Journal of Environmental Radioactivity 141 (2015) 44–50

Contents lists available at ScienceDirect

Journal of Environmental Radioactivity

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

Estimation of radon prone areas through binary classification, part 2: radon prone geologies

P. Bossew

German Federal Office for Radiation Protection, Köpenicker Allee 120-130, 10318 Berlin, Germany

ARTICLE INFO

Article history:
Received 7 August 2014
Received in revised form
24 November 2014
Accepted 30 November 2014
Available online

Keywords:
Radon prone geology
Binary classification
Radon mapping

ABSTRACT

A radon prone geology is one for which the probability is increased that in a house built on it, elevated indoor Rn concentration will be encountered, or that its Rn potential will be increased. Labelling geological units as Rn prone or not can be an important support in deciding whether a geographical or administrative region in which that geological unit occurs, should be called Rn prone area, possibly in absence of other predictors. In this article a method is proposed which, given a set of geological classes, sorts the classes into Rn prone and non-Rn prone classes depending on a classification criterion which one can choose according the purpose. The method is computationally simple and is demonstrated on the example of Germany.

1. Introduction

It has been known for a long time that houses built over certain geologies, notably granite, black shale or permeable rock or sediments are prone to elevated – sometimes indeed very high – indoor Rn concentrations. Other geological grounds, typically most limestone (except karst), sandy alluvial sediments or sandstones not bearing organic matter are usually low in Rn. Much literature (not to be reviewed here because of its sheer amount) deals with the relation between geology and Rn related quantities. The reason for this interest is that geological information about an area is easily available in most instances, and it can serve as a "cheap" predictor of the presence of a Rn hazard. (Here the term geology is used in a wider sense, including base rock and overburden, and possibly soil properties or even tectonic and geo-hydrological features. In the context of the example discussed here, however, the available geological information has been derived from a base rock map.)

Suppose the case that a region has scarcely or not all been sampled for Rn related quantities. Such quantities are usually indoor Rn concentration or the geogenic Rn potential (RP). If measurements of these quantities are not available, one may look for other data, which allow predicting, if only roughly, whether the region may be Rn prone.

Since one main physical control of indoor Rn is geology (in the wider sense), it is obvious that one may attempt classifying geological units for whether they cause a Rn hazard.[1] Here a method is presented to do this, depending on a given criterion. Such criterion can be exceedance of a mean indoor Rn concentration above a threshold, or exceedance of a certain probability threshold that a concentration threshold is exceeded, etc. Calibration of the model is performed in regions with sufficient data, but if the same geology occurs in a region without (sufficient) data, one may use geology as predictor whether this region is Rn prone or not, until better data are available. Qualifying a region through geology may actually motivate surveys of indoor Rn or RP.

In a first attempt in this sense (Gruber et al., 2012), a geogenic Rn map of parts of Europe was prepared based on geological units. They were sorted into four levels according to their mean RP. The problem with the approach used was the large variability of the RP within geological units, which makes grouping difficult and sometimes ambiguous.

The method presented in this article attempts the same thing, but appears to be more robust and the result seems to be more clearly related to the criteria, which calibrate or "gauge" the grouping. It is based on ROC (receiver operating characteristic) analysis, which has been proposed for Rn classification in a

[1] Anthropogenic control factors that are at least as important as geology are construction type, floor level and ventilation. However, regional trend is mainly caused by varying geology, while other factors contribute a "noise" component, since for example construction types are regionally stronger randomly mixed than geology. Therefore they do not contribute to the trend as clearly (if at all) as geology does.

E-mail address: pbossew@bfs.de.

http://dx.doi.org/10.1016/j.jenvrad.2014.11.020
0265-931X/

previous paper (Bossew, 2014a, referred to as "part 1", in the following). The basics of the technique are not repeated here. Instead, the focus is set to the extension of the method towards optimally grouping a *categorical predictor* (geology). Some exemplary results are shown and the reliability of such grouping is addressed.

The result is an optimal assemblage of a given set of geological classes into two sub-classes, "Rn prone geologies" (RPG) and "non-RPG". Optimality is to be understood as optimizing a defined loss function or "score" in ROC space. The approach is of "non-parametric", classification type, as opposed to the "parametric" one in Gruber et al. (2012). Among its features is that classification error rates are side results and that the relative importance of first and second kind errors can be set, which is difficult for other methods. To my knowledge, the method is novel in Rn studies. The results are compared with the ones from an alternative method which tests the "radon proneness" of geological classes independently, instead of assemblages of classes.

2. Methods

2.1. Data

German data were used for demonstrating the method. *Geogenic radon potential* (RP) is calculated from soil Rn and permeability as $RP := C(\text{soil})/(-\log_{10}k - 10)$ (modified from Neznal et al., 2004), C(soil) the Rn concentration in soil air in kBq/m^3 and k permeability in m^2, determined after the Kemski protocol (Kemski et al., 2002). There are nearly 4000 RP values distributed over Germany.

For indoor Rn concentrations, about 15,000 values (strongly spatially clustered) from ground floor living rooms in houses with full basements were used.

