Environmental Pollution 271 (2021) 116356

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

# Residential radon – Comparative analysis of exposure models in Switzerland☆

Danielle Vienneau [a, b, *], Seçkin Boz [a, b], Lukas Forlin [a, b], Benjamin Flückiger [a, b], Kees de Hoogh [a, b], Claudia Berlin [c], Murielle Bochud [d], Jean-Luc Bulliard [d], Marcel Zwahlen [c], Martin Röösli [a, b]

[a] Swiss Tropical and Public Health Institute, Basel, Switzerland
[b] University of Basel, Switzerland
[c] Institute of Social and Preventive Medicine, Bern, Switzerland
[d] Centre for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

*Article history:*
Received 13 July 2020
Received in revised form
15 December 2020
Accepted 16 December 2020
Available online 23 December 2020

*Keywords:*
Radon
Household
Modelling
Exposure

ABSTRACT

Residential radon exposure is a major public health issue in Switzerland due to the known association between inhaled radon progeny and lung cancer. To confirm recent findings of an association with skin cancer mortality, an updated national radon model is needed. The aim of this study was to derive the best possible residential radon prediction model for subsequent epidemiological analyses. Two different radon prediction models were developed (linear regression model vs. random forest) using ca. 80,000 measurements in the Swiss Radon Database (1994–2017). A range of geographic predictors and building specific predictors were considered in the 3-D models (x,y, floor of dwelling). A five-fold modelling strategy was used to evaluate the robustness of each approach, with models developed (80% measurement locations) and validated (20%) using standard diagnostics. Random forest consistently outperformed the linear regression model, with higher Spearman's rank correlation (51% vs. 36%), validation coefficient of determination ($R^2$ 31% vs. 15%), lower root mean square error (RMSE) and lower fractional bias. Applied to the population of 5.4 million adults in 2000, the random forest resulted in an arithmetic mean (standard deviation) of 75.5 (31.7) Bq/m$^3$, and indicated a respective 16.1% and 0.1% adults with predicted radon concentrations exceeding the World Health Organization (100 Bq/m$^3$) and Swiss (300 Bq/m$^3$) reference values.


## 1. Introduction

Residential radon is the second leading cause of lung cancer, after smoking, accounting for an estimated 3–14% of cases (Darby et al., 2005; WHO, 2009). Based on this known association between inhalation and lung cancer, in particular for smokers (Barros-Dios et al., 2012), radon is recognised as an important public health issue in Switzerland. Previous estimates for the country indicated an excess 230 lung cancer deaths per year due to radon exposure (Menzler et al., 2008).

Radon is a ubiquitous radioactive gas, formed by the decay of uranium that naturally occurs in granitic and metamorphic rocks. In a country like Switzerland, with crystalline and karstic rocks underlying the mountainous regions (Kropat et al., 2015a), radon concentrations in some areas can exceed current reference levels. To generally monitor the population exposure, the Federal Office of Public Health (FOPH) maintains official radon measurements carried out across the country since 1994 in the Swiss Radon Database (Barazza et al., 2018). These measurements of long-term radon gas concentration have been used extensively to support radon risk mapping to detect areas likely to exceed reference levels (FOPH, 2018; Kropat et al., 2015b).

While measurements are also the preferred method to evaluate exposure in many health studies, this is only feasible in smaller populations e.g. Barbosa-Lorenzo et al. (2016), Darby et al. (2005) and Krewski et al. (2006). Alternatively, larger studies such as the American Cancer Society CPS-11 cohort (Turner et al., 2012) have

☆ This paper has been recommended for acceptance by Payam Dadvand.

* Corresponding author. Department of Epidemiology and Public Health Swiss Tropical and Public Health Institute Socinstrasse 57, CH-4051, Basel, Switzerland
*E-mail address:* danielle.vienneau@swisstph.ch (D. Vienneau).

https://doi.org/10.1016/j.envpol.2020.116356

**Abbreviations**

| | |
|---|---|
| EGID | unique building identifier in the GWR |
| GWR | Federal Register of Buildings and Dwellings |
| LM | linear regression model |
| RF | random forest model |
| SEP | socio-economic position |
| SNC | Swiss National Cohort |

either used radon maps, thus an ecological measure of exposure, or as in the Danish studies (Bräuner et al., 2015) regression-based approaches to model radon for households. Thus in addition to monitoring activities, exposure models to estimate individual-level, indoor radon concentrations in homes are needed for epidemiological investigation.

In radon prone areas, exposure to radon progeny mainly occurs indoors after radon infiltrates buildings from the ground through unsealed and soil basements, or via cracks and openings in basement floors and walls. The influencing factors of indoor radon concentrations, with a focus on Switzerland, are detailed in Kropat et al. (2014) and Hauri et al. (2012), and briefly described here. As the main determinant of radon concentration, geology, specifically lithological units, are usually grouped based on similar properties to reduce the number of classes and increase statistical power and interoperability in radon prediction models (Kropat et al., 2015a). Permeability of the surface is largely determined by the soil texture, though geologic faults and elevation can act as indicators. Other important determinants include the building type, characteristics and materials, as these can influence building permeability, as well as floor of residence because concentrations of radon gas are highest in the lower levels (WHO, 2009). Finally, individual behaviours such as home heating and window opening for ventilation also play a role (Groves-Kirkby et al., 2015; Kropat et al., 2015a; Steck et al., 2019).

We previously developed and validated a Swiss-wide residential radon prediction model (Hauri et al., 2012), and subsequently applied it in the Swiss National Cohort (SNC) studies on skin cancer mortality (Vienneau et al., 2017) and childhood cancer risk (Hauri et al., 2013). Research into the health effects of residential radon in Switzerland is ongoing and the existing cohorts have longer follow up periods; as such, there is a need for an updated exposure model considering the extended follow up of the SNC.

The specific aims of this study were to extend the temporal coverage, by incorporating new measurements to coincide with the SNC follow up 2000–2016, and to investigate alternative statistical approaches to generate the best possible spatially resolved radon prediction model for Switzerland. Two modelling approaches were implemented and compared: multiple linear regression (LM), similar to our previous radon prediction model (Hauri et al., 2012), vs. the machine learning random forest (RF) model. Models from both approaches were applied to the 5.4 million adults in the SNC (i.e. over 20 years old in the year 2000) to gain insight into concordance between the estimates. The overarching goal was to select the best model to take forward to subsequent epidemiological analyses.

