Year : 2014

PREDICTIVE ANALYSIS AND MAPPING OF INDOOR RADON CONCENTRATIONS IN SWITZERLAND

Georg Kropat

Department of radiology - CHUV

PREDICTIVE ANALYSIS AND MAPPING OF INDOOR RADON CONCENTRATIONS IN SWITZERLAND

PhD thesis in Life Sciences

presented to

Faculty of Biology and Medicine
University of Lausanne

by

Georg Kropat

M.Sc. in Physics

Jury

Prof. Philippe Maeder, President (CHUV-UNIL)
Prof. Francois Bochud, PhD Supervisor (CHUV-UNIL)
Dr. Sébastien Baechler, PhD Co-supervisor (Federal Office of Public Health, Switzerland)
Dr. Peter Bossew, Expert (German Federal Office for Radiation Protection, Berlin)
Mauro Gandolla, Expert (Università della Svizzera italiana)

Lausanne, 2014

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur Prof. Philippe **Maeder**
Directeur de thèse	Monsieur Prof. François **Bochud**
Co-directeur de thèse	Monsieur Dr Sébastien **Baechler**
Experts	Monsieur M. Mauro **Gandolla**
	Monsieur Dr Peter **Bossew**

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Georg Kropat

physicien diplômé de l' Université de Leipzig, Allemagne

intitulée

PREDICTIVE ANALYSIS AND MAPPING OF INDOOR RADON CONCENTRATIONS IN SWITZERLAND

Lausanne, le 9 février 2015

pour La Doyenne
de la Faculté de Biologie et de Médecine

Prof. Philippe Maeder

Table of contents

Acknowledgements

PhD theses are known to be a tough challenge in life. I am therefore grateful that I was surrounded by so many great people during this period at the Institute of Radiation Physics (IRA) in Lausanne.

My first thank you goes to my co-supervisor Sébastien Baechler for his enthusiasm and commitment to this project. When it comes to leadership, people often say that listening is one of the most important skills. Sébastien's ability to listen was one of the biggest helps during my PhD. Being a great scientist and an exemplary leader he always had the right instinct and an open ear for new ideas. Without his expertise I would not have had that much fun throughout this work.

Furthermore, I would like to thank my supervisor Francois Bochud for his scientific guidance and for his dedicated help during the writing of articles. Francois is an excellent physicist who gave me important advice in a variety of complex issues. I really enjoyed the working atmosphere at the IRA, this is also due to the fact that Francois is doing a great job as the director of the institute. It was a big pleasure to work with Francois.

A special thanks goes to the division of radiological risks at the Federal Office of Public Health. The help of Christophe Murith, Martha Palacios, Fabio Barazza and Walther Gfeller in scientific issues, organization of data and creating contacts was crucial for the success of this project.

Moreover, I want to thank Michel Jaboyedoff for his bright advice in geological and statistics related questions. Michel assisted in numerous meetings and helped us with many creative ideas to progress with the project. A big thank you goes to the experts of the jury of my PhD thesis. I am particularly grateful for Peter Bossew's encouragement and brilliant scientific support. Peter's help had a considerable impact on the outcome of this work. At the same time I thank Mauro Gandolla for his advice in many practical issues. Mauro's field expertise helped a lot to make sense of the data. Thank you also to the jury president Phillip Maeder for taking the time to evaluate this work.

Then, I want to thank the group of radiation protection at the IRA, especially to Nicole Meyer who helped me a lot with ^{222}Rn measurements and Jérôme Damet for his nice collaboration and his encouragement especially in the last phase of my PhD thesis.

In addition to that I want to say thank you to Gernot Butterweck from the Paul-Scherrer-Institute in Villigen for his advice and the efforts he put into our project of the Politrack calibration.

A very special thank you goes to Claude Bailat for his brilliant scientific advice in all measurement related issues and for many groovy jam sessions. It's great to meet fun people at work that also have a passion for music.

I also want to say thank you to Claude Collet for his friendly collaboration in the soil gas measurement campaign in La-Chaux-de-Fonds and his commitment in the organization of the scientific visit of our Montenegrin colleagues. A big thank you also to Harouna Wa Mbengi for his efforts to realize the soil gas measurements. I appreciated Harouna a lot as a colleague.

Moreover, I want to thank Manuel Santos for his genius ideas and skillful ability with the development of technical devices. At the same time thank Jean-Pascal Laedermann for many interesting discussions as well as for his dedicated help with the compute cluster. Thank you also to Pascal Froidevaux for his kind advice in environmental related questions and to Milan Beres for providing us with important geological data.

A big thank you goes to all the colleagues at the IRA with whom I shared so many great moments during the last four years. The lunch time and the after-work beer with awesome colleagues double the fun at work. I want to particularly thank Alexandre Ba, Julien Ott and Damien Racine for their kind corrections of my French writing attempts. Furthermore, a special thanks goes to Nicole Tille and her team in the secretary for their nice support during my PhD.

Finally, a big thank you to my mother Barbara and to my brothers Eckart, Christopher and Marcel. Your listening and personal support meant a lot to me during good and bad times. You are the best.

This work was partly financed by the Swiss Federal Office of Public Health.

Abstract

It is estimated that around 230 people die each year due to radon (^{222}Rn) exposure in Switzerland. ^{222}Rn occurs mainly in closed environments like buildings and originates primarily from the subjacent ground. Therefore it depends strongly on geology and shows substantial regional variations. Correct identification of these regional variations would lead to substantial reduction of ^{222}Rn exposure of the population based on appropriate construction of new and mitigation of already existing buildings. Prediction of indoor ^{222}Rn concentrations (IRC) and identification of ^{222}Rn prone areas is however difficult since IRC depend on a variety of different variables like building characteristics, meteorology, geology and anthropogenic factors.

The present work aims at the development of predictive models and the understanding of IRC in Switzerland, taking into account a maximum of information in order to minimize the prediction uncertainty. The predictive maps will be used as a decision-support tool for ^{222}Rn risk management. The construction of these models is based on different data-driven statistical methods, in combination with geographical information systems (GIS).

In a first phase we performed univariate analysis of IRC for different variables, namely the detector type, building category, foundation, year of construction, the average outdoor temperature during measurement, altitude and lithology. All variables showed significant associations to IRC. Buildings constructed after 1900 showed significantly lower IRC compared to earlier constructions. We observed a further drop of IRC after 1970. In addition to that, we found an association of IRC with altitude. With regard to lithology, we observed the lowest IRC in sedimentary rocks (excluding carbonates) and sediments and the highest IRC in the Jura carbonates and igneous rock. The IRC data was systematically analyzed for potential bias due to spatially unbalanced sampling of measurements. In order to facilitate the modeling and the interpretation of the influence of geology on IRC, we developed an algorithm based on *k*-medoids clustering which permits to define coherent geological classes in terms of IRC. We performed a soil gas ^{222}Rn concentration (SRC) measurement campaign in order to determine the predictive power of SRC with respect to IRC. We found that the use of SRC is limited for IRC prediction.

The second part of the project was dedicated to predictive mapping of IRC using models which take into account the multidimensionality of the process of ^{222}Rn entry into buildings. We used kernel regression and ensemble regression tree for this purpose. We could explain up to 33% of the variance of the log-transformed IRC all over Switzerland. This is a good performance compared to former attempts of IRC modeling in Switzerland. As predictor variables we considered geographical coordinates, altitude, outdoor temperature, building type, foundation, year of construction and detector type. Ensemble regression trees like random forests allow to determine the role of each IRC predictor in a multidimensional setting. We found spatial information like geology, altitude and coordinates to have stronger influences on IRC than building related variables like foundation type, building type and year of construction. Based on kernel estimation we developed an approach to determine the local probability of IRC to exceed 300 Bq/m^3. In addition to that we developed a confidence index in order to provide an estimate of uncertainty of the map. All methods allow an easy creation of tailor-made maps for different building characteristics.

Our work is an essential step towards a ^{222}Rn risk assessment which accounts at the same time for different architectural situations as well as geological and geographical conditions. For the communication of ^{222}Rn hazard to the population we recommend to make use of the probability map based on kernel estimation. The communication of ^{222}Rn hazard could for example be implemented via a web interface where the users specify the characteristics and coordinates of their home in order to obtain the probability to be above a given IRC with a corresponding index of confidence. Taking into account the health effects of ^{222}Rn, our results have the potential to substantially improve the estimation of the effective dose from ^{222}Rn delivered to the Swiss population.

1. Introduction

1.1. Radon risk

^{222}Rn is a radioactive gas which is a decay product of uranium (^{238}U) and radium (^{226}Ra). ^{238}U occurs everywhere in nature in form of traces and consequently ^{222}Rn can be found everywhere as well. Different types of rocks bear different amounts of ^{222}Rn depending on their amounts of ^{238}U and ^{226}Ra. In soil gas ^{222}Rn can easily reach activity concentrations above 100 kBq/m^3 (Dubois 2005; Neznal 2005; Kemski et al. 2006). In the outdoor air ^{222}Rn is strongly diluted and occurs hence only in negligible amounts (4-41 Bq/m^3 (Vaupotič et al. 2010)). However, in closed environments like buildings ^{222}Rn can occur in substantial concentrations above 1000 Bq/m^3 (Dubois 2005). Thoron (^{220}Rn) is an isotope of radon that is presumed to have similar health effects than ^{222}Rn. Due to lack of ^{220}Rn measurements in Switzerland, this study only addresses the ^{222}Rn problematic. Hence, with the term radon we refer exclusively to ^{222}Rn throughout this study.

Most of the ^{222}Rn which a human inhales is immediately exhaled again. However, the short-lived decay products of ^{222}Rn, in particular the α-emitting radionuclides ^{218}Po and ^{214}Po, can be deposited in the lungs and interact with biological tissue (Figure 1). The impact of one α particle can already lead to DNA damage of cell. Thus, a threshold indoor ^{222}Rn concentration (IRC) for cancer risk is commonly not hypothesized.

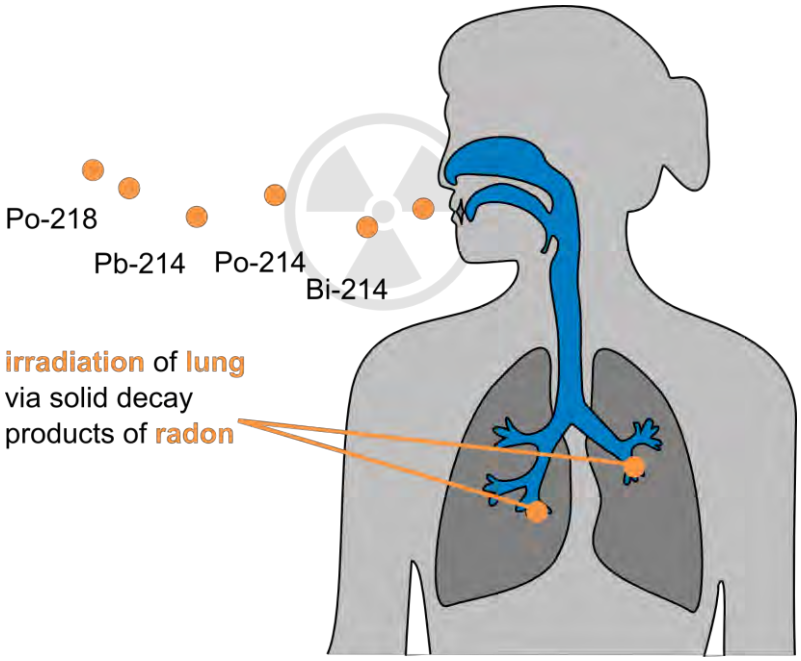


Figure 1 Decay products of ^{222}Rn in the human respiratory tract. ^{222}Rn itself is exhaled directly after inhalation and is thus not considered to be the direct cause of lung cancer

The association between ^{222}Rn and lung cancer has been studied the first time in ore miners from the region of Schneeberg in the Ore Mountains in eastern Germany. Lung cancer occurring in miners caused by ^{222}Rn was therefore known as “Schneeberg disease”. Since then several studies have been carried out in uranium miners and showed significant association between lung cancer and ^{222}Rn concentrations. It was therefore necessary to investigate the health effect of IRC on the general population. The results of initial attempts to study this effect showed however too much variation in between studies. For this purpose pooled studies have been carried out in

order to combine the results of the former studies. These studies have been performed in North America, China and Europe. The American study was a combined analysis of 7 north American case-control studies (Krewski et al. 2005). They found an increase of lung cancer risk of 11% (95% confidence interval 0-28%) per 100 Bq/m^3 of IRC based on a total of around 3700 cases and around 5000 controls. In China (Lubin et al. 2004) a pooled analysis of two studies with a total of around 1000 cases and around 2000 controls showed an increase of lung cancer risk of 13% (95% confidence interval 1-36%) per 100 Bq/m^3. The European study was a pooling of 13 former case-control studies with a total of around 7000 cases and 14000 controls (Darby et al. 2005). The increase of risk in lung cancer was 8% (95% confidence interval 3-16%) per 100 Bq/m^3 measured IRC. However after correction for regression dilution due to IRC measurement uncertainty, the increase of risk resulted to 16% (95% confidence interval 5-31%) per 100 Bq/m^3.

The three studies showed similar risk estimates and provide overwhelming evidence for an association between lung cancer and IRC in the general population. (Zeeb and Shannoun 2009) estimate that after correction for random errors in IRC measurements in the Chinese and the North American study, a weighted average of the three studies could yield an increase in risk of 20% per 100 Bq/m^3. This result is however only given informally in (Zeeb and Shannoun 2009). A pooled combination of all three studies will give a more precise estimate of the association between lung cancer and IRC after considering the random error in IRC measurements.

(Menzler et al. 2008) estimated the population attributable fraction (PAF) in Switzerland for lung cancer due to IRC to 8.3%. The PAF is the percentage reduction of cases of a disease after the elimination of a given risk factor (Rockhill et al. 1998). (Menzler et al. 2008) assumed an average IRC in Switzerland of 78 Bq/m^3 and based their analysis on the increase in lung cancer risk of 16% per 100 Bq/m^3 determined by (Darby et al. 2005). This results to around 230 deaths each year in Switzerland due to lung cancer caused by IRC. The PAF varied substantially among cantons. The highest PAFs were obtained in the cantons Jura (15.5%), Ticino (15.1%) and Neuchatel (14.4%). However, due to larger populations, the highest numbers of deaths due to lung cancer caused by IRC were estimated in Zürich (36), Bern (29) and Vaud (24).

The creation of maps with a higher spatial resolution that can be used for different buildings types can substantially improve such calculations.

1.1.1. Machine learning

The entry and propagation of ^{222}Rn concentrations in buildings are complex processes which are difficult to describe by parametric models. IRC data are usually not collected by trained professionals but by the home owners. Furthermore IRC are influenced by a variety of different variables like building characteristics that are in themselves qualitative information that cannot be included into a model that is derived from first principles. In order to model IRC it is thus inevitable to make use of empirical simplifications to represent the IRC data generating process. Traditional parametric models make usually strong assumptions on the interaction among the involved variables and their distributions, like the postulation of linear relation between IRC and predictor variables and their normal distribution. These assumptions are however not met in the case of IRC.

Machine learning is a field of research that proposed a huge variety of different methods in the last 30 years in order to overcome these limitations. Machine learning algorithms aim at extracting information and learning from data without making too strong assumptions about the data generating process and the involved variables.

Two major paradigms exist among machine learning methods: Supervised and unsupervised learning. Supervised learning consists of learning rules to assign labels or values to an outcome variable based on previous observations. Consider for example, an algorithm that is supposed to detect faces of members of the Swiss Federal Council among many other random portrait photos. In order to correctly classify a photo, the algorithm has to get information about the faces of the councilors prior to the classification task. In order to teach the algorithm the rules of how to assign an image to a councilor, the algorithm is trained with existing photos of Swiss Federal Council members. Training the algorithm based on existing photos is analogous to having a supervisor who tells the algorithm how faces of the councilors look like.

Unsupervised learning describes the process of creating variable assignment rules based on patterns in the data. An unsupervised algorithm could for example analyze a huge amount of portrait photos of members of the Swiss Federal Council without having information about which photo belongs to which councilor. However, defining a similarity measure for the faces on the photos, the algorithm can group the photos according to the councilor who is shown on it. Hence, the algorithm is grouping the photos without having prior information about which councilor belongs to which photo. This is in analogy to the situation where no supervisor is present who is telling the algorithm how to assign the images.

Unsupervised learning creates new rules and can be used to find structure in data. Supervised learning learns and generalizes rules which have been observed previously and is typically used for prediction tasks (Figure 2).

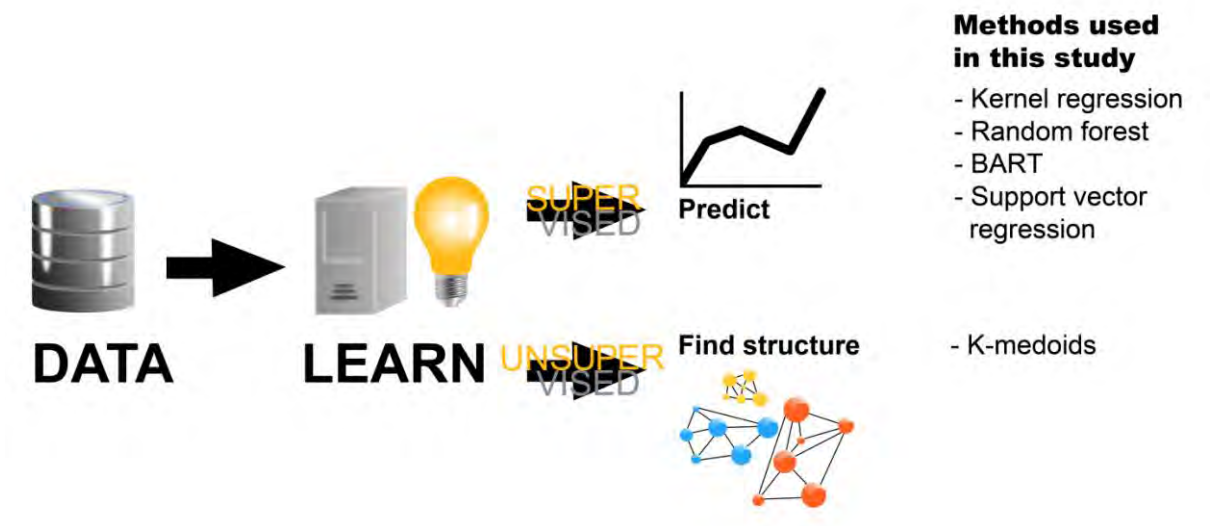


Figure 2 Scheme of supervised and unsupervised machine learning. The corresponding methods we used in this study are listed on the right of the figure

In order to model and understand the data generating process of IRC in Switzerland we used different supervised and unsupervised learning methods in this work.

1.1.2. Predictive analysis via supervised learning algorithms

In the following section several supervised learning methods are presented that were used throughout this PhD project.

1.1.2.1. Kernel regression

The prediction of an dependent variable y based on a set of independent variables $\vec{x}$ can be formulated in the following equation (Racine and Li 2004):

$$y = g(\vec{x}) + \varepsilon \tag{1}$$

$g(\cdot) = E(y \mid \vec{x})$ is the expected value of y conditional on $\vec{x}$ and ε a random error. $E(y \mid \vec{x})$ can be estimated by the approximation of the conditional probability density function $f(y \mid \vec{x})$ of y given $\vec{x}$ via the use of kernel functions. This results to the following weighted average

$$E(y \mid \vec{x}) = \frac{\sum_{i=1}^{N} Y_i K(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i)}{\sum_{i=1}^{N} K(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i)} \quad \text{(Racine and Li 2004)} \tag{2}$$

Each observation Y_i is weighted by a kernel function $K(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i)$. The choice of $K(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i)$ in this study is described in (Kropat et al. 2015b). $\vec{\sigma}$ and $\vec{\lambda}$ are vectors of smoothing parameters for continuous and categorical variables respectively and $\vec{X}_i$ is the i th observations of the predictor variables $\vec{x}$. $\vec{x}$ indicates the point in the predictor space at which y is estimated. A derivation of formula (2) can be found in Annex A1.

Kernel density estimation has the benefit that many conditional statistical indicators, like conditional mean values, probabilities and quantiles, can be easily estimated. However, the bandwidth estimation can be computationally more intensive compared to the other methods used in this study.

1.1.2.2. Random forests and Bayesian additive regression trees (BART)

Ensemble regression trees are methods that perform averaging over several regression tree models. The regression trees that are the basis of the models used in this PhD thesis are binary trees that aim at partitioning the predictor space in rectangular regions. For each region a simple model for the outcome variable y is calculated. In the case of random forests the model consists of the arithmetic mean of all observations in the rectangular predictor region. In BART, the estimate of y in a given predictor space region consists of drawing a random number from a Gaussian distribution. The regional model is accepted or not based on a posterior ratio criterion. A more comprehensive explanation of BART can be found in Annex A2.

The generation of regression tree ensembles is based on stochastic approaches. In random forests the regression tree ensemble is produced by a special form of bootstrap aggregation. In BART the tree ensemble is generated based on a Metropolis Hastings algorithm.

We found random forests to have the best predictive performance of all supervised learning methods of this work. Random forests are computationally fast and in addition to that a powerful tool in order to analyze the importance of the involved predictor variables. BART provides good predictability as well and has the advantage that it yields directly an estimate of the prediction uncertainty.

1.1.2.3. Support vector regression (SVR)

SVR is a supervised learning method that is based on the principle of regularization. Regularization is the principle of keeping the complexity of a model possibly small. This means in practice that large effect sizes are penalized during the model selection process (Cherkassky and Mulier 2007). Roughly speaking the principle of SVR can be divided into two parts: The statement of a linear model which is chosen based on regularization constraints. In order to cope with non-linearity, the predictor variables are transformed into a higher dimensional space by means of kernel functions.

The penalization of complexity keeps the risk of overfitting small in SVR. This makes SVR particularly suitable for problems with few observations but many variables. However compared to kernel regression and ensemble regression trees SVR has the drawback that categorical variables have to be taken into account as dummy variables. This can substantially increase the dimensionality of the model which is at the cost of prediction performance.

1.1.2.4. Revealing structure in data via unsupervised learning

In order to find a classification of lithological units in terms of their IRC characteristics, we performed *k*-medoids clustering of lithological units based on partitioning around medoids (PAM). A medoid of a point ensemble is a point which has the minimum average distance to all the other ensemble points (Xu and Wunsch 2008).

PAM is an iterative two step procedure. Prior to the initiation of the algorithm, a number k of clusters is defined. The algorithm is initiated by random assignment of k points as cluster medoids. Then each point of the dataset is assigned to its nearest medoid. In this manner k clusters are created. In the following step for each cluster the medoid points are recalculated. The two steps of reassigning points to its nearest medoid and the re-estimation of the new cluster medoids is repeated until convergence.

k-medoids has the advantage that it only needs information on the distance between points. No absolute position in a vector space is needed. This is very useful for the clustering of lithological units in this study, since we defined the pair wise distance between lithological units via Kolmogorov distances of the corresponding IRC distributions. Furthermore, k-medoids is very robust to outliers since the position of a medoid is given by a data-point. The variation of the position of a medoid is hence limited by the position of the points in the data set.

1.2. The goal of the study

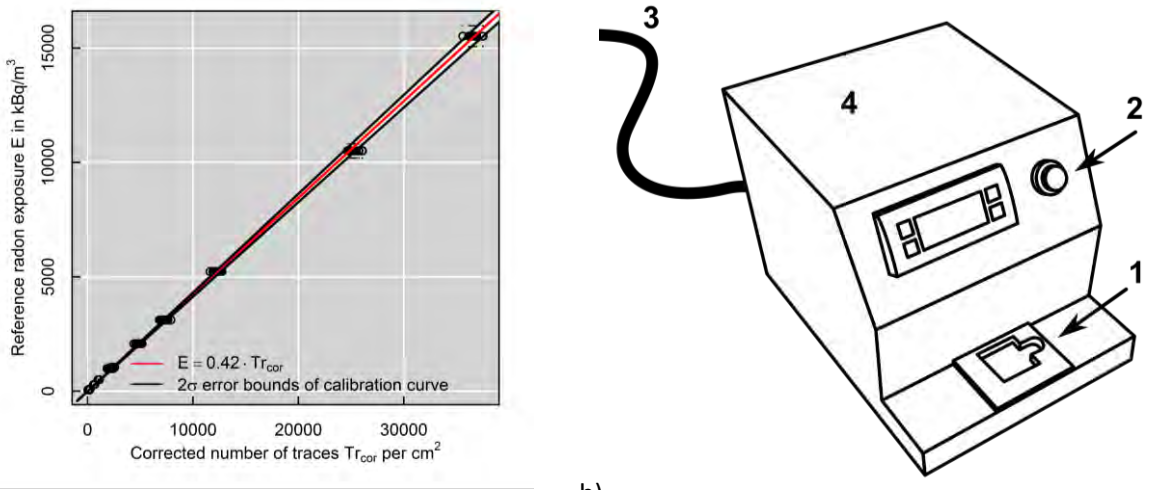
The present work aims at the development of models for the prediction and the understanding of IRC in Switzerland, taking into account a maximum of information in order to minimize the prediction uncertainty. The predictive maps will be used as a decision-support tool for radon risk management. The construction of these models is based on different data-driven statistical methods, in combination with geographical information systems (GIS).

2. Brief summary of principle results

We calibrated and characterized a solid state nuclear track detector (SSNTD) reader system (Kropat et al. 2015a) to understand the comparability of different IRC measurement systems, the uncertainty influence on IRC from the measurement system and the role of ^{222}Rn concentrations during the transport of an SSNTD between measurement laboratory and measured buildings. (Kropat et al. 2014) gives an insight into the univariate influences of several potential radon determinants. The analyses were carried out by visualizing simultaneously the possible bias due to spatially inhomogenous distributions of IRC sampling. (Kropat et al. 2015b) reports on modeling IRC based on kernel estimation. The approach permits to develop predictive as well as probability maps. In addition to that we developed a confidence index map in order to provide an uncertainty indication for a local IRC probability estimation. Moreover, (Kropat et al. 2015c) describes predictive IRC mapping using ensemble regression tree in order to improve predictability of kernel regression and other former approaches (Hauri et al. 2012). Ensemble regression trees provide powerful tools to analyze the influence of a predictor variable in a multidimensional setting. Furthermore, based on unsupervised learning, we found an approach to group lithological units according to their IRC characteristics. Finally, we carried out a field study in order to assess the potential of SRC as an IRC predictor. A short description of this study can be found in Annex A3.

2.1. Calibration of the Politrack® system based on CR39 solid state nuclear track detectors for passive indoor radon concentration measurements (Kropat et al. 2015a)

The biggest part the Swiss IRC data base consists of measurements carried out with SSNTD. We calibrated a SSNTD reader system and developed tools to monitor the stability of the system in collaboration with the Paul-Scherrer Institute (PSI), Switzerland. The reference exposures for the calibration were performed in the radon chamber of the Secondary Calibration Laboratory at the PSI. In order to calculate the calibration curve and the corresponding uncertainty we used a Monte Carlo fitting procedure (Figure 3a). To monitor the long term stability of the system, we developed a device to produce reference SSNTDs which are accompanied to the SSNTD development and reading procedure of each read out series (Figure 3b). We determined the characteristic limits for the detection of a potential drift of the system based on ISO Standard 11929. The overall uncertainty of the system was determined to 8%. To compare the performance of the system and our calibration, we performed a comparison measurement in 30 Swiss schools with commercially available SSNTDs from the manufacturer Landauer Nordic. Both systems showed good accordance. Finally we explored the potential influence of IRC concentrations during transport. For this purpose we exposed 40 SSNTDs welded in plastic bags to an exposure of about 15000 kBq h m^{-3} in the PSI. Our results indicate sufficient air tightness of the detector packaging.



a) b)

Figure 3a) Calibration curve of the Politrack SSNTD reader system. b) Device to produce reference SSNTDs in order to monitor long term changes of the Politrack system (1 Shutter to screen Am-241 source; the reference SSNTD is placed here for irradiation, 2 button to start exposure, 3 compressed air to open shutter, 4 body housing with shutter mechanics)

2.2. Major influencing factors of indoor radon concentrations in Switzerland (Kropat et al. 2014)

Before being able to model and to map IRC we determined the main factors that influence this quantity through univariate analysis (Kropat et al. 2014). We had about 212 000 IRC in about 136 000 buildings at our disposal. We took into account the variables foundation type, year of construction and building type, altitude, average outdoor temperature during the measurement and the lithology. The principle approach was to graphically compare the 95% confidence intervals of the classes of each variable. In order to assess bias due to spatially inhomogeneous distributions of sampling, we created a map of the spatial distribution of IRC sampling for each class of each variable. The amount of spatial clustering in the IRC data was measured via fractal dimension. We found significant associations between IRC and all covariables taken into consideration. Electret detectors revealed 35% higher IRC measurements than track detectors. Concerning building type, we found IRC of apartments to be considerably lower than detached buildings. In addition to that, concrete foundation showed the lowest IRC among all foundation types. Buildings constructed after 1900 revealed remarkably reduced IRC compared to older buildings. We observed a further decrease after 1970. Moreover, the univariate analyses showed a significant association between IRC and outdoor temperature estimates. The lowest IRC occurred at the highest temperatures. Furthermore, IRC are associated with altitude. With respect to lithology, IRC in carbonate rock in the Jura Mountains are by a factor 2 higher than in carbonate rock in the Alps. The lowest IRC we observed in sedimentary rock and sediment and the highest IRC in igneous rock and in carbonate rock from the Jura Mountains. A map of the lithological units can be found in Figure 4a and the mean IRC for each lithological unit in Figure 4b. In order to study IRC differences within buildings we compared IRC between different floor levels. Finally, we produced a probability map to assess the risk of exceeding an IRC of 300 Bq/m^3 using basic geostatistical techniques

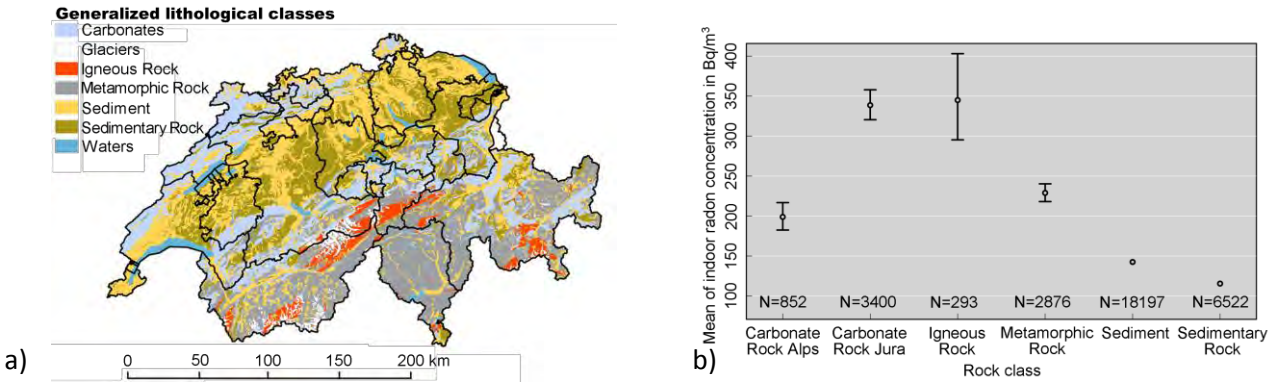


Figure 4 a) Spatial distribution of generalized lithological classes in Switzerland b) Mean IRC and 95% confidence intervals versus generalized lithological classes

2.3. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: an application to Switzerland (Kropat et al. 2015b)

In this article we used kernel estimation in order to create maps indicating the local mean IRC as well as the probability to exceed an IRC of 300 Bq/m^3 (Figure 5a) for specific building characteristics. Furthermore we developed a confidence index in order to indicate the confidence on a local probability value (Figure 5b). After an update of the IRC data, we had about 240 000 IRC measurements available in around 150 000 buildings. Based on the results of the univariate analysis we took into account the predictor variables: building type, foundation type, year of construction, detector type, geographical coordinates, altitude, temperature and lithology. Categorical and continuous variables were accounted for by the choice of appropriate kernels.

The kernel regression yielded an R^2 of 28%. We evaluated the importance of each variable in the model based on the corresponding bandwidth value and explored the mapping properties of the method on different spatial scales. In addition to that, the method allows obtaining different maps for different architectural characteristics. Maps produced for detached buildings with concrete foundation revealed substantially smaller IRC than maps for farms with earth foundation.

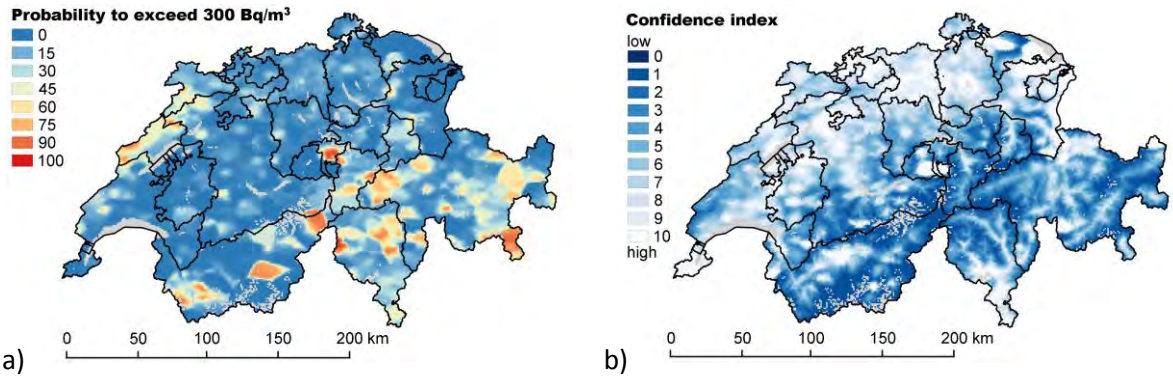


Figure 5a) Map of the local probability to exceed 300 Bq/m^3 b) Confidence index for the probability to exceed 300 Bq/m^3. Both maps correspond to detached houses, with concrete foundation, built between 1900-1970, Gammadata SSNTDs and an outdoor temperature of 3.5°

2.4. Improved predictive mapping of IRC using ensemble regression trees based on automatic clustering of geological units (Kropat et al. 2015c)

In this study we took advantage of ensemble regression trees in order to improve the predictability of modeling compared to kernel regression (Kropat et al. 2015b) and other approaches that were used in Switzerland (Hauri et al. 2012). For this purpose we used the same data set as in the kernel regression part (Kropat et al. 2015b) and performed data driven modeling based on random forests and BART. Random forests were able to explain 33% of the variation in IRC data and BART 29%. Since BART is based on posterior sampling of regression tree ensembles, the prediction uncertainty can directly be obtained by calculating the standard deviation of the posterior sample. Random forests provide a convenient way to evaluate the importance of a predictor variable in a multidimensional setting. We found that building related variables have a less important influence on IRC than location/ geology related variables.

It is common to group lithological classes into subgroups in order to obtain models which are easier to interpret. Many of these approaches group lithological classes based on major rock types like igneous rock, sedimentary rock, metamorphic rock etc. However the IRC characteristics can be very different for lithological classes within major rock types. Our approach directly groups lithological classes according to their IRC characteristics. We calculated the Kolmogorov distance between IRC distributions of all pairs of lithological classes. The Kolmogorov distance served as similarity metric in order to perform *k*-medoids clustering (Figure 6a). The explained variance of the original classes was 6.5% whereas the explained variance of the regrouped lithological units resulted to 6.3%. Mapping the regrouped lithological units yields a detailed representation of lithological areas in terms of their IRC characteristics (Figure 6b).

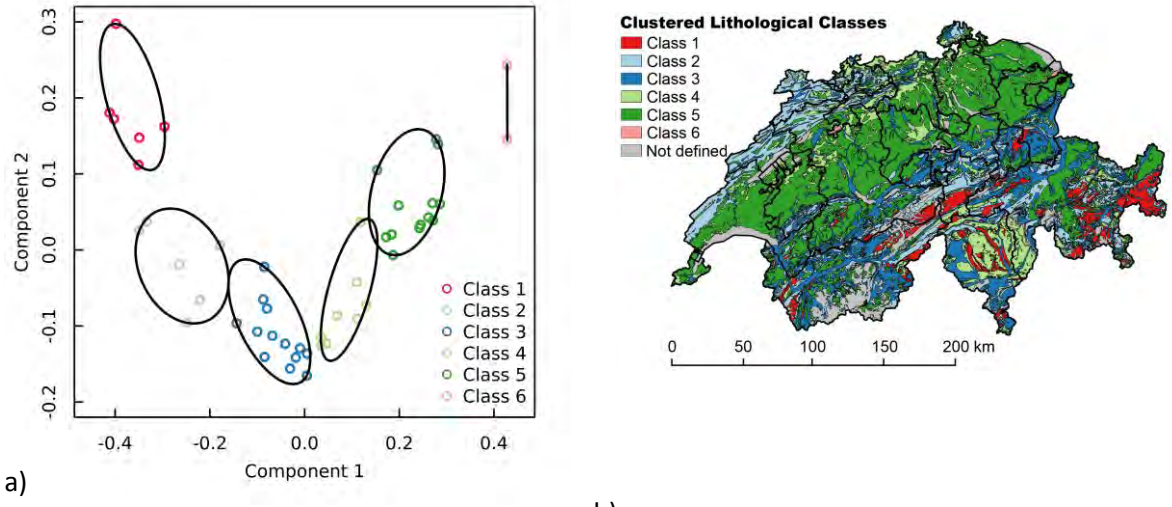


Figure 6a) Multidimensional scaling representation of Kolmogorov distances between IRC distributions of lithological classes. The different groups resulting from k-medoid clustering are indicated by different colors b) Map of clustered lithological units. Areas of lakes, glaciers or of lithological units for which not enough IRC measurements were available are indicated in grey

3. Discussion

Taking into account a maximum of information, we developed a variety of approaches in order to improve the current Swiss IRC mapping (Figure 7).

3.1. IRC data

IRC measurements in Switzerland are highly clustered. The 63 950 buildings we examined in (Kropat et al. 2014) have a fractal dimension of 1.15 whereas a dimension of 2 is expected for a perfectly random spatial distribution. After declustering, the fractal dimension of the IRC measurements resulted to 1.42 with 32 151 buildings. This implies that there is still some clustering present. However stronger declustering would reduce the number of observations, such that not enough measurements would be available for the analyses. The IRC measurements have a mean value of 198 Bq/m^3 for the unclustered and 189 Bq/m^3 for the declustered case. In both situations the arithmetic means are higher than 78 Bq/m^3 which is the value obtained by (Menzler et al. 2008). This is due to the fact that this study used a population weighted approach. Urban areas that usually have lower IRC are hence given a stronger weight. Since the aim of this work is mapping, prediction and analysis of IRC over the whole of Switzerland we favor a mean that takes all geographical regions equally into account.

The log-normal hypothesis for the IRC distribution in Switzerland is rejected by a Kolmogorov-Smirnoff test. Fitting a gamma distribution to the log-transformed IRC yields a slightly better Kolmogorov distance (0.04) than the normal distribution (0.055). Clustering and inhomogeneous spatial distributions may be the cause of the departure from log-normality. The Kolmogorov distance is however indicating that the empirical cumulative distribution of the log-transformed IRC does not have a larger deviation than 6% of the fitted normal distribution. With regard to statistical methods assuming normal distributions, we considered the log IRC distribution to be sufficiently close to normality.

We found electrets to overestimate IRC substantially compared to solid state nuclear track detectors (SSNTD). Electret detectors consist of an electrostatically charged teflon disk enclosed in a permeable plastic housing. The α-particles originating from ^{222}Rn ionize the air within the plastic housing. The ionized air causes a discharge of the Teflon disk. After the measurement, the IRC can be estimated by the voltage difference of the electret between the beginning and the end of the measurement. Dust and humidity can lead to an unforeseen discharge of the electret which results in overestimation of IRC. Taking into account the IRC of the whole country shows an overestimation of SSNTD. This can be explained by the observation that regions with a tendency of higher IRC like the canton Ticino and the Jura Mountains have been more strongly sampled with SSNTD compared to electrets. Restricting the study area to the Swiss Plateau shows that electrets actually overestimate IRC compared to SSNTD. This finding is in accordance with earlier observations (FOPH 2011).

3.2. Building characteristics

Architectural characteristics of measured buildings substantially influence IRC. We found earth foundation to be associated with higher IRC than concrete foundations. This can be explained by the fact that earth foundations are more permeable and hence more prone to let radon enter into the building. Surprisingly we found concrete foundation built into buildings after construction to have higher IRC than the earth foundation or built-in concrete foundations. It may be that buildings with high IRC are more likely to have a concrete foundation that was built after construction as mitigation measure. However our results implicate that this method does not sufficiently

reduce IRC. As (Hauri et al. 2012) we found that IRC decrease with the year of a buildings construction. Buildings constructed before 1900 revealed the highest IRC. We suppose that the ^{222}Rn emanation of natural stones is the cause of this observation, since this type of building material was more often involved in construction before 1900 (Gunby et al. 1993). We observed furthermore a drop in IRC after 1970, which we attribute to a change in building regulation after the oil crisis in the 1970ties (Burkart et al. 1984). Since 1970 we could not conclude a change of IRC. We also observed differences in IRC means associated with building type. Apartment buildings had the lowest IRC followed by detached buildings and farms. We found the building type to be related to building age as well as foundation type.

