Journal of Environmental Radioactivity 147 (2015) 51–62

Contents lists available at ScienceDirect

Journal of Environmental Radioactivity

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

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*

A R T I C L E I N F O

Article history:
Received 20 February 2015
Received in revised form
30 April 2015
Accepted 7 May 2015
Available online 28 May 2015

Keywords:
Indoor radon
Mapping
Predictive modeling
k-medoids clustering
Random forest
Bayesian additive regression trees

A B S T R A C T

Purpose: According to estimations around 230 people die as a result of radon exposure in Switzerland. 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: The automated classification groups lithological units well in terms of their IRC characteristics. Especially the IRC differences in metamorphic rocks like gneiss are well revealed by this method. The maps produced by random forests soundly represent the regional difference of IRCs 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 influence of a variable evaluated by random forests shows that building characteristics are less important predictors for IRCs 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 IRCs. Automatic clustering of lithological units complements 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 soil gas radon measurements 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, 2009a), it is strongly dependent on the underlying geology (Appleton and Miles, 2010; Bossew et al., 2014; Cinelli et al., 2009; Dubois et al., 2007; Friedmann and Gröller, 2010). Thus, indoor radon concentrations (IRCs) 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

* Corresponding author. 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@gmail.com (G. Kropat).

http://dx.doi.org/10.1016/j.jenvrad.2015.05.006
0265-931X/

achieved through the appropriate construction of new buildings and the mitigation of already existing buildings (Groves-Kirkby et al., 2008). The prediction of IRCs and identifying zones at risk is however difficult and still subject to scientific debate (Friedmann and Bossew, 2010).

IRCs are known to be subject to several sources of variance. Building architecture is an important factor. In several studies the influence of the building substructure on IRCs was observed to be significant (Burkart et al., 1984; Kropat et al., 2014; Mäkeläinen et al., 2001). Construction materials are known to play an important role in the occurrence of elevated IRCs in buildings (Demoury et al., 2013; Girault and Perrier, 2012; Gunby et al., 1993). Furthermore, the ventilation and heating habits of people are an important factor of IRCs (Gunby et al., 1993). This leads to visible seasonal effects (Bossew and Lettner, 2007; Denman et al., 2007; Groves-Kirkby et al., 2010; Singh et al., 2002; Trevisi et al., 2010). In winter time, buildings are generally less ventilated than in summer. In addition to that, indoor/outdoor pressure differences due to heating in winter can lead to stack effects. IRC dependencies on weather conditions also have been observed (Miles, 2001). Finally, one of the most discussed determinants of IRCs is the geology subjacent to the concerned buildings. IRCs are known to be strongly dependent on geological parameters like uranium content, permeability of the ground as well as soil properties (Appleton et al., 2011; Bossew et al., 2008; Buchli and Burkart, 1989; Gundersen and Schumann, 1996; Neznal, 2005; Singh et al., 2002). Many studies have suggested generalizing the geological units in order to simplify the interpretation and to improve the statistical power of IRC models (Bossew et al., 2008; Kemski et al., 2001; Kropat et al., 2015, 2014; Miles and Appleton, 2005; Smethurst et al., 2008; Tondeur et al., 2014). 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 superficial deposits and bedrock e.g. obtained via airborne gamma ray spectrometry (Ielsch et al., 2010; Smethurst et al., 2008). However, few data-driven approaches exist to classify geologies in terms of their radon characteristics. Just recently (Bossew, 2015) developed a two class approach for the determination of radon prone and non-radon prone geologies.

The aim of this study was twofold: we developed a data driven method to group lithological units according to their IRC characteristics. 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. We could not account for the radon exposure during transport. This has to be regarded as an increase of the uncertainty of the measurement. However since the time of transport is generally substantially shorter than the time of overall exposure we consider the uncertainty contribution due to transport as small. 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 and in living rooms. About 30% of IRC measurements from the raw database corresponded to this criterion. Like many other studies, we carried out analysis, mapping and validation on log-transformed IRCs in order to avoid the influence of extreme values and to stay comparable to other approaches (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. Detector types

The IRC measurements used in this study were mainly performed with alpha track and electret detectors (Zeeb and Shannoun, 2009b). 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 that differed by vendor and detector type. The most prevalent detector types in the raw data were Gammadata, E-Perm, Radtrak and Miam.

2.1.3. Coordinates

The building coordinates were either owner-declared or indicated by the Swiss Federal Statistical Office (FSO). As a quality control, we checked that all building coordinates were not in regions that 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.

In order to make the geographical orientation in Switzerland easier for the reader, Fig. 1 indicates all locations, cantons and geological regions that 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, including bedrock and drift geology alike 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 individuals that carried out the measurement.