Development of Predictive Model for Radon-222 Estimation in the Atmosphere using Stepwise Regression and Grid Search Based-Random Forest Regression

Omodele E. Olubi[a,b], Ebenezer O. Oniya[a,*], Taoreed O. Owolabi[a]

[a]*Physics and Electronics Department, Adekunle Ajasin University, Akungba-Akoko, 342111, Ondo State, Nigeria.*
[b]*Achievers university, P.M.B 1030, Owo, Ondo State, Nigeria.*

Abstract

This work develops predictive models for estimating radon (^{222}Rn) activity concentration in the regression (SWR). The developed models employ meteorological parameters which include the temperature, pressure, relative and absolute humidity, wind speed and wind direction as descriptors. Experimental data of radon concentration and meteorological parameters from two observatories of the Korea Polar Research Institute in Antarctica (King Sejong and Jang Bogo) have been employed in this work. The performance of the developed models was assessed using three different performance measuring parameters. On the basis of root mean square error (RMSE), the GS-RFR shows better performance over the SWR. An improvement of 64.09 % and 15.19 % was obtained on the training and test datasets, respectively at King Sejong station. At the Jang Bogo station, an improvement of 75.04 % and 28.04 % was obtained on the training and test datasets, respectively. The precision and robustness of the developed models would be of significant interest in determining the concentration of radon (^{222}Rn) activity concentration in the atmosphere for various physical applications especially in regions where field measuring equipment for radon is not available or measurements have been interrupted.

1. Introduction

The importance of radon (Rn-222), the only gaseous member of the U-238 series, has been of interest to scientists since the twentieth century when it was first suspected to be a caus-ative agent for lung cancer among miners. The radioactive gas has been a significant subject of research among health and environmental scientists having been characterized as a potential indoor source of air pollution. Its subsequent classification as a carcinogen has led to investigation and monitoring of the indoor concentration of the gas in several countries of the world [1-8]. The source of the noble gas is from the decay of Ra-226 in bedrock and soil and migrates through soil pores by gas-phase diffusion and advection to the surface and its sink process is by radioactive decay [9, 10].

Due to some important characteristics of radon as a tracer of atmospheric processes, there has been a growing interest in recent decades in monitoring environmental radon. Being a noble gas, it is not chemically reactive with other elements.

*Corresponding author tel. no: +2348035033421
Email address: ebenezer.oniya@aaua.edu. (Ebenezer O. Oniya)

Its relative solubility in water and non-attachment to aerosols makes it highly insusceptible to dry or wet atmospheric removal processes. Its half-life of 3.82 days is comparable to the life times of short-lived environmental pollutants (e.g NO_x, SO_2, CO, O_3, CH_4) and atmospheric residence times of water and aerosols [10].

The noble gas has become very useful as a tracer of the influence of the terrestrial environment on air mass composition. Some areas of application of ground-based radon observation include atmospheric, pollution studies and climatic studies [11-16]. Observed anomalous behaviour of radon in soil and groundwater during earthquake events has been employed as a precursor for impending earthquakes [17, 18].

Despite the progress that has been made in radon instrumentation, access to data on atmospheric radon concentration is still to a large extent, lacking in the public domain. Africa for instance, has only one mention of a radon observatory in the literature; an ANSTO-developed detector installed at a Global Atmospheric watch (GAW) station at Cape Point, South Africa [10]. Ground based radon measurement methods have not been applied to study atmospheric processes as have been done in Europe. As a matter of fact, the only radon time series characterization to have been reported was published recently for the first time on the continent [19]. In the unavailability of measuring equipment, a theoretical approach to developing predictive models for radon concentration in the atmosphere may be a viable step in generating synthetic data for the noble gas, using machine learning to train available experimental data. Theoretical models have been developed by several researchers in the literature to predict radon behaviour and concentration for various conditions and applications [17, 18, 20-22]. Several studies in the literature have reported the variation of atmospheric radon and its progenies with changes in meteorological parameters like temperature, pressure, humidity and windspeed [23-25]. [26] used these meteorological variables as independent predictors in the development of a multiple linear regression model for estimation and prediction of the time series of radon and thoron progeny concentrations.

Random forest (RF) methodology is a machine learning technique developed by Leo Breiman and is useful for classification and prediction problems [27]. Its algorithm operates by sampling small divisions of the data, grows a tree predictor that is randomized on each small division, then aggregates these predictors together. It applies bootstrap aggregation and random feature selection to individual classification or regression trees for prediction [28]. Apart from the speed and ease of implementation of random forests, their predictions are remarkably accurate, with the ability to process a very large input data whilst dealing with overfitting. They also perform well with small to medium data [29]. Their good predictive abilities have made them highly applicable to regression and classification problems in the atmospheric sciences [30-31].

