Romanian Reports in Physics, Vol. 65, No. 2, P. 524–544, 2013

MODELING THE ^{222}Rn AND ^{220}Rn PROGENY CONCENTRATIONS IN ATMOSPHERE USING MULTIPLE LINEAR REGRESSION WITH METEOROLOGICAL VARIABLES AS PREDICTORS

FLORIN SIMION[1, 2], VASILE CUCULEANU[2, 3], ELENA SIMION[1, 4], ANTON GEICU[5]

[1] National Environmental Protection Agency, Radioactivity Laboratory, 294 Splaiul Independentei, RO-060031, Bucharest, Romania
[2] Faculty of Physics, University of Bucharest, Magurele, RO-077125 Bucharest, Romania
[3] Academy of Romanian Scientists, 54 Splaiul Independentei, RO-050094, Bucharest, Romania
[4] Faculty of Chemistry, University of Bucharest, Bucharest, Romania
[5] National Administration of Meteorology, Bucharest, Romania

Received July 24, 2012

Abstract. The present paper deals with the study of the multiple linear regression model for the estimation and prediction of the time series of radon and thoron progeny concentrations in atmosphere. The general purpose of multiple linear regression model is to find the linear relationship between a dependent (or explained) variable and several independent (or predictor) variables. Radon and thoron progeny data measured in Bacău station of the National Environmental Radioactivity Survey Network, coordinated by National Environmental Protection Agency, are modeled, at different time scales, by making use of the multiple linear regression with meteorological parameters as current and lagged independent variables. The daily average values of radon and thoron progeny concentrations for different months, as dependent variables, have been modeled as function of the following meteorological variables: air temperature, soil temperature, atmospheric pressure, relative humidity, precipitation and wind speed, as predictor variables. The collinearity and multicollinearity of independent variables have been analyzed. The estimation performances of the model have been checked by using the regression statistics: multiple correlation coefficient, coefficient of determination, *F*-test values, level of significance (*p*-value) and sum squared residuals. The prediction performances have been quantified by means of the sum squared residuals, correlation coefficient and the relative error specific to each time interval from prediction period. Also, there has been studied the prediction performances, considering different time periods on which the estimation regression has been built.

Key words: radon, thoron, progeny concentration, multiple linear regression, statistics.

1. INTRODUCTION

Radon (^{222}Rn) and thoron (^{222}Rn) are inert radioactive gases produced as decay products of ^{226}Ra and ^{224}Ra from the decay series of ^{238}U and ^{234}Th that are present in all terrestrial materials. Due to exhalation process these gases enter the atmosphere where they are subject to all atmospheric physical processes. The atmospheric concentrations of ^{222}Rn, ^{220}Rn and their daughters are functions of exhalation rate from the soil, intensity of vertical mixing and of horizontal transport and radioactive decay. Moreover, progeny concentrations are affected by the depletion due to fallout, washout, rainout, and other scavenging effects. It is known that the physical processes mentioned above (excepting the radioactive decay) are dependent on certain meteorological parameters specific to the planetary boundary layer: air temperature, soil temperature, atmospheric pressure, wind speed, precipitations and relative humidity [1, 2, 3]. That is why these meteorological parameters have been used as independent variables in the multiple linear regression model.

There is a constant interest in the study of radon that can be explained, on the one hand, by its radiological impact on human health and on the other hand, by its widespread use as an environmental tracer [4–12]. The ^{222}Rn having a lifetime comparable with the ventilation time of the planetary boundary layer, may be considered as a sensitive tracer of the vertical transport within the entire boundary layer, while ^{220}Rn , due to its very short lifetime, may be regarded as tracer only for the lower part of the surface layer, namely, the transition layer (few meters) between the laminar roughness layer and the turbulent surface layer [13]. Thus, the dynamics of the time series of ^{222}Rn and ^{220}Rn progeny concentrations will trace the dynamics of the meteorological variables with their physical interactions taking place during the measurement period of the respective time series.

In this paper were investigated the performances of the multiple linear regression model for estimation and prediction of the time series of radon and thoron progeny concentrations making use of the meteorological parameters as current and lagged independent variables [14]. The following meteorological variables were used as current and lagged independent variables of the regression model: air temperature, soil temperature, temperature gradient between air and soil, relative humidity, atmospheric pressure, precipitation, wind speed. Regression model has been used to study the pollutant dynamics in the planetary boundary layer [15, 16, 17].

In order to obtain reliable results, the collinearity and multicollinearity of predictors have to be analyzed. Study of the collinearity is based on the analysis of correlation matrix of the predictor variables. In case of multicollinearity, two or more predictor variables are highly linearly correlated.

The regression model performances for the estimation of the time series values are quantified by the following regression statistics: multiple correlation coefficient (R), coefficient of determination (R^2), p-level (significance level), F (test), sum squared residuals (SSR) [18]. In case of prediction, the model performance was quantified by means of correlation coefficient R, p-value, SSR and the relative error specific to each time interval from prediction period in order to have a measure of error dynamics in the prediction process.

With regard to the applicability, the developed model can be used for the following purposes: (i) to test and quantify the physical relationship intensity between radon/thoron/progeny concentrations (or other elements) and meteorological variables for any site of interest; (ii) to fill the missing data periods in the time series of radon/thoron/progeny concentrations by making use of meteorological variables from the respective period [19]; (iii) to predict the concentration values in a site where meteorological data exist but the measurements of the radon/thoron /progeny concentrations have been interrupted because of diverse reasons (e.g, economic). Because we had at our disposal only daily values of meteorological data, the modeling approach uses only this type of data. In order to model the influence of the diurnal variation of predictors on progeny concentrations it is necessary to use daytime and nighttime average predictors. As far as the maritime influence on progeny concentration is concerned this can be physically simulated by including in the variable sets the wind direction. The influence of the Black See on Bacau site is negligible. The processed data regarding the radon and thoron progeny concentrations were measured in the year 1996 by the Environmental Radioactivity Monitoring Station Bacău, located near Bacău city [12]. This city is situated at low altitude (about 200 m) in the eastern part of Romania. The meteorological were measured in the same area. The choice of this year is justified by the fact that does not exist missing data for both radon and thoron progeny concentrations and corresponding meteorological data.

2. DATA AND METHODS

2.1. MEASURING METHOD OF RADON AND THORON PROGENY

It is known that atmospheric turbulent diffusion is determining close activity concentration distributions for radon and progeny within surface layer, for different turbulent mixing conditions [20]. Thus one may consider that the *short*-lived progeny of ^{222}Rn are in secular equilibrium with outdoor ^{222}Rn in most atmospheric conditions [21]. For ^{220}Rn progeny, irrespective of the turbulent diffusion intensity, there is no equilibrium with ^{220}Rn, except for the case when a strong inversion within the atmospheric layer of about 1 m above soil occurs or for the case of an intense mixing in a layer of about 100 m against the ground [22].

