International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:6, No:1, 2012

Modelling Indoor Air Carbon Dioxide (CO_2) Concentration using Neural Network

J-P. Skön, M. Johansson, M. Raatikainen, K. Leiviskä and M. Kolehmainen

***Abstract*—**The use of neural networks is popular in various building applications such as prediction of heating load, ventilation rate and indoor temperature. Significant is, that only few papers deal with indoor carbon dioxide (CO_2) prediction which is a very good indicator of indoor air quality (IAQ). In this study, a data-driven modelling method based on multilayer perceptron network for indoor air carbon dioxide in an apartment building is developed. Temperature and humidity measurements are used as input variables to the network. Motivation for this study derives from the following issues. First, measuring carbon dioxide is expensive and sensors power consumptions is high and secondly, this leads to short operating times of battery-powered sensors. The results show that predicting CO_2 concentration based on relative humidity and temperature measurements, is difficult. Therefore, more additional information is needed.

***Keywords*—**Indoor air quality, Modelling, Neural networks

I. Introduction

INDOOR Air Quality (IAQ) is a widely researched topic, because of its impacts on occupant's health. Symptoms like e.g. eye dryness, running nose, headache and dizziness are experienced by occupants in a building. Sick building syndrome (SBS) is a combination of ailments and usually it is related to poor indoor air quality [1]. About half of the studies concerning non-residential and non-industrial buildings present that the risk of the SBS decreased substantially, if ventilation rates were increased, so that carbon dioxide CO_2 concentrations were reduced below 800 ppm [2], indicating better IAQ. As a whole, linking symptoms and IAQ of building occupants has been a very difficult task.

The concentration of CO_2 in indoor air is generally used as a surrogate for ventilation rate and concentration below 1000 ppm is widely recommended. For the temperature, the Finish guideline value is 21°C and for the relative humidity it is 20-60 % during the heating season [3].

J-P. Skön is with the Department of Environmental Science, Research Group of Environmental Informatics, University of Eastern Finland, Kuopio, Finland (e-mail: jukka-pekka.skon@uef.fi)

M. Johansson is with the Department of Environmental Science, Research Group of Environmental Informatics, University of Eastern Finland, Kuopio, Finland (e-mail: markus.johansson@uef.fi)

M. Raatikainen is with the Department of Environmental Science, Research Group of Environmental Informatics, University of Eastern Finland, Kuopio, Finland (e-mail: mika.raatikainen@uef.fi)

K. Leiviskä is with the Control Engineering Laboratory, University of Oulu, Oulu, Finland (e-mail: kauko.leiviska@oulu.fi)

M. Kolehmainen is with the Department of Environmental Science, Research Group of Environmental Informatics, University of Eastern Finland, Kuopio, Finland (e-mail: mikko.kolehmainen@uef.fi)

In addition, increased interest in energy efficiency is thought to affect negatively on indoor air quality. For instance, in Nature there are discussions about low-energy buildings and their relation to carbon emissions [4], as well as on the use of biological indicators for IAQ [5]. In Science, there are articles discussing about using and extending smart grids for energy efficiency [6], sustainability [7], and the relationships between healthiness and the environment [8].

Neural networks have been used in the prediction of indoor air quality e.g. feedforward backpropagation [9, 10], recurrent neural networks [11], fuzzy neuro systems [12] and model comparison [13]. There are also previous studies on forecasting outdoor air quality parameters using computational methods [14, 15, 16].

This study aims to explore the applicability of multilayer perceptron (MLP) network to predict CO_2 concentration in indoor air using measurements of relative humidity and temperature.

II. Materials and Methods

A. Data Collection

The case study was conducted in an apartment building located in Kuopio, Finland, from May to October 2011. The building has been built in 1973. Indoor air quality data was collected continuously in 8 apartments from 4 bedrooms and 6 living rooms, using an energy consumption and indoor air quality monitoring system [17]. Measurements were taken every 10 seconds.

