Analysing Events and Anomalies in Indoor Air Quality Using Self-Organizing Maps

Jukka-Pekka Skön[1], Markus Johansson[1], Mika Raatikainen[1], Ulla Haverinen-Shaughnessy[2], Pertti Pasanen[1], Kauko Leiviskä[3] and Mikko Kolehmainen[1]

[1]Department of Environmental Science, University of Eastern Finland, P.O. BOX 1627
70211 Kuopio, Finland
jukka-pekka.skon@uef.fi
markus.johansson@uef.fi
mika.raatikainen@uef.fi
pertti.pasanen@uef.fi
mikko.kolehmainen@uef.fi

[2]Department of Environmental Health, National Institute for Health and Welfare, P.O. BOX 95
70701 Kuopio, Finland
ulla.haverinen-shaughnessy@thl.fi

[3]University of Oulu, Control Engineering Laboratory, P.O. BOX 4300
90014 Oulu, Finland
kauko.leiviska@oulu.fi

ABSTRACT

This paper presents an overview of an indoor air quality data analysis, using self-organizing maps. The aim of the study was to research quality variations in indoor air, and the suitability of the used method for large scale data analysis of indoor air quality. Research was conducted in a six floor apartment building (built in the 80'ies) located in Kuopio, Finland, from January to May 2011. Quality data were collected continuously in 6 apartments from 10 rooms at the 2nd and the 6th floors, using an energy consumption and indoor air quality monitoring system. Three of the six research apartments were located on the 2nd floor and the other three on the 6th floor. At first the indoor air quality data were modelled using the SOM-algorithm. Next, the neuron reference vectors of the formed map were clustered to reveal dominating elements of each territory of the map. The results indicated that the method presented in this paper is an efficient way to analyse indoor air quality. Results indicated also that problems with indoor air quality occur more often during the wintertime in buildings utilizing mechanical exhaust ventilation. In particular, elevated CO_2 concentrations indicate poor air quality in the bedrooms.

Keywords: Clustering, Indoor air quality, Self-organizing map, Neural networks

Mathematics Subject Classification: 62-07, 62H30

Computing Classification System: I.5

1. INTRODUCTION

Indoor Air Quality (IAQ) is a widely researched topic because of its manifold impacts to occupant's health. Heating, ventilation, and air conditioning (HVAC) systems, as well as building materials used, are related to measured IAQ parameters, including relative humidity, temperature, carbon dioxide (CO_2) concentration, and volatile organic compounds (VOC). Recently, development in the areas of novel data logging units and processing massive databases allows switching to continuous measurements, which are bound to be more reliable than single (or short time) measurements. The use of large data-sets leads to higher standard of research (Skön et al., 2011).

The concentration of CO_2 in indoor air is generally used as a surrogate for ventilation rate. Indoor CO_2 concentration below 1000 ppm is generally recommended, however, it does not guarantee that the ventilation rate is always adequate. Indeed, the recommended limit of 1000 ppm is regarded as indicative of ventilation rates that are acceptable with respect to body odours (Apte et al., 2000). For the temperature, the Finnish guideline value is 21 ºC and for the relative humidity it is 20-60 % during the heating season (Asumisterveysohje, 2003).

It is challenging to characterize the indoor CO_2 concentration, because in addition to the outdoor CO_2 concentration, the concentration indoors is depending on both the ventilation rate and occupancy, both varying as a function of time. Grab samples or other short-term measurements may be inadequate to provide information about the long-term ventilation conditions in buildings (Daisey et al., 2003).

About half of the studies concerning non-residential and non-industrial buildings suggest that the risk of the sick building syndrome symptoms (SBS) decreased substantially if ventilation rates were increased so that CO_2 concentrations were reduced below 800 ppm (Seppänen et al., 1999), indicating better IAQ. In addition, there are numerous other factors affecting indoor air quality, like the occupant density, ventilation, as well as maintenance and cleaning practices (Sofuoglu et al., 2011).

In past years there has been increasing concern about the health effects of indoor air quality (Jones, 1999). Reasons for concern may not be exaggerated, considering that approximately 90 % of our time is spent indoors (Salthammer, 2011). There are studies relating poor indoor air quality to health effects such as asthma and allergic reactions. Children may be more sensitive to the possible health effects of poor indoor air quality than adults.

