International Journal of Science, Environment and Technology, Vol. 4, No 2, 2015, 318 – 325

ISSN 2278-3687 (O)
2277-663X (P)

DATA ANALYSIS MODELS OF INDOOR POLLUTION

Simona Skorupskaitė[1], Rūta Sidaravičiūtė[2], Dainius Martuzevičius[2] and Tomas Ruzgas[1*]

[1]Department of Applied Mathematics, Kaunas University of Technology, Lithuania
[2]Department of Environmental Technologies, Kaunas University of Technology, Lithuania
E-mail: tomas.ruzgas@ktu.lt ([*]*Corresponding Author*)

Abstract: Chemical air pollution may be several times higher in our living environment than in the busiest street. Usually just outdoor air quality is repeatedly tracked. The public is being immediately informed in the case of hazard, while pollution monitoring in private properties is exclusively its residents concern. This paper presents data analysis models, designed for problems of indoor pollution reduction with lightened decision making and motivation.
The models were developed using applied statistics methods (discriptive statistics, analysis of variance, nonparametric tests, correlation, cluster analysis, and decision trees) as well as multidimensional reporting and visualisation. The developed system is used for the analysis of indoor pollution of East European city (Kaunas, Lithuania). It enables to analyse the relationships and dependencies between various factors (e.g. number of residents, CO_2, NO_2, benzene, toluene, ethylbenzene, xylene, formaldehyde, radon) and use the results for the effective decision making.
Keywords: Indoor pollution, clustering, decision tree.

Introduction

People spend an average 90% of their time indoors, thus indoor air quality (IAQ) has an important influence on human wellbeing and health (Klepeis et al., 2001; Tan et al., 2013). Because of relatively long time people spend indoors the bigger exposure to air pollutants occurs. IAQ is also one of the main indoor environment parameters contributing to human satisfactory indoor environment (Frontczak & Wargocki, 2011). About 66% of the Lithuanian population lives in multi-family houses built before 1990 and most of the current buildings still rely on natural ventilation which not always ensures adequate removal of indoor pollutants (Prasauskas, 2014). It is widely known that for a large number of pollutants indoor, pollution level is higher than outdoor. Potential indoor sources of air pollution are numerous, including building material and equipment, adhesives (continuous generation of pollutants); cleaning and personal care products, tobacco smoke (intermittent generation of pollutants); bio-contaminants, combustion processes such as cooking and heating, ventilation

*Received Feb 25, 2015 * Published April 2, 2015 * www.ijset.net*

practice (Jones, 1999; Billionnet et al., 2011; Tan et al., 2013). Major of indoor pollutants are recognized to be volatile organic compounds (VOCs) as well as formaldehyde, nitrogen dioxide (NO_2), carbon dioxide (CO_2), particulate matters (PM), biocontaminants (Jones, 1999).

Continuous exposure to indoors pollutants may influence bigger negative effects to human‘s health, this may be true especially for chronic illnesses that often manifest symptoms long after the initial exposures (Hawthorne et al., 1986). The most sensitive effects of VOCs inhalation are neurological and irritative to respiratory tract, although formaldehyde, benzene are classified in Group 1 of human carcinogens by the International Agency for Research on Cancer (IARC, 2004). Darby et al. (2005) and Krewski et al. (2006) claim, radon is second highest reason of lung cancer after smoking.

20 multi-family buildings, located in urban area of Kaunas city were selected for the study of the IAQ in Lithuania. The average age of selected buildings was 39 years. Average five apartments per house located in different part of a building were selected during heating seasons 2011 – 2013 (December 2011 – March 2012 and November 2012 – April 2013). The measured indoor air gaseous pollutants included carbon dioxide (CO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), radon (Rn), formaldehyde and BTEX (benzene, toluene, ethylbenzene, xylenes). During the gaseous pollutants measurements indoor temperature, relative humidity and effectiveness of ventilation were recorded (Prasauskas, 2014). The objective of this paper is to determine the existance or non existance of dependance as well as the strength of relantionship between descriptive factors and concentration of acomodation compounds. The dataset has 95 observations. Each apartment has levels for flooring, type of stove, type of ventilation, floor number, distance from a road, number of residents, indoor maintenance – wear out of furniture and repair.

Methods and Models

Observed data has been used to analyse and make assumptions about the required information for decition making. Thus, the proposed statistical analysis models are:

1. Reports of descriptive statistics with OLAP application.
2. Goodness of fit hypothesis testing to determine distribusion of values

(*CO_2*, *benzene*, *toluene,* etc.).

