COMPARISON OF MULTIVARIATE CLASSIFICATION AND REGRESSION METHODS FOR THE INDOOR RADON MEASUREMENTS

by

Dimitrije M. MALETIĆ, Vladimir I. UDOVIČIĆ*, Radomir M. BANJANAC, Dejan R. JOKOVIĆ, Aleksandar L. DRAGIĆ, Nikola B. VESELINOVIĆ, and Jelena Z. FILIPOVIĆ

Institute of Physics, University of Belgrade, Belgrade, Serbia

Scientific paper
DOI: 10.2298/NTRP1401017M

We present the results of a test usage of multivariate methods, as developed for data analysis in high-energy physics and implemented in the toolkit for multivariate analysis software package, in our analysis of the dependence of the variation of indoor radon concentration on climate variables. The method enables the investigation of the connections of the wide spectrum of climate variables with radon concentrations. We find that multivariate classification and regression methods work well, giving new information and indications, which may be helpful in further research of the variation of radon concentration in indoor spaces. The method may also lead to considerable prediction power of the variations of indoor radon concentrations based on the knowledge of climate variables only.

Key words: radon, multivariate analysis, climate parameter

INTRODUCTION

Radon is a unique natural element since it is a gas, noble and radioactive in all of its isotopes. As noble gases, radon isotopes are mobile and can travel significant distances within the ground and through the atmosphere. Being radioactive, radon makes for about 55% of the annual effective dose received by average non-professional. Indoor radon concentrations vary significantly due to a large number of factors, which include the local geology, soil permeability, building materials and lifestyle characteristics, climate parameters and the exchange rate between indoor and outdoor air. Since both the climate parameters and air exchange rates may significantly vary during a day, it is important to investigate their correlation with short-term variations of indoor radon concentrations. In the past somewhat unusual climate parameters, such as wind speed and cloud cover, were occasionally considered, using a multivariate method [1-3]. We start this analysis with the maximum of 18 climate parameters and use and compare 12 different multivariate methods.

Variations of radon concentration were studied in our laboratory [4] in many details since 1999 [5-8]. Several climate variables, like air temperature, pressure, and humidity were considered [8, 9]. We now make further advance and try to use all publicly available climate variables monitored by, in our case, nearby automatic meteorological station (Automatic Meteorological Station Belgrade-south, Banjica-Trošarina, 44°45'16"N, 20°29'21"E). We want to find the appropriate method out of the wide spectrum of multivariate analysis methods that are developed for the analysis of data from high-energy physics experiments to analyze our measurements of variations of radon concentrations in indoor spaces.

FORMULATION OF THE PROBLEM

The demand for detailed analyses of large amount of data in high-energy physics resulted in wide and intense development and usage of multivariate methods. Many of multivariate methods and algorithms for classification and regression are already integrated into the analysis framework ROOT [10], more specifically, into the toolkit for multivariate analysis (TMVA) [11]. We use these multivariate methods to create, test and apply all available classifiers and regression methods implemented in the TMVA in order to find the method that would be the most appropriate

* Corresponding author; e-mail: udovicic@ipb.ac.rs

and yield maximum information on the dependence of indoor radon concentrations on the multitude of climate variables.

The first step is to calculate and rank the correlation coefficients between all the variables involved, what will help in setting up and testing the framework for running the various multivariate methods contained in the TMVA. Although these correlation rankings will later be superseded by method-specific variable rankings, they are useful at the beginning of the analysis.

The next step is to use and compare the multivariate methods in order to find out which one is best suited for classification (division) of radon concentrations into what would be considered acceptable and what would be considered increased concentration in indoor spaces. Main aim is to find out which method can, if any, on the basis of input climate variables only, give an output that would satisfactorily close match the observed variations of radon concentrations. This would enable the creation of the "radon alarm" using only the multivariate classification of the now widely available records of climate variables. Towards this aim, this work should be considered a preliminary one, for the number of specific cases that should be studied in this way should be much larger, to comprise the multitude of possible representative situations that occur in real life.