In general, indoor environment can be regarded as an environment where it is difficult to assess occupants' exposure to indoor air pollutants because of the spatial and temporal variations in the substance spectrum. However, nowadays there are data available for a large number of substances. Therefore, it is possible to make recommendations regarding good indoor air quality based on statistically derived reference values and toxicologically based guideline values (Salthammer, 2011).

Recently increased interest in energy efficiency is thought to affect negatively indoor air quality. For instance, in Nature there are discussions about low-energy buildings and their relation to carbon emissions (Butler, 2008), as well as on the use of biological indicators for IAQ (Armstrong and Spiller, 2010). In Science, there are articles discussing about using and extending smart grids for energy efficiency (Gershenfeld et al., 2010), sustainability (Jackson, 2007), and the relationships between healthiness and the environment (Holdren, 2007).

Neural networks have been used in the prediction of indoor air quality: feedforward backpropagation (Sofuoglu, 2007, Xie et al., 2009), recurrent neural networks (Kim et al., 2010), and fuzzy neuro systems (Alhanafy et al., 2010). There are also previous studies on outdoor air quality using computational methods (e.g. Kolehmainen et al. 2000, Kolehmainen et al. 2001, and Niska et al. 2003). This paper describes methodology used on the indoor air quality data analysis, including data processing chain, pre-processing the raw data, basic idea of self-organizing map (SOM) and clustering. Some results are also presented regarding testing the methodology in a case study building.

2. METHODOLOGY

2.1. Collection of indoor air quality data

The case study was conducted in an apartment building built in 1980's. The six storey building is located in Kuopio, Finland. Indoor air quality data; temperature, relative humidity and CO_2 concentration were collected continuously (from January to May 2011) from six apartments, a total of ten rooms at the 2nd and the 6th floors using a system for energy consumption and indoor air quality monitoring (Skön et al., 2011). Monitoring system is developed as a part of the Finnish AsTEKa-project and overview of the monitoring system is presented in Figure 1. The apartments were distributed evenly on both floors. Measurements were taken at ten seconds intervals.

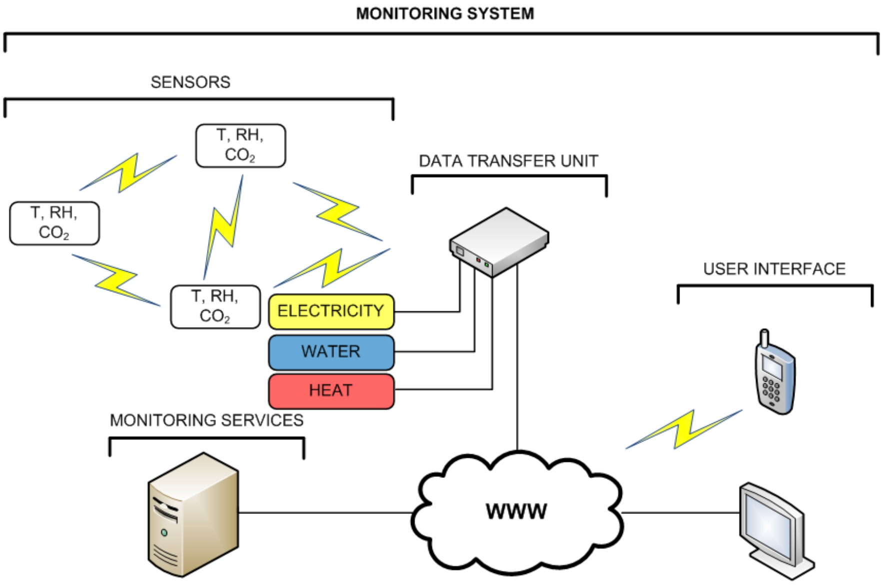

Figure 1. Overview of the monitoring system (modified from Skön et al., 2011).

