A first version of a Pan-European Indoor Radon Map

Javier Elío[1], Giorgia Cinelli[2], Peter Bossew[3], José Luis Gutiérrez-Villanueva[4], Tore Tollefsen[2], Marc De Cort[2], Alessio Nogarotto[5], Roberto Braga[5]

[1]Geology, School of Natural Sciences, Trinity College, Dublin, Ireland
[2]European Commission, Joint Research Centre (JRC), Ispra, Italy
[3]German Federal Office for Radiation Protection, Berlin, Germany
[4]Radonova Laboratories AB, Uppsala, Sweden
[5]Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Università di Bologna, Italy

Correspondence to: Giorgia Cinelli (Giorgia.CINELLI@ec.europa.eu)

Abstract. A hypothetical Pan-European Indoor Radon Map has been developed using summary statistics estimated from 1.2 million indoor radon samples. In this study we have used the arithmetic mean (AM) over grid cells of 10 km x 10 km to predict a mean indoor radon concentration at ground-floor level of buildings in the grid cells where no or few data (N < 30) are available. Four interpolation techniques have been tested: inverse distance weighted (IDW); ordinary kriging (OK); collocated cokriging with uranium concentration as secondary variable (CoCK); and regression kriging with topsoil geochemistry and bedrock geology as secondary variables (RK). Cross-validation exercises have been carried out to assess the uncertainties associated with each method. Of the four methods tested, RK has proved to be the best one for predicting mean indoor radon concentrations; and by combining the RK predictions with the AM of the grids with 30 or more measurements, a Pan-European Indoor Radon Map has been produced. This map represents a first step towards a European radon exposure and, further on, a radon dose map.

1 Introduction

Radon (Rn) is the major contributor to the ionizing radiation dose received by the general population, being the second cause of lung cancer death after smoking (WHO, 2009). Worldwide radon exposure is linked to an estimated 222,000 out of the 1.8 million lung-cancer cases reported per year (Gaskin et al., 2018), and in Europe alone it has been estimated that 18,000 lung-cancer cases per year are induced by radon (Gray et al., 2009). Since lung-cancer survival rates after five years are as low as below 20% (Cheng et al., 2016), a reduction in radon exposure will have a significant positive impact on the health of the general population. In this context, the EU recently revised and consolidated the Basic Safety Standards Directive (Council Directive 2013/59/EURATOM), which aims to reduce the number of radon-induced lung cancer cases.

The main sources of radon indoors are the surrounding subsoils on which buildings are located, the ground water used in the building, and the building materials (Cothern and Smith, 1987). Consequently, radon is present everywhere. The likelihood of having high indoor radon concentration may, however, be higher in some areas than others. Radon maps are

therefore an essential tool at large scale and give very good indications of the problem, helping policy-makers to design cost-effective radon action plans (Gray et al., 2009). Importantly, because of high local variability, large-scale Rn maps do not inform about Rn concentration in a particular building. Instead, this requires measuring in that building.

In 2006, the EU's Joint Research Centre (JRC) launched a long-term project to map radon at the European level (Tollefsen et al., 2014). For more than ten years now, the JRC has been developing a European Atlas of Natural Radiation (Cinelli et al., 2019). It includes maps of the natural radioactive levels of: i) annual cosmic-ray dose; ii) indoor radon concentration; iii) uranium, thorium and potassium concentration in soil and in bedrock; iv) terrestrial gamma dose rate; and v) soil permeability. Digital versions of these maps are available from a JRC website (https://remon.jrc.ec.europa.eu/About/Atlas-of-Natural-Radiation), and updated at irregular intervals when new data become available. The objectives of this Atlas are: (1) to increase public knowledge of natural ionizing radiation; (2) to analyse the level of natural radioactivity caused by different sources; (3) to produce a better estimate of the annual dose to which the general population is exposed; and (4) to compare natural and artificial sources (Cinelli et al., 2019).

The European Indoor Radon Map (EIRM) displays the annual average indoor radon concentration (Rn; ^{222}Rn) measured on ground floor of buildings over 10 km x 10 km grid cells (Dubois et al., 2010). Based on input-data specifications stipulated by the JRC, European countries provide summary statistics estimated over 10 km x 10 km grid cells without communicating the original data, thus guaranteeing data privacy confidentiality for the individual house owners. As a result, the European indoor radon dataset contains the following parameters: the arithmetic mean and standard deviation of the indoor radon measurements (AM_z and SD_z) and the log-transformed data (AM_lnz and SD_lnz); the median (Med), the minimum (Min), and the maximum (Max) values; as well as the total number (N) of dwellings sampled in each grid cell (Tollefsen et al., 2014).

The dataset underlying the EIRM represents a huge amount of work. At the time of writing (end-2018), 32 countries (EU and non-EU Member States alike) have contributed data, and information from almost 1.2 million dwellings has been aggregated into 28,468 grid cells. Since some cells overlap between countries, 28,203 of these grid cells were filled by 1 country, while 262 and 3 grids were filled by two and three countries, respectively (i.e. border areas which share the same grid) (version: 29-09-2018). However, there are still a large number of grid cells over European land territory with no data, and the number of measurements per grid cell varies widely, from many with only 1 measurement up to a single one with 23,993 dwellings sampled (Table 1). Evaluating the radon exposure to European citizens would therefore require another ten years, or more, if it had to be done based on indoor radon measurements over each grid cell.

Interpolation techniques are therefore essential at this stage to predict a mean indoor radon concentration in the grid cells for which no or few data are available, and thus develop a Pan-European Indoor Radon Map. We have tested four interpolation techniques: two that use solely indoor radon concentration measurements, viz. inverse distance weighted (IDW) and ordinary kriging (OK); and another two which also take into account geological information, viz. collocated cokriging with uranium concentration in topsoil as secondary variable (CoCK) and regression kriging with topsoil geochemistry and bedrock geology as secondary variables (RK). Cross-validation exercises were carried out to assess the uncertainties associated

with each method. The map generated here is a hypothetical indoor Rn map in the sense that it estimates the mean per 10 km × 10 km grid cell under the assumption that there are dwellings in the grid cell. In some remote areas (mountains, extreme Northern Europe), however, this may not be the case in reality. The final map represents a first step towards a European radon exposure and, further on, a radon dose map. Furthermore, it may assist European countries in developing their respective National Indoor Radon Maps.

2 Methods

2.1 Indoor radon data

The primary dataset used to predict the mean per grid cell with no or few data is the one of arithmetic means (AM_z). The AM was assigned to the centre of each grid cell, and predictions were carried out only in grid cells where U, Th, and K_2O concentrations also were available (46,000 grid cells; version 28-05-2018, Pantelić et al., 2018). Data from grid cells filled by more than one country (i.e. points with the same coordinates) were merged, and the summary statistics recalculated according to Eq. 1-10:

$$AM = \frac{S}{N} \quad (1)$$

$$SD = \sqrt{\frac{SQ - \frac{s^2}{N}}{N-1}} \quad (2)$$

$$Med = \sqrt{\prod_{i=1}^{n} Med_i} \text{ (Approximation)} \quad (3)$$

$$Min = Min[Min_i] \quad (4)$$

$$Max = Max[Max_i] \quad (5)$$

$$N = \sum_{i=1}^{n} N_i \quad (6)$$

$$S = \sum_{i=1}^{n} S_i \quad (7)$$

$$S_i = AM_i \cdot N_i \quad (8)$$

$$SQ = \sum_{i=1}^{n} SQ_i \quad (9)$$

$$SQ_i = SD_i \cdot (N_i - 1) + \frac{S_i}{N_i} \quad (10)$$

where "i" is the number of countries that filled the grid. The values for the log-transformed data (AM_lnz and the SD_lnz) were estimated with the same equations as used for the AM and the SD, but with the ln values provided by each country (i.e. AM_lnz, and SD_lnz).