In order to be able to use the multivariate classification, the set of input events (values for climate variables for each measurement) used, have to be split into those that correspond to the signal (the radon concentrations that are considered increased) and to the background (consisting of radon concentrations that are declared acceptable). This splitting of the set of input events is for the purposes of this preliminary analysis performed at the limiting value of 40 Bq/m^3. This value is used for most of the analyses, and is selected because this splitting ensures maximum employment of multivariate comparison methods, and this particular value reflects the fact that in our test case the statistics on higher radon concentration values are lower. For the purposes of setting of a sort of a "radon alarm", the value of radon concentration that should be used for splitting of input events is the value for radon concentration recommended by World health organization of 100 Bq/m^3. The method of multivariate regression, however, does not require preliminary splitting of input events, and is therefore a more general one.

EXPERIMENTAL DATA

There are many methods available for measurement of radon concentrations in air. According to the integrating measurement time, these may be divided into the long-term and short-term ones. The first are mostly performed with passive integrating measuring devices based on nuclear track detectors, which are due to their low cost, simplicity, and wide availability well suited for simultaneous collection of data from a large number of measurement points and are thus used in large radon mapping projects. The second group comprises the methods that are performed with more complex and more expensive passive or active (with pumped air sampling) devices. For the short-term measurements of radon concentration in a single-family dwelling house in Belgrade, Serbia, we use the SN1029 radon monitor (manufactured by the Sun Nuclear Corporation, NRSB approval-code 31822). The device consists of two diffused junction photodiodes as a radon detector, and is furnished with sensors for temperature, barometric pressure and relative humidity. The user can set the measurement intervals from 30 minutes to 24 hours. It was set to record simultaneously the radon concentration, temperature, atmospheric pressure and relative humidity.

The selected house to measure the temporal variations of radon concentration is a typical one-family detached dwelling house built with standard construction materials such as brick, concrete, and mortar. The house is thermally insulated with Styrofoam. During the period of measurements (summer), the house was naturally ventilated and air conditioning was used during the hottest days. The indoor radon measurements were performed in the living room, where family spends anything from 16 up to 24 hours during the working days of the week. Radon monitor was measuring radon concentration, temperature, pressure, and humidity at 2 hour intervals, starting from the 3rd of June till the 3rd of July and from the 18th of July till the 11th of August 2013.

The values of climate variables, which will be correlated with radon monitor results, are obtained from a modern automatic meteorological station located some 400 m (GPS coordinates) away from the house where the radon monitor was placed. The wide set of climate variables were used, for the measurements of which were performed at 5 minute intervals during June, July, and August 2013. The fifteen climate parameters used are: outdoor air temperature, pressure and humidity, solar irradiance, wind speed at the height of 10 m above the ground, precipitation, evaporation, and underground temperature and humidity at the depths of 10-30 and 50 cm.

The second site used for the tests is our own ground level laboratory [1], which is air-conditioned and only rarely accessed, thus having much more stable indoor conditions than the dwelling house described. The measurements were performed during September and October 2012. Measurements of climate parameters that will be combined with radon measurements in this case come from the different, and somewhat older automatic metrological station, located about 4 km from the laboratory where the radon monitor was taking data.

MULTIVARIATE METHODS

The TMVA provides a ROOT-integrated environment for the processing, parallel evaluation and application of multivariate classification and multivariate regression methods. All multivariate methods in TMVA belong to the family of "supervised learning" algorithms. They make use of training events, for which the desired output is known, to determine the mapping function that either describes a decision boundary (classification) or an approximation of the underlying functional behavior defining the target value (regression). All MVA methods see the same training and test data. The correlation coefficients of the input variables are calculated and displayed, and a preliminary ranking is derived (which is later superseded by method-specific variable rankings). For standalone use of the trained classifiers, TMVA also generates lightweight C++ response classes that do not depend on TMVA or ROOT, neither on any other external library. As will be demonstrated, the two most important multivariate methods for our purposes are the boosted decision trees (BDT) and the artificial neural networks (ANN) methods.

