Trace and evaluation of the dependencies of Radon222 on air pressure and temperature using machine learning tools

Josephine S. Kramer [1]

Abstract

The pollution by radon is nowadays known as one of the most significant pollution in Switzerland. Furthermore, it is a unsolved problem. Usually, Radon 222 is a problem for soils and accumulation in sediments. Nevertheless, its increase in concentration can also be affected not from human activity, but from natural weather conditions. In the case of convective weather regimes, the formation of radon is enforced. For this research work, the dependencies of Radon on air pressure and seasons were looked at. For this purpose linear regression was compared with random forest regression and later evaluated. The statistical errors (root mean squared error, mean absolute error and mean squared error) were calculated to make a statement. Both regression methods performed well, nevertheless random forest has a minimal less error. Even if the random regressor acts well, the data has properties of a time series, which is why linear regression would be in general easier to use.

1. Introduction

Radon-222 is naturally emitted from land surfaces. The only sink of this noble gas in the atmosphere is radioactive decay. Its half-life of 3.8 days provides for large concentration differences between the planetary boundary layer and free tropospheric air (Veveva et al., 2009), making it a good tracer for recent land contact of air masses sampled at the high-altitude observatory Jungfraujoch. Switzerland is because of its geology exposed to a high radon concentration. It is striking that especially the regions in the Alps suffer of high Radon concentration. One reason for this is the change of air pressure and temperature with altitude. Exactly with this correlation is this work about. It is expexcted, from previous research that at high atmospheric pressure in the cold time of the year (November to March) inversion weather conditions are very stable (Jinzhao et al., 2013) and radon hardly comes from the planetary boundary layer to the Jungfraujoch.

The aim of this research paper is to understand the dependencies of Radon 222 on air pressure and temperature. For this work, the given data was analysed using machine learning tools. In the results chapter the founding are described and furthermore explained in the discussion.

2. Methods

2.1. Study Site

The data came from current and previous Radon 222 measurements of the measuring station on the Jungfraujoch, 3463 meters above sea level. This high alpine research station provides interesting data on climate (especially greenhouse gases). In this case, data from January 2019 to April 2021 were looked at, measured half-hourly, consisting of the date, the air pressure and the radon 222 concentration (measured half hourly). In total there are exactly 36 697 data points, but without blank measurements.

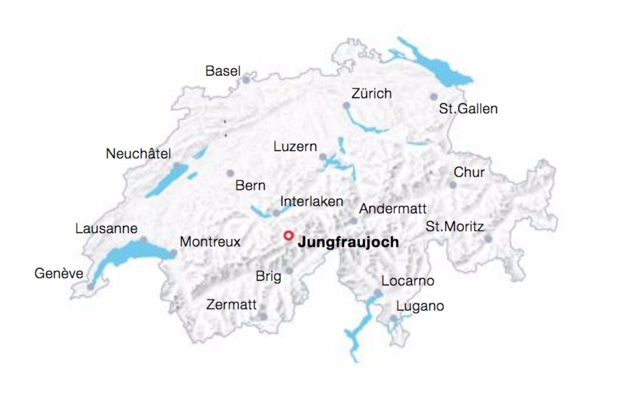


Figure 1. The location of the study site in Switzerland. Source: jungfraujoch.ch.

[1] Department of Environmental Science, University of Lausanne, Switzerland. Correspondence to: Josephine S. Kramer <josephine.kramer@unil.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Table 1. Test and train split for the here used Radon 222 Data.

SPLIT	VALUE
TRAIN	29597
TEST	7400

2.2. Data

In this case, data from January 2019 to April 2021 were looked at, consisting of the date and time, the air pressure (in ppm) and the Radon 222 concentration (in Bq m-3 STP) measured half hourly. In total there are exactly 36 697 data points, but without blank measurements. The data were provided in the form of an Excel file. Since blank measurements are taken every few days, this file had to be carefully edited before the actual analysis could begin. Blanks had values below zero, and since these would distort average values, they were not considered further.