In the study area (i.e. area with topsoil geochemistry data) there are 25,367 grid cells with indoor radon measurements (Figure 1). The distribution of the AM is approximately lognormal (Figure 2), with values ranging from 1 to 10,116 Bq m^{-3}. The summary statistics are shown in Table 2. Nominal concentrations below 10 Bq m^{-3} are unrealistic from the point of view

of true occurrence and measurement possibility, but this could not be verified in this context. The impact of such errors on the result is probably negligible.

2.2 Interpolation techniques

A mean (over a 10 km × 10 km grid cell) radon concentration at ground-floor level was estimated at 1 m off the grid centroid, to which the AMs in the input database are referenced. Predictions were therefore carried out at locations slightly different from the ones of the data. The reason is that we wanted to avoid exact interpolations. To some extent, indoor radon variations at small scale can be taken into account this way.

2.2.1 Inverse Distance Weighted (IDW) Interpolation

The Inverse Distance Weighted (IDW) Interpolation technique estimates a weighted average at an unsampled point ($\hat{Z}_o$) according to its distance (d_i) to the sampled points (Z_i):

$$\hat{Z}_0 = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} Z_i}{\sum_{i=1}^{n} \frac{1}{d_i^p}} \text{ if } d_i > 0; \text{ otherwise } (d_i=0): \hat{Z}_0 = Z_i \quad (11)$$

where "p" is the inverse distance weighting power (idp) which represent "the degree to which the nearer points are preferred over more distant points" (Bivand et al., 2008). IDW assumes that, on average, nearby points are more similar to each other than more distant points (Li and Heap, 2008), and therefore the weights for the closer ones are higher than the weights for distant points.

The result is highly influenced by the inverse distance weighting power chosen. An optimal value of p which minimizes a loss function L, p_{opt} = argmin L(data, target locations; p), can be found for example by k-fold cross validation. The loss function has to be defined by the user, and a common choice is the Root-Mean-Square Error (RSME) (Janik et al., 2018). In our case the optimal idp was found to be 1.5 (Figure 3), and interpolations of the AM were carried out using the observations within a distance of 1,000 km, and a minimum and maximum number of nearest observations was set to 5 and 75, respectively.

2.2.2. Ordinary Kriging (OK)

Trans-Gaussian kriging using Box-Cox transforms (function krigeTg in R software, package "gstat" and "MASS"; Gräler et al., 2016; Kendall et al., 2016; Pebesma, 2004; R Core Team, 2018; Venables and Ripley, 2002) was performed with the arithmetic mean. The normal transformation of data (X) with the transformation parameters lambda (λ) follow Eq. 12 (Box and Cox, 1964):

$$\phi_\lambda^{-1} = \begin{cases} \frac{X^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(X) & \lambda = 0 \end{cases} \quad (12)$$

Predictions are carried out over the transformed data, and then unbiased back-transformed to the original scale using the Lagrange multiplier (Eq. 13-15; Cressie, 1993; Varouchakis et al., 2012):

$$\hat{Z}(S_o) = \phi\left(\hat{Y}_{OK}(S_0)\right) + \phi''(\hat{\mu})\left(\frac{\sigma^2_{OK}(S_0)}{2} - m\right) \quad (13)$$

$$\phi(x) = \begin{cases} (x \cdot \lambda)^{\frac{1}{\lambda}} & \lambda \neq 0 \\ e^x & \lambda = 0 \end{cases} \quad (14)$$

$$\phi''(x) = \begin{cases} (1-\lambda)(x \cdot \lambda + 1)^{\frac{1}{\lambda}-2} & \lambda \neq 0 \\ e^x & \lambda = 0 \end{cases} \quad (15)$$

where $\hat{Z}(S_o)$ is the ordinary kriging predictor on the original scale, $\hat{Y}_{OK}(S_0)$ is the ordinary kriging predictor on the transformed scale data, $\sigma^2_{OK}(S_0)$ the ordinary kriging variance, $\hat{\mu}$ the mean estimate at each location, and m the Lagrange multiplier of the OK system for each location (Kozintsev et al., 1999).

The variogram was modelled with two components: a Matérn model (Minasny and McBratney, 2005; Pardo-Iguzquiza and Chica-Olmo, 2008) up to a distance of 50 km; and an exponential model up to 1,500 km (Figure 4). The very low kappa (0.15) points to high "roughness" of the field. Predictions were then carried out with observations within a distance of 1,000 km, and using a minimum and a maximum number of nearest observation of 5 and 75, respectively.

2.2.3 Collocated CoKriging (CoCK) with uranium as secondary variable

Collocated Cokriging (CoCK) is a special case of cokriging where only the direct correlation between the primary (e.g. AM_z) and the secondary variables (e.g. U) is used, ignoring the direct variogram of the secondary variable and the cross variograms. It simplified the cokriging equations although the secondary variable must be sampled at all prediction points (Bivand et al., 2008). The method is a simplification of the physical reality, because the dependence structure between covariates is more complex, as they result from different physical processes.

We performed the CoCK with uranium concentration in topsoil as secondary variable since radon is generated in the uranium decay series (Cothern and Smith, 1987), and a positive correlation between uranium and indoor radon is therefore expected. The analysis was carried out with the data log-transformed and then back-transformed to original scale (AM_z) with Eq. 16-17 (where μ is the kriging prediction and σ the kriging variance):

$$E[X] = e^{(\mu + \frac{\sigma^2}{2})} \quad (16)$$

$$var[X] = e^{(2\mu + \sigma^2)} \cdot (e^{\sigma^2} - 1) \quad (17)$$

The uranium map (Figure 5a; Tollefsen et al., 2016), part of the European Atlas of Natural Radiation, has been created using approximately 5,000 topsoil data from GEMAS and FOREGS datasets (i.e. GEMAS: Geochemical Mapping of Agricultural and Grazing Land soil in Europe, GEMAS 2008; and FOREGS: Geochemical Atlas of Europe developed by the Forum of European Geological Surveys, FOREGS 2005). Uranium explains about 7.75% of the total indoor radon variability (correlation coefficient = 0.2783; Figure 5b). As in the previous cases, a maximum distance of 1,000 km and a minimum and maximum number of nearest observation of to 5 and 75, respectively, were used in the predictions.

2.2.4 Regression Kriging (RK)

Regression Kriging (RK) is a two-step interpolation technique: first, a regression estimation of the dependent variable (e.g. AM_z) is performed against secondary variables (e.g. geogenic factors); and, second, an analysis of the spatial distribution of the residual is carried out using geostatistical methods (i.e. OK; Pásztor et al., 2016). The final estimates are the sums of the regression estimates and the ordinary kriging estimates of the residuals (Di Piazza et al., 2015). The analysis was also carried out with the log-transformed data, and directly back-transformed with the same equation as in CoCK.

The technique applied in the regression step can vary (Li and Heap, 2008); here, we have performed a linear regression using topsoil geochemistry (i.e. U and K_2O) and geology (i.e. 1:5 Million International Geological Map of Europe – IGME 5000; Asch, 2003) as secondary variables. The IGME 5000 has been developed by the German Federal Institute for Geosciences and Natural Resources; this European GIS project involved more than 40 European and adjacent countries, covering an area from the Caspian Sea in the east, to the Mid-Ocean Ridge in the west, and from Svalbard Islands in the north to the southern shore of the Mediterranean Sea. The aim of the project was to develop a GIS underpinned by a geological database. The original IGME map presents 178 lithostratigraphic units that were reduced to 28 lithological units (Figure 6). Based on ANOVA tests ran on an extensive Italian geological database, Nogarotto et al. (2018) demonstrated that lithology alone has a large effect on geochemical variations of key elements (U, Th, K_2O), regardless of the tectono-stratigraphic position of a given unit. It is therefore assigned the prevalence unit to each grid of 10 km x 10 km (Figure 6).