The Grid Search (GS) is one of the algorithms for hyperparameter optimisation and tuning of models with an expectation of the most accurate results. With a specified subset of the hyperparameters space of the training algorithm, the algorithm conducts a search with the aim of producing the best combination of parameters yielding the most remarkable results. To apply a grid search, boundaries need to be specified because some parameters within the algorithm's parameter space may contain unlimited values. The high dimensional space problem with grid search algorithms is easily resolved with parallelization of the of the process since the hyperparameters are usually not dependent on each other [32].

Multiple stepwise regression is efficient in the selection of contributing factors used in establishing models that can do statistical prediction. The critical objective it sets to achieve is to discover the most cordial relationship between predictor variables that would accurately forecast the predicted variable. The regression process begins with the input of the mostly contributing predictor variable to the prediction model. Additional variables are continuously added as long as they are of any essence to the regression equation. [33, 34].

This present work develops stepwise regression (SWR) and grid search-based random forest regression (GS-RFR) models through which radon concentration can be estimated and predicted using much more available meteorological parameters (air temperature (AT), atmospheric pressure (AP), absolute humidity (AH), relative humidity (RH), wind direction (WD) and wind speed (WS) as predictors. A comparison is also made between both models in terms of performance. The proposed model will help not only to predict radon concentration, it may also help to generate estimated or synthetic radon data that can approximate measured data, for regions that lack measuring instruments for atmospheric radon but have access to meteorological data. It will also help estimate radon data for sites where measurements have been interrupted.

2. Theory

2.1. Description of Random Forest Regression

A random forest is described, according to [35], to consist of N regression trees that are randomized also referred to as a family. For any individual (i-th) tree, the predicted value at the query point $\mathbf{y}$ can be represented as $m_n(\mathbf{x}; \Theta_i, D_n)$, where $\Theta_1, \ldots, \Theta_N$ are independent random variables that are not dependent on D_n. Resampling of the training set is first done using Θ before individual trees are grown. The finite forest estimate for regression as a result of the combination of the trees is

$$m_{N,n}(\mathbf{x}; \Theta_1, \ldots \Theta_N, D_n) = \frac{1}{N}\sum_{i=1}^{N} m_n(\mathbf{x}; \Theta_i, D_n) \qquad (1)$$

In the case of classification, the random forest classification makes use of the majority vote among the classification trees. The forest estimate for classification is

$$m_{N,n}(\mathbf{x}; \Theta_1, \ldots \Theta_N, D_n) = \begin{cases} 1 \text{ if } \frac{1}{N}\sum_{i=1}^{N} m_n(\mathbf{x}; \Theta_i, D_n) > \frac{1}{2} \\ 0 \text{ if otherwise} \end{cases} \qquad (2)$$

2.2. Description of Grid Search optimization

The implementation of the grid search technique involves upper and lower bound vectors $V = V_1, V_2, \ldots, V_q$ and $W = W_1, W_2, \ldots, W_q$ respectively, defined for each component of hyperparameters $H = H_1, H_2, \ldots, H_q$ where q is the number of hyperparameters. The optimization and parameter search procedure involves taking n equally spaced points within the search space over an interval of the form $[V_i, W_i]$ which includes of V_i and W_i. The algorithm searches through n^q possible points and a selection of the optimum values results, following the evaluation of each grid point in space [36].

2.3. Stepwise Regression

Based on the forward and backward selection, stepwise regression is a self-determining process for in the selection of independent variables. Multiple linear regression (MLR) has the form

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \cdots + \beta_p X_p + \varepsilon \tag{3}$$

In equation (3), Y is the output variable and $X_1, X_2, X_3\ldots$ are predictor variables. β_i are regression parameters, β_o is an intercept and ε is the random error term. The process is summarised below;

1. If after the performance of simple multiple linear regression of n predictor variables, all the variables show remarkable significance, then the whole model containing all n variables is adopted.
2. If results show otherwise, simple n-variable linear regression is performed with each of the predictor variables and the process selects the variable which gives lowest p-value for t-test.
3. A subsequent $n-1$ variable regression is performed taking the selected variable in step 2 as common.
4. Step 3 is repeated with each significant variable becoming added to the model in a stepwise manner. The test for significance by stepwise regression can be applied at two levels. The first being for addition of variables and the second, for removal of variables [37].