The collected IAQ data consisted of continuous measurements of temperature, relative humidity and CO_2 concentration in the study building. Measured variables and their ranges are presented in Table I.

TABLE I
Data variables and their range

Variable	Range
Temperature [°C]	20.7-27.7
Relative humidity [%]	15.9-62.8
Carbon dioxide [ppm]	341.0-998.9

The size of the collected data matrix was 1270916 rows, 31 (including measurement time) variables in columns.

B. Multilayer Perceptron (MLP)

Multilayer perceptrons have been used successfully to solve

classification, regression and function approximation problems. Multilayer perceptron models are capable on modelling highly non-linear and complex problems, through the topology of the network, as presented in a simple form in Figure 1.

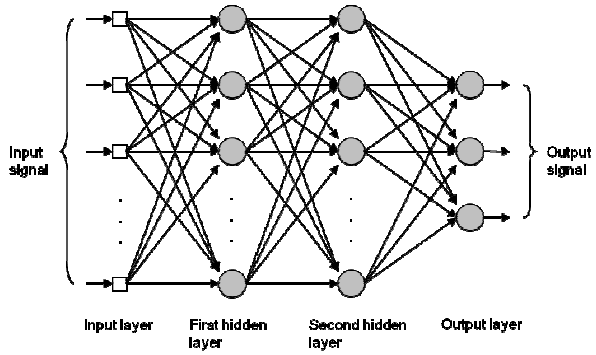

Fig. 1 The structure of a multilayer perceptron with two hidden layers (modified from [18])

MLP networks consist of groups of interconnected nodes arranged in different layers, such as the input layer, hidden layers, and the output layer. The purpose of the input layer is to distribute inputs to the first hidden layer, where the mathematical processing task takes place. It summarizes the inputs based on predefined weights, processes them by a transfer function and transfers the result to the next layer, which is usually an output layer, as a linear combination. Finally, the output layer receives the information from the last hidden layer. The network outputs are calculated by a transfer function, which can be e.g. hyperbolic or sigmoid [18].

C. Modelling Carbon Dioxide Concentration using MLP

The data was processed and modelled under a Matlab-software platform (Mathworks, Natick, MA, USA) according to Figure 2. At the beginning, the indoor air quality data was pre-processed. This means removing outliers, scaling the data and extracting the features using time window of 60 minutes. Extracted features are presented in Table II, where n is the total number of data samples, x_i is the ith measurement, $\bar{x}_i$ is the mean of the measurements, and σ is the standard deviation, respectively. RH means relative humidity, T temperature and CO_2 carbon dioxide.

TABLE II
DEFINITIONS OF THE EXTRACTED FEATURES FROM THE INDOOR AIR QUALITY DATA

Features	**Definitions**	**Variable number**
Minimum	$\min_{i=1}^{n} x_i$	3 (RH), 4 (T)
Maximum	$\max_{i=1}^{n} x_i$	5 (RH), 6 (T)
Kurtosis	$\frac{n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4}{\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)^2}-3$	7 (RH), 8 (T)
Skewness	$\frac{\sqrt{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^3}{\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)^{\frac{3}{2}}}$	9 (RH), 10 (T)
Standard deviation	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$	11 (RH), 12 (T)
Average	$\frac{1}{n}\sum_{i=1}^{n} x_i$	13 (RH), 14 (T), 21 (CO_2)
Median	The number separating the higher half of a sample from the lower half.	15 (RH), 16 (T)
Root mean square	$\sqrt{\frac{1}{n}\left(x_1^2+x_2^2+\cdots+x_n^2\right)}$	17 (RH), 18 (T)
Sum	$\sum_{i=1}^{n} x_i$	19 (RH), 20 (T)

The size of the final data matrix used was 78077 rows, 21 variables in columns. Variable number 1 is measurement time, variable number 2 is the room ID and the rest of the variables are calculated features. No outliers or missing values were found. The data was scaled using variance scaling, defined as:

$$x_i' = \frac{x_i - \bar{x}}{\sigma_x}, \sigma_x \neq 0, \quad (1)$$

where $\bar{x}$ is the average of values in vector x and σ_x denotes the standard deviation of those values. Thus, variance scaling not only equalizes the effect of variables having a different range; it also reduces the effect of possible outliers in the data.