## 2. Materials and methods

To predict indoor radon concentrations for all residential dwellings in Switzerland, using both a linear regression and random forest approach, we combined information from the census-based SNC, measurements from the extensive Swiss Radon Database and ancillary data including e.g. the building registry and environmental data. Specific inputs from each of these sources are described below.

### 2.1. Swiss National Cohort

The SNC links the Swiss census with data on births, mortality and emigration (Bopp et al., 2009; Spoerri et al., 2010), and with the recently introduced Registry Based Census and annual structural surveys from 2010 onward. The overall objective in developing a new radon prediction model for Switzerland was to derive exposure estimates for the extended SNC follow-up.

Due to compulsory participation, nearly all persons residing in Switzerland at the time of the census are represented, i.e. 98.6% of the 7.3 million inhabitants in 2000 (Renaud, 2004). The SNC includes information for individuals (e.g. age, sex), households (e.g. floor of dwelling) and buildings (e.g. x,y coordinates, period of construction, building type). The SNC was approved by the Ethics Committees of the Cantons of Zurich and Bern.

### 2.2. Radon database

#### 2.2.1. Description

The Federal Office for Public Health is responsible for radon protection as mandated by the Radiation Protection Ordinance (https://www.admin.ch/opc/en/classified-compilation/20163016/index.html). This includes approving radon measurement providers, and maintaining the central database that holds these official measurements obtained through accredited laboratories. The Swiss Radon Database, covering the period 1994 to 2017, was obtained from the FOPH. The database is briefly explained here (Barazza et al., 2018).

In total 235,585 measurements were available, collected in buildings across the country. The database includes location-specific information: community (from which we determined the larger Swiss canton); x,y coordinate for the building and for most measurements the unique building identifier (EGID); floor on which the measurement was taken (categorical “floor of dwelling”: basement, ground floor [reference], first floor, second floor, third floor and above); and room type (categorical: living room [reference], study, dining room, child’s room, kitchen, bedroom). It also contains measurement specific information including: dosimeter type (entered into the models as unique categories to control for known differences between devices (Kropat et al., 2014): Gamma-data [reference], Altrac, AT-100, Elektret, Miam and Radtrak); measurement period via a start and end date; if radon remediation was performed (categorical: no [reference], yes); and quality indicators including measurement error.

Prior to 2005, radon sampling was conducted across the whole country via random selection of buildings within communities. The strategy changed in 2005, such that radon monitoring focussed in radon prone areas (FOPH, 2011; Kropat et al., 2014). For this reason, an indicator for measurement epoch (categorical: before January 01, 2005 [reference], after January 01, 2005) was created. Additionally, since the updated ordinance came into force on January 01, 2018 only measurements with a duration of at least 90 days during the heating season are approved and included in the Swiss Radon Database. Data collected prior to 2018, however, remain valid if the measurement duration was at least one month (personal communication: F. Barazza, FOPH, October 2020).

### 2.3. Data cleaning

Data cleaning was conducted according to the scheme shown in Fig. 1 and elaborated in Figure S1. As the aim was to model exposure

in residential rooms where people spend a substantial amount of time, and to enable prediction at home locations requiring precision in geocoding, many measurements did not meet inclusion criteria. The first "Quality cleaning" step excluded measurements: in rooms designated as uninhabited (i.e. not heated), in non-residential buildings, or in room types with low occupancy rates (i.e. bathrooms, hallways, hobby rooms, cellars, other) (46%). It also excluded measurements of potentially less comparable quality i.e. taken using less common dosimeters (i.e. <1000 occurrences in the whole database, not validated by authorities, or those considered unreliable due to excessively short or long measurement duration (i.e. outside the range of 28–180 days) (8.7%). Next, the "Coordinate cleaning" step removed measurements with obvious invalid x,y coordinates and those falling outside Switzerland (7.0%). The last "GIS cleaning" step used key predictors from our Swiss-wide GIS (described below) and the Federal Register of Buildings and Dwellings (GWR) (FSO, 2018) to identify any remaining measurements that could not be used in a predictive model. This included measurements with recorded elevation above that of the highest Swiss village (>1800 m), those not linkable to the building registry or building footprints, or those falling within waterbodies (2.1%). In total, 85,473 (36.3%) measurements were retained after the three-step data cleaning process. This data retention rate is in line with other studies using these data (Kropat et al., 2015a, 2017).

A small proportion of dwellings had multiple measurements within the same room. In brief, replicates were separated from the main pool of measurements, and then through random selection one was assigned back into the main pool (n = 3267) and the remaining measurements were discarded (Fig. 1). Finally, the main pool of 79,598 measurements were randomly partitioned (by canton, epoch, dosimeter type, radon remediation and floor of dwelling) into 20% subsets from which five unique 80:20 datasets were created (referred to as set 1–5). This was done to facilitate a five-fold modelling strategy to evaluate the robustness of each approach, with each 80% subset used for training (i.e. model building) and the 20% reserved subset used for validation.

### 2.4. Ancillary data

Information on additional predictors not available in the Swiss Radon Database were obtained from national databases. Building data were from the above mentioned GWR, maintained by the