2.2. Data processing and pre-processing the raw data

In the beginning of the data processing (Figure 2), the raw IAQ data were pre-processed for data analysis. First, this means removing outliers and secondly, variance scaling of the data. After removing outliers the data matrix consists variables between 0 ºC ≤ T ≤ 50 ºC, 0 % ≤ RH ≤ 100 % and 0 ppm ≤ CO_2 ≤ 2000 ppm. The solid data matrix is then modelled by self-organizing map (SOM). Reference vectors of the SOM are easier to analyse than original measurement vectors due to the reduced number of data. Next, the reference vectors are classified to clusters by K-means clustering method. Finally, the clusters are analysed with respect to IAQ events or anomalies using asset expertise.

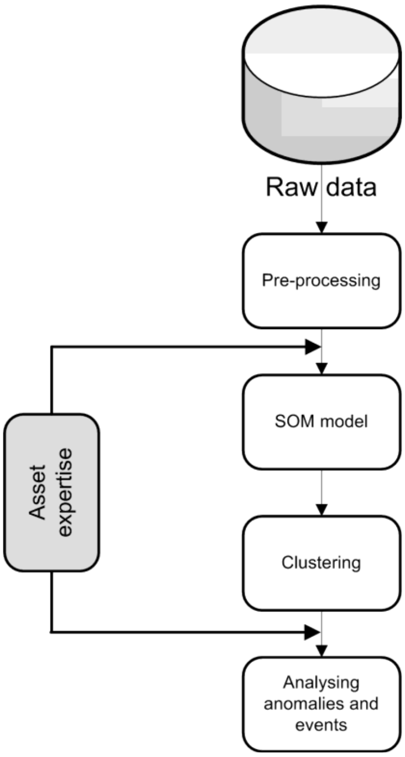

Figure 2. A diagram of the data processing.

The collected IAQ data consisted of continuous measurements of temperature, relative humidity and CO_2 concentrations in different apartments. Measured variables and their ranges are presented in Table 1. The size of the collected data matrix was 1 591 380 rows, 31 variables in columns. Missing data or outliers comprised only 0.6 percent of the data, and these rows were removed. The size of the final data matrix used was 1 581 992 rows, 31 variables in columns.

Table 1. Variables and their range.

Variable	**Range**
Temperature [ºC]	21.0-28.9
Relative humidity [%]	5.9-62.2
Carbon dioxide [ppm]	300.1-1975.4

2.3. Self-organizing map (SOM)

The self-organizing map (SOM) is a neural network algorithm developed by Teuvo Kohonen in the early 1980s, and its common purpose is to facilitate data analysis by mapping n-dimensional input vectors to the neurons, for example in a two dimensional lattice (Kohonen, 2001). In this lattice, the input vectors with common features result in the same or neighbouring neurons, preserving the topological order of the original data. The SOM learning process is unsupervised: no a priori classifications for the input vectors are needed. A large variety of SOM-based applications have been developed. The common application fields of SOM have been, for example machine vision, signal processing, exploratory data analysis and pattern recognition (Kohonen, 2001).

The training of SOM results in a topological arrangement of output neurons, each of which has a special reference vector describing its hits, or input vectors. Each neuron of the SOM is defined on one hand by this reference vector, which has the same dimensionality as the input vectors, and on the other hand by its location. The reference vector can be defined as follows:

$$r_m = (r_{m1}, r_{m2}, \dots, r_{mn}), \quad (m = 1, \dots, M), \tag{1}$$

where *n* is the number of variables, and *M* refers to the number of neurons in the map.

In the beginning of the training, the SOM must be initialized. In linear initialization, the SOM is initialized along the map dimensions, according to the greatest eigenvectors of the training data. In random initialization, the map is initialized by using arbitrary values for the reference vector. The use of linear initialization results in an ordered initial state for reference vectors instead of arbitrary values generated by random initialization (Kohonen, 2001).

The Best Matching Unit (BMU) is the neuron being at the smallest Euclidean distance from the input vector:

$$\beta(x_i, R) = \arg\ \min_j \left\| x_i - r_j \right\|, \tag{2}$$

where *β* is the index of the BMU, x_i denotes the input vector, and *R* includes the reference vectors of the SOM.