The differences in bandwidths of the kernel regression for the variables building type, foundation and year of construction compared to the maximum bandwidths show that the kernel regression accounted for these variables. This is also visible in the mapping results for different architectural situations. For detached buildings with concrete foundation built between 1970 and 1990 we found considerably different mapping results compared to farms with earth foundation built between 1900 and 1970. The kernel bandwidth provides a useful tool in order to interpret the influence of a variable on the model. A drawback is however, that bandwidths depend on whether the corresponding variable is continuous or categorical as well as on the number of classes in a categorical variable.

In order to analyze the effect of floor levels on IRC we compared the IRC of a given floor level to the IRC of the basement individually for each building. We found a nearly linear relationship between the mean IRC ratio and the floor level. In order to analyze the similarity of IRC within buildings, we compared pair-wise IRC between basement and ground floor as well as first floor and second floor. We found that IRC between basement and ground floor are less correlated than IRC between first floor and second floor. One reason for this may be that IRC in higher floor levels are due to ^{222}Rn emanation from building materials, which would result in similar IRC, independent of the floor level. However one has to bear in mind that in 37% the measurements in the second floor were above 100 Bq/m^3. An IRC of more than 100 Bq/m^3 is however not likely to be caused by building materials (Schuler et al. 1991). Furthermore, transfer via electrical conduits may cause IRC in upper floor levels. This would explain a similarity in IRC for different upper floors.

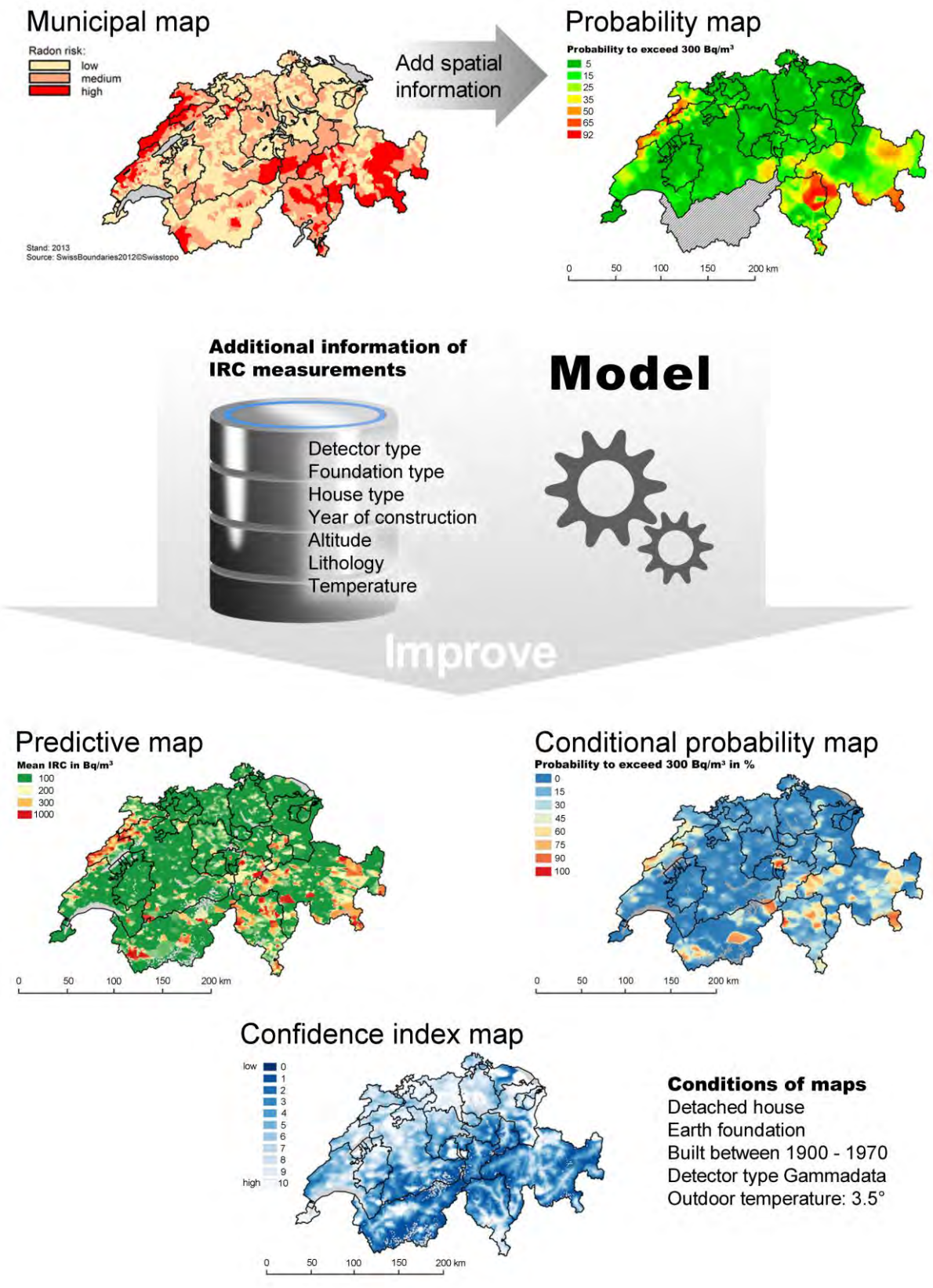


Figure 7 Scheme of the mapping improvement by accounting for coordinates and building characteristics in order to obtain more adapted and detailed maps

3.3. Lithology

Like many other studies, we found IRC to be related to geological information (Gunby et al. 1993; Bossew et al. 2008; Smethurst et al. 2008; Appleton and Miles 2010). Igneous rock showed the highest concentration. This can be explained by the fact, that igneous rock like granites is rich in uranium (Schön 2004). The second largest lithological class is carbonate in the Jura Mountains which consists mainly of limestone. Limestone is known to be subject to weathering, which is also called karstification. Karstification can result in large cave systems that facilitate the propagation of ^{222}Rn (Sajó-Bohus et al. 1997). Limestone is however not known to be particularly rich in ^{238}U or ^{226}Ra (2-4 ppm (von Gunten et al. 1996)). A common hypothesis for higher IRC in the Jura Mountains is that ^{222}Rn originates from the crystalline basement subjacent to the limestone layer (see Figure 8) Due to the high permeability of the karstified limestone ^{222}Rn can easily move to the surface. This would explain our observation that IRC are less elevated in carbonates in the Alps than in carbonates in the Jura Mountains since the sedimentary rock layers are thinner in the Jura Mountains than in the Alps.

Nevertheless, the hypothesis that IRC in the Jura Mountains originate from the crystalline basement is controversial since the migration time through the limestone layer is assumed to be considerably longer than the ^{222}Rn half life of 3.8 days (Parriaux et al. 2010).

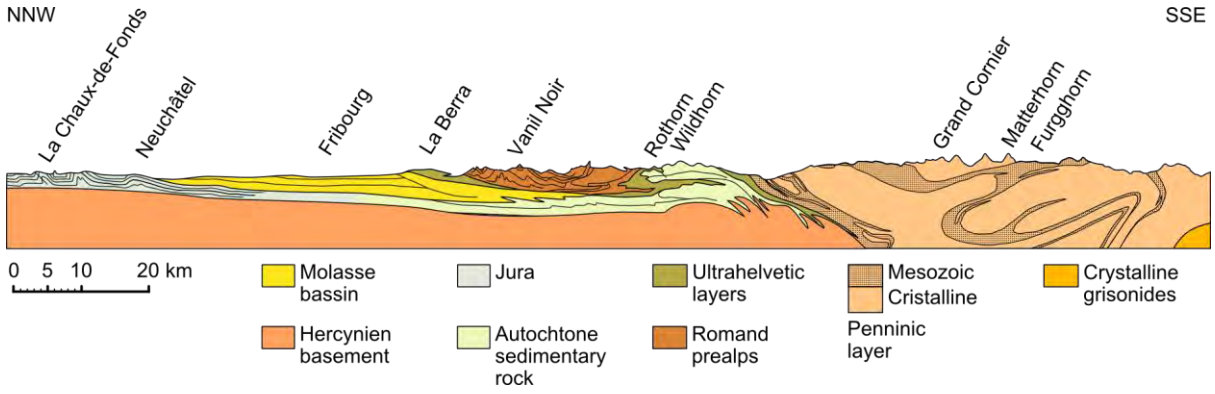


Figure 8 Geological cut of Switzerland from north to south (adopted from (Wepf 1934)). The Jura layer is thinner than the autochtone sedimentary rock layer.

Another hypothesis that could explain the high IRC in limestone may be ^{238}U enrichment of the limestone in consequence of weathering (von Gunten et al. 1996; Parriaux et al. 2010). (von Gunten et al. 1996) observed remarkable disequilibria between uranium and its daughter products ^{230}Th and ^{226}Ra in limestone rich areas of the Jura Mountains. In a soil that is rich in carbonate, uranium is soluble when it is complexed by carbonate and is therefore washed out of the soil. On the other hand, the uranium daughter products ^{230}Th and ^{226}Ra do not form such soluble carbonate complexes and are hence immobile. This leads to the fact that only uranium is washed out in the karstic system, leaving a disequilibrium between ^{234}U and ^{230}Th in the upper soil layer as shown in Figure 9. However uranium is only soluble in the oxidation state (VI). In an anoxic aqueous environment, as it can occur in greater depth in the karstic system, U(VI) can be reduced to a U(IV) compound (uraninite). U(IV) compounds are not soluble and precipitate. This leads consequently to uranium enrichments in areas of the karst were reducing conditions can be found. The ^{222}Rn originating from this uranium enrichment can easily propagate in the porous

karst system and hence increases the probability of high IRC in buildings built on the top of limestone (Parriaux et al. 2010).

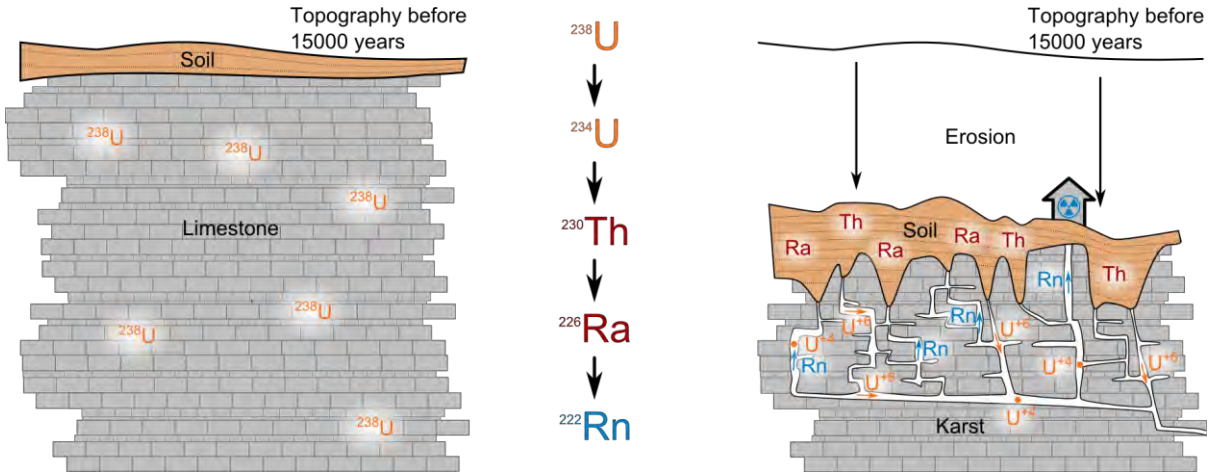


Figure 9 Uranium enrichment in karst due to weathering: Uranium is washed out during erosion into the karstic system. ^{230}Th and ^{226}Ra accumulate in the upper soil layer since they don't form soluble carbonate complexes

Carbonate rock is a sedimentary rock. However, in this work we separated carbonate rock from other sedimentary rocks due to its particular IRC characteristic. The lowest lithological classes in terms of IRC were sediments and sedimentary Rocks. This was also found in other studies (Bossew et al. 2008; Hunter et al. 2009). It has however to be noted that sediments can originate from a variety of different rock types. Therefore local areas with higher IRC in sediments and sedimentary Rocks are not unusual (Bossew et al. 2008). Finally, metamorphic rock shows intermediate IRC. This is well in line with the fact that metamorphic rock can be derived from many different geological formations, igneous rock and sediments alike.

The automatic classification of lithological units in (Kropat et al. 2015c) has the ability to create meaningful groups of lithological units only based on their similarities in IRC distribution. Cluster 1 is primarily found in the Alps and contains mainly two-mica gneisses, granites and porphyrites. Cluster 2 and 3 can be found in the Jura Mountains as well as in the Alps. Both classes contain carbonates. However, for the most part the Jura Mountains are covered by cluster 2 and cluster 3 appears rather in the Alps. The clustering algorithm distinguishes hence automatically between carbonates in the Alps and carbonates in the Jura Mountains. In cluster 3 also metamorphic and sedimentary rocks can be found. Consequently, cluster 3 groups lithological classes with moderate IRC. Finally, cluster 4 and 5 represent essentially lithological units in the Swiss Plateau containing sedimentary rock and sediment. The clustering method is particularly useful to distinguish between metamorphic rocks with different IRC characteristics. As described above, metamorphic rocks can originate from a variety of different rock types. This leads to a high variation of IRC within this rock class. Our clustering method puts the metamorphic rocks in the appropriate classes and can especially point out the IRC characteristics of different metamorphic rock types in the Alps.

The kernel regression yielded a bandwidth of $\lambda = 0.7$ with a maximum bandwidth of $\lambda_{max} = 0.86$ for lithology. This indicates that the kernel regression accounts for this variable, which is also visible in the IRC mapping where the lithological units are clearly observable on a local scale. In the kernel estimation we took lithology as an unordered categorical variable into account. However some countries defined a geogenic ^{222}Rn potential for their

geological units. In future work this information could be taken into account by using kernels for ordered categorical units.

3.4. Altitude

Finally we found IRC to be related to altitude. We do not expect a causal relationship between altitude and IRC, but we consider altitude rather to serve as a proxy variable for geological information as well as for meteorology. Igneous rock in the Alps and carbonate in the Jura Mountains show high IRC and are at the same time located at higher altitudes. On the other hand, sediment and sedimentary rock are attributed to lower IRC and occur predominantly at lower altitudes in the Swiss Plateau.

3.5. Mapping

We performed mapping based on several methods: spatial aggregation, kernel estimation, regression tree ensembles and clustering of lithological units.

Based on spatial aggregation we mapped the probability to exceed an IRC of 300 Bq/m^3 (Kropat et al. 2014). This map represents well the regional differences of IRC in Switzerland. Compared to the existing radon risk map of Switzerland (FOPH 2013) this map has the advantage that it is not bound to municipality boundaries. We consider it hence being more neutral with respect to political questions. However this map does not differentiate between different building characteristics.

Based on kernel estimation, we modeled and mapped IRC by taking into account all together the spatial relationships between IRC observations, building characteristics, measurement conditions and geological information. Also with this method the well known tendencies between the major geological areas Alps, Swiss Plateau and Jura Mountains have been well represented. At the municipal level, we could show that kernel estimation improves the spatial detail compared to former mapping approaches. In addition to that, the lithological units are clearly visible in the mapping results on a local scale. This indicates directly that kernel estimation takes into account the relation between IRC measurements as well as lithological information. For the coordinates we chose kernels that consider bandwidths that are constant over the whole range of the SN and EW direction. For future work we suggest however to consider methods which can adapt bandwidths to different regions, in order to account for locally different IRC trends.

The mapping of the probability to exceed an IRC of 300 Bq/m^3 based on kernel estimation revealed similar spatial trends as the predictive map based on kernel regression. In the areas where few measurements have been carried out, for example the Alps, the mapping results have however to be interpreted carefully. For this purpose we developed a confidence index map that can be accompanied with the probability map. This map has a strong relation to the topography of Switzerland, which is not surprising since at higher altitudes the density of IRC sampling is lower than at lower altitudes. The estimation uncertainty is consequently higher. However, uncertainty differences are well represented between areas at similar altitude and different densities of IRC sampling. We also developed a method to validate probability estimations. For this purpose we created a grid of 5 *km* x 5 *km* and compared the mean estimated probability to exceed 300 Bq/m^3 with the actual proportion of measurements above 300 Bq/m^3 of an independent test set. This validation yielded an R^2=78%. The kernel estimation shows hence a good agreement with the actual observed proportion. However, the reader has to bear in mind that this is based on a comparison of spatially aggregated statistics. We expect a pointwise comparison to

be more prone to unexplained variance. Obtaining the observation of a proportion is however difficult for a spatial point because it is not very common to have several buildings at the same location.

IRC mapping based on ensemble regression trees yields similar global patterns than kernel estimation. Like kernel estimation, ensemble regression trees provide the possibility to create maps specific to particular building characteristics. However BART has the specific advantage that prediction uncertainties can easily be mapped. Similarly to the confidence index map based on kernel estimation the prediction uncertainty map form BART shows a strong association to topography.

Finally, the results from lithology clustering can be used to distinguish regions according to their radon characteristics (Figure 6b). This map has the advantage that the geometry of the lithological units determines the spatial structure of the map. (Kemski et al. 2009) consider lithological units as spatial support for mapping the geogenic radon potential (GRP) as being the optimal choice since the GRP is primarily determined by geological conditions.

3.6. Predictive models

Kernel regression could explain 28% of variation in IRC data all over Switzerland (in terms of variance). Restricting the data only to farms resulted in a considerably higher explained variance of 38%. In the univariate analyses, we observed farms to have higher IRC than other building types. At the same time we assume that many uncontrollable error influences, for example IRC during transport, get less important with higher IRC. Consequently, the data is less prone to unexplainable variance and can better be modeled. BART yielded an R^2=29% and random forests an R^2=33% for IRC measurements from all over Switzerland. In Table 1 a comparison of the different methods used in this work can be found. Only taking into account the municipality as predictor already yields an R^2 of 20%. It can be expected that houses are generally not homogeneously distributed within a municipality but rather closely clustered. Calculating the mean value per municipality is hence similar to *k* nearest neighbor estimation and explains therefore a substantial part of the spatial IRC variation. We found random forests to perform best in terms of predictability. For different prediction tasks an R^2 of 33% may appear to be small. However, referring to (Hauri et al. 2012), an R^2 of 33% can be considered as a good performs in comparison to IRC modeling studies from other countries. The sources of unexplained variance in IRC data are discussed in the following section.

Table 1 Comparison of different IRC mapping and prediction methods that were used formerly and throughout this work. This table considers IRC measurements from all over Switzerland

	Prediction	Predictability (R^2)	Probability estimation	Uncertainty mapping
Municipality mapping	X	20%	-	-
Linear modelling (Hauri 2012)	X	20%	-	-
Kernel estimation	X	28%	x	x
Bayesian additive regression trees	X	29%	-	x
Random forest	X	33%	-	-

3.7. Uncertainty

The calibration procedure of the SSNTD resulted in a combined expanded uncertainty of 8% (k=2). We also found that the plastic transport bags are sufficiently air tight to avoid error influences during the mailing to the home owners. However when the detectors are sent back, the plastics bags are usually not welded. That means that ^{222}Rn concentrations during transport from the buildings to the measuring laboratory can influence the final ^{222}Rn measurement. This leads to variance which can practically not be explained by statistical modeling although the travelling time is usually much smaller than the actual measuring time. Therefore, we consider the IRC distribution within buildings to be the strongest source of unexplained IRC variation. Within a room, the positioning of the IRC detector can have crucial consequences on the measurement. We experienced in field observations, that the IRC nearby electrical sockets or wall cracks are substantially higher than elsewhere in the room. Further studies to quantify this influence are necessary in order to understand the limitations of IRC modeling and prediction. We expect it however to be substantially higher than the uncertainty due to etching and readout of SSNTD detectors.

3.8. Influence of thoron during long term IRC measurements

Another isotope of radon, coming from building material and commonly called thoron (^{220}Rn), was not considered in the present study although it is known to have potentially two effects: it can bias the estimation of IRC and it can contribute to the annual effective dose delivered to the Swiss population. However, estimating retrospectively the influences of ^{220}Rn on long term IRC measurements is difficult, but its presence should be more thoroughly considered in future IRC measurement campaigns.

4. Conclusion and perspectives

Two main questions arise when it comes to national radon risk communication: What effects does radon have on human's health and how can local radon hazard be estimated. This work addresses the estimation of local radon hazard. We had a large dataset of Swiss IRC measurement available. Since IRC are driven from a multitude of determinants, in-depth understanding of the multidimensional nature of the IRC generating process is necessary in order to localize radon hazard appropriately. In our work we addressed this issue via both univariate and multidimensional analysis and modeling of Swiss IRC.

As depicted in Figure 7 we are now able to improve the existing radon communication by taking into account a variety of IRC influencing variables in order to produce tailor-made IRC maps for Switzerland. Our results are the first element for the development of a risk map. Taking into account the health effects of radon, the radiation dose to the population due to ^{222}Rn can be more accurately estimated by considering for example the local population densities and the stratification for different building types.

Finally, for the risk communication to the public in Switzerland, we suggest to use the probability map proposed in (Kropat et al. 2015b). We consider probabilities to be easier to communicate than mean values due to a much more intuitive concept of contingency. Mean values are often interpreted as a deterministic value and lay persons may disregard the fact that a mean value is only an indicator for a statistical distribution which itself is subject to randomness. On the other hand a probability could be communicated as being the fraction among several buildings that exceed a certain IRC. In order to provide an assessment of uncertainty of the probability maps we developed a confidence index which can be communicated additionally to the probability estimates. Local

probabilities to exceed a certain IRC and the corresponding confidence index could be communicated to the public via a web interface where users can specify the properties and coordinates of their homes (Figure 10).

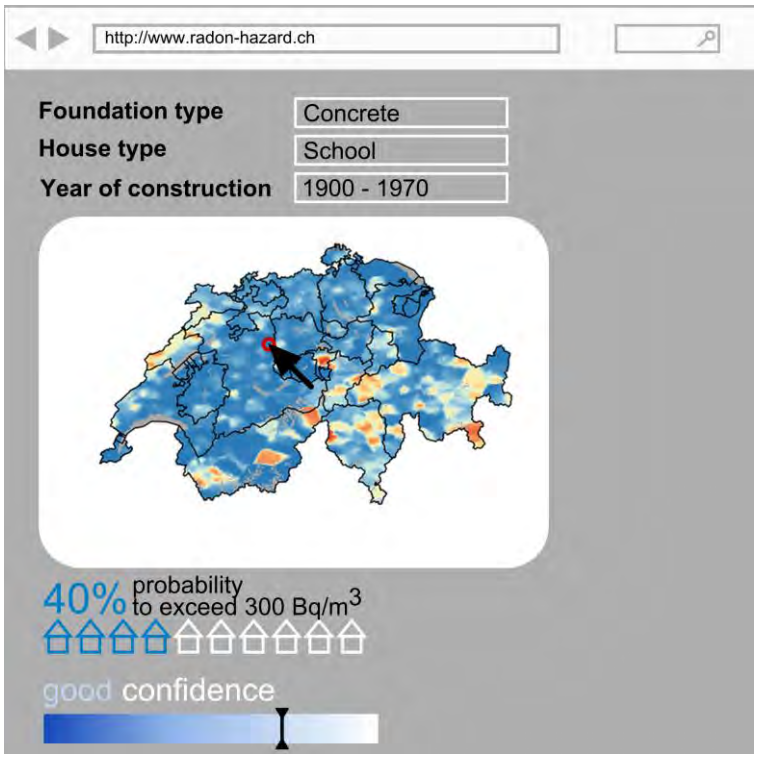


Figure 10 Mockup of a possible website to communicate the probability to exceed a certain IRC for given building characteristics. The probabilities could be accompanied by a confidence estimate.

Annex

A1. Kernel density estimation

The following section aims at giving more detail for the interested reader on the theoretical principle of the estimation of the conditional expectation $E(y|\vec{x})$ via kernel density estimation. $E(y|\vec{x})$ was introduced in equation (2) in the introduction section.

In order to derive equation (2), the conditional expectation $E(y|\vec{x})$ can be written as

$$E(y|\vec{x}) = \int_{-\infty}^{\infty} y f(y|\vec{x}) dy \quad \text{(Racine 2008)} \tag{3}$$

$f(y|\vec{x})$ is the conditional probability density function of y given $\vec{x}$. The estimation of $f(y|\vec{x})$ is the core problem of kernel regression and can be carried out via kernel density estimation. The conditional probability density $f(y|\vec{x})$ can be expressed via the following equation:

$$f(y|\vec{x}) = \frac{f(y,\vec{x})}{f(\vec{x})} \tag{4}$$

Where $f(y,\vec{x})$ is the joint probability function of $\vec{x}$ and y and $f(\vec{x})$ is the marginal probability density of $\vec{x}$. The joint probability density $f(y,\vec{x})$ can be estimated by:

$$\hat{f}(y,\vec{x}) = \frac{1}{Nh}\sum_{i=1}^{N} K_h(\vec{x},y,h,\vec{X}_i,Y_i) \quad \text{(Racine and Li 2004)} \tag{5}$$

Where K is a kernel function, N is the number of observations, h a smoothing parameter (bandwidth), and $\vec{X}_i$ and Y_i are the ith observation of $\vec{x}$ and y respectively.

The choice of K depends on the types of variables taken into account. IRC are influenced by a variety of variables. These variables can either be continuous or categorical. For both cases different kernels have to be taken into account.

A1.1. Kernel for continuous variables

For continuous variables we assumed a Gaussian kernel w in this work

$$w\left(\frac{x_t^c - X_{t,i}^c}{\sigma_t}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_t^c - X_{t,i}^c}{\sigma_t}\right)^2} \quad \text{(Specht 1991)} \tag{6}$$

Where $X_{t,i}^c$ represents the ith instance of the variable x_t^c and σ_t its bandwidth.

In the case of several continuous variables, the kernels of each variable can be combined in a product kernel

$$W\left(\vec{x}^c,\vec{X}_i^c,\vec{\sigma}\right)=\prod_{t=1}^{p}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_t^c-X_{t,i}^c}{\sigma_t}\right)^2} \tag{7}$$

Where $\vec{X}_i^c$ is the vector of the ith observation of the vector $\vec{x}^c$ of all independent continuous variables. p stands for the number of continuous variables and $\vec{\sigma}$ for the ensemble of bandwidths of $\vec{x}^c$. We illustrated the univariate case of a probability density function $f\left(x\right)$ by means of Gaussian kernels in Figure 11. The small orange lines on the x-axis represent 6 observations of the variable x. A kernel is placed at each observation point (orange dashed lines). The sum over all kernels (see equation (5)) results in the blue line which is the estimation of $f\left(x\right)$. In order to facilitate the illustration we scaled the kernels by the number of observations.

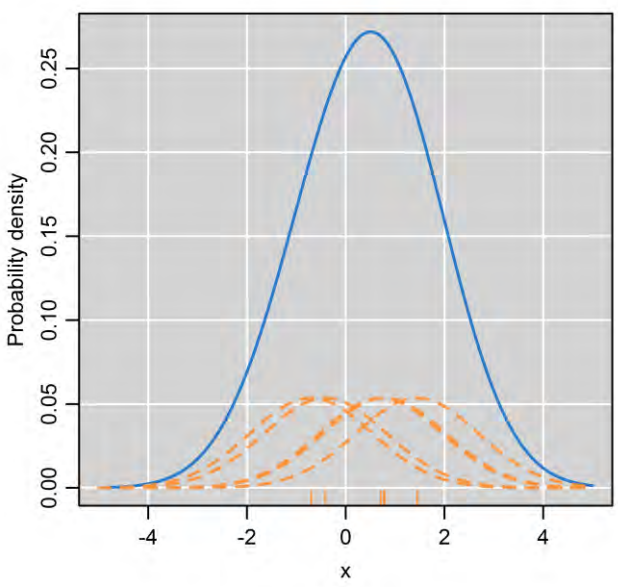


Figure 11 Probability density estimation based on Gaussian kernels. The blue line represents the sum at each x over all single kernels (orange dashed line). The observations of x are indicated as orange bars on the x-axis.

A1.2. Kernel for categorical variables

In the case of categorical variables we assumed a product $L\left(\vec{x}^d,\vec{X}_i^d,\vec{\lambda}\right)$ as described in (Kropat et al. 2015b). $\vec{\lambda}$ represents the vector of bandwidths for the different categorical variables and $\vec{X}_i^d$ is the ith observation of the vector of categorical variables $\vec{x}^d$.

In order to obtain $K\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)$, the kernel for continuous and categorical variables can simply be combined as

$$K\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)=W\left(\vec{x}^c,\vec{X}_i^c,\vec{\sigma}\right)L\left(\vec{x}^d,\vec{X}_i^d,\vec{\lambda}\right) \tag{8}$$

The final probability density estimate for continuous and categorical variables is given by

$$\hat{f}\left(y,\vec{x}\right)=\frac{1}{N}\sum_{i=1}^{N}K\left(\vec{\sigma},\vec{\lambda},\vec{z},\vec{Z}_i\right) \tag{9}$$

Where $\vec{z}=\left(y,x_1^c,\ldots,x_p^c,x_1^d,\ldots,x_k^d\right)$ and $\vec{Z}_i=\left(Y,X_{i,1}^c,\ldots,X_{i,p}^c,X_{i,1}^d\ldots,X_{i,\mathrm{k}}^d\right)$.

Given the estimate $\hat{f}\left(y,\vec{x}\right)$, $f\left(y\,|\,\vec{x}\right)$ can be estimated by

$$\hat{f}\left(y\,|\,\vec{x}\right)=\frac{\sum_{i=1}^{N}K\left(\vec{\sigma},\vec{\lambda},\vec{z},\vec{Z}_i\right)}{\int_{-\infty}^{\infty}f\left(y,\vec{x}\right)dy} \tag{10}$$

A1.3. Conditional expectation as a weighted average

(10) yields an estimate for $f\left(y\,|\,\vec{x}\right)$ in equation (3) and $f\left(y,\vec{x}\right)$ can be estimated by equation (5). Integrating over equation (3) leads to

$$E\left(y\,|\,\vec{x}\right)=\frac{\sum_{i=1}^{N}Y_iK\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)}{\sum_{i=1}^{N}K\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)} \quad \text{(Racine and Li 2004)} \tag{11}$$

$E\left(y\,|\,\vec{x}\right)$ is hence a weighted sum over all observations Y_i with weights $K\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)$. The regression parameters in this case are the bandwidths $\vec{\sigma}$ and $\vec{\lambda}$ which can be found via leave-one-out cross validation.

A2. BART

In the following we describe the principle of BART in more detail than it was done in (Kropat et al. 2015c). For further detail we refer the interested reader to the publications (Chipman et al. 1998) and (Chipman et al. 2010).

BART is an ensemble method which approximates $E\left(y\,|\,x\right)$ by averaging over several regression trees. The corresponding regression model can be formulated as

$$Y=\sum_{i=1}^{m}g\left(x;T_j,M_j\right)+\varepsilon \tag{12}$$

Where $M_j=\left\{\mu_{ij}\right\}$ represents the set of the terminal node parameters μ_{ij} and T_j the structure of the j th tree. $g\left(\cdot\right)$ is the functional implementation of a single tree and ε a normally distributed random error ($\varepsilon\sim\mathrm{N}\left(0,\sigma^2\right)$). BART differs from random forests in the manner in which each tree and the ensemble of trees is constructed. BART finds the best trees and the ensemble of trees by defining a prior for each important parameter of the tree

ensemble. Using a Metropolis-Hasting algorithm, a posterior distribution for the tree ensembles is generated. Resulting statistics like the mean or the mode of the posterior sample can finally be used to perform the prediction. The standard deviation of the posterior sample gives an estimate of the uncertainty of the prediction.

The prior specification problem is simplified by the statement of independency of the priors

$$\begin{aligned} p\left(\left(T_1,M_1\right),\ldots,\left(T_m,M_m\right),\sigma\right) &= \left[\prod_j p\left(T_j,M_j\right)\right] p\left(\sigma\right) \\ &= \left[\prod_j p\left(M_j \mid T_j\right) p\left(T_j\right)\right] p\left(\sigma\right) \end{aligned} \tag{13}$$

And

$$p\left(M_j \mid T_j\right) = \prod_i p\left(\mu_{ij} \mid T_j\right) \tag{14}$$

This means that the trees $\left(T_j,\ M_j\right)$ within one tree ensemble are independent of each other and of σ (Chipman et al. 1998).

The construction of $p\left(T_j\right)$ consists of 3 parts: The first is the definition of the probability of a node at depth $d\left(=0,1,2,\ldots\right)$ to be non-terminal

$$\alpha\left(1+d\right)^{\beta}, \alpha \in \left(0,1\right), \beta \in \left[0,\infty\right) \tag{15}$$

The second and the third are the distributions of the splitting variable assignments and the splitting rules at each non-terminal node. Both distributions are uniform. That means that each variable has equal probability to be chosen for the split at one node, and that each value of the chosen variable is equally probable to be chosen for the splitting rule.

The prior $p\left(\mu_{ij} \mid T_j\right)$ of the end node parameters consists of the conjugate normal distribution $N\left(\mu_\mu,\sigma_\mu^2\right)$. μ_μ and σ_μ are the hyperparameters of the prior. For the sake of computational simplicity the data Y is shifted and rescaled such that the prior $p\left(\mu_{ij} \mid T_j\right)$ is centered around zero $\left(\mu_\mu = 0\right)$ and $k\sqrt{m}\sigma_\mu = 0.5$.

$$\mu_{ij} \sim N\left(0,\sigma_\mu^2\right) \tag{16}$$

Where $\sigma_\mu = 0.5 / k\sqrt{(m)}$ and k is the variable for the scaling of Y and m the number of trees. A default choice of $k = 2$ is recommended. It can also be determined by cross-validation. For the number of trees m a default value of $m = 200$ is recommended (Chipman et al. 2010).

For the prior $p\left(\sigma\right)$ the inverse chi-square distribution $\sigma^2 \sim \nu\lambda / \chi_\nu^2$ is used. ν and λ are also estimated on the data (see (Chipman et al. 2010)).

To avoid exhaustive calculations, the posterior distribution

$$p\left(\left(T_1,M_1\right),\ldots,\left(T_m,M_m\right),\sigma\,|\,y\right) \tag{17}$$

can be sampled from a Markov Chain Monte Carlo (MCMC) algorithm.

Assuming that $T_{(j)}$ is the ensemble of all trees except T_j and $M_{(j)}$ the set of corresponding end node parameters, the MCMC algorithm can be implemented as a Gibbs sampler. Often the direct sampling from a joint probability distribution $f\left(x,y\right)$ is difficult in practice. To obtain a sample of $f\left(x,y\right)$ one can also perform successive draws from the conditional distribution $f\left(x\,|\,y\right)$ and $f\left(y\,|\,x\right)$ when these distributions are known. This principle is called Gibbs sampling (Casella and George 1992).

A draw from the posterior distribution in equation (17) can hence be obtained by performing successive draws of $\left(T_j,\,M_j\right)$ from

$$\left(T_j,M_j\right)|\,T_{(j)},M_{(j)},\sigma,y \tag{18}$$

with $j=1,\ldots,m$ and a following draw of σ from

$$\sigma\,|\,T_1,\ldots,T_m,M_1,\ldots,M_m,y \tag{19}$$

The draw of σ is just a draw from an inverse gamma distribution.

$\left(T_j,\,M_j\right)$ can be obtained by assuming that the $p\left(\left(T_j,M_j\right)|\,T_{(j)},M_{(j)},\sigma,y\right)$ is dependent on $\left(T_{(j)},M_{(j)},\sigma,y\right)$ via the residuals of the fit excluding the j th tree

$$R_j=y-\sum_{k\neq j}g\left(x;T_k,M_k\right) \tag{20}$$

This simplifies (18) to

$$\left(T_j,M_j\right)|\,R_j,\sigma \tag{21}$$

The posterior

$$p\left(T_j\,|\,R_j,\sigma\right)\sim p\left(T_j\right)\int p\left(R_j\,|\,M_j,T_j,\sigma\right)p\left(M_j\,|\,T_j,\sigma\right)dM_j \tag{22}$$

can be achieved in closed form, since the prior for M_j was chosen to be a conjugate prior.

The Gibbs sampler samples $\left(T_j,\,M_j\right)$ in two steps

$$T_j \mid R_j, \sigma \tag{23}$$

$$M_j \mid T_j, R_j, \sigma \tag{24}$$

The draw of T_j is realized by a Metropolis Hastings algorithm described in (Chipman et al. 1998) and M_j is drawn from a normal distribution. Having obtained a sample of several ensembles of $\left(T_j, M_j\right)$ from the posterior distribution, predictions of y can simply be obtained by calculating the mean over the sample.

A3. SRC measurements

In the frame work of a master thesis in collaboration with the University of Fribourg we performed a local soil gas measurement campaign in the town of La-Chaux-de-Fonds to evaluate SRC as a possible predictor for IRC (Wa MBengi and Collet 2013). For this purpose we measured the SRC in the vicinity of 53 buildings were long term IRC measurements had been carried out previously. In order to obtain a possibly even distribution of SRC measurements we chose around 150 buildings via random declustering from around 1700 existing IRC measurements in La-Chaux-de-Fonds. Due to time limitations we only completed SRC measurements at 53 of the 150 buildings. In Figure 12 the spatial distribution of the SRC measurements is shown with the corresponding lithological classes.

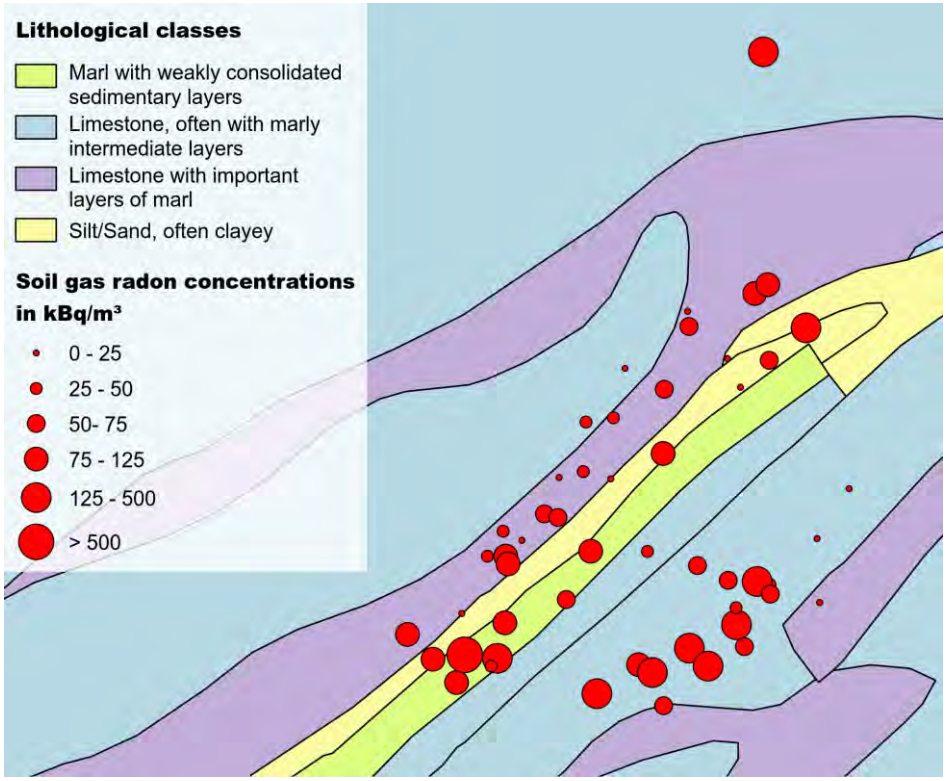


Figure 12 SRC measurements in La Chaux-de-Fonds with the corresponding lithological units

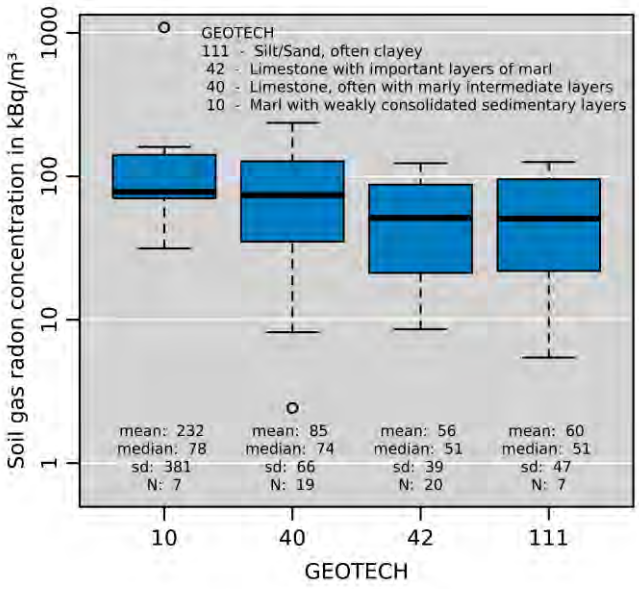


Figure 13 Distributions of the SRC for different lithological classes. The description of the GEOTECH Id is given in the legend

The average of the 53 SRC measurements resulted to around 90 kBq/m^3. The SRC distributions for different lithological classes are shown in Figure 13. The class "Marl with weakly consolidated sedimentary layers" appears to have slightly higher SRC. In particular in this class we observed an extreme value of more than 1000 kBq/m^3. However a Kruskal-Wallis analysis of variance resulted to a *p*-value of 13% and a χ^2= 5.63. This finding does not support the hypothesis of different SRC medians for different lithological classes. Nevertheless, we performed 8 SRC measurements in the Swiss Plateau resulting to an average of around 35 kBq/m^3, which is considerably lower than the SRC average obtained in La-Chaux-de-Fonds. We expect therefore regional SRC trends within Switzerland. In order to make final conclusions on this issue more SRC measurements are necessary. This was however not in the scope of this work.