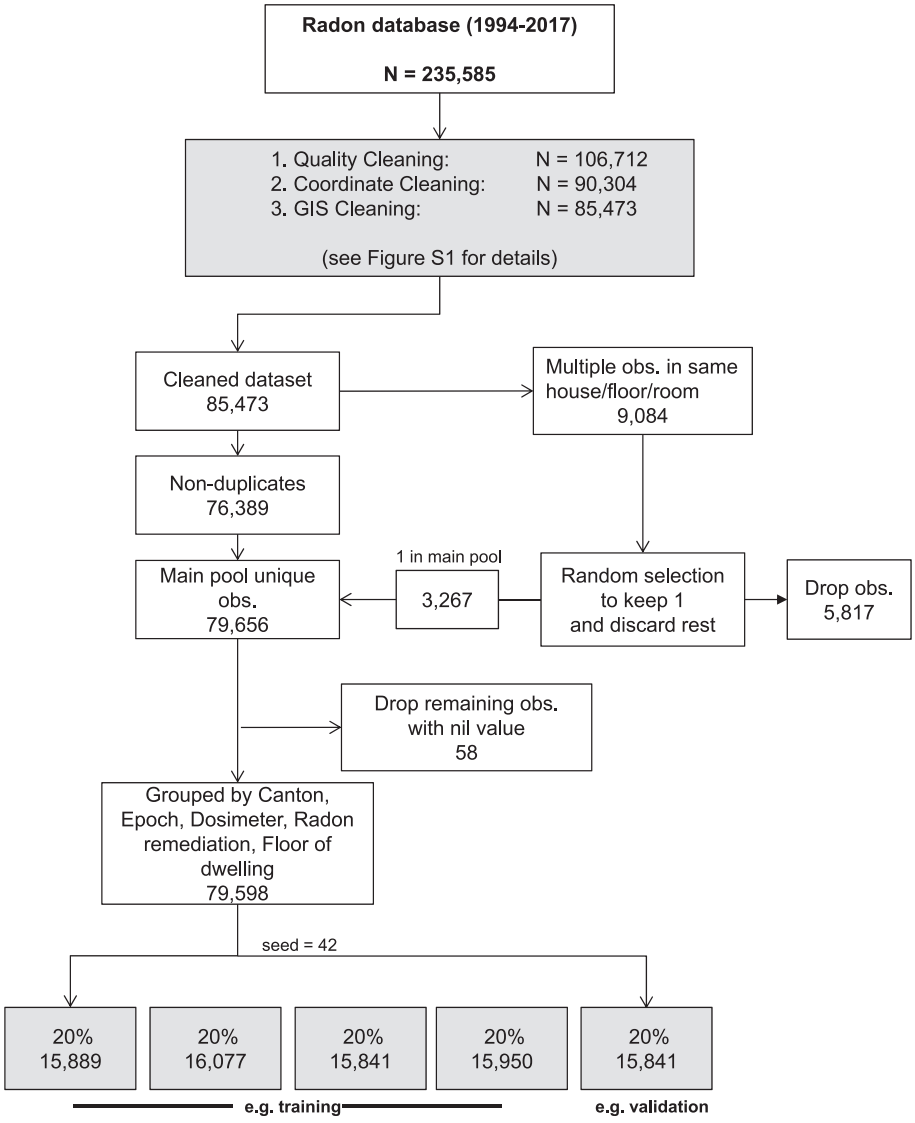


**Fig. 1.** Radon data cleaning scheme.

Federal Statistical Office. Period during which the building was constructed (recoded to categories "construction period": before 1919 [reference], 1919–1945, 1946–1960, then each decade [e.g. 1961–1970] until 2017); building type (categorical: single family home [reference], apartment, farmhouse, other, unknown); and total number of floors in the building were extracted. Basic information about the buildings were also available in the Swiss Radon Database, but was considered less reliable. We therefore linked the GWR data to the measurements by the unique building identifier EGID. In cases where EGID was not known, the x,y coordinates for the measurement location were used to link to the nearest building in the GWR (ArcGIS 10.6 Near command, with a 50 m threshold). As mentioned above, measurements with obvious mistakes in the x,y coordinates were removed during the "Coordinate cleaning" stage (Figure S1).

The other potential predictors derived from available GIS databases. Soil texture was reclassified based on the 1:1,000,000 European Soil Database (EC, 2004; Panagos, 2006) (categorical: medium grained [reference], fine grained, coarse grained, other (organic) soil type, unknown). For lithology, we used the reclassified lithology derived by Kropat et al. (2015a) based on the 1:500,000 geology map for Switzerland (Swisstopo, 2005) (categorical: carbonate rock alps [reference], sediments, carbonate rock jura, igneous rock, metamorphic rocks, sedimentary rocks excluding carbonates, and others). Additionally, the original 1:500,000 geology map was used to calculate distance to the nearest geological fault line (categorical "fault distance": 0–100 m [reference], 100–500 m, 500–1000 m, >1000 m), and determine classes for depth to aquifer (categorical "Groundwater": Productive, changeable or marginal useable (reference), see Table S1 for others) and hydrogeology (categorical: Productive groundwater partly outside valley plain [reference], see Table S1 for others). Continuous elevation above sea level was taken from the 25 m digital elevation model for Switzerland (Swisstopo, 2004), and community-level urbanisation data was obtained from the Federal Statistical Office (FSO, 2012) (categorical "area type": urban [reference], peri-urban, rural). Finally, continuous terrestrial radiation levels (in nSvh) were extracted from the Swiss radiation map with a spatial resolution of $2 \times 2$ km (ENSI; Rybach et al., 2002). This terrestrial radiation layer represents natural radiation, mainly composed of radionuclides of uranium and thorium decay chains, and excludes cosmic and artificial radiation.

### 2.5. Exposure modelling

Both a linear regression (LM) and random forest (RF) approach were tested in developing the residential radon prediction models. Modelling was done following a five-fold strategy (i.e. 80% subset for model building and 20% for validation) to evaluate the robustness of each approach. Radon concentrations were ln-transformed prior to modelling. Start season was offered to the models as a continuous variable. It was calculated as a cosine function (with a value of 0 on 1st January and 1 on 1st July) on the basis of the start date (Julian day) for each measurement according to Equation (1) (Röösli et al., 2006). It was thus used to control for season in which the measurement began.

The linear regression (LM) model was developed following a similar approach as used in our previous radon prediction model (Hauri et al., 2012). Important potential predictors were first identified from the literature. Each was tested in univariate analyses, and pruned if the *p-value* > 0.05. We also tested potential interactions (e.g. between building construction period and floor of dwelling), and non-linear associations for altitude and terrestrial radiation modelled as natural and b-splines with 3, 4 and 5 degrees of freedom (df). Next a multivariable regression model including all remaining potential predictors was developed, and those predictors with a *p-value* > 0.05 were subsequently removed. Finally, supervised forward selection was performed to define the final model by including the variables, in turn. We used the Akaike Information Criterion (AIC) to decide the final model. AIC was evaluated at each step and predictors only retained if the AIC decreased by five. This stepwise model was also used to determine relative variable importance *post hoc* based on incremental increase in adjusted $R^2$.