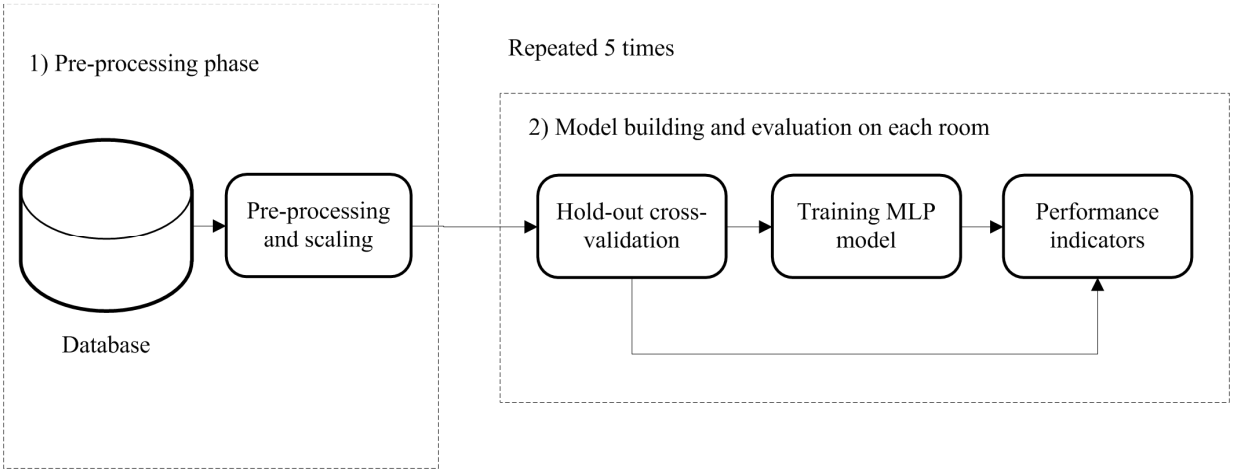

Fig. 2 Main stages of the building and evaluating MLP model for predicting CO_2 concentration

After pre-processing, the input variables of the MLP model were selected using correlation analysis. Selected varibles and their delayed values were used in training the MLP model. Delaying horizon was set to 1, 2, 3, 24, 25, 26, 168, 169 and 170 hours. The model parameters were selected based on experience and knowledge. The parameters used were 10 hidden neurons in a hidden layer, the back-propagation learning was based on the Levenberg-Marquardt algorithm, the performance function was regularized mean squared error, hyperbolic sigmoid tangent was used for the hidden layers and linear for the output layer.

MLP models room-specific performance indicators were estimated by repeating model training 5 times on each model, using hold-out cross-validation [19] (Figure 2). The used method is the simplest way to validate the goodness of a model. In this approach the data was divided into two sets; the training set and the validation set (hold-out set). The training data set consisted measurements of relative humidity and temperature 9 rooms and rest of the data was used as a validation data.

Performance of the models was based on four indicators, namely Index of Agreement (IA) [20], Coefficient of Determination (R^2) [20], Root Mean Square Error (RMSE) and their statistics (mean ± S.D). Here P_i denotes a predicted element and O_i equals to observed element and $\overline{O}$ is the symbol for the average of observations. Index of Agreement is a measure which can be used to describe the goodness of a model:

$$IA = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} \quad (2)$$

Coefficient of Determination (R^2) is defined as follows:

$$R^2 = \frac{\sum_{i=1}^{N}\left(P_i - \overline{O}\right)^2}{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2} \quad (3)$$

R^2 is an index measuring the proportion of variation explained by the model.