3. Analysis of variance

- Analysis of factors for *number of residents*, *floor*, *distance from a road*, *ventiliation type*, *stove type*, *flooring type*, *maintenante* and furniture impact for CO_2, NO_2, *benzene*, *toluene*, *ethylbenzene*, *xylene*, *formaldehyde* and *radon* concentracion differences of means.

4. Correlation analysis

- Analysis of correlation between air pollutants concentracions.

5. Cluster analysis

- Classification of apartments by origin of air pollutants: combustion emissions, construction materials, cleaning agents.

6. Decision trees

- Classification of apartments by *number of residents* classes.

The purpose of analysis of variance is to determine differences of *maintenance*, *furniture*, *number of residents*, *stove type*, *distance form the road*, *floor levels* means for NO_2, CO_2, *benzene*, *toluene*, *xylene*, *ethyl-benzene*, *formaldehyde* and *radon* concentrations.

Flats clustering using the EM algorithm. If the density function of the random vector X has q maxima, it can be approximated by a mixture of q unimodal densities:

$$f(x) = \sum_{k=1}^{q} p_k f_k(x). \tag{1}$$

Let the distribution of the random vector X depend on a random variable v that takes on the values 1,...,q. It is interpreted as the number of class the observed object belongs to. In the classification theory quantities $p_k = \mathbf{P}\{v=k\}$ are called *a priori* probabilities that the observed object belongs to the k^{th} class, and quantities $\pi_k(x) = \mathbf{P}\{v=k|X=x\}$ are *a posteriori* probabilities. The function f_k is treated as the conditional density of X as $v=k$. By the term soft clustering of a sample we refer to the estimation of the values $\pi_k(X(t))$ for all $k = 1,..., q$, $t = 1,..., n$. A sample is hard-clustered if estimators $\hat{v}(1),...,\hat{v}(n)$ of $v(1),\ldots, v(n)$ are indicated where $v(t)$ denotes the class number of the *flat* $X(t)$.

The mixture of Gaussian distributions is the most popular model in the clusterisation theory and practice. Therefore, in this section we assume that $f_k(x)$ are Gaussian densities with means M_k and covariance matrices R_k. Let $f(\theta,x)$ denote the right-hand side of equation (1), where $\theta = ((p_k, M_k, R_k)\ k=1,\ldots,q)$. Since

$$\pi_k(x) = \frac{p_k f_k(x)}{f(\theta, x)},\ k = \overline{1,q}. \tag{2}$$

the estimators of *a posteriori* probabilities are obtained as usual by the "plug-in" method which replaces the unknown parameter vector θ on the right side of (2) by its maximum

likelihood estimate $\theta^* = \arg\max_{\theta} L(\theta)$, $L(\theta) = \prod_{t=1}^{n} f(\theta, X(t))$. The EM algorithm, an iterative procedure most frequently used to find this estimate, was also applied in this study.

Let $\hat{\pi}_k = \hat{\pi}_k^{(r)}$ be the estimates obtained after r cycles of the iterative procedure. Then a new estimate $\hat{\theta} = \hat{\theta}^{(r+1)}$ is defined by the equalities

$$\hat{p}_k = \frac{1}{n}\sum_{t=1}^{n} \hat{\pi}(X(t)),$$

$$\hat{M}_k = \frac{1}{n\hat{p}_k}\sum_{t=1}^{n} \hat{\pi}(X(t)) \cdot X(t),$$

$$\hat{R}_k = \frac{1}{n\hat{p}_k}\sum_{t=1}^{n} \hat{\pi}(X(t)) \cdot [X(t) - \hat{M}_k] \cdot [X(t) - \hat{M}_k]'$$

for all $k = 1,\ldots,q$. Inserting $\hat{\theta}^{(r+1)}$ into the right-hand side of expression (2) we get $\hat{\pi}^{(r+1)}(X(t))$, $k = \overline{1,q}$, $t = \overline{1,n}$. Using this iterative procedure a non-decreasing sequence $L(\hat{\theta}^{(r)})$ it obtained, however its convergence to the global maximum depends on the initial value $\hat{\theta}^{(0)}$ (or $\hat{\pi}^{(0)}$). The simplest solution of the initial value selection problem is a random start technique: the EM procedure is repeated many times from the random starting value $\hat{\pi}^{(0)}$. The result with the maximal value of $L(\hat{\theta})$ is selected as final. The methodology of consecutive extraction of the mixture components (Rudzkis & Radavicius, 1995) can be also applied as well.