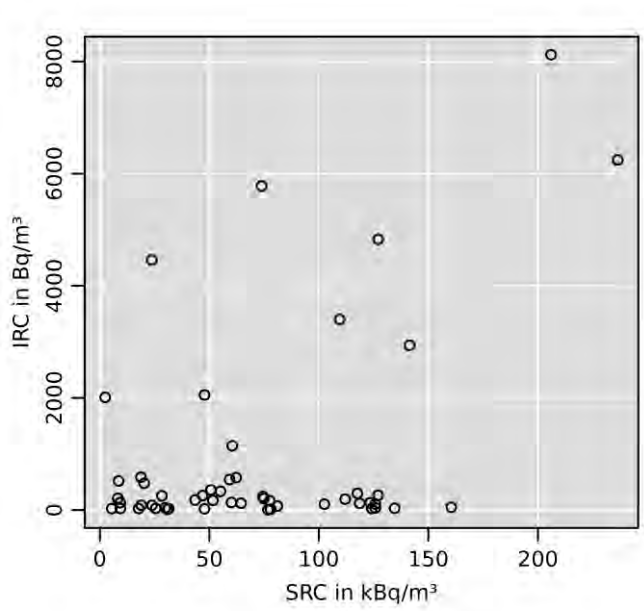


Figure 14 IRC versus SRC. Each SRC was measured in the vicinity of the building of the corresponding IRC measurement

Figure 14 shows the IRC measurements versus the SRC measurements for each building. Table 2 presents the mean and median IRC for SRC measurements below and above 100 kBq/m^3. The difference in mean IRC appears to be substantial. For the difference in median IRC an inverse relationship can be observed. This is a typical observation for log-normal distributions were the mean value can be strongly influenced by extreme values.

	Mean IRC (Bq/m^3)	Median IRC (Bq/m^3)
<= 100 kBq/m^3	594	176
> 100 kBq/m^3	1583	130

Table 2 Mean and median IRC for SRC below and above 100 kBq/m^3 measurements

A linear regression analysis between the log transformed IRC and the log transformed SRC yielded a *p*-Value of 64%. We considered the SRC measurements above 1000 kBq/m^3 as outlier and did not include it in this analysis. The regression analysis does reveal a significant relationship between IRC and SRC. In order to study the SRC as potential IRC predictor further measurements SRC have to be carried out in different regions of Switzerland. This part of our work aimed at a first insight and understanding of possible relationships between SRC and IRC.

References

Appleton JD, Miles JCH. A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk. J Environ Radioact. 2010 Oct;101(10):799–803.

Bossew P, Dubois G, Tollefsen T. Investigations on indoor Radon in Austria, part 2: Geological classes as categorical external drift for spatial modelling of the Radon potential. J Environ Radioact. 2008 Jan;99(1):81–97.

Burkart W, Wernli C, Brunner HH. Matched Pair Analysis of the Influence of Weather-Stripping on Indoor Radon Concentration in Swiss Dwellings. Radiat Prot Dosimetry. 1984 Jan 1;7(1-4):299–302.

Casella G, George EI. Explaining the Gibbs Sampler. Am Stat. 1992 Aug 1;46(3):167–74.

Cherkassky V, Mulier FM. Support Vector Machines. Learn Data Concepts Theory Methods. John Wiley & Sons; 2007. p. 404–66.

Chipman HA, George EI, McCulloch RE. Bayesian CART Model Search. J Am Stat Assoc. 1998 Sep;93(443):935–48.

Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2010 Mar;4(1):266–98.

Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, et al. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ. 2005 Jan 29;330(7485):223.

Dubois G. An overview of radon surveys in Europe. European Commission; 2005.

FOPH. Comparison of the Measurement Results Obtained by Radosure with Other Detectors. Federal Office of Public Health; 2011.

FOPH. Radon risk in Switzerland [Internet]. Federal Office of Public Health FOPH; 2013. Available from: http://www.bag.admin.ch/themen/strahlung/00046/11952/index.html?lang=en

Gunby JA, Darby SC, Miles JCH, Green BMR, Cox DR. Factors affecting indoor radon concentrations in the United Kingdom. Health Phys. 1993;64(1):2–12.

Von Gunten HR, Surbeck H, Rössler E. Uranium Series Disequilibrium and High Thorium and Radium Enrichments in Karst Formations. Environ Sci Technol. 1996 Mar 1;30(4):1268–74.

Hauri DD, Huss A, Zimmermann F, Kuehni CE, Röösli M. A prediction model for assessing residential radon concentration in Switzerland. J Environ Radioact. 2012 Oct;112:83–9.

Hunter N, Muirhead CR, Miles JCH, Appleton JD. Uncertainties in radon related to house-specific factors and proximity to geological boundaries in England. Radiat Prot Dosimetry. 2009 Jan 8;136(1):17–22.

Kemski J, Klingel R, Siehl A, Valdivia-Manchego M. Radon risk prediction in Germany based on gridded geological maps and soil gas measurements. 8th Int Work Geol Asp Radon Risk Mapp Prague Czech Repub [Internet]. 2006 [cited 2014 Sep 9]. p. 26–30. Available from: http://www.kemski-bonn.de/downloads/Prag2006_Kemski_RnRiskPred.pdf

Kemski J, Klingel R, Siehl A, Valdivia-Manchego M. From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. Environ Geol. 2009 Feb 1;56(7):1269–79.

Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, et al. Residential Radon and Risk of Lung Cancer: A Combined Analysis of 7 North American Case-Control Studies. Epidemiology. 2005;16(2):137–45.

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios M, et al. Major influencing factors of indoor radon concentrations in Switzerland. J Environ Radioact. 2014 Mar;129:7–22.

Kropat G, Baechler S, Bailat C, Barazza F, Bochud F, Damet J, et al. CALIBRATION OF THE POLITRACK® SYSTEM BASED ON CR39 SOLID STATE NUCLEAR TRACK DETECTORS FOR PASSIVE INDOOR RADON CONCENTRATION MEASUREMENTS. Submitted to Radiat Prot Dosim. 2015 a;

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios (Gruson) M, et al. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland. Sci Total Environ. 2015 b Feb 1;505:137–48.

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios (Gruson) M, et al. Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. To be submitted to Environ Int. 2015 c;

Lubin JH, Wang ZY, Boice JD, Xu ZY, Blot WJ, De Wang L, et al. Risk of lung cancer and residential radon in China: Pooled results of two studies. Int J Cancer. 2004 Mar 10;109(1):132–7.

Menzler S, Piller G, Gruson M, Rosario AS, Wichmann H-E, Kreienbrock L. POPULATION ATTRIBUTABLE FRACTION FOR LUNG CANCER DUE TO RESIDENTIAL RADON IN SWITZERLAND AND GERMANY: Health Phys. 2008 Aug;95(2):179–89.

Neznal M. Permeability as an important parameter for radon risk classification of foundation soils. Ann Geophys [Internet]. 2005 [cited 2013 Dec 10]; Available from: http://www.earth-prints.org/handle/2122/894

Parriaux A, Turberg P, Gandolla M. Géologie et santé publique: contamination au radon. 2010 [cited 2013 Dec 10]; Available from: https://www.abms.com.br/site/links/radonio_pdf1.pdf

Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data. J Econ. 2004 Mar;119(1):99–130.

Racine JS. Nonparametric Econometrics: A Primer. Found Trends® Econ. 2008;3(1).

Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health. 1998 Jan 1;88(1):15–9.

Sajó-Bohus L, Greaves ED, Pálfalvi J, Urbani F, Merlo G. Radon concentration measurements in Venezuelan caves using SSNTDS. Radiat Meas. 1997;28(1–6):725–8.

Schön JH. Natural Radioactivity of Rocks. Phys Prop Rocks Fundam Princ Petrophysics. Elsevier; 2004. p. 107 – 132.

Schuler C, Crameri R, Burkart W. Assessment of the Indoor Rn Contribution of Swiss Building Materials. Heal Phys March 1991. 1991;60(3):447–51.

Smethurst MA, Strand T, Sundal AV, Rudjord AL. Large-scale radon hazard evaluation in the Oslofjord region of Norway utilizing indoor radon concentrations, airborne gamma ray spectrometry and geological mapping. Sci Total Environ. 2008 Dec 15;407(1):379–93.

Specht DF. A general regression neural network. Neural Networks IEEE Trans. 1991;2(6):568–76.

Vaupotič J, Kobal I, Križman M. Background outdoor radon levels in Slovenia. Nukleonika. 2010;55(4):579–82.

Wa MBengi H, Collet C. Mesure et cartographie de la concentration de radon dans le sol en suisse romande, cas d'étude : La Chaux-de-Fonds. 2013.

Wepf B. Geologischer Führer der Schweiz - Guide géologique de la suisse. Basel: Société géologique suisse; 1934.

Xu R, Wunsch D. Clustering. John Wiley & Sons; 2008.

Zeeb H, Shannoun F. Health effects of radon. WHO Handb Indoor Radon Public Heal Perspect [Internet]. Geneva: World Health Organization (WHO); 2009 [cited 2013 Dec 10]. p. 3–20. Available from: http://www.who.int/iris/handle/10665/44149

Papers

The following papers have been attached to this document:

Kropat G, Baechler S, Bailat C, Barazza F, Bochud F, Damet J, et al. CALIBRATION OF THE POLITRACK® SYSTEM BASED ON CR39 SOLID STATE NUCLEAR TRACK DETECTORS FOR PASSIVE INDOOR RADON CONCENTRATION MEASUREMENTS. Accepted for publication in Radiat Prot Dosim. 2015 a;

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios M, et al. Major influencing factors of indoor radon concentrations in Switzerland. J Environ Radioact. 2014 Mar;129:7–22.

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios (Gruson) M, et al. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland. Sci Total Environ. 2015 b Feb 1;505:137–48.

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios (Gruson) M, et al. Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. To be submitted to Environ Int. 2015 c;

Radiation Protection Dosimetry (year), Vol. 0, No. 0, pp. 0–0
DOI: 10.1093/rpd/nc0000

CALIBRATION OF THE POLITRACK® SYSTEM BASED ON CR39 SOLID STATE NUCLEAR TRACK DETECTORS FOR PASSIVE INDOOR RADON CONCENTRATION MEASUREMENTS

G. Kropat[1,*], S. Baechler[2], C. Bailat[1], F. Barazza[2], F. Bochud[1], J. Damet[1], N. Meyer[1], M. Palacios (Gruson)[2], G. Butterweck[3]
[1]Institute of Radiation Physics, Lausanne, Switzerland
[2]Swiss Federal Office of Public Health, Berne, Switzerland
[3] Paul Scherrer Institute, Villigen, Switzerland

Received month date year, amended month date year, accepted month date year

Swiss national requirements for measuring radon gas exposures demand a lower detection limit of 50 kBq h m^{-3}, representing the Swiss concentration average of 70 Bq m^{-3} over a one-month period. A solid state nuclear track detector (SSNTD) system (Politrack, Miam, Italy) has been acquired to fulfill these requirements. This work is aimed at the calibration of the Politrack system with traceability to international standards and the development of a procedure to check the stability of the system. 275 SSNTDs were exposed to 11 different radon exposures in the radon chamber of the Secondary Calibration Laboratory at the Paul Scherrer Institute, Switzerland. The exposures ranged from 50 kBq h m^{-3} to 15000 kBq h m^{-3}. For each exposure of 20 detectors, 5 SSNTDs were used to monitor possible background exposures during transport and storage. The response curve and the calibration factor of the whole system were determined using a Monte Carlo fitting procedure. A device to produce CR39 samples with a reference number of tracks using an Am-241 source was developed for checking the long term stability of the Politrack system. The characteristic limits for the detection of a possible system drift were determined following ISO Standard 11929.

INTRODUCTION

Radon is a radioactive gas that is known to be the most important cause of lung cancer after smoking. Most of the radon exposure of the public takes place in closed environments at home or at work (Effects of Ionizing Radiation, 2009). To effectively manage radon risk, radon concentrations have to be estimated via reliable measurements. The most reliable radon measurements are long term measurements that integrate between 3 to 12 months of exposure (WHO, 2009). There is a variety of different devices to measure radon concentrations. SSNTD have proven to be cost effective and therefore particularly suitable for large scale national radon surveys. However practical questions arise when putting into place a SSNTD device: How to keep track of possible drifts of the system in order to warrant a stable quality of results and how do radon concentrations during transport and storage influence the result of SSNTD readings?

INSTRUMENTATION AND METHODS

Solid state nuclear track detectors (SSNTD)

Heavily ionizing particles leave trails of damage on most insulating materials (Fleischer et al., 1965). These trails can be made visible with a microscope by etching the material with a chemical reagent which attacks preferentially the damaged trails. Detectors that work on this principle are called "Solid State Nuclear Track Detectors (SSNTD)". A SSNTD for α-particles can be build with the polymer CR39 as detector material. We used CR39 films of a thickness of 1mm and a size of 25 *mm* x 25 *mm* and etched them, according to the manufacturer recommendation, in a 6.25M NaOH bath for 1 hour at 98°*C* after exposure.

SSNTD reader system

After etching, the tracks of the α-particles can be counted with a light microscope. The Politrack reader system consists of a microscope equipped with a CCD camera and a 4x magnifying objective that can be moved in z direction and a SSNTD stage that can be moved in xy direction. The images are sent to a computer via firewire and analyzed by a program written in LabView by the manufacturer. The LabView software returns the number of counted tracks per cm^2 and the sum of the area of all detected tracks. Each α–track is detected separately by a pattern recognition algorithm. At higher exposures the probability increases that two or more α–tracks overlap. This leads to a saturation effect. The algorithm is not capable to distinguish between two overlapping tracks. Since the

*Corresponding author: georg.kropat@chuv.ch

probability to obtain overlapping tracks is difficult to model, track overlapping was corrected by the empirical formula

$$Tr_{cor} = \frac{Tr_{net}}{1 - m \cdot A_{Tr}} \quad (1)$$

where m is an empirical factor to be determined by a fit procedure and Tr_{net} is the difference between the counted tracks and the background. A_{Tr} is the area of all detected tracks. The area was corrected by the mean area measured on the background SSNTDs.

CALIBRATION

We exposed 275 SSNTDs at 11 different exposures in the radon chamber at the Paul Scherrer Institute in Villigen, Switzerland (Schuler, 1998). The exposures took place consecutively at 3 different concentration levels: 1000 *Bq m*$^{-3}$, 5000 *Bq m*$^{-3}$ and 20000 *Bq m*$^{-3}$. The different exposures within each concentration level were realized by different exposure times resulting in the exposures: 46, 96, 289, 502, 987, 1046, 2077, 3112, 5230, 10506 and 15520 *kBq h m*$^{-3}$, with an expanded uncertainty ($k=2$) ranging between 1.4% and 2.3%. After exposure, the SSNTDs were shipped back to the Institute of Radiation Physics, Lausanne and stored until the exposure of the last SSNTDs finished. The SSNTDs were etched at 3 batches. In order to keep track of background exposure during transport and storage each exposure level was accompanied by 5 transport SSNTDs.

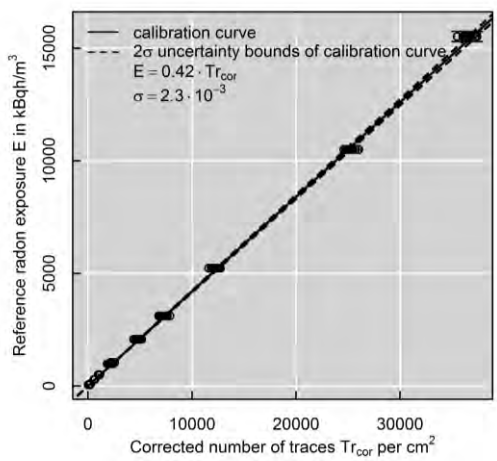


Figure 1. Calibration curve of reference exposures versus corrected tracks.

Taking into account the correction of equation (1), the following model can be fitted to the background corrected track counts Tr_{net}, A_{Tr} and the reference exposure E to obtain the empirical factors c and m, where c is the calibration factor.

$$E = c \frac{Tr_{net}}{1 - m \cdot A_{Tr}} \quad (2)$$

To take into account the uncertainty due to random errors of the exposure and the track reading, we used a Monte Carlo fit procedure according to (JCGM, 2008). For this purpose we repeated least squares fitting 10000 times by adding each time a random error ε_{Exp} to each exposure value with

$$\varepsilon_{Exp} \in N(0, u_{Exp}) \quad (3)$$

u_{Exp} is the standard uncertainty estimated for the reference exposure of each SSNTD.

Furthermore, for each of the 10000 repeated fittings we added a random error ε_{Tr} to the track counts Tr_{net} in order to account for the reading uncertainty with

$$\varepsilon_{Tr} \in N(0, u_{Tr}) \quad (4)$$

u_{Tr} depends on the number of tracks of each readout and was given by the manufacturer.

We took the arithmetic mean and the standard deviation of c and m over the 10000 fitting results as best estimates for the expected values and the corresponding standard uncertainties of c and m.

Figure 1 shows the calibration curve that we fitted by the Monte Carlo fit procedure. For simplicity, we plotted the exposure E versus Tr_{cor}. The uncertainty u_{Tr} due to read out of the system was given by the vendor for each SSNTD read out, and ranged from 0.7% to 13%. The expanded uncertainties ($k=2$) of the reference exposures, corresponding to u_{Exp}, ranged between 1.4% and 2.3%.

The Monte Carlo fit procedure yielded a calibration factor c = 420.0 10^{-3} *kBq h m*$^{-3}$ *cm*^{2} with an uncertainty of u_c = 2.3·10^{-3} *kBq h m*$^{-3}$ *cm*^{2} (0.6%) and a track correction factor m = 5.55 *cm*$^{-2}$ with a standard uncertainty of u_m = 0.07 *cm*$^{-2}$ (2.7%). Hence the overall uncertainty of c an m results in 0.93%. The uncertainty u_m contributes only to a small amount, since A_{Tr} is generally very small. 0.93% indicates a small overall uncertainty for the calibration factor and is due to the fact that the uncertainty is attributed to random errors, which have a very small influence on the final fit result, since we used a relatively large number of SSNTDs for the calibration. For simplicity we did not assume an intercept in equation (1). The Monte Carlo fit procedure would however allow to calculate a covariance between the intercept and the slope of the calibration curve which could be used to further improve the uncertainty estimation.

We assumed a maximum u_{Exp} of 1.15% as systematic standard uncertainty contribution from the calibration of the radon chamber.

LONG TERM STABILITY MONITORING

To monitor possible drifts of the system we developed and built a device to produce reference CR39 films at a reproducible exposure level. For this purpose we used

an Am-241 source with an activity of 320 Bq. To control the exposure time of the Am-241 source we shielded the source with an automatic shutter.
The distributions of track counts for several reference exposure series are presented as boxplot in Figure 2. The boxplot represents the median, first and third quartile. The whiskers represent 1.5 times the interquartile range of the distribution. For further readings on boxplots we refer to (Diez et al., 2012). Each batch has been etched separately. We annotated the basic statistical indicators for each batch on the boxplot. The uncertainty of the track counts is attributable to the randomness of the Am-241 decay, to variation in the etching procedure as well as to read out uncertainties.

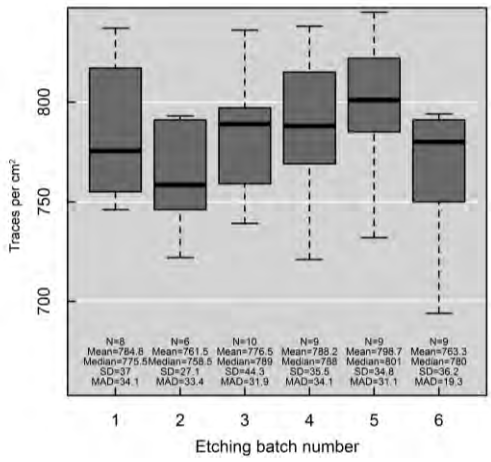


Figure 2. Track count distributions of separately etched SSNTD batches used for reference exposure series.

We determined the decision threshold for potential drifts according to ISO 11929 on international standards (ISO, 2010). The decision threshold Tr^*_{net} can be determined by the equation

$$Tr^*_{net} = q_{1-\alpha}\,\tilde{u} \qquad (5)$$

Where $\tilde{u}$ is the standard uncertainty of the average number of tracks of the irradiated reference SSNTDs and $q_{1-\alpha} = 2.33$ the (1-α)−quantile of the standardized normal distribution. We chose $\alpha = 1\%$ and assumed the relative decision threshold Tr^*_{net} as uncertainty contribution resulting from long term stability monitoring of the system. For the irradiation of 10 reference SSNTDs with the Am-241 source for 20 *s*, we observed a mean of ~780 cm^{-2} with a standard uncertainty of *14* cm^{-2}. This results in a decision threshold for a possible drift of 32 cm^{-2} and hence gives an uncertainty contribution on Tr_{net} of 4% due to long term stability monitoring.

OVERALL UNCERTAINTY OF THE SYSTEM

Table 1 shows the uncertainty budget of the estimated exposure *E*. The summation in quadrature leads to an expanded combined uncertainty (*k=2*) of around 9%

Uncertainty component	**Relative standard uncertainty**
Long term stability monitoring	4%
Reference exposure values	1.15%
Monte Carlo fit procedure	0.93%

Table 1. Uncertainty budget of estimated exposure *E*

By far the largest uncertainty contribution results from the long term stability monitoring with 4%. This is reasonable, since the etching procedure is a process that is difficult to control.

LEAKAGE OF SSNTD WRAPPINGS

To control the air tightness of the SSNTD packaging, we exposed 40 SSNTD welded up in plastic bags to the highest exposure of 15520 $kBq\ h\ m^{-3}$. Some of the weld seams showed small defects. To keep control of a possible background contribution we used 10 SSNTDs welded up in plastic bags that were exposed simultaneously to ambient air. After exposure we distinguished between SSNTDs packed in plastic bags with defects (Weld defect), with no defects (No weld defect) and background SSNTDs (Transport SSNTD) and compared their distributions via boxplots.
The results are shown in Figure 3. A Kruskal-Wallis analysis of variance comparing the medians of the 3 groups yielded a $\chi^2 = 1.93$ and a *p*-value of 38%.This result does not support the hypothesis of leakage of detector wrappings. This holds for regular weld seams as well as for weld seams exhibiting little defects. Since we observed 41 cm^{-2} on the 10 transport SSNTDs that were exposed to ambient air in plastic bags, we assumed a general background correction of all readings of 40 cm^{-2} for the comparison measurements described in the next section.

COMPARISON MEASUREMENT

In order to compare our calibration with commercially available SSNTDs we carried out a comparison study with our SSNTDs and SSNTDs well established on the market (Landauer Nordic, former Gammadata). For this purpose we distributed 30 SSNTDs of each type pairwise in Swiss schools and measured the radon concentration for about 3 months.
The comparison of measurements with Politrack SSNTDs with our calibration factor and Gammadata SSNTDs shows a mean difference of 14.1 $kBq\ h\ m^{-3}$ (Figure 4). That corresponds to a concentration of 6.5 $Bq\ m^{-3}$ for a measurement over 90 days. This is indicating a good accordance of both systems.

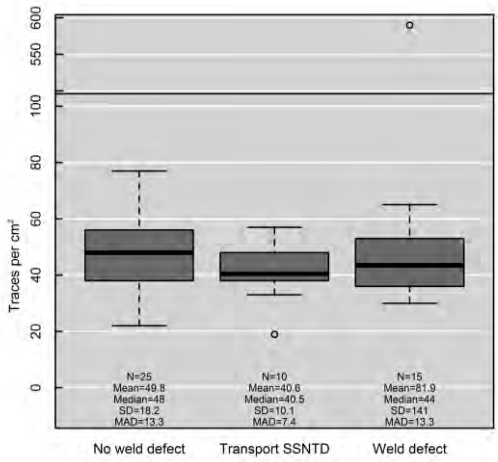


Figure 3. Boxplot of track count distributions of SSNTDs that were packed in plastic bags with and without weld seam defects. The boxes represent the median, first and third quartile. The whiskers represent 1.5 times the interquartile range of the distribution. Data points that lie outside of the whiskers are drawn on the plot as extreme values. The plastic bags with the SSNTDs were exposed to 15520 *kBq h m*$^{-3}$. The track count distribution of the SSNTDs that were only exposed to background radon concentrations is indicated as "Transport SSNTD".

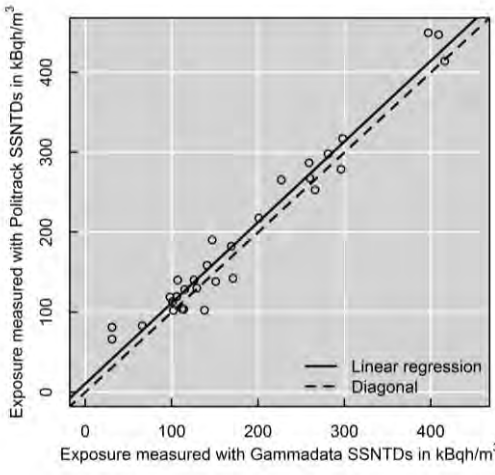


Figure 4. Comparison of Politrack with Gammadata SSNTDs.

CONCLUSION

We calibrated and characterized a SSNTD reader system for the measurement of indoor radon concentrations. In addition to that, we developed a procedure to keep track of long term stability of the system. The system is ready for routine use.

We achieved an overall uncertainty of around 9% for the system. A comparison in school buildings with SSNTDs of the manufacturer Landauer Nordic yielded consistent results with our calibration.

Our results indicate that the weld of the SSNTD transport plastic bags is sufficiently air tight for shipping. After an exposure of packaged detectors to about 15000 *kBq h m*$^{-3}$ during 31 days we did not find significantly higher mean track counts than for detectors that were only exposed to background irradiation. This also holds for plastic bags having defects in the weld seam.

ACKNOWLEDGEMENT

The presented work was partly funded by the Swiss Federal Office of Public Health.

REFERENCES

1. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR). *Effects of Ionizing Radiation*. *UNSCEAR 2006 Report to the General Assembly*. United Nations. New York (2006).
2. World Health Organization. *WHO Handbook on Indoor Radon: A Public Health Perspective*. WHO, ISBN 978 92 4 154767 3. http://www.who.int/ionizing_radiation/env/radon/en/index1.html (2009).
3. Fleischer, R.L., Price, P.B., Walker, R.M. *Solid-State Track Detectors: Applications to Nuclear Science and Geophysics.* Annu. Rev. Nucl. Sci. 15, 1–28. doi:10.1146/annurev.ns.15.120165.000245 (1965)
4. Schuler, Ch. *Das Referenzlabor für Radongas-Konzentrationsmessungen am PSI*. PSI-Bericht Nr. 98-08. ISSN 1019-0643. Paul Scherrer Institut. Villigen, Switzerland (1998)
5. International Standardization Organisation (ISO). *ISO 11929:2010 Determination of the characteristic limits (decision threshold, detection limit and limits of the confidence interval) for measurements of ionizing radiation -- Fundamentals and application*. ISO. Geneva. (2010)
6. Joint Committee for Guides in Metrology (JCGM). *Evaluation of Measurement Data–Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement"–Propagation of distributions using a Monte Carlo method.* JCGM 101 (2008).
7. Diez, D.M., Barr, C.D., Cetinkaya-Rundel, M. *Inference for categorical data*, OpenIntro Statistics. CreateSpace independent publishing platform, 263–314, (2012).

Journal of Environmental Radioactivity 129 (2014) 7–22

Contents lists available at ScienceDirect

Journal of Environmental Radioactivity

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

Major influencing factors of indoor radon concentrations in Switzerland

Georg Kropat [a,*], Francois Bochud [a], Michel Jaboyedoff [c], Jean-Pascal Laedermann [a], Christophe Murith [b], Martha Palacios [b], Sébastien Baechler [a]

[a] Institute of Radiation Physics, University Hospital Center of Lausanne (CHUV), Rue du Grand-Pré 1, 1007 Lausanne, Switzerland
[b] Swiss Federal Office of Public Health, Schwarzenburgstrasse 165, 3003 Berne, Switzerland
[c] Center for Research on Terrestrial Environment, University of Lausanne, Quartier UNIL-Mouline, 1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 19 August 2013
Received in revised form
16 November 2013
Accepted 22 November 2013
Available online

Keywords:
Indoor radon
Univariate analysis
Building characteristics
Outdoor temperature
Lithology

ABSTRACT

Purpose: In Switzerland, nationwide large-scale radon surveys have been conducted since the early 1980s to establish the distribution of indoor radon concentrations (IRC). The aim of this work was to study the factors influencing IRC in Switzerland using univariate analyses that take into account biases caused by spatial irregularities of sampling.
Methods: About 212,000 IRC measurements carried out in more than 136,000 dwellings were available for this study. A probability map to assess risk of exceeding an IRC of 300 Bq/m^3 was produced using basic geostatistical techniques. Univariate analyses of IRC for different variables, namely the type of radon detector, various building characteristics such as foundation type, year of construction and building type, as well as the altitude, the average outdoor temperature during measurement and the lithology, were performed comparing 95% confidence intervals among classes of each variable. Furthermore, a map showing the spatial aggregation of the number of measurements was generated for each class of variable in order to assess biases due to spatially irregular sampling.
Results: IRC measurements carried out with electret detectors were 35% higher than measurements performed with track detectors. Regarding building characteristics, the IRC of apartments are significantly lower than individual houses. Furthermore, buildings with concrete foundations have the lowest IRC. A significant decrease in IRC was found in buildings constructed after 1900 and again after 1970. Moreover, IRC decreases at higher outdoor temperatures. There is also a tendency to have higher IRC with altitude. Regarding lithology, carbonate rock in the Jura Mountains produces significantly higher IRC, almost by a factor of 2, than carbonate rock in the Alps. Sedimentary rock and sediment produce the lowest IRC while carbonate rock from the Jura Mountains and igneous rock produce the highest IRC. Potential biases due to spatially unbalanced sampling of measurements were identified for several influencing factors.
Conclusions: Significant associations were found between IRC and all variables under study. However, we showed that the spatial distribution of samples strongly affected the relevance of those associations. Therefore, future methods to estimate local radon hazards should take the multidimensionality of the process of IRC into account.

1. Introduction

Radon is a naturally occurring radioactive noble gas that is a decay product of uranium. The decay products of radon are known to cause lung cancer through their accumulation in the lungs. In outdoor air, radon is strongly diluted (Vaupotic et al., 2010). However, in environments with low air exchange, such as buildings, radon concentrations are generally higher and can lead to a considerable health threat. In Switzerland, about 230 cancer deaths per year are attributable to radon (Menzler et al., 2008).

Radon concentrations in houses originate from the underlying geology, building materials and domestic water supplies. The geological parameters controlling IRC are mainly the uranium content of the ground and its permeability (Johner and Surbeck, 2001). In Switzerland, three main different geological areas are generally considered. The Alps in the south of Switzerland are

* Corresponding author. Tel.: +41 21 314 82 96; fax: +41 21 314 82 99.
E-mail address: georg.kropat@chuv.ch (G. Kropat).

0265-931X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jenvrad.2013.11.010

dominated by granites and gneisses. Some of the variants of these rocks are known to be rich in uranium (Schön, 2004). In northwest Switzerland, the landscape is formed by the Jura Mountains, which are characterized by a high abundance of carbonate rock. Carbonate rock is subject to strong weathering, also called karstification. The karstification of carbonate rock results in a highly permeable cave system, which facilitates the transport of radon gas. Karstic regions are therefore known to be radon prone areas (Vaupotic et al., 2001). The Swiss Plateau is the lower part of the country. This area is located between the Jura Mountains and the Alps, is covered mainly by quaternary sediments containing partially glacial deposits originating from the Alps and is not considered an area with high radon potential. Those quaternary deposits overlay the molasse sedimentary rock (mainly detrital sediments, sandstones, shales, etc.) (Trümpy, 1980). Most of Switzerland's population resides along the Swiss Plateau.

Apart from geology, IRC are subject to several other variables, a reality which has made it difficult to develop reliable predictive models up to now. These variables can be grouped into 3 categories: spatial variability (geology, lithology, pedology) (Bossew et al., 2008; Cinelli et al., 2011; Friedmann and Bossew, 2010; Ielsch et al., 2010; Kemski et al., 2009; Miles and Appleton, 2005; Tapia et al., 2006), temporal variability (meteorology, anthropogenic influences) (Bossew and Lettner, 2007; Burke et al., 2010; Denman et al., 2007; Groves-Kirkby et al., 2006; Miles, 2001) and architectural characteristics of the structures concerned (building age, floor level, foundation, building material, building type, room type of measurement) (Friedmann, 2005; Friedmann and Groeller, 2010; Girault and Perrier, 2012; Kemski et al., 2009).

The aim of this work was to identify the relevant factors influencing IRC in Switzerland. For this purpose, univariate analyses of IRC were performed for each variable under study and maps showing the density of IRC measurement were computed for each class of variable in order to account for potential spatial biases. Using univariate analyses, we explore separately the effect of each variable on IRC while it is known that many variables may contribute to determine IRC. However, univariate analyses are more easily interpretable and may help to understand the major controls of IRC.

2. Data and methods

2.1. Data

The IRC data used in this study originates from the radon database of the Swiss Federal Office of Public Health (FOPH). The database consists of 211,714 measurements carried out in 136,401 Swiss dwellings.

The sampling strategy of the IRC data used in this study changed over the last 30 years. Initially the criterion was to obtain a minimal number of randomly measured buildings in each municipality. This strategy changed towards sampling of houses with potentially high IRC values. However, cantons with radon-prone areas tended to be more active with respect to IRC sampling.

2.1.1. Measurement characteristics

The measurements were taken with passive electret or alpha track detectors (Kotrappa et al., 1990; Nikolaev and Ilić, 1999). The detectors were sent to homeowners, who then set them out to expose them over a time period of about 3 months. Homeowners were asked to fill in a questionnaire containing details about the concerned building and the measurement conditions. We chose only those measurements which were taken in the basement, the ground-, the first- and the second floor. It was further recorded whether the room where the measurement was taken was inhabited during the measurement period. Finally, starting and ending date of the radon detector exposition were indicated. All measurements were taken between 1981 and 2012.

2.1.2. Building characteristics

The questionnaire also asked for information about the building characteristics. The relevant variables are the geographical coordinates in the Swiss geographical coordinate system (CH1903) and the building type. The database contains 9 different types of building. For convenience, we grouped them into four major types: "Apartment", "Detached House", "Farm" and "School". Furthermore, the database contains information about the type of foundation of the corresponding building. We limited our analysis to cases in which the type of foundation was uniquely indicated, which results in the three following types: "Concreted", "Concreted afterwards" and "Earth foundation". Finally, we analyzed IRC with respect to the year of the building's construction.

2.1.3. Outdoor temperature data

The Federal Office of Meteorology and Climatology "MeteoSwiss" provides access to the daily mean outdoor temperatures at 125 stations evenly distributed over Switzerland for the last decades. We downloaded the daily mean outdoor temperatures for the last 30 years (MeteoSwiss, 2013).

2.1.4. Lithological data

The lithological data we used in this study originate from the map "Lithologisch-petrografische Karte der Schweiz-Lithologie-Hauptgruppen 1:500,000" (SGTK, 2000). The map is vectorized, on a scale of 1:500,000 and consists of 70 lithological classes.

2.2. Data preprocessing

2.2.1. Coordinate corrections

The geographical coordinates of each building in the radon database of the FOPH were often not reliably indicated by the building-owners. Most of the buildings are registered in the central database of the Swiss Federal Statistical Office (FSO). This building registry provides a unique building ID with which the exact coordinate of each building can be determined. For those analyses for which the spatial distributions of the measurements were relevant we included all buildings for which this building ID was available. The altitude above sea level of each building was sampled from the digital elevation model "DHM25" which is provided by the Federal Office of Topography Swisstopo. This digital elevation model has a resolution of 25 m.

2.2.2. Random declustering and fractal dimension

Due to different cantonal sampling strategies and local differences of population density in Switzerland, the measurements of IRC in the database of the FOPH exhibit a strong spatial clustering. This leads to biased estimates especially in the estimation of IRC characteristics of geological units. To reduce this bias, we applied random declustering (Kanevski and Maignan, 2004) by creating a grid of 500 m $\times$ 500 m over all of Switzerland and by randomly sampling 6 houses in each grid cell. The degree of clustering of the point set was measured by its fractal dimension resulting from sandbox counting (Kanevski and Maignan, 2004).

2.2.3. Attribution of IRC to buildings

In about 87% of the considered houses one measurement was taken in inhabited rooms. 2 measurements of inhabited rooms were available in 10% of the houses. With the exception of the floor level analysis, we chose the maximum IRC in inhabited rooms as the unique IRC for each house. This resulted in a total of 32,151 measurements. When certain variables were unknown (e.g.

unknown year of construction), we excluded the measurement from the analysis of the corresponding variable.

2.2.4. Distribution of IRC

Many authors refer to the fact that IRC follows a lognormal distribution (Bossew, 2010; Nero et al., 1986; Tuia and Kanevski, 2008). However, Janssen and Stebbings (1992) found that the log-transformed IRC can more accurately be described by a gamma distribution. We explored the distribution of the log-transformed IRC by fitting gamma and normal probability density functions to it. The probability density function of the normal distribution is given by:

$$f(x;\mu,\sigma) = 1/(2\pi)\exp\Big(-(x-\mu)^2/(2\sigma)\Big) \quad (1)$$

The probability density function of the gamma distribution is given by

$$f(x;k,\theta) = x^{k-1}e^{-x/\theta}/\Big(\theta^k\Gamma(k)\Big) \quad (2)$$

where k is called the shape and θ the scale parameter. $\Gamma(k)$ is the gamma function given by

$$\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}\mathrm{d}t \quad (3)$$

The goodness of fit was measured and compared by means of Kolmogorov distances between the real and the fitted distributions.

2.3. Probability mapping

To map a local probability estimation to exceed the recommended reference level of 300 Bq/m^3 (WHO, 2009), we created a 1 km × 1 km grid covering all of Switzerland. For each pixel, we searched the 50 IRC measurements nearest to the center of the pixel within a radius of 20 km. That means the measurement sites could also be located outside of the pixel. As the probability to exceed 300 Bq/m^3 we calculated the percentage of measurements higher or equal to 300 Bq/m^3.

$$P\Big(\mathrm{Rn} > 300\mathrm{Bq/m^3}\Big) = N_{\mathrm{Rn}>300\mathrm{Bq/m^3}}/50 \quad (4)$$

$N_{\mathrm{Rn}>300\mathrm{Bq/m^3}}$ is the number of houses within the next nearest 50 neighboring houses with a radius of 20 km to the center of the pixel which exceed an IRC of 300 Bq/m^3. In regions with very low sampling density this method doesn't give meaningful results. Therefore we excluded the area of the canton Valais from the mapping.

2.4. Statistical analysis

The fact that many authors reported the distribution of IRC to be close to log-normality requires a careful determination of the expectation value and its confidence intervals.

Confidence intervals of the log-normal mean can be directly estimated by assuming that the arithmetic sample mean of a log-normal distribution follows a normal distribution for infinitely large sample sizes. However, the arithmetic sample mean of a log-normal distribution converges very slowly to a normal distribution (Land, 1972).

Therefore, we used the estimator (Shen et al., 2006):

$$\overline{v} = \exp\Big(\overline{x} + s^2/2\Big) \quad (5)$$

for the expectation value of the IRC, where $\overline{x}$ is the arithmetic sample mean of the log-transformed IRC and s^2 the sample variance of the log-transformed IRC. We calculated the approximate confidence intervals at a given α-level according to a method proposed by D. R. Cox (Land, 1972) by

$$\exp\Big(\overline{x} + s^2/2 \pm z_{1-\infty/2}\gamma\Big) \quad (6)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and

$$\gamma^2 = s^2/n + s^4/(2(n+1)) \quad (7)$$

This method was reported to give good results for moderate to large samples (Land, 1972; Zhou and Gao, 1997). We chose $\alpha = 0.05$ in order to estimate the two-sided 95% confidence intervals of the log-normal mean.

Since most of the variables considered in this study are categorical, the principal method is to calculate mean values and the 95% confidence intervals of the IRC in each class and to graphically compare the intervals of the classes. If the 95% confidence intervals of two class means do not overlap, the probability can be considered as lower than 5% that the mean of one class lies in the 95% interval of the other class. The mean values of the two classes are hence regarded as being significantly different at a 95%-level. Moreover, we carried out Kruskal–Wallis one-way analysis of variance to quantify the association between IRC and each variable. Kruskal–Wallis one-way analysis of variance tests whether the IRC medians of the classes of an independent variable are significantly different or not. The resulting test statistic is denoted as K. IRC measurements with missing attributes were sorted out.

Additionally, for each class, we computed a map representing the spatial aggregation of the number of measurements on a grid of 5 km × 5 km. This allows detecting biases caused by spatially irregular sampling. To explicitly quantify the uniformity of the spatial distribution of samples, we performed χ^2-test for the corresponding distributions of each class of each variable. The reduced statistic

$$\chi^2_{\mathrm{red}} = \frac{1}{\mathrm{df}}\sum_{i=1}^{n}(E_i - O_i)^2/E_i \quad (8)$$

gives information about the uniformity of the spatial distribution of samples. The null hypothesis is that the spatial distribution of sampling is uniform throughout the cells taken into account. $E_i = N/n$ is the expected frequency and O_i is the observed frequency of samples in cell i. N is the number of samples in the corresponding class, n the number of cells and $df = n - 1$ is the number of degrees of freedom. The p-values were obtained by calculating the probability of a χ^2-distribution with degrees of freedom df to obtain the non reduced statistic χ^2. The null hypothesis was rejected at the 0.05 significance level.