Root Mean Square Error (RMSE) is defined as follows:

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}(P_i - O)^2\right)^{\frac{1}{2}} \quad (4)$$

RMSE is the estimated standard deviation of the errors. If the RMSE is small relative to the variation in the data, then the R^2 is near to 1 and the data are concentrated close to the fitted model. Both R^2 and RMSE measure the goodness-of-fit of the model in their own way.

III. RESULTS

Input variables of the MLP model input were selected using correlation analysis (Figure 3). Variables which correlated with mean CO_2 were 5 (max RH), 6 (max T), 9 (Skewness RH), 14 (Average T), 16 (Median T), 18 (RMS T) and 20 (Sum T). Negative linear relationship was reasonable ($0.2 < R < 0.3$) between mean CO_2 and selected variables.

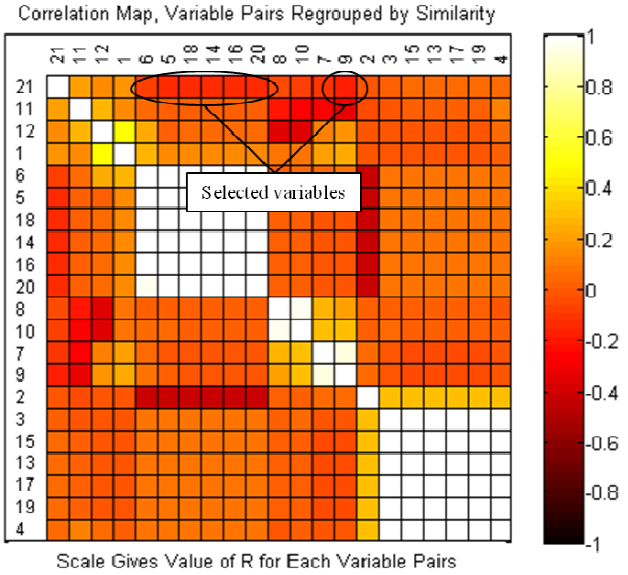

Fig. 3 Correlation map of extracted features (variables); Variable pairs are regrouped by similarity using k-means clustering

Averages and standard deviations of model performance indicators are presented in Table III. The results indicated, that predicting CO_2 concentration, based on calculated features utilizing on relative humidity and temperature measurements, is difficult. However, it can be seen that the best model performances can be found when predicting living rooms CO_2 concentration.

TABLE III
STATISTICS (MEAN ± S.D.) OF THE MLP MODEL PERFORMANCE
BD MEANS BEDROOM AND LR LIVING ROOM

Model	IA	R^2	RMSE
1 (BD)	0.68 ± 0.02	0.23 ± 0.02	83.14 ± 1.92
2 (LR)	0.60 ± 0.12	0.22 ± 0.12	177.45 ± 102.04
3 (BD)	0.66 ± 0.02	0.28 ± 0.05	175.45 ± 6.88
4 (LR)	0.67 ± 0.03	0.24 ± 0.04	144.22 ± 4.33
5 (LR)	0.76 ± 0.01	0.39 ± 0.02	122.85 ± 5.37
6 (BD)	0.40 ± 0.01	0.00 ± 0.00	258.68 ± 13.97
7 (LR)	0.54 ± 0.00	0.11 ± 0.00	193.89 ± 2.34
8 (BD)	0.58 ± 0.03	0.27 ± 0.04	189.16 ± 10.15
9 (LR)	0.70 ± 0.01	0.32 ± 0.02	122.26 ± 2.08
10 (LR)	0.63 ± 0.01	0.31 ± 0.02	174.77 ± 2.74

The performance was also visualized using the scatter plot (Figure 4) and time series plot (Figure 5) of the predicted versus observed mean CO_2 concentrations.