Boosted decision trees

BDT has been successfully used in high energy physics analysis for example by the MiniBooNE experiment [12]. In BDT, the selection is done on a majority vote on the result of several decision trees. Decision tree consists of successive decision nodes, which are used to categorize the events in sample as either signal or background. Each node uses only a single discriminating variable to decide if the event is signal-like "goes right" or background-like "goes left". This forms a tree like structure with "baskets" at the end (leave nodes), and an event is classified as either signal or background according to whether the basket where it ends up has been classified as signal or background during the training. Typically, BDT is constructed of a forest of such decision trees. The (final) classification for an event is based on a majority vote of the classifications done by each tree in the forest. However, the advantage of the straightforward interpretation of the decision tree is lost. In many academic examples with more complex correlations or real life examples, the BDT often outperform the other techniques. More detailed information about training can be found in [11].

Artificial neural networks

An artificial neural network (ANN) [13] is most generally speaking any simulated collection of interconnected neurons, with each neuron producing a certain response at a given set of input signals. By applying an external signal to some (input) neurons the network is put into a defined state that can be measured from the response of one or several (output) neurons.

ANN in TMVA belong to the class of multilayer perceptrons (MLP), which are feed-forward neural networks. The input layer contains as many neurons as input variables used in the MVA. The output layer contains a single neuron for the signal weight. In between the input and output layers are a variable number of k hidden layers with arbitrary numbers of neurons.

All neuron inputs to a layer are linear combinations of the neuron output of the previous layer. The transfer from input to output within a neuron is performed by means of an "activation function". In general, the activation function of a neuron can be zero (deactivated), one (linear), or non-linear. The ANN used for our purposes uses a sigmoid activation function. The transfer function of the output layer is usually linear.

RESULTS

We comment on the results of our analyses divided into cases that differ by the size of the set of climate parameters used, by the indoor space studied, and by the methods of analysis used.

First, we intercompare the multivariate methods used for classification of radon concentrations by using the full set of climate variables as described in previous sections.

We are using the input events (set of climate variables for each measurement) to train, test and evaluate the 12 multivariate methods implemented in TMVA. The graph presenting the receiver operating characteristic (ROC) for each multivariate method (fig. 1) may be considered as the most indicative in comparing the different methods used for classification of radon concentrations using climate variables. On this graph one can read the dependence of background rejection on signal efficiency. The best method is the one that holds maximum value of background rejection for highest signal efficiency, *i. e*. the best method has ROC curve closest to the upper right corner on the graph presented in fig. 1. It turns out that the method best suited for our purpose is the BDT method. This means that BDT gives most efficient classification of input events. This is seen in fig. 2, which shows the distribution of BDT classification method outputs for input signal and background events. The second best method is the implementation of ANN MLP.

In fig. 3, one can see the values of signal and background efficiency and significance. Significance, calculated as