The measuring method of radon and thoron progeny activity concentrations is based on the aerosol aspiration on glass fiber filters with high filtration efficiency (96–99%), using high volume aerosol samplers with aspiration head located at 2 meters above the ground. The filter activity has been measured using a low background total beta counter and a ^{90}(Sr/Y) reference standard for determination of equipment detection efficiency. The counter background was between 2.5 and 6 counts per minute. The sampling time interval was 5 hours. The filters have been measured immediately after sampling (3 min. after), for 1000 seconds, after 20 hours, for 3000 seconds and after 5 days, for 3 000 seconds. The first two measurements provide filter activities necessary for the determination of the radon and thoron progeny concentrations and the last measurement indicates the presence of artificial radioactivity in the atmosphere. In 1996, Environmental Radioactivity Monitoring Station (ERMS) Bacău had a 4 aspiration program: 02:00–07:00 (aspiration A1), 08:00-13:00 (aspiration A2), 14:00-19:00 (aspiration A3) and 20:00-01 (aspiration A4). The measured data were monthly reported to the National Reference Radioactivity Laboratory from National Environmental Protection Agency-Bucharest.

2.2. RADON AND THORON TIME SERIES

The data used for this study are averaged daily values obtained from the 4 aspirations, for both ^{222}Rn and ^{220}Rn progeny. The time series of daily data were used to develop the multiple linear regression model for each month for both estimation and prediction of the progeny concentrations.

In Fig. 1, the daily average values of ^{222}Rn and ^{220}Rn progeny concentrations and meteorological variables for the months of January and July are presented.

From Fig. 1 it can be clearly seen that the dynamics of ^{222}Rn and ^{220}Rn progeny concentrations is determined by the time variations of the meteorological (predictor) variables. Thus, in the figure for the month of January the significant decrease of progeny concentrations during the first five days of January may be explained by the presence of precipitation diminishing the exhalation rates during this period, corroborated with the increasing wind speeds that is intensifying the atmospheric dilution of air born substances. The maxima of progeny concentrations from January 18 are mainly determined by the atmospheric calm (wind speed equals to zero) that is theoretically diminishing the atmospheric eddy diffusion until the level of molecular clusters.

The minimum of the progeny concentrations from July 10 is obviously caused by three simultaneous physical processes: the intense precipitation, the high wind speed and very low soil temperature that are dominant processes as compared with the high value of relative humidity that usually determines an increase of progeny concentrations.

On the other hand, the minimum from July 17 (Fig. 1), may be due to the high wind speed corroborated with the very low relative humidity.

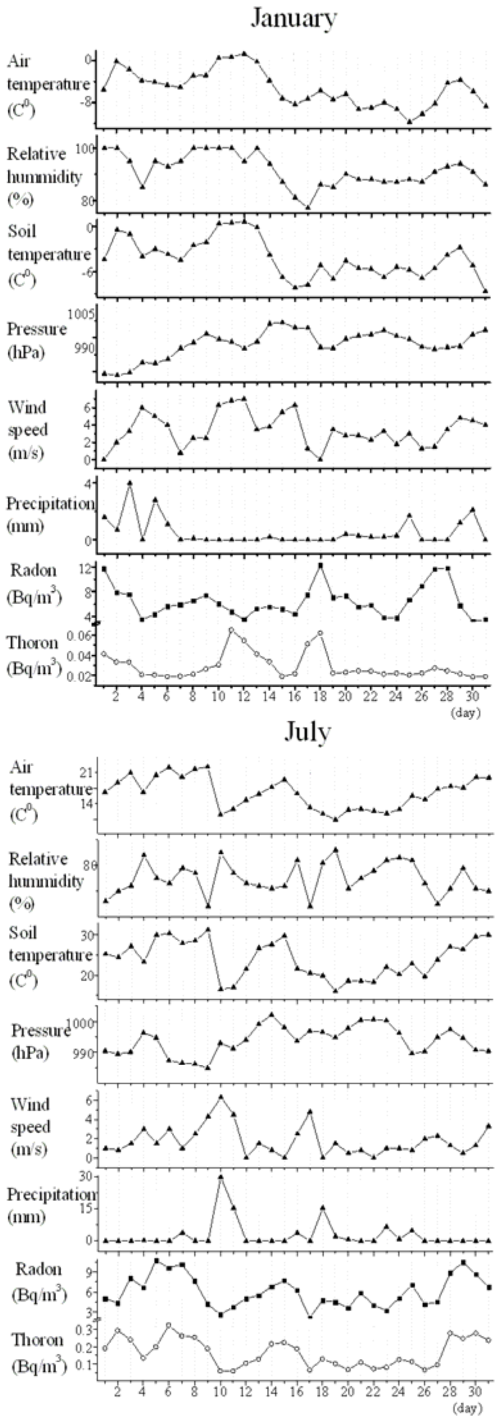

Fig. 1 – ^{222}Rn and ^{220}Rn progeny concentrations and meteorological variables in January and July.

2.3. METEOROLOGICAL DATA

The meteorological data were determined with hourly frequency by Bacău Weather Station, coordinated by National Administration of Meteorology. The meteorological parameters were recorded according with the climatologically monitoring program of the National Administration of Meteorology. Air temperature was determined using a thermograph, placed at 2 m above ground in the station yard, atmospheric pressure was determined using a barograph, placed the office room (precision of ± 0.1 hPa), wind speed was recorded at 10 m above ground, the volume of precipitation was determined with rain gauge and air relative humidity was recorded in the same conditions as the temperature using a hygrograph. The hourly values read on the thermogrammes, barogrammes and hygrogrammes, were previously adjusted taking into account the corresponding values obtained at the four terms of observation (01:00, 07:00, 13:00, 19:00), using: thermometer (Hg) precision of ± 0.1°C, barometer (Hg) precision of ± 0.1 hPa, respectively hygrometer (precision of ± 1%).

3. MULTIPLE LINEAR REGRESSION

Linear regression was the first type of regression analysis that was rigorously studied and extensively used in practical applications, because it is easier to fit the linear models and their statistical characteristics are easier to be determined. Linear regression is widely used in science to describe possible relationships between variables. It ranks as one of the most important tools used in different disciplines.

3.1. ESTIMATION METHOD OF THE TIME SERIES DATA USING CURRENT AND LAGGED PREDICTOR VARIABLES

The estimation of progeny concentrations is made by using the analytical expression of the multiple linear regression with the corresponding meteorological variables from the time period on which the regression is defined. Multiple linear regression model assumes a linear relationship between the progeny concentration as dependent variable and the meteorological data as predictor variables.

Therefore, the dependent variable $y(t)$ is associated with the current predictor variables $\{x_i(t)\}_{i=1,2,..n}$ by the following relationship [18]:

$$y(t)= \beta_0 + \beta_1 x_1(t) + \beta_2 x_2(t)......+ \beta_n x_n(t), \quad (1)$$

where β_0 is the regression constant, $\beta_1\ldots.. \beta_n$ are the regression coefficients when only current variable are used. If current and lagged independent variables are used, the regression relationship becomes:

$$y(t) = \gamma_0 + \gamma_1 x_1(t) + \gamma_2 x_2(t) .... + \gamma_n x_n(t) + \sum_{k=1}^{m}\sum_{i=1}^{n} \delta_{ik} x_{ik} , \qquad (2)$$

where the variables $x_{ik} = \{x_i (t-k)\}_{i=1...n}$ are the lagged variables of order k and m is the maximum order of these variables. A lagged variable of order k means a variable delayed with k days if the values of variables are defined as daily average data. The quantity γ_0 is the regression constant, γ_j are the regression coefficients of current variables, δ_{ik} are the regression coefficients of lagged predictor variables.