2.4. Performance measuring parameters

Three performance measuring parameters were used to assess the developed models namely correlation coefficient (CC), root mean square error (RMSE), mean absolute error (MAE). Correlation coefficient is defined as

$$CC = \frac{\sum_{i=1}^{N} \left(Y_i{}^* - \overline{Y^*} \right) \left(Y_i - \overline{Y} \right)}{\sqrt{\sum_{i=1}^{N} \left(Y_i{}^* - \overline{Y^*} \right)^2} \sqrt{\sum_{i=1}^{N} \left(Y_i - \overline{Y} \right)^2}} \tag{4}$$

where where $Y_i{}^*$ and Y_i are the mean values of the predicted and actual outputs.
RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_i - Y_i*)^2} \tag{5}$$

N represents the number of samples contained in the dataset
MAE is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |Y_i - Y_i*| \tag{6}$$

3. Methods

3.1. Description of sites

The data used in this work was published by [38], being data measured in two Korea Antarctic Research Program stations namely King Sejong (KSG) and Jang Bogo (JBS). Measurements have been done jointly with the Australian Nuclear Science and Technology Organisation (ANSTO). The Korea Polar Research Institute (KOPRI) operates the KSG station (62.217° S, 58.783° W). The station functions as a regional World Meteorological Organisation (WMO) Global Atmospheric Watch (GAW) station. The JBS is 10 m above sea level with coordinates (74.623°S, 164.228°E). A detailed geographical description of the sites is seen in [39].

3.2. Radon and Meteorological Data

At JBS, radon measurements have been made using a 1200 L two-filter dual-flow-loop radon detector custom built by ANSTO. Installed approximately 40 m east of the main station, air is sampled at 55 L min^{-1} through 50 mm high-density polyethylene pipe from approximately 6 m above ground level. In order to avoid thoron (^{220}Rn; t$_{0.5}$ = 55.6 s) from entering into the pipe and contaminating sampled air, a 400 L delay volume is coupled within the sampling line. At approximately 170 m from the radon detector, meteorological data was collected from a 10 m tower with instrumentation composed of a sonic anemometer, temperature and humidity probe, barometer and a windspeed logger. In post processing, all observations are aggregated to hourly values [39]. A radon detector similar in operation to that in JBS but with a volume of 1500 L was used for radon data collection in KSG with meteorological data collected from a nearby observation system [40]. The dataset used was measured between December and February with 1818 and 1955 datapoints for JBS and KSG respectively.

Table 1 shows the statistical analysis from JBS and KSG. The mean, standard deviation and range are presented. While the mean and range describe the content of the dataset, the correlation coefficients depict the level of linear relationship between the target and predictor variables. Both tables indicate weak correlation between A_{Rn} and the descriptors indicating that a purely linear regression model may be insufficient to represent the relationship between the descriptors and target.

3.3. Computational methodology
3.3.1. SWR model

A flow chart of the stepwise process is presented in figure 1. Whenever a variable x is added in each step, all the predictor variables in the model are assessed for their significance p. If it has been reduced below a specified threshold.

Table 1: Statistical analysis of dataset from JBS

	Correlation Coefficient	Mean	Standard Deviation	Range
A_{Rn} (Bq/m^3)	1	0.937	0.743	5.213
WS (m/s)	-0.32	3.723	3.284	17.600
WD (^{o})	-0.27	167.766	112.667	359.700
AT (^{o}C)	0.22	-3.055	3.814	19.300
RH (%)	-0.11	57.524	16.121	72.400
AP (hPa)	-0.16	982.704	7.370	36.700
AH (g/m^3)	0.073	2.317	0.822	4.05

Table 2: Statistical analysis of dataset from KSG

	Correlation Coefficient	Mean	Standard Deviation	Range
A_{Rn} (Bq/m^3)	1	63.344	30.921	314.060
WS (m/s)	-0.05	7.043	3.463	18.600
WD (^{o})	-0.09	235.993	96.123	358.500
AT (^{o}C)	0.40	-0.023	1.447	11.600
RH (%)	0.12	84.583	8.763	41.200
AP (hPa)	-0.02	987.753	9.369	43.000
AH (g/m^3)	0.39	4.111	4.111	4.070
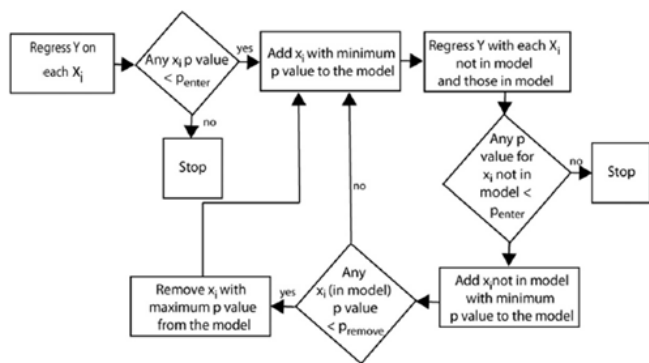
Figure 1: The stepwise regression flow chart