For statistical analysis and plotting we used the statistical software R (R Core Team, 2012), for spatial analyses the R packages "sp" (Bivand et al., 2008) and "RANN" (Kemp and Jefferis, 2012), and for support vector regression the R package "e1071" (Dimitriadou et al., 2012). For mapping, we used the GIS applications QGIS (Quantum GIS Development Team, 2012) and GRASS (GRASS Development Team, 2012).

2.4.1. Type of radon detector

The influence of the type of radon detector on IRC measurements was analyzed by comparing the mean values and 95% confidence intervals for track and electret detectors. We produced a map with the spatial distribution of samples for track as well as for electret detectors. To investigate the effect of spatial sampling we carried out the analysis for two spatial domains: the Swiss Plateau and whole Switzerland.

2.4.2. Type of foundation

To analyze the influence of foundation on the IRC, we calculated the mean values and the confidence intervals for the foundation classes "Concrete", "Concreted afterwards" and "Earth". For each class, we created a map representing the spatial distribution of samples. The Canton of Ticino is nearly not represented in this analysis because the foundation type has been registered in less than 7% of all measured houses in this canton.

2.4.3. Year of construction

To quantify the impact of the building age on the IRC we grouped the years of construction into 4 classes: <1900, 1901–1970, 1971–1990, >1990. For older houses the information of the year of construction can be inaccurate. This inaccuracy is compensated by the classification into 4 broad classes. For each class, we calculated the arithmetic mean and the corresponding confidence intervals. The classes of years of construction are supposed to be representative for substantial changes in building and construction norms in Switzerland.

2.4.4. Type of building

To examine the relationship between IRC and building types, we calculated mean values and confidence intervals for the following classes: schools, farms, detached houses and apartment buildings. The spatial distribution of samples was visualized by a map for each class.

2.4.5. Floor levels

Since most radon gas enters the house by coming up from the ground (Åkerblom et al., 1984) and because the basement is often the less ventilated level in a building (Dessau et al., 2005) one may expect that IRC would be higher on the lower floors. To explore this, we compared IRC on different floors within buildings. For this analysis, we chose only buildings were IRC had been measured both in the basement and on another floor during the same period of time and in inhabited rooms. Finally, we could calculate the average ratio of the IRC on the second-, first- or ground floor and the concentration in the basement. The average ratios between IRC of the corresponding floor to IRC in the basement were plotted versus the corresponding floor number.

Additionally, we investigated the correlation of measurements, which were carried out in the same building but at different floor levels. For this complementary analysis, we selected all buildings for which both measurements in the basement and on the ground floor had been carried out in inhabited rooms and during the same period of time. Furthermore, we chose all houses in which the first and second floor IRC had been measured over the same time period. For both datasets we calculated Spearman's rank correlation coefficients of the IRC between both corresponding floors and created scatter plots of the log-transformed IRC of both floors.

2.4.6. Altitude

We investigated the relationship of IRC to altitude by sub setting IRC measurements into 5 classes of altitudes in meters: "<350", "350–500", "500–700", "700–900" and ">900". We chose this classification to represent the lower part of the Canton Ticino ("<350") and to consider two more or less balanced classes for the Swiss Plateau ("350–500", "500–700") as well as for the mountainous regions of Switzerland ("700–900" and ">900"). For each class, we calculated the mean value of the measured IRC and the confidence intervals for this mean value. The mean value was plotted versus the altitude classes. Additionally, we calculated the local sample sizes for each class and presented it graphically to get an insight into the influence of the spatial distribution on the resulting mean IRC.

2.4.7. Outdoor temperatures during measurement

Using the period of measurement registered in the Swiss database for each single measurement enabled us to estimate the mean outdoor temperature for each measurement by using support vector regression. A comprehensive introduction to support vector regression is given by Smola and Schölkopf (2004) and Cherkassky and Mulier (2007). We used the temperature measurement from the MeteoSwiss stations to interpolate the outdoor temperatures at the location of each building. We considered only meteorological stations situated between 193 m and 1069 m altitude, which corresponds to the minimum and the 97.5 percentile of the altitude of all buildings taken into account in this study. We chose the 97.5 percentile to avoid leverage points of temperatures measured at higher altitudes which are nearly not populated.

As independent variables of the support vector regression, we took longitude, latitude and altitude into consideration. An important parameter of support vector regression is the cost parameter. The cost parameter determines the amount of over fitting respectively under fitting of the data. If the cost parameter is high, the support vector regression tends to over fit the data; if it is small it tends to under fit. We optimized the cost parameter of the support vector regression for each single day by 5-fold cross-validation which yielded an R^2 of 57% for the year 2000. We took into account only cost parameters between 0.1 and 10 with a resolution of 0.1. We carried out interpolations for each single day of the IRC measurement period. Hence, we calculated the mean outdoor temperature by averaging the estimated daily mean outdoor temperatures over the whole period. Finally, we grouped outdoor temperatures arbitrarily into 4 classes "<0 °C", "0–5 °C", "5–10 °C" and ">10 °C" and calculated the mean outdoor temperature for each class. The mean outdoor temperature in the Swiss Plateau is around 10 °C. We chose this classification to obtain one class above the mean temperature of the Swiss Plateau. However, most of the measurements have been performed during the heating period. To distinguish these classes by different degrees of cold, we chose "<0 °C" as a class to represent very cold temperatures, "0–5 °C" moderately cold temperatures and "5–10 °C" for rather mild temperatures.

2.4.8. Lithology

Since the lithological database consists of 70 different classes, interpreting it with respect to IRC is cumbersome. Therefore, we generalized the lithological classes into 6 classes according to our assumptions regarding their expected level for elevated radon potential: (1) sediment, i.e. quaternary deposits which correspond to loose surficial material mostly deposited after the last glaciation; (2) carbonate rock Jura, i.e. limestone and associated carbonate rock in the Jura Mountains; (3) carbonate rock Alps, i.e. limestone and associated carbonate rock in the Alps (4) igneous rock, i.e. granites; (5) metamorphic rocks which are rocks transformed through the formation of the Alps, and (6) sedimentary rocks which are not carbonates (e.g. sandstones, conglomerates etc.). Fig. 1 shows a map of the 6 generalized lithological classes. To investigate the influence of the lithology on IRC we plotted the arithmetic mean values and its confidence intervals versus the generalized lithological units.

3. Results

3.1. Distribution of IRC

The undeclustered point set yields a fractal dimension of 1.15 with 63,950 houses whereas after random declustering the fractal dimension results in 1.42 with 34,297 houses. Knowing that pure random locations would yield to a fractal dimension of 2.0, the value 1.42 implies that some clustering is still present in the data.

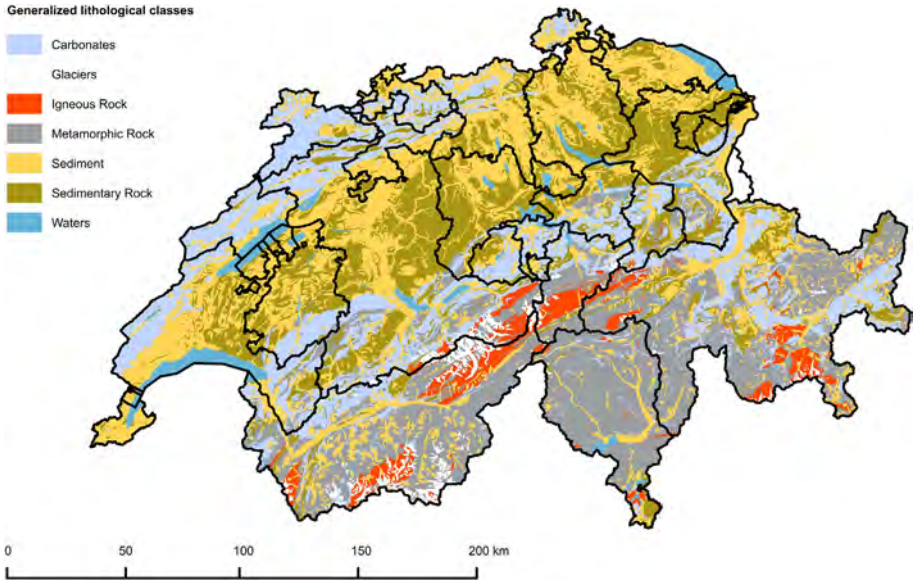


Fig. 1. Generalized lithological classes.

Since we only considered houses with measurements in inhabited rooms, the final dataset is composed of 32,151 houses in the declustered and 60,298 houses in the undeclustered case. Table 1 gives relevant statistical parameters of the distribution of IRC of the Swiss radon database for declustered as well as for undeclustered IRC measurements. The skewness of the declustered log transformed IRC is 0.6 and the kurtosis 4.11.

Fig. 2 displays the log-transformed distribution of the declustered IRC. We fitted a gamma and a normal probability density function to the log-transformed IRC. The fit of the gamma distribution yields a shape parameter of 20.78 and a rate of 4.54. The lognormal distribution was fitted with a mean of 4.58 and a standard deviation of 1.01. The hypotheses that the log-transformed IRC are drawn from a gamma or a normal distribution are rejected by a Kolmogorov–Smirnoff test in both cases with p-values smaller than $2.2 \cdot 10^{-16}$. The Kolmogorov distance for the gamma distribution is 0.040 and for the normal distribution 0.055.

3.2. Probability mapping

Fig. 3a presents the local probability to exceed an IRC of 300 Bq/m^3 in inhabited rooms. The probability ranges from 0 to 0.92. The probability to exceed 300 Bq/m^3 is clearly higher in the Jura Mountains, in the Canton of Ticino in the South of the Alps and in parts of the Canton of Grisons in the South-East of the Alps. In the very mountainous Canton of Valais in the South-West of Switzerland, the sampling density was too low to perform reliable probability mapping. Therefore we did not calculate the pixel values in this area.

Table 1
Summary statistics of indoor radon concentrations in Switzerland.

	Declustered data	Undeclustered data
Number of houses	32,151	60,298
Arithmetic mean (Bq/m^3)	189	198
Geometric mean (Bq/m^3)	98	107
Median (Bq/m^3)	87	95
Standard deviation (Bq/m^3)	439	434
Geometric standard deviation	2.8	2.7
Interquartile range (Bq/m^3)	119	131
Median absolute deviation (Bq/m^3)	68	74
Fraction of houses above 300 Bq/m^3	0.13	0.14
Fraction of houses above 1000 Bq/m^3	0.03	0.03

3.3. Statistical analysis

The hypothesis of uniform spatial distribution of samples was rejected by the χ^2-test for each class of each variable.

3.3.1. Type of radon detector

Fig. 4a shows the mean IRC of track and electret detectors. The red set corresponds to measurements which were taken over the whole of Switzerland ($K = 17.9$, $p < 5 \cdot 10^{-4}$) and the black one to

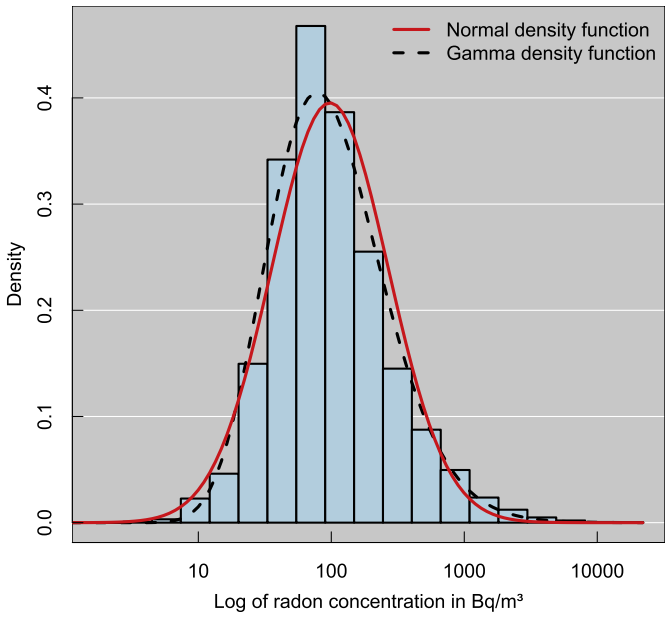


Fig. 2. Distribution of log-transformed IRC with fitted Gaussian and Gamma density.

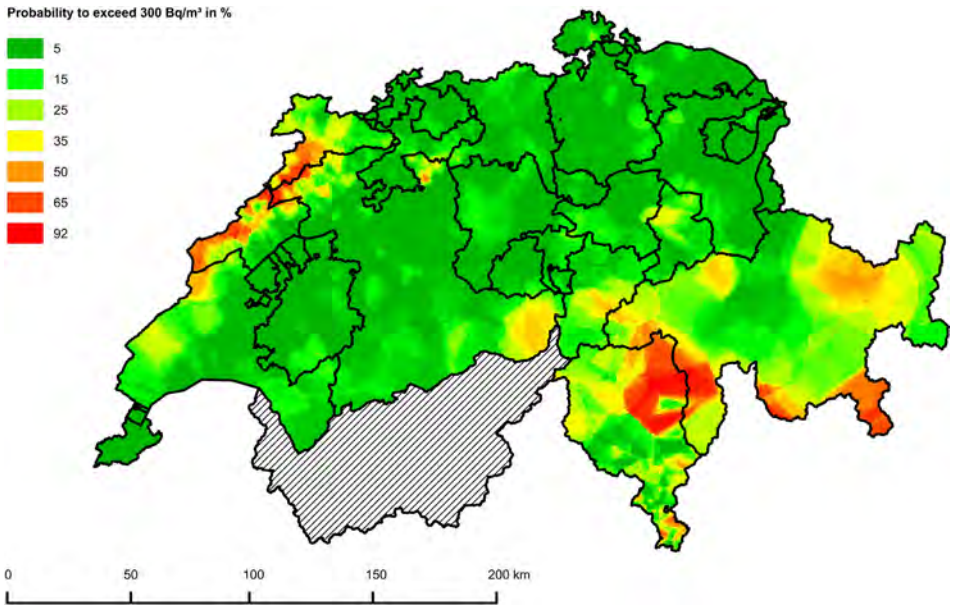


Fig. 3. Map of Switzerland indicating the local probability to exceed 300 Bq/m.[3].

measurements taken only in the Swiss Plateau ($K = 162.9$, $p < 2 \cdot 10^{-16}$). The confinement to the Swiss Plateau results in an inversion of the effect of both detector types on the IRC. For the all-of-Switzerland data, a strong discrepancy in spatial distribution can be observed in Fig. 4b. The electret detectors show some oversampling in the Swiss Plateau and the track detectors have a considerable share in the Jura Mountains and the canton Ticino. Fig. 4c shows the spatial distribution of samples restricted to the Swiss Plateau. The spatial distribution of sampling for both types of detectors has a low deviation from uniformity.

3.3.2. Type of foundation

The mean IRC for different foundation types are shown in Fig. 5 ($K = 295.9$, $p < 2 \cdot 10^{-16}$). Concrete foundations reveal the lowest mean values followed by earth foundations. The highest class relates to concrete slabs which were added after the construction. The spatial distributions of samples for all classes don't show considerable differences.

3.3.3. Year of construction

This analysis included 27,878 buildings since the year of construction was not indicated for all houses. Fig. 6 shows the mean values of IRC depending on the year of construction ($K = 337.8$, $p < 2 \cdot 10^{-16}$). For the first three classes of years, a significant decrease of mean IRC can be observed. Between the classes "(1970, 1990]" and "(1990, 2011]" the mean IRC are not significantly different.

3.3.4. Type of building

Fig. 7a shows the mean IRC for the building type "School", "Apartment building", "Detached house" and "Farm" ($K = 111.9$, $p < 2 \cdot 10^{-16}$). The IRC in apartment buildings are significantly below those of detached houses and farms. The estimation of the mean value for schools is very uncertain due to lack of measurements. Nearly no samples for farms are available in the canton Ticino. The spatial distribution of samples for detached houses is highly non-uniform and shows a considerable oversampling in the canton of Ticino.

3.3.5. Floor levels

Fig. 8 shows how IRC is dependent on the floor level. The mean ratios of the basement concentrations and other floor levels are plotted versus the corresponding floor levels. The graph reveals a nearly linear dependency between the mean ratios and the floor level. Since few houses ($N = 58$) were measured at the same period of time in the basement and on the second floor, the confidence intervals are very large and a significant difference between the first and second floor cannot be observed. Note that, surprisingly, 37% of all houses exceed 100 Bq/m^3 in the second floor.

Fig. 9a shows the log-transformed concentrations measured on the ground floor versus the log-transformed concentrations measured in the basement. The black line indicates the case of equality of IRC between the two floors. The Spearman's rank correlation coefficient equals $\rho = 0.65$. About a quarter of the houses have higher concentrations on the ground floor than in the basement. Fig. 9b shows the log transformed IRC on the first floor versus the log transformed concentrations on the second floor. In this case the Spearman's rank correlation coefficient equals $\rho = 0.89$. The correlation coefficients show that concentrations in the basement are much less correlated to concentrations on the ground floor than concentrations on the first floor to concentrations on the second floor.

3.3.6. Altitude

Fig. 10a shows the mean IRC depending on the altitude ($K = 1530.9$, $p < 2 \cdot 10^{-16}$). The mean IRC of class "(900, 2235]" differs considerably from the other classes. The spatial distribution of IRC measurements in each altitude class is shown in Fig. 10b. The spatial distribution of samples in the Swiss Plateau clearly decreases with higher altitude. Class "<350" reveals that the spatial distribution of samples concentrates mainly on the northern part of the Jura Mountains and the southern part of Ticino.

3.3.7. Outdoor temperature during measurement

The mean values for each outdoor temperature class are shown in Fig. 11a ($K = 180.7$, $p < 2 \cdot 10^{-16}$). A tendency of decreasing IRC with increasing outdoor temperature can be observed. However the mean value of the class "(5, 10]" is higher than the IRC mean value of "(0, 5]". Fig. 11b shows the spatial distribution of IRC

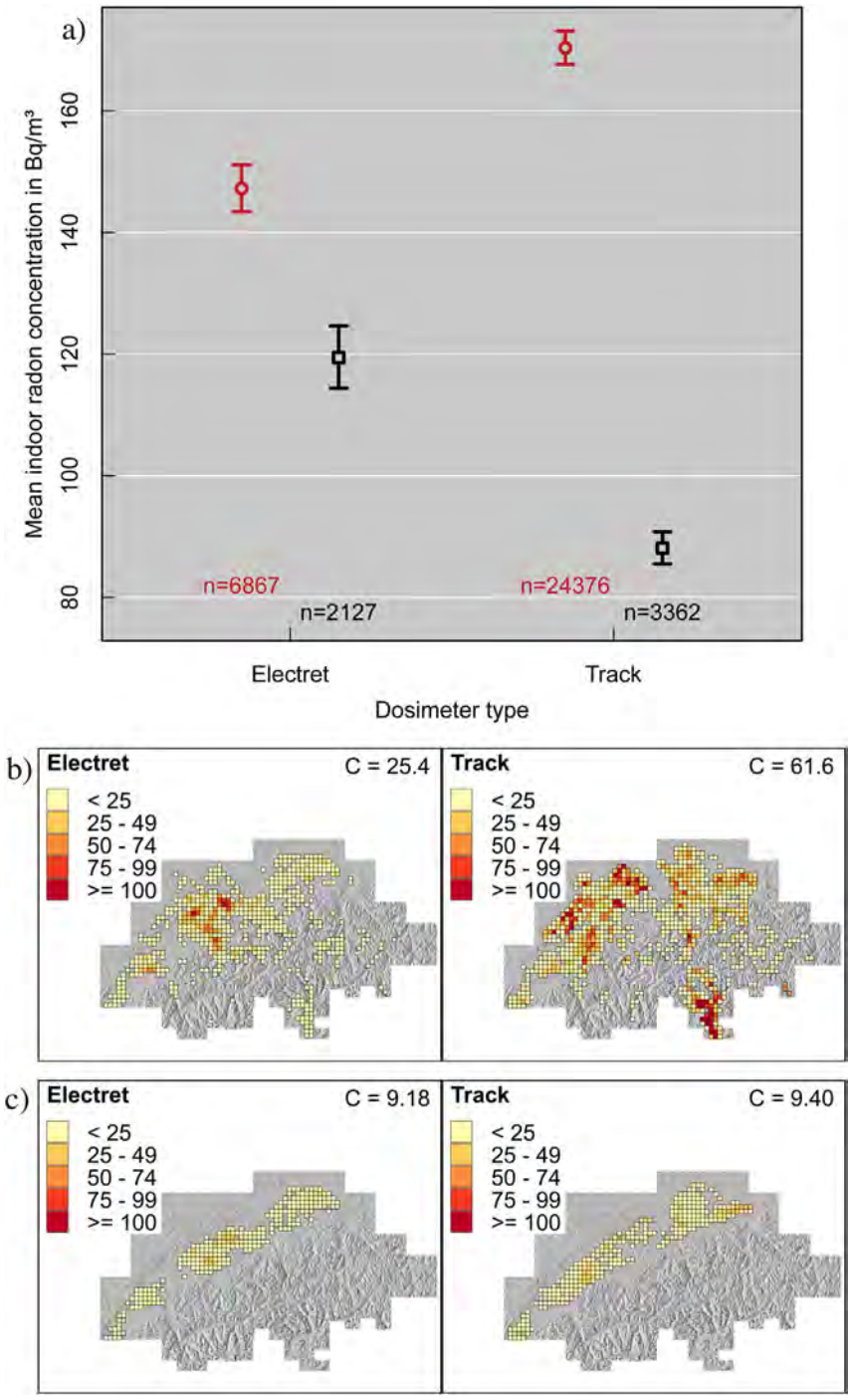


Fig. 4. a) Mean IRC versus radon detector type. Red circles: radon detectors located throughout Switzerland. Black squares: Measurements carried out exclusively in the Swiss Plateau. Spatial distribution of samples by radon detector class carried out b) in the whole of Switzerland and c) exclusively in the Swiss Plateau. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

measurements for the corresponding outdoor temperature class. In both outdoor temperature classes "(−5, 0]" and "(10, 23]", the density of samples in the canton of Ticino in the south of the Alps is smaller compared to the outdoor temperature classes "(0, 5]", "(5, 10]". The outdoor temperature classes "(0, 5]" and "(5, 10]" show clear non-uniformity of spatial distribution of samples.

3.3.8. Lithology

Fig. 12a shows the means and confidence intervals of IRC grouped by 6 different lithological classes ($K = 1086.3$, $p < 2 \cdot 10^{-16}$). Igneous rock and carbonate rock in the Jura Mountains demonstrated the highest IRC. Sediment and sedimentary rock have the lowest IRC. Carbonate rock in the Alps and carbonate rock in the Jura Mountains show a significant difference.

4. Discussion

The purpose of this work was to answer the following question: Which factors influence IRC in Switzerland and how are IRC samples corresponding to these factors spatially distributed?

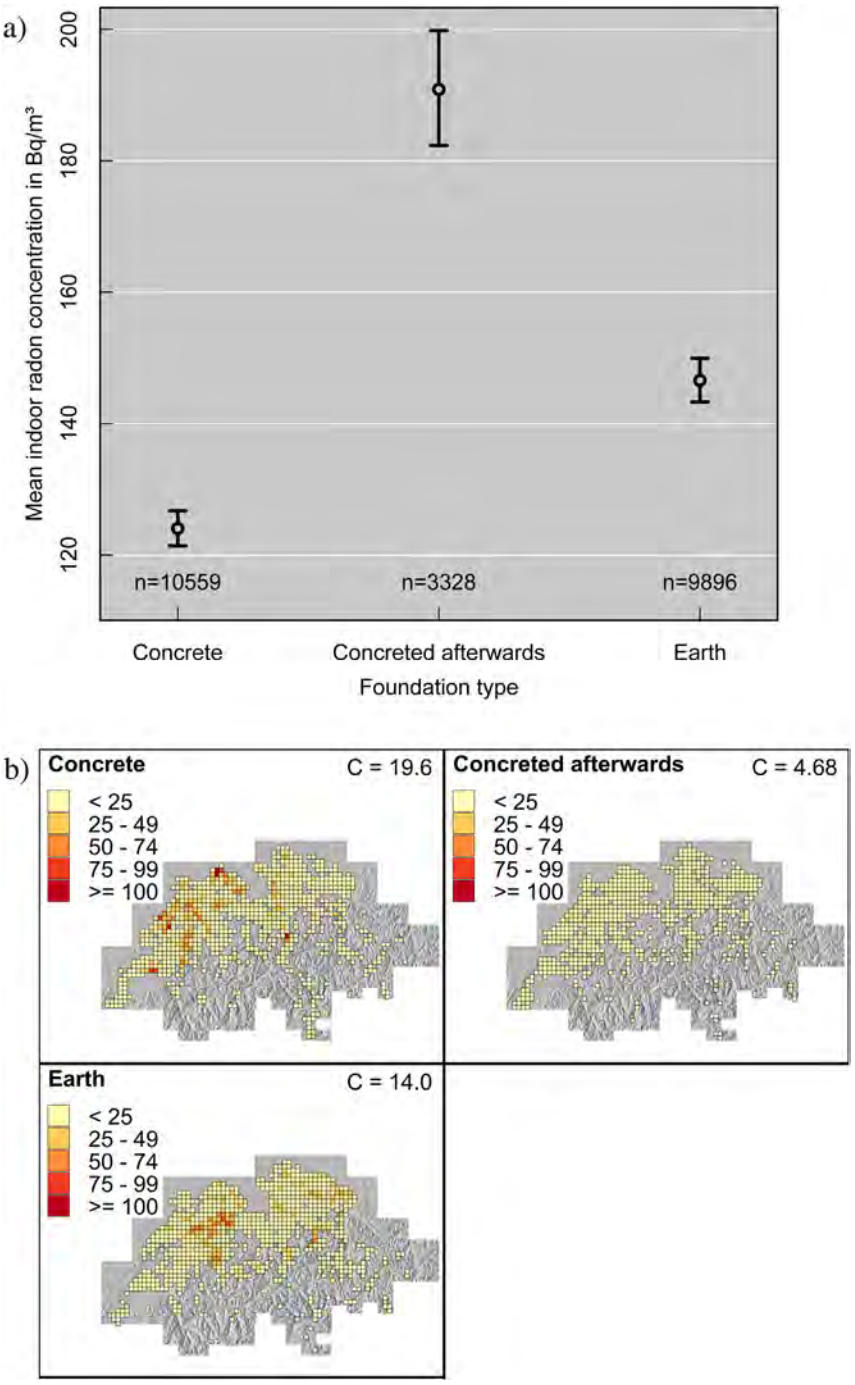


Fig. 5. a) Mean IRC versus building foundation. b) Spatial distribution of samples by class of foundation.

Since IRC measurements in Switzerland are taken autonomously by each canton, the spatial distribution of samples of the IRC data is very irregular. This leads to local over- or under representation of sampling. To reduce bias due to local over- or under sampling, we applied random declustering on the IRC dataset. Table 1 shows that the arithmetic mean for the undeclustered IRC dataset is higher (198 Bq/m^3) than the arithmetic mean in the declustered case (189 Bq/m^3). This difference can be explained by the sampling strategy of the Swiss authorities to favor measurements in radon-prone areas. The arithmetic mean value of the declustered dataset of 189 Bq/m^3 is large compared to the mean of 78 Bq/m^3 reported by Menzler et al. (2008). This is due to the fact that the mean value in the present study is not a population weighted mean, which would give a bigger weight to IRC of densely populated areas in Switzerland. Most of Switzerland's population is located in the Swiss Plateau, which has rather low IRC as can be seen in Fig. 3a. Therefore, population weighted mean of IRC in Switzerland is smaller than the unweighted mean. Since this study aims to explore variables that influence IRC, we favored a mean that takes all sampled regions equally into account. The quantity estimated here is the spatial mean.

Since log-transformed IRC were often observed to follow a distribution close to a normal or a gamma distribution, we used a Kolmogorov–Smirnov test to compare the distribution of the log-

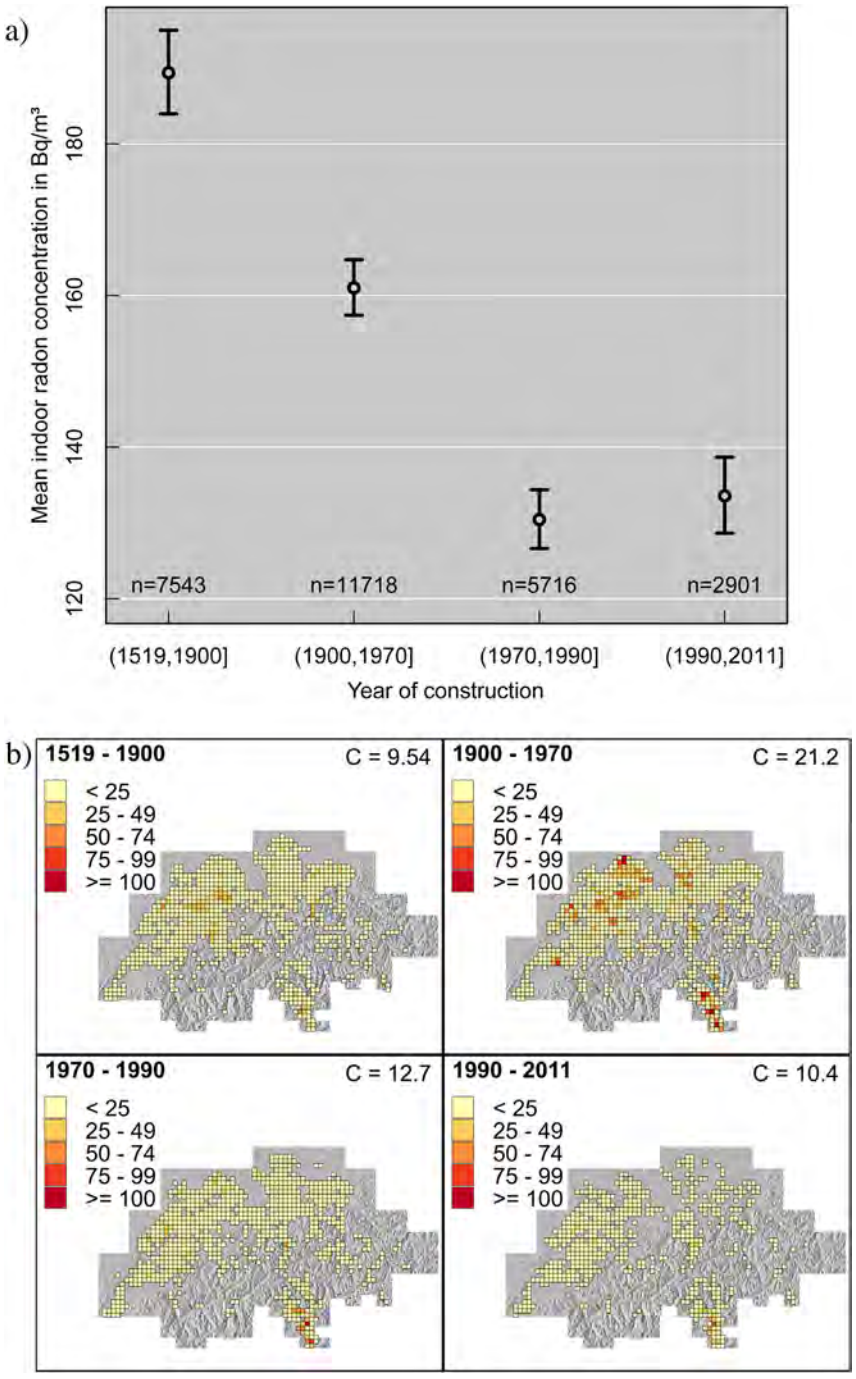


Fig. 6. a) Mean IRC versus year of construction. b) Spatial distribution of samples by class of years of construction.

transformed IRC with a normal and a gamma distribution. The Kolmogorov–Smirnoff test rejects the hypotheses that the log-transformed IRC in Switzerland follow normal or gamma distributions. Nevertheless, the Kolmogorov–Smirnov distance is slightly smaller for the gamma distribution, indicating a better goodness-of-fit with the gamma distribution. This may be due to the fact that the gamma distribution is more flexible with regard to skewness. The skewness of 0.6 indicates a deviation from the symmetry of the normal distribution. The departure from log-normality may be attributable to clustering, which is indicated by the low fractal dimension of 1.42 of the IRC dataset. On the other hand, a more rigorous random declustering of the spatial distribution of the IRC data would be at the expense of data size which would result in less accurate estimates for all analyses in this study. A deeper insight into the reasons for the deviation from log-normality is not in the scope of this work and will be subject to future investigations. Nevertheless, as depicted in Fig. 2, the distribution of IRC in our study is reasonably close to log-normality, thus justifying the use of statistical methods assuming normality with log-transformed IRC.

A significant influence on the dispersion of IRC measurements is the measurement process itself. We observed substantial differences between electret and track detectors in the estimation of IRC (see Fig. 4a). Electret detectors measure the concentration of alpha

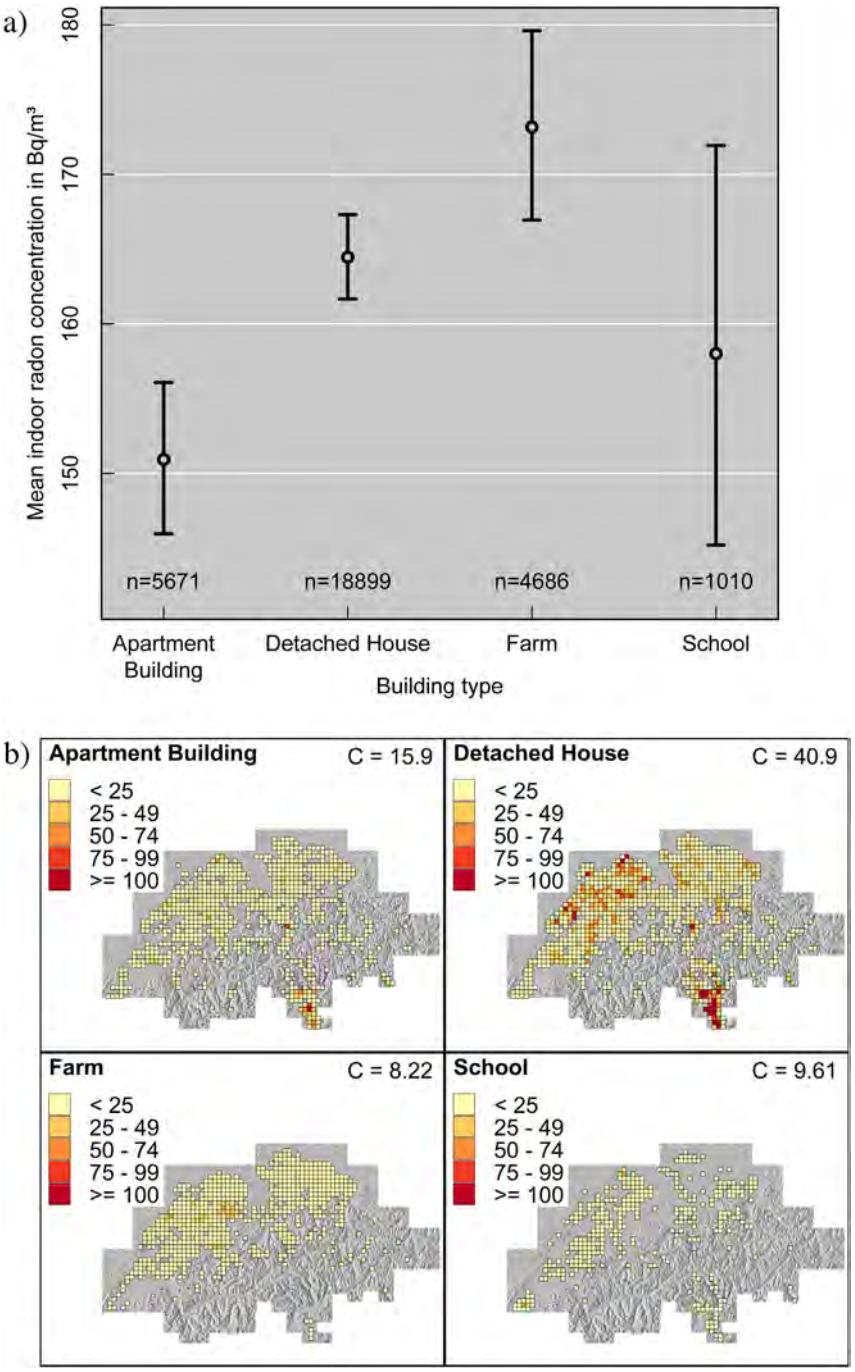


Fig. 7. a) Mean IRC versus type of building. b) Spatial distribution of samples by class of type of building.

emitters in the air with a dielectric plastic film, which is discharged by the impact of alpha particles. Higher air humidity and dirt can lead to a faster discharge of the electret detector consequently resulting in an overestimation of alpha particles in the air. This overestimation of IRC by electrets is in accordance with earlier findings, where several electret and track etch detectors were compared pair wise in the same houses over a three month period (Federal Office of Public Health, 2011). Consequently, IRC measurements carried out with electret detectors have to be interpreted carefully and models which aim to predict IRC have to account for the differences of electrets compared to other detectors. The fact that the dataset from the Swiss Plateau reproduces these findings rather than the dataset from the whole of Switzerland is due to the more uniform spatial distribution of samples and a more homogeneous geology in the Swiss Plateau. Track detectors have a considerable higher share in the canton Ticino and the Jura Mountains compared to electret detectors (see Fig. 4b). This may explain the higher mean of IRC measurements for track detectors in the case of the all-of-Switzerland data.

Building characteristics also contribute to the variability of IRC measurements. The influence of the type of foundation on IRC is represented in Fig. 5. Buildings with earth foundations are significantly more prone to radon than buildings with concrete foundations. Indeed, earth foundations are less sealed against radon entry

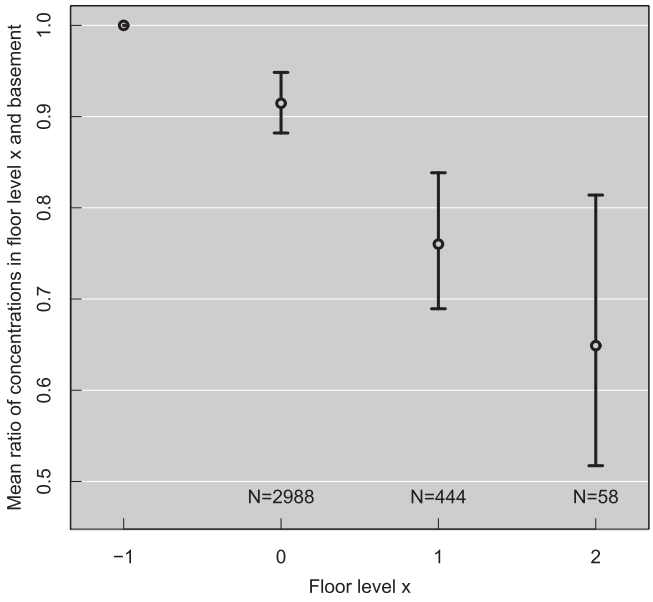


Fig. 8. Ratio of IRC at a given floor level in a building to IRC in the basement of the same building.

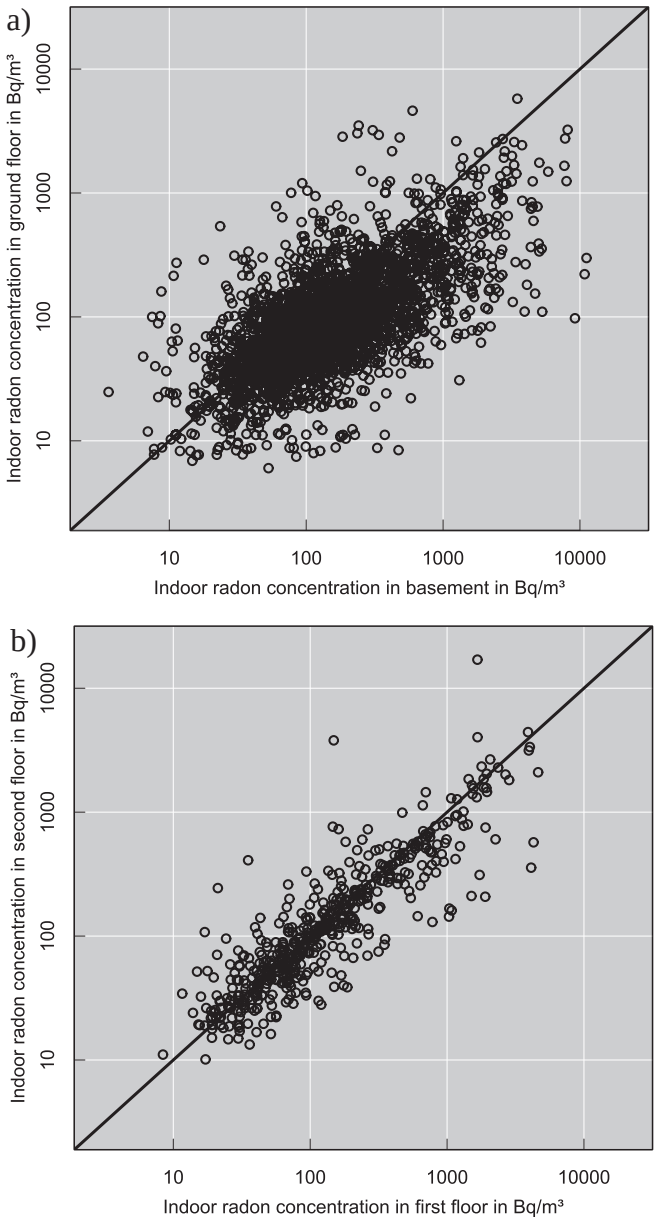


Fig. 9. IRC in different floor levels of same building while same measurement period: a) ground floor versus basement b) second floor versus first floor.

from the subsoil than concrete foundations. Concrete foundations built into the house after construction result in higher IRC and therefore suggest that this method is not effective to avoid radon infiltration. The spatial distribution of measurements is similar for all classes (see Fig. 5b). However, it is likely that buildings in which the concrete foundation was built afterwards were particularly radon prone before. Hauri et al. (2012) showed that the year of construction of a building is one of the most influencing variables with respect to IRC. Fig. 6a shows a significant difference between IRC measured in buildings built before 1900, between 1900 and 1970, and after 1970. The drop after the 1970s can be explained by new building regulations put in place following the oil crisis in the 1970s which resulted in better insulation against subsoil (Burkart et al., 1984). Gunby et al. (1993) found that elevated IRC in buildings built before 1900 may be caused by the use of stones as a building material. However, since information about building materials is not provided in the Swiss radon database, we could not further study this variable. Moreover, the type of building influences IRC. In particular, apartment buildings do have significantly lower IRC than detached houses and farms. Indeed, apartment buildings are rather newer constructions with better tightening of the bottom slab which prevents radon entries. On the other hand, measurements for apartment buildings were carried out at higher floor levels compared to detached houses with 35% versus 22% of measurements taken above the ground floor, respectively. Moreover, spatial distributions of samples for the different classes are different as can be seen in Fig. 7b. This leads to substantial biases in the estimation of IRC mean values for the different building types.