The regression coefficients β_0 ,$\{ \beta_i \}_{i=1...n}$, γ_0 , $\{ \gamma_i \}_{i=1...n}$, $\{\delta_{ik} \}_{i=1...n,\, k=1....m}$ are determined making use of the least-square method. The analytical expressions (1) and (2) constitute the mathematical basis of the multiple linear regression for the estimation and prediction of the radon/thoron progeny concentrations in the atmosphere.

The performances of this regression model are quantified by the following statistics:

- R-Multiple correlation coefficient. R is measuring the degree of correlation between ^{222}Rn or ^{222}Rn progeny concentrations and the ensemble of meteorological data
- R^2 -Squared multiple correlation coefficient. R^2 is quantifying the proportion of the variation of dependent variable that is explained by the independent variables
- F-Test – This is a global test of significance for the ensemble of coefficients. This statistic has F-distribution and is used to test the null hypothesis stating that all regression coefficients β_j, γ_j or δ_{ik} are equal to zero and the alternative hypothesis stating that at least one of these coefficients is different of zero. Large values of the F-test provide evidence against the null hypothesis
- p-value – The p-value characterizes the significance level so that a null hypothesis to be rejected or accepted against an alternative hypothesis. The p-value is computed making use of the t-Student distribution for the regression coefficients and F distribution for the F-test for overall model. Generally, a p-value < 0.05 indicates that the null hypothesis may be rejected and the alternative one has to be accepted
- SSR-Sum Squared Residual Residuals quantifies the deviation between the progeny concentrations estimated or predicted by the multiple regression and the measured values of the respective concentrations. The residual value indicates how close the concentrations simulated by the regression model are as against the measured concentrations.

3.2. PREDICTION METHOD OF THE PROGENY CONCENTRATIONS

The prediction is achieved by using the analytical expression of regression obtained for a previous time period, with predictor variables from the time interval on which the prediction is done. This means that the regression coefficients are kept constant and the meteorological variables from the prediction interval are used. Supposing that a number of N data coresponding to N time intervals from the observed data series were used for achieving the model equations (1) and (2), one may use these expressions to predict the progeny concentrations on the time intervals N+1, N+2, …, by making use of the meteorological variables from the period intended for prediction. For example, in order to predict the progeny concentrations on the time interval N+1, $[\tilde{y}]_{N+1}$, making use of current variables, the folowing equation is utilized:

$$[\tilde{y}]_{N+1} = \beta_0 + \beta_1[x_1]_{N+1} + \beta_2 [x_2]_{N+1} \cdots + \beta_n [x_n]_{N+1}, \qquad (3)$$

where $[x_1]_{N+1}$, $[x_2]_{N+1}$ …. $[x_n]_{N+1}$, are the values of current variables in the time interval N+1. In case of current and lagged variables, the predicted value of dependent variable $[\tilde{y}]_{N+1}$, has the expression:

$$[\tilde{y}]_{N+1} = \tilde{y}_0 + \tilde{y}_1[x_1]_{N+1} + \tilde{y}_2 [x_2]_{N+1} .... + \tilde{y}_n[x_n]_{N+1} + \sum_{k=1}^{m}\sum_{j=1}^{n} \tilde{\delta}_{ik} \left[x_{ik} \right]_{N+1}, \qquad (4)$$

where $[x_{ik}]_{N+1,}$ are the values of the independent lagged variables for the time intervals N+1.

As for criteria for prediction performances, we have used the SSR and correlation coefficient (R) between predicted and observed progeny concentrations, the corresponding level of significance (p-value) and the relative error specific to each predicted concentration value.

3.3. COLLINEARITY AND MULTICOLLINEARITY

Collinearity is a linear relationship between two predictor variables. Two collinear variables explaining the same part of variation for the dependent variable determines less significance in the case of individual terms of partial regression coefficients. In this case the regression coefficient values become more dependent on the distribution and values of errors. The analysis of collinearity of predictors reveals the independent variables having correlation coefficients close to 1 (0.99 and larger). Multicollinearity refers to a situation in which two or more predictor variables in a multiple regression model are highly linearly related. The existence of multicollinearity inflates the variances of the parameter estimates. That may result, in lack of statistical significance of individual independent variables while the overall model may be strongly significant. Multicollinearity may also result in wrong signs and magnitudes of regression coefficient estimates, and consequently in incorrect conclusions about relationships between independent and dependent variables.

The main statistic for multicollinearity diagnostic is the Variance Inflation Factor (VIF) defined as [23]: VIF=$1/(1-R_j^2)$, where R^2_j is the coefficient of determination when variable x_j is regressed on the $j-1$ remaining independent variables. A variable is considered to be problematic if its VIF is larger than 10.0 [23].

When multicollinearity is present, theory and physical considerations should be used to choose the best variables to be kept in the model.

3.4. SELECTION OF THE INDEPENDENT VARIABLES

In order to build the regression model the following current meteorological predictors have been used: air temperature (T_a), soil temperature (T_s), relative humidity (H), atmospheric pressure (Pa), wind speed (W), precipitation (Pr), temperature gradient between soil and atmosphere ($Gr = T_a - T_s$), and corresponding lagged variables of order 1: $T_{a1} = Ta(t-1)$, T_{s1}, H_1, P_{a1}, W_1, Pr_1, Gr_1. To select the sets of predictor variables the following considerations have been taken into account: (i) the correlation coefficient between predicted and measured data to be as great as possible; (ii) the SSR between measured and predicted data to be as little as possible; (iii) the VIF not to be greater than 10; (iv) no negative values for predicted data. According to the above criteria, the following sets of independent variables have been utilized for estimation and prediction of the progeny concentrations: $\mathbf{S_1}$: $\{T_a, H, P_a, W, Pr\}$; $\mathbf{S_2}$:$\{T_s, H, P_a, W, Pr\}$; $\mathbf{S_3}$: $\{Ta, T_s, H, P_a, W, Pr\}$; $\mathbf{S_4}$: $\{Ta, H, Pa, W, P, Ta_1, H_1, Pa_1, W_1, Pr_1\}$; $\mathbf{S_5}$:$\{Gr, H, Pa, W, Pr\}$; $\mathbf{S_6}$:$\{Gr, H, P_a, W, Pr, Gr_1, H_1, Pa_1, W_1, Pr_1\}$; $\mathbf{S_7}$: $\{T_s, H, P_a, W, Pr, T_{s1}, H_1, P_{a1}, W_1, Pr_1\}$.

For individual month differing sets ($\mathbf{S_1} - \mathbf{S_6}$) was used (see Tables 3 and 4). Dependence on gradient proves that during some periods of the year the air concentrations of both gases and their progeny are mainly determined by the atmospheric convection and thermal inversion processes in the planetary boundary layer, in case of ^{222}Rn, respectively, in the lower part of the surface layer, in case of ^{220}Rn. Another reason for selecting the gradient was the fact that the use of both T_a and Ts, due to their high collinearity, has led to negative values of predicted concentrations in certain months where a pronounced variation of these variables has occurred in comparison with the previous month where the regression expressions have been defined. By examining the sets of variables for ^{222}Rn and ^{220}Rn progeny, it may note that T_a is present as independent variable only for ^{222}Rn, while for ^{220}Rn only T_s. Air temperature being highly correlated with turbulent diffusion intensity in boundary layer, the presence of the variable T_a proves that atmospheric eddy diffusion is one of the major physical process determining the ^{222}Rn progeny concentrations in certain months. In case of ^{220}Rn progeny, the presence of T_s in all months (directly or by Gr) indicate that the exhalation rate is the main process contributing to the ^{220}Rn progeny concentrations, the atmospheric eddy diffusion not being fully developed in the air layer near the ground.