The procedure is therefore: i) to fit a linear model to the data (Figure 7a and Table 3), the total indoor radon variance explained by U, K_2O, and simplified geology is 20.24 % (7.75%, 7.88%, and 4.61% respectively); ii) to analyse the spatial distribution of residuals, Ordinary Kriging (Figure 7b); iii) to predict a radon value (i.e. log(AM_z)) in each grid using the linear model, and add the residual predictions; and iv) to back-transform to the original scale with the equations described in the previous section (Eq. 16-17; where μ is the linear model prediction plus the ordinary kriging prediction of the residuals, and σ is the kriging variance).

2.2.5 Cross-validation

The performances of the different methods were assessed by 5x10-Fold Cross-Validation, and by Moving Windows Cross-Validation (Kasemsumran et al., 2006). For the 5x10-fold cross-validation method, data were randomly split into 10 subgroups and predictions were carried out 10 times, each time one group is used for validation and 9 for modelling the variable of interest (i.e. AM_z) at the validation locations. This process is then repeated 5 times, obtaining a total of 50 realizations. The Moving Windows Cross-Validation (MWCV) was carried out with cell sizes of 200 km x 200 km (total number of windows = 197). Grid cells within a window are removed, and an AM is predicted with the rest, then errors are calculated and the process is repeated with another window until all windows are covered. Some restrictions to the validation set were used to avoid errors during kriging methods (i.e. number of grids in the validation set higher than one; var(log[U]) > 0; and geological units of the validation set must also be in the model set).

The accuracy of the different methods was assessed using six indicators: the mean absolute error (MAE), the root-mean-square error (RMSE), the root-mean-squared log error (RMSLE), the index of agreement (IA), the percentage bias (PB), and the coefficient of determination (R^2) (Eq. 18-23).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Z_i - X_i| \quad (18)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i - X_i)^2} \quad (19)$$

$$RMSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(Z_i + 1) - \log(X_i + 1))^2 \quad (20)$$

$$IA = 1 - \frac{\sum_{i=1}^{n}(Z_i - X_i)^2}{\sum_{i=1}^{n}(|X_i - \bar{X}| - |Z_i - \bar{X}|)^2} \quad (21)$$

$$PB = 100\ \frac{\sum_{i=1}^{n}(Z_i - X_i)}{\sum_{i=1}^{n}X_i} \quad (22)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Z_i - X_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (23)$$

where Z_i and X_i are the predicted and measured values in the validation location (S_i), n the number of points in the validation group, and $\bar{X}$ the mean of X_i. MAE and RMSE are commonly used for assessing model performance; however, they may be influenced by outliers (Chen et al., 2017). RMSLE, on the contrary, is less sensitive to outliers and preferable when there is a large range in the values (Janik et al., 2018). These parameters are positive values, and the closer they are to 0, the better is the model fit. IA is a standardized measure of the degree of model prediction error; it varies from 0 (no agreement at all) to 1 (perfect match). PB (%) measures the average tendency of having larger/smaller predicted values than the observed ones. The optimal value is 0, and positive/negative values indicate over/under-estimation bias (Janik et al., 2018). Finally, R^2 is a measure of how well the model fits a data set; a perfect model has $R^2 = 1$ (Alexander et al., 2015).

3 Results and discussion

3.1 Cross-validation

The 5 x 10-Fold Cross-Validation (Figure 8 and Table 4) shows that geostatistical techniques (i.e. OK, CoCK, RK), which take into account the spatial autocorrelation of the data, generally perform better (i.e. lower MAE, RMLSE; and higher R^2) than IDW. However, they have a tendency to overestimate bias (PB > 0). Then, geostatistical results are slightly improved when geological information is added. The model which has the highest R^2 is RK (median = 0.2462), followed by CoCK (0.2460), OK (0.2377). RK also is the geostatistical technique with higher IA (0.6014) and lower PB (2.513), and it has similar MAE and RMSLE as OK and CoCK (around 47 and 0.36, respectively).

Similar results are obtained in the MWCV exercise (Table 5). Geostatistical techniques (i.e. OK, CoCK, RK) also have the highest R^2 and the lowest MAE and RMLSE. However, in these cases the RK bias is close to 0 (PB = -0.98), while OK and CoCK overestimate the values. MWCV also suggests that results are slightly improved when geogenic factors are taken into account: e.g. R^2 increases from 0.3457 (OK), to 0.3512 (CoCK), and then to 0.3687 (RK); and the highest IA is

obtained with RK (0.4531). However, similar MA, RMSE, and RMSLE also were found (around 54, 136 and 0.48, respectively) which indicates the difficulty of predicting an average indoor radon concentration even when secondary variables are added.

3.2 Indoor radon predictions

Radon predictions with the different methods range from minimum values of 1 - 4 Bq m^{-3} to up to 10,116 Bq m^{-3}, while the mean values are in the order of 95 – 105 Bq m^{-3} (Table 6). The very high value of an AM (i.e. 10,116 Bq m^{-3}) seems improbable, although the grid is in a region with uranium deposits and former uranium mines (border region between Spain and Portugal). This cell has only two measurements (i.e. 9,726 and 10,507 Bq m^{-3}), so that the level of reliability of this extremely high AM is therefore low and it would probably decrease if the number of data were increased. In this sense, IDW interpolation, which gives an exact interpolation when the distance between the predicted and measured points is zero, estimates a value that is the arithmetic mean (i.e. 10,116 Bq m^{-3}). Nevertheless, when the spatial autocorrelation between cells is considered (i.e. OK, CoCK and RK), the predicted values, although also high, are reduced to 2,500 - 2,800 Bq m^{-3}. These latter values may be more realistic; and are similar to average values found in some villages of the region (i.e. 1,851 Bq m^{-3} in Villar de la Yegua, Spain; Sainz et al. 2010). However, this effect shows the difficulties with predicting an AM when the number of measurements in a grid cell is low. Geostatistical techniques may help to overcome some of these limitations, although the reliability of data because of different numbers of measurements (e.g. grids with only 1 or 2, and other with more than 20-30 measurements) is still a problem. Nor is it clear whether in an "anomalous area" such as the one cited above, where the geological conditions are particular, the covariance function (or the variogram) which has been estimated from all data, still applies. One can assume that in such a region second-order stationarity is violated. But the accuracy of local prediction depends very much on the local covariance model.

Small differences may be appreciated in the predictions of the different interpolation techniques (Figure 9). IDW and OK are methods that rely on the Rn data only, while CoCK and RK use additional predictors (i.e. geology, U and K_2O concentration in the ground) as secondary variables. The first type is weak in areas with no conditioning data as it simply interpolates between existing ones, ignoring physical reality in these areas (e.g. South Italy, North-East Germany). Including it is the rationale of the second type; practically, missing conditioning data of the primary variable (Rn) are substituted by functions of the secondary variables. Although certainly more reasonable in the physical sense, this type is analytically more complicated.

The influence of geogenic factors on indoor radon is well known, and normally used for radon mapping (e.g. Casey et al., 2015; Elío et al., 2017; Pásztor et al., 2016; Scheib et al., 2013; Tondeur et al., 2014). In our cases, an ANOVA analysis (Table 3) shows that the total indoor radon variance explained by U, K_2O, and geology is about 20% (7.75%, 7.88% and 4.61%, respectively). Uranium is a source of radon in soil, and thus a positive association with indoor radon is expected (e.g. Appleton et al., 2008; Ferreira et al., 2018). However, the Pearson's correlation coefficient between indoor radon and uranium concentration in topsoil is relatively low (r = 0.2783), which implies that CoCK estimations with U as secondary variable are

still mainly based on the primary variable (i.e. AM; Rocha et al., 2012). Therefore, although CoCK performs slightly better than OK, spatial predictions are similar (Figure 9).

Geology is associated with both uranium/radon source, and with physical properties which permit the release of radon from the soil matrix and its transport in the environment (e.g. mineralogy, porosity, permeability, etc.). The total indoor radon variance explained by geology is normally in the order of 5% - 25% (Appleton and Miles, 2010; Borgoni et al., 2014; Miles and Appleton, 2005; Tondeur et al., 2014; Watson et al., 2017), although it depends on the geological scale map (i.e. increase with the scale; Appleton and Miles, 2010). A 4.64% of indoor radon variation explanation is therefore reasonable, taking into account that we used a simplified 1:5 million geological map, and that data are averaged over grids of 10 km x 10 km.