2.3. Split and Analysis of Data

After the Excel file was successfully processed, the dataset was imported into GitHub. The data was plotted to see if any dependencies could be determined visually at first glance. Afterwards the test and train split was done to create test and train set (0.2 for the test set, and 0.8 for the train set). Now the linear regression was calculated. For this, first the dependent and the independent variable had to be determined. Radon is dependent in this case since it probably changes according to air pressure (but air pressure does not change according to radon concentration). So, radon was always used as” y” and air pressure as “x”. Thus it should be possible to see the dependence of radon in relation to air pressure. To see how the test set performs, it was now calculated how it predicts values and how they deviate from the real values. The next step was to evaluate the errors. For this purpose, the root mean square error, absolute mean error and mean squared error were calculated.

Moreover, the Random Forest Regressor was used to see a difference according to the regression. It also calculated the three statistical errors and later looked to see where they differed. To check if the result might be better with changed depth and number of branches, these were adjusted and again the errors were calculated. The two different regressions were then plotted, just to see a visually difference, and then discussed.

3. Results

3.1. Train and test split

It can be seen (Table 1) that the test split makes less than 30 percent of the data. It will be now used to predict the data.

3.2. Seasonality

Figure 2 shows a seasonal trend in Radon for the first 400 days of the dataset. It is striking that after nearly 365 days the course repeats itself.

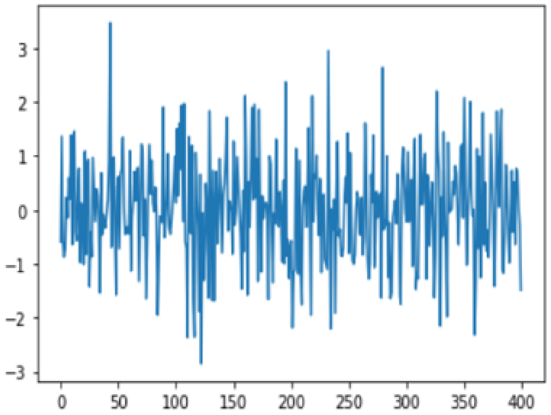
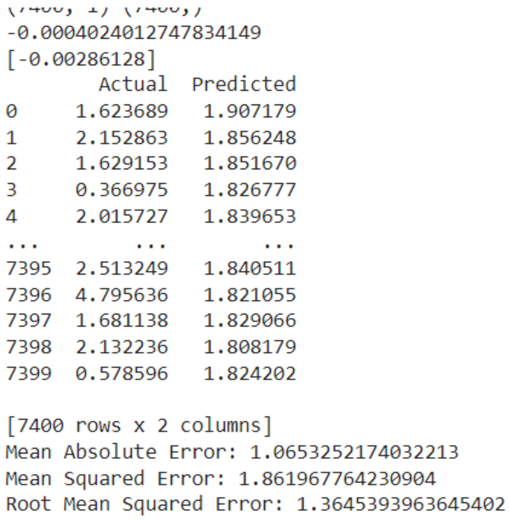


Figure 2. The seasonaliyt of Radon (y-axis in Bq) for the first 400 days of the data series.

3.3. Linear Regression

The predicted values that were modified for the linear regression (Figure 3) showed mostly similar value, meaning the test set should perform well for the majority of values.



```
(7400, 1) (7400,)
-0.0004024012747834149
[-0.00286128]
         Actual  Predicted
0      1.623689   1.907179
1      2.152863   1.856248
2      1.629153   1.851670
3      0.366975   1.826777
4      2.015727   1.839653
...         ...        ...
7395   2.513249   1.840511
7396   4.795636   1.821055
7397   1.681138   1.829066
7398   2.132236   1.808179
7399   0.578596   1.824202

[7400 rows x 2 columns]
Mean Absolute Error: 1.0653252174032213
Mean Squared Error: 1.861967764230904
Root Mean Squared Error: 1.3645393963645402
```

Figure 3. The predicted values and the real ones in comparison.

Furthermore, after applying the linear regression, three statistical errors were calculated to make a statement about the

Table 2. Statistical errors for the linear regression.

Error	Value
Mean absolute Error	1.07
Mean squared Error	1.89
Root Mean square Error	1.38

Table 3. Statistical errors for the Random Forest Regression.