The effect of the floor level on the IRC was studied by comparing IRC of different floor levels in the same houses. This method avoids biases that would occur by just calculating the mean value of the IRC for each floor. Fig. 8 shows that the mean ratios between the basement IRC and the other floors are significantly different except between the first and the second floor. Nevertheless, the difference between the ground floor and the basement is rather small. This is in accordance to the fact that in 26% of cases, IRC on the ground floor are higher than in the basement. The fact that the correlation between IRC in the basement and on the ground floor are much lower than the correlation between concentrations on the first floor and second floor can be interpreted in three ways. First, IRC on higher floors are predominantly determined by the exhalation of IRC by the building materials, thus leading to similar IRC in all floor levels. However, this hypothesis does not seem realistic since 37% of the houses exceed concentrations of 100 Bq/m^3 on the second floor, a concentration which can rarely be found to be produced by building materials in literature (Schuler et al., 1991). A second hypothesis is that IRC on the higher floor levels are transferred by electrical conduits. This would result in rather similar IRC on different higher floor levels. Finally, the fact that indoor–outdoor air exchange and anthropogenic influences are more similar between higher floors than between basement and ground floor, may also explain the higher correlation between first and second floor. Detailed investigations of radon pathways within houses remain a topic for future studies.

IRC show significant association with altitude in Switzerland (see Fig. 10a). This relationship can be explained by the relationship

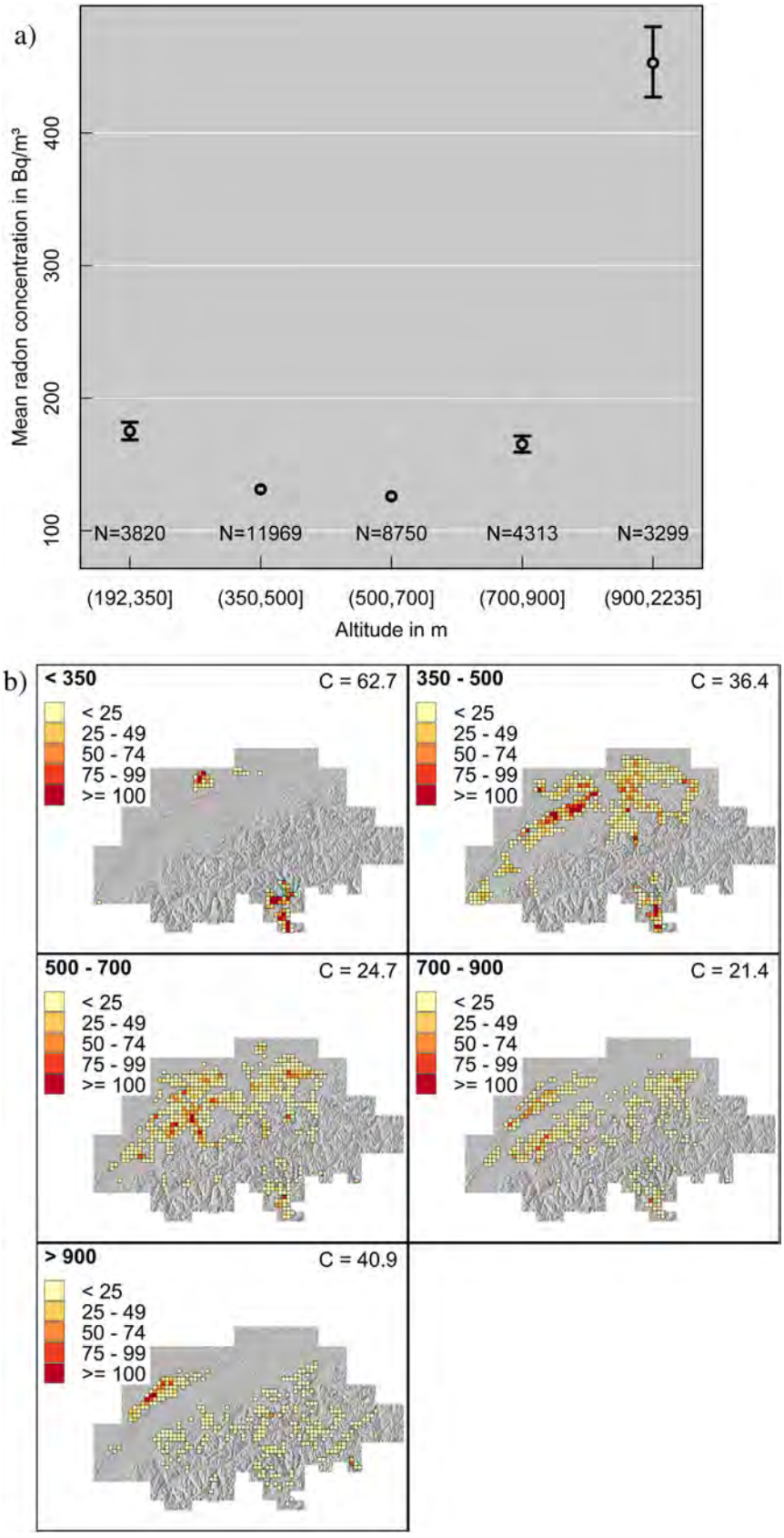


Fig. 10. a) Mean IRC versus altitude b) Spatial distribution of samples stratified by altitude.

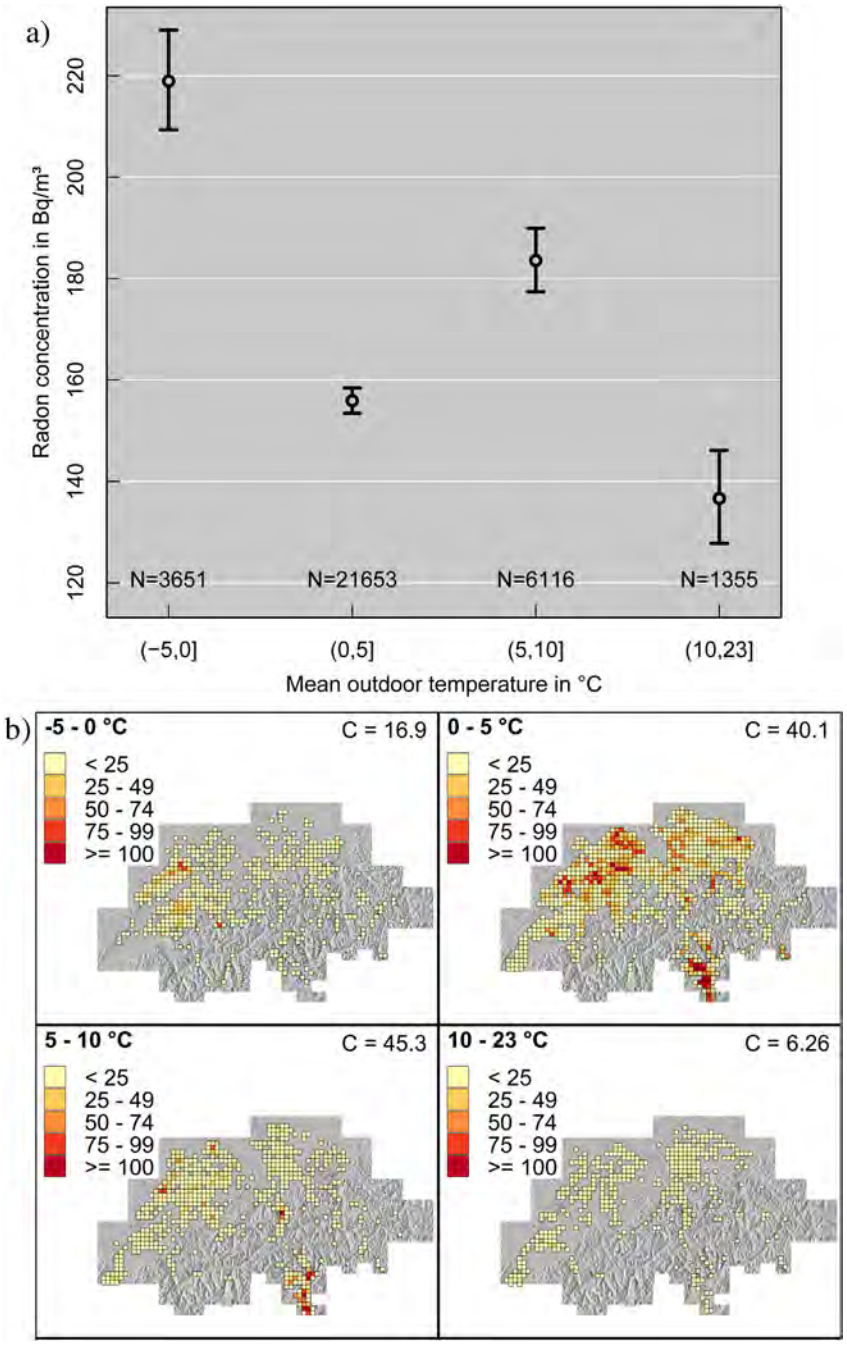


Fig. 11. a) Mean IRC versus outdoor temperature b) Spatial distribution of samples stratified by outdoor temperature.

of altitude to outdoor temperature and geology. The tendency of measuring lower IRC at higher outdoor temperatures (see Fig. 11a) can be attributed to the higher ventilation during warmer periods. However, this result has to be interpreted with caution since outdoor temperatures are often strongly associated with altitude. Since geology is also dependent on altitude, the covariation of IRC with outdoor temperature can be partly explained by the dependence of IRC on geology. The Alps with granites and gneisses as well as the Jura Mountains with a high abundance of karst lie at a higher altitude than the Swiss Plateau that is dominated by quaternary glacial sediments and sedimentary rock. Fig. 12a reveals that geology has a significant effect on IRC. Clearly carbonate rock and igneous rock have higher IRC than sedimentary rock and sediment. Gneisses in Switzerland and felsic igneous rock are rich in uranium (Schön, 2004). This explains the high IRC mean in igneous rock in Fig. 12a as well as the high probability of exceeding 300 Bq/m^3 in the Alps, a region where felsic igneous rock and gneisses are abundant. The high IRC with carbonate rock in the Jura Mountains can be explained by a widespread karstification of the carbonate rock in this area compared to carbonate rock in the Alps. Since both the Alps and the Jura Mountains are at higher altitudes, IRC are obviously altitude-dependent (see Fig. 10a). The mean IRC in metamorphic rock is centered between the other lithological classes. This is not surprising since metamorphic rock can originate from a variety of different rocks which are not necessarily rich in uranium (e.g. slate, paragneiss, etc).

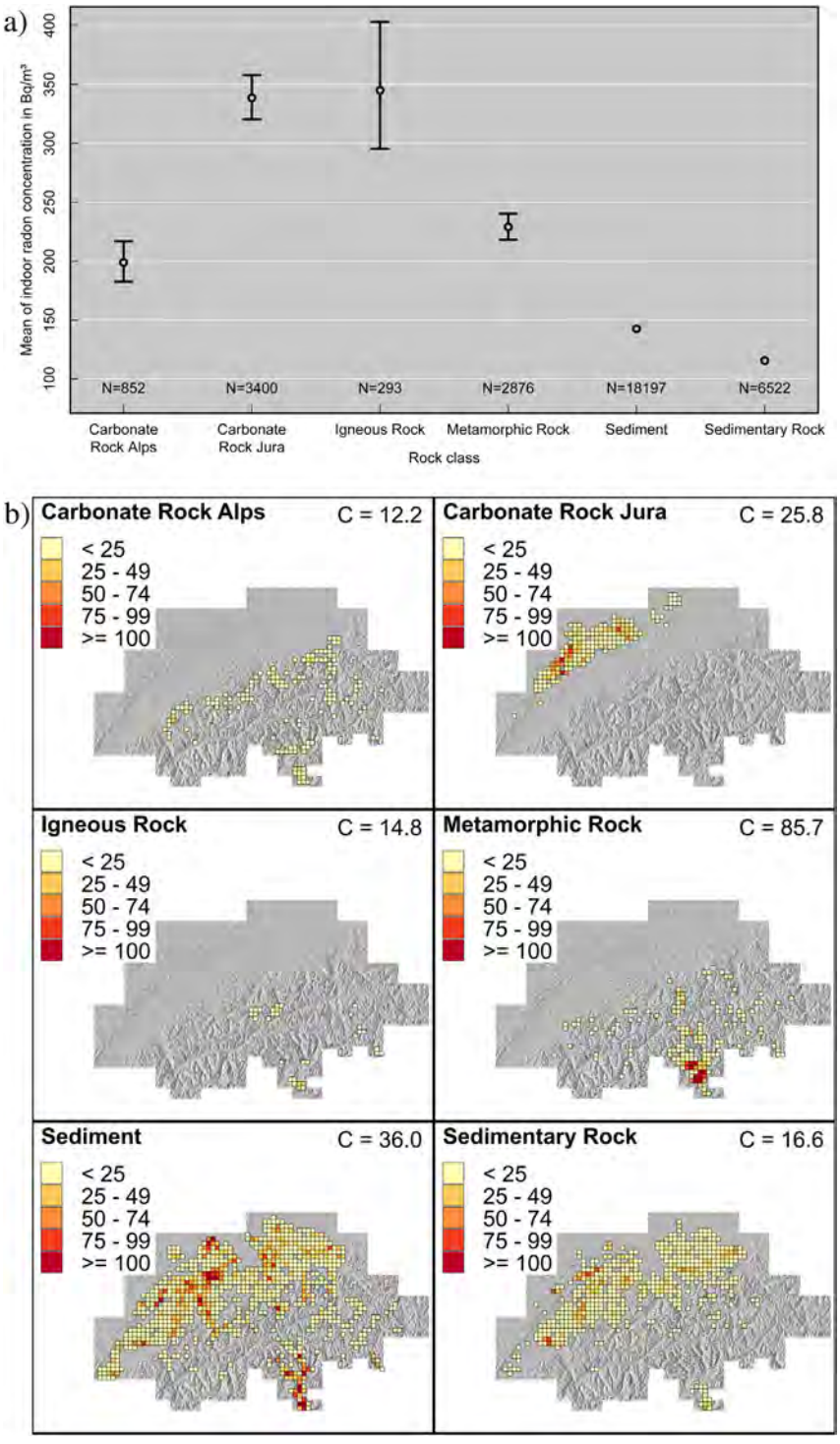


Fig. 12. a) Mean IRC versus lithology b) Spatial distribution of samples by class of lithology.

It is important to note that IRC measurements in this study were taken by house owners and not by trained personnel. We may expect that this would have increased the random error rather than the systematic error. A larger random error of a measurement broadens the confidence intervals of the estimation of the mean value of the measured quantity. However, in most of our analyses, enough measurements were available in order to provide sufficient confidence on the estimations of mean values.

Furthermore, it has to be mentioned, that we considered the maximum IRC measurements in inhabited rooms per house. Taking the mean value could lead to different results. Nevertheless, in less than 3% of the considered houses more than 2 measurements were available in inhabited rooms. Taking the mean value of 2 measurements still leaves a high uncertainty on the estimation. According to the Swiss legislation, we considered the maximum measured IRC per house in inhabited rooms which is the legally binding measurement to be compared with the reference level. In

addition, the maximum IRC provides information about the potential IRC that a house can reach and we think that it is more relevant to compare directly this value with potentially influencing variables.

The IRC measurements in this study haven't been corrected for seasonal variations. The FOPH applies a correction factor on IRC in Switzerland (Piller and Johner, 1998). Those correction factors have been calculated based on measurements carried out in summer and in winter. However the amount of summer measurements available in Switzerland is very small compared to the amount of winter measurement. The correction factors derived from this data are hence subject to strong uncertainty. To minimize additional errors in our measurements we preferred to use the uncorrected measurements for our analysis.

The fact that the sampling strategy changed over time may lead to bias of the estimate of IRC mean values, which cannot be adjusted by declustering. Nevertheless, the mapping of the spatial distribution of samples permits to illustrate the lack of uniformity of sampling and to assess the possible bias which could be produced by different sampling criteria.

Finally, it has to be born in mind that we did not correct IRC measurements for possible errors due to thoron concentrations.

5. Conclusions

This study contributes to a better understanding of IRC in Switzerland. We found significant relationships between IRC and all variables taken into consideration. Our findings are in general accordance with previous studies (Andersen et al., 2007; Appleton and Miles, 2010; Gunby et al., 1993; Hauri et al., 2012; Hunter et al., 2009; Khan, 2000; Papaefthymiou et al., 2003; Žunic et al., 2007). However, few studies explored IRC datasets with such a high spatial sampling density. Additionally, we mapped the spatial distribution of samples for each class of each variable. This provides a visual tool to understand possible biases caused by unbalanced spatial distributions of samples. Based on our results, it will be possible to develop models for the prediction of local IRC considering for geographical coordinates and most relevant associated variables. For instance, we observed that geology has a significant influence on IRC. Sedimentary rock and sediment have clearly lower IRC than the carbonate rock in the Jura and igneous rock. On the other hand, we found that carbonate rock in the Jura Mountains shows a different radon characteristic than carbonate rock in the Alps. Consequently, and unsurprisingly, 3D geological information is a potential variable that could be used to predict IRC. Data driven machine-learning techniques, such as Random Forests or kernel density estimation, are interesting nonlinear methods to carry out highly accurate prediction by combining categorical as well as continuous predictors. However, prediction accuracy is often in tradeoff with the interpretability of predictors since the interdependency of predictors can be very complex. Therefore we stressed the simple univariate relationships between potential predictors and IRC in the present study without taking into account interactions between variables. This work is essential for the development of more complex predictive models to map IRC in Switzerland. The follow-up work of our project will be the exploration of the potential of data driven machine-learning techniques to analyze, map and predict IRC in Switzerland based on the univariate analyses of this study.

Acknowledgments

We thank Milan Beres of Swisstopo for providing us with the geological data that was used in this study. This study was supported by the Federal Office of Public Health Switzerland.

References

Åkerblom, G., Andersson, P., Clevensjö, B., 1984. Soil gas radon – a source for indoor radon daughters. Radiat. Prot. Dosim. 7, 49–54.
Andersen, C.E., Raaschou-Nielsen, O., Andersen, H.P., Lind, M., Gravesen, P., Thomsen, B.L., Ulbak, K., 2007. Prediction of ^{222}Rn in Danish dwellings using geology and house construction information from central databases. Radiat. Prot. Dosim. 123, 83–94.
Appleton, J.D., Miles, J.C.H., 2010. A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk. J. Environ. Radioact. 101, 799–803.
Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., 2008. Applied Spatial Data Analysis with R. Springer, NY.
Bossew, P., 2010. Radon: exploring the log-normal mystery. J. Environ. Radioact. 101, 826–834.
Bossew, P., Lettner, H., 2007. Investigations on indoor radon in Austria, part 1: seasonality of indoor radon concentration. J. Environ. Radioact. 98, 329–345.
Bossew, P., Dubois, G., Tollefsen, T., 2008. Investigations on indoor radon in Austria, part 2: geological classes as categorical external drift for spatial modelling of the radon potential. J. Environ. Radioact. 99, 81–97.
Burkart, W., Wernli, C., Brunner, H.H., 1984. Matched pair analysis of the influence of weather-stripping on indoor radon concentration in Swiss dwellings. Radiat. Prot. Dosim. 7, 299–302.
Burke, Ó., Long, S., Murphy, P., Organo, C., Fenton, D., Colgan, P.A., 2010. Estimation of seasonal correction factors through Fourier decomposition analysis-a new model for indoor radon levels in Irish homes. J. Radiolog. Prot. 30, 433.
Cherkassky, V.S., Mulier, F., 2007. Learning from Data: Concepts, Theory, and Methods. John Wiley & Sons.
Cinelli, G., Tondeur, F., Dehandschutter, B., 2011. Development of an indoor radon risk map of the Walloon region of Belgium, integrating geological information. Environ. Earth Sci. 62, 809–819.
Denman, A.R., Crockett, R.G.M., Groves-Kirkby, C.J., Phillips, P.S., Gillmore, G.K., Woolridge, A.C., 2007. The value of seasonal correction factors in assessing the health risk from domestic radon – a case study in Northamptonshire, UK. Environ. Int. 33, 34–44.
Dessau, J., Gagnon, F., Lévesque, B., Prévost, C., Leclerc, J., Belles-Isles, J., 2005. Le radon au Québec-Évaluation du risque à la santé et analyse critique des stratégies d'intervention. INSPQ.
Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, D., 2012. e1071: Misc Functions of the Department of Statistics. TU Wien.
Federal Office of Public Health, 2011. Comparison of the Measurement Results Obtained by Radosure with Other Detectors.
Friedmann, H., 2005. Final results of the Austrian radon project. Health Phys. 89, 339–348, 310.1097/1001.HP.0000167228.0000118113.0000167227.
Friedmann, H., Bossew, P., 2010. Selected statistical problems in spatial evaluation of Rn related variables. Nukleonika 55, 429–432.
Friedmann, H., Groeller, J., 2010. An approach to improve the Austrian radon potential map by Bayesian statistics. J. Environ. Radioact. 101, 804–808.
Girault, F., Perrier, F., 2012. Estimating the importance of factors influencing the radon-222 flux from building walls. Sci. Total Environ. 433, 247–263.
GRASS Development Team, 2012. Geographic Resources Analysis Support System (GRASS GIS) Software. Open Source Geospatial Foundation.
Groves-Kirkby, C.J., Denman, A.R., Crockett, R.G.M., Phillips, P.S., Gillmore, G.K., 2006. Identification of tidal and climatic influences within domestic radon time-series from Northamptonshire, UK. Sci. Total Environ. 367, 191–202.
Gunby, J.A., Darby, S.C., Miles, J.C.H., Green, B.M.R., Cox, D.R., 1993. Factors affecting indoor radon concentrations in the United-Kingdom. Health Phys. 64, 2–12.
Hauri, D.D., Huss, A., Zimmermann, F., Kuehni, C.E., Röösli, M., 2012. A prediction model for assessing residential radon concentration in Switzerland. J. Environ. Radioact. 112, 83–89.
Hunter, N., Muirhead, C.R., Miles, J.C.H., Appleton, J.D., 2009. Uncertainties in radon related to house-specific factors and proximity to geological boundaries in England. Radiat. Prot. Dosim. 136, 17–22.
Ielsch, G., Cushing, M.E., Combes, P., Cuney, M., 2010. Mapping of the geogenic radon potential in France to improve radon risk management: methodology and first application to region Bourgogne. J. Environ. Radioact. 101, 813–820.
Janssen, I., Stebbings, J.H., 1992. Gamma distribution and house ^{222}Rn measurements. Health Phys. 63.
Johner, H.U., Surbeck, H., 2001. Soil gas measurements below foundation depth improve indoor radon prediction. Sci. Total Environ. 272, 337–341.
Kanevski, M., Maignan, M., 2004. Analysis and Modelling of Spatial Environmental Data. EPFL Press.
Kemp, S.E., Jefferis, G., 2012. RANN: Fast Nearest Neighbour Search.
Kemski, J., Klingel, R., Siehl, A., Valdivia-Manchego, M., 2009. From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. Environ. Geol. 56, 1269–1279.
Khan, A.J., 2000. A study of indoor radon levels in Indian dwellings, influencing factors and lung cancer risks. Radiat. Measure. 32, 87–92.
Kotrappa, P., Dempsey, J.C., Ramsey, R.W., Stieff, L.R., 1990. A practical E-PERM (electret passive environmental radon monitor) system for indoor ^{222}Rn measurement. Health Phys. 58, 461–467.
Land, C.E., 1972. An evaluation of approximate confidence interval estimation methods for lognormal means. Technometrics 14, 145–158.
Menzler, S., Piller, G., Gruson, M., Rosario, A.S., Wichmann, H.-E., Kreienbrock, L., 2008. Population attributable fraction for lung cancer due to residential radon

in Switzerland and Germany. Health Phys. 95, 179–189, 110.1097/1001.HP.0000309769.0000355126.0000309703.
MeteoSwiss, 2013. Federal Office of Meteorology and Climatology. MeteoSwiss.
Miles, J.C.H., 2001. Temporal variation of radon levels in houses and implications for radon measurement strategies. Radiat. Prot. Dosim. 93, 369–375.
Miles, J.C.H., Appleton, J.D., 2005. Mapping variation in radon potential both between and within geological units. J. Radiolog. Prot. 25, 257.
Nero, A.V., Schwehr, M.B., Nazaroff, W.W., Revzan, K.L., 1986. Distribution of airborne radon-222 concentrations in U.S. homes. Science 234, 992–997.
Nikolaev, V.A., Ilić, R., 1999. Etched track radiometers in radon measurements: a review. Radiat. Measure. 30, 1–13.
Papaefthymiou, H., Mavroudis, A., Kritidis, P., 2003. Indoor radon levels and influencing factors in houses of Patras, Greece. J. Environ. Radioact. 66, 247–260.
Piller, G., Johner, I., 1998. Classification of radon areas in Switzerland. Radiat. Prot. Dosim. 78, 7–9.
Quantum GIS Development Team, 2012. Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project.
R Core Team, 2012. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
Schön, J., 2004. Physical Properties of Rocks: Fundamentals and Principles of Petrophysics. Elsevier.
Schuler, C., Crameri, R., Burkart, W., 1991. Assessment of the indoor Rn contribution of Swiss building materials. Health Phys. 60, 447–451.
SGTK, 2000. Lithologisch-petrografische Karte der Schweiz-Lithologie-Hauptgruppen 1:500000. Schweizerische Geotechnische Kommission.
Shen, H.P., Brown, L.D., Zhi, H., 2006. Efficient estimation of log-normal means with application to pharmacokinetic data. Stat. Med. 25, 3023–3038.
Smola, A., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.
Tapia, R., Kanevski, M., Maignan, M., Gruson, M., 2006. Comprehensive multivariate analysis of indoor radon data in Switzerland. In: International Workshop on the Geological Aspects of Radon Risk Mapping.
Trümpy, R.S.G.K., 1980. Geology of Switzerland: a Guide-book. Wepf & Co., Basel; New York.
Tuia, D., Kanevski, M., 2008. Indoor radon distribution in Switzerland: lognormality and extreme value theory. J. Environ. Radioact. 99, 649–657.
Vaupotic, J., Hunyadi, I., Baradacs, E., 2001. Thorough investigation of radon in a school with elevated levels. Radiat. Measure. 34, 477–482.
Vaupotic, J., Kobal, I., Krizman, M.J., 2010. Background outdoor radon levels in Slovenia. Nukleonika 55, 579–582.
WHO, 2009. WHO Handbook on Indoor Radon – a Public Health Perspective.
Zhou, X.H., Gao, S., 1997. Confidence intervals for the log-normal mean. Stat. Med. 16, 783–790.
Žunic, Z.S., Yarmoshenko, I.V., Birovljev, A., Bochicchio, F., Quarto, M., Obryk, B., Paszkowski, M., Čeliković, I., Demajo, A., Ujić, P., Budzanowski, M., Olko, P., McLaughlin, J.P., Waligorski, M.P.R., 2007. Radon survey in the high natural radiation region of Niška Banja, Serbia. J. Environ. Radioact. 92, 165–174.

Science of the Total Environment 505 (2015) 137–148

Contents lists available at ScienceDirect

Science of the Total Environment

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

Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland

Georg Kropat [a,*], Francois Bochud [a], Michel Jaboyedoff [c], Jean-Pascal Laedermann [a], Christophe Murith [b], Martha Palacios (Gruson) [b], Sébastien Baechler [a,b]

[a] Institute of Radiation Physics, Lausanne University Hospital, Rue du Grand-Pré 1, 1007 Lausanne, Switzerland
[b] Swiss Federal Office of Public Health, Schwarzenburgstrasse 165, 3003 Berne, Switzerland
[c] Faculty of Geosciences and Environment, University of Lausanne, GEOPOLIS — 3793, 1015 Lausanne, Switzerland

HIGHLIGHTS

- Kernel regression was used to map indoor radon concentration in Switzerland.
- Our model explains 28% of the variations of radon concentration data.
- Maps were generated considering different architectural elements and geology.
- Maps showing the local probability to exceed 300 Bq/m3 were proposed.
- We developed a confidence index to assess the reliability of the probability map.

ARTICLE INFO

Article history:
Received 29 June 2014
Received in revised form 10 September 2014
Accepted 22 September 2014
Available online xxxx

Editor: Pavlos Kassomenos

Keywords:
Indoor radon
Kernel regression
Predictive mapping
Probability mapping
Switzerland
Geology

ABSTRACT

Purpose: The aim of this study was to develop models based on kernel regression and probability estimation in order to predict and map IRC in Switzerland by taking into account all of the following: architectural factors, spatial relationships between the measurements, as well as geological information.
Methods: We looked at about 240 000 IRC measurements carried out in about 150 000 houses. As predictor variables we included: building type, foundation type, year of construction, detector type, geographical coordinates, altitude, temperature and lithology into the kernel estimation models. We developed predictive maps as well as a map of the local probability to exceed 300 Bq/m^3. Additionally, we developed a map of a confidence index in order to estimate the reliability of the probability map.
Results: Our models were able to explain 28% of the variations of IRC data. All variables added information to the model. The model estimation revealed a bandwidth for each variable, making it possible to characterize the influence of each variable on the IRC estimation. Furthermore, we assessed the mapping characteristics of kernel estimation overall as well as by municipality. Overall, our model reproduces spatial IRC patterns which were already obtained earlier. On the municipal level, we could show that our model accounts well for IRC trends within municipal boundaries. Finally, we found that different building characteristics result in different IRC maps. Maps corresponding to detached houses with concrete foundations indicate systematically smaller IRC than maps corresponding to farms with earth foundation.
Conclusions: IRC mapping based on kernel estimation is a powerful tool to predict and analyze IRC on a large-scale as well as on a local level. This approach enables to develop tailor-made maps for different architectural elements and measurement conditions and to account at the same time for geological information and spatial relations between IRC measurements.

1. Introduction

Radon is a radioactive gas that is known to be the second leading cause of lung cancer after smoking (Zeeb and Shannoun, 2009a). In Switzerland, radon accounts for about 230 deaths per year (Menzler

* Corresponding author at: Institute of Radiation Physics, University Hospital Center of Lausanne (CHUV), Rue du Grand-Pré 1, 1007 Lausanne, Switzerland. Tel.: + 41 21 314 82 96; fax: + 41 21 314 82 99.
E-mail address: georg.kropat@chuv.ch (G. Kropat).

http://dx.doi.org/10.1016/j.scitotenv.2014.09.064
0048-9697/

et al., 2008). Many of these deaths could be avoided by effective radon prevention.

Exposure to radon mainly takes place in closed environments like buildings at home or at work. Since most of the radon enters from the ground into a building, radon-exposure prevention consists of remediation measures and changing ventilation habits of the inhabitants.

The entry and the behavior of IRC inside houses are complex processes driven by several variables: building characteristics, anthropogenic influences, and geological and meteorological factors. Much of the variability of IRC is therefore related to variables which can change among neighborhood houses and on the time scale of hours (Groves-Kirkby et al., 2006; Miles, 2001).

The management of radon risk requires reliable predictive maps. If radon can be predicted in local areas, the local population would be informed about their health risk. This would motivate homeowners to take measurements and remediate their buildings, as well as to consider radon preventive measures in new constructions. Predictive mapping of indoor radon concentrations (IRC) is however difficult, since IRC can strongly vary on small spatial ranges (Borgoni et al., 2011; Cinelli et al., 2011; Zhu et al., 2001).

The approaches which exist to estimate the local radon hazard can be classified into two major groups. The first is that the radon hazard can be estimated using prior knowledge of the radon characteristics of the local geology. For example, uranium concentrations can serve as a proxy to determine the radon characteristics of a geological unit (Ielsch et al., 2010). An advantage of this approach is that no additional radon measurements are needed, a process which can be very costly and time consuming. However, mapping radon availability without taking into account radon measurements requires a very detailed knowledge about the local geology since uranium concentrations for the same geological unit can be very different at different locations (Schön, 2004). In addition to the amount of uranium, soils can be characterized by their amount of radon gas in order to estimate local radon hazards (Kemski et al., 2001). Fortunately, radon soil gas measurements do give direct information about the radon footprint of a geological unit (Barnet et al., 2010; Kemski et al., 2009). Nevertheless, it is difficult to transform uranium concentrations of the ground or radon soil gas measurements into an estimation of IRC, since IRC is additionally influenced by the architecture of the corresponding houses, permeability of the ground, the habits of inhabitants and meteorological variables (Bossew and Lettner, 2007; Cucoş (Dinu) et al., 2012; Groves-Kirkby et al., 2006; Kropat et al., 2014; Miles, 2001; Žunic et al., 2007).

A second approach is to directly estimate the local radon hazard from existing IRC measurements. Since natural radon concentrations are only of concern in closed environments, using IRC data to estimate local radon availability is the most direct way to estimate this health threat. Most national radon maps in Europe are based on local IRC averages on a grid or administrative boundaries (Dubois, 2005). Other studies report IRC mapping based on geological units (Appleton and Miles, 2010; Appleton et al., 2011; Drolet et al., 2014; Friedmann and Gröller, 2010; Smethurst et al., 2008). Furthermore, many authors made use of geostatistical techniques to interpolate IRC (Bossew et al., 2008, 2014; Cinelli et al., 2011; Dubois et al., 2007; Raspa et al., 2010; Zhu et al., 2001). However, few studies on IRC mapping were published which take the building characteristics and the measurement conditions into account (Borgoni et al., 2011; Pegoretti and Verdi, 2009).

The aim of this study was to develop models to predict and map IRC by accounting for the following all at the same time: the spatial relationships between IRC measurements and knowledge of measurement conditions, building characteristics of corresponding houses and geological information.

2. Data and methods

2.1. Data

2.1.1. IRC data

The IRC data used in this study are long-term IRC measurements carried out with passive radon detectors from the early 1980s up to now. For this purpose a laboratory sends the detectors with an enclosed questionnaire and an instruction sheet to the households. The homeowners expose the detectors over a given time period and send them back to the responsible laboratory with the completed questionnaire. The questionnaire indicates the address of the building, a unique ID of the Swiss national building registry, the coordinates, the altitude, the building type, the year of construction, if the building was previously remediated, the foundation type, the floor level of measurements, the room type of measurement, whether the measured room was inhabited during the measurement period or not, the exact dates of beginning and ending of the measurement period and the detector type.

In the beginning, the IRC measurements were taken randomly to obtain at least a minimum number of measurements per municipality. However, over time this strategy changed to targeted sampling of radon prone areas. The raw database available before data preprocessing consisted of 238 769 measurements in 148 458 houses. Analogous to many other studies, we performed all modeling and validations on log-transformed IRC measurements, in order to diminish the influence of extreme values (Andersen et al., 2007; Borgoni et al., 2011; Bossew et al., 2008; Cinelli et al., 2011; Dubois et al., 2007; Hauri et al., 2012; Zhu et al., 2001).

2.1.2. Coordinates

Many buildings of the IRC database don't possess correct coordinates. Often the coordinates of the houses are just the center of the municipality, in several cases the coordinates were outside the geographical area of Switzerland or on the top of mountains where no buildings exist. We only used measurement for which the coordinate was clearly indicated as the location of the house. To reduce possible errors we checked that within a distance of 100 m, there is actually a house of the Swiss building registry. The range of the east–west coordinate (EW-Coord) was 486295 m to 830883 m and 75881 m to 294237 m for the north–south coordinate (NS-Coord). The whole study was carried out in the Swiss coordinate system CH1903.

2.1.3. Detector type

In earlier work we found that detector types can have a substantial influence on the result of an IRC measurement (Kropat et al., 2014). In this study detector types were electret or track detectors from different manufacturers. Electret detectors may be subject to unforeseen discharge due to dust or air humidity. This can lead to overestimation of IRC. To account for differences in detector types, we took each detector type from each manufacturer as a unique class, which resulted in 13 detector type classes.

2.1.4. Floor level and inhabitation

In a previous study we found floor levels to have an important influence on IRC. The Swiss legislation only considers IRC measurements carried out in inhabited rooms as legally binding. Since most of the measurements have been carried out in the ground floor, we restricted this study to IRC measurements carried out in inhabited rooms of the ground floor. Around 50% of the measurements in the raw data base corresponded to this condition.

2.1.5. Year of construction

The year of construction is associated with several other variables like building materials and the method of construction. These variables have been shown to have a significant influence on IRC in several previous studies (Cucoş (Dinu) et al., 2012; Hauri et al., 2012; Kropat et al., 2014). Since detailed information about building materials and

the method of construction was not available, we took the year of construction into account as a proxy variable. However, in many cases homeowners only gave a vague estimation of a building's year of construction. To account for this uncertainty, we merged the years of construction into 4 classes rather than attributing the exact year of construction to each building. Since building materials changed over time we merged all buildings built before 1900 into the class "(1499, 1900)". The assumption that before 1900 natural stone was considerably more used in construction is in accordance with Gunby et al. (1993). As a second class we merged together all buildings between 1900 and 1970 into the class "(1900, 1970)". The methods of constructing buildings before and after 1970 considerably differ since the petrol crisis in 1970 led to substantial changes in building regulations. One consequence of the new building regulations was, for example, better insulation of the buildings against the subsoil (Burkart et al., 1984). Furthermore, we joined all buildings built between 1970 and 1990 into the class "(1970, 1990)". We assume that prefabricated houses were built differently before 1990. Furthermore, energy-saving construction techniques developed substantially in the last 20 years (Frei, 2013).

2.1.6. Building type

We merged the various building types into four classes: "Apartment building", "Detached houses", "Farm", "School" and "Other". In an earlier study we found these classes to have significantly different IRC mean values (Kropat et al., 2014).

2.1.7. Foundation type

It is very well known that the type of foundation has an influence on IRC of a building (Jelle, 2012; Mäkeläinen et al., 2001). Since we observed in a previous study (Kropat et al., 2014) that concrete foundation, earth foundations, and foundations that were concreted after construction show significantly different IRC, we took these classes into account resulting in the foundation type classes: "Concrete", "Concreted afterwards", "Earth" and "Other".

2.1.8. Altitude

According to earlier observations, the altitude is related to IRC in Switzerland (Kropat et al., 2014). This can be explained by the fact that geology depends on the altitude in Switzerland. For example igneous rocks like granites have a higher abundance at higher altitude in the Alps than at lower altitudes in the Swiss Plateau. However, the Swiss Plateau is mainly characterized by quaternary deposits. To account for this information, we included this variable into our models. To correct the altitude information of the measured buildings we used a digital elevation model (swisstopo, 2004) on a grid resolution of 25 m. To avoid uncertainty of the altitude indication from the questionnaire we sampled the altitudes of each house at the corresponding coordinates from the digital height model of Switzerland (swisstopo, 2004). The altitude in the final data set ranged from 193 m to 2434 m with 5% and 95% quantiles at 255 and 1122 m respectively.

2.1.9. Outdoor temperature

To estimate the outdoor temperatures we downloaded daily mean temperatures of 125 temperature stations which are uniformly distributed all over Switzerland. We downloaded the data for the last 30 years (MeteoSwiss, 2013). To calculate the mean outdoor temperature over the period of each measurement we interpolated daily mean temperatures of Swiss weather stations for each day of the measurement. Finally, we calculated the arithmetic mean of the estimated daily mean outdoor temperatures over the whole period of time. We used the same method to calculate outdoor temperatures as in an earlier study (Kropat et al., 2014). To interpolate the daily outdoor temperatures we used support vector regression (Cherkassky and Mulier, 2007; Smola and Schölkopf, 2004) by taking into account coordinates and altitudes as predictors. The support vectors were learned on the basis of the data of 125 weather stations. Support vector regression is a method that aims to find the flattest function $f\left(\vec{x}\right)$ that has no larger deviation than ε from the training observations y. In our case $\vec{x}$ are the independent variable coordinates and altitude and $f\left(\vec{x}\right)$ is an estimator of the temperature y. Keeping $f\left(\vec{x}\right)$ as flat as possible for a given training error maintains the best generalization properties to predict unknown temperatures. At the same time increasing flatness of the function may result in underfitting of the data, which consequently increases the training error. The goal is to find the optimum between flatness of function and training errors. To control this tradeoff a cost-parameter c is introduced. If the cost parameter c small, $f\left(\vec{x}\right)$ tends to underfit the data. If c is high $f\left(\vec{x}\right)$ tends to overfit the data. The cost parameter was optimized for each day by 5-fold-cross validation which resulted in an R^2 of 57% in the year 2000.