The positive correlation between indoor radon and potassium is, however, not evident. K_2O may be related with clay content in soils (e.g. Barré et al., 2008; Milošević et al., 2017; Tarvainen et al., 2005), and although the permeability of wet clays is low, it may increase when soils are dried (Petersell et al., 2005) as a consequence of building a house (Barnet et al., 2008). This hypothesis should be tested, since clay soils are normally considered as low risk although its radium concentration may be high (Maestre and Iribarren, 2018). We have decided, however, to include this parameter into the model since previous studies have shown a positive association between indoor radon and K_2O/clay (Forkapic et al., 2017).

Back-transform predictions to the original scale is a critical point of lognormal and trans-Gaussian kriging. OK as given in this study solves this problem by using the Lagrange multiplier in the back-transformation. However, the E[X] and Var[X] for CoCK and RK are biased, unless the true mean is known (although for RK it should be zero by definition). These equations should also use the Lagrange multiplier which appears in the kriging system (Chilès and Delfiner, 1999; Matheron, 1974); but unfortunately in common geostatistical packages this parameter is not accessible, and it is not easy to estimate it. Another problem with lognormal kriging is that ill-assessment of the kriging SD leads to large errors in E[X] and Var[X] due to exponentiation, so that variogram parameters must be estimated very carefully (Armstrong and Boufassa, 1988). Deviations from stationarity and uni- as well as multivariate lognormality are also critical (Cressie, 1993; Roth, 1998). On the other hand, in highly skewed quantities (as is typical for Rn and in fact for many positive-definite environmental quantities such as concentrations) there seems to be little choice but to work with transformed (e.g. log, Box-Cox, Nscore) variables.

Finally, a theoretical problem, if using kriging-type interpolators, may be the fact that input data are actually cell or grid means (blocks in geostatistical language), treated as point samples. The change of support problem, which is particularly unpleasant in lognormal kriging, may be alleviated since the target supports are also the same. We regard input data as point data at the cell centre, and estimate points at other locations that again represent cells of the same size. However, the theoretical aspect remains to be clarified in more depth. Taking into account all of these limitations and weaknesses, the solution demonstrated here however represents an acceptable compromise between mathematical exactness, numerical tractability, and complexity of the physical realm.

4 A Pan-European indoor radon map

We would like to produce a Pan-European Indoor Radon Map by minimising data processing, and therefore we prefer to estimate the radon average directly by indoor radon measurements carried out at each grid (i.e. AM_z). However, if the number of measurements were low, the uncertainty of this value could be high. In this sense, if dwellings were randomly selected and therefore representative, which is the condition for unbiased estimates of the mean and other statistics, and the sample size large, the mean value and the confidence interval would be (Eq. 24):

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \pm t_{(1-\frac{\alpha}{2},n-1)} \frac{s}{\sqrt{n}} \quad (24)$$

An additional condition for the validity of the confidence interval is statistical independence of the data. For large n, due to the Central Limit Theorem, $t_{1-\alpha/2;n-1}$ can be approximated by the normal score $x_{1-\alpha/2}$.

The confidence interval decreases when the sample size increases. In our cases (Figure 10), the relative (to the mean) $CI_{95\%}$ ($\alpha = 0.05$) for sample size of about 30-40 data is around $\pm$5%, and generally lower than 15 - 30%. Therefore, although the assumption of data independence is not valid (i.e. there is spatial correlation between indoor radon measurements which can be modelled by the variogram), 30 measurements seems reasonable for obtaining a good estimation of the radon exposure in a specific grid (Figure 11). However, if sampled dwellings were highly clustered, the AM could be not representative of the radon exposure in a grid even with high numbers of indoor radon measurements.

For the final Pan-European Indoor Radon Map (Table 7 and Figure 12), we use therefore the AM of the grid cells with 30 or more measurements (Figure 11), and the value predicted by RK (Figure 9) in the cells with less than 30 measurements. Indoor radon concentration ranges from 3 to 2,662 Bq m^{-3}, with a mean value of 98 Bq m^{-3}. The standard deviation may be calculated with the SD of the measurements carried out in the grids with 30 or more data, and with the kriging standard deviation of the RK (i.e. grids with less than 30 measurements). It ranges from 1 to 3,233 Bq m^{-3}, with RSD from 3% to up to 1,101%. The map appears "noisy" to varying degrees between different regions. The reason is that in regions with more conditioning, Rn data, local variability of the estimate is higher than in regions with sparse or without data, where the estimate is based essentially on geology and geochemistry. These covariates are much smoother on the scale available to us than Rn data, where available. Would we dispose of denser geochemical data and higher resolution geological maps, also these regions would appear noisier.

5 Conclusions

After more than 10 years of collecting and processing Rn data, with the support of 32 European countries, we could cover approximately 50% of the continent with 10 km × 10 km grids containing the mean indoor radon concentration in ground floors of dwellings. However, completing the European Indoor Radon Map still requires a significant effort by the participating countries, since a robust estimation of the radon exposure in a grid of 10 km x 10 km involves at least 30 indoor radon measurements and, at the time of writing this article, most of the grids cells sampled (78%) have less than 20 dwellings tested.

Interpolation techniques which take advantage of the contiguity of Rn seen as spatial random field, may help to overcome some of the present limitations, and would permit estimating radon exposure at European scale until the coverage of all Europe with indoor radon measurements has strongly improved.

Of the four methods tested in this study, Regression Kriging (RK), using a simplified geological map and the topsoil concentration of U and K_2O, has proven to be the best one for predicting mean indoor radon concentrations over grids of 10 km x 10 km (i.e. arithmetic mean, ground floor). By combining RK predictions with empirical average values (AM) in grids with 30 or more measurements, a Pan-Europe Indoor Radon Map has been created. The map represents the average value of indoor radon concentration on ground floor, and thus it is not representative of the radon exposure to European citizens since most people do not live on ground floor (Cinelli et al., 2019). However, it is the first step towards a radon exposure and, further on, a dose map. Based on demographic and sociological databases, we plan to develop models which allow correcting from ground-floor dwellings to the real situation, accounting for the building stock (Bossew et al., 2018).

The Pan-European Indoor Radon Map is not a finished map, and it will be upgraded as new data become available. In future versions a larger scale of the geological map (e.g. scale 1:1 million), as well as other geogenic factors which may influence the indoor radon concentration (e.g. soil units, aquifer types) would be included in the model. Furthermore, the influence of anthropogenic factors and those which may affect building characteristics and living styles (e.g. average temperatures, annual precipitation, altitude, etc.) will be analysed. Finally, machine-learning techniques are viewed as promising methods for modelling the AM since kriging-type predictions come to an end if many (inter-correlated) predictors are involved.

Acknowledgements

We wish to thank all the national competent authorities, universities and laboratories who have provided, and continue to provide, indoor radon data to the JRC (see https://remon.jrc.ec.europa.eu/About/Atlas-of-Natural-Radiation/Indoor-radon-AM/Indoor-radon-concentration). The sole responsibility of this publication lies with the authors. The European Commission and national competent authorities are not responsible for any use that may be made of the information contained herein.

References

Alexander, D. L. J., Tropsha, A. and Winkler, D. A.: Beware of R 2 : Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models, J. Chem. Inf. Model., 55(7), 1316–1322, doi:10.1021/acs.jcim.5b00206, 2015.

Appleton, J. D. and Miles, J. C. H.: A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk., J. Environ. Radioact., 101(10), 799–803, doi:10.1016/j.jenvrad.2009.06.002, 2010.