The BMU and the group of its neighbouring neurons can be trained using the following update rule: (Kohonen, 2001):

$$r_m(k+1) = r_m(k) + h_{\beta m}(k)[x_i - r_m(k)], \tag{3}$$

where *k* is the iteration round and *m* signifies the index of the neuron that is updated. The new reference vector becomes a weighted average of the original data samples assimilated to it. The neighbourhood function is often assumed to be Gaussian (Kohonen, 2001):

$$h_{\beta m}(k) = \alpha(k)\exp\left(-\frac{\|v_\beta - v_m\|^2}{2\sigma^2(k)}\right), \tag{4}$$

where $v_\beta$and v_m are the location vectors for the corresponding nodes, α refers to the factor of learning rate, and σ(k) defines the width of the kernel.

In summary, the training of the SOM proceeds as follows: 1) finding the BMU for one input vector according to the minimum Euclidean distance, 2) moving the reference vector (using the update rule) of the BMU towards that input vector, 3) moving the reference vectors (using the update rule) of neighbouring neurons towards that input vector, 4) Repeating steps 1-3 for the next input vector until all input vectors have been used, 5) Repeating steps 1-4 until the algorithm converges, 6) Finding the final BMU for each input vector according to the Euclidean distance.

2.4. Clustering

The K-means clustering is a well-known non-hierarchical cluster algorithm (MacQueen, 1967). The basic version begins by randomly picking K cluster centers, assigning each point to the cluster whose mean is closest in the sense of Euclidean-distances. The next steps involve computing the mean vectors of the points assigned to each cluster, and using these as new clusters in an iterative approach. They are determined by minimizing the sum of squared errors:

$$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - m_k)^2, \tag{5}$$

where x_i is a vector representing the *i*th data point and m_k is the centroid of the data points in C_k.

The number of clusters in the case specific application may not be known a priori. In the K-means algorithm the number of clusters has to be predefined. It is common that the algorithm is applied with different number of clusters, and the best solution is selected using a validity index (Davies and Bouldin, 1979) or asset expertise.

2.4. Methods in practice

The IAQ data were coded into inputs for the self-organizing map. All the input values were normalized by variance scaling before training the map. After that, a SOM having 100 neurons in a 10 x 10 hexagonal grid was constructed. Linear initialization and batch training algorithm were used in the training of the map. The Gaussian function was used as the neighbourhood function. The map was taught with 10 epochs and the initial neighbourhood had the value of 6. The SOM Toolbox version 2.0 (Aalto University, Laboratory of Computer and Information Science) was used in the data analysis under a Matlab-software platform (Mathworks, Natick, MA, USA).

The K-means algorithm was used for clustering the trained map, or precisely, to cluster the reference vectors. The Davies-Bouldin index (DBI) was used to evaluate the clustering. After clustering, the desired reference vector elements of clustered neurons were visualised in a two-dimensional space to reveal the possible interactions between variables.

3. RESULTS

Average values of measured CO_2 concentrations are presented in Figure 3 and all average values of measured parameters are presented in Table 2. Table 2 shows that average IAQ characterized by thermal conditions and CO_2 concentrations is the worst in room 8, which is a living room (8LR). For temperature, the Finnish guideline value is 21 °C and for relative humidity it is 20-60 % during the heating season (Asumisterveysohje, 2003).

Table 2: Averages of the variables at the 2^{nd} and the 6^{th} floor.

Floor 2	**6BR**	**6LR**	**8LR**	**10BR**	**10LR**
%RH	20.6	21.3	19.9	20.6	20.8
°C	24.2	24.5	25.8	24.0	23.6
CO_2	767.4	599.8	871.0	664.0	564.2
Floor 6	**26BR**	**26LR**	**28LR**	**30BR**	**30LR**
%RH	18.2	19.0	26.6	18.2	18.3
°C	24.5	24.4	24.1	23.8	23.7
CO_2	688.7	546.6	615.3	591.5	506.6

In this case, Figure 3 shows that there is a difference in enhance indoor environmental quality (IEQ) between the 2^{nd} and the 6^{th} floor indoor air quality. For example, the average values of measured CO_2 variables are lower on the 6^{th} floor.