$$\frac{N(\text{signal})}{\sqrt{N(\text{signal}) \quad N(\text{background})}}$$

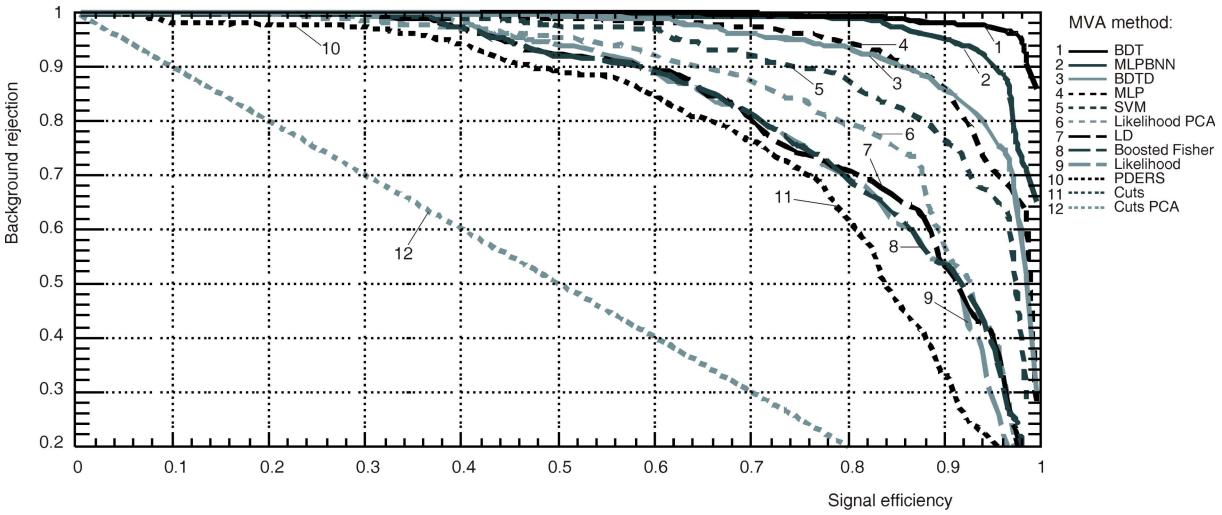

Figure 1. ROC for all multivariate methods used for classification of radon concentration using climate variables

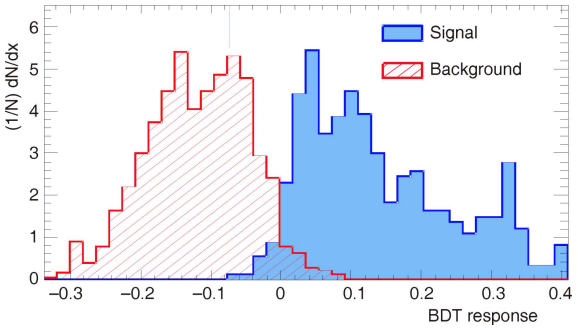

Figure 2. Distribution of BDT classification method outputs for input signal and background events

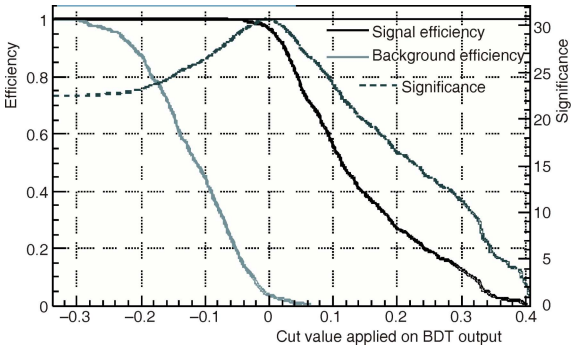

Figure 3. Cut efficiency and optimal cut value of BDT classification MVA method

can be used as the value for comparison of various multivariate methods, and also for comparison of method efficiencies for different sets of input variables. The significance of the BDT method with full set of input climate variables turns out to be 30.6. Ranking of the BDT input variables (tab. 1.) is derived by counting how often the variables are used to split decision tree nodes, and by weighting each split occurrence by the separation it has achieved and by the number of events in the node. As seen from tab. 1, temperature of the soil at the depth of 10 cm appears to be by far the most important variable.