Appleton, J. D., Miles, J. C. H., Green, B. M. R. and Larmour, R.: Pilot study of the application of Tellus airborne radiometric and soil geochemical data for radon mapping., J. Environ. Radioact., 99(10), 1687–97, doi:10.1016/j.jenvrad.2008.03.011,

2008.

Armstrong, M. and Boufassa, A.: Comparing the robustness of ordinary kriging and lognormal kriging: Outlier resistance, Math. Geol., 20(4), 447–457, doi:10.1007/BF00892988, 1988.

Asch, K.: The 1 : 5 Million International Geological Map of Europe and Adjacent Areas: Development and Implementation of a GIS-enabled Concept, Geologisches Jahrbuch SA, Schweizerbart Science Publishers, Stuttgart, Germany., 2003.

Barnet, I., Pacherová, P., Neznal, M. and Neznal, M.: Radon in geological environment: Czech experience - Special Papers No. 19, Czech Geological Survey., 2008.

Barré, P., Montagnier, C., Chenu, C., Abbadie, L. and Velde, B.: Clay minerals as a soil potassium reservoir: Observation and quantification through X-ray diffraction, Plant Soil, 302(1–2), 213–220, doi:10.1007/s11104-007-9471-6, 2008.

Bivand, R. S., Pebesma, E. J. and Gómez-Rubio, V.: Applied Spatial Data Analysis with R, edited by R. Gentleman, G. Parmigiani, and K. Hornik, Springer New York, New York, NY., 2008.

Borgoni, R., De Francesco, D., De Bartolo, D. and Tzavidis, N.: Hierarchical modeling of indoor radon concentration: how much do geology and building factors matter?, J. Environ. Radioact., 138, 227–237, doi:10.1016/j.jenvrad.2014.08.022, 2014.

Bossew, P., Cinelli, G., Tollefsen, T., Cort, M. De, Gruber, V., García-Talavera, M., Elío, J. and Villanueva, J. G.: From the European indoor radon concentration map to a European indoor radon dose map, in VI Terrestrial Radioisotopes in Environment. International Conference on Environmental Protection, Social Organization for Radioecological Cleanliness, Veszprém, Hungary., 2018.

Box, G. E. P. and Cox, D. R.: An analysis of transformations, J. R. Stat. Soc. Ser. B (Methodological, 26(2), 211–252, doi:10.2307/2287791, 1964.

Casey, J. A., Ogburn, E. L., Rasmussen, S. G., Irving, J. K., Pollak, J., Locke, P. A. and Schwartz, B. S.: Predictors of indoor radon concentrations in Pennsylvania, 1989-2013, Environ. Health Perspect., 123(11), 1130–1137, doi:10.1289/ehp.1409014, 2015.

Chen, C., Twycross, J. and Garibaldi, J. M.: A new accuracy measure based on bounded relative error for time series forecasting, PLoS One, 12(3), 1–23, doi:10.1371/journal.pone.0174202, 2017.

Cheng, T. Y. D., Cramb, S. M., Baade, P. D., Youlden, D. R., Nwogu, C. and Reid, M. E.: The international epidemiology of lung cancer: Latest trends, disparities, and tumor characteristics, J. Thorac. Oncol., 11(10), 1653–1671, doi:10.1016/j.jtho.2016.05.021, 2016.

Chilès, J.-P. and Delfiner, P.: Geostatistics – Modeling Spatial Unvertainty, John Wiley., 1999.

Cinelli, G., Tollefsen, T., Bossew, P., Gruber, V., Bogucarskis, K., De Felice, L. and De Cort, M.: Digital version of the European Atlas of natural radiation, J. Environ. Radioact., 196(August 2017), 240–252, doi:10.1016/j.jenvrad.2018.02.008, 2019.

Cothern, R. and Smith, J.: Environmental radon, edited by C. R. Cothern and E. J. Smith, Plenum Press, New York., 1987.

Cressie, N.: Statistics for Spatial Data, Wiley, New York. [online] Available from:

http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471002550.html, 1993.

Dubois, G., Bossew, P., Tollefsen, T. and De Cort, M.: First steps towards a European atlas of natural radiation: Status of the European indoor radon map, J. Environ. Radioact., 101(10), 786–798, doi:10.1016/j.jenvrad.2010.03.007, 2010.

Elío, J., Crowley, Q., Scanlon, R., Hodgson, J. and Long, S.: Logistic regression model for detecting radon prone areas in Ireland, Sci. Total Environ., 599–600, 1317–1329, doi:10.1016/j.scitotenv.2017.05.071, 2017.

Ferreira, A., Daraktchieva, Z., Beamish, D., Kirkwood, C., Lister, T. R., Cave, M., Wragg, J. and Lee, K.: Indoor radon measurements in south west England explained by topsoil and stream sediment geochemistry, airborne gamma-ray spectroscopy and geology, J. Environ. Radioact., 181, 152–171, doi:10.1016/j.jenvrad.2016.05.007, 2018.

FOREGS: Geochemical Atlas of Europe, [online] Available from: http://weppi.gtk.fi/publ/foregsatlas/index.php (Accessed 28 January 2018), 2005.

Forkapic, S., Maletić, D., Vasin, J., Bikit, K., Mrdja, D., Bikit, I., Udovičić, V. and Banjanac, R.: Correlation analysis of the natural radionuclides in soil and indoor radon in Vojvodina, Province of Serbia, J. Environ. Radioact., 166, 403–411, doi:10.1016/j.jenvrad.2016.07.026, 2017.

Gaskin, J., Coyle, D., Whyte, J. and Krewksi, D.: Global Estimate of Lung Cancer Mortality Attributable to Residential Radon, Environ. Health Perspect., 126(5), 1–8, doi:10.1289/EHP2503, 2018.

GEMAS: Geochemical mapping of agricultural and grazing land soil, [online] Available from: http://gemas.geolba.ac.at/ (Accessed 28 January 2018), 2008.

Gräler, B., Pebesma, E. and Heuvelink, G.: Spatio-Temporal Interpolation using {gstat}, R J., 8 (1), 204–218 [online] Available from: http://journal.r-project.org/archive/2016-1/na-pebesma-heuvelink.pdf, 2016.

Gray, A., Read, S., McGale, P. and Darby, S.: Lung cancer deaths from indoor radon and the cost effectiveness and potential of policies to reduce them., BMJ, 338, a3110, 2009.

Janik, M., Bossew, P. and Kurihara, O.: Machine learning methods as a tool to analyse incomplete or irregularly sampled radon time series data, Sci. Total Environ., 630, 1155–1167, doi:10.1016/j.scitotenv.2018.02.233, 2018.

Kasemsumran, S., Du, Y. P., Li, B. Y., Maruo, K. and Ozaki, Y.: Moving window cross validation: A new cross validation method for the selection of a rational number of components in a partial least squares calibration model, Analyst, 131(4), 529–537, doi:10.1039/b515637h, 2006.

Kendall, G. M., Miles, J. C. H., Rees, D., Wakeford, R., Bunch, K. J., Vincent, T. J. and Little, M. P.: Variation with socioeconomic status of indoor radon levels in Great Britain: The less affluent have less radon, J. Environ. Radioact., 164, 84–90, doi:10.1016/j.jenvrad.2016.07.001, 2016.

Kozintsev, B., Kozintsev, S. and Kedem, B.: Kriging In Splus, [online] Available from: http://www.math.umd.edu/~bnk/bak/Splus/kriging.html (Accessed 28 January 2018), 1999.

Li, J. and Heap, A. D.: A Review of Spatial Interpolation Methods for Environmental Scientists, Geosci. Aust. Rec. 2008/23, GeoCat# 68(2008/23), 137, doi:http://www.ga.gov.au/image_cache/GA12526.pdf, 2008.

Maestre, C. R. and Iribarren, V. E.: The radon gas in underground buildings in clay soils. The plaza balmis shelter as a

paradigm, Int. J. Environ. Res. Public Health, 15(5), doi:10.3390/ijerph10051004, 2018.

