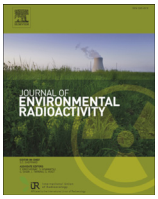
Journal of Environmental Radioactivity 143 (2015) 100–109

Contents lists available at ScienceDirect

Journal of Environmental Radioactivity

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



Log-normality of indoor radon data in the Walloon region of Belgium

Giorgia Cinelli [a], [*], François Tondeur [b]

[a] European Commission, Joint Research Centre (JRC), Institute for Transuranium Elements (ITU), Nuclear Security Unit, Via Enrico Fermi 2749, 21027 Ispra, VA, Italy
[b] ISIB, Haute Ecole P.-H. Spaak, Rue Royale 150, 1000 Brussels, Belgium

ARTICLE INFO

Article history:
Received 12 November 2014
Received in revised form
5 February 2015
Accepted 14 February 2015
Available online 9 March 2015

Keywords:
Indoor radon
Geology
Log-normal distribution
Walloon region of Belgium

ABSTRACT

The deviations of the distribution of Belgian indoor radon data from the log-normal trend are examined. Simulated data are generated to provide a theoretical frame for understanding these deviations. It is shown that the 3-component structure of indoor radon (radon from subsoil, outdoor air and building materials) generates deviations in the low- and high-concentration tails, but this low-C trend can be almost completely compensated by the effect of measurement uncertainties and by possible small errors in background subtraction. The predicted low-C and high-C deviations are well observed in the Belgian data, when considering the global distribution of all data.

The agreement with the log-normal model is improved when considering data organised in homogeneous geological groups. As the deviation from log-normality is often due to the low-C tail for which there is no interest, it is proposed to use the log-normal fit limited to the high-C half of the distribution. With this prescription, the vast majority of the geological groups of data are compatible with the log-normal model, the remaining deviations being mostly due to a few outliers, and rarely to a "fat tail". With very few exceptions, the log-normal modelling of the high-concentration part of indoor radon data is expected to give reasonable results, provided that the data are organised in homogeneous geological groups.

1. Introduction

Indoor radon is recognised as one of the major indoor pollutants, the second cause of lung cancer (WHO, 2009). The presence of radon in the indoor atmosphere is highly variable from one house to the other, and it must be considered as a random variable to be studied with statistical tools. Among them, it is essential to choose a statistical distribution fitting the data as well as possible, in particular to allow various predictions to be made on the basis of a limited number of data. For example, the percentage of houses having a radon concentration higher than some reference level is often used as the indicator of the risk of indoor radon pollution in a given area.

It is known that the indoor radon data can be approximately described by a log-normal distribution (Nero, 1988), as can be expected from the central-limit theorem (EOM, 2012) for a variable which is the product of many independent random factors. In most cases, the interest is focussed on the high-concentration side of the distribution. Deviation of high-concentration data from log-normality, more precisely an excess of high values (the "fat tail effect") was noted and studied by several authors. We shall only refer here to a few recent works, a more systematic list being given in table 2 of Bossew (2010). Tuia and Kanevski (2008) examined the quality of the fit to the Swiss data using the log-Gumbel distribution developed for extreme-value theory, and conclude that it often better performs for describing the high-value tail of indoor radon data. Bossew (2010) suggested that the LN could be adequate for local samplings of Austrian data, and thus useful for mapping, the deviations from log-normality increasing with the size of the sampled area. For Murphy and Organo (2008) who use the Irish data, the log-normal model can provide good estimates, but it is necessary to discard the outliers.

These authors did not consider any stratification of the data, e.g. according to geology. Kies et al. (1996) found a somewhat better agreement with the LN for data grouped according to the two geographic–geological regions of the grand Duchy of Luxembourg. Andersen et al. (2001) include some geological information, but shift the radon concentration data by 8 Bq/m^3 in order to improve their log-normality (a correction which obviously is mostly

---

[*] Corresponding author. Tel.: +39 0332 786620; fax: +39 0332 785466.
E-mail address: giorgia.cinelli@jrc.ec.europa.eu (G. Cinelli).

http://dx.doi.org/10.1016/j.jenvrad.2015.02.014
0265-931X/

important for low concentrations). Cinelli et al. (2009) considered data from the Walloon region of Belgium, grouped according to the geological age of the local formation. They confirmed the observation of a “fat tail”, but concluded that log-Gumbel does not perform better than the simple log-normal distribution. Toth et al. (2006) considered another, more complex classification of Hungarian data, concluding that homogeneous classes are well described by the log-normal, although the global sampling is not. However, the logic of their classification is not really explained, and the fact that geology is not considered for houses built with bricks looks a bit strange. Cinelli and Tondeur (2010) noted that strong deviations from log-normality also exist in the low-concentration tail, which can influence the parameters of the distribution and worsen the fit to the high-C tail. The low data bear no special interest, as they do not correspond to an unacceptable health risk. Moreover, they are the most uncertain in terms of relative measurement errors. It would be a paradox if the evaluation of the high-concentration tail was partially determined by the specificities and uncertainties of the low side.