Now we compare the multivariate methods for classification of radon concentration by using the minimum set of climate variables that would give similar results as when using the full set. While searching for the best multivariate method for radon classification indoors in this situation, we found that the BDT method again gives the best result, with the significance of 29.6 as compared to 30.6, when all the available climate variables for training and testing of multivariate methods are used. The climate variables chosen for training and testing in this case were: outdoor air temperature, humidity and pressure, outdoor soil temperature at the depth of 10 cm, differences of

Table 1. Ranking of BDT input variables

Variable	Variable importance
Temperature of soil at depth of 10 cm	1.37e-01*
Outside air temperature	7.40e-02
Evaporation	7.16e-02
Outside air pressure	7.16e-02
P (outside) – P (radon monitor)	6.51e-02
Outside air humidity	6.40e-02
H (outside) – H (radon monitor)	6.12e-02
T (outside) – T (radon monitor)	5.79e-02
Humidity of soil at depth of 10 cm	5.74e-02
Solar irradiance	5.16e-02
Temperature of soil at depth of 20 cm	4.99e-02
Temperature of soil at depth of 50 cm	4.68e-02
Temperature of soil at depth of 30 cm	4.46e-02
Humidity of soil at depth of 20 cm	4.31e-02
Wind speed at height of 10 m	3.87e-02
Humidity of soil at depth of 30 cm	3.41e-02
Humidity of soil at depth of 50 cm	3.13e-02
Precipitation	0.00e+00

*1.37e-01 read as $1.37\ 10^{-1}$

outdoor and indoors temperature, and the indoors humidity and pressure. One important caveat is in place here. It concerns the possibility that the two sets of instruments (for indoor and outdoor measurements) are not identically calibrated, what may especially be the case when two different groups or institutions conduct the indoor and outdoor measurements. It is estimated that these instrumental effects do not influence significantly the results of this study. In the case of calibration of MVA classification method, we need radon monitor apparatus indoors and apparatus for P, H, and T measurements outdoors and an apparatus for measurement of the outdoor soil temperature with the sensor positioned at the soil depth of 10 cm. While aiming at setting a "radon alarm" in this case, we thus have to have two apparatuses for P, H, and T measurements, indoor and outdoor, and an apparatus for measurement of outdoor soil temperature with the sensor positioned at the depth of 10 cm.

Next we compare the uses of multivariate methods for classification of radon concentration indoors when using the simplest possible set of climate variables. The climate variables used for training and testing were: outdoor air temperature, pressure and humidity, and differences of outdoor and indoor temperature, pressure and humidity. That means that we need to have two devices for measurement and recording of temperature, pressure and humidity, both indoors and outdoors at the same time. For calibration and testing of multivariate methods, in case of using this set of climate variables we would need one radon monitor indoors, and an apparatus for measurement of P, H, and T outdoors. For the purpose of seting the radon alarm, we would need to have two apparatuses for P, H, and T measurement. The best multivariate method for radon classification indoors in this case is also BDT method. The resulting significance is 28.2 as compared to 30.6 what we get when using the full set of available climate variables for training and testing of multivariate methods. This testifies that when we drop out many climate parameters in this case of analysis the resulting significance decreases notably, but still leaving MVA classification work good.

We also compared the multivariate methods for classification of radon concentration using the simplest set of climate variables in our Ground level laboratory, which is, as said, an air-conditioned and only seldom accessed space. The climate data are provided by the 4 km away and somewhat older automatic meteorological station. The methods are still found to work satisfactorily – the resulting significance of the BDT method now being 27.6 as compared to 28.2, obtained with the simplest set of variables in the case of the actively inhabited dwelling. The climate variables, requirements for training and testing are the same as in the previous case.

We also tested the simple set of only outdoor measured climate variables consisting of the outdoor air temperature, pressure and humidity, and the outdoor soil temperature at the depth of 10 cm. This means that the devices for measurement and recording of outdoor temperature, pressure and humidity as well as the device for measurement and recording of the outdoor soil temperature at depth of 10 cm are required. The resulting significance is now 27.2 as compared to 30.6 when using the full set of available climate variables, and 28.19 when using the two apparatuses for P, H, and T measurements.

