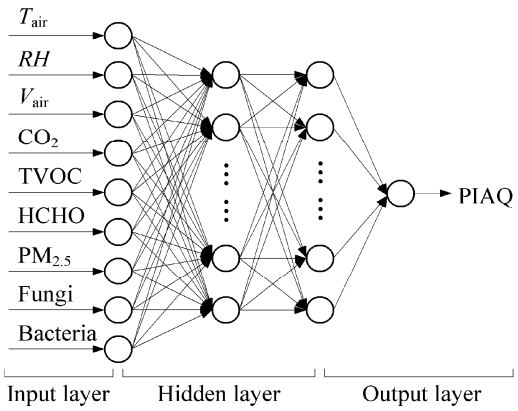

Figure 2. ANN architecture for IAQ prediction.

V. RESULTS AND DISCUSSION

Structures of constructed networks and their performance levels were listed in Table 2. Optimum trained ANN structures were selected according to the minimum RMSE and maximum R^2 values of the test set. Table 2 indicated that R^2 values of the test set ranged from 0.33 for ANN4 to 0.69 for ANN3, all RMSE values were close, taking values around 8%. Considering only the better performing networks, those with testing R2 more than 0.5 (ANN1 and ANN3), all two models were consistent in having a problem predicting low and high ends of the PIAQ range. Although ANN3 produced higher RMSE value, it performed better than the ANN1 model at both ends of the PIAQ range. Therefore, the structure of the best performing network in this study was 9-9-9-1 (ANN3).

Training and testing performance of the best performance network (ANN3) were shown in Fig. 3 and Fig. 4, which were provided with a red straight line indicating perfect prediction. Multiple linear regression analyses were conducted to compare the results of ANN modeling. PIAQ was considered as the dependent variable; nine IAQ

variables were considered as the independent variables. R^2 value for the model was 0.17. The ANOVA test showed that the model was significant only at p=0.15. Among nine independent variables, only HCHO (p=0.05) and bacteria (p=0.08) were significant at significance level of 0.10. In addition, the t-statistic p-values for the remaining variables were more than 0.40. Residuals showed no patterns for any of independent variables. The results obtained from the ANN model (R^2=0.69) were encouraging compared to the constructed multiple linear regression model (R^2=0.17).

TABLE II. ANN STRUCTURE OPTIMIZATION

ANN model	Network structure	Training		Testing	
		R^2	*RMSE*	R^2	*RMSE*
ANN1	9-9-1	0.61	6.9	0.55	8.1
ANN2	9-18-1	0.50	7.4	0.47	8.9
ANN3	9-9-9-1	0.77	7.2	0.69	8.8
ANN4	9-18-18-1	0.42	6.5	0.33	7.7

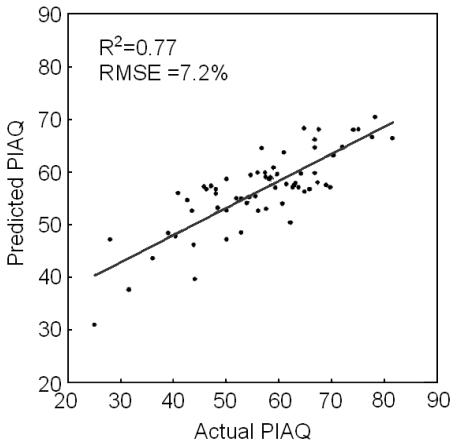

Figure 3. ANN-predicted PIAQ vs. actual PIAQ for training.

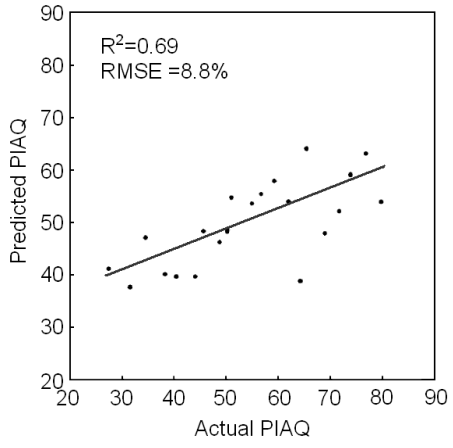

Figure 4. ANN-predicted PIAQ vs. actual PIAQ for testing.