Geological classification is essentially based on a simplified scheme proposed by Kemski et al. (2001, 2009), slightly modified by merging some groups (e.g. all Mesozoic) and splitting up others. All data are stored in the German Rn databank "BURG". A list of the geological classes is given in the Annex, Table 2. The German territory is divided into 10 km × 10 km cells, to each of which (with a few exceptions) a geological class is assigned, see Fig. 5 in the Annex. (Ill-assignment of geology to cells, or geological misclassification of cells, is possible, but this is not subject of this study.)

2.2. Binary classification through ROC analysis for categorical predictors

The reader who is not familiar with the technique of ROC analysis is asked to consult part 1 of the article (Bossew, 2014a) first for an explanation how this technique works and where terminology is introduced. Here we attempt to resolve the following question.

Given a number (k) of geological classes, how can they be optimally assembled into two groups, so that the first group represents "*Rn prone geologies*" and the second group "non-Rn prone geologies", according to a given criterion. A "Rn prone geology" (RPG) is one for which a Rn related quantity shows higher levels than for "non-Rn prone geology". Among such quantities are the mean indoor Rn level, the probability that indoor Rn concentration exceeds a threshold or reference level, or the RP.

With k classes of a category (here the category is geology, its levels are the geological classes), one subset of these k classes is assigned to set U_0, the rest to set U_1. If a class (j; $j = 1 \ldots k$) belongs to U_0, let $g(j) = 0$, otherwise $g(j) = 1$. One assemblage can be written ($g(k)$, $g(k-1)$, …, $g(1)$), which can be read as a binary number whose value is $b = g(k)2^{k-1} + g(k-1)2^{k-2} + \ldots + g(1)2^0$. For example, for $k = 5$, the binary number [01001] means that classes 2, 3 and 5 belong to U_0 and classes 1 and 4 to U_1 (reading the sequence from right to left). Since a k-digit binary number runs from 0 to $\sum_{(i=0 \ldots k-1)}2^{i-1} = 2^k - 1$, this is the number of possible assemblages of the classes into U_0 and U_1.

In the ROC algorithm, instead of a number z which designs the threshold above which a cell characterized by continuous variable Z (for example its mean Rn potential RP) belongs into U_0, otherwise into U_1, we now use the index b, which is the value of the binary sequence ($g(k)$, …, $g(1)$). By varying b from 0 to $2^k - 1$, all possibilities of sorting of the classes into U_0 and U_1, are included.

For each b, all cells are checked for their geological class and the value of the standardizing variable or criterion, for example AM(C) > c or prob(C > c) > p, C = indoor Rn concentration, c a concentration and p a probability threshold. If the geology of the cell belongs to U_0 (defined by b) and the criterion is positively fulfilled, AM(C) > c, the cell adds to TP (true positive). If the cell belongs to U_0 but AM(C) < c, then it adds to FP (false positive), and so on. For each value b the FPR and TPR are calculated as usual, making one point in the TPR vs. FPR "ROC" graph. (FPR and TPR denote the false and true positive rates, defined FPR = false positives divided by observed negatives, FPR = FP/(FP + TN), and TPR = true positives divided by observed positives, TPR = TP/(TP + FN). The plot TPR vs. FPR is the ROC graph.)

Since now the "classifier" b is not a continuous quantity as the RP threshold, rp (part 1 of the article), there is no ROC curve, but a more or less dense scatter plot. However, statistics such as the Y score and others can be calculated the same way, and their optimization (maximum Y, minimum d01, etc.) yields an optimal b, which represents the optimal sequence of sorting the geological classes into U_0 (Rn prone geologies) and U_1 (non-Rn prone geologies). This sequence is recovered by back translating b into binary.

The scatter plot is symmetrical against FPR → 1 − FPR & TPR → 1 − TPR because each sequence appears as a "negative copy", for example [01001] and [10110].

A discussion of the reliability of an assemblage into RPG and non-RPG is given in the Results section after the presentation of some examples.

An obvious drawback of this otherwise simple method is that the computational effort increases exponentially with the number of classes (k). For 13 classes, $b_{\max} = 2^{13} - 1 = 8191$, and the number ($N$) of cells available in Germany (1428 cells occupied with at least one sample of indoor Rn concentration) computation takes a few seconds on an average PC using a QB64 Basic compiler. In each of the $N \cdot b_{\max} = 11{,}696{,}748$ instances (in this example) the program decides into which of the four fields of the truth table the instance belongs.

2.3. Classification criteria

For demonstrating the method, we use the following criteria for classifying the geologies.

CRIT1: AM(C in cell) > 100 Bq/m^3
CRIT2: prob(C > 100 Bq/m^3 in cell) > 0.1
CRIT3: GM(RP in cell) > 32

The cells are again 10 km × 10 km grid cells, identical to the ones used to produce the German RP map, and also identical to the grid used for the European indoor Rn map (e.g., Dubois et al., 2010; Tollefsen, 2014; Gruber et al., 2013). CRIT1 and CRIT2 are the same as used in part 1 of the article.

The probability in CRIT2 is estimated from empirical data in the following way, differently from how it has been done in part 1 of the article. It can be shown that for Z ~ Normal,

$$\text{prob}(Z > z) = t_{n-1}\, k\sqrt{n/(n+1)}$$