Comparison of multivariate methods for classification of radon concentration indoors

The difference between this case and the previous one with the full set of climate variables is that input events are now split at the value of radon concentration of 100 Bq/m^3, which is the recommended limiting value between the acceptable and increased radon concentration by the World Health Organization (WHO). Previous method had a cut on the value of 40 Bq/m^3, which was found to insure maximum employment of multivariate classifications. This particular value reflects the fact that the statistics on higher radon concentrations are getting progressively lower. In tab. 2, we present the significance and the signal and background efficiency for several best multivariate classifier methods. Again, the BDT (and BDT decorrelated) multivariate method shows the best performance in classifying the events into the categories of increased and acceptable concentrations.

Figure 4 shows the distribution of BDT classification method outputs for input signal and background events. These figures again demonstrate that classification methods work well *i. e*., that the separation of signal and background works very good. Also, the significance value for BDT is higher for higher cut values for splitting of input events. Interestingly, it appears that other multivariate methods also give better results under these new conditions.

Regression methods

Regression is the approximation of the underlying functional behavior defining the target value. We tried to find the best regression method that will give

Table 2. Significance, signal, and background efficiency for several best multivariate classifier methods in the case of imposed limiting value of 100 Bq/m^3

Classifier	S/sqrt(S + B)	EffSig	EffBkg
BDT	31.1	0.97	0.01
BDTD	30.9	0.98	0.03
MLPBNN	30.6	0.95	0.02
MLP	30.0	0.93	0.04
SVM	29.6	0.93	0.05

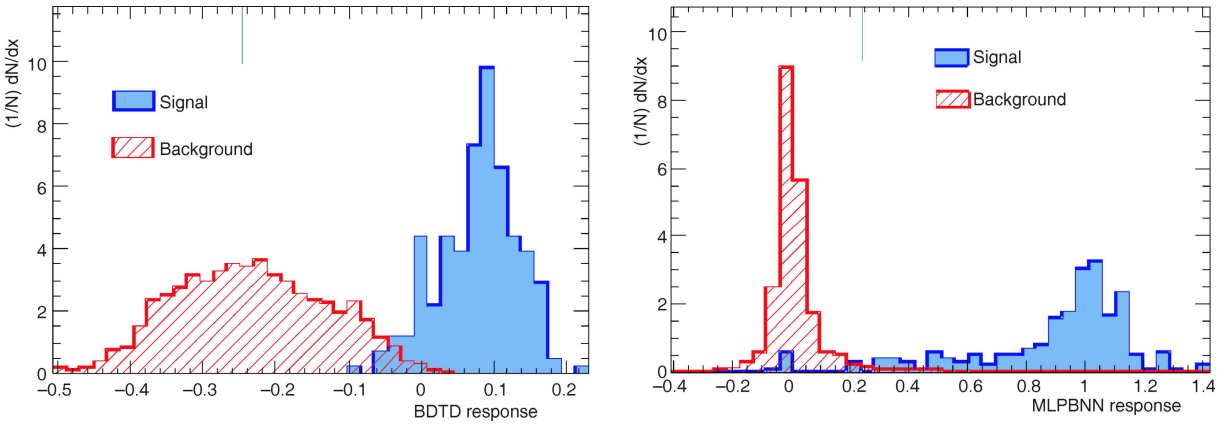

Figure 4. Distribution of BDT and ANN MLP classification method outputs for input signal and background events