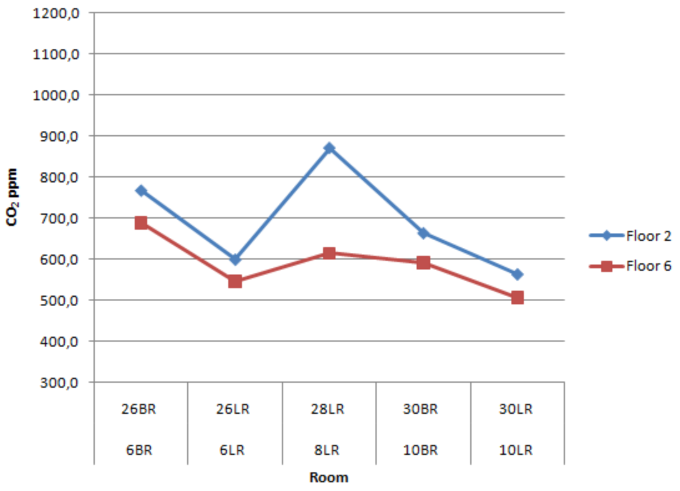

Figure 3. Average values of the measured CO_2 concentrations.

SOM component planes are presented in Figure 4, which indicates that relative humidity (RH) variables correlate. Thus, relative humidity seems to be similar in all rooms. In addition, the location of the apartment does not have an effect on relative humidity. However, there are major differences seen in the temperature readings and CO_2 concentrations between the rooms.

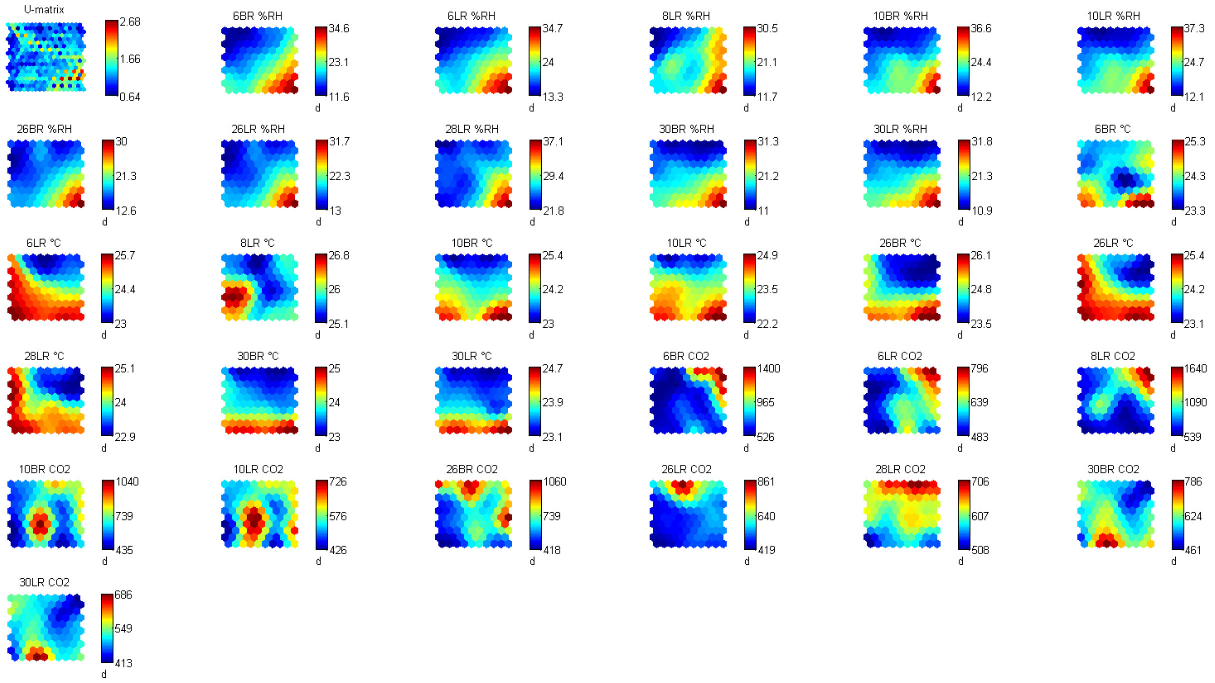

Figure 4. U-matrix and component planes from the SOM on the indoor air quality data. Rooms are marked as follows: living room is LR and bedroom is BR. 2nd floor rooms are named 6BR, 6LR, 8LR, 10 BR and 10LR. 6th floor rooms are named 26BR, 26LR, 28LR, 30BR and 30LR.