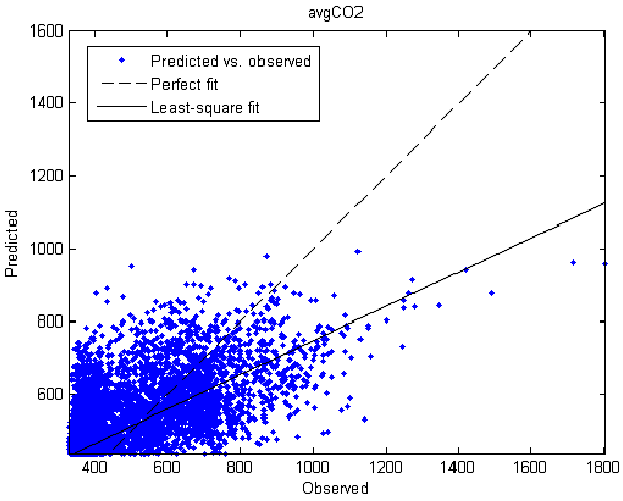

Fig. 4 Mean CO_2 concentrations (observed versus predicted) obtained as a result of one of five MLP model 5 (LR). The dashed line gives the perfect fit and the solid line the fitting using least-squares

In Figures 4 and 5 it can be seen that the prediction accuracy is reasonable in normal situations, but in exceptional circumstances the model cannot predict correctly.

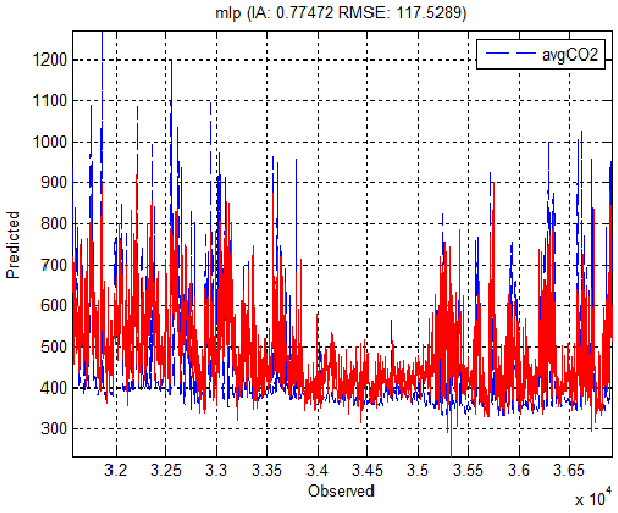

Fig. 5 Time series plot of observed versus predicted mean CO_2 concentration obtained as a result of one of five MLP model 5 (LR)

IV. DISCUSSION

In this study we tested the MLP model for predicting mean CO_2 concentrations in ten rooms. Overall, it seems that predicting CO_2 is challenging, if it is only based on measurements on relative humidity and temperature. At first, we tried to model mean CO_2 concentration, using means of relative humidity and temperature as model inputs, but results were poor (not presented here). Mean values of index of agreements were lower than 0.5. Therefore we decided to calculate several features to attain further information concerning the dependences. After that the performance of

MLP models was tested using selected varibles (7). Performance indicators IA, R^2 and RMSE (Table III) show that the goodness and fit of the model were reasonable on models 5 (LR) and 7 (LR). It seems that predicting living rooms CO_2 concentration is easier probably due to small variations in CO_2 concentration.

Thus, it seems to be very difficult to build up a reliable and generalizable prediction model using only relative humidity and temperature as input variables. If the model generalization ability and prediction accuracy were good, it could be implemented as a soft sensor to make predictions of CO_2 concentration.

V. Conclusion

Today, buildings are more airtight and energy efficient, which can have an effect on indoor air quality. Therefore, the developing new affordable and reliable indoor air quality sensors (e.g. soft sensors) is important. The results presented in this paper show, that prediction of mean CO_2 concentration is difficult, if it is based only on measurements of relative humidity and temperature. Further study is needed to improve the model accuracy.

In the future, the study will be expanded to several apartment buildings and more additional information is needed as model input e.g. information on presence and electricity consumption, to improve the goodness of the model.