The final estimations of the temperature ranged between −4.5 °C and 21.2 °C with the 5% and 95% quantiles at −0.6 °C and 8.8 °C respectively.

2.1.10. Lithology

The information about the local lithological characteristics was extracted from a lithological map "Lithologic/petrographic map of Switzerland – Lithologic main groups 1:500 000" (SGTK, 2000) which is vectorized on a scale of 1:500 000. The lithological data consists of 70 classes. In many classes no IRC measurements have been taken. This poses problems with respect to mapping since the corresponding lithological classes would appear as holes on the final map. Therefore, we merged the original 70 lithological classes into 6 generalized lithological classes, which we supposed to be reasonable with respect to IRC concentrations: carbonate rock in the Jura mountains ("Carbonate Rock Jura"), carbonate rock in the alps ("Carbonate Rock Alps"), sediments ("Sediments"), sedimentary rock which is not carbonate rock ("Sedimentary Rock") (e.g. Sandstone, Conglomerate), metamorphic rock ("Metamorphic Rock") and igneous rocks ("Igneous Rock"). The same classification was used in Kropat et al. (2014).

2.2. Statistical modeling

2.2.1. Kernel regression

For the reader who is only interested in a very brief description of the kernel regression method, we refer to the last part of this chapter (Kernel regression in a nutshell).

2.2.1.1. Regression model. The issue of predicting IRC based on variables like measurement conditions, geological data can be stated in a common regression model (Racine and Li, 2004):

$$y = g\left(\vec{x}\right) + \varepsilon. \tag{1}$$

The variable y is the outcome variable of the model and $\vec{x}$ is a vector of all the predictors taken into account. $g(\cdot) = E\left(y|\vec{x}\right)$ is the conditional expectation of y given $\vec{x}$ and ε is a random error term. The conditional expectation $E\left(y|\vec{x}\right)$ can be written as (Racine and Li, 2004)

$$E\left(y|\vec{x}\right) = \frac{\sum_{i=1}^{N} Y_i K\left(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i\right)}{\sum_{i=1}^{N} K\left(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i\right)} \tag{2}$$

$K\left(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i\right)$ is a so called kernel function which weights each observation Y_i depending on the predictor variables $\vec{X}_i$ observed at each

observation, the predictor variables $\vec{x}$ indicating the point at which $E\left(y|\vec{x}\right)$ is estimated and vectors of smoothing parameters $\vec{\sigma}$ for continuous and $\vec{\lambda}$ for categorical variables. Each continuous variable is represented by an entry in $\vec{\sigma}$ and each categorical variable by an entry in $\vec{\lambda}$. We chose the optimal $\vec{\sigma}$ and $\vec{\lambda}$ based on cross-validation. $K\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)$ is a product kernel. That means that $K\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)$ is composed by the product of two kernels W for continuous and L for categorical variables.

$$K\left(\vec{\sigma},\vec{\lambda},\vec{x},\vec{X}_i\right)=W\left(\vec{x}^c,\vec{X}_i^c,\vec{\sigma}\right)L\left(\vec{x}^d,\vec{X}_i^d,\vec{\lambda}\right) \tag{3}$$

$\vec{X}_i^c$ is indicating the ith observations of the continuous predictors $\vec{x}^c$ and $\vec{X}_i^d$ corresponds to the ith observations of the discrete predictors $\vec{x}^d$.

2.2.1.2. Kernel for continuous variables. In this study we assumed a Gaussian kernel w for continuous variables (Specht, 1991)

$$w\left(\frac{x_t^c-X_{t,i}^c}{\sigma_t}\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_t^c-X_{t,i}^c}{\sigma_t}\right)^2} \tag{4}$$

where σ_t stands for the bandwidth and $X_{t,i}^c$ for the ith observation of the tth continuous variable x_t^c. The Gaussian kernel implies a certain spatial correlation structure in the case of the coordinates. Several other types of kernels could be taken into account, for example Gaussian or Epanechnikov kernels at higher orders. We chose the Gaussian kernel as described in Eq. (4) in this study since it performed best in preliminary tests and also because it is the most common one, which makes the approach more accessible for other researchers.

Since we took into account several continuous variables we considered a product kernel

$$W\left(\vec{x}^c,\vec{X}_i^c,\vec{\sigma}\right)=\prod_{t=1}^{p}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_t^c-X_{t,i}^c}{\sigma_t}\right)^2} \tag{5}$$

where p is the number of continuous variables. This product kernel does not account for covariance terms between two variables. We think however, that the simple product of two kernels should sufficiently model the correlation of two variables within the joint probability distribution function underlying the data generating process.

2.2.1.3. Kernel for categorical variables. For categorical variables we assumed the kernel (Aitchison and Aitken, 1976)

$$l\left(x_t^d,X_{t,i}^d,\lambda_t\right)=\begin{cases}1-\lambda_t & \text{if } X_{t,i}^d=x_t^d\\ \dfrac{\lambda_t}{c_t-1} & \text{if } X_{t,i}^d\neq x_t^d\end{cases} \tag{6}$$

where λ_t is the bandwidth, c_t the number of categories and $X_{t,i}^d$ the ith observation of the tth categorical variable x_t^d.

Since we dealt with several categorical variables we combined the kernels to a product kernel

$$L\left(\vec{x}^d,\vec{X}_i^d,\vec{\lambda}\right)=\prod_{t=1}^{k}l\left(x_t^d,X_{t,i}^d,\lambda_t\right) \tag{7}$$

where k is the number of categorical variables.

2.2.1.4. The role of bandwidths. The optimal bandwidth acts as a measure of importance of a predictor variable. The value of the bandwidths determines the weight that each observation gives to the prediction. Is the kernel weight $K(\cdot)$ in Eq. (2) the same for all observations, the prediction consists simply in the arithmetic mean of all observations.

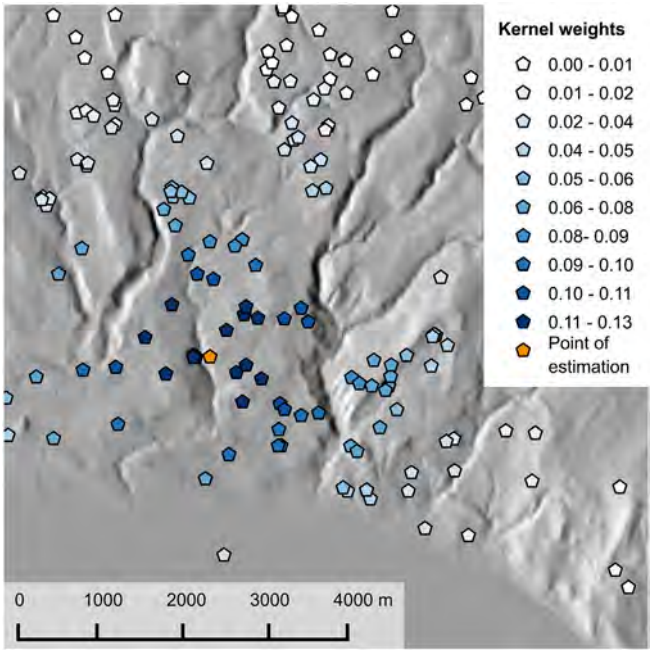


Fig. 1. Illustration of the decrease of kernel weights for each IRC measurement with increasing distance to the point of estimation (orange pentagon). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For categorical variables the bandwidth λ can attain a maximum value $\lambda_{max}=(c_t-1)/c_t$ for which the kernel weight $l(\cdot)$ is the same, no matter if an observation has the same class as the point of prediction or not. If $\lambda<\lambda_{max}$, the kernel weight $l(\cdot)$ takes on different values for each observation depending on the observation class. In this case the categorical predictor variable makes different predictions for different classes and consequently has an influence on the outcome variable.

A similar reasoning can be done for continuous variables. A larger optimal bandwidth σ results in a slower decrease of the kernel weight $w(\cdot)$ in dependence of the corresponding continuous variable. That means for a σ that is much larger than the extent of the continuous variable, the kernel weights $w(\cdot)$ are nearly the same for all observations. Consequently, the related variable doesn't influence the prediction. Note that after a distance of σ, the kernel weights $w(\cdot)$ decrease to 61% of its maximum value.

2.2.1.5. Kernel regression in a nutshell. Kernel regression aims at predicting a variable y by taking into account variables (predictor variables) that have an influence on y. In essence kernel regression consists of calculating a weighted average over previous observations of y. For each predictor variable a weighting function (kernel function) is multiplied to each observation of y. In the case of continuous variables this kernel function decreases with the distance between the point of prediction and the point of observation. In the case of categorical variables the kernel function changes depending on whether a point of estimation has the same class as the point of the observation or not. The kernel functions are controlled by bandwidths. Fig. 1 shows several IRC measurements (blue pentagons) which serve to predict a point were no measurement has been taken out (orange pentagon). To estimate the IRC value at the coordinate of the orange pentagon, kernel regression calculates a weighted average over the IRC values of the blue pentagon. The contribution (weight) of each IRC measurement to the average is indicated by different shades of blue. The weights decrease with the distance to the point of estimation (orange pentagon). The kernel bandwidth controls the decrease of the weights with the distance. In

this case the bandwidths are $\sigma_{EW} = 1922$ m and $\sigma_{SN} = 1514$ m. In the case of the Gaussian kernel function the weight decreases to 61% of its initial value after a distance of σ. That means that, in the special case of Fig. 1, the IRC measurement decreases to 61% of its actual value in EW-direction after around 2 km. The bandwidth σ serves hence as a measure of the range of an IRC measurement.

2.2.2. Probability estimation

The probability to exceed a certain IRC in a house can be obtained by the estimation of the conditional cumulative distribution function $F\left(y|\vec{x}\right)$.

2.2.2.1. Kernel estimation of conditional cumulative distribution function. The conditional cumulative distribution function (CDF) can be estimated in the following way (Li et al., 2013):

$$\hat{F}\left(y|\vec{x}\right) = \frac{1}{N}\sum_{i=1}^{N} I(Y_i \le y) K\left(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i\right) / \hat{f}\left(\vec{x}\right) \tag{8}$$

where $I(A)$ is an indicator function which equals to 1 when A is true and 0 otherwise. The optimal bandwidths $\vec{\sigma}$ and $\vec{\lambda}$ can be found via cross validation.

2.2.2.2. Confidence index and validation of conditional cumulative distribution function. Due to the nature of IRC sampling in Switzerland, the distribution of IRC samples is not homogenous all over Switzerland. It is therefore to be expected that the conditional probability estimate does not have the same reliability everywhere. To quantify the local reliability we developed a confidence index of the conditional cumulative distribution function.

The standard error of a proportion is known to take the following form (Diez et al., 2012)

$$SE = \sqrt{\frac{P(1-P)}{N}} \tag{9}$$

where P is the actual probability which is estimated by the proportion, and N is the sample size. The uncertainty of the probability estimate depends hence on the probability P and the sample size N. In the case of kernel estimation of conditional cumulative distribution functions it is however difficult to determine the sample size from which a local estimation $\hat{F}_a\left(y|\vec{x}\right)$ has been obtained. Instead of N we propose therefore to use the kernel sum

$$KS\left(\vec{x}\right) = \sum_{i=1}^{N} K\left(\vec{\sigma}, \vec{\lambda}, \vec{x}, \vec{X}_i\right) \tag{10}$$

as an equivalent for the local sample size, assuming, that the estimate should be more accurate by increasing $KS\left(\vec{x}\right)$.

Defining a new uncertainty estimate

$$UE\left(y|\vec{x}\right) = \sqrt{\frac{\hat{F}\left(y|\vec{x}\right)\left(1-\hat{F}\left(y|\vec{x}\right)\right)}{KS\left(\vec{x}\right)}} \tag{11}$$

we obtain an estimate which is in itself not a standard error but which we expect to take appropriately into account the probability estimate and the density of samples contributing to the local estimate. To define a confidence index from the uncertainty estimates, we determined the percentiles at 0, 10, 20, …, 100% of the obtained uncertainty estimates and numbered them decreasingly from 10 to 0. Hence, a confidence index of 10 corresponds to the least uncertainty on a map. In other words, there are no probability estimations on the probability map which would vary less if one would re-estimate the probability map with a similar set of IRC measurements. Whereas a confidence index of 5 indicates that half of the probabilities estimated elsewhere would vary more than the corresponding probability estimate.

For the kernel probability estimation we used around 50% of the IRC measurements that were available after data preprocessing as training data. The other 50% were used as test data for the validation of the kernel probability estimates. For this purpose we created a 5 km × 5 km grid over the whole of Switzerland and calculated the actual proportion of measurements above 300 Bq/m^3 based on the test data. Grid cells covering less than 20 measurements were excluded from the procedure. Furthermore we calculated the mean value of the kernel probability estimation for each grid cell based on the training data. By comparing both estimates in each grid cell, we obtained the amount of variance of the proportion estimates (test data) that is explained by the variation of the kernel probability estimates (training data).

2.2.2.3. Computational tools. For mapping and data analysis we used the open source tools R (R Core Team, 2014), Quantum Gis (QGIS Development Team, 2014) and GRASS (GRASS Development Team, 2012). The kernel bandwidth estimation was carried out on a 96 core compute cluster with the R package “npRmpi” (Hayfield and Racine, 2008). However, we did only use a part of the computational power of the cluster and want to point out that in the case of most national radon surveys the calculations can be done with much less performing hardware. For the temperature calculations we used the R package “e1071” (Meyer et al., 2014).

3. Results

Fig. 2a shows the current national IRC map of Switzerland that is based on municipal average IRC (FOPH, 2013). Fig. 2b displays the measurements that we used for mapping the model estimation in this study. After data preprocessing and removal of missing values, the number of IRC observations resulted to 72638. Around 8% of which were smaller or equal to 30 Bq/m^3, which is the minimum detectable IRC for track detectors (Zeeb and Shannoun, 2009b). The spatial density of measurement in the Alps is much smaller than in the Swiss Plateau or the Jura Mountains. The following maps were all calculated at an outdoor temperature of 3.5 °C and for the detector type Gammadata. 3.5 °C corresponds to the mean outdoor temperature of the measurements taken into account for analysis and Gammadata to the mode of detector types of the IRC data set. Two maps of the IRC kernel prediction are shown in Fig. 3. Each map corresponds to a different combination of building characteristics. Fig. 3a shows the map for detached houses with earth foundation built between 1900 and 1970. Fig. 3b corresponds to apartment buildings with concrete foundation, which were built between 1970 and 1990. Fig. 3b reveals clearly lower IRC than Fig. 3a. Both maps in Fig. 3 show that IRC are substantially higher in the Alps and in the Jura Mountains. Comparing Figs. 2b and 3 reveals that in alpine regions, where few IRC measurements were available, the IRC estimation in certain areas is dominated by a few measurements. Fig. 4 demonstrates the IRC mapping in the community of St. Imier. Fig. 4a shows the indoor radon map of Switzerland based on municipal IRC means. Fig. 4b reveals the measurements that have been used from this community to estimate the kernel bandwidths. The result of the kernel regression for St. Imier is shown in Fig. 4c for the classes: detached houses, concrete foundation and year of construction 1970–1990, Fig. 4d reveals the result of the kernel regression for classes: farms, earth foundation and year of construction 1900–1970 and Fig. 4e shows the lithological classes of this area. A clear difference in radon concentrations can be seen in Fig. 4c and d and the shape of the lithological units can be clearly observed in the kernel regression result.

Fig. 5a shows the result of the five-fold cross validation for the whole data set. The IRC measurements are plotted versus the IRC predictions. The kernel regression explains about 28% of the variance in the log-transformed IRC data. Restricting the test sets only to buildings of the type “Farms” results to an explained variance of 38%. The kernel

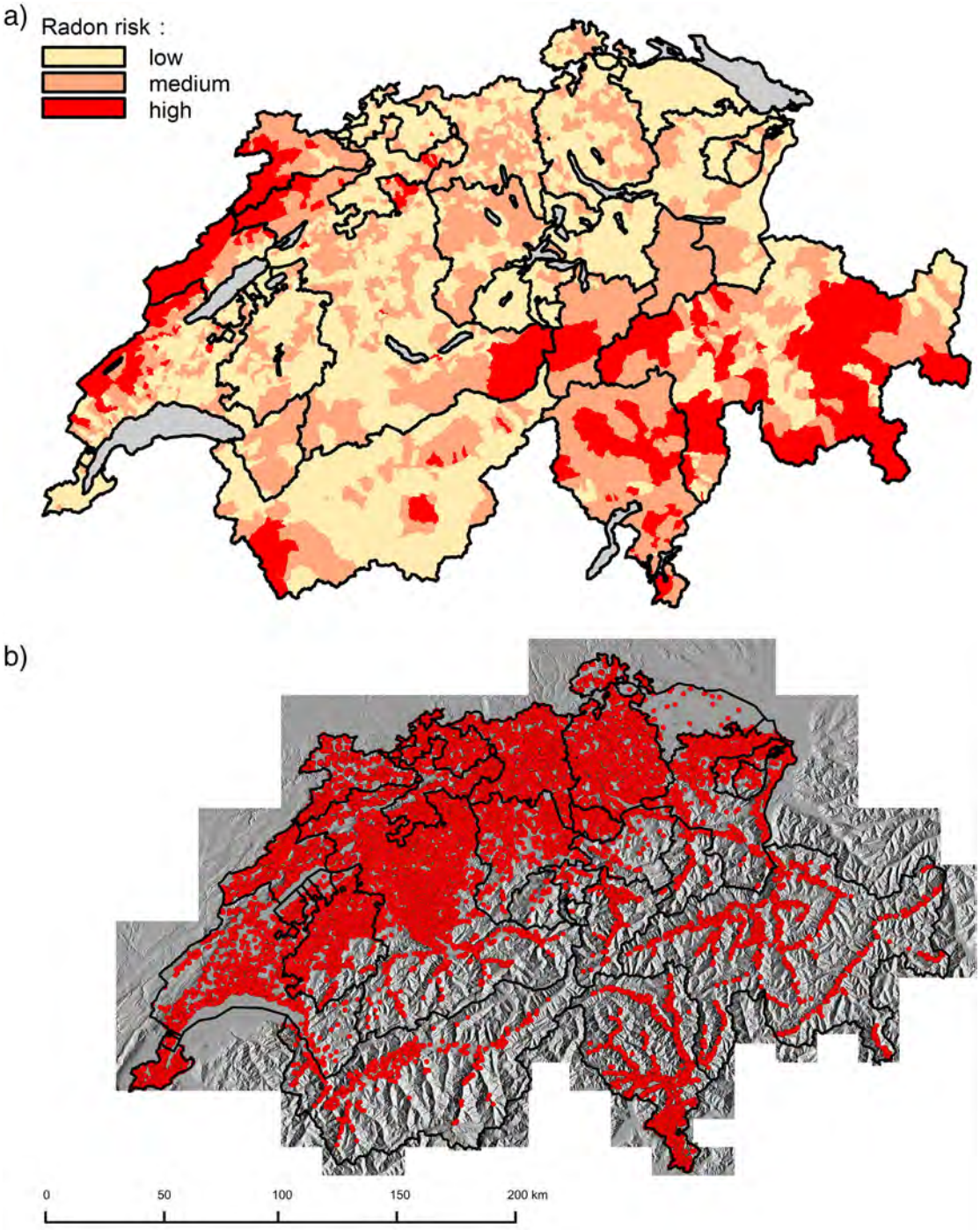


Fig. 2. a) Current IRC map of Switzerland (FOPH, 2013). (Radon risk in terms of the arithmetic municipality mean: low < 100 Bq/m^3, medium 100–200 Bq/m^3, high > 200 Bq/m^3), b) IRC measurements used to perform mapping in this study.

bandwidths estimated from the data are shown in Table 1. In the case of categorical variables the bandwidths are all considerably smaller than the maximum bandwidths. The bandwidths of the coordinates are in the range of about 1 km. The probability to exceed 300 Bq/m^3 is mapped in Fig. 6 for the classes of detached houses, concrete foundation and year of construction 1900–1970. The pattern of higher radon availability in the Jura Mountains and in the Alps can be observed in this map as well. The validation of the probability estimate resulted to an R^2 of 78%. Finally, Fig. 7 shows the confidence index of the probability estimate in Fig. 6. It can be seen that the reliability is higher in the valleys of the Alps compared to the regions of higher altitude.

4. Discussion

The aim of this study was to develop models to predict and map IRC by accounting at the same time for the spatial relationships between IRC measurements and knowledge of measurement conditions, building characteristics of corresponding houses and geological information. We developed predictive maps as well as a map indicating the local probability to exceed 300 Bq/m^3. Our predictive model could explain 28% of the variance of the IRC data. The validation of the probability estimation yielded an R^2 of 78%. Finally, we developed a confidence index in order to assess the local reliability of the probability map. These maps are appropriate to communicate IRC risk in Switzerland.

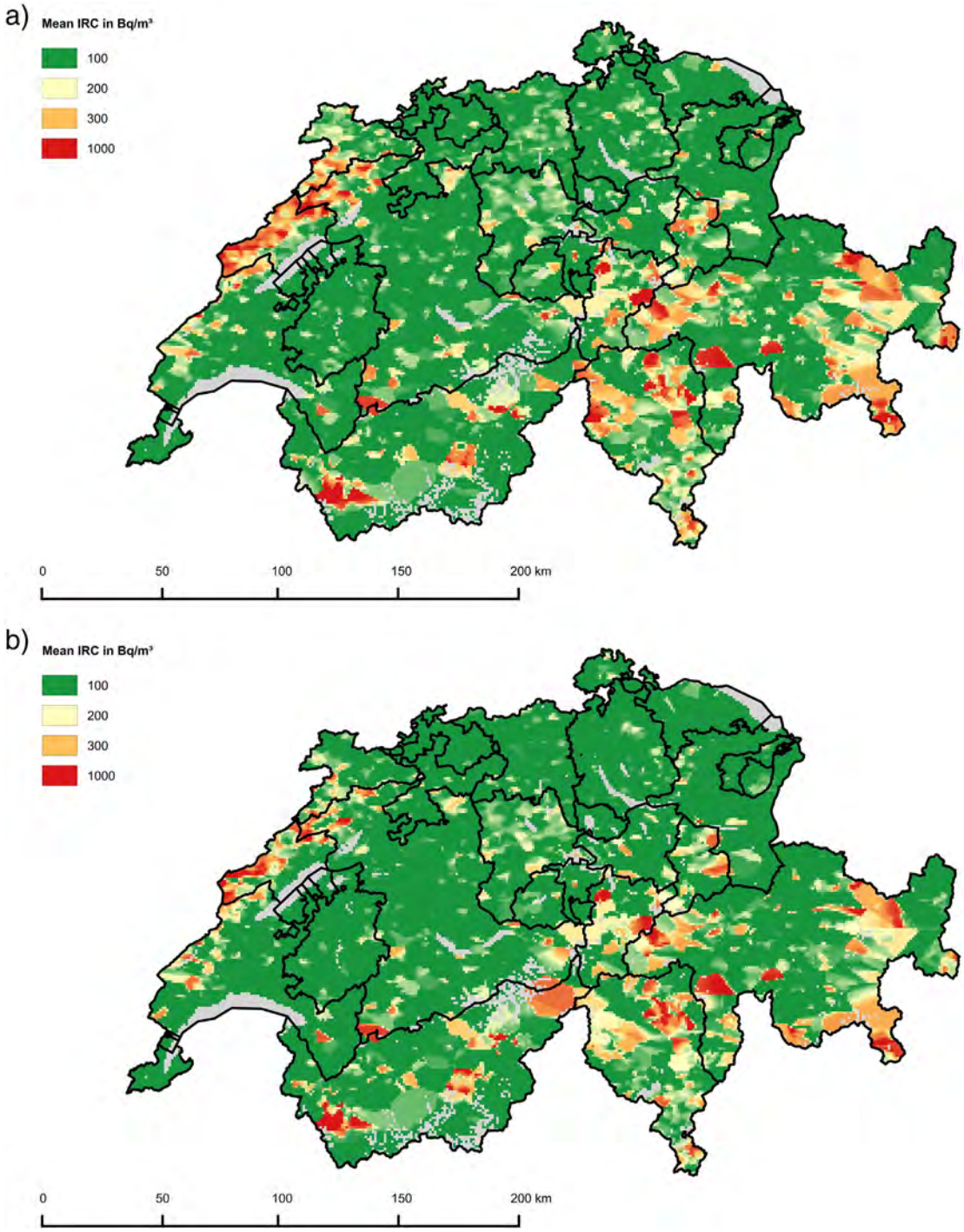


Fig. 3. Predictive map of IRC for the classes: a) detached houses, earth foundation, year of construction 1900–1970, detector type: Gammadata, outdoor temperature: 3.5 °C b) apartment building, concrete foundation, year of construction: 1970–1990, detector type: Gammadata, outdoor temperature: 3.5 °C.

Our results reproduce well the regional IRC differences which have been observed in earlier studies. The difference of IRC in the Alps, the Swiss Plateau and the Jura Mountains appears clearly in Fig. 3. This trend can be explained by major geological differences between these regions as mentioned by Hauri et al. (2012) and Kropat et al. (2014). However, the predictions in our model are only partly driven by the lithology variable. This can be seen in the corresponding bandwidth of 0.7 compared to the maximum bandwidth of 0.86. This is not surprising since the lithological data was only available on a very coarse scale, which results consequently to a higher geological misclassification rate. However, the lack of spatial detail of the lithological data is compensated by the spatial relationship between the IRC measurements. The bandwidth of the variable EW-Coord of 1081 m shows that the weight of measurements decreases to around 61% after this distance, the bandwidth in north–south direction was 719 m. This leads to the conclusion that radon data contains more information in north–south direction than in east–west direction, which may be due to stronger variations of the spatial trend of IRC between the Jura Mountains, the Swiss Plateau and the Alps. In regions that exhibit a stronger variation of spatial trend in EW-direction, like the canton of Ticino, the chosen bandwidth is maybe only suboptimal. We chose a global bandwidth for the sake of simplicity of the approach. However in future work local adaptive bandwidths may be considered. The kernel weights of the altitude decrease to 61% after a difference in elevation of about 164 m. A

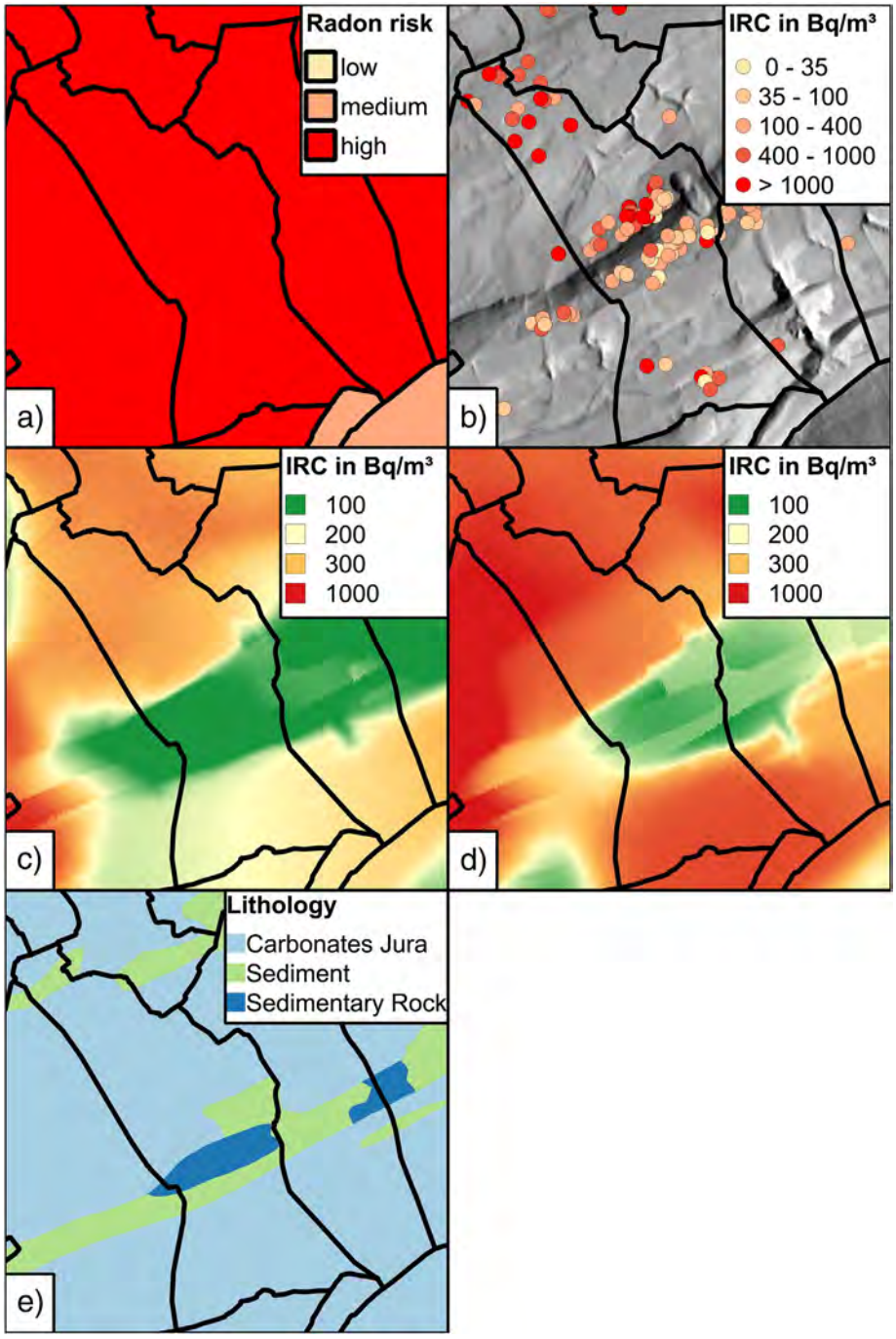


Fig. 4. Community of St. Imier: a) Swiss radon map b) radon measurements c) kernel regression radon map for classes: detached houses, concrete foundation, year of construction 1970–1990, detector type: Gammadata, outdoor temperature: 3.5 °C d) kernel regression radon map for classes: farms, earth foundation, year of construction 1900–1970, detector type: Gammadata, outdoor temperature: 3.5 °C e) lithological classes.

bandwidth of the altitude of 164 m indicates some predictive power regarding the fact that 90% of the data lie within 866 m. As discussed in Kropat et al. (2014) this may be attributable to the association between geology and altitude. Taken as a whole, the interpretation of the bandwidths suggests that coordinates and altitude act as a proxy for the lacking spatial detail in lithological data. The kernel regression comes in handy here, since very detailed information of local lithology would necessitate much more data to assure a reliable estimate of the regional IRC footprint.

Like many other countries, Switzerland possesses a radon risk map based on municipal IRC mean values (Fig. 2a). Overall, we obtained similar spatial IRC patterns. However, a strong benefit from kernel regression based methods is that trends within municipals can be accounted for. To illustrate this, we chose a municipality, in which IRC measurements exhibit a visible trend (Fig. 4b). Kernel regression accounts for this trend, resulting in more cartographical detail as can be seen in Fig. 4c and d. Fig. 4c and d maps two contrary cases: due to easier radon entry we expected different IRC characteristics for detached houses with concrete foundation built between 1970 and 1990 compared to farms with earth foundation built between 1900 and 1970. The differences of the bandwidth of the variables “Building type”, “Foundation” and “Year of construction” to their corresponding maximal bandwidth indicate a considerable difference in IRC estimation for the two abovementioned cases. This difference emerges clearly in Fig. 4c and d.

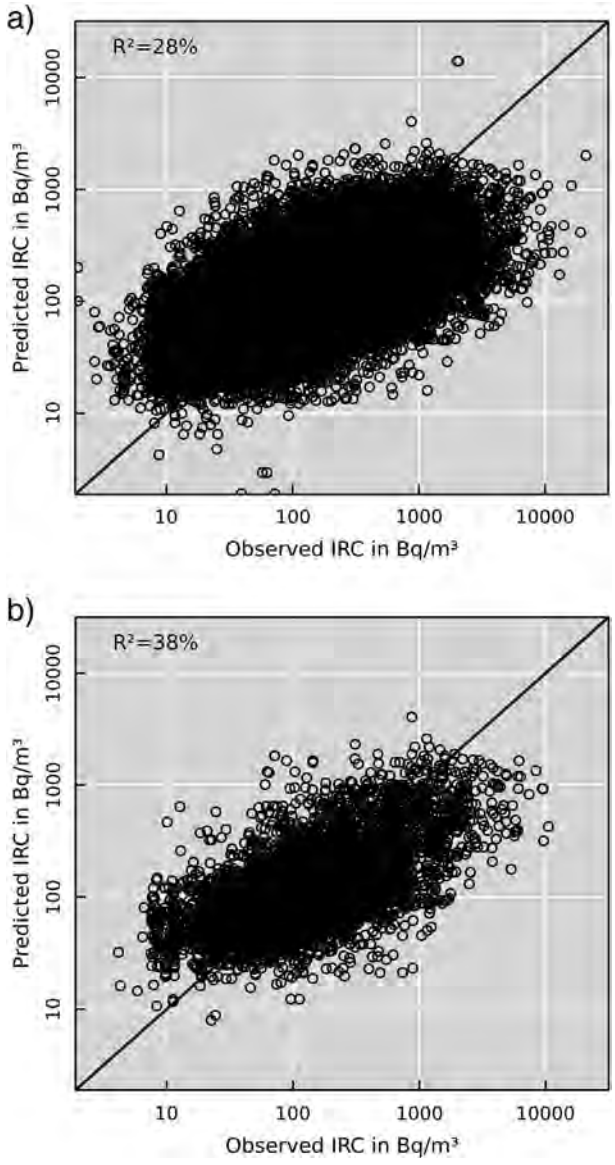


Fig. 5. Scatterplots of predicted versus measured radon concentrations: a) whole data set b) test sets only consisting of measurements in farms.

Furthermore, the structure of the lithological units is visible in the kernel regression results. With a bandwidth of $\lambda = 0.7$ compared to $\lambda_{max} = 0.86$, geology introduces information into the model. In other words, for the prediction at a given geographical coordinate with a given geology, measurements with the same distance and architectural elements contribute to a different extent depending on their underlying geology. For example, the prediction in a lithological unit of sedimentary rock is influenced by a measurement carried out in the same lithological unit with a kernel weight of 0.3. A measurement with the same distance in the lithological unit of carbonate rock, however, contributes only with a weight of 0.14. This result illustrates the benefit of kernel regression to account at the same time for the spatial relationship between IRC measurements as well as for geological information.

Table 1
Bandwidths of each variable taken into account.

Variable	λ or σ	c_t	λ_{max}
Building type	0.46	5	0.8
Foundation	0.46	4	0.75
Year of construction	0.31	4	0.75
Detector type	0.59	12	0.92
EW-Coord	1081	–	–
NS-Coord	719	–	–
Temperature	5.78	–	–
Altitude	164	–	–
Lithology	0.7	6	0.86

Countries that defined a geogenic radon potential, see Gruber et al. (2013), Kemski et al. (2001) and Neznal et al. (2004), could consider for the variable geology to take into account kernels for ordered categorical variables, see for example (Wang and van Ryzin (1981). This would avoid that geological units are treated as being equally dissimilar. However, since a geogenic radon potential is not defined in Switzerland and for simplicity of the approach we chose a kernel for unordered categorical variables for the lithology as described in Eq. (6).

Since 90% of the outdoor temperature estimations can be found within a range of 8.2 °C, the bandwidth of 5.8 °C for the outdoor temperature indicates that our models weakly account for outdoor temperature. Seasonal variations of IRC have often been observed (Arvela, 1995; Bossew and Lettner, 2007; Groves-Kirkby et al., 2010; Tapia et al., 2006). This can be explained by the fact, that houses are more often closed in winter than in summer. A further driving factor can be also pressure difference between indoor and outdoor. This pressure difference should be stronger at lower outdoor temperatures. Two arguments could explain the weak temperature influences in our model. Either the temperature influence is very different all over Switzerland, such that IRC variance due to temperature is overlaid by another source of variance, for example geology. Another is that our temperature estimations are subject to uncertainty. Including atmospheric pressure differences into IRC models could be revealing and possibly improve the predictability further. Unfortunately, this information was not available from the questionnaires. Hence, it might be interesting in future work to develop methods for the estimation of atmospheric pressure difference of existing IRC measurements.

Fig. 5a reveals, that the kernel regression model can explain 28% of the variation across all IRC data. This indicates that a large part of the data variation remains unexplained. It is visible in Fig. 5a that the model tends to overestimate low concentrations and underestimate high concentrations, which indicates a smoothing effect of the model. We suppose smoothing effects to be inherent in modeling of spatial random variables with high variance on small spatial scales as it was observed for IRC in earlier studies (Cinelli et al., 2011; Tapia et al., 2006). Another reason for the amount of unexplained variance is that not enough variables are available to explain more variance and that the variables taken into account are themselves subject to uncertainty like for example the house coordinates. Furthermore, it may be that the kernel regression model does not sufficiently account for the interaction effects between the variables. Moreover, the entry and the distribution of IRC in buildings may be a too complex process to be predicted with very high accuracy. Hauri et al. (2012) modeled IRC in Switzerland based on linear models and could explain 20% of the variation of IRC. This indicates that IRC in Switzerland are particularly difficult to model due to a complex geology and to different regional approaches to sample IRC, which leads to some heterogeneity of IRC sampling. We observed that 38% of the variation of IRC in farms can be explained by the kernel regression model. This remarkable difference to the predictability of the case of inclusion of all building classes may be explained by the fact that IRC measurements in farms are generally higher than in other houses (Kropat et al., 2014). Consequently, in farms, fewer IRC measurements are near the detection limit, which results in a better predictability.

It has been shown that the probability to exceed a certain value is a good way to complement risk communication instead of solely stating the mean value (Ibrekk and Morgan, 1987). Fig. 6 shows that kernel probability estimation is a useful tool to develop IRC probability maps. As can be expected, the spatial structure of this map is similar to the spatial structure of the predictive maps in Fig. 3. Probabilities near 100% to exceed IRC of 300 Bq/m^3 in Fig. 6 occur mainly in regions where only

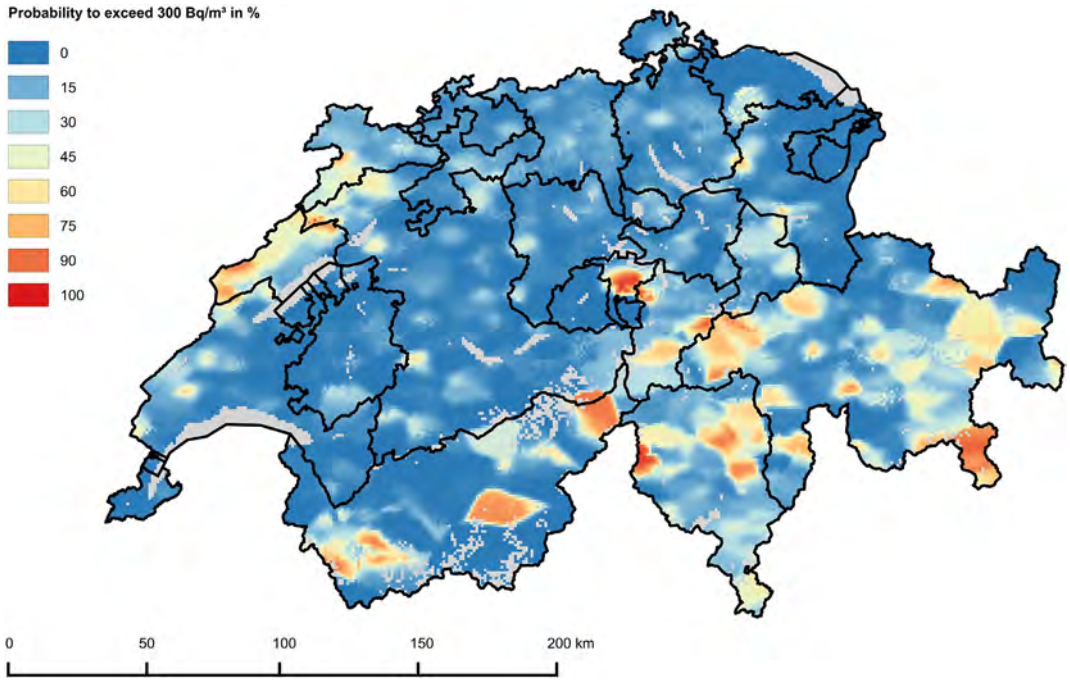


Fig. 6. Map of the local probability to exceed 300 Bq/m^3.

few measurements have been carried out. This means that the estimate in these regions is less reliable. Regions that are known for high IRC like the Jura Mountains and which have high sampling densities at the same time show probabilities closer to 50%. In order to evaluate the reliability of the probability estimate in Fig. 6 we created a confidence index map (Fig. 7). The Swiss topography is clearly visible on this map, which is reasonable, since the sampling density strongly decreases with altitude. However, regions with a similar altitude and different spatial densities of sampling are pointed out by the map. Finally, the variation in local sample size is more visible on this map than the variation in the probability estimate. This is because the magnitude of change of the sample size is much larger than the magnitude of change of the probability. The explained variance of 78% is indicating a good accordance between observed and estimated probability values on a grid of 5 km × 5 km. The reader has to bear in mind, that the explained variance in this case is not based on a point wise comparison of probability estimation and observation, but rather on the comparison of spatial aggregates. However, a point wise observation of probability is not feasible in our view since several houses cannot be on the same point.