SOM was clustered according to the reference vectors by using the K-means algorithm (Figure 5). In this case, 8 clusters seemed to be adequate. The properties of the found clusters are:

- Cluster 1 describes the situation when the relative humidity is low and the temperature is at the normal level. Overall, CO_2 concentration seemed to be at the recommended level. Season is the early spring.
- Cluster 2 is similar to cluster 1. Only the room temperatures are slightly higher. Season is the early spring.
- Cluster 3 is common in the early summer, when the sun warms the indoor air, and relative humidity is at normal level.
- Cluster 4 is common in wintertime / cold season, when relative humidity is low (i.e. air is very dry).
- Cluster 5 is similar to cluster 4. In addition, it represents the situation when the CO_2 concentration is high in some apartments. This may be due to inadequate ventilation e.g. due to limiting ventilation in order to conserve energy, or due to high occupant density in this apartment or these apartments.
- Cluster 6 describes the situation where the indoor temperatures are high and the relative

humidity is below the recommended levels. It is typical for the early spring.

- Cluster 7 is similar to cluster 5. Only the CO_2 concentration is higher in different rooms.
- Cluster 8 represents the situation when indoor air quality is mainly good with respect to all measured parameters.

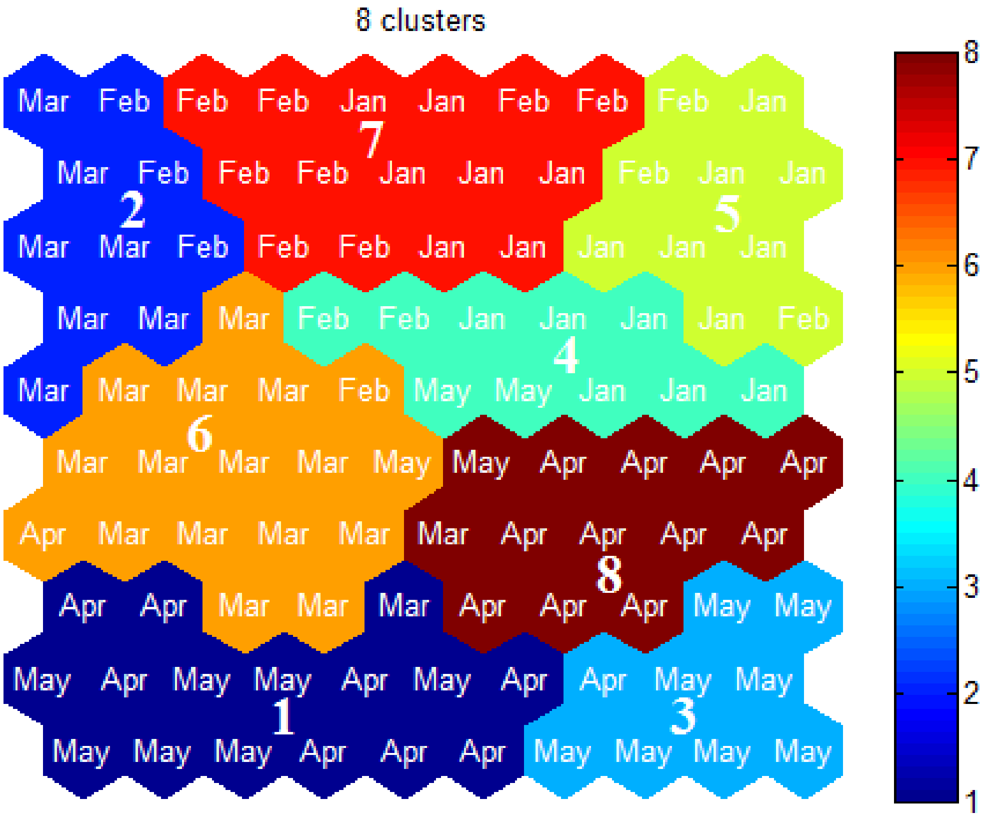

Figure 5. SOM using indoor air quality data of an apartment building.