The random forest (RF) was selected as an alternative approach, as used in similar recent studies (Kropat et al., 2015a; Nikkilä et al., 2020). RF is a machine learning technique based on decision trees. In brief, a large number of trees are generated in an ensemble, with each tree developed from an independent yet identically distributed subset of the data. In a RF for regression, as applied here, the average from all individual trees determines the final prediction (Breiman, 2001). The same predictors as in the final LM model were used, and factors were handled as unordered covariates. The model was run with a maximum of 500 trees, and up to 4 variables per split at each node according to the square root of the maximum number of variables. Variable importance was recorded from the models. This is an internal estimate from the RF, based on the reduction in sum of squared errors after a random permutation.

Moran's I was used to evaluate spatial autocorrelation in the residuals calculated for the training datasets, with a *p-value* < 0.05 indicating spatial autocorrelation. Each derived model was applied to the respective, independent validation dataset in the five-fold strategy. Model evaluation included calculating $R^2$ via linear regression of the modelled vs. measured ln-transformed radon concentrations, the root mean square error (RMSE) and fractional bias. Fractional bias is a measure of agreement of the mean observed and predicted values; a value of 0.05, for example, represents an over-prediction of 5% (see Vienneau et al. (2009) for equations). Predictions for the five independent validation sets were also aggregated in order to evaluate concordance, by quintiles, using the weighted Kappa statistic.

Finally, models were applied to all adults in the SNC, using their residential address for year 2000 (x,y coordinates and by floor of dwelling). Elevation at residential addresses was capped at 1800m, which is the elevation of the highest Swiss village. The range of values for terrestrial radiation at residential addresses vs. the measurement locations were the same, thus capping of the values at the residential addresses was not needed. Final predictions were computed by averaging predictions of the five respective sets for each approach. Concordance between the predictions from the two approaches was determined using the Spearman rank correlation, as well as the weighted Kappa statistic based on quintiles.

$$start\ season = \frac{\cos\left(\mathrm{pi}*2*\left(\frac{\text{Day of Year}}{366} - \left(\mathrm{floor}\left(\frac{\text{Day of Year}}{366}\right)\right) - 0.5\right)\right) + 1}{2} \quad (1)$$

Analyses were conducted in R version 3.5.2 using the following main packages: linear regression (stats), random forest (ranger (Wright and Ziegler, 2017)), GIS functionality (sf and raster). ArcGIS 10.6 was used for GIS database management and mapping.

## 3. Results

A detailed description of the measured radon concentrations for key predictor variables is presented in Table S1. Median measured radon concentrations across Switzerland taken after 2005, when the strategy focussed more on high risk areas, were higher than before 2005 (96.0 vs. 88.3 Bq/m3). Figure S2 illustrates the broad spatial coverage of monitored locations across the populated areas of Switzerland. Median measured radon concentrations were higher in single family homes (95.1 Bq/m$^3$) compared to apartments (85.1 Bq/m$^3$) and, as illustrated in Figure S3, concentrations decreased with increasing floor of dwelling (from median 121.1 Bq/m$^3$ in the inhabited basements to 57.6 Bq/m$^3$ on the third floor and above). Median measured concentrations were also lower in newer compared to older buildings (~77–83 Bq/m$^3$ for construction after 1971 vs. ~104–107 Bq/m$^3$ for construction before 1960). This may be related to the construction of both higher and better sealed buildings in the more recent years. Measurements across the country were obtained using a selection of comparable passive dosimeters, regularly tested in accordance with an ordinance for measuring devices for ionizing radiation. To pass the testing, the standard deviation of several dosimeters of the same type and the deviation from the reference value have to be within ± 20%. An analysis of variance (ANOVA p-value <0.001) and post-hoc TukeyHSD test on the measurements included here indicated a difference between means for some combinations (see boxplots in Figure S4). There was also some variation in terms of measurement duration (median 95.0 days) and measurement error reported by the laboratory (median 15.0 Bq/m3) (Figure S5).

The composition of the LM is presented in Table S2. The largest contribution to predicted radon concentration is from the intercept; after exponentiation, this equated to ~100–105 Bq/m$^3$. All other variables contribute <3 Bq/m$^3$ to the prediction. Most of the retained predictors were as expected based on the literature, and included floor of dwelling, building construction period, lithology (geological class), groundwater productivity, elevation and soil texture. In univariate analyses, natural spines (5 df) for altitude and terrestrial radiation showed slightly stronger associations than the linear terms; however, the sensitivity analysis replacing the linear with non-linear terms, and including the significant interaction term for building construction period and floor of dwelling, did very little to improve the LM (Table S3, set 1 sensitivity). The simpler model with linear terms was thus used for all subsequent comparisons. Table 1 shows the ranking of variables by importance in the derived RF model (Figure S6 shows the five sets). According to the RF the most influential predictors were the continuous variables elevation followed by start season and terrestrial radiation. The table also shows the relative order of variable importance in the LM, based on incremental increase in adjusted $R^2$. The continuous variables were ranked much lower than in the RF, suggesting some non-linearity in these variables that, while not sufficiently captured by splines with 5 df, was better modelled by the flexibility of the RF. Following the continuous variables, Table 1 shows the next highest ranked variables in the RF were building construction period and canton. These variables ranked the two highest in the LM. Similar rankings between the two approaches, through further down the list, were also obtained for groundwater productivity and building type. Interestingly, the categorical variable for floor of dwelling ranked fourth for the LM but only tenth for the RF.

**Table 1**
Random forest (RF) variable importance in model building ranked highest to lowest, compared to order variables entered linear model (LM).

| Variable | RF | | LM |
|---|---|---|---|
| | Variable importance[a] | Rank | Rank |
| Elevation | 10153 | 1 | 10 |
| Start season | 6396 | 2 | 11 |
| Terrestrial radiation | 5173 | 3 | 8 |
| Construction period | 3715 | 4 | 2 |
| Canton | 3322 | 5 | 1 |
| Groundwater | 2199 | 6 | 5 |
| Building type | 1948 | 7 | 6 |
| Lithology | 1833 | 8 | 3 |
| Fault distance | 1628 | 9 | 13 |
| Floor of dwelling | 1523 | 10 | 4 |
| Area type | 1172 | 11 | 12 |
| Soil texture | 1076 | 12 | 7 |
| Measurement epoch | 853 | 13 | 9 |

[a] Variable importance is the increase in node purity (IncNodePurity). The value shown is here is derived by averaging variable importance from set 1-5.

