Journal of Environmental Radioactivity 129 (2014) 121–132

Contents lists available at ScienceDirect

Journal of Environmental Radioactivity

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

Determination of radon prone areas by optimized binary classification

P. Bossew

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

ARTICLE INFO

Article history:
Received 23 August 2013
Received in revised form
10 December 2013
Accepted 19 December 2013
Available online 10 January 2014

Keywords:
Radon prone area
Geogenic radon potential
ROC graph

ABSTRACT

Geogenic radon prone areas are regions in which for natural reasons elevated indoor radon concentrations must be expected. Their identification is part of radon mitigation policies in many countries, as radon is acknowledged a major indoor air pollutant, being the second cause of lung cancer after smoking. Defining and estimating radon prone areas is therefore of high practical interest.

In this paper a method is presented which uses the geogenic radon potential as predictor and thresholds of indoor radon concentration for defining radon prone areas, from which thresholds for the geogenic radon potential are deduced which decide whether a location is flagged radon prone or not, in the absence of actual indoor observations.

The overall results are different maps of radon prone areas, derived from the geogenic radon map, and depending (1) on the criterion which defines what a radon prone area is; and (2) on the choice of score whose maximization defines the optimal classifier. Such map is not the result of a transfer model (geogenic to indoor radon), but of the optimization of a classification rule. The method is computationally simple but has its caveats and statistical traps, some of which are also addressed.

1. Introduction

The basic concept of geogenic radon prone area (RPA) is a region where for natural, i.e. geogenic reasons elevated indoor radon (Rn) levels or an elevated probability of their occurrence must be expected. Geogenic factors are high radium concentrations in rock and soil, high permeability due to the coarse texture of the ground or due to fracturing of rocks, hydrological peculiarities and others. Knowing RPAs can assist allocating resources more efficiently for denser surveys, remediation of affected houses and regional implementation of stricter building codes for new houses. Therefore considerable work is being invested into methods of estimating such regions from observed data of geogenic quantities and/or indoor Rn measurements.

Although RPA as understood here is, by its very concept, a geogenic entity independent of anthropogenic factors, such as house type or living habits, its quantification – that is, decision whether a certain location is flagged RPA or not – should be linked to quantities related to Rn risk. The latter is often quantified through a proxy, namely the probability that indoor Rn concentration exceeds a threshold. (More precisely the chance of lung cancer depends on exposure to Rn progenies which is however more difficult to measure, and which is substituted by the average Rn concentration, or an exceedance probability.) Since indoor Rn concentration is to a high degree controlled by anthropogenic factors, the practically defined RPA also contains these factors, contrary to its concept. One tries to reduce the influence of these factors by standardizing the indoor Rn values used for "calibration" of what is a RPA, e.g. by restricting to ground floor rooms in houses with basement (to be applied in this study), by recalculating indoor values to standard conditions (Austrian approach; Friedmann, 2005), or to restricting to ground floor rooms in single-family houses which have not been remediated with respect to Rn (Belgian approach, Cinelli et al., 2011; B. Dehandschutter, personal comm.).

Several methods for defining RPAs in practice have been proposed. As spatial supports administrative or geological units or grid cells are common. Criteria to flag them as RPA include exceedance of mean indoor concentrations, or high probabilities that indoor Rn concentration thresholds are exceeded, or geological units which are known for high Rn potential (the geogenic property which leads to high indoor concentration in houses which allow Rn infiltration), etc. Quantities known to be related to indoor Rn, such as soil Rn (as in this article), geochemistry or dose rate may substitute indoor Rn for assessing RPAs.

In this article a binary classification based on ROC (receiver operating characteristic) curve analysis is proposed. It establishes an optimal classification of the predictor (radon potential) by optimizing the number of cells that have been classified correctly according to a target criterion, related to indoor Rn concentration. The optimum is defined by a score derived from the ROC statistic.

E-mail addresses: pbossew@bfs.de, peter.bossew@reflex.at.

0265-931X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jenvrad.2013.12.015

2. Materials and methods

2.1. Data

In Germany the following data are available currently (2013): a set of about 4000 measurements of Rn concentration in soil air together with soil permeability; about 40,000 measurements of indoor Rn concentrations, ca. 15,000 of which in ground floor rooms of houses with basement; and geology on 1:1 M resolution, or finer. The soil Rn sampling points are reasonably distributed over most of the territory, though being strongly clustered in certain regions. The indoor samples originate from various, mainly uncoordinated surveys. They are highly clustered and their representativeness is questionable in some cases. Some regions are very poorly covered. It is therefore difficult to derive estimates of RPAs directly from indoor Rn data, although of course they give an indication about which regions are affected.