Distribution of CO_2 variable of the 8 clusters has been examined in more detail in Figure 6. For clusters 5, 6 and 7, many situations can be found when the CO_2 concentration exceeds 1200 ppm.

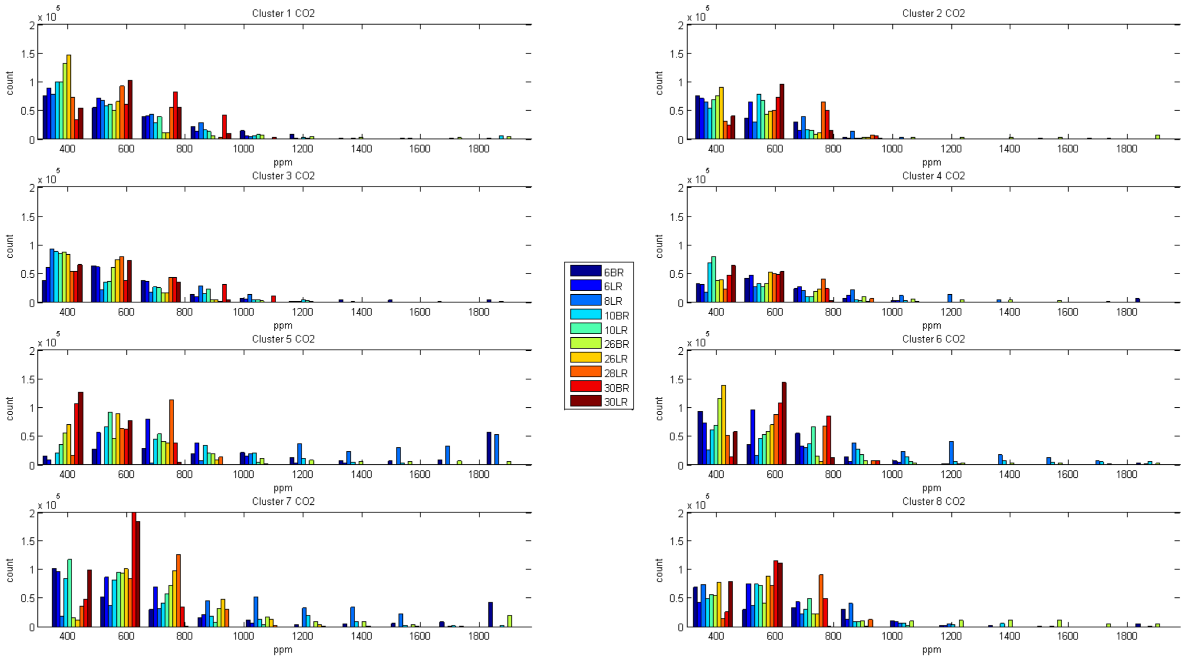

Figure 6. CO_2 variable distributions of the found clusters.

4. DISCUSSION AND CONCLUSION

In this study we tested the applicability of the neural network modelling method for analysing IAQ in an apartment building. The results indicate that the approach presented in this paper is a useful way to analyse anomalies and events on indoor air quality data. The SOM-based methodology can reveal dependencies between data variables relatively fast and easily. Several facts make the use of the SOM-based method useful in the analysis of indoor air quality data, because of easy visualisation, well-known and trusted tools in the Matlab-software platform. First of all, SOM-based method is good when the clustering behaviour of the data is not known before the data analysis.

The results also indicate, that the time of the year, as presented in Figure 5, describes well prevailing real indoor air quality from January to May in this building equipped with mechanical exhaust ventilation. In addition, the results indicate, that high CO_2 concentrations can increase above the recommended levels, thus mechanical exhaust ventilation may be commonly inadequate especially in the wintertime. Higher CO_2 concentrations are not only measured in the bedrooms, but are also found in the living rooms. For example, indoor air quality can be improved using mechanical supply and exhaust ventilation. In some cases, indoor CO_2 concentrations were below outdoor air CO_2 concentrations, which can be explained by drifting or inaccuracy of used measurement devices (Skön et al., 2011).