3.4. Building of GS-RFR model

The computation of the GS-RFR model development was achieved using PYTHON software. The radon concentration and the descriptors, which include (air temperature (AT), atmospheric pressure (AP), absolute humidity (AH), relative humidity (RH), wind direction (WD) and wind speed (WS), after randomization, was partitioned into training and testing sets in the ratio of 8:2 respectively. The RFR model development was done with the training set, while the general predictive ability of the model was assessed using the 20% test set. A helpful purpose for randomization is that it enhances computation efficiency by ensuring unbiased spread of data during both the training or testing phase. The performance algorithm was optimized through an optimum selection of hyperparameters using grid search (GS) with cross validation. Table 3 below shows the hyperparameters that were tuned as suggested in the literature [41, 42]. During the hyperparameters tuning, the 5-fold cross validation was used as the fitness function. Verification of the RF model with the optimum hy-

perparameters was carried out on the testing set.
*

4. 4. Results and Discussion

4.1. Comparison of Performance between the SWR and GS-RFR

For the two datasets, the performance of the two models developed by SWR and GS-RFR is depicted in figure 1. The predictive capabilities of the two models were assessed using the performance measuring parameters: correlation coefficient (CC), root mean square error (RMSE) and mean absolute error (MAE).

Tables 4 and 5 shows the estimated predictive performance for the two regression methods based on correlation coefficient, root mean square error and mean absolute error. Figure 2 compares the performance of the test set on the models developed using the data from KSG and JBS. The figures show better performance by the GS-RFR model over the more traditional SWR. Considering RMSE, an improvement of 64.09 % and 15.19 % was obtained on the training and test datasets, respectively at KSG while at JBS, an improvement of 75.04 % and 28.04 % was obtained on the training and test datasets, respectively (Table 6). The optimum hyperparameters for the RFR algorithm for each dataset is summarized in Table 7.

Table 4. Predictive performance of the two regression models in terms of Correlation Coefficient (CC), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for KSJ dataset.

The GS-RFR model presents the smallest RMSE on the two datasets employed. It also achieved the highest correlation coefficient on both training and test sets. The plots in figure 3 show the correlation between predicted and measured values of radon concentration. It can be seen that the GS-RFR model

Table 3: Hyperparameters description

No	Hyperparameters	Definition
1	Max_depth	The maximum depth of Decision trees (DT).
2	Min_samples_split	The minimum number of samples for the split
3	Min_samples_leaf	The minimum number of samples at the leaf node
4	n_estimators	The number of trees in the forest of the model
5	Max_features	The number of features considered during the selection of the best splitting

Table 4: Predictive performance of the two regression models in terms of Correlation Coefficient (CC), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for KSJ dataset