4. RESULTS AND DISCUSSIONS

4.1. COLLINEARITY ANALYSIS

In order to detect the collinearity, the correlation matrices of the independent variables have to be examined. That is why the correlation matrices containing the correlation coefficients (R) and corresponding level of significance (*p* value) have been computed for the following predictor variables: T_a, T_s, H, P_a, W, Pr, Gr, T_{a1}, T_{s1}, H_1, P_{a1}, W_1, Pr_1, Gr_1, by making use of a specialized software [24, 25] and Excel program.

The values of correlation characteristics related to the current variables for all months, reveal a very strong correlation between air temperature and soil temperature with a minimum of 0.716 in May and a maximum of 0.991 in December with very high level of significance ($4.5\cdot10^{-27} \div 5.9\cdot10^{-6}$).

Table 1

Correlation matrices for December

Variables	T_a	T_s	H	P_a	W	Pr	Gr
T_a	1						
T_s	0.991	1					
	$p \le 2.1e^{-28}$						
H	0.197	0.239	1				
	$p \le 0.287$	$p \le 0.194$					
P_a	-0.319	-0.349	-0.367	1			
	$p \le 0.079$	$p \le 0.054$	$p \le 0.042$				
W	-0.007	-0.012	-0.178	-0.162	1		
	$p \le 0.972$	$p \le 0.947$	$p \le 0.338$	$p \le 0.383$			
Pr	0.067	0.112	0.309	-0.395	-0.011	1	
	$p \le 0.719$	$p \le 0.549$	$p \le 0.09$	$p \le 0.027$	$p \le 0.955$		
Gr	-	-	-0.218	0.089	0.037	-0.282	1
	-	-	$p \le 0.237$	$p \le 0.635$	$p \le 0.842$	$p \le 0.123$	
Variables	T_{a1}	T_{s1}	H_1	P_{a1}	W_1	Pr_1	Gr_1
T_a	0.935	-	0.431	-0.361	-0.168	0.081	-
	$p \le 4.5e^{-15}$	-	$p \le 0.015$	$p \le 0.046$	$p \le 0.365$	$p \le 0.666$	-
T_s	-	0.904	0.480	-0.38	-0.23	0.113	-
	-	$p \le 1.0e^{-12}$	$p \le 0.006$	$p \le 0.034$	$p \le 0.214$	$p \le 0.546$	-
H	0.079	0.079	0.531	0.036	-0.387	0.106	0.025
	$p \le 0.672$	$p \le 0.672$	$p \le 0.002$	$p \le 0.849$	$p \le 0.031$	$p \le 0.572$	$p \le 0.895$
P_a	-0.167	-0.168	-0.504	0.730	0.118	-0.265	-0.045
	$p \le 0.369$	$p \le 0.366$	$p \le 0.004$	$p \le 3.0e^{-06}$	$p \le 0.526$	$p \le 0.148$	$p \le 0.811$
W	0.089	0.11	0.139	-0.461	0.221	0.173	-0.14
	$p \le 0.634$	$p \le 0.554$	$p \le 0.455$	$p \le 0.009$	$p \le 0.232$	$p \le 0.351$	$p \le 0.451$
Pr	0.089	0.089	0.228	-0.293	-0.312	0.485	0.033
	$p \le 0.633$	$p \le 0.635$	$p \le 0.216$	$p \le 0.109$	$p \le 0.086$	$p \le 0.006$	$p \le 0.861$

Table 1 (continued)

Gr	-	-	-0.176	0.004	0.361	-0.191	0.179
	-	-	$p \le 0.342$	$p \le 0.984$	$p \le 0.046$	$p \le 0.302$	$p \le 0.336$
Variables	$\boldsymbol{T_{a1}}$	$\boldsymbol{T_{s1}}$	$\boldsymbol{H_1}$	$\boldsymbol{P_{a1}}$	$\boldsymbol{W_1}$	$\boldsymbol{Pr_1}$	$\boldsymbol{Gr_1}$
$\boldsymbol{T_{a1}}$	1						
$\boldsymbol{T_{s1}}$	-	1					
$\boldsymbol{H_1}$	0.269	0.302	1				
	$p \le 0.143$	$p \le 0.099$					
$\boldsymbol{P_{a1}}$	-0.351	-0.377	-0.398	1			
	$p \le 0.052$	$p \le 0.036$	$p \le 0.026$				
$\boldsymbol{W_1}$	-0.022	-0.026	-0.178	-0.153	1		
	$p \le 0.906$	$p \le 0.891$	$p \le 0.337$	$p \le 0.412$			
$\boldsymbol{Pr_1}$	0.080	0.130	0.355	-0.417	-0.021	1	
	$p \le 0.668$	$p \le 0.486$	$p \le 0.049$	$p \le 0.019$	$p \le 0.911$		
$\boldsymbol{Gr_1}$	-	-	-0.175	0.095	0.021	-0.369	1
	-	-	$p \le 0.346$	$p \le 0.61$	$p \le 0.909$	$p \le 0.04$	

The explanation consists in the fact that low atmosphere and Earth surface being in direct contact there is a permanent exchange of thermal energy, mainly from soil to atmosphere that determines close values and similar time dependencies of temperatures for both subsystems. These independent variables could generate (multi)collinearity difficulties for the regression model.

With regard to the current and lagged variables that have been used for building up the regression model, the correlation coefficients do not exceed those provided only by the current variables. As an example, from the Table 1 containing the correlation matrices of current and lagged variables for the month of December having the highest correlation coefficients, one may notice that the maximum values of the correlation coefficients between current and lagged variables are those of T_a and T_{a1} (0.935) and of T_s and T_{s1} (0.904). The maximum value of R for the used lagged variables is between Pr_1 and P_{a1}, namely, –0.417. Because the only variables which could generate collinearity difficulties are T_a and T_s, the gradient has been introduced as predictor variable in certain months and the sets of variables containing T_a and T_s that have generated VIF values greater than 10 have been disregarded.

4.2. MULTICOLLINEARITY ANALYSIS

As it was previously mentioned the main statistic for multicollinearity diagnostic is the variance inflation factor. This statistic has been computed for the sets of current and lagged variables selected for each month according to the criteria discussed in section 3.4.