The model performance and validation statistics for the LM and RF, as well as modelled radon levels at the measurement locations, are presented in Table 2 for set 1 (Table S3 shows all sets with only marginal variations across sets). Overall the RF consistently outperformed the LM, with higher validation $R^2$ (0.31 vs. 0.15), lower RMSE and lower fractional bias. Modelled radon concentrations from both models produced almost identical geometric means, while the RF produced a slightly lower median and wider interquartile range (Table 2). Diagnostics for model training and validation are shown in Figures S7-S8. Over prediction in the higher range and under prediction in the lower range is less severe in the RF. Predictions were made for each of the 5 independent validation sets, and merged to produce one complete validation dataset of all measurement locations (n = 79,598). The agreement between measured vs. modelled radon concentration, by quintile, was fair for both approaches. Spearman's rank correlation across all measurement locations was higher for the RF (0.51) than the LM (0.36) approach (Figure S9). Finally, Moran's I on the residuals from model building revealed spatial autocorrelation in the LM but not the RF (Figure S10).

Both approaches were applied at the dwellings of the 5.4 million adults in Switzerland at year 2000 to predict radon concentrations. Details on selection of the SNC study population for which radon exposure was assigned, indicating why some individuals were excluded, are outlined in Table S4. Final predicted radon concentrations for the cohort were calculated by averaging the predictions from the five LM sets and five RF sets. For these final cohort predictions, the arithmetic mean predicted radon concentrations were approximately 5 Bq/m$^3$ lower for the LM (69.8 Bq/m$^3$) compared to the RF (75.5 Bq/m$^3$), a pattern that held across most of the exposure distribution (Table 3). The histogram in Figure S11 comparing these final predictions highlights the paucity of predictions in the lowest range, and long right-tail in the distribution for the RF predictions. The RF estimated a respective 16.1% and 0.1% adults living in houses with predicted radon concentrations exceeding the WHO (100 Bq/m$^3$) and Swiss (300 Bq/m$^3$) reference values for radon concentrations in buildings; for the LM the proportions were 12.1% and 0.004%. Despite these differences, the overall agreement between the predictions for the cohort from the two approaches was high with a weighted Kappa of 0.77 and a Spearman's rank correlation of 0.84 (Figure S12). The spatial patterns of the individual-level radon predictions aggregated to community-level were mapped, and broadly indicated similar patterns for both approaches. Areas with the highest predicted radon concentrations tended to be in the mountainous regions of the Alps and Jura (Fig. 2).

**Table 2**
Model performance and validation, comparison of linear model (LM) vs. random forest (RF) for set 1.

| Set | Model | N | $R^2$ | RMSE (ln Bq/m$^3$) | Fractional Bias (ln Bq/m$^3$) | Modelled radon concentrations (Bq/m$^3$) | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Geo Mean | Median | IQR |
| **Linear model (LM)** | | | | | | | | |
| Set1 | Training | 63,757 | 0.14 | 0.81 | 0.000 | 101.10 | 100.60 | 81.4–126.3 |
| | Validation | 15,841 | 0.15 | 0.82 | 0.000 | 101.29 | 100.74 | 81.0–127.0 |
| **Random forest (RF)** | | | | | | | | |
| Set1 | Training | 63,757 | 0.30 | 0.73 | 0.000 | 101.25 | 95.82 | 72.8–132.8 |
| | Validation | 15,841 | 0.31 | 0.74 | −0.001 | 101.52 | 96.08 | 72.8–133.8 |

Geo Mean = geometric mean; IQR = interquartile range.

**Table 3**
Distribution of predicted radon concentrations for the 5.4 million adults in the SNC (prediction at home address at year 2000).

| Set | Predicted radon levels (Bq/m3) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Geo Mean | Mean | SD | P5 | P25 | P50 | P75 | P95 |
| Linear model (LM) | | | | | | | | |
| Set1 | 64.89 | 69.72 | 27.47 | 34.04 | 50.88 | 65.16 | 83.43 | 119.40 |
| Set2 | 64.63 | 69.71 | 28.04 | 33.15 | 50.30 | 65.43 | 83.85 | 120.59 |
| Set3 | 64.47 | 69.45 | 27.83 | 33.13 | 50.41 | 64.94 | 83.34 | 119.64 |
| Set4 | 65.17 | 70.32 | 28.48 | 33.24 | 50.77 | 65.60 | 84.51 | 122.15 |
| Set5 | 64.66 | 69.73 | 28.08 | 33.06 | 50.39 | 65.33 | 83.81 | 120.62 |
| Final LM[a] | 64.78 | 69.79 | 27.94 | 33.32 | 50.58 | 65.32 | 83.80 | 120.39 |
| **Random Forest (RF)** | | | | | | | | |
| Set1 | 71.58 | 76.38 | 31.24 | 40.16 | 57.29 | 70.57 | 87.88 | 129.91 |
| Set2 | 68.38 | 73.67 | 32.33 | 38.02 | 52.93 | 67.03 | 85.98 | 130.87 |
| Set3 | 70.89 | 75.94 | 32.30 | 41.30 | 55.54 | 68.41 | 88.61 | 131.93 |
| Set4 | 70.92 | 76.16 | 33.06 | 40.34 | 54.98 | 69.06 | 88.59 | 133.88 |
| Set5 | 69.98 | 75.30 | 32.73 | 39.22 | 53.96 | 68.40 | 88.14 | 132.68 |
| Final RF[a] | 70.53 | 75.49 | 31.74 | 40.62 | 55.18 | 68.91 | 87.69 | 130.91 |

Geo Mean = geometric mean, Mean = arithmetic mean, P = percentile.
[a] Final predictions derived by averaging predictions from set 1-5.

## 4. Discussion

We explored two algorithms for modelling indoor residential radon concentrations in Switzerland, with random forest, a machine learning approach, substantially outperforming the more standard multivariable linear regression (model validation $R^2$ = 0.31 vs. 0.15; Spearman's rank correlation = 0.51 vs. 0.36). Models were developed using the Swiss Radon Database including measurements collected over more than 20 years in a variety of building types and locations across the country. Selecting only those measurements from occupied rooms in residential dwellings, to reflect our aim to build a model for long-term average exposure in the household, provided ca. 80,000 available measurements. This also allowed for implementation of a five-fold modelling strategy, demonstrating model robustness.