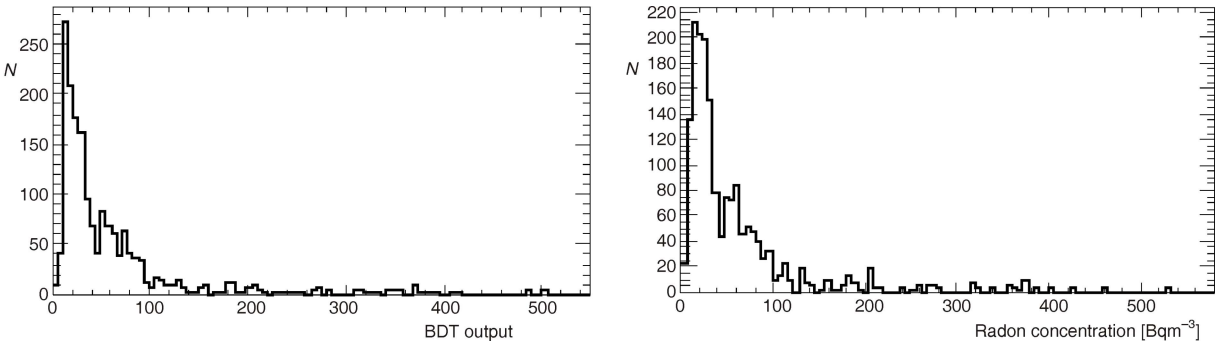

Figure 5. Distribution of radon concentrations and outputs from BDT multivariate method for regression of radon concentration using all climate variables

output values (predicted radon concentration) closest to the actual radon concentration that corresponds to specific input climate variables. The best multivariate regression method is found to be BDT, and the second one is MLP, same as in case of multivariate classifiers. Figure 5 presents the distribution of radon concentrations and outputs from the BDT multivariate method from regression of radon concentration using all climate variables.

To best way to estimate the quality of the method is to look at the differences between the output values from BDT multivariate regression method and the values of measured radon concentrations (fig. 6). The figure indicates the satisfactory predictive power of multivariate regression methods as applied for prediction of variations of indoor radon concentrations based on the full set.

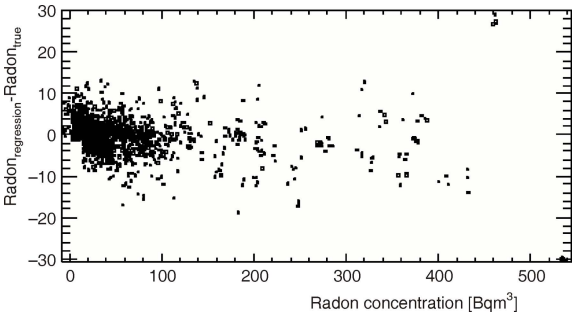

Figure 6. Difference of outputs from BDT multivariate regression method and radon concentrations, *vs.* radon concentration

CONCLUSIONS

The first test of multivariate methods developed for data analysis in high-energy physics and implemented in the TMVA software package applied to the analysis of the dependence of indoor radon concentration variations on climate variables demonstrated the potential usefulness of these methods. It appears that the method can be used with sufficient reliability for prediction of the increase of indoor radon concentrations above some prescribed value on the basis of monitored set of climate variables only. Surprisingly, this set of climate variables does not have to include too many of those which are nowadays widely available. To confirm these promising preliminary findings more case studies of similar character are required.

ACKNOWLEDGEMENT

The authors wish to thank Prof. Ivan Aničin for constant interest and support. This work is supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under project numbers III 43002 and OI171002.

AUTHOR CONTRIBUTIONS

The idea for this paper came as a result of discussions of V. I. Udovičić, R. M. Banjanac, D. R. Joković, and D. M. Maletić. Gathering climate data and MVA analysis was done by D. M. Maletić. V. I. Udovičić performed indoor radon measurements. Writing of the paper was done by D. M. Maletić and V. I. Udovičić. A. L. Dragić gave idea about using MVA methods in cosmic and radon measurements. N. B. Veselinović analyzed and validated climate data. J. Z. Filipović helped with MVA analysis. D. R. Joković helped with data analysis and paper technical preparation.