Error	Value
Mean absolute Error	1.06
Mean squared Error	1.85
Root Mean square Error	1.36

preciseness of the linear regression (Table 2).

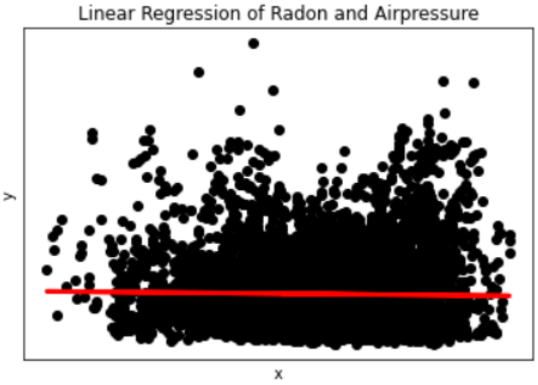


Figure 4. The linear regression of Radon 222 with air pressure. Presented in a scatter plot. Xaxis shows xtest values (Air pressure) and yaxis shows ytest values (Radon).

3.4. Random forest

For the random forest also the errors were calculated (Table 3). Again, x and y train set and predicted values were calculated first. However, it is important to note that no n-estimators, features or maximum depth of the random forest were created. Table 3 shows similar values as for the linear regression (Table 2).

In comparison to the linear regression, the mean absolute error was about 0.01 smaller, the mean square error was 0.04 smaller, the root mean square error was 0.02 smaller. It can be concluded that the errors are slightly smaller with the Random Forest Regressor, but not really significant, but almost identical.

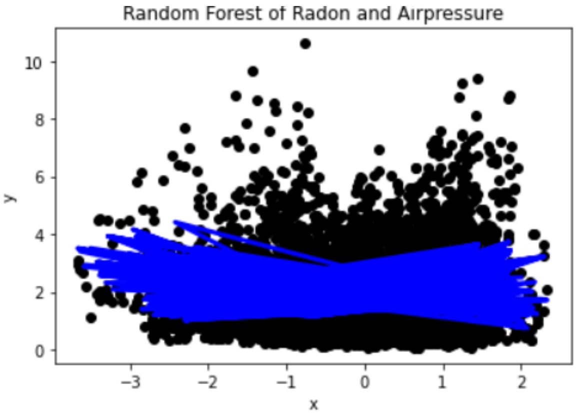


Figure 5. The Random Forest Regression of Radon 222 with air pressure. Presented in a scatter plot. Xaxis shows xtest values (Air pressure) and yaxis shows ytest values (Radon)

Table 4. Statistical errors with modified/ improved Random forest Regression.

Error	Value
Mean squared Error	1.81
Root Mean square Error	1.35

Moreover, the random forest was modified, to see if with the adjustment of depth and n-estimators the results would get better. It was tested with several values (see in the code) and the best ones were carried out with this constellation: nestimators = 300, maxfeatures = 'sqrt', maxdepth = 7, randomstate = 18). Thus, the nestimators were significantly increased which leads to a more accurate analysis. The mean squared error now showed a value of 1.81 (see Table 5) and the root mean square error a value of 1.35. Thus, especially the mean squared error has improved after the modifications, although not by large values.

Graphically, differences between linear regression (Figure 3) and Random Forest Regression (Figure 4) can be seen. Values have the same Scala, the same distribution, but the regression acts different. While the linear regression seems to fit the more usual, average values well, the Random Forest tries to reach also extreme values.

4. Discussion

The scientific founding relates to previous research (Genthon et al., 1995). It can be said that Radon flows faster when there is a bigger difference in pressure between the high-pressure soil and the low pressure air (Vevea et al., 2010). This principle of pressure differences is the main driving force that causes radon levels to change. Moreover, the assumption, that not only a seasonal cycle, but also a daily cycle can be observed, was verified. The reason for this is that in the morning when the inversion is burned off, the near-surface radon is growing in the convective layer (Xie et al., 2013).