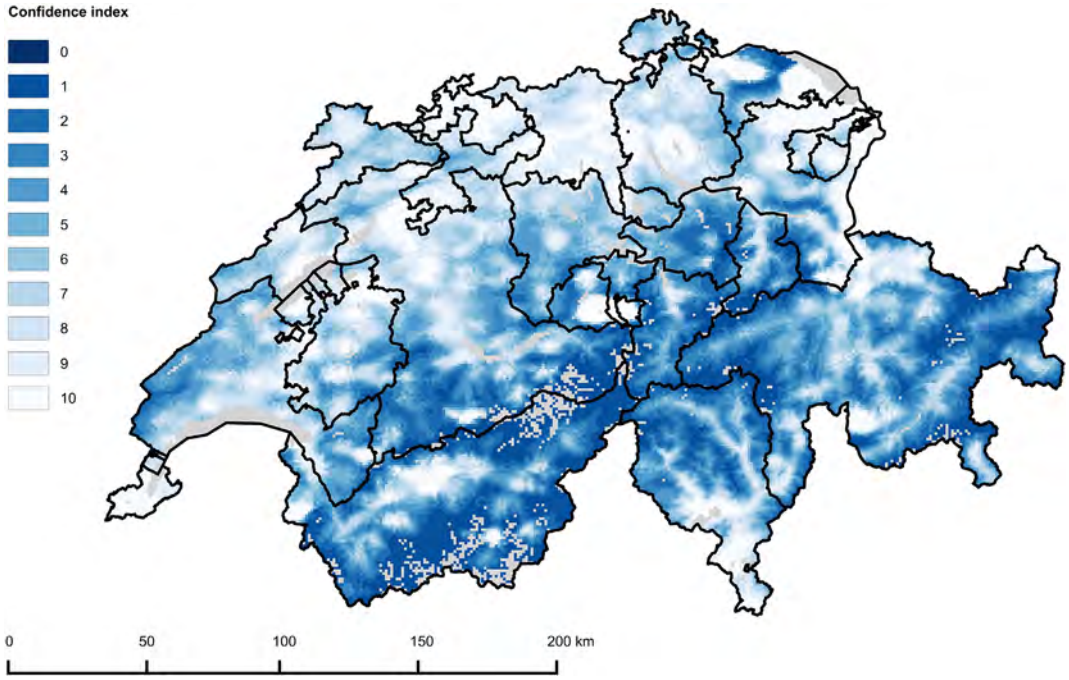


Fig. 7. Confidence index of the probability to exceed 300 Bq/m^3. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

In regions with few measurements like the Alps, the map in Fig. 3 shows coarse IRC patterns. This may be due to the fact that in these regions the IRC estimation of larger areas is dominated by just one measurement. However, for the communication of local radon risk we recommend making use of the IRC probability map. In this case the probability estimation is accompanied by a qualitative confidence index (Fig. 7), which prevents stakeholders from putting too much weight into potentially erroneous estimations.

Finally, we have to admit that the bandwidths of categorical and continuous variables do not directly quantify the predictive power of a variable and can thus not be directly compared. Nevertheless, the bandwidths provide useful information about how influential each predictor is.

Many studies were published, proposing IRC mapping based on geological boundaries (Appleton and Miles, 2010; Drolet et al., 2014; Friedmann and Gröller, 2010; Smethurst et al., 2008). Taking into account geological information as a main determinant for IRC is a natural choice since radon can be expected to come predominantly from the underlying ground into houses. On the other hand, spatial interpolation procedures like kriging have the advantage that they account for information about IRC stemming from the spatial correlation among neighbored measurements (Cinelli et al., 2011; Dubois et al., 2007; Zhu et al., 2001). IRC variation on small spatial scales can hence be better represented. However, Kemski et al.'s (2009) argument that interpolation techniques like kriging do not account for boundaries of geological units, which they consider being the primary source of geogenic radon. Bossew et al. (2008) address this issue by modeling radon potentials based on kriging by taking into account geological classes as external drift. Moreover Borgoni et al. (2011) use linear models to fit IRC based geological information and building characteristics and perform kriging on the resulting residuals. Finally Pegoretti and Verdi (2009) use a weighted k-nearest neighbor approach to fit IRC also based on geological information and building characteristics. We demonstrated an approach with the benefit that predictive as well as probability maps can be obtained for different architectural situations. The method takes into account geological information as well as the spatial relation among IRC measurements. Furthermore the method allows considering meteorological variables and it does not require linear relationships between IRC and predictor variables. In addition to that our approach has the advantage that a confidence index map can be easily obtained in order to evaluate local uncertainties about probability estimations.

5. Conclusions

Our results clearly show that kernel regression is a versatile tool to map, predict and analyze IRC on both national and municipal levels. The main advantage of kernel regression is to provide tailor-made maps for different architectural elements and measurement conditions. In addition to that, kernel regression takes into account at the same time geological information and spatial relations between IRC measurements. A big advantage of kernel regression is that is does not require linear relationships between IRC and predictor variables. Furthermore we found a reliable way to map the probability to exceed a given IRC. As in the case of kernel regression, the method allows the production of probability maps for different architectural elements. The future of this work consists in developing models that can explain a larger amount of the IRC variance. For that, it will be interesting to integrate other potentially relevant predictors like ventilation characteristics of houses, building materials, uranium content and permeability of the ground as well as radon soil gas concentration. Other meteorological variables like change of atmospheric pressure or precipitation levels could be also of interest as potential predictors. Furthermore, we need to explore other machine learning methods like random forests that account more strongly for interactions among predictor variables.

References

Aitchison J, Aitken CGG. Multivariate binary discrimination by the kernel method. Biometrika 1976;63(3):413–20. [Dec 1].

Andersen CE, Raaschou-Nielsen O, Andersen HP, Lind M, Gravesen P, Thomsen BL, et al. Prediction of ^{222}Rn in Danish dwellings using geology and house construction information from central databases. Radiat Prot Dosimetry 2007;123(1):83–94. [Jan 1].

Appleton JD, Miles JCH. A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk. J Environ Radioact 2010;101(10):799–803. [Oct].

Appleton JD, Miles JCH, Young M. Comparison of Northern Ireland radon maps based on indoor radon measurements and geology with maps derived by predictive modelling of airborne radiometric and ground permeability data. Sci Total Environ 2011;409(8):1572–83. [Mar 15].

Arvela H. Seasonal variation in radon concentration of 3000 dwellings with model comparisons. Radiat Prot Dosimetry 1995;59(1):33–42. [Jan 3].

Barnet I, Pacherová P, Preusse W, Stec B. Cross-border radon index map 1:100 000 Lausitz–Jizera–Karkonosze–Region (northern part of the Bohemian Massif). J Environ Radioact 2010;101(10):809–12. [Oct].

Borgoni R, Tritto V, Bigliotto C, De Bartolo D. A geostatistical approach to assess the spatial association between indoor radon concentration, geological features and building characteristics: the case of Lombardy, Northern Italy. Int J Environ Res Public Health 2011;8(5):1420–40.

Bossew P, Lettner H. Investigations on indoor radon in Austria, part 1: seasonality of indoor radon concentration. J Environ Radioact 2007;98(3):329–45.

Bossew P, Dubois G, Tollefsen T. Investigations on indoor radon in Austria, part 2: geological classes as categorical external drift for spatial modelling of the radon potential. J Environ Radioact 2008;99(1):81–97. [Jan].

Bossew P, Žunić ZS, Stojanovska Z, Tollefsen T, Carpentieri C, Veselinović N, et al. Geographical distribution of the annual mean radon concentrations in primary schools of Southern Serbia — application of geostatistical methods. J Environ Radioact 2014;127:141–8. [Jan].

Burkart W, Wernli C, Brunner HH. Matched pair analysis of the influence of weatherstripping on indoor radon concentration in Swiss dwellings. Radiat Prot Dosimetry 1984;7(1–4):299–302. [Jan 1].

Cherkassky V, Mulier FM. Support vector machines. Learn data concepts theory methods. John Wiley & Sons; 2007. p. 404–66.

Cinelli G, Tondeur F, Dehandschutter B. Development of an indoor radon risk map of the Walloon region of Belgium, integrating geological information. Environ Earth Sci 2011;62(4):809–19. [Feb 1].

Cucoș (Dinu) A, Cosma C, Dicu T, Begy R, Moldovan M, Papp B, et al. Thorough investigations on indoor radon in Băița radon-prone area (Romania). Sci Total Environ 2012;431:78–83. [Aug 1].

Diez DM, Barr CD, Cetinkaya-Rundel M. Inference for categorical data. OpenIntro Stat [internet]. 2nd ed. CreateSpace Independent Publishing Platform; 2012. p. 263–314 [cited 2014 Sep 1, Available from: http://www.openintro.org/stat/textbook.php].

Drolet J-P, Martel R, Poulin P, Dessau J-C. Methodology developed to make the Quebec indoor radon potential map. Sci Total Environ 2014;473–474:372–80. [Mar 1].

Dubois G. An overview of radon surveys in Europe. European Commission; 2005.

Dubois G, Bossew P, Friedmann H. A geostatistical autopsy of the Austrian indoor radon survey (1992–2002). Sci Total Environ 2007;377(2–3):378–95. [May 15].

FOPH. Radon risk in Switzerland [internet]. Federal Office of Public Health FOPH; 2013 [Available from: http://www.bag.admin.ch/themen/strahlung/00046/11952/index.html?lang=en].

Frei K. Zehn Jahre Minergie. NOVA — Z Für Wärmedämmung Solarenergie [Internet]; 2013. p. 26 [Available from: http://www.flumroc.ch/de/aktuell/zeitschrift_nova.php#].

Friedmann H, Gröller J. An approach to improve the Austrian radon potential map by Bayesian statistics. J Environ Radioact 2010;101(10):804–8. [Oct].

GRASS Development Team. Geographic resources analysis support system (GRASS GIS) software [internet]. Open Source Geospatial Foundation; 2012 [Available from: http://grass.osgeo.org].

Groves-Kirkby CJ, Denman AR, Crockett RGM, Phillips PS, Gillmore GK. Identification of tidal and climatic influences within domestic radon time-series from Northamptonshire, UK. Sci Total Environ 2006;367(1):191–202. [Aug 15].

Groves-Kirkby CJ, Denman AR, Phillips PS, Crockett RGM, Sinclair JM. Comparison of seasonal variability in European domestic radon measurements. Nat Hazards Earth Syst Sci 2010;10(3):565–9. [Mar 26].

Gruber V, Bossew P, Cort MD, Tollefsen T. The European map of the geogenic radon potential. J Radiol Prot 2013;33(1):51. Mar 1.

Gunby JA, Darby SC, Miles JCH, Green BMR, Cox DR. Factors affecting indoor radon concentrations in the United Kingdom. Health Phys 1993;64(1):2–12.

Hauri DD, Huss A, Zimmermann F, Kuehni CE, Röösli M. A prediction model for assessing residential radon concentration in Switzerland. J Environ Radioact 2012;112:83–9. [Oct].

Hayfield T, Racine JS. Nonparametric econometrics: the np package. J Stat Softw 2008;27(5). [Internet, available from: http://www.jstatsoft.org/v27/i05/].

Ibrekk H, Morgan MG. Graphical communication of uncertain quantities to nontechnical people. Risk Anal 1987;7(4):519–29.

Ielsch G, Cushing ME, Combes P, Cuney M. Mapping of the geogenic radon potential in France to improve radon risk management: methodology and first application to region Bourgogne. J Environ Radioact 2010;101(10):813–20. [Oct].

Jelle BP. Development of a model for radon concentration in indoor air. Sci Total Environ 2012;416:343–50. [Feb 1].

Kemski J, Siehl A, Stegemann R, Valdivia-Manchego M. Mapping the geogenic radon potential in Germany. Sci Total Environ 2001;272(1–3):217–30. [May 14].

Kemski J, Klingel R, Siehl A, Valdivia-Manchego M. From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. Environ Geol 2009;56(7):1269–79. [Feb 1].
Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios M, et al. Major influencing factors of indoor radon concentrations in Switzerland. J Environ Radioact 2014;129:7–22. [Mar].
Li Q, Lin J, Racine JS. Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. J Bus Econ Stat 2013;31(1):57–65.
Mäkeläinen I, Arvela H, Voutilainen A. Correlations between radon concentration and indoor gamma dose rate, soil permeability and dwelling substructure and ventilation. Sci Total Environ 2001;272(1–3):283–9. [May 14].
Menzler S, Piller G, Gruson M, Rosario AS, Wichmann H-E, Kreienbrock L. Population attributable fraction for lung cancer due to residential radon in Switzerland and Germany. Health Phys 2008;95(2):179–89. [Aug].
MeteoSwiss. MeteoSwiss. Federal Office of Meteorology and Climatology; 2013.
Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: misc functions of the Department of Statistics (e1071), TU Wien [internet]; 2014 [Available from: http://CRAN.R-project.org/package=e1071].
Miles J. Temporal variation of radon levels in houses and implications for radon measurement strategies. Radiat Prot Dosimetry 2001;93(4):369–75. [Jan 2].
Neznal M, Neznal M, Matolin M, Barnet I, Miksova J. The new method for assessing the radon risk of building sites. Czech Geol Surv Spec Pap [Internet]; 2004. p. 16. [cited 2014 Sep 1, Available from: http://www.geology.cz/spec-papers/obsah/no16/16-1.pdf].
Pegoretti S, Verdi L. Machine learning for the analysis of indoor radon distribution, compared with ordinary kriging. Radiat Prot Dosimetry 2009;137(3–4):324–8. [Jan 12].
QGIS Development Team. QGIS geographic information system [internet]. Open Source Geospatial Foundation; 2014 [Available from: http://qgis.osgeo.org].
R Core Team. R: a language and environment for statistical computing [internet]. Vienna, Austria: R Foundation for Statistical Computing; 2014 [Available from: http://www.R-project.org/].
Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data. J Econ 2004;119(1):99–130. [Mar].
Raspa G, Salvi F, Torri G. Probability mapping of indoor radon-prone areas using disjunctive kriging. Radiat Prot Dosimetry 2010;138(1):3–19. [Jan 1].
Schön JH. Natural radioactivity of rocks. Phys Prop Rocks Fundam Princ Petrophysics. Elsevier; 2004. p. 107–32.
SGTK. Lithologisch-petrografische Karte der Schweiz-Lithologie-Hauptgruppen 1: 500000. Schweizerische Geotechnische Kommission. 2000.
Smethurst MA, Strand T, Sundal AV, Rudjord AL. Large-scale radon hazard evaluation in the Oslofjord region of Norway utilizing indoor radon concentrations, airborne gamma ray spectrometry and geological mapping. Sci Total Environ 2008;407(1):379–93. (Dec 15).
Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14(3):199–222. (Aug 1).
Specht DF. A general regression neural network. Neural Netw IEEE Trans 1991;2(6):568–76.
swisstopo. DHM25. The digital height model of Switzerland [internet]. Federal Office of Topography; 2004 [Available from: http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/height/dhm25.html].
Tapia R, Kanevski M, Maignan M, Gruson M. Comprehensive multivariate analysis of indoor radon data in Switzerland. 8th Int Work "Geological Asp Radon Risk Mapping" Prague; 2006. p. 26–30.
Wang M-C, van Ryzin J. A class of smooth estimators for discrete distributions. Biometrika 1981;68(1):301–9. [Apr 1].
Zeeb H, Shannoun F. Health effects of radon. WHO handb indoor radon public heal perspect [internet]. Geneva: World Health Organization (WHO); 2009a. p. 3–20 [cited 2013 Dec 10, Available from: http://www.who.int/iris/handle/10665/44149].
Zeeb H, Shannoun F. Radon measurements. WHO handb indoor radon public heal perspect [internet]. Geneva: World Health Organization (WHO); 2009b. p. 21–40 [cited 2013 Dec 10, Available from: http://www.who.int/iris/handle/10665/44149].
Zhu H, Charlet J, Poffijn A. Radon risk mapping in southern Belgium: an application of geostatistical and GIS techniques. Sci Total Environ 2001;272(1–3):203–10. [May 14].
Žunic ZS, Yarmoshenko IV, Birovljev A, Bochicchio F, Quarto M, Obryk B, et al. Radon survey in the high natural radiation region of Niška Banja, Serbia. J Environ Radioact 2007;92(3):165–74.

Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units

Georg Kropat[a], Francois Bochud[a], Michel Jaboyedoff[c], Jean-Pascal Laedermann[a], Christophe Murith[b], Martha Palacios (Gruson)[b], Sébastien Baechler[a,b]

[a]Institute of Radiation Physics, Lausanne University Hospital, Rue du Grand-Pré 1, 1007 Lausanne, Switzerland

[b]Swiss Federal Office of Public Health, Schwarzenburgstrasse 165, 3003 Berne, Switzerland

[c] Faculty of Geosciences and Environment, University of Lausanne, GEOPOLIS - 3793, 1015 Lausanne, Switzerland

Corresponding Author:
Georg Kropat
Institute of Radiation Physics
University Hospital Center of Lausanne (CHUV)
Rue du Grand-Pré 1

1007 Lausanne
Tel.: + 41 21 314 82 96
Fax : + 41 21 314 82 99
Email: georg.kropat@chuv.ch

Abstract

Purpose: In Switzerland each year approximately 230 people die as a result of radon exposure. This public health concern makes reliable indoor radon prediction and mapping methods necessary in order to improve risk communication to the public. The aim of this study was to develop an automated method to classify lithological units according to their radon characteristics and to develop mapping and predictive tools in order to improve local radon prediction.

Method: About 240 000 indoor radon concentration (IRC) measurements in about 150 000 buildings were available for our analysis. The automated classification of lithological units was based on k-medoids clustering via pair-wise Kolmogorov distances between IRC distributions of lithological units. For IRC mapping and prediction we used random forests and Bayesian additive regression trees (BART).

Results: k-medoids clustering of lithological units based on pair-wise Kolmogorov distances of IRC distribution accounts well for the IRC characteristics of lithological units. Especially the heterogeneity in metamorphic rocks like gneiss is well resolved by this method. The maps produced by random forests soundly represent the regional difference of IRC in Switzerland and improve the spatial detail compared to existing approaches. We could explain 33% of the variations in IRC data with random forests. Additionally, the variable importance evaluated by random forests shows that building characteristics are less important predictors for IRC than spatial/geological influences. BART could explain 29% of IRC variability and produced maps that indicate the prediction uncertainty.

Conclusion: Ensemble regression trees are a powerful tool to model and understand the multidimensional influences on IRC. Automatic clustering of lithological units increases the value of this method by facilitating the interpretation of radon properties of rock types. This study provides an important element for radon risk communication. Future approaches should consider taking into account further variables like indoor and outdoor pressure differences as well as more detailed geological information.

1. Introduction

Radon is a natural radioactive gas that is known to be the most important cause of lung cancer after smoking. In Switzerland, about 230 people die each year as a result of radon exposure (Menzler et al. 2008). Many of these deaths could be avoided if public radon exposure could be effectively reduced.

Radon exposure is mainly of concern in closed environments like buildings. Since radon mainly enters a building from the ground (Zeeb and Shannoun 2009b), it is strongly dependent on the underlying geology (Dubois et al. 2007; Cinelli et al. 2009; Appleton and Miles 2010; Friedmann and Gröller 2010; Bossew et al. 2014). Thus, IRC vary strongly from region to region (Friedmann et al. 1996; Manic et al. 2006). If these regional differences can be identified correctly, substantial reductions of radon exposure to the population can be achieved through the appropriate construction of new buildings, the mitigation of already existing buildings and local smoking cessation campaigns (Groves-Kirkby et al. 2008, 2011). The prediction of IRC and identifying zones at risk is however difficult and still subject to scientific debate (Friedmann and Bossew 2010).

IRC are known to be subject to several sources of variance. Building architecture is an important factor. In several studies the influence of the building underground on IRC was observed (Burkart et al. 1984; Mäkeläinen et al. 2001; Kropat et al. 2014). Construction materials are known to play an important role for the occurrence of IRC in buildings (Gunby et al. 1993; Girault and Perrier 2012; Demoury et al. 2013). Furthermore, the ventilation habits of people are an important factor of IRC (Gunby et al. 1993). This leads to visible seasonal effects (Singh et al. 2002; Bossew and Lettner 2007; Denman et al. 2007; Groves-Kirkby et al. 2010; Trevisi et al. 2010). In winter time, buildings are generally less ventilated than in summer. In addition to that, indoor/outdoor pressure differences can lead to stack effects. IRC dependencies on weather conditions also have been observed (Miles 2001). Finally, one of the most discussed determinants of IRC is the geology subjacent to the

concerned buildings. IRC is known to be strongly dependent on geological parameters like uranium content, permeability of the ground as well as soil properties (Buchli and Burkart 1989; Gundersen and Schumann 1996; Singh et al. 2002; Neznal 2005; Bossew et al. 2008; Appleton et al. 2011). However, finding geological units which are appropriate to represent the local IRC characteristic is still subject to debate in the scientific community (Tondeur et al. 2014). Many studies have suggested generalizing the geological units in order to simplify analysis and modeling of IRC with respect to geological information (Kemski et al. 2001; Miles and Appleton 2005; Bossew et al. 2008; Smethurst et al. 2008; Kropat et al. 2014, 2015). Most of these approaches were based on the generalization into standard geological categories like metamorphic, igneous and sedimentary rock as well as quaternary geology. Others took into account the direct information of the uranium content of rocks obtained via airborne gamma ray spectrometry (Smethurst et al. 2008).

The aim of this study was twofold: We developed a data driven method to classify lithological units based on their similarity in terms of IRC distribution. Furthermore, we performed IRC prediction and mapping based on ensemble regression trees by accounting for the following variables: building coordinates, altitude, building type, foundation type, year of construction, detector type, clustered lithological units and temperature.

2. Methods

2.1. Data and predictor variables

2.1.1. IRC data

In Switzerland, long term IRC measurements have been carried out since the early 1980s resulting in a total of 238,769 measurements in 148,458 buildings. In the beginning the sampling strategy was to target radon prone areas. It then evolved in order to reach a minimum number of samples per municipality. To perform the measurements, local laboratories sent IRC detectors to the homeowners.

Upon reception, the homeowners exposed the detectors in their buildings and sent them back once the measurement period was completed. The mean duration of measurements was about 3 months. The measurements were accompanied by a questionnaire in which the homeowner gave details about measurement conditions and architectural characteristics of the measured building. IRC measurements in Switzerland only have a legal implication if they have been carried out in an inhabited room. Hence we restricted our study to measurements that were carried out in inhabited rooms on the ground floor of the concerning buildings. Most of the measurements in inhabited rooms were carried out in ground floors. About 30% of IRC measurements from the raw data base corresponded to this criterion. Like many other studies , we carried out analysis, mapping and validation on log-transformed IRC in order to avoid the influence of extreme values and to stay comparable to other approaches (Zhu et al. 2001; Andersen et al. 2007; Dubois et al. 2007; Bossew et al. 2008; Borgoni et al. 2011; Cinelli et al. 2011; Hauri et al. 2012).

2.1.2. Detector types

The IRC measurements used in this study were mainly performed with alpha track and electret detectors (Zeeb and Shannoun 2009a). We observed in an earlier study that IRC measurements substantially differ between these two detector types (Kropat et al. 2014). To account for this fact, we considered detector types as an IRC predictor variable. Finally, the variable "Detector type" contained 12 classes which differed by vendor and detector type.

2.1.3. Coordinates

The classification of the coordinates in the IRC database consists of three classes: municipal coordinates, buildings coordinates and coordinates determined by the Swiss Federal Statistical Office (FSO). The FSO coordinates are the most reliable ones, since the data quality is directly guaranteed by the FSO. The municipal coordinates indicate the center of the buildings municipality. These coordinates are the least precise because a single value defines the location of all the building of the

concerned municipality. Finally, the building coordinates are the coordinates which have been indicated by the building owners or by the laboratory conducting the corresponding IRC measurement. Hence the building coordinates are more prone to uncertainty than the FSO coordinates. In this study we only used FSO as owner-declared coordinates. As a quality control, we checked that all building coordinates were not in regions which are actually not populated such as lakes or mountain summits. Furthermore, we verified that each building coordinate was in the attributed municipality and that there was at least one building from the national building registry (FSO 2014) in a vicinity of 100 m. The coordinates were reported in the Swiss coordinate System CH1903.

In order to make the geographical orientation in Switzerland easier for the reader, Figure 1 indicates all locations, cantons and geological regions which are mentioned in this article .

2.1.4. Altitude

The homeowners were asked to indicate the altitude of the measured buildings. To reduce uncertainty of the altitude indication we sampled the altitude for each building from a digital elevation model with a resolution of 25 m (swisstopo 2004) based on the building coordinates.

2.1.5. Lithology

To determine the underlying lithology of each building we used a vector map of lithological classes in Switzerland (SGTK 2000). This map contains about 70 lithological classes and is vectorized on a scale of 1:500 000. Based on the building coordinates we sampled the lithological class for each building from this map.

2.1.6. Building type

We divided the type of building into the following 5 classes: “Detached Houses”, “Apartment Building”, “Farm”, “School” and “Other”. Information about the building type was obtained from the questionnaire filled in by the homeowner.

2.1.7. Year of construction

We showed in an earlier study that IRC are associated with a building's year of construction (Kropat et al. 2014). We suppose that the construction of buildings built before 1900 more often involved natural stones than the constructions after 1900 (Gunby et al. 1993). Furthermore, we assume that building regulations changed substantially after 1970 following the oil crisis, a situation that resulted in better insulation of buildings against subsoil (Burkart et al. 1984). Finally, we presume that processes for constructing readymade buildings as well as energy saving measures have changed since 1990 (Frei 2013). Following this logic, we divided the year of construction into the classes: "(1499 – 1900]", "(1900 – 1970]", "(1970 – 1990]", "(1990 – 2012]". The same classification was used in (Kropat et al. 2014, 2015). Information about a building's year of construction was obtained from the questionnaire filled in by the homeowner.

2.1.8. Foundation type

Since a building's foundation type is known to influence IRC (Mäkeläinen et al. 2001; Jelle 2012; Kropat et al. 2014), we divided the foundation types into 4 classes: "Concrete", "Concreted afterwards", "Earth", "Other". Information about the foundation type was obtained from the questionnaire filled in by the homeowner.

2.1.9. Outdoor temperature

In an earlier study, we found that IRC was associated with outdoor temperature (Kropat et al. 2014). In this study, we used the same method to estimate the mean outdoor temperature. For this purpose, we used the daily mean temperatures of the last 30 years for about 125 Swiss weather stations that we downloaded from the database of the Federal Office of Meterology and Climatology MeteoSwiss (MeteoSwiss 2013). To infer the mean temperature of each day of an IRC measurement we interpolated the mean temperatures of the weather station for the corresponding days based on support vector regression. We used the coordinates and the altitude of a building as predictor variables. A

detailed description about support vector regression can be found in (Smola and Schölkopf 2004; Cherkassky and Mulier 2007). Support vector regression is a method used to model the relation between an outcome variable y and some predictor variables $\vec{x}$ with $y = f(\vec{x}) + \eta$. $f(\vec{x})$ represents the functional relation between $\vec{x}$ and y and η a random error. For this study, y corresponds to the daily mean temperatures and $\vec{x}$ to a vector containing the coordinates and the altitude. For the sake of simplicity we describe the principle of support vector regression in general for the case of a one-dimensional predictor variable x. The goal of support vector regression is to approximate the function $f(\cdot)$ by accounting at the same time for several conditions: data observations (x_i, y_i) do not contribute to the estimation procedure when $|f(x_i) - y_i| < \varepsilon$ for a given $\varepsilon > 0$. The region around $f(\cdot)$ with $|f(x_i) - y_i| < \varepsilon$ is also called ε-insensitive tube. A further restriction is that observations y_i outside of the ε-insensitive tube shall not have a larger distance than $|\xi_i| + \varepsilon$ to $f(x_i)$. Finally $f(\cdot)$ is supposed to be as flat as possible. The optimum of $f(\cdot)$ under these conditions can be found by means of Lagrange formalism and the Karush-Kuhn-Tucker-theorem (Cherkassky and Mulier 2007). The tradeoff between the flatness of $f(\cdot)$ and the fit contributions of observations y_i lying far away from $f(x_i)$ is controlled by the cost parameter C. A large C strongly penalizes large deviations $|\xi_i|$. Observations y_i with large $|\xi_i|$ have hence a stronger influence on the determination of $f(\cdot)$. If C is small, more importance is given to the flatness of the curve. C is hence a parameter that determines the tradeoff between over- and underfitting of $f(\cdot)$. We determined C for each day via 5-fold-cross validation. For example, the explained variance for the year 2000 resulted in average to 57%. The mean outdoor temperature of a measurement was finally determined by predicting the mean outdoor temperature of each day of an IRC measurement and calculating the temperature average over all days.

2.2. Clustering of lithological units

The IRC characteristics of a lithological unit depend on a variety of different parameters like the type, fracturing, dissolution as well as the uranium and radium content of the rock (Smethurst et al. 2008; Kemski et al. 2009; Appleton et al. 2011). These parameters can locally vary themselves for a given lithological unit. It is hence not trivial to group lithological units according to their IRC characteristics based only on prior assumptions of their lithological properties. In this work we developed a method to group lithological units based on their similarity in IRC characteristics. This could be done simply by comparing their mean values. This however does not account for the difference in variability of the IRC between several lithological units. We expect for example that, due to rock fracturing and dissolution, the variability in karstic regions is higher than in quaternary deposits. In this study we grouped the lithological units according to their similarity in IRC distribution measured via the Kolmogorov distance (Györfi et al. 1996). The Kolmogorov distance measures the maximum difference in probability of two cumulative distribution functions. Based on this similarity measure a similarity matrix can be produced for each lithological unit. We performed the analysis based on the average IRC in inhabited rooms of the ground floor for each building. Original lithological units, for which less than 30 buildings were available, were excluded from the clustering procedure. The resulting similarity matrix can be used to produce clusters of similar lithological units. In this work we used a k-medoids algorithm to find the most suiting clusters (Xu and Wunsch 2008). We chose the number of clusters based on the predictability of the resulting clusters. The smallest number of clusters above which the predictability did not increase was chosen as the final cluster number. Furthermore, we created a map of the obtained clustered lithological units.

3. Ensemble regression trees

3.1. Random forests

Random forests are a special case of regression trees. In order to explain the concept of random forests we first explain regression trees. Then we clarify the concept of random forests which is an ensemble method of regression trees. For the interested reader, a more comprehensive explanation of these concepts can be found in (Cutler et al.; Hastie et al. 2009).

3.1.1. Regression trees

Regression trees are a method to estimate the regression function $g(\vec{x}) = E(y \mid \vec{x})$, where $E(\cdot)$ is the conditional expectation of a dependent variable y given a number of predictor variables $x_1, \ldots, x_p$ represented as vector $\vec{x} = (x_1, \ldots, x_p)$. The basic principle used to estimate $g(\vec{x})$ is called recursive partitioning. It consists of partitioning the space of possible values of $\vec{x}$ into smaller subspaces and to calculate a simple model for y in each of these subspaces. In our case, the model was simply the mean value of observations y_i in each of the different subspaces of $\vec{x}$. The choice of the subspaces works by recursively partitioning $\vec{x}$. The principle of recursive partitioning can be represented as a tree structure. The tree starts with a root node. A random subset of the predictor variables $\{x_1, \ldots x_p\}$ is chosen. Each variable x_j of the random subset is split into two regions, such that the mean values $\bar{y}_{region1}$ and $\bar{y}_{region2}$ of y in the new subspaces fit best to the observations y_i in these regions. The variable x_j with the best fit is chosen, to create two new nodes at the split point determined before. The same procedure is then subsequently repeated for the newly created nodes until the previously defined size of the tree is achieved.

3.1.2. The principle of random forests

As the name suggests, random forests take advantage of the ensemble of several regression trees. The principle of random forests is to grow regression trees, as described in the previous paragraph, on several numbers N of bootstrap samples of the data. Hence for each bootstrap sample, a different estimation of $g(\vec{x})$ is obtained. The prediction of y is simply obtained by averaging over the N different estimations of $g(\vec{x})$. At each bootstrap sampling step, about one third of all observation is left out of the estimation process. The left out data is called out-of-bag sample.

3.1.3. Variable importance

The out-of-bag data plays an important role in random forests, since it can be used to estimate the importance of each predictor variable. From the out-of-bag sample the prediction error can be calculated. The importance of a variable can be estimated in two steps. First, the random forests are trained and the prediction error of each tree is calculated using the out-of-bag sample as test set. The prediction errors are then averaged over all trees. In the second step, the variable of interest is randomly permuted and the random forests are computed again with the corresponding out-of-bag error estimate. The difference between the first and the second out-of-bag error estimate quantifies the importance of a variable. If the out-of-bag error estimate is higher after random permutation of the variable of interest, the variable adds information to the model. If the out-of-bag error estimate is the same after permutation, the corresponding variable does not have an importance within the model.

This measure of importance is however biased. Categorical variables with many classes are more likely to produce a good criterion value just by chance compared to categorical variables with fewer classes (Kononenko 1995; Hothorn et al. 2006b; Strobl et al. 2007). Moreover, the aggregation scheme based on bootstrap resampling of observations introduces an additional bias (Strobl et al. 2007). This problem can be overcome by using a statistical test as split criterion at each node of a tree, instead of maximizing an information measure like R^2 etc. A detailed description of this procedure is given in

(Hothorn et al. 2006b). The bias due to bootstrap resampling can be avoided just by using resampling without replacement instead of bootstrapping.

3.2. Bayesian Additive Regression Trees (BART)

The principle of BART is described in detail in (Chipman et al. 1998) and (Chipman et al. 2010). The following is a synthesis of the method described in these two works.

Like random forests, BART approximates $E(y \mid x)$ by averaging over an ensemble of trees. This can be stated as the following regression model

$$Y = \sum_{i=1}^{m} g\left(x; T_j, M_j\right) + \varepsilon \tag{0.1}$$

Where T_j describes the structure, M_j is the set of the terminal node values of the j th tree of the ensemble. ε is a random error with $\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$ and $g(\cdot)$ the functional representation of a single tree. The fundamental difference with random forests is the way in which trees are grown and how the final ensemble of trees is determined.

A single tree $\left(T_j, M_j\right)$ in BART is stochastically grown according to a predefined prior distribution. The grown tree is then adapted to the data by changing the internal tree structure.

The prior distribution of a single tree $\left(T_j, M_j\right)$ consists of 3 elements: a prior $p\left(T_j\right)$ of the tree structure, a prior $p\left(M_j\right)$ of the terminal node values and a prior $p(\sigma)$ of the residual variance.

The prior $p\left(T_j\right)$ of the tree structure is composed of a uniform distribution for the splitting rule at each internal node, a uniform distribution of the variable choice at each internal node and the

probability of a node to be nonterminal. The probability of a node to be nonterminal is controlling the size of the tree.

The prior $p\left(M_j\right)$ of the terminal node values is a normal distribution with zero mean. The variance of the normal distribution depends inversely on the number of trees in the BART ensemble. Hence, the larger the number of trees in an ensemble, the smaller the contribution of each terminal node to the sum over all trees. $p\left(M_j\right)$ has consequently a shrinkage effect on the tree ensemble, which prevents overfitting of the model.

The prior $p(\sigma)$ of the residual standard deviation σ is an inverse chi-square distribution.

Contrary to random forests, the BART model is determined not by fitting the data but rather by fitting the residuals R_j between the model and the data. In other words, each tree of the model is subsequently modified by taking into account the residuals R_j between the data and the ensemble of all other trees except the tree $\left(T_j, M_j\right)$.

$$R_j = y - \sum_{k \neq j} g\left(x; T_k, M_k\right) \tag{0.2}$$

Each tree T_j depends hence on the other trees of the ensemble via R_j. A single tree can be modified via one of the four following procedures: growing a terminal node, pruning a pair of terminal nodes, swapping the splitting rules between a parent and a child and changing the splitting rule of a non-terminal node. Whether the modification of a tree T_j improves the tree ensemble is assessed via the ratio between the posterior probabilities of the modified and the unmodified tree.

The posterior distribution of the tree parameters $\left(T_j, M_j\right)$ given the residuals R_j of the model is derived based on the Bayes' theorem by accounting for the above described priors and the likelihood

to obtain residuals R_j given a set of tree parameters $\left(T_j, M_j\right)$. For a more complete description of the posterior and the likelihood we refer to (Chipman et al. 1998) and (Chipman et al. 2010).

If the modification of $\left(T_j, M_j\right)$ augments the posterior ratio, it is accepted. If the modification does not augment the posterior ratio, it is only accepted with a certain probability. By applying this procedure subsequently to every tree of the ensemble, a Metropolis-Hastings algorithm is implemented. The periodical repetition of this principle produces a posterior sample of tree ensembles. The mean of the posterior sample finally provides the estimation of $E\left(y \mid x\right)$ and can be used to perform predictions. The standard deviation of the posterior sample gives an estimate of the uncertainty of the prediction.

3.3. Software

For data analysis and visualization we used the opensource software tools R (R Core Team 2014), QGis (QGIS Development Team 2014) and GRASS (GRASS Development Team 2012). We used the R package e1071 (Meyer et al. 2014) to implement support vector regression, randomForest (Liaw and Wiener 2002) and cforest (Hothorn et al. 2006a, 2006b; Strobl et al. 2008) for random forests, BayesTree for BART (Chipman and McCulloch 2009) and cluster (Maechler et al. 2014) to implement k-medoids clustering .

4. Results

After preprocessing, the number of IRC measurements resulted in 72 460 in 63 076 buildings. 48 of the original 69 lithological classes were represented in the IRC data. We marked lithological classes as undefined when not enough measurements were available. Figure 2a) shows the IRC distributions of the different lithological classes that were taken into account. Figure 2b) illustrates the results of the k-medoids clustering of the lithological units by means of multidimensional scaling. We calculated the

predictability of the clustering results for different numbers of clusters via cross validation. Six clusters resulted in about 6.3% of explained variance of the IRC data. More than 6 clusters did not improve the predictability. The original lithological classification could explain 6.5% of the variability of IRC. The resulting map of the clustered lithological units can be seen in Figure 3. Original lithological units, for which no more than 30 measurements were available, were excluded from the clustering procedure and are indicated as grey surfaces on Figure 3. Detailed information about the clustering results can be found in Table A1. On average, 19% of the polygons of an original lithological unit contained an IRC measurement. In case of decimal numbers we rounded up the number of measurements per 10 km^2. The original lithological units were covered with an average IRC density of sampling of about 10 buildings per 10 km^2.

Figure 4 shows the IRC map resulting from random forests modeling. The Jura Mountains in the north-west of Switzerland and the Swiss Alps exhibit a substantially higher tendency of IRC than the Swiss Plateau. Areas of lakes, glaciers or areas in which lithological units were not defined were indicated with grey color. The 5-fold cross-validation of the modeling is illustrated in Figure 5 and results in an R^2 of 33%. The variable importance measures for the biased and unbiased random forests can be obtained in Table 1. The mapping results of the BART algorithm is shown in Figure 6. Figure 7 maps the local posterior standard deviation resulting from BART. BART could explain 29% of the variability of the IRC data.

5. Discussion

The aim of this study was to develop an automatic method to classify lithological units based on their similarity in IRC distributions and to map and predict IRC by means of ensemble regression tree algorithms.

The automatic lithological classification procedure clearly points out the geological areas, Alps, Jura Mountains and Suisse Plateau (Figure 3). Cluster 1 is found only in the Swiss Alps and contains

lithological classes like igneous rock and gneiss. Both igneous rock and gneiss have been associated to elevated IRC and soil gas radon concentrations (Bossew et al. 2008; Kemski et al. 2009; Barnet et al. 2010). In the eastern part of the Canton of Graubünden a substantial area is covered by Cluster 1. This part is actually consisting of dolomite rocks. As can be seen in Table A1, the sampling density in this lithological unit is very low and only few of the polygons of this lithological unit have been measured. The result in this area must, therefore, be interpreted cautiously. The Jura Mountains are mainly covered by cluster 2. Cluster 2 and 3 can be found in the Jura Mountains as well as in the Alps. Both classes contain several classes of carbonates. As can be seen in Figure 2b, Cluster 2 is intermediate between Cluster 1 and 3. In an earlier study we already observed that carbonate rocks in the Jura Mountains bear higher IRC than carbonate rocks in the Alps (Kropat et al. 2014). Carbonate rocks, especially limestone, are subject to weathering, which is also called karstification. Karst formations are often characterized by large cave systems that facilitate the propagation of radon gas (Savoy et al. 2011). Few explanations exist however for high uranium or radium abundances in carbonate rock. One explanation for the IRC difference between carbonate rocks in the Jura Mountains and the Alps may be that carbonate rock in the Jura Mountains are closer to the crystalline basement than carbonate rock in the Alps. Consequently, the pathways for radon gas originating from the radium content in the crystalline basements are shorter in the Jura Mountains than in the Alps. The fact that Cluster 2 is between Cluster 1 and Cluster 3 fits well into this theory. The Swiss Plateau is dominated by clusters 4 and 5. As can be seen in Table A1 these groups consist mainly of quaternary deposits. The grouping of lithological units accounts well for the heterogeneity within major classes of rocks. At the edges of the Riviera and the Leventina valley, for example, we found gneisses with mica of homogenous appearance. Gneiss is a metamorphic type of rock that can have its origin in a variety of rock types—for example granite rock that can be high in uranium content (Quindós Poncela et al. 2004) as well as sedimentary rock that may have a lower uranium content. Hence, gneiss can vary substantially in its impact on IRC. Putting gneisses into a group “Metamorphic rocks” or even putting all different gneisses into a group “Gneiss” would indicate the Riviera and the Leventina valleys as regions with

moderate IRC, even if this region actually has much higher IRC. Cluster 6 consists only of marl and is found in the Suisse plateau. The number of lithological units and the number of measurements in this cluster is, however, too small to draw a final conclusion. Like any other classification methods, the approach reported here is prone to misclassification. As already mentioned, we excluded lithological units with less than 30 buildings. Nevertheless, for future work maybe this criterion should be shifted to 100 buildings, to obtain more robust results. Since the approach is based on IRC measurements, inhabited zones are better represented by this method.