Matheron, G.: Effet proportionnel et lognormalité ou: le retour du serpent de mer, , 45, 1974.

Miles, J. C. H. and Appleton, J. D.: Mapping variation in radon potential both between and within geological units, J. Radiol. Prot., 25(3), 257–276, doi:10.1088/0952-4746/25/3/003, 2005.

Milošević, M., Logar, M., Kaluderović, L. and Jelić, I.: Characterization of clays from Slatina (Ub, Serbia) for potential uses in the ceramic industry, Energy Procedia, 125, 650–655, doi:10.1016/j.egypro.2017.08.270, 2017.

Minasny, B. and McBratney, A. B.: The Matérn function as a general model for soil variograms, Geoderma, 128(3–4), 192–207, doi:10.1016/j.geoderma.2005.04.003, 2005.

Nogarotto, A., Cinelli, G. and Braga, R.: U, Th and K concentration in bedrock data: validity of geological grouping (country study Italy)., 2018.

Pantelić, G., Eliković, I., Živanović, M., Vukanac, I., Nikolić, J., Cinelli, G. and Gruber, V.: Literature review of Indoor radon surveys in Europe, Publications Office of the European Union (JRC114370), Luxembourg., 2018.

Pardo-Iguzquiza, E. and Chica-Olmo, M.: Geostatistics with the Matern semivariogram model: A library of computer programs for inference, kriging and simulation, Comput. Geosci., 34(9), 1073–1079, doi:10.1016/j.cageo.2007.09.020, 2008.

Pásztor, L., Szabó, K. Z., Szatmári, G., Laborczi, A. and Horváth, Á.: Mapping geogenic radon potential by regression kriging, Sci. Total Environ., 544, 883–891, doi:10.1016/j.scitotenv.2015.11.175, 2016.

Pebesma, E. J.: Multivariable geostatistics in S: The gstat package, Comput. Geosci., 30(7), 683–691, doi:10.1016/j.cageo.2004.03.012, 2004.

Petersell, V., Åkerblom, G., Ek, B.-M., Enel, M., Mõttus, V. and Täht, K.: Radon risk map of Estonia: Explanatory text to the Radon Risk Map Set of Estonia at scale of 1:500 000 (SSI Report 2005:16 - SGU Dnr. c., 2005.

Di Piazza, A., Conti, F. Lo, Viola, F., Eccel, E. and Noto, L. V.: Comparative analysis of spatial interpolation methods in the Mediterranean area: Application to temperature in Sicily, Water (Switzerland), 7(5), 1866–1888, doi:10.3390/w7051866, 2015.

R Core Team: R: A language and environment for statistical computing, [online] Available from: https://www.r-project.org/, 2018.

Rocha, M. M., Yamamoto, J. K., Watanabe, J. and Fonseca, P. P.: Studying the influence of a secondary variable in collocated cokriging estimates, An. Acad. Bras. Cienc., 84(2), 335–346, doi:10.1590/S0001-37652012005000017, 2012.

Roth, C.: Is Lognormal Kriging Suitable for Local Estimation?, , 30(8), 999–1009, doi:10.1023/A:1021733609645, 1998.

Sainz, C., Gutierrez-Villanueva, J. L., Fuente, I., Quindos, L., Soto, J., Arteche, J. L. and Quindos Poncela, L. S.: Two significant experiences related to radon in a high risk area in Spain, Nukleonika, 55(4), 513–518, 2010.

Scheib, C., Appleton, J., Miles, J. and Hodgkinson, E.: Geological controls on radon potential in England, Proc. Geol. Assoc., 124(6), 910–928, doi:10.1016/j.pgeola.2013.03.004, 2013.

Tarvainen, T., Reeder, S. and Albanese, S.: Database management and map production (K - Potassium), Geochemical atlas Eur. part, 2005.

Tollefsen, T., Cinelli, G., Bossew, P., Gruber, V. and De Cort, M.: From the European indoor radon map towards an atlas of natural radiation, Radiat. Prot. Dosimetry, 162(1–2), 129–134, doi:10.1093/rpd/ncu244, 2014.

Tollefsen, T., De Cort, M., Cinelli, G., Gruber, V. and Bossew, P.: 04. Uranium concentration in soil. European Commission, Joint Research Centre (JRC) [Dataset]: http://data.europa.eu/89h/jrc-eanr-04_uranium-concentration-in-soil, [online] Available from: http://data.europa.eu/89h/jrc-eanr-04_uranium-concentration-in-soil, 2016.

Tondeur, F., Cinelli, G. and Dehandschutter, B.: Homogenity of geological units with respect to the radon risk in the Walloon region of Belgium, J. Environ. Radioact., 136, 140–151, doi:10.1016/j.jenvrad.2014.05.015, 2014.

Varouchakis, E. A., Hristopulos, D. T. and Karatzas, G. P.: Improving kriging of groundwater level data using nonlinear normalizing transformations—a field application, Hydrol. Sci. J., 57(7), 1404–1419, doi:10.1080/02626667.2012.717174, 2012.

Venables, W. N. and Ripley, B. D.: Modern Applied Statistics with S, Fourth Edition, Springer, New York., 2002.

Watson, R. J., Smethurst, M. A., Ganerød, G. V., Finne, I. and Rudjord, A. L.: The use of mapped geology as a predictor of radon potential in Norway, J. Environ. Radioact., 166, 341–354, doi:10.1016/j.jenvrad.2016.05.031, 2017.

WHO: WHO Handbook on Indoor Radon: A Public Health Perspective, edited by H. Zeeb and F. Shannoun, World Health Organization, France, 2009.

Table 1: Number of dwelling sampled by grid cells of 10x10 km in the study area

Dwellings	Number of grids
N = 1	6,643
1 < N ≤ 5	9,064
5 < N ≤ 10	3,306
10 < N ≤ 20	3,161
20 < N ≤ 30	1,896
30 < N ≤ 23,993	4,398
TOTAL	28,468

Table 2: Summary statistics of indoor radon data (AM_z) after merged border grids (N = 25,367).

	Min.	Q1	Median	Mean	Q3	Max
AM [Bq m^{-3}]	1	40	71	103	123	10,116
SD [Bq m^{-3}]	0	20	47	89	100	6,873
RSD [%]	0	45	67	72	92	370

Table 3: ANOVA table for indoor radon concentration

	df	Sum Sq.	Mean Sq.	F Value	Pr(>F)	
Log(U)	1	1457.8	1457.77	2461.303	<2.2e-16	***
Log(K2O)	1	1483.6	1483.58	2504.891	<2.2e-16	***
Sim. Geology	27	868.1	32.15	54.228	<2.2e-16	***
Residuals	25337	15006.5	0.59			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4: 5 x 10-Fold Cross-validation results

Method	MAE	RMSE	RMSLE	IA	PB	R^2
IDW	50.07	113.44	0.4189	0.5755	-0.346	0.2352
OK	46.98	112.10	0.3728	0.5680	4.785	0.2377
CoCK	46.62	111.64	0.3711	0.5741	5.326	0.2460
RK	47.41	111.73	0.3744	0.6014	2.513	0.2462

Table 5: Moving windows cross-validation results

Method	MAE	RMSE	RMSLE	IA	PB	R^2
IDW	57.756	138.33	0.5457	0.4116	-1.899	0.1001
OK	53.926	136.60	0.4765	0.4142	3.758	0.3457
CoCK	53.990	136.28	0.4870	0.4033	2.851	0.3512
RK	55.573	136.41	0.4863	0.4531	-0.980	0.3687

Table 6: Summary indoor radon predictions (AM, ground floor)

Method	Min	Q1	Median	Mean	Q3	Max	SD
IDW	1	52	84	105	129	10,116	115.14
OK	4	52	80	95	120	2,546	67.40
CoCK	3	51	79	95	121	2,768	69.33
RK	3	51	79	98	123	2,661	73.81