Decision trees are generated with ID3 and CART algorithms invented by Quinlan (1986) and Breimanet al. (1984). Both of them uses the particular amount of uncertainty as a split criteria. Every iteration ID3 measures the entropy $H(S)$, then the attribute with the smallest entropy is used for dataset split. Algorithm uses attributes for splitting only once.

$$H(S) = -\sum_{x \in X} p(x) log_2(x)$$

where S is the dataset of a particular iteration, X is set of S classes, $p(x)$ is the proportion of S belonging to class x. The idea of entropy value is that when $H(S) = 0$ all elements of S belong to one class.

Instead of entropy, as a measure of a quality of a split, CART uses gini index. If attribute takes q different classes for a dataset S, then Gini index of S is defined:

$$Gini(S) = 1 - \sum_{k=1}^{q} p_i^2$$

where p_k is the probability of S belonging to class k. Gini coefficient has main idea that it equals to one if S has no split, and index has its maximum if a probability distribution of p is uniform. In other words, if elements are labelled random accordingly to the distribution of labels, Gini index measures the frequency of an element to be incorrectly labelled.

Examples of Results

The apartments are classified into homogenous groups that are called clusters. The selection of clustering variables are made according to their correlations (orthogonal variables forms stable clusters). Thus, the apartments are clustered by:

- CO_2, benzene, formaldehyde, NO_2
- Benzene, formaldehyde, radon

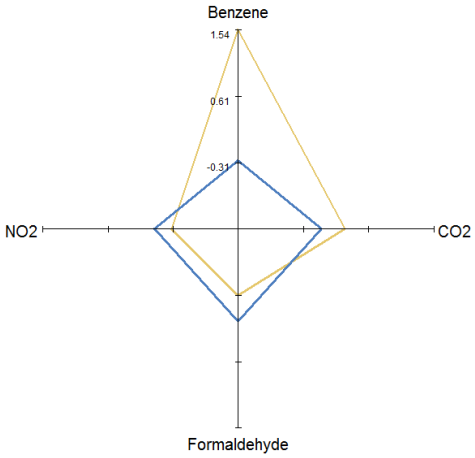

Figure1: Star plot of cluster centers. Clustering variables: CO_2, benzene, formmal dehyde, NO_2

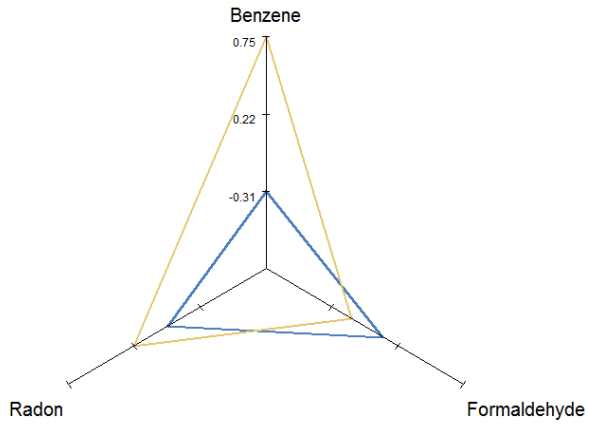

Figure2: Star plot of cluster centers. Clustering variables: benzene, formaldehyde, radon.

Obtained two types of apartments: (figure 1 and figure 2):

- Indoor air in the first type is characterized by high benzene concentration. CO_2 concentration found to be higher in the same group either.
- Second type of apartments have higher formaldehyde amount that is observed with higher NO_2 concentration and significantly lower concentration of benzene.

CO_2, NO_2, temperature and humidity amounts are used to classify apartments to different number of residents groups. Decision trees are constructed with ID3 and CART algorithms with training sample that is 70% of all apartments. ID3 classifies apartments by humidity value (figure 3), while CART uses all variables for classification (figure 4). The accuracy is

estimated for the rest of 30% of all apartments. ID3 and CART classifies with respectively 74% and 69% accuracy. Thus, different number of residents classes can be formed by observed humidity.

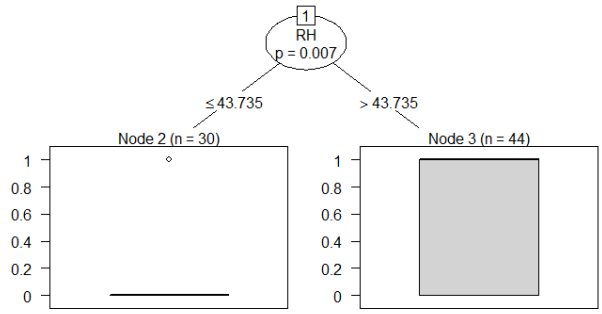

Figure 3: Decision tree for number of residents, ID3

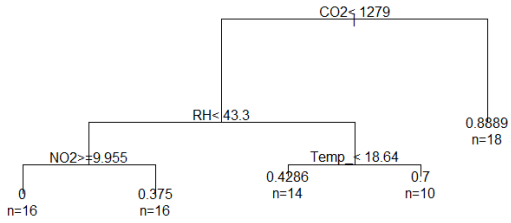

Figure 4: Decision tree for number of residents, CART

Conclusions and Recommendations

Application and combination of descriptive statistics, analysis of correlation, analysis of clustering and other statistical methods lead up statistical models of indoor pollution that can be used for determination of various factors relationships and dependences.