Random Forest and linear regression yielded almost the same values. The Random Forest Regressor was minimally better, especially after adjusting its parameters (nestimater, etc.). Talking about the statistical errors, the Mean Squared Error (MSE) shows how close a fitted line is to data points. For every data point, the vertically distance from the point to the corresponding y value on the curve fit (the error), is taken and then and squared. They were here 1.89 (linear regression) and 1.81 (random forest modified). According the root mean squared error (RMSE) it can be said, the lower the RMSE, the better a given model is able to fit a dataset. The mean absolute error is the average difference between the observations (true values) and model output (predictions) that were also shown in Figure 2. As it can be seen, the predictions are not always, but mostly similar to the true values. All statistical errors weren't next to zero, meaning the regression between air pressure and radon is not as big as expected. This can be also verified visually. The plots (Figure 4 and 5) showed connections between the points to the curve, but because the Radon values seems to differ a lot, it is difficult to see a really strong regression. The regression seems to fit the most common values. The sense of the calculation of statistical errors is that it's a way to assess how well a regression model fits a dataset. However, since there are unexpectedly many outlier values here, the regression line connects only the most frequently occurring values. This is were the scientific connection can be made. It has been seen by just plotting the data, that Radon differs according to the seasons. Meaning, the dependence on this is stronger than just the one on air pressure. However, since there are unexpectedly many outlier values here, the regression line connects only the most frequently occurring values. This was somewhat surprising, since actually the air pressure also changes according to season. However, it does not change as much as radon, and since radon changes a lot in different times of the year (sometimes values above 10 Bq, sometimes only 0.5 Bq) and air pressure does not change as much, the regression of the two variables is not as strong as expected. Nevertheless, it is definitely present.

In general, it is believed that Random Forest actually performs more effectively on large data sets. Here, however, the differences between the regressions are very small, and since the data set is based on time, i.e., strictly speaking, on time series, linear regression is more appropriate. This has the advantage that it is much simpler and faster to run compared to the Random Forest. For classifications or more complicated variables, however, it is probably more effective to use the Random Forest. Statistical errors showed satisfactory results and showed how important they are to evaluate the performance of the model. In any case, it can be said that machine learning tools have been successfully tested on the dataset. In climate data, it is not yet commonplace to use machine learning, often because the mechanisms may not be known to the scientists. However, it should be mentioned how well the algorithms have worked on the dataset and that it can be quite helpful for the analysis.

5. Conclusion

Radon 222 definitely shows an annual cycle, most likely even a daily one. Also the influence of air pressure could be shown, but it is less than expected. However, this can also be due to the data, and in other summers and winters it looks somewhat different. Nevertheless, this research shows how to successfully use machine learning tools for climate data and may become of immense importance in the future (climate change...).

6. Acknowledgement

This research work would not have been possible without the data from Franz Conen, University of Basel, who is an expert in radon research. Last but not least, the International Atomic Energy Agency and co-sponsored by the World Meteorological Organization provided interesting results about Radon 222, who helped me to understand the connections.

7. References

Genthon et al (1995) Radon 222 as a comparative tracer of transport and mixing in two general circulation models of the atmosphere.JGR Atmospheres. Vol. 110, Issue D2. 2849-2866. doi/abs/10.1029/94JD02846

Jinzhao et al (2013) The Migration of Radon in Different Air Pressure Experimental Study and the Average Velocity Estimation. Energy Procedia 39. 443 – 453. doi: 10.1016/j.egypro.2013.07.235

Vevea et al (2010) Variation of short-lived beta radionuclide (radon progeny) concentrations and the mixing processes in the atmospheric boundary layer. Journal of Environmental Radioactivity 101. 538–543. doi:10.1016/j.jenvrad.2009.08.008

Xie et al (2013) Radon dispersion modeling and dose assessment for uranium mine ventilation shaft exhausts under neutral atmospheric stability.Journal of Environmental Radioactivity 129. 57e62. http://dx.doi.org/10.1016/j.jenvrad.2013.12.003

The link for the code can be found here: https://github.com/josephinekramer/2022$_ML_Earth_Env_Sci/blob/main/$

Radon$_final_analysis_machine_learning_course_Josephine_{S_{Kramer.ipynb}}$