Table 7: Summary indoor radon at European scale

	Min	Q1	Median	Mean	Q3	Max	SD
AM (Bq m^{-3})	2.8	50.8	78.7	97.1	122.2	2661.4	76.1
SD (Bq m^{-3})	1.1	28.0	45.0	61.7	73.4	3232.7	76.7
RSD (%)	2.9	44.9	60.9	60.3	67.2	1101.0	22.8

Figure 1: Arithmetic mean (AM_z) over 10 km x 10 km grid cells (Bq m^{-3}) and Relative Standard Deviation (RSD = AM/SD)

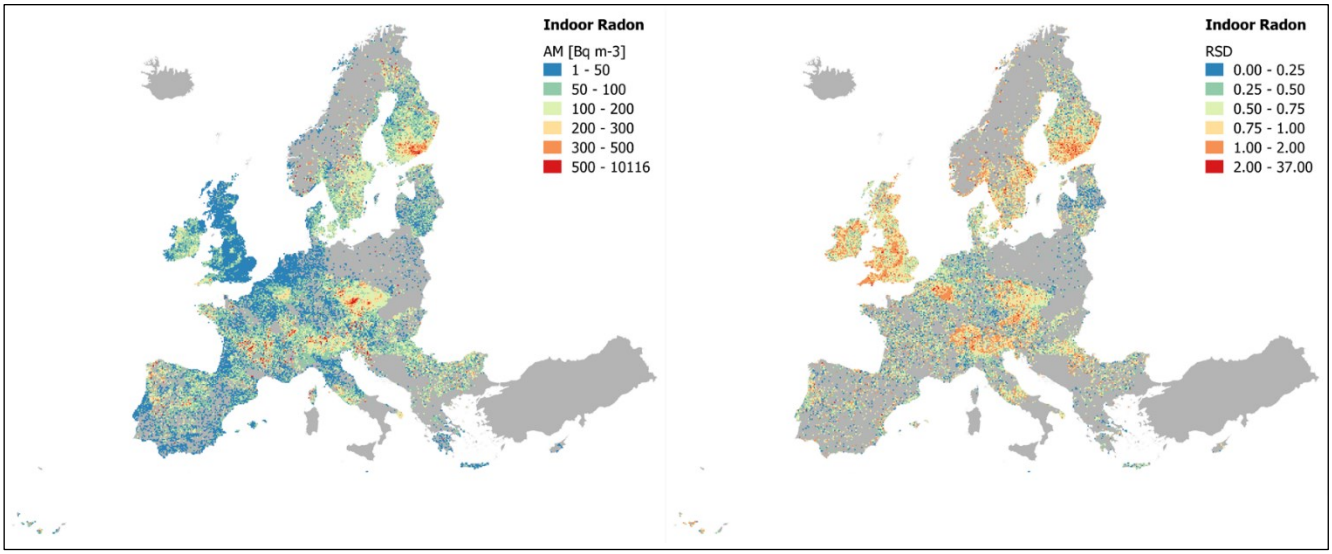


Figure 2: Histogram and q-q plot of average indoor radon concentration (AM_z) in ground floor of dwellings

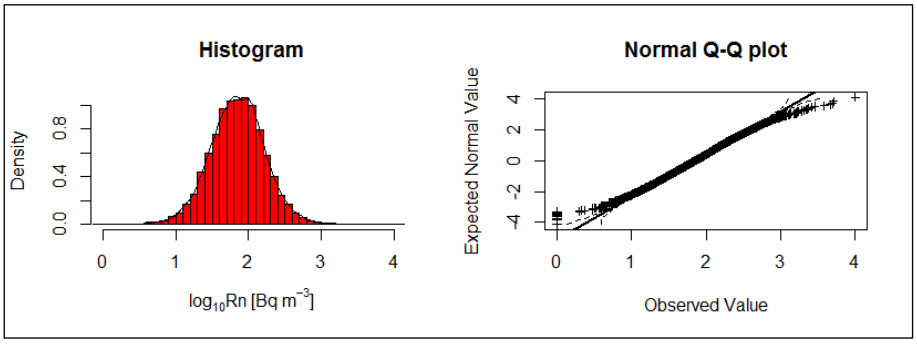


Figure 3: Inverse distance weighting power (idp) optimization

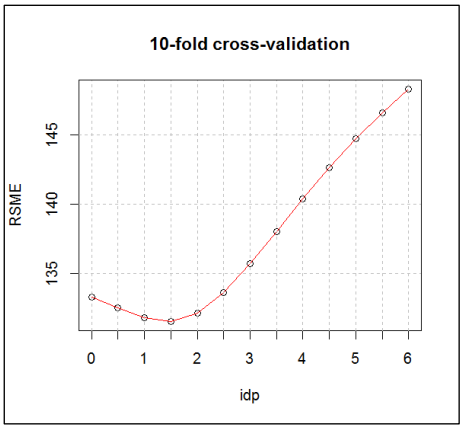


Figure 4: Model variogram (blue line; green dots are pairs of points up to a distance of 50km, and red points up to 1500 km), and 100 variograms from random permutations of the data (grey lines)

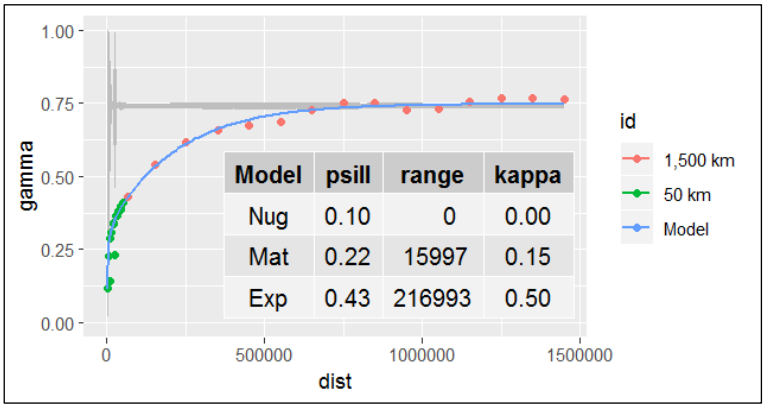


Figure 5: a) Uranium concentration in topsoil (Max = 9.73 mg km-1; Tollefsen et al., 2016), and b) scatterplot between indoor radon and uranium concentration in topsoil

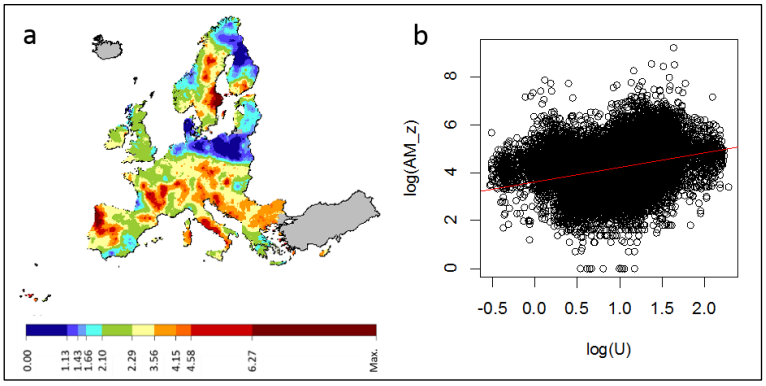


Figure 6: Simplified geology map with geological units defined on lithology basis (Nogarotto et al. 2018). The base geological map is the IGME (Asch 2003)