REFERENCES

[1] Baciu, A. C., Radon and Thoron Progeny Concentration Variability in Relation to Meteorological Conditions at Bucharest (in Romania), *Journal of Environmental Radioactivity, 83* (2005), 2, pp. 171-189
[2] Simon, E., *et al*., Estimation and Prediction of the Outdoor ^{222}Rn and ^{220}Rn Progeny Concentrations Usin Meteorological Variables, *Rom. Journ. Phys., 58* (2013), Suppl., pp. S262-S272
[3] Cuculeanu, V., *et al*., Dynamics, Deterministic Nature and Correlations of Outdoor ^{222}Rn and ^{220}Rn Progeny Concentrations Measured at Baciu, *Journal of Environmental Radioactivity, 102* (2011), 7, pp. 703-712
[4] Dragić, A., *et al*., The New Setup in the Belgrade Low-Level and Cosmic-Ray Laboratory, *Nucl Technol Radiat, 26* (2011), pp. 181-192
[5] Antanasijević, R., *et al*., Radon Measurements During the Building of a Low-Level Laboratory, *Radiat. Meas., 31* (1999), 1-6, pp. 371-374
[6] Banjanac, R., Indoor Radon Measurements by Nuclear Track Detectors: Applications in Secondary Schools, *Facta Universitas, Series: Physics*, *Chemistry and Technology, 4* (2006), 1, pp. 93-100
[7] Udovičić, V., *et al*., Radon Problem in an Underground Low-Level Laboratory, *Radiat. Meas., 44* (2009), 9-10, pp. 1009-1012
[8] Udovičić, V., Radon Time-Series Analysis in the Underground Low-Level Laboratory in Belgrade (in Serbian), *Radiation Protection Dosimetry, 145* (2-3), (2011), pp. 155-158
[9] Mihailović, D. T., *et al*., A Complexity Measure Based Method for Studying the Dependance of ^{222}Rn Concentration Time Series on Indoor Air Temperature and Humidity, *Applied Radiation and Isotopes, 84* (2014), pp. 27-32
[10] Brun, R., Rademakers, F., ROOT – An Object Oriented Data Analysis Framework, *Nucl. Inst. Meth. in Phys. Res., A 389* (1997), 1-2, pp. 81-86
[11] Hoecker, A., *et al*., TMVA Users Guide – Toolkit for Multivariate Data Analysis, PoS ACAT 040 (2007), http://arxiv.org/abs/physics/070303
[12] Yang, H.-J., Roe, B. P., Zhu, J., Studies of Boosted Decision Trees for MiniBooNE Particle Identification, *Nucl. Instrum. Meth., A555* (2005), 1-2, pp. 370-385
[13] Rojas, R., Neural Networks, Springer-Verlag, Berlin, 1996

Received on September 23, 2013
Accepted on March 10, 2014

Димитрије М. МАЛЕТИЋ, Владимир И. УДОВИЧИЋ, Радомир М. БАЊАНАЦ, Дејан Р. ЈОКОВИЋ, Александар Л. ДРАГИЋ, Никола Б. ВЕСЕЛИНОВИЋ, Јелена З. ФИЛИПОВИЋ

ПОРЕЂЕЊЕ МУЛТИВАРИЈАНТНИХ МЕТОДА ПРИ КЛАСИФИКАЦИЈИ И РЕГРЕСИЈИ РЕЗУЛТАТА МЕРЕЊА РАДОНА У ЗАТВОРЕНИМ ПРОСТОРИЈАМА

Представљамо резултате тестирања коришћења мултиваријантних метода, развијених за анализу података у физици високих енергија и имплементираних у програмском пакету за мултиваријантну анализу – у нашем проучавању зависности варијација концентрације радона у затвореним просторијама и климатских варијабли. Мултиваријантни методи омогућавају испитивање повезаности широког спектра климатских варијабли и концентрације радона, и онда када међу њима нема значајних корелација. Показали смо да мултиваријантни методи за класификацију и регресију раде добро, дајући као резултат нове информације и индикације које би могле бити корисне у даљем изучавању варијација концентрације радона у затвореним просторијама. Коришћењем ових метода, моћи ће да се дође до релативно добре моћи предвиђања концентрација радона, користећи само податке климатских варијабли.

Кључне речи: радон, мултиваријантна анализа, климатски параметар