The mapping of IRC based on regression trees represents well the regional difference of IRC which were observed in earlier Swiss IRC mapping attempts (FOPH 2013; Kropat et al. 2015). Furthermore, the map reveals spatial detail beyond municipal limits. In regions like the Canton of Thurgau few measurements have been carried out. This leads to linear artifacts in south-north direction, which is due to the binary nature of the regression trees underlying random forests and BART. Without the ensembling of several trees we expect this effect to be even stronger. The Alps are also a region with a low density of IRC sampling. Here the linear artifacts are less pronounced. This may be due to the high spatial variation of altitude in this area. Since IRC are related to altitude, the spatial variation of IRC mapping is consequently higher than in flatter regions. Overall, the linear artifacts are not very prominent on the map, which is due to the good sampling coverage over a large part of Switzerland.

We found the explained variance of random forests with 33% to be superior to that of BART with 29%. Also the computation time was considerably faster. Compared to earlier Swiss IRC prediction models, random forests considerably improve the explained variance ((Hauri et al. 2012): 20%; (Kropat et al. 2015): 28%). In addition, the great advantage of BART is that it provides a direct uncertainty measure of the prediction. Figure 7 clearly points out areas of high altitude in the Alps as being areas of high prediction uncertainty. This can be explained because in this region the sampling density is very low.

The variable importance measure of the biased random forests tends to put more weight on continuous variables and categorical variables with many classes (Table 1). This effect seems to be less pronounced in the unbiased method. The variables with the strongest importance are clearly the variables related to the location of the buildings like lithology, the coordinates and the altitude. After location related variables, the detector type was found to strongly influence the IRC concentration. This is a finding which is in accordance with an earlier finding from a univariate study (Kropat et al. 2014) where the authors hypothesized that contrary to other detectors, electret detector can be biased by the presence of dust or humidity. We found that variables related to architectural characteristics have less influence on the IRC. Finally, the two variables with the lowest importance are building type and outdoor temperature. However, a substantial influence of outdoor temperature was observed in an earlier univariate IRC analyses (Kropat et al. 2014). Possibly the formerly observed outdoor temperature effect was due to the spatial inhomogeneity of IRC sampling. The difference between location-based and construction-related variables can be explained by the assumption that construction related variables are much less precise than location based variables. The year of construction, the foundation and building type only give limited information about the properties of a building.

6. Conclusions

This study contributes to a better understanding of the multi factorial determinants of public radon hazards as well as the radon properties of geological units. Furthermore, here we propose data driven modeling techniques which have the potential to substantially improve the creation of national radon risk maps.

We found ensemble regression trees to be a powerful tool for assessing the importance of a predictor variable in a multidimensional setting. We observed that building-related variables have a less important influence on IRC than location/ geology related variables.

Based on state-of-the-art clustering, we found a method which enables the creation of a coherent definition of geological classes in terms of their radon characteristics. Combining this method with ensemble regression trees leads to models which considerably improve the predictability compared to former studies carried out on Swiss IRC.

National radon risk communication is subject to two main issues; what are the effects of radon on human health and how can radon hazards be localized? This study addresses the localization of radon hazards and proposes modeling approaches which can in the future be combined with the health effects of radon in order to perform appropriate national radon risk communication. Future improvements of these approaches could be obtained by including more detailed geological information as well as more explanatory variables like pressure differences and the uranium content of the ground.

References

Andersen CE, Raaschou-Nielsen O, Andersen HP, Lind M, Gravesen P, Thomsen BL, et al. Prediction of 222Rn in Danish dwellings using geology and house construction information from central databases. Radiat Prot Dosimetry. 2007 Jan 1;123(1):83–94.

Appleton JD, Miles JCH. A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk. J Environ Radioact. 2010 Oct;101(10):799–803.

Appleton JD, Miles JCH, Young M. Comparison of Northern Ireland radon maps based on indoor radon measurements and geology with maps derived by predictive modelling of airborne radiometric and ground permeability data. Sci Total Environ. 2011 Mar 15;409(8):1572–83.

Barnet I, Pacherová P, Preusse W, Stec B. Cross-border radon index map 1:100 000 Lausitz – Jizera – Karkonosze – Region (northern part of the Bohemian Massif). J Environ Radioact. 2010 Oct;101(10):809–12.

Borgoni R, Tritto V, Bigliotto C, De Bartolo D. A Geostatistical Approach to Assess the Spatial Association between Indoor Radon Concentration, Geological Features and Building Characteristics: The Case of Lombardy, Northern Italy. Int J Environ Res Public Health. 2011;8(5):1420–40.

Bossew P, Dubois G, Tollefsen T. Investigations on indoor Radon in Austria, part 2: Geological classes as categorical external drift for spatial modelling of the Radon potential. J Environ Radioact. 2008 Jan;99(1):81–97.

Bossew P, Lettner H. Investigations on indoor radon in Austria, Part 1: Seasonality of indoor radon concentration. J Environ Radioact. 2007;98(3):329–45.

Bossew P, Žunić ZS, Stojanovska Z, Tollefsen T, Carpentieri C, Veselinović N, et al. Geographical distribution of the annual mean radon concentrations in primary schools of Southern Serbia – application of geostatistical methods. J Environ Radioact. 2014 Jan;127:141–8.

Buchli R, Burkart W. Influence of subsoil geology and construction technique on indoor air 222Rn levels in 80 houses of the central Swiss Alps. Health Phys. 1989 Apr;56(4):423–9.

Burkart W, Wernli C, Brunner HH. Matched Pair Analysis of the Influence of Weather-Stripping on Indoor Radon Concentration in Swiss Dwellings. Radiat Prot Dosimetry. 1984 Jan 1;7(1-4):299–302.

Cherkassky V, Mulier FM. Support Vector Machines. Learn Data Concepts Theory Methods. John Wiley & Sons; 2007. p. 404–66.

Chipman HA, George EI, McCulloch RE. Bayesian CART Model Search. J Am Stat Assoc. 1998 Sep;93(443):935–48.

Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2010 Mar;4(1):266–98.

Chipman HA, McCulloch RE. Baysian Methods for Tree Based Models [Internet]. 2009. Available from: http://cran.r-project.org/src/contrib/Archive/BayesTree/

Cinelli G, Tondeur F, Dehandschutter B. Statistical analysis of indoor radon data for the Walloon region (Belgium). Radiat Eff Defects Solids. 2009;164(5-6):307–12.

Cinelli G, Tondeur F, Dehandschutter B. Development of an indoor radon risk map of the Walloon region of Belgium, integrating geological information. Environ Earth Sci. 2011 Feb 1;62(4):809–19.

Cutler A, Cutler DR, Stevens JR. Random Forest. Ensemble Mach Learn - Methods Appl [Internet]. [cited 2014 Sep 19]. p. 157 – 175. Available from: http://www.springer.com/engineering/computational+intelligence+and+complexity/book/978-1-4419-9325-0

Demoury C, Ielsch G, Hemon D, Laurent O, Laurier D, Clavel J, et al. A statistical evaluation of the influence of housing characteristics and geogenic radon potential on indoor radon concentrations in France. J Environ Radioact. 2013 Dec;126:216–25.

Denman AR, Crockett RGM, Groves-Kirkby CJ, Phillips PS, Gillmore GK, Woolridge AC. The value of Seasonal Correction Factors in assessing the health risk from domestic radon—A case study in Northamptonshire, UK. Environ Int. 2007 Jan;33(1):34–44.

Dubois G, Bossew P, Friedmann H. A geostatistical autopsy of the Austrian indoor radon survey (1992–2002). Sci Total Environ. 2007 May 15;377(2–3):378–95.

FOPH. Radon risk in Switzerland [Internet]. Federal Office of Public Health FOPH; 2013. Available from: http://www.bag.admin.ch/themen/strahlung/00046/11952/index.html?lang=en

Frei K. Zehn Jahre Minergie. NOVA - Z Für Wärmedämmung Solarenergie [Internet]. 2013;(26). Available from: http://www.flumroc.ch/de/aktuell/zeitschrift_nova.php#

Friedmann H, Bossew P. Selected statistical problems in spatial evaluation of Rn related variables. Nukleonika. 2010;Vol. 55, No. 4:429–32.

Friedmann H, Gröller J. An approach to improve the Austrian Radon Potential Map by Bayesian statistics. J Environ Radioact. 2010 Oct;101(10):804–8.

Friedmann H, Zimprich P, Atzmüller C, Hofmann W, Lettner H, Steinhäusler F, et al. The Austrian Radon Project. Environ Int. 1996;22, Supplement 1:677–86.

FSO. Eidgenössisches Gebäude- und Wohnungsregister. Swiss Federal Statistical Office; 2014.

Girault F, Perrier F. Estimating the importance of factors influencing the radon-222 flux from building walls. Sci Total Environ. 2012 Sep 1;433:247–63.

GRASS Development Team. Geographic Resources Analysis Support System (GRASS GIS) Software [Internet]. Open Source Geospatial Foundation; 2012. Available from: http://grass.osgeo.org

Groves-Kirkby CJ, Denman AR, Phillips PS, Crockett RGM, Sinclair JM. Comparison of seasonal variability in European domestic radon measurements. Nat Hazards Earth Syst Sci. 2010 Mar 26;10(3):565–9.

Groves-Kirkby CJ, Denman AR, Phillips PS, Tornberg R, Woolridge AC, Crockett RGM. Domestic radon remediation of U.K. dwellings by sub-slab depressurisation: Evidence for a baseline contribution from constructional materials. Environ Int. 2008 Apr;34(3):428–36.

Groves-Kirkby CJ, Timson K, Shield G, Denman AR, Rogers S, Phillips PS. Lung-cancer reduction from smoking cessation and radon remediation: A preliminary cost-analysis in Northamptonshire, UK. Environ Int. 2011 Feb;37(2):375–82.

Gunby JA, Darby SC, Miles JCH, Green BMR, Cox DR. Factors affecting indoor radon concentrations in the United Kingdom. Health Phys. 1993;64(1):2–12.

Gundersen LCS, Schumann RR. Mapping the radon potential of the United States: Examples from the Appalachians. Environ Int. 1996;22, Supplement 1:829–37.

Györfi L, Vajda I, Meulen E van der. Minimum kolmogorov distance estimates of parameters and parametrized distributions. Metrika. 1996 Dec 1;43(1):237–55.

Hastie T, Tibshirani R, Friedman J. Random Forests. Elem Stat Learn [Internet]. Springer New York; 2009 [cited 2014 Sep 19]. p. 587–604. Available from: http://link.springer.com/chapter/10.1007/978-0-387-84858-7_15

Hauri DD, Huss A, Zimmermann F, Kuehni CE, Röösli M. A prediction model for assessing residential radon concentration in Switzerland. J Environ Radioact. 2012 Oct;112:83–9.

Hothorn T, Buehlmann P, Dudoit S, Molinaro A, Van Der Laan M. Survival Ensembles. Biostatistics. 2006 a;7(3):355–73.

Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. J Comput Graph Stat. 2006 b;15(3):651–74.

Jelle BP. Development of a model for radon concentration in indoor air. Sci Total Environ. 2012 Feb 1;416:343–50.

Kemski J, Klingel R, Siehl A, Valdivia-Manchego M. From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. Environ Geol. 2009 Feb 1;56(7):1269–79.

Kemski J, Siehl A, Stegemann R, Valdivia-Manchego M. Mapping the geogenic radon potential in Germany. Sci Total Environ. 2001 May 14;272(1–3):217–30.

Kononenko I. On Biases in Estimating Multi-Valued Attributes. Morgan Kaufmann; 1995. p. 1034–40.

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios (Gruson) M, et al. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland. Sci Total Environ. 2015 Feb 1;505:137–48.

Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios M, et al. Major influencing factors of indoor radon concentrations in Switzerland. J Environ Radioact. 2014 Mar;129:7–22.

Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. 2014.

Mäkeläinen I, Arvela H, Voutilainen A. Correlations between radon concentration and indoor gamma dose rate, soil permeability and dwelling substructure and ventilation. Sci Total Environ. 2001 May 14;272(1–3):283–9.

Manic G, Petrovic S, Vesna M, Popovic D, Todorovic D. Radon concentrations in a spa in Serbia. Environ Int. 2006 May;32(4):533–7.

Menzler S, Piller G, Gruson M, Rosario AS, Wichmann H-E, Kreienbrock L. POPULATION ATTRIBUTABLE FRACTION FOR LUNG CANCER DUE TO RESIDENTIAL RADON IN SWITZERLAND AND GERMANY: Health Phys. 2008 Aug;95(2):179–89.

MeteoSwiss. MeteoSwiss. Federal Office of Meteorology and Climatology; 2013.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien [Internet]. 2014. Available from: http://CRAN.R-project.org/package=e1071

Miles J. Temporal Variation of Radon Levels in Houses and Implications for Radon Measurement Strategies. Radiat Prot Dosimetry. 2001 Jan 2;93(4):369–75.

Miles JCH, Appleton JD. Mapping variation in radon potential both between and within geological units. J Radiol Prot. 2005 Sep 1;25(3):257.

Neznal M. Permeability as an important parameter for radon risk classification of foundation soils. Ann Geophys [Internet]. 2005 [cited 2013 Dec 10]; Available from: http://www.earth-prints.org/handle/2122/894

QGIS Development Team. QGIS Geographic Information System [Internet]. Open Source Geospatial Foundation; 2014. Available from: http://qgis.osgeo.org

Quindós Poncela LS, Fernández PL, Gómez Arozamena J, Sainz C, Fernández JA, Suarez Mahou E, et al. Natural gamma radiation map (MARNA) and indoor radon levels in Spain. Environ Int. 2004 Feb;29(8):1091–6.

R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available from: http://www.R-project.org/

Savoy L, Surbeck H, Hunkeler D. Radon and CO2 as natural tracers to investigate the recharge dynamics of karst aquifers. J Hydrol. 2011 Sep 6;406(3–4):148–57.

SGTK. Lithologisch-petrografische Karte der Schweiz-Lithologie-Hauptgruppen 1:500 000. Schweizerische Geotechnische Kommission. 2000.

Singh S, Kumar A, Singh B. Radon level in dwellings and its correlation with uranium and radium content in some areas of Himachal Pradesh, India. Environ Int. 2002 Apr;28(1–2):97–101.

Smethurst MA, Strand T, Sundal AV, Rudjord AL. Large-scale radon hazard evaluation in the Oslofjord region of Norway utilizing indoor radon concentrations, airborne gamma ray spectrometry and geological mapping. Sci Total Environ. 2008 Dec 15;407(1):379–93.

Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004 Aug 1;14(3):199–222.

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008 Jul 11;9(1):307.

Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007 Jan 25;8(1):25.

swisstopo. DHM25 The digital height model of Switzerland [Internet]. Federal Office of Topography; 2004. Available from: http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/height/dhm25.html

Tondeur F, Cinelli G, Dehandschutter B. Homogenity of geological units with respect to the radon risk in the Walloon region of Belgium. J Environ Radioact. 2014 Oct;136:140–51.

Trevisi R, Caricato A, D'Alessandro M, Fernández M, Leonardi F, Luches A, et al. A pilot study on natural radioactivity in schools of south-east Italy. Environ Int. 2010 Apr;36(3):276–80.

Xu R, Wunsch D. Partitional Clustering. Clustering. John Wiley & Sons; 2008. p. 63 – 110.

Zeeb H, Shannoun F. Radon measurements. WHO Handb Indoor Radon Public Heal Perspect [Internet]. Geneva: World Health Organization (WHO); 2009a [cited 2013 Dec 10]. p. 21–40. Available from: http://www.who.int/iris/handle/10665/44149

585 Zeeb H, Shannoun F. Radon prevention and mitigation. WHO Handb Indoor Radon Public Heal Perspect [Internet]. Geneva: World Health Organization (WHO); 2009b [cited 2013 Dec 10]. p. 41 – 56. Available from: http://www.who.int/iris/handle/10665/44149

Zhu H., Charlet J., Poffijn A. Radon risk mapping in southern Belgium: an application of geostatistical and GIS techniques. Sci Total Environ. 2001 May 14;272(1–3):203–10.

590

7. Tables

Table 1 Variable importance obtained by biased and unbiased random forests

Variable	Number of levels	Random Forests	CForest
Building type	5	3274	0.09
Foundation type	4	2584	0.1
Year of construction	4	3009	0.14
Detector type	13	4234	0.18
Coordinate X	-	13758	0.2
Coordinate Y	-	12462	0.18
Temperature	-	10312	0.06
Altitude	-	12903	0.2
Clustered lithological units	6	4064	0.21

595

8. Figure captions

Figure 1 Map of main geological regions in Switzerland and locations of cantons and valleys which are mentioned in this study. The black lines indicate the cantonal and national boundaries.

Figure 2 a) Boxplot of IRC distributions within different lithological classes b) Multidimensional scaling representation of Kolmogorov distances between IRC distributions of lithological classes. The different groups resulting from k-medoid clustering are indicated by different colors.

Figure 3 Map of clustered lithological units. Areas of lakes, glaciers or of lithological units for which not enough IRC measurements were available are indicated in grey.

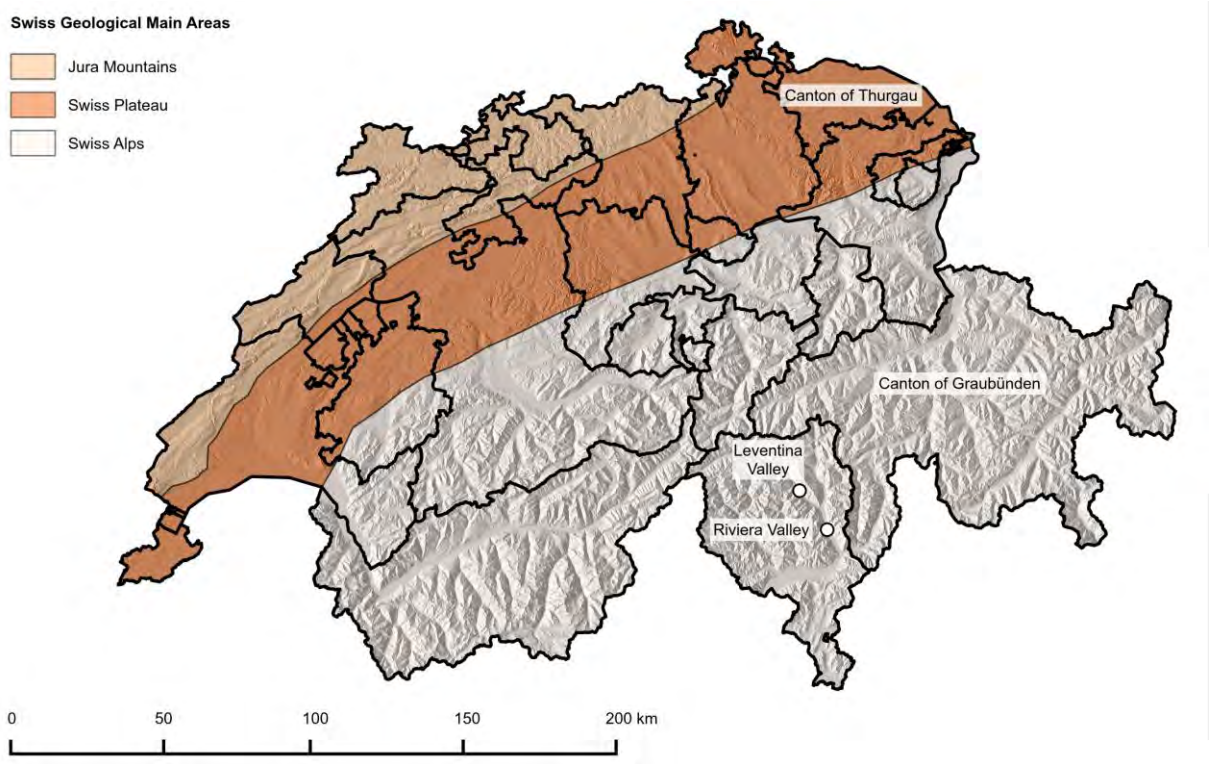
Figure 4 Mapping of IRC by means of random forests.

Figure 5 Comparison of observed and predicted IRC obtained by 5-fold cross validation of IRC modeling via random forests

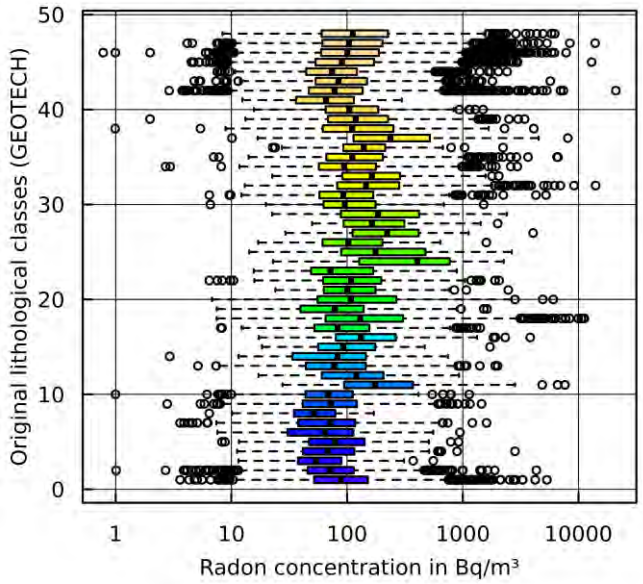
Figure 6 IRC mapping with BART. The estimation corresponds to the posterior sample mean for each pixel of the map.

Figure 7 Mapping of uncertainty estimate obtained from BART. The uncertainty estimate corresponds to the posterior sample standard deviation of each pixel.

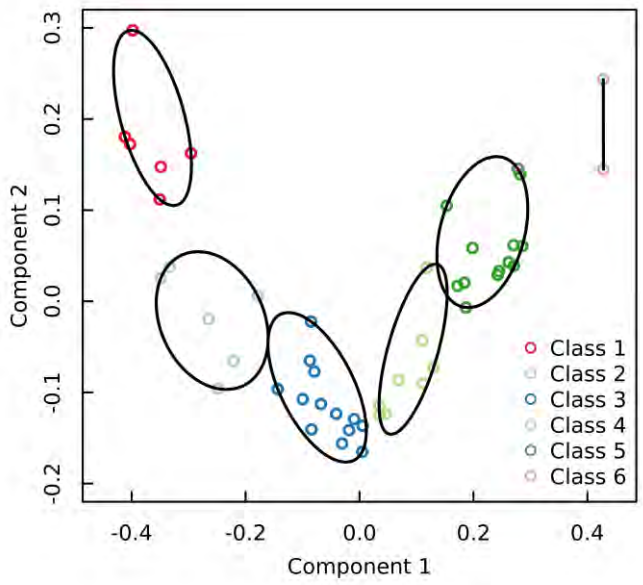
Figure 1


Swiss Geological Main Areas
Jura Mountains
Swiss Plateau
Swiss Alps
Canton of Thurgau
Canton of Graubünden
Leventina Valley
Riviera Valley
0
50
100
150
200 km

615 Figure 2



a)



b)

620

Figure 3

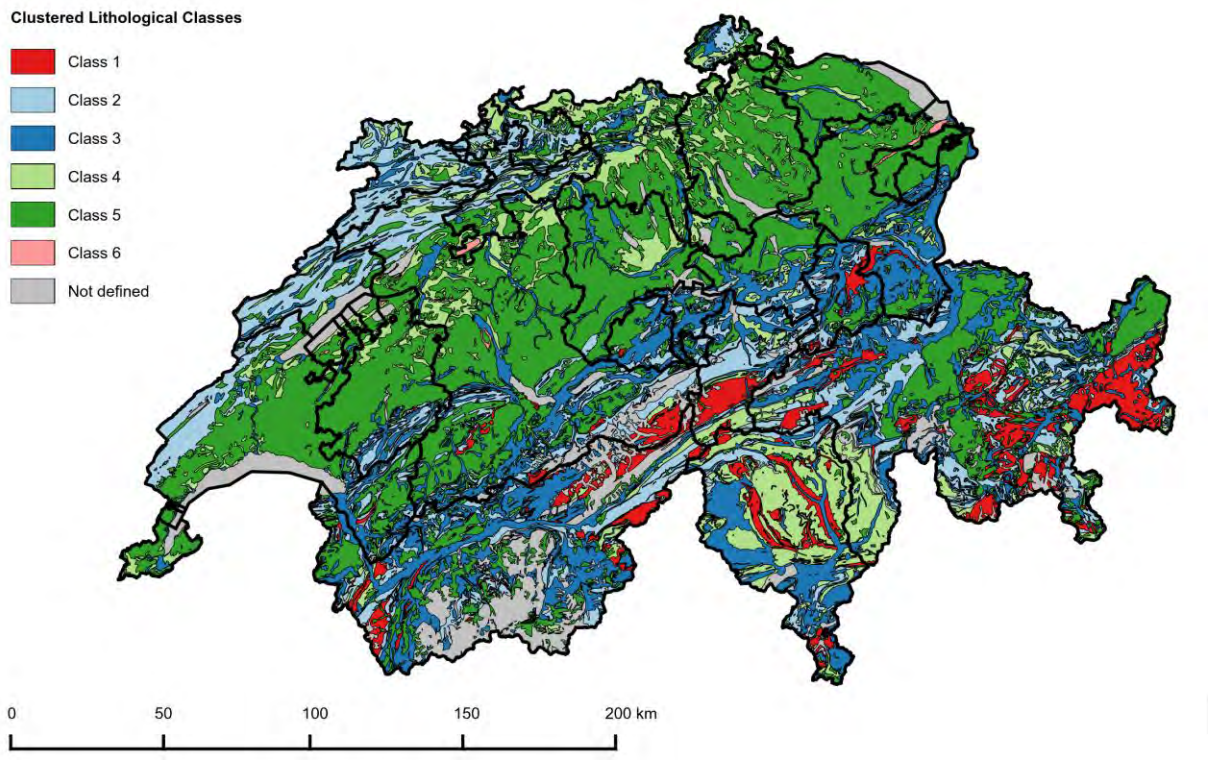
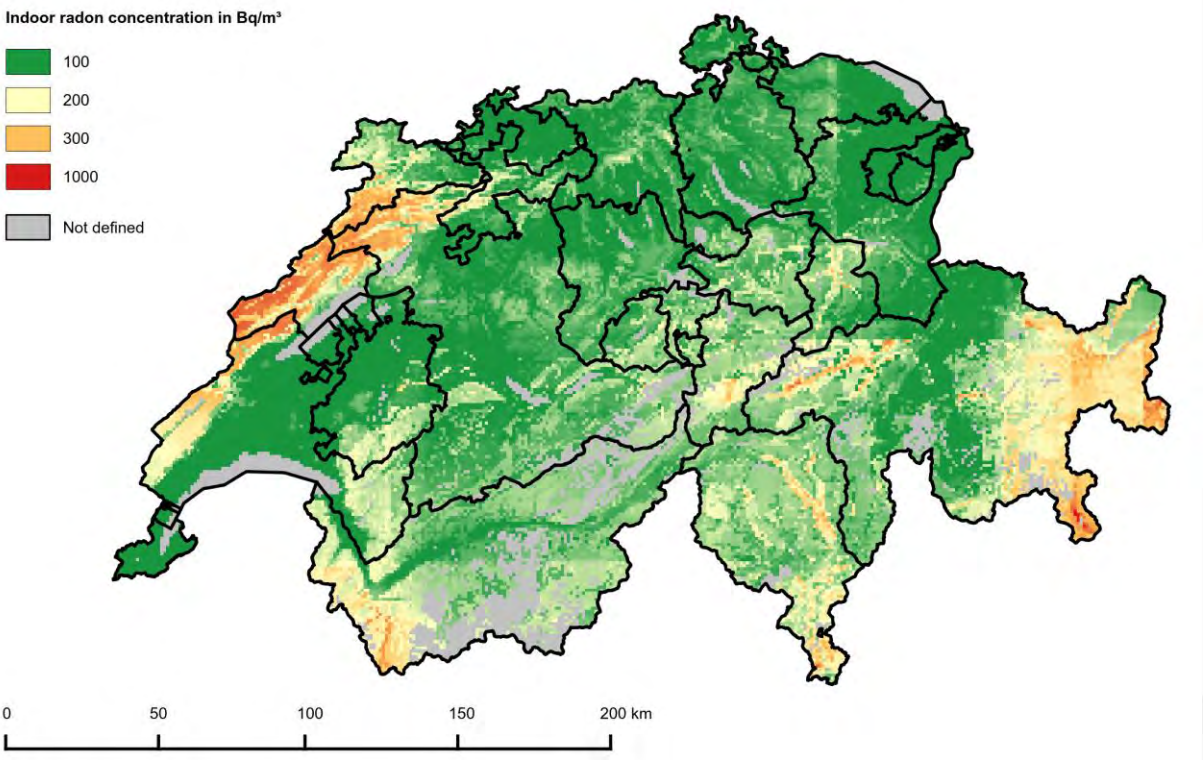
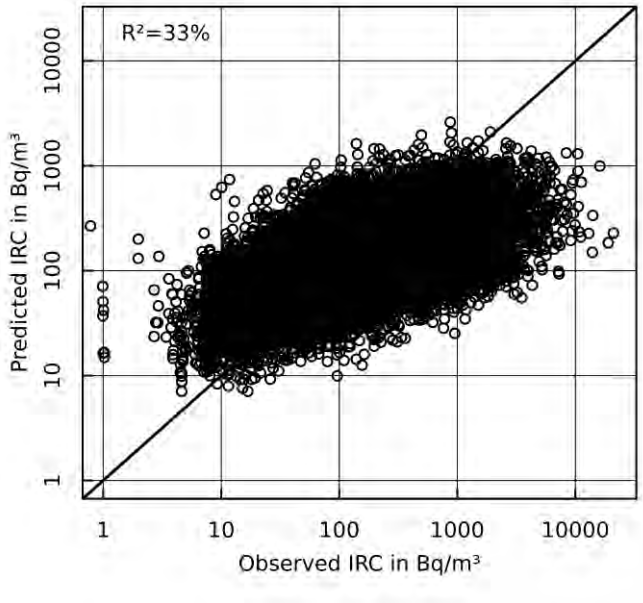

Clustered Lithological Classes
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Not defined
0
50
100
150
200 km

Figure 4


Indoor radon concentration in Bq/m³
100
200
300
1000
Not defined
0
50
100
150
200 km

625

Figure 5


R²=33%
Predicted IRC in Bq/m³
Observed IRC in Bq/m³
1
10
100
1000
10000

630 Figure 6

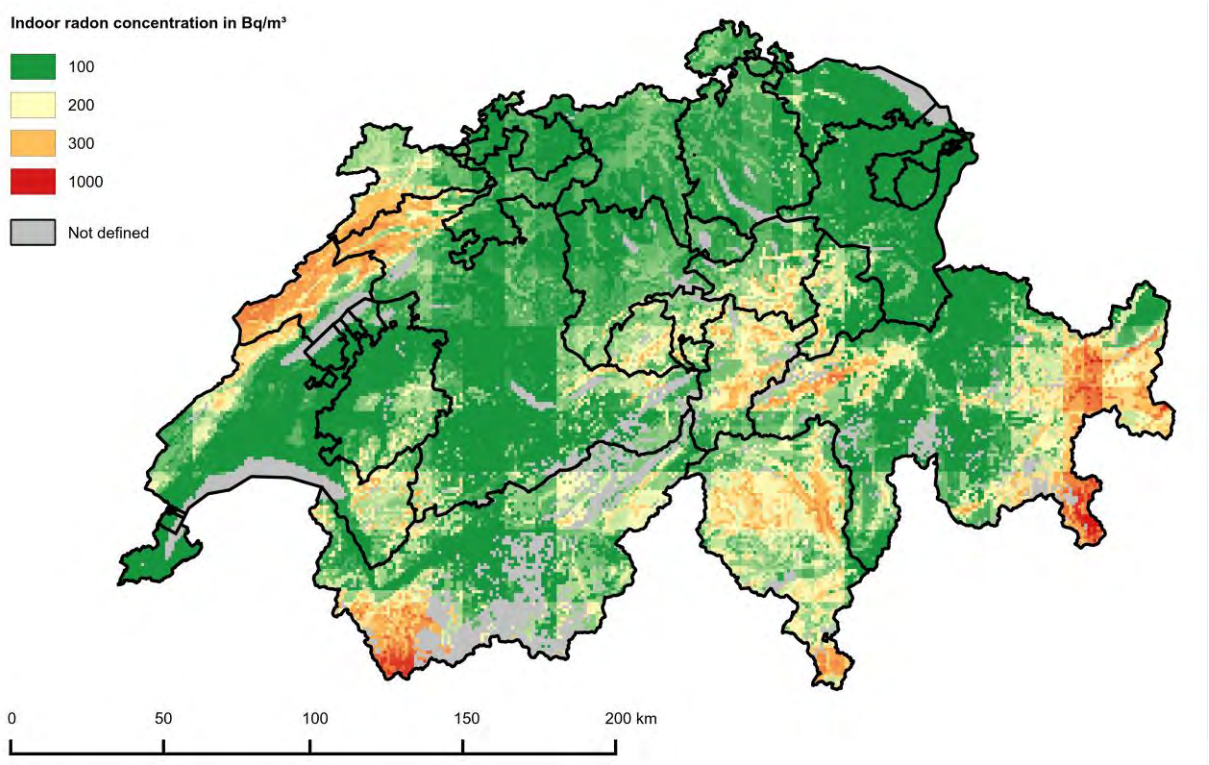
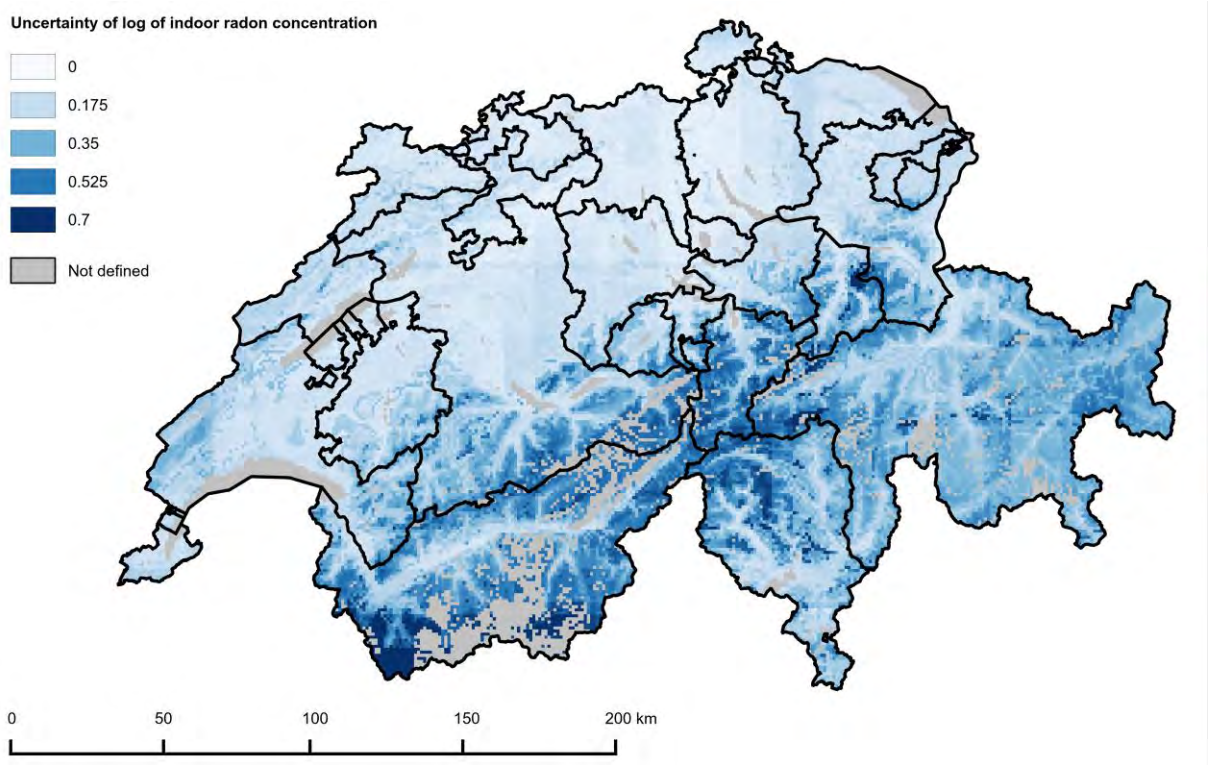

Indoor radon concentration in Bq/m³
100
200
300
1000
Not defined
0
50
100
150
200 km

Figure 7


Uncertainty of log of indoor radon concentration
0
0.175
0.35
0.525
0.7
Not defined
0
50
100
150
200 km

635

9. Annex

Table A1 Reclassification of lithological units

GEOTECH	LITH_PET	LITH_PET_GENERAL	LITH_CLUSTER	N	Fraction of measured polygons in %	Number of measurements per 10 km^2
50	Granites with transitions in quartz diorites and quartz syenites	Igneous Rock	1	125	3	1
54	Porphyrites and porphyrite-tuff	Igneous Rock	1	236	12	22
22	Conglomerates to breccias with arkoses and sandstones	Sedimentary Rock	1	162	4	2
47	Dolomite rocks partly with lime-layers	Carbonates	1	187	3	1
46	Lime-breccias or lime-conglomerates	Carbonates	1	39	25	16
65	Two-mica- to biotite-gneisses, feldspar-rich, predominantly in homogeneous formation	Metamorphic Rock	1	252	11	6
40	Limestones, often with marly intermediate layers	Carbonates	2	5838	11	5
61	Gneisses with abundant feldspar, developed under sericite, epidote and chlorite formation	Metamorphic Rock	2	98	7	1
60	Gneisses with abundant feldspar	Metamorphic Rock	2	1536	6	8
52	Quartz-porphyries	Igneous Rock	2	39	8	3
35	Lime-phyllites to lime-mica-slates	Metamorphic Rock	2	187	5	2
64	Sericite-rich conglomerates and breccias	Sedimentary Rock	2	135	21	13
130	Pebbles and sands, partly with clayey or silty layers	Sediment	3	10566	34	26
42	Limestones with important layers of marl	Carbonates	3	1189	9	3
132	Scree and talus slope	Sediment	3	1051	12	7
131	Sands, pebbles, stones and boulders	Sediment	3	3633	38	54
43	Helvetic siliceous limestones	Carbonates	3	124	9	2
30	Clay-slates to phyllites with enclaves of sandstones and breccias to conglomerates	Metamorphic Rock	3	72	10	2
44	Sand-limes to pebble-limes with layers of marl slates	Sedimentary Rock	3	382	12	3
48	Dolomites and cellular limes	Carbonates	3	52	8	3
80	Amphibolites with transitions in diorites and in hornblende bearing gneisses	Metamorphic Rock	3	258	5	4
63	Sericite-chlorite-gneisses to -slates	Metamorphic Rock	3	1292	12	7
56	Limestones to lime-marbles	Carbonates	3	219	4	10
68	Biotite- to muscovite-rich gneisses, partly chlorite bearing, partly with lime-silicate rocks or quartzite layers (hornfels)	Metamorphic Rock	3	1327	5	24
66	Two-mica- to biotite-gneisses, feldspar-rich, laminated	Metamorphic Rock	3	113	16	2
120	Pebbles and sands	Sediment	4	8005	37	22
10	Marls with weakly consolidated sandstone-, conglomerate-, or cobble-layers	Sedimentary Rock	4	3350	30	16
38	Limestones with dolomite layers	Carbonates	4	327	25	6
111	Silts to silty sands, often clayey, mostly lime-bearing	Sediment	4	755	24	7
62	Biotite- to muscovite-rich gneisses and mica-slates	Metamorphic Rock	4	2014	8	4
45	Clauconite-bearing quartz-sandstone with echinoderm clasts	Sedimentary Rock	4	35	9	3
33	Marl-slate to lime-phyllite with layers of volcanic tuff-sandstones	Metamorphic Rock	4	106	8	11
59	Two-mica- to biotite-gneisses feldspar-rich of changeful structure	Metamorphic Rock	4	530	25	6
110	Silts to silty sands, with pebbles, sands and boulder	Sediment	5	10638	23	9
11	Predominantly lime-bearing, porous sandstones with marl-layers	Sedimentary Rock	5	2762	44	14
112	Clayey silts to clays with sand-layers	Sediment	5	1260	40	13
21	Sandstone- and marl-layers with layers of weakly to moderately consolidated conglomerates	Sedimentary Rock	5	820	43	8
13	Marl and slate-clays with lime-, dolomite-, sandstone-banks	Sediment	5	367	21	4
16	Sandstone with marl-layers	Sedimentary Rock	5	510	23	8
15	Marl with layers of stronger consolidated sandstones	Sedimentary Rock	5	220	16	4
20	Conglomerates with sandstone- and marl-layers	Sedimentary Rock	5	942	38	7
14	Ferriferous clays	Sedimentary Rock	5	126	23	46
32	Firm, compact sandstones with layers of marl-slate and lime-phyllites	Sedimentary Rock	5	84	11	1
81	Green slate with transitions in basic igneous rock, eclogite	Metamorphic Rock	5	39	3	1
31	Marl-slate to lime-phyllite with enclaves of sandstones	Metamorphic Rock	5	542	10	3
41	Limestone, often marlish	Carbonates	5	328	13	5
17	Marls with layers of firm, sandy coquina	Sedimentary Rock	6	52	100	15
12	Marls with layers of shell rich sandstones with shell-breccia	Sedimentary Rock	6	152	50	33