One clear advantage of the RF over the LM is that no assumption needs to be made about the data distribution and dependencies. Tree based learning methods have also been shown useful for modelling complex relationships and interactions between variables, leading to strong predictive performance (Breiman, 2001; De'ath, 2007). We tested non-linear continuous terms and several interactions in our LM and found no meaningful improvement in model performance. Thus, in our evaluation the RF proved better suited to capture the complex structures within the radon measurement data. While the performance statistics generally remain rather low – even with the RF – and a large amount of uncertainty remained, they are in line with the existing literature for similar models (Hauri et al., 2012; Kropat et al., 2015a, 2015b; Nikkilä et al., 2020). As in many applications, exposure measurement errors are unavoidable, but in our modelling the prediction error is expected to be mainly Berkson type error. This occurs when a group's average is assigned to individuals of the group (Armstrong, 1998; Heid et al., 2004), but assignment to a group (e.g. type or floor of dwelling in which an individual resides) is not influenced by the error. In general, Berkson type error does not bias the regression coefficients (Hauri et al., 2012).

To compare the structure of the models from the two approaches is not possible. Still the variable rankings were clearly different (Table 1), with the RF assigning more statistical importance to the continuous vs. categorical variables despite the known importance of some of the latter. The RF flexibly models non-linear relationships and interactions, and then presumably gives higher weight to these variables. Assuming a linear or simple function, the LM rank is perhaps the better way to evaluate the relevance of any single variable. Thus, following the LM ranking, the canton, construction period, lithology and floor of dwelling remain important predictors. Elevation was also found to be significant in prior Swiss models, as was some form of seasonality (Hauri et al., 2012; Kropat et al., 2014). Measurement start period was used in Hauri et al. (2012), while ambient temperature was used in the studies by Kropat (Kropat 2014, 2015a; 2015b). As highlighted in a recent Canadian study, the relevance of season may be context specific, further challenging the use of applying a seasonal correction to evaluate an annual average (Stanley et al., 2019). We *a priori* removed the temporal correction similar to Kropat et al. (2014). Though not substantial, we did observe some difference between measurements by season (spring vs. winter). A recent study in Switzerland also examined indoor radon after energy retrofit in 154 buildings. They found replacement of windows, in particular, significantly increased indoor radon concentrations. Assuming individuals modulate their ventilation practices over the year, we might expect some influence of season.

A recent study in Finland used the more classical log-linear regression and compared it to RF and deep neural networks to produce prediction models for residential radon separately for houses and apartments (Nikkilä et al., 2020). The latter approaches were considered exploratory. Similar to our findings, the fit from their random forest was best ($R^2$ 0.28 and 0.23 for houses and apartments, respectively). Their RF compared to their classical model, however, resulted in a modest improvement in $R^2$ of 3–7% as opposed to the 16% (31% vs. 15%) we found.

Modelling of indoor radon concentrations and/or probability of exceeding reference values remains an ongoing research and public health activity in Switzerland, providing an opportunity to more directly compare our latest RF and LM to previous country-wide models. Kropat et al. (2014) used a basic geostatistical approach to map local probability of exceeding 300 Bq/m$^3$, the current Swiss reference level, at a 1 km resolution. The same authors applied kernel regression, an approach that explicitly accounts for spatial relationships between measurements, to similar input data. Based on five-fold cross validation, the kernel regression model explained 28% of the variation in measured concentrations; however, it was found to respectively over and underestimate the low and high