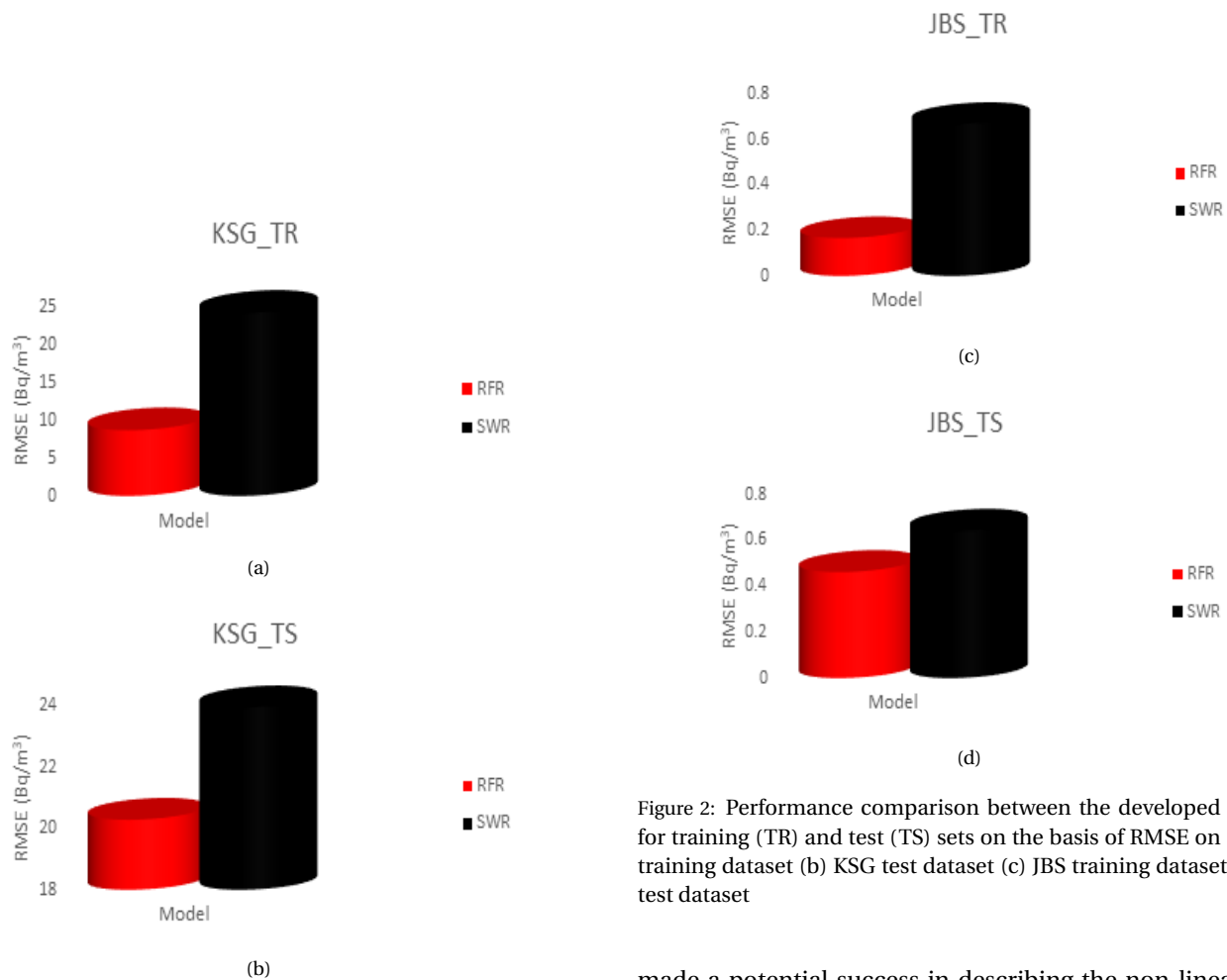| Method	CC		RMSE (Bq/m^3)		MAE (Bq/m^3)	
	Training	Test	Training	Test	Training	Test
GS-RFR	0.99	0.83	8.57	20.26	5.38	14.07
SWR	0.64	0.61	23.86	23.89	0.01	0.10

(a)

(b)

(c)

(d)

Figure 2: Performance comparison between the developed models for training (TR) and test (TS) sets on the basis of RMSE on (a) KSG training dataset (b) KSG test dataset (c) JBS training dataset (d) JBS test dataset

made a potential success in describing the non-linear relationship between atmospheric radon concentration and influencing meteorological parameters considering strong correlation coefficients it achieved.

5. Conclusion

In this work, modelling of atmospheric radon as done using the more traditional stepwise regression (SWR) and a novel

Table 5: Predictive performance of the two regression models in terms of Correlation Coefficient (CC), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for KSJ dataset

Method	CC		RMSE (Bq/m^3)		MAE (Bq/m^3)	
	Training	Test	Training	Test	Training	Test
GS-RFR	0.98	0.82	0.17	0.46	0.11	0.32
SWR	0.61	0.66	0.66	0.63	0.06	6553

Table 6: Improvement of GS-RFR over SWR in this study
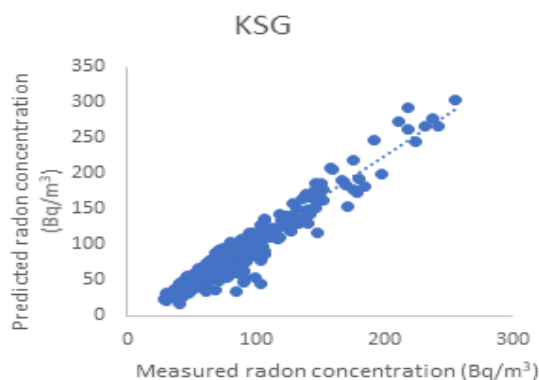
JBS		KSG	
Training set	Test set	Training set	Test set
75.04 %	28.04 %	64.09 %	15.19 %

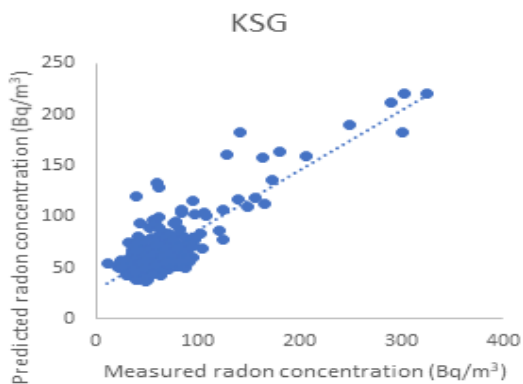Table 7: Selected optimum hyperparameters after the grid search