Nowadays, buildings are more energy efficient and airtight, which can have adverse effects on indoor air quality. Therefore, developing new IAQ monitoring systems and new methods for analysing the data are important. The results presented in this paper show, that the applied SOM-based neural network method is an efficient way to analyze indoor IAQ data. In the future, the study will be expanded to several apartment buildings, and the measurement period will also be extended. In addition, other application possibilities of neural network modelling or Fuzzy Logic and Linguistics will be explored in the field of energy efficiency and healthy housing.

ACKNOWLEDGMENT

This study was done as a part of the Finnish AsKo-project (*Asuinrakennusten korjaus- ja täydennysrakentamisen vaikutukset asumisen energiatehokkuuteen ja sisäilman laatuun (in Finnish)*; The effects of renovation and complimentary construction on energy efficiency and indoor air quality). For financial support, the authors would like to thank The Housing Finance and Development Centre of Finland (ARA). In addition, the authors would like to thank Niiralan Kulma Oy for good cooperation.

REFERENCES

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

Apte, M. G., Fisk, W. J. and Daisey, J. M., 2000, Association between indoor (CO_2) concentrations and sick building syndrome symptoms in US Office Buildings: an analysis of the 1994-1996 BASE study data, *Indoor Air* **10**, 246-257.

Armstrong, R. and Spiller, N.,2010, Synthetic biology: Living quarters, *Nature* **467**, 916-918.

Asumisterveysohje, 2003, *Sosiaali- ja terveysministeriön oppaita* **2003:1**, Sosiaali- ja terveysministeriö, Oy Edita Ab, Helsinki (in Finnish).

Butler, D., 2008, Architects of a Low-energy Future, *Nature* **452**, 520-523.

Daisey, J. M., Angell, W. J. and Apte, M. G., 2003, Indoor air quality, ventilaition and health symptoms in schools: an analysis of existing information, *Indoor Air* **13**, 53-64.

Davies, D. and Bouldin, D. A., 1979, Cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1(2)***,* 224-227.

Gershenfeld, N., Samouhos, S. and Nordman, B., 2010, Intelligent infrastructure for energy efficiency, *Science* **372**, 1086-1088.

Holdren, J. P., 2007, Energy and sustainability, *Science* **315(5813)**, 737.

Jackson, R. J., 2007, Environment meets health, again, *Science* **315(5817)**, 1337.

Jones, A. P., 1999, Indoor air quality and health, *Atmospheric Environment* **33**, 4535-4564.

Kim, M.H., Kim, Y.S., Lim, J.J., Kim, J.T., Sung, S.W. and Yoo, C.K., 2010, Data-driven prediction model of indoor air quality in an underground space, *Korean Journal of Chemical Engineering* **27(6)**, 1675-1680.

Kohonen, T., 2001, *Self-organizing maps*, 3rd ed., Springer-Verlag, Berlin Heidelberg.

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

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

MacQueen, J., 1967, Some methods for classification and analysis of multivariate observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability,* 281-297.

Niska, H., Hiltunen, T., Kolehmainen, M. and Ruuskanen, J., 2003, Hybrid models for forecasting air pollution episodes, *International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA'03)*, Springer-Verlag, Roanne, France, 80-84.

Salthammer, T., 2011, Critical evaluation of approaches in setting indoor air quality guidelines and reference values, *Chemosphere* **82**, 1507-1517.

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

Skön, J-P., Kauhanen, O. and Kolehmainen, M., 2011, Energy consumption and air quality monitoring system, *Proceedings of 7th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2011)*, Adelaide, Australia, *163-167*.

Sofuoglu, S.C. 2007. Application of artificial neural networks to predict prevalence of building-related symptoms in office buildings, *Building and Environment* **43(6)**, 1121-1126.

Sofuoglu, S. C., Aslan, G., Inal, F. and Sofuoglu, A., 2011, An assessment of indoor air concentrations and health risks of volatile organic compounds in three primary schools, *International Journal of Hygiene and Environment Health* **214**, 36-46.

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