VI. CONCLUSIONS

In this study, an ANN model was developed to predict IAQ using six indoor air pollutants and three indoor comfort variables as input variables and the occupant symptom metric (PIAQ) as output variable. Back-propagation feed-forward networks with the hyperbolic tangent function as the transfer function were shown to predict IAQ with R^2 >0.50. R^2 values for the ANN model were higher than those of the multiple linear regression model, which indicated that the present approach could be a new promising methodology to predict IAQ in complex situations.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial supports from National Key Technologies R&D Program of China (task no. 2006BAJ02A10).

REFERENCES

[1] Arnold K, "Sick building syndrome solutions," Professional Safety, vol. 46, 2001, pp. 43–44.

[2] Seppanen O, Fisk WJ, and Mendell MJ, "Association of ventilation rates and CO_2 concentrations with health and other responses in commercial and institutional buildings," Indoor Air, vol. 9, 1999, pp. 226–252.

[3] Teeuw KB, Vandenbroucke-Grauls CM, and Verhoef J, "Airborne gramnegative bacteria and endotoxin in sick building syndrome. A study in Dutch governmental office buildings," Archives of Internal Medicine, vol. 154, 1994, pp. 2339–2345.

[4] 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, vol. 8, 1998, pp. 140–152.

[5] 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, vol. 11, 2001, pp. 72–86.

[6] 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, vol. 13, 2002, pp. 296–304.

[7] Jaakkola JJ, and 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, vol. 15, 1989, pp. 163–168.

[8] Skov P, Valbjorn O, and DICS Group, "The Danish town hall study—A one-year follow-up: indoor air'90," Indoor air'90: Proceedings of the fifth international conference on indoor air quality and climate, Toronto, 1990, pp. 787–791.

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

[10] Hodgson MJ, and Collopy P, "Symptoms and the microenvironment in the sick building syndrome: a pilot study," Proceedings of IAQ'90 the human equation: Human health and comfort, ASHRAE, Atlanta, GA, 1990.

[11] Sofuoglu SC, and Moschandreas DJ, "The link between symptoms of office building occupants and in-office air pollution: the indoor air pollution index," Indoor Air, vol. 13, 2003, pp. 332–343.

[12] Moschandreas DJ, and Sofuoglu SC, "The indoor environmental index and its relationship with symptoms of office building occupants," Journal of the Air and Waste Management Association, vol. 54, 2004, pp. 1440–1451.

[13] Gardner MW, and Dorling SR, "Neural network modeling and prediction of hourly NOx and NO_2 concentrations in urban air in London," Atmospheric Environment, vol. 33, 1999, pp. 709–719.

[14] Wang W, Lu WZ, Wang XK, and Leung AYT, "Prediction of maximum daily ozone level using combined neural network and statistical characteristics," Environment International, vol. 29, 2003, pp. 555–562.

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

[16] Chelani AB, Rao CVC, Phadke KM, and Hasan MZ, "Prediction of sulphur dioxide concentration using artificial neural networks," Environmental Modeling and Software, vol. 17, 2002, pp. 161–168.

[17] Sofuoglu SC, Tayfur G, Birgili S, and Sofuoglu A, "Forecasting ambient air SO2 concentrations using artificial neural networks," Energy Sources Part B, vol. 1, 2006, pp. 127–136.

[18] Perez P, Trier A, and Reyes J, "Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile," Atmospheric Environment, vol. 34, 2000, pp. 1189–1196.

[19] Engel A, "Complexity of learning in artificial neural networks," Theoretical Computer Science, vol. 265, 2001, pp. 285–306.

[20] Burton LE, Baker B, Hanson G, Girman JG, Womble SE, and McCarthy JF, "Baseline information on 100 randomly selected office buildings in the United States (BASE): gross building characteristics," Proceedings of healthy buildings, 2000, pp. 151–156.

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