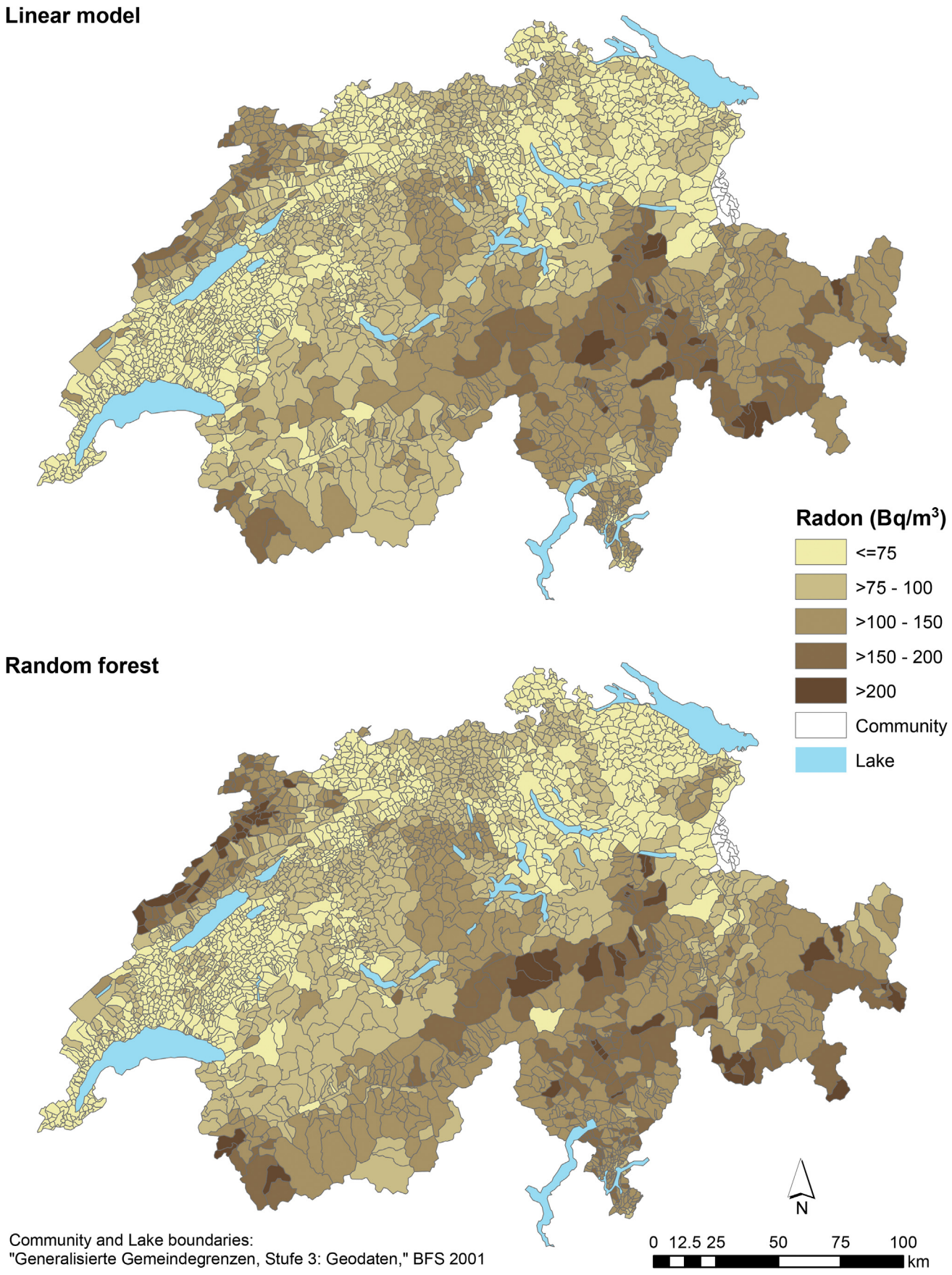


**Fig. 2.** Predicted arithmetic mean radon concentrations (Bq/m$^3$) at community using the linear model (top) and random forest (bottom).

concentrations (Kropat et al., 2015b). These same authors also developed a RF and Bayesian additive regression trees (BART), which explained a respective 33% and 29% of the variation in measured concentrations in five-fold cross validation (Kropat et al., 2015a).

Our group's own previous attempts included a radon prediction model by Hauri et al. (2012) which used a similar multivariable log-linear regression model that performed slightly better than the LM presented here. It included many of the same predictors, specifically geology (tectonic unit), soil texture, degree of urbanisation, dwelling type, year of construction of the building, and floor of the dwelling. It was determined to be robust (Spearman's rank correlation was 0.45 [95% CI: 0.44–0.46) for model development, and 0.44 [0.42–0.46] in an independent validation dataset), with a validation $R^2$ of 0.19. Since this previous LM radon prediction model, the years of data collection and number of suitable radon measurements to support modelling of residential radon has doubled in Switzerland. This large increase in measurements may partly explain our small reduction in $R^2$ and Spearman's rank correlation compared to Hauri et al. (2012). Though not of the same magnitude, examples from air pollution modelling have shown that regression models based on a larger number of measured values give lower $R^2$ but more robust models than those using less (Basagaña et al., 2012; Wang et al., 2013).

Perhaps more relevant is a recent field study from Iowa USA measuring radon in 76 rooms of 38 houses that found modest correlation of long-term radon measurements taken in different seasons ($R^2 = 0.30$), partly explained by seasonal differences in occupant behaviour (windows closed all winter vs. kept open all summer leading to substantially different radon concentrations) (Steck et al., 2019). Their work also supports the notion that multiple locations and time-integrated measurements in homes are necessary for a good and accurate evaluation of long-term, radon-related dose. Residential radon concentrations are influenced by a variety of factors, from geology to personal ventilation practices, and the indoor concentrations can differ substantially from home to home. As highlighted in the Steck et al. (2019) field study, this makes radon exposure particularly challenging to model and emphasizes that achieving high performing models is not a good benchmark to evaluate success. This is why we put more emphasis on the robustness of the models, demonstrated through the highly consistent results across all analyses using a five-fold validation strategy. Given that many of the predictors were ecological in nature, and only broad control for canton was included, mainly to account for differences in measurement strategy, we may have expected residual spatial autocorrelation in both approaches. Interestingly, however, we found no evidence of residual spatial autocorrelation in the RF models giving more support for selecting the RF as our preferred model.

A major limitation of our study – and contributing factor to the unexplained uncertainty – is the lack of specific information in national registries to model the known variations in indoor radon concentration, in addition to the lack of individual behaviours and ventilation habits that can influence radon levels in the home. Unlike the Finnish study, the building registry did not include details such as availability or use of mechanical ventilation, type of building material or physical size (i.e. area and volume) of individual dwellings (Nikkilä et al., 2020). Despite the additional predictors, however, the Finnish models had very similar performance to ours. Information on type of foundation would have also been useful, as not all buildings in Switzerland have sealed basements. A survey on radon remediation indicated that 25–30% of remedial work relates to installing concrete floors or other measures to tighten the interface between basements and living spaces (Barazza et al., 2018). Note that we also used shorter measurement durations than three months recommended by the WHO (WHO, 2009). The median duration was 95 days; however ~20% of the included measurements over our study period of 1994–2017 only had a duration of one to three months, which was in line with the Swiss guidelines at that time. This may result in imprecise annual estimates for a given location and thus increase the random error in our model. In our statistical modelling, however, they would not result in a bias towards over or underestimation of the radon levels and still be more informative than absence of any data.

Applied to the home addresses of all adults in Switzerland for the census year 2000 (often used as the baseline year in SNC epidemiological analyses), we estimated the arithmetic mean (median) population radon exposure to be 75.5 (68.9) Bq/m$^3$. Similar to the previous Swiss-wide residential radon prediction model (Hauri et al., 2012), predicted radon concentrations exceeded the current Swiss reference level of 300 Bq/m$^3$ for only a small proportion (0.1%) of adults. Though less than the ~26% estimated by Hauri et al. (2012), our results also suggest that predicted radon concentrations exceeded the WHO reference level of 100 Bq/m$^3$ for a considerable 16.1% of adults.

## 5. Conclusions

Random forest is an attractive approach for modelling residential radon exposure. This analysis illustrated the gains in performance of RF compared to the more classical LM, and the advantages of machine learning to more flexibly capture complex structures within the measurement data. The RF produced highly robust predictions at the validation locations and when estimating exposure at unmeasured locations. Still, a large amount of uncertainty remains due to the general challenges in modelling radon without detailed information on dwellings and personal behaviours.

## Authors' contributions

DV, MR study concept; DV, MR study design; DV, BF, LF, MR model development; DV, BE, SB model prediction; DV, MR. KdH, MB, MZ, JB data interpretation; DV write and revise manuscript; all review and comment on manuscript.

## Funding

This work was supported by the Swiss National Science Foundation (grant nos. 3347CO-108806, 33CS30_134273 and 33CS30_148415) and Swiss Oncology (grant no. KFS-4116-02-2017).