After real data analysis the following results were obtained:

- The number of residents have an impact to CO_2 concentration: the measured average concentration of CO_2 for apartments with two and less residents is 846 µg/m^3, the concentration average for more than two residents – 1124 µg/m^3. Factor impacts 9% of the concentration value.
- The type of stove impacts the concentration of NO_2 (9%): the average of NO_2 was 14.84 µg/m^3 in gas stove apartments, while other stove type apartments had an average of 8.32 µg/m^3.
- Benzene concentration has relationship with indoor maintenance. Not renovated apartments with old (more than 5 years) furniture has the highest values of benzene concentration (the average is 8.045 µg/m^3). Otherwise, recently renovated and holding new furniture apartments has the average of 3.348 µg/m^3 concentration.
- Toluene, ethylbenzene, xylene and formaldehyde concentrations are dependent on the distance from a road and the floor level. Highest concentrations of these compounds were observed in apartments that are further from a street and has the level up to 3rd floor (the average respectively is 13.84 µg/m^3, 3.26 µg/m^3, 8.61 µg/m^3, 30.45 µg/m^3). Lower

compounds concentrations were measured in apartments that are closer to a street and has higher than 3rd floor level (the average respectively is 4.82 $\mu g/m^3$, 0.78 $\mu g/m^3$, 2.43 $\mu g/m^3$, and 20 $\mu g/m^3$).

IAQ should be a priority in all respects. Since main sources of many pollutants are indoors, the source control, such as the use of low-emission of household products, the assurance of good ventilation especially during the cooking and combustion processes, could be an effective way to control indoor pollution. On the other hand, the quality of outdoor air entering the room should be considered.

References

[1] Billionnet C., Gay E., Kirchner S., Leynaert B., Annesi-Maesano I. (2011). Quantitative assessments of indoor air pollution and respiratory health in a population-based sample of French dwellings. *Environmental Research, 111*, 425-434.

[2] Breiman L. Friedman J.H., Olshen R.A., Stone C.J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

[3] Darby S., Hill D., Auvinen A. et al. (2005). Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case–control studies. *BMJ*, 330, 223-227.

[4] Frontczak M., Wargocki P. (2011). Literature survey on how different factors influence human comfort in indoor environments. *Building and Environment, 46*, 922-937.

[5] Hawthorne A. R., Gammage R. B., Dudney C. S. (1986). An indoor air quality study of 40 east Tennessee homes, *Environmental international, 12*, 221-239.

[6] IARC(2004). Overall evaluation of Carcinogenicity to Humans, Formaldehyde [50-00-0], Monographs Series, 88. International Agency for Research on Cancer, Lyon, France.

[7] Jones A.P. (1999). Indoor air quality and health. *Atmospheric Environment, 33*, 4535-4564.

[8] Klepeis N.E., Nelson W.C., Ot W. R., Robinson J.P., Tsang A.M., Switzer P., Behar J.V., Hern S.C., Engelmann W.H. (2001). The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Analysis and Environmental Epidemiology, 11*, 231-252.

[9] Krewski D., Lubin J.H., Zielinski J.M., et al. (2006). A combined analysis of North American case–control studies of residential radon and lung cancer. *J Toxicol Environ Health A*, *69* (*7*), 533-597.

[10] Prasauskas T., Martuzevicius D., Krugly E., Ciuzas D., Stasiulaitiene I., Sidaraviciute R.,

Kauneliene V., Seduikyte L., Jurelionis A., Haverinen-Shaughnessy U. (2014). Spatial and temporal variations of particulate matter concentrations in multifamily apartment buildings. *Building and Environment, 76*, 10-17.

[11] Quinlan J.R. (1986). Induction of Decision Trees. *Machine Learning*, *1*, 81-106.

[12] Rudzkis R., Radavicius M. (1995). Statistical Estimations of a Mixture of Gaussian Distributions. *Acta Applicandae Mathematicae, 38,* 37-54.

[13] Tan C.C.L., Finney K.N., Chen Q., Russell N.V., Sharifi V.N., andSwithenbank J. (2013) Experimental Investigation of Indoor Air Pollutants in Residential Buildings, *Indoor and Built Environment*, *22*(*3*), 471-489.