Table 2

Variance inflation factors for estimation and prediction variables

^{222}Rn progeny multicollinearity data						**^{220}Rn progeny multicollinearity data**					
January				**July**		**January**				**July**	
S_4	**VIF**	**S_3**	**VIF**	**S_1**	**VIF**	**S_7**	**VIF**	**S_5**	**VIF**	**S_7**	**VIF**
T_a	5.26	T_a	7.19	T_a	1.11	T_s	3.57	Gr	1.35	T_s	2.17
H	3.13	T_s	9.09	H	1.43	H	2.44	H	1.28	H	2.50
P_a	9.09	H	4.00	P_a	1.20	P_a	7.69	Pa	1.79	P_a	3.85
W	2.56	Pr	1.56	W	1.15	W	2.56	W	1.30	W	1.47
Pr	1.72	P_a	2.00	Pr	1.52	Pr	1.72	Pr	1.54	Pr	1.85
T_{a1}	5.00	W	1.79			T_{s1}	3.45			T_{s1}	2.17
H_1	3.33					H_1	2.38			H_1	3.13
P_{a1}	10.00					P_{a1}	8.33			P_{a1}	4.35
W_1	2.70					W_1	2.70			W_1	1.54
Pr_1	1.85					Pr_1	1.82			Pr_1	2.22

By analyzing the obtained results the following conclusions can be drawn: (i) there are no VIF values exceeding the upper limit indicating the presence of multicollinearity, namely, 10; (ii) the only VIF attaining the upper value of 10 is that for January in case of ^{222}Rn, for the variable P_{a1} having a correlation coefficient of 0.852 ($p \leq 7.0e^{-10}$) with P_a in this month; (iii) in February, in case of ^{222}Rn, the presence of T_a and T_s ($R = 0.955$, $p \leq 2.5 \cdot 10^{-11}$) among predictor variables has induce a value of 9.09 for the VIF corresponding to T_s, under the limit defining the presence of multicollinearity; (iv) in case of ^{220}Rn the only greater VIF values are those of January for the variable P_a, 7.69 and P_{a1}, 8.33.

Examples of VIF values for the sets of independent variables, for the months of January and July, are presented in the Table 2. Generally, one may say that the selection criteria we have used provided reliable predictor variables for building up the regression model.

The study performed for whole year, showed that multicollinearity is not present for none of the considered predictors and that all the predictors can be used to estimate and predict adequately the concentrations of ^{222}Rn and ^{220}Rn progeny.

4.3. ESTIMATION OF ^{222}Rn and ^{220}Rn PROGENY CONCENTRATION

Making use of the dedicated software [24] and Excel, the regression coefficients are determined by the least square method.

The specific statistic parameters used to quantify the regression model performances are those presented in Section 3.1: multiple R, R^2, p-value, F-test and SSR.

In the Table 3 the regression statistics for the estimation of the ^{222}Rn and ^{220}Rn progeny concentrations for each month, are presented.

Table 3

Regression statistics for ^{222}Rn and ^{220}Rn progeny estimation

	Regression statistics for ^{222}Rn progeny						**Regression statistics for ^{220}Rn progeny**					
month	S_i	R	R^2	p-value	F-test	SSR	S_i	R	R^2	p-value	F-test	SSR
Dec '95	S_4	0.808	0.652	$5.7e^{-3}$	3.754	114.8	S_7	0.726	0.527	$6.1e^{-2}$	2.228	0.006
Jan '96	S_3	0.711	0.505	$5.8e^{-3}$	4.083	100.1	S_5	0.595	0.354	$5.9e^{-2}$	2.518	0.007
Feb '96	S_2	0.770	0.593	$5.5e^{-4}$	6.706	90.9	S_6	0.627	0.393	$3.7e^{-1}$	1.168	0.006
Mar '96	S_6	0.755	0.571	$3.0e^{-2}$	2.658	72.5	S_7	0.703	0.495	$9.5e^{-2}$	1.962	0.003
Apr '96	S_5	0.712	0.507	$2.9e^{-3}$	4.945	27.7	S_6	0.861	0.742	$7.5e^{-4}$	5.471	0.090
May '96	S_3	0.818	0.668	$7.7e^{-5}$	8.065	15.2	S_2	0.619	0.383	$2.6e^{-2}$	3.106	0.175
Jun '96	S_1	0.894	0.800	$8.8e^{-5}$	7.593	19.2	S_7	0.807	0.651	$8.5e^{-3}$	3.541	0.106
Jul '96	S_1	0.897	0.805	$3.6e^{-5}$	8.264	32.7	S_7	0.934	0.873	$1.0e^{-6}$	13.75	0.026
Aug '96	S_6	0.749	0.561	$3.6e^{-2}$	2.556	52.8	S_2	0.734	0.538	$1.1e^{-3}$	5.827	0.086
Sep '96	S_5	0.624	0.390	$2.7e^{-2}$	3.069	86.5	S_6	0.739	0.546	$5.8e^{-2}$	2.285	0.077
Oct '96	S_5	0.614	0.377	$2.8e^{-2}$	3.029	158.8	S_2	0.702	0.493	$3.1e^{-3}$	4.860	0.146
Nov '96	S_6	0.703	0.495	$1.2e^{-1}$	1.859	127.1	S_5	0.618	0.382	$3.2e^{-2}$	2.970	0.374

One may notice that in case of ^{222}Rn progeny, the multiple correlation coefficients has the greatest values in December '95 and in the period May-July with R values equal to 0.808 and 0.818, 0.894, 0.897 with very high level of significance ($p < 0.01$). This proves that the corresponding predictor variables are describing quite well the variability of radon progeny in these months. It is interesting to mention that for the months having the greatest R values, the air temperature is used in the sets of independent variables. This fact indicates that ^{222}Rn being a tracer of entire planetary boundary layer is more sensitive to the air temperature than to the gradient. It seems that the elimination of the T_a from the sets of independent variables in certain months in order to avoid the multicollinearity consequences (*e.g.*, the negative values for predicted concentrations), has made the regression model less sensitive to the variability of ^{222}Rn progeny.

With regard to residual value indicating how close the concentrations simulated by the regression model are as against the measured concentrations, one may see that this statistic has the lowest value in May (15.200) even if the multiple correlation coefficient has no greatest value: R value is indicating the degree of similarity of time dependence of the measured and simulated data and SSR is indicating how close the two time distributions are. The greatest SSR value (158.839) is associated to the month of October which has the least R (0.614), too. Referring to the R^2 values, one may notice that in case of June and July around 80% from the variation of radon progeny concentration can be explained by the set of variables S_1, while in case of November only around 38% can be explained by the set of variables S_5.

As far as the ^{220}Rn progeny are concerned, the greatest R values are obtained when soil temperature is contained in the variable sets used for obtaining the analytical expression of regression, what is indicating that the exhalation rate is a dominant physical process determining the thoron progeny concentration level. For instance, by making use of the set S_7 of current and lagged variables including T_s, the R value for July is equals to 0.934, with a very high level of significance, $p \leq 1.0e^{-6}$ (almost 100% confidence) and a F-test value equals to 13.759. According to R^2 value more about 87 % from the variation of thoron in this month can be described by the independent variables included in the set S_7. The least value of R is for January, *i.e.*, 0.595, with a level of significance equals to 0.059 (around 94% confidence), using the set of variable S_5 that includes the gradient as current variable. The corresponding value of R^2 indicates that only around 35% of the variation of thoron in January can be explained by the predictor variables of the set S_5.

From Table 3, one may remark that the best results for estimation are obtained in summer months for both ^{222}Rn (R varies between 0.749 ÷ 0.897) and ^{220}Rn (R varies between 0.734 ÷ 0.934) progeny concentrations.

Below, there are given the multiple linear regression equations for the month of July for estimation of ^{222}Rn and ^{220}Rn progeny concentrations:

$$^{222}\text{Rn}: Rn(t) = -111.9164+0.6081\cdot T_a(t) + 0.1805\cdot H(t) + 0.0662\cdot T_s(t) + 0.0919\cdot P_a(t) - 0.3480\cdot W(t) - 0.00068\cdot Pr(t)$$

$$^{220}\text{Rn}: Tn(t) = -0.2570 + 0.0170\cdot T_a(t) + 0.00180\cdot H(t)+0.0043\cdot T_s(t)-0.00014\cdot P_a(t) - 0.0120\cdot W(t) + 0.00140\cdot Pr(t)$$

and in Fig. 2 the corresponding estimated time series as compared with measured values, are presented.