Acknowledgment

This research was done as a part of the Finnish AsKo-project (*Asuinrakennusten korjaus- ja täydennysrakentamisen vaikutukset asumisen energiatehokkuuteen ja sisäilman laatuun*; The effects of renovation and complimentary construction on energy efficiency and indoor air quality). For financial support, the authors would like to thank the Ministry of the Environment.

References

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

[2] O. A. Seppänen, W. J. Fisk and M. J. Mendell, "Association of ventilation rates and CO_2 concentrations with health and other responses in commercial and institutional buildings," *Indoor Air*, vol. 9, pp. 226-252, 1999.

[3] Asumisterveysohje, *Sosiaali- ja terveysministeriön oppaita*, Sosiaali- ja terveysministeriö, Oy Edita Ab, Helsinki, 2003 (in Finnish).

[4] D. Butler, "Architects of a Low-energy Future," *Nature*, 452, pp. 520-523, Apr. 2008.

[5] R. Armstrong and N. Spiller, "Synthetic biology: Living quarters," *Nature*, 467, pp. 916-918, Oct. 2010.

[6] N. Gershenfeld, S. Samouhos, and B. Nordman: "Intelligent Infrastructure for energy efficiency," *Science*, vol. 372, pp.1086-1088, Feb. 2010.

[7] R. J. Jackson, "Environment Meets Health, Again," *Science*, 315(5817), pp.1337, Mar. 2007.

[8] J. P. Holdren, "Energy and Sustainability," *Science*, 315(5813), pp. 737, Feb. 2007.

[9] S. C. Sofuoglu, "Application of artificial neural networks to predict prevalence of building-related symptoms in office buildings," *Building and Environment*, vol. 43, pp. 1121-1126, 2007.

[10] H. Xie, F. Ma and Q. G. Bai, "Prediction of indoor air quality using artificial neural networks," *Fifth International Conference on Natural Computation (ICNC '09)*, vol. 2, pp. 414-418, 2009.

[11] M. H. Kim, Y. S. Kim, J. J. Lim, J. T. Kim, S. W. Sung and C. K. Yoo, "Data-driven prediction model of indoor air quality in an underground space," *Korean Journal of Chemical Engineering*, vol. 27, pp. 1675-1680, 2010.

[12] T. E. Alhanafy, F. Zaghlool and A. S. El Din Moustafa, "Neuro fuzzy modeling scheme for the prediction of air pollution," *Journal of American Science*, vol. 6, pp. 605-616, 2010.

[13] T. Lu and M. Viljanen, "Prediction of indoor temperature and relative humidity using neural network models: model comparison," *Neural Computing & Applications*, vol.18, pp. 345-357, 2009

[14] M. Kolehmainen, H. Martikainen, T. Hiltunen, and J. Ruuskanen, "Forecasting air quality parameters using hybrid neural network modelling," *Environmental Monitoring and Assessment*, vol. 65, pp. 277-286, 2000.

[15] M. Kolehmainen, H. Martikainen and J. Ruuskanen, "Neural networks and periodic components used in air quality forecasting," *Atmospheric Environment*, vol. 35, pp. 815-825, 2001.

[16] H. Niska, T. Hiltunen, M. Kolehmainen and J. Ruuskanen, **"**Hybrid models for forecasting air pollution episodes," *International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA'03)*, University Technical Institute of Roanne, France April 23-25, 2003.

[17] J-P. Skön, O. Kauhanen and M. Kolehmainen, "Energy Consumption and Air Quality Monitoring System," *Proceedings of the 7th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 163-167, Adelaide, Australia Dec. 6-9, 2011.

[18] S. Haykin, "Neural Networks–A Comprehensive Foundation," 2nd ed., New Jersey: Prentice-Hall Inc., 1999.

[19] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, vol. 30, pp. 271–274, 1998.

[20] C. J. Willmott, "Some Comments on the Evaluation of Model Performance," *Bulletin American Meteorological Society*, vol. 63, pp.1309-1313, 1982.