Hyperparameters	JBS	KSG
Max_depth	2000	2000
Min_samples_split	2	3
Min_samples_leaf	1	1
n_estimators	650	50
Max_features	3	3

(c)

(d)

(a)

(b)
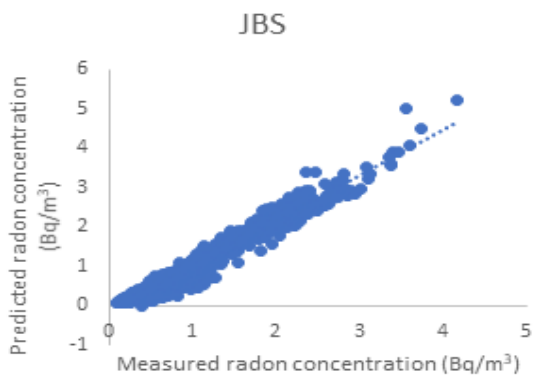
Figure 3: Performance comparison between the developed models for training (TR) and test (TS) sets on the basis of RMSE on (a) KSG training dataset (b) KSG test dataset (c) JBS training dataset (d) JBS test dataset.

grid search based random forest regression (GS-RFR). Datasets from two radon stations in Antarctica were used in the building of the models. Important factors such as air temperature, atmospheric pressure, absolute humidity, relative humidity, wind direction and wind speed were used as predictors. Comparing both models, the results show that the GS-RFR model performed better on both datasets in the training and testing phases. It presents a respective training and test improvement of 64.09 % and 15.19 % on one dataset and 75.04 % and 28.04 % on the other. Atmospheric radon data, which is finding more relevance today in the atmospheric sciences, is still scarce and not readily available. The precision and robustness of the developed models would be of significant interest in determining the concentration of radon (^{222}Rn) activity concentration in the atmosphere for various physical applications especially in regions where field measuring equipment for radon is not available but have meteorological parameters are.

Acknowledgment

References

[1] O. S. Ajayi, E. O. Owoola, O. E. Olubi & C. G. Dike, "Survey of indoor radon levels in some universities in southwestern Nigeria", Radiation Protection Dosimetry **87** (2019) 34. https://doi.org/10.1080/15275922.2016.1230909

[2] J. Chen & K. L. Ford, "A study on the correlation between soil radon potential and average indoor radon potential in Canadian cities", Journal of Environmental Radioactivity, **166** (2017) 152. https://doi.org/10.1016/j.jenvrad.2016.01.018

[3] C. Grossi, A. Àgueda, F. R. Vogel, A. Vargas, M. Zimnoch, P. Wach, J. E. Martín, I. López-Coto, J. P. Bolívar, J. A. Morguí & X. Rodó, "Analysis of ground-based 222Rn measurements over Spain: Filling the gap in southwestern Europe", Journal of Geophysical Research **121** (2016) 11,021. https://doi.org/10.1002/2016JD025196

[4] I. Lázár, E. Tóth, G. Marx, I. Cziegler & G. J. Köteles, "Effects of residential radon on cancer incidence", Journal of Radioanalytical and Nuclear Chemistry, **258** (2003) 519.

[5] A. M. Maghraby, K. Alzimami & M. Abo-Elmagd, "Estimation of the residential radon levels and the population annual effective dose in dwellings of Al-kharj, Saudi Arabia", Journal of Radiation Research and Applied Sciences, **7** (2014) 577. https://doi.org/10.1016/j.jrras.2014.09.013

[6] W. J. Mccarthy, R. Meza, J. Jeon & S. H. Moolgavkar, "Chapter 6: Lung cancer in never smokers: Epidemiology and risk prediction models", Risk Analysis **32** (2012) 69. https://doi.org/10.1111/j.1539-6924.2012.01768.x

[7] V. T. Rasmussen, "Determining the mean year value of radon in the indoor air", MATEC Web of Conferences, **282** (2019) 02001. https://doi.org/10.1051/matecconf/201928202001

[8] K. Walczak, J. Olszewski, P. Politański & M. Zmyślony, "Occupational exposure to radon for underground tourist routes in Poland: doses to lung and the risk of developing lung cancer", International Journal of Occupational Medicine and Environmental Health, **30** (2017) 687. https://doi.org/10.13075/ijomeh.1896.00987

[9] F. Giustini, G. Ciotoli, A. Rinaldini, L. Ruggiero & M. Voltaggio, "Mapping the geogenic radon potential and radon risk by using Empirical Bayesian Kriging regression: A case study from a volcanic area of central Italy", Science of The Total Environment, **661** (2019) 449. https://doi.org/10.1016/j.scitotenv.2019.01.146