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We thank Martha Palacios, Fabio Barazza, Christophe Murith and Daniel Storch at the Federal Office for Public Health for providing the Swiss Radon Database and expert advice. We also thank the Swiss Federal Statistical Office for providing mortality and census data and for the support which made the Swiss National Cohort and this study possible. We also acknowledge the members of the Swiss National Cohort Study Group: Matthias Egger (Chairman of the Executive Board), Adrian Spoerri and Marcel Zwahlen (all Bern), Milo Puhan (Chairman of the Scientific Board),

Matthias Bopp (both Zurich), Martin Röösli (Basel), Murielle Bochud (Lausanne) and Michel Oris (Geneva).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2020.116356.

## References


Armstrong, B.G., 1998. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup. Environ. Med. 55, 651–656.

Barazza, F., et al., 2018. A national survey on radon remediation in Switzerland. J. Radiol. Prot. 38, 25.

Barbosa-Lorenzo, R., et al., 2016. Residential radon and cancers other than lung cancer: a cohort study in Galicia, a Spanish radon-prone area. Eur. J. Epidemiol. 31, 437–441.

Barros-Dios, J.M., et al., 2012. Residential radon exposure, histologic types, and lung cancer risk. A case–control study in galicia, Spain. Cancer Epidemiology Biomarkers & Prevention 21, 951.

Basagaña, X., et al., 2012. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmos. Environ. 54, 634–642.

Bopp, M., et al., 2009. Cohort Profile: the Swiss National Cohort—a longitudinal study of 6.8 million people. Int. J. Epidemiol. 38, 379–384.

Bräuner, E.V., et al., 2015. Residential radon exposure and skin cancer incidence in a prospective Danish cohort. PloS One 10, e0135642.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Darby, S., et al., 2005. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ 330, 223.

De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88, 243–251.

EC, 2004. ESDB v2.0: the European Soil Database. European Commission and the European Soil Bureau Network.

ENSI, Swiss Radiation Map. Confederate Nuclear Safety Inspectorate (ENSI), Swiss Geophysical Commission (SGPK), Institute for Geophysics, ETH Zürich.

FOPH, 2011. National Action Plan Concerning Radon 2012-2020. Federal Office of Public Health.

FOPH, 2018. Radon Map of Switzerland. Federal Office of Public Health, Division of Radiological, Radiological Risk Section.

FSO, 2012. Stadt/Land-Typologie. Federal Statistical Office.

FSO, 2018. Federal Register of Buildings and Dwellings. Federal Statistical Office.

Groves-Kirkby, C.J., et al., 2015. A critical analysis of climatic influences on indoor radon concentrations: implications for seasonal correction. J. Environ. Radioact. 148, 16–26.

Hauri, D., et al., 2013. Domestic radon exposure and risk of childhood cancer: a prospective census-based cohort study. Environ. Health Perspect. 121, 1239–1244.

Hauri, D.D., et al., 2012. A prediction model for assessing residential radon concentration in Switzerland. J. Environ. Radioact. 112, 83–89.

Heid, I.M., et al., 2004. Two dimensions of measurement error: classical and Berkson error in residential radon exposure assessment. J. Expo. Sci. Environ. Epidemiol. 14, 365–377.

Krewski, D., et al., 2006. A combined analysis of North American case-control studies of residential radon and lung cancer. J. Toxicol. Environ. Health 69, 533–597.

Kropat, G., et al., 2014. Major influencing factors of indoor radon concentrations in Switzerland. J. Environ. Radioact. 129, 7–22.

Kropat, G., et al., 2015a. Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. J. Environ. Radioact. 147, 51–62.

Kropat, G., et al., 2015b. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: an application to Switzerland. Sci. Total Environ. 505, 137–148.

Kropat, G., et al., 2017. Modeling of geogenic radon in Switzerland based on ordered logistic regression. J. Environ. Radioact. 166, 376–381.

Menzler, S., et al., 2008. Population attributable fraction for lung cancer due to residential radon in Switzerland and Germany. Health Phys. 95, 179–189.

Nikkilä, A., et al., 2020. Predicting residential radon concentrations in Finland: model development, validation, and application to childhood leukemia. Scand. J. Work. Environ. Health 46, 278–292.

Panagos, P., 2006. The European soils database. GEOconnexion International Magazine 5, 32–33.

Renaud, A., 2004. Coverage Estimation for the Swiss Population Census 2000: Estimation Methodology and Results. Swiss Statistics Methodology Report. Swiss Federal Statistical Office, Neuchâtel, p. 147.

Röösli, M., et al., 2006. Sleepless night, the moon is bright: longitudinal study of lunar phase and sleep. J. Sleep Res. 15, 149–153.

Rybach, L., et al., 2002. Radiation doses of Swiss population from external sources. J. Environ. Radioact. 62, 277–286.

Spoerri, A., et al., 2010. The Swiss National Cohort: a unique database for national and international researchers. Int. J. Publ. Health 55, 239–242.

Stanley, F.K.T., et al., 2019. Radon exposure is rising steadily within the modern North American residential environment, and is increasingly uniform across seasons. Sci. Rep. 9, 18472.

Steck, D.J., et al., 2019. Spatial and temporal variations of indoor airborne radon decay product dose rate and surface-deposited radon decay products in homes. Health Phys. 116, 582–589.

Swisstopo, 2004. Digital Height Model DHM25. Federal Office of Topography.

Swisstopo, 2005. Geological Map of Switzerland 1:500,000. Federal Office of Topography.

Turner, M.C., et al., 2012. Radon and nonrespiratory mortality in the American cancer society cohort. Am. J. Epidemiol.

Vienneau, D., et al., 2009. A GIS-based method for modelling air pollution exposures across Europe. Sci. Total Environ. 408, 255–266.

Vienneau, D., et al., 2017. Effects of radon and UV exposure on skin cancer mortality in Switzerland. Environ. Health Perspect. 125, 067009.

Wang, M., et al., 2013. Evaluation of land use regression models for NO2 and particulate matter in 20 European study areas: the ESCAPE project. Environ. Sci. Technol. 47, 4357–4364.

WHO, 2009. WHO Handbook on Indoor Radon: A Public Health Perspective. World Health Organization, Geneva.

Wright, M.N., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Software 1 (Issue 1), 2017.