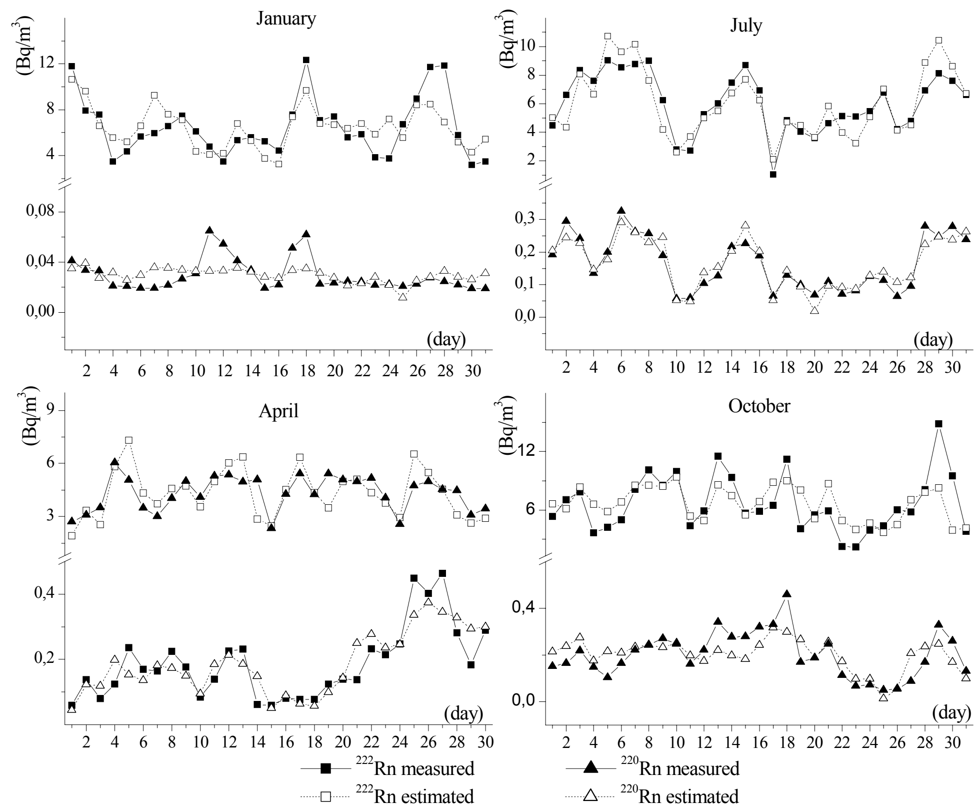

Fig. 2 – Measured and estimated progeny concentrations in January, April, July and October.

From Fig. 2 containing representative months for each season, it can be clearly seen, that the estimation values generated by the multiple regression model, are very close to the measured values. This means that the considered predictor variables may describe well the radon and thoron concentration dynamics in the respective months.

4.4. PREDICTION OF ^{222}Rn and ^{220}Rn Progeny Concentration

The prediction procedure is based on the regression analytical expression obtained with the data from the previous month or other period, using the independent variable data from the time interval on which the prediction is performed.

In Table 4, the statistics for the prediction of the ^{222}Rn and ^{220}Rn progeny concentrations using regression build up on one month interval, are presented.

Referring to the prediction characteristics in the case of ^{222}Rn progeny, one can observe that the greatest values of *R* are for the months January, June, July and August. The explanation consists in the fact that the corresponding multiple regressions build up in previous months (December 1995 for January, May for June and so on) have the greatest values of multiple *R* (Table 3) meaning that these regressions can describe well enough the time variation of ^{222}Rn progeny even in the next months with the condition that predictor variables not to differ significantly as against the previous months. As an example can be considered the case of statistics for the ^{222}Rn progeny prediction for November: the regression built up in October has the multiple *R* value equals to 0.614 (Table 3) and corresponding correlation coefficient for November is equals to 0.090 (Table 4).

Table 4

Regression statistics for ^{222}Rn and ^{220}Rn progeny concentration prediction

Month	**^{222}Rn progeny**				**^{220}Rn progeny**			
	S_i	**Prediction characteristics**			**S_i**	**Prediction characteristics**		
		R	***p*-value**	**SSR**		***R***	***p*-value**	**SSR**
January	S_4	0.715	$5.1e^{-6}$	105.245	S_7	0.126	0.498	0.020
February	S_3	0.590	$6.8e^{-4}$	173.782	S_5	0.363	0.052	0.009
March	S_2	0.585	$5.0e^{-4}$	172.810	S_6	0.185	0.319	0.009
April	S_6	0.195	0.300	286.988	S_7	0.604	$3.7e^{-4}$	0.675
May	S_5	0.557	0.001	52.245	S_6	0.340	0.060	0.628
June	S_3	0.644	$1.1e^{-4}$	62.822	S_2	0.426	0.018	0.302
July	S_1	0.752	$8.4e^{-7}$	83.769	S_7	0.711	$6.0e^{-6}$	0.137
August	S_1	0.771	$2.8e^{-7}$	108.808	S_7	0.507	0.003	0.697
September	S_6	0.280	0.133	229.960	S_2	0.231	0.218	0.216
October	S_5	0.290	0.113	242.036	S_6	0.352	0.051	0.342
November	S_5	0.090	0.634	363.382	S_2	0.370	0.044	0.916
December	S_6	0.260	0.158	320.700	S_5	0.209	0.259	1.576

This means that the multiple regression obtained making use of the independent variables from October is not generating reliable predictions for November because the independent variables in this month are significantly different of those from October (*e.g.*, in October the average air temperature is 9.64^0C, average soil temperature is 10.01^0C, total precipitation is 16.5 mm, while in November the air temperature is 6.97^0C, soil temperature is 6.71^0C and total precipitation is 41.2 mm). As it may notice from Tables 3 and 4, in case of ^{220}Rn progeny, this limitation of the prediction procedure is more pronounced. Thus, in December '95 the multiple R is 0.726 and the corresponding correlation coefficient between predicted and measured data in January is 0.126 with a low level of significance ($p \leq 0.498$). Similar situations are specific to March, May and the period September–December. The only months having reasonable correlation coefficients between predicted and measured data with high level of significance are April (0.604, $p \leq 3.7\cdot10^{-4}$), July (0.711, $p \leq 6.0\cdot10^{-6}$) and August (0.507, $p \leq 3.0\cdot10^{-3}$), the corresponding regression statistics being (0.703, $p \leq 9.5\cdot10^{-2}$) for March, (0.807, $p \leq 8.5\cdot10^{-3}$) for June and (0.934, $p \leq 1.0\cdot10^{-6}$) for July. The comparison between prediction characteristics of ^{222}Rn progeny and ^{220}Rn progeny shows that ^{220}Rn progeny concentrations are much more sensitive to the variability of the independent variables for prediction with respect to those for estimation, because the ^{220}Rn and progeny have a much lower air concentrations.