[10] W. Zahorowski, S. D. Chambers & A. Henderson-Sellers, "Ground based radon-222 observations and their application to atmospheric studies" Journal of Environmental Radioactivity **76** (2004) 3. https://doi.org/10.1016/j.jenvrad.2004.03.033

[11] S. D.Chambers, D. Scott, W. Zahorowski, A. G. Williams, J. Crawford & A. D. Griffiths, "Identifying tropospheric baseline air masses at Mauna Loa Observatory between 2004 and 2010 using Radon-222 and back trajectories: Radon-Derived Mauna Loa Baseline Events", Journal of Geophysical Research: Atmospheres **118** (2013) 992. https://doi.org/10.1029/2012JD018212

[12] D. Desideri, C. Roselli, L. Feduzi & M. Assunta Meli, "Monitoring the atmospheric stability by using radon concentration measurements: A study in a Central Italy site", Journal of Radioanalytical and Nuclear Chemistry, **270** (2006) 523. https://doi.org/10.1007/s10967-006-0458-1

[13] A. D. Griffiths, F. Conen, E. Weingartner, L. Zimmermann, S. D. Chambers, A. G. Williams & M. Steinbacher, "Surface-to-mountaintop transport characterised by radon observations at the Jungfraujoch", Atmospheric Chemistry and Physics **14** (2014) 12763. https://doi.org/10.5194/acp-14-12763-2014

[14] A. Podstawczyńska & S. D. Chambers, "Radon-based technique for the analysis of atmospheric stability – a case study from Central Poland", Nukleonika **63** (2018) 47. https://doi.org/10.2478/nuka-2018-0006

[15] R. Vecchi, F. A. Piziali, G. Valli, M. Favaron & V. Bernardoni, "Radon-based estimates of equivalent mixing layer heights: a long-term assessment", Atmospheric Environment **197** (2019) 150. https://doi.org/10.1016/j.atmosenv.2018.10.020

[16] A. G. Williams, S. D. Chambers, F. Conen, S. Reimann, M. Hill, A. D. Griffiths & J. Crawford, "Radon as a tracer of atmospheric influences on traffic-related air pollution in a small inland city", Tellus B: Chemical and Physical Meteorology, **68** (2016) 30967. https://doi.org/10.3402/tellusb.v68.30967

[17] A. Pasini, R. Salzano & A. Attanasio, "Modeling Radon Behavior for Characterizing and Forecasting Geophysical Variables at the Atmosphere–Soil Interface", In: *Sengupta D. (eds) Recent Trends in Modelling of Environmental Contaminants*, Springer, New Delhi,2014. https://doi.org/10.1007/978-81-322-1783-1_9

[18] B. Zmazek, L. Todorovski, S. Džeroski, J. Vaupotič & I. Kobal, "Application of decision trees to the analysis of soil radon data for earthquake prediction", Applied Radiation and Isotopes **58** (2003) 697. https://doi.org/10.1016/S0969-8043(**??**)00094-0

[19] R. Botha, C. Labuschagne, A. G. Williams, G. Bosman, E. G. Brunke, A. Rossouw & R. Lindsay, "Characterising fifteen years of continuous atmospheric radon activity observations at Cape Point (South Africa)", Atmospheric Environment, **176** (2018) 30. https://doi.org/10.1016/j.atmosenv.2017.12.010

[20] K. M. Ajayi, K. Shahbazi, P. Tukkaraja & K. Katzenstein, "A discrete model for prediction of radon flux from fractured rocks", Journal of Rock Mechanics and Geotechnical Engineering **10** (2018) 879. https://doi.org/10.1016/j.jrmge.2018.02.009

[21] A. V. Glushkov, O. Yu Khetselius, V. V. Buyadzhi, Y. V. Dubrovskaya, I. N. Serga, E. V. Agayar & V. B. Ternovsky, "Nonlinear chaos-dynamical approach to analysis of atmospheric radon 222Rn concentration time series", Indian Academy of Sciences – Conference Series **1** (2017) 61. https://doi.org/10.29195/iascs.01.01.0025

[22] A. Pasini & F. Ameli, "Radon short range forecasting through time series preprocessing and neural network modeling: Radon Short Range Forecasting", Geophysical Research Letters **30** (2003) 1. https://doi.org/10.1029/2002GL016726

[23] M. Janik & P. Bossew, "Analysis of simultaneous time series of indoor, outdoor and soil air radon concentrations, meteorological and seismic data", Nukleonika **61** (2016) 295. https://doi.org/10.1515/nuka-2016-0049