The idea is therefore to use the geogenic quantities – geology as a categorical variable with area support and soil Rn as a continuous one with point support – as predictors and indoor Rn as "calibration" data. Based on these geogenic data one should be able to decide whether a location – in practice an area, either a grid cell or a geological polygon – is labelled RPA or not, given a certain definition of RPA.

2.2. The radon potential

From the geogenic data a radon potential (RP) is defined and a RP map of Germany created. The RP definition follows a proposal by Neznal et al. (2004):

$$\mathrm{RP} := \mathrm{C(soil)}/(-{}_{10}log(\mathrm{k}) - 10),$$

with C(soil) is the Rn concentration in soil determined by the "Kemski protocol" (Kemski et al., 2002) in kBq/m^3 and k, the permeability (same protocol) in m^2. The RP is essentially, for medium permeability, proportional to the advective flux component normalized to the pressure gradient across an interface (e.g. ground – house). For very high and low permeability the RP is smoothed against this quantity.

The RP is mapped for Germany on a 10 km × 10 km grid by a method described in Bossew (2013a,b); put shortly, geological classes are used as deterministic predictors on top of which cell estimation is performed by conditional (to the point values of RP) simulation. This yields cell-wise statistics of the RP (the local ccdf, or conditional cumulative distribution function), from which statistics such as the expectation or confidence intervals and exceedance probabilities can be derived. An alternative would be ordinary kriging which would however not allow estimating local distributions. (Other, more complicated varieties of kriging would.)

Whereas 3506 cells of area 10 km × 10 km are available which contain a value of the RP, covering the whole of Germany (357,121 km^2) with a few exceptions, the numbers of cells which contain a minimum number of indoor data is much smaller, as a result of their strong spatial clustering. While 1428 cells with at least one observation are available, only 407 are for at least 10, and 126 cells have at least 30 observations.

2.3. Classification by ROC analysis

The classification of cells x* containing estimates RP(x*) is performed as follows.

(1) In those cells where enough indoor Rn data C are available, calculate *empirical* statistics of C, such as arithmetical or geometrical mean or the empirical probability to exceed a threshold, e.g. prob(C > 100 Bq/m^3) = number of C > 100 divided by total number of samples in the cell.
(2) Establish a criterion CRIT for a cell to be labelled RPA, such as: E[C] > 100 Bq/m^3; or: E[prob(C > 100 Bq/m^3)] > 10%, where E denotes the expectation.
(3) For each cell x* decide whether the criterion is met for the empirical estimates. This results in a binary coding of the cells, positive/negative, or {1,0}.
(4) For a given value rp of the quantity RP, decide for each cell x*, whether RP(x*) ≥ rp or RP(x*) < rp. This results in another binary coding of the cells.
(5) Next, the two codings or classifications are compared, and the value rp_0 which classifies RP(x*) such that the coincidence becomes optimal, is determined. To this end one calculates the following statistics:
   - True positives, TP: number of cells classified or "predicted" as RPA through RP and through C;
   - False positives, FP: number of cells classified as RPA though RP, but not through C;
   - True negatives, TN: number of cells classified as non-RPA through both RP and C;
   - False negatives, FN, number of cells classified a non-RPA through RP, although they are RPA as classified by C. The logic is summarized in Fig. 1.
(6) Evidently, TP + FP + TN + FN = total number of cells. From these numbers several statistics can be derived, in particular:
   - True positive rate, or "sensitivity", also "recall":

$$\mathrm{TPR} := \mathrm{TP}/(\text{observed positives}) = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}),$$

   and
   - False positive rate, or 1-"specificity":

$$\mathrm{FPR} := \mathrm{FP}/(\text{observed negatives}) = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$$

TPR shall be as high, while FPR as low as possible; they cannot be chosen independently, however. Therefore one has to find a trade-off considered optimal, after some score.

(7) Plotting TPR against FPR results in a graph called ROC, or receiver operating characteristic, an example of which (for the data discussed here) is shown in Fig. 2. The shape of the ROC graph depends on the chosen criterion CRIT; points of the graph correspond to values of the threshold rp for classification according to the RP. By tuning the value rp the classification can be optimized such as to minimize misclassifications (essentially by false positives or negatives), yielding an optimum, rp_0. The diagonal indicates a random association between the classifications, whereas a perfect classification would be a point at (0,1).
(8) Several scores derived from the ROC have been proposed whose optimization lead to an optimal classifier rp_0. The corresponding point on the curve can be considered the best

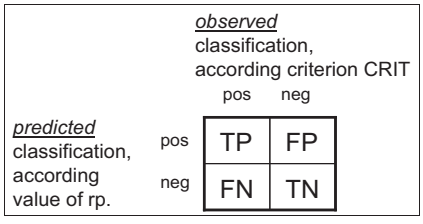

Fig. 1. Classification table and definition of true and false positives and negatives (TP, FP, TN, FN).