Because the multiple regressions from the months January, April, July and October have been used to predict the progeny concentrations in the next months, in Fig. 3 there are given the time distributions of the predicted air concentrations for February, May, August and November and in Fig. 4 the corresponding relative error for each time interval from the time series. A general feature of these figures is the more pronounced closeness of the predicted and measured concentrations, namely, the lowest relative errors, in the first days of the respective month, what is indicating that the independent and dependent variables in this period have a time variation close to that of previous month. From Fig. 4 one may see that the relative errors associated to thoron progeny prediction are much greater than those of radon progeny, especially in the second part of the months.

The linear character of multiple regression, possibly, has determined the occurrence of the negative values of progeny concentrations in some time intervals at the end of prediction period. This aspect is obvious, particularly, in case of thoron progeny (August 31, November 25) because their concentrations having low values are more sensitive to the internal non-linearity of the atmospheric dynamics.

Another objective of this paper was the study of the prediction performances, considering different time periods on which the estimation regression has been built.

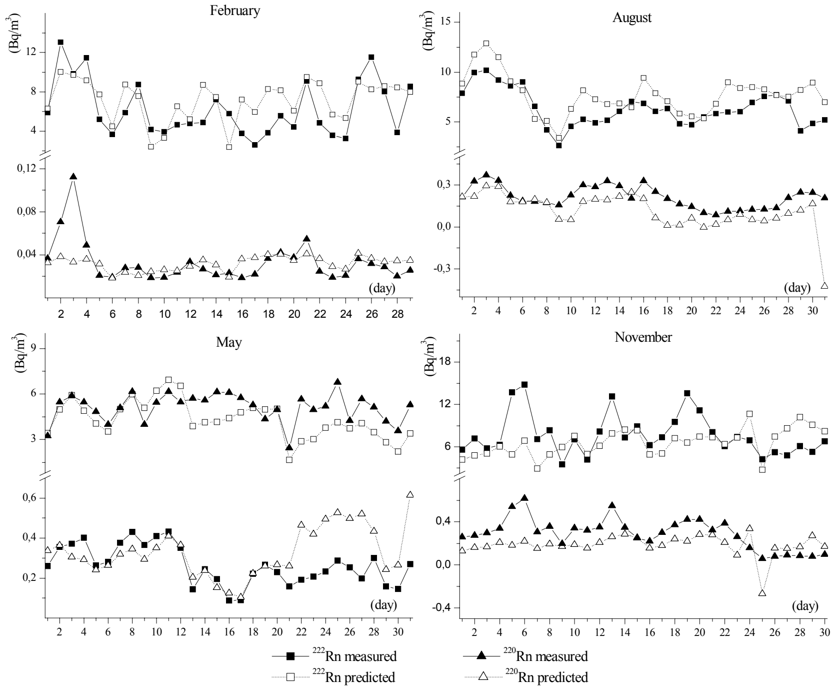

Fig. 3 – Measured and predicted progeny concentrations in February, May, August and November.

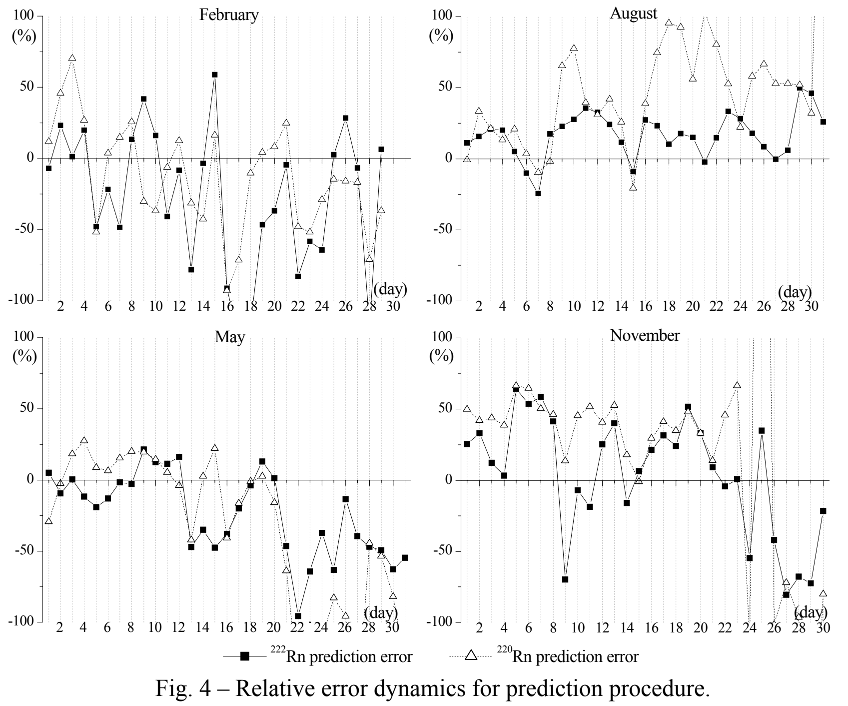

Fig. 4 – Relative error dynamics for prediction procedure.

In Table 5 there are presented the statistics of ^{222}Rn and ^{220}Rn progeny concentration prediction for following time periods on which the corresponding regressions have been built up: one month, two months, three months and one year.

Table 5

Statistics of ^{222}Rn and ^{220}Rn progeny prediction for different time period of the regression building-up

Month*	June			December		
	R	***p*-value**	**SSR**	***R***	***p*-value**	**SSR**
Statistics of ^{222}Rn progeny prediction						
1	0.644	$1.1e^{-4}$	62.823	0.260	$1.6e^{-1}$	320.660
2	0.582	$6.9e^{-4}$	73.546	0.674	$2.8e^{-5}$	118.278
3	0.443	$1.4e^{-2}$	114.884	0.653	$6.0e^{-5}$	127.507
12	0.622	$2.2e^{-4}$	124.804	0.606	$2.8e^{-4}$	165.448
Statistics of ^{220}Rn progeny prediction						
1	0.426	$1.8e^{-2}$	0.302	0.209	$2.6e^{-1}$	1.576
2	0.662	$5.9e^{-5}$	0.367	0.441	$1.3e^{-2}$	0.785
3	0.658	$6.7e^{-5}$	0.364	0.296	$1.1e^{-1}$	0.748
12	0.668	$4.7e^{-5}$	0.261	0.187	$3.1e^{-1}$	0.183

* The length of the periods of the regression building-up.

In case of ^{222}Rn progeny, for the month of June, the consideration of periods for regression building up longer than one month does not contribute to the improvement of the prediction statistics, while for December, periods for regression building up longer than one month generate better statistics: greater R (0.606 ÷ 0.674), very high level of significance ($2.8\cdot10^{-5}$ ÷ $2.8\cdot10^{-4}$) and much lower SSR (118.278 ÷ 165.448), as compared with corresponding statistics of the one month case ($R = 0.260$, $p \le 1.6\cdot10^{-1}$, SSR = 320.660).

As far as the ^{220}Rn progeny are concerned, in case of June, the longer periods for regression building up have generated better statistics: greater R(0.658 ÷ 0.668) and higher level of significance($p \le 4.7\cdot10^{-5}$ ÷$6.7\cdot10^{-5}$); for the month of December, the longer periods for regression building up determine low improvement of the correlation coefficient but with low level of significance.