[24] G. Mentes & I. Eper-Pápai, "Investigation of temperature and barometric pressure variation effects on radon concentration in the Sopronbánfalva Geodynamic Observatory, Hungary" Journal of Environmental Radioactivity, **149** (2016) 64.

https://doi.org/10.1016/j.jenvrad.2015.07.015

[25] K. Singh, M. Singh, S. Singh, H. S. Sahota & Z. Papp, "Variation of radon (Rn) progeny concentrations in outdoor air as a function of time, temperature and relative humidity", Radiation Measurements **39** (2005)213. https://doi.org/10.1016/j.radmeas.2004.06.015

[26] F. Simion, V. Cuculeanu, E. Simion & A. Geicu, "Modeling the 222Rn and 220Rn progeny concentrations in atmosphere using multiple linear regression with meteorological variables as predictors", Romanian Reports in Physics **65** (2013) 524.

[27] L. Breiman, "Random forests", Machine Learning **45** (2001)5. https://doi.org/10.1201/9780429469275-8

[28] R. Xu, *Improvements to random forest methodology* Dissertation (Doctor of Philosophy) Iowa State University, 2013. http://lib.dr.iastate.edu/etd/13052/

[29] G. Biau, "Analysis of a Random Forests Model", Journal of Machine Learning Research, 13 (2012) 1063.

[30] A. C. Keller & J. M. Evans, "Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10", Geoscientific Model Development **12** (2019) 1209. https://doi.org/10.5194/gmd-12-1209-2019

[31] A. Masih, "Application of Random Forest Algorithm to Predict the Atmospheric Concentration of NO2", 2019 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (US-BEREIT), 252–255.

[32] P. Liashchynskyi & P. Liashchynskyi, "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS", ArXiv (2019)1–11.

[33] T. Gao & L. Xie, *Multivariate regression analysis and statistical modeling for summer extreme precipitation over the Yangtze River basin, China,* Advances in Meteorology, 2014. https://doi.org/10.1155/2014/269059

[34] M. Lin, J. Tao, C. Y. Chan, J. J. Cao, Z. S. Zhang, L. H. Zhu & R. J. Zhang, "Regression analyses between recent air quality and visibility changes in megacities at four haze regions in china", Aerosol and Air Quality Research **12** (2012) 1049. https://doi.org/10.4209/aaqr.2011.11.0220

[35] G. Biau & E. Scornet, "A random forest guided tour", Test **25** (2016)197. https://doi.org/10.1007/s11749-016-0481-7

[36] S. M. I. Shamsah & T. O. Owolabi, "Empirical method for modeling crystal lattice parameters of A2XY6 cubic crystals using grid search-based extreme learning machine", Phys. J. Appl **128** (2020) 185106 . https://doi.org/10.1063/5.0024595.

[37] R. Silhavy, P. Silhavy & Z. Prokopova, "Evaluating subset selection methods for use case points estimation", Information and Software Technology **97** (2018) 1. https://doi.org/10.1016/j.infsof.2017.12.009

[38] S. Hong, *Radon 222 and meteorological time series at Jang Bogo and King Sejong Station, Antarctica, in 2015-2016,* Pangae, 2017. https://doi.org/https://doi.org/10.1594/PANGAEA.879451.

[39] S. D. Chambers, T. Choi, S. J. Park, A. G. Williams, S. B. Hong, L. Tositti, A. D. Griffiths, J. Crawford & E. Pereira, "Investigating Local and Remote Terrestrial Influence on Air Masses at Contrasting Antarctic Sites Using Radon-222 and Back Trajectories", Journal of Geophysical Research: Atmospheres, 122(2017)13525. https://doi.org/10.1002/2017JD026833.

[40] S. D. Chambers, S. B. Hong, A. G. Williams, J. Crawford, A. D. Griffiths & S. J. Park, "Characterising terrestrial influences on Antarctic air masses using Radon-222 measurements at King George Island", Atmospheric Chemistry and Physics **14** (2014) 9903. https://doi.org/10.5194/acp-14-9903.

[41] B. T. Pham, C. Qi, L. S. Ho, T. Nguyen-Thoi, N. Al-Ansari, M. D. Nguyen, H. D. Nguyen, H. B. Ly, H. Van Le & I. Prakash, "A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil", Sustainability (Switzerland) **12** (2020) 1. https://doi.org/10.3390/su12062218.

[42] C. Qi, Q. Chen, A. Fourie & Q. Zhang, "An intelligent modelling framework for mechanical properties of cemented paste backfill", Minerals Engineering **123** (2018) 16. https://doi.org/10.1016/j.mineng.2018.04.010