All these works have in common a purely empirical approach, with no underlying physical model. But deviations from log-normality can be expected on a theoretical basis because indoor radon is not a pure product of random factors, but the sum of three largely independent components, each of which is the product of several factors: radon from the subsoil under the house, radon from the building materials, and radon from outdoors. The three components are different from the point of view of their mean value as well as for their variability between the houses. Deviations from log-normality are also expected if the distribution is bimodal or multimodal, for example if the data set includes data from a low-risk region and from a high risk region, without the full spectrum of intermediate situations. Recently, Daraktchieva et al. (2014) showed that the distribution of indoor radon data in the UK can be reproduced as the sum of six log-normal modes, after subtracting a constant outdoor contribution. However, they do not distinguish between the contributions of the building materials and of the subsoil, and do not consider separately the different geological formations.

The goal of the present work is to see to what extent these considerations may help to understand the deviations with respect to the LN, and to better specify in which conditions this distribution can be used or not. The Belgian data will be analysed globally, but also separately for the low- and high-concentration data. We shall also examine how the quality of the log-normal fit can be improved by separating the data in homogenous geological groups, for which the subsoil contribution could be assumed to be unimodal. Simulated data will first be calculated, providing a theoretical frame which can help to understand the empirical results.

2. Materials and methods

2.1. Indoor radon database

The indoor radon database we use for the Walloon region of Belgium was described in (Cinelli et al., 2010). More data were included since then, and the database now includes 18772 data for the Walloon region, 75% being long-term data (LT) and 25% short-term data (ST), all being measured on the ground floor. The area of the region is 16,844 km^2, and the average sampling density is thus slightly above 1 data/km^2.

For each house in which a radon measurement is available, the database includes the geographical coordinates (Belgian Lambert system 1972), the radon concentration on the ground floor, and the local geological unit determined with the digital geological map (GSB). Unfortunately, this old geological map does not allow a precise and automated determination of lithology. The lithological information is thus absent of the database.

2.2. Division in geological units

As shown in a recent work (Tondeur et al., 2014), the geological series is often the most appropriate division for radon risk mapping in the Wallon region, but a second division according to the geographical region or massif is nearly always necessary for obtaining a total of 35 geological units with a reasonable degree of uniformity for the radon risk. Even so, some of them remain inhomogeneous geological units. They are listed in Appendix 1.

Quaternary is not included in this classification, because of specific problems explained in our recent work (Tondeur et al., 2014).

2.3. Log-normality tests

Considering the natural logarithm of the ground floor radon concentration ln(C) as the variable, the deviation of its distribution with respect to the normal one will be studied by several means:

- The quantile–quantile plot (qq-plot), which should be linear for a normal distribution; a special attention will be given to the outliers, which we shall define as data more than 1 LSD (logarithmic standard deviation) apart from the log-normal prediction,
- The Shapiro–Francia normality test (Shapiro and Francia, 1972) with Royston (1993) algorithm, well-adapted for the samples including many data,
- The root-mean-square deviation between the actual distribution of ln(C) and the normal one with the same mean and standard deviation. We shall particularly consider the ratio of the RMS deviation to the logarithmic standard deviation RMS/LSD.

This threefold approach is followed because the traditional tests (like Shapiro–Francia) give an answer to a question “Are the data compatible with a normal distribution, taking into account the fluctuations related to the random sampling?”, which is not the practical one. It assumes that the distribution is normal, and this ideal distribution is expected to show up when the number of data becomes large enough. But the actual distribution is not exactly normal, and this will appear more and more clearly when the number of data is increased, as noted by Bossew (2010). Thus the large samplings never fulfill such tests, even when they can be fitted by a log-normal with a reasonable accuracy.

Our practical question is: “How good is the fit of the data by a normal distribution, thus how good could it perform when used to predict unknown data?”. This is well expressed by the RMS deviation.

2.4. Simulations

To facilitate the interpretation of the deviations of the qq-plot from linearity, simulated distributions were calculated to display the consequences of:

- The three-component structure of indoor radon (subsoil, building materials, outdoor) instead of a single log-normally distributed component,
- A possible multimodal distribution,
- Inexact reporting and encoding of data lower than the minimal detectable activity: one of the four laboratories which participated to the measurements included in the database reported