5. CONCLUSIONS

In the present paper there have been studied the statistical and physical aspects of the modeling of the ^{222}Rn and ^{220}Rn progeny concentrations using multiple linear regression with current and lagged variables as predictors. The regression statistics for the progeny concentration estimation show that the greatest values of multiple R with high level of significance, have been obtained in the months with the air temperature included in the sets of predictor variables: December '95, May, June and July. This proves that radon having the lifetime comparable with the ventilation time of entire planetary boundary layer is very sensitive to the air temperature which induces the vertical transport of the air masses, this being the dominant physical process determining the time variation of the concentrations of atmospheric radon and its daughters. In case of thoron progeny, the regression statistics indicate that the greatest values of multiple *R* have been obtained for the month of July, where the set of variables contain the soil temperature as current and lagged variable. This fact reveals the importance of the soil temperature for the dynamics of the exhalation rate which is the dominant physical process causing time variation of the thoron and progeny concentrations. The analysis of the progeny concentration prediction has shown that the corresponding statistics are mainly dependent on the degree of variation of the independent variables from the month on which the prediction is performed, as against the same variables from the month on which the regression has been built up. This degree of variation being, generally, minimum at the beginning of the prediction period, the relative errors have the lowest values in the early days of this period. The limitation of the prediction procedure related to the degree of variation mentioned above, is more pronounced in case of the ^{220}Rn progeny concentration prediction, because the thoron progeny concentrations being much lower than those of radon progeny are more sensitive to the variations of the dependent and independent variables used in the estimation and prediction procedures.

The study regarding the influence of the time period length on which the estimation regression has been built up, on the prediction performances, has indicated that multiple linear regressions built up on longer periods can improve the prediction performances of these regressions in certain months. The explanation consists in the fact that a longer period for prediction building-up contains much more information on the variability of both the independent and dependent variables which can be similar to those from the month under prediction process. The negative values of thoron progeny concentrations occurred at the end of the prediction period can be explained by the linear nature of the considered regression which cannot sufficiently well simulate the nonlinear dynamics of atmosphere.

*Acknowledgments***.** Our colleagues from Environmental Radioactivity Monitoring Station (ERMS) Bacau are acknowledged for their effort in sample collection. One author, F. SIMION, was supported by the strategic grant POSDRU/88/1.5/S/56668 – priority axis nr.1, co-financed by the European Social Found within the Sectorial Operational Program for Human Resources Development 2007–2013.

REFERENCES

1. C.R. Hosler, *Meteorlogical effects on atmospheric concentrations of radon (Rn^{222}), RaB(Pb^{214}), and RaC(Bi^{214}) near the ground,* Monthly Weather Review, **94**, *2*, 89–99 (1966).
2. M. Wilkening, *Radon in atmospheric studies: a review*, Second Special Symposium, Bhabha Atomic Research Center Bombay, CONF-810153-3 DE93 003345, India, 1981.

3. J. Porstendorfer, G. Butterweck, A. Reineking, *Diurnal variation of the concentrations of Radon and its short-lived daughters in the atmosphere near the ground*, Atmos. Environ., **25A**, *3/4*, 709–713 (1991).
4. MC O'Riodan, Miles, J.C. (Eds), *Radon 2000- Proceedings of a Conference held in London March 26–27*. Radiat. Prot. Dosim. **42**, *3*, 147–262 (1992).
5. United Nations Committee on the Effects of Atomic Radiation (UNSCEAR), *Ionizing radiation: sources and biological effects, Report to the General Assembly* (Annex D), New York, 1982.
6. V. Cuculeanu, A. Lupu, E. Sütö, *Fractal dimensions of the outdoor radon isotopes time series*. Env. Int., **22**(Suppl.1), 171–179 (1996).
7. V. Cuculeanu, A. Lupu, *Deterministic chaos in atmospheric radon dynamics*, J. Geophys. Res., **106**, *D16*, 17961–17968 (2001).
8. J.P. McLaughlin, S.E. Simopoulos, F. Steinhäusler (Eds), Seventh International Symposium on the *Natural Radiation Environment* (NRE VII), Rhodes, Greece, 20–24 May 2002, 2005.
9. C.A. Baciu, *Radon and thoron progeny concentration variability in relation to meteorological conditions at Bucharest (Romania)*, J. Environ. Radioact., **83**, *2*, 171–189 (2005).
10. A.S. Paschoa, F. Steinhäusler, (Eds)., The 8th International Symposium *Natural Radiation Environment* – (NRE VIII) Buzios, Rio de Janeiro, Brazil, 7–12 October 2007. AIP Conference Proceedings 1034, New York.
11. C. Papastefanou, *Radon decay product aerosols in ambient air*, Aerosols and Air Qual., Res. **9**, 385–393 (2009).
12. V. Cuculeanu, F. Simion, E. Simion, A. Geicu,. *Dynamics, deterministic nature and correlations of outdoor ^{222}Rn and ^{220}Rn progeny concentrations measured at Bacău* (*Romania*), J. Environ. Radioact., **102**, 703–712 (2011).
13. G. Butterweck, A. Reineking, J. Kersten, J. Porstendörfer. *The use of the natural radioactive noble gases radon and thoron as tracers for the study of turbulent exchange in the atmospheric boundary layer*, Atmos. Environ., **28**, 1963–1969 (1994).
14. J.B. Cromwell, M.J. Hannan, W.C. Labys, M. Terreza, *Multivariate tests for time series models.* SAGE Publications, Inc 1994.
15. M.C. Hubbard, W.G. Cobourn, *Development of a regression model to forecast ground-level ozone concentration in Louisville*, KY. Atmos.Environ., **32**, *14/15*, 2637–2647 (1998).
16. C. Perrino, A. Pietrodangelo, A. Febo, *An atmospheric stability index based on radon progeny measurements for the evaluation of primary urban pollution*, Atmos. Environ. **35**, 5235–5244 (2001).
17. A. Vlachogianni, P. Kassomenos, A. Karppinen, S. Karakitsios, J. Kukkonen, *Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM_{10} in Athens and Helsinki*, Science of the Total Environment, **409**, 1559–1571 (2011).
18. D.S. Wilks, *Statistical methods in atmospheric sciences: an introduction*, Academic Press, 1995.
19. E. Kovac-Andric, J. Brana, V. Gvozdic, *Impact of meteorological factor on ozone concentrations modeled by time series analysis and multivariate statistical methods*, Ecological Informatics, **4**, 117–122 (2009).
20. J.Porstendorfer, *Properties and behaviour of radon and thoron and their decay products in air.* J.Aerosol Sci., **25**, *2*, pp. 219–263 (1994).
21. M. Schmidt, R. Graul, H. Sartorius, I. Levin, *Carbon dioxide and methane in continental Europe: a climatology, and ^{222}Radon-based emission estimates*, Tellus, **48(B)**, 457–473 (1996).
22. W. Jacobi, K. André, *The vertical distribution of Radon 222, Radon 220 and their Decay products in the Atmosphere*. J. Geophys. Res., **68**, *13*, 3799–3814 (1963).
23. P.A. Rogerson, *A Statistical Method for the Detection of Geographic Clustering*, 2001.
24. P. Wessa. *Free Statistics Software*, Office for Research Development and Education, Version 1.1.23-r7, URL http://www.wessa.net/ (2012).
25. J. M. Chambers, *Linear models*, eds. J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole, 1992.