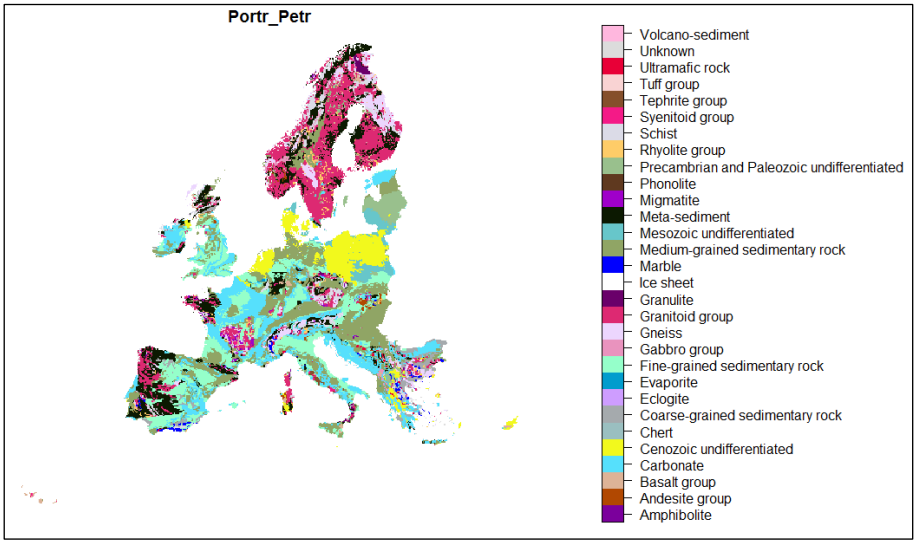


Figure 7: a) Linear model and b) variogram of residuals

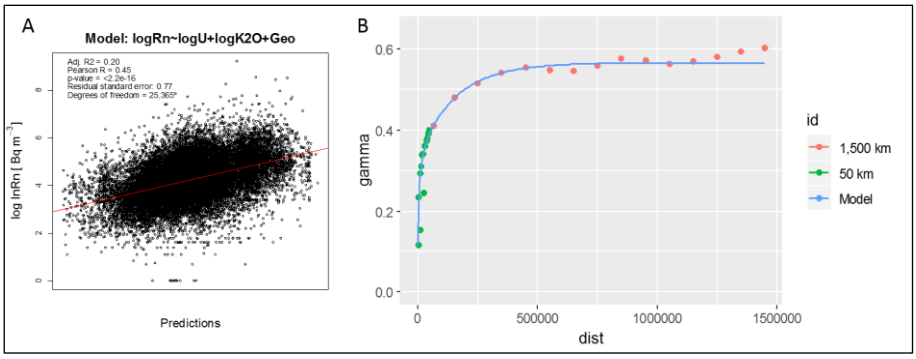
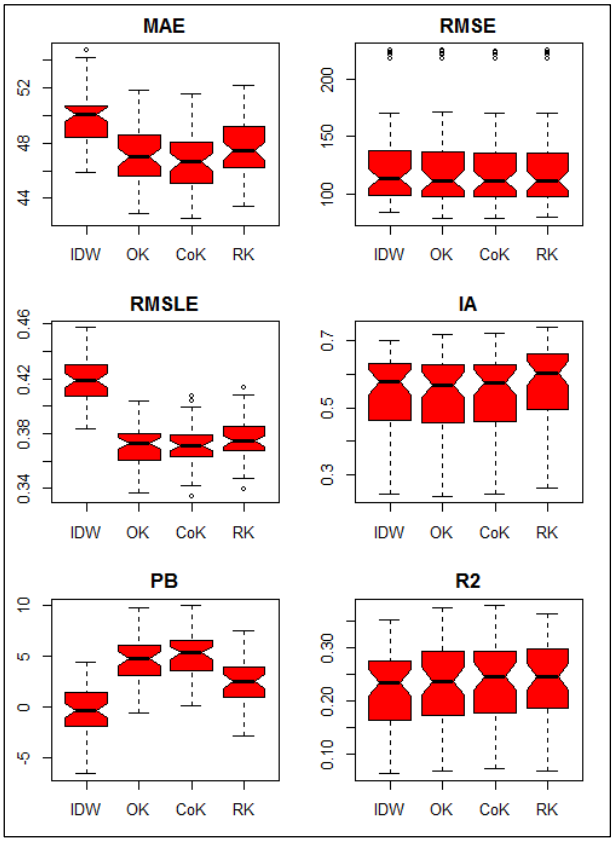


Figure 8: Boxplot of the 5x10 fold cross-validation results



5

Figure 9: Indoor radon predictions (AM [Bq m^{-3}], ground floor)

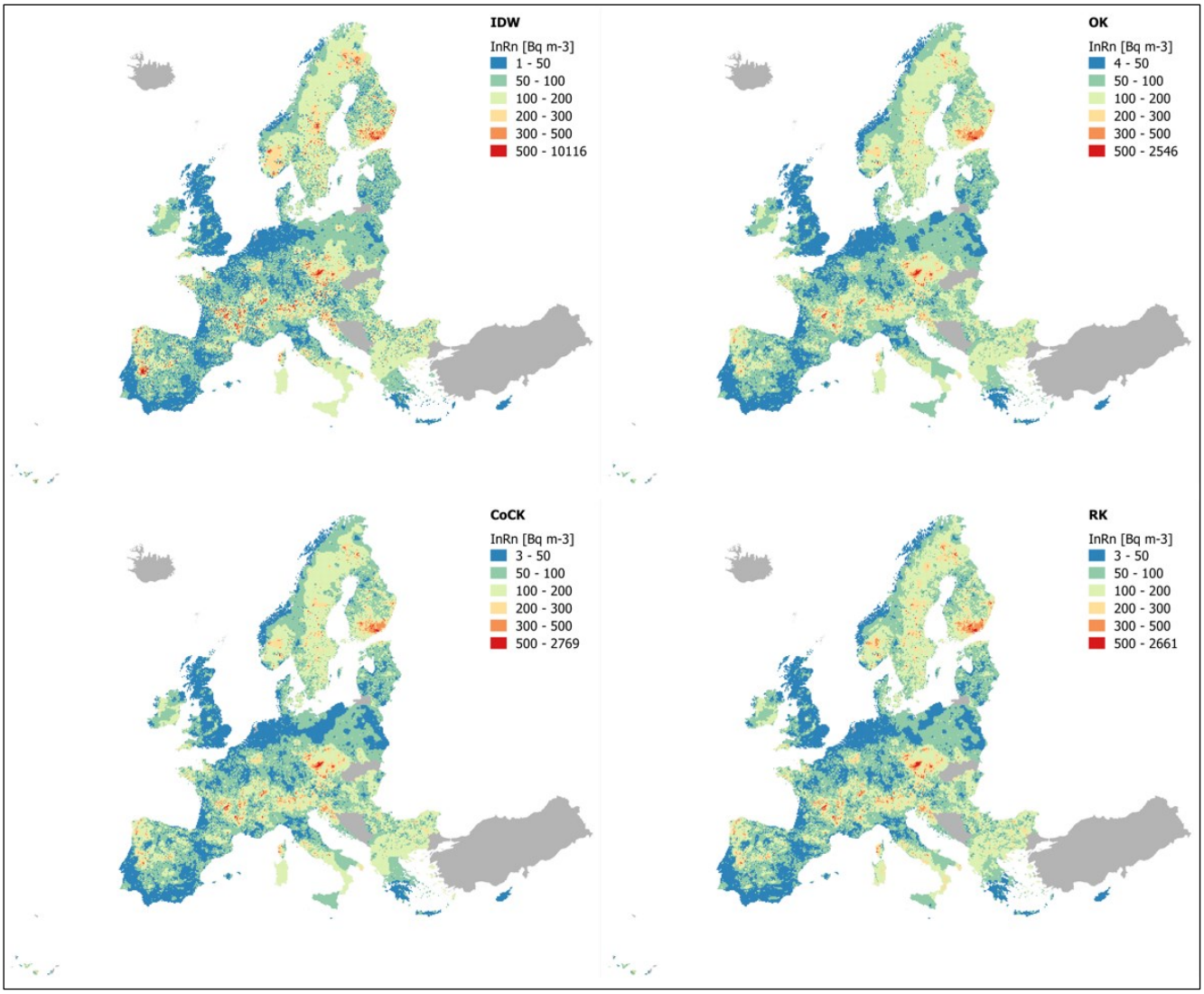


Figure 10: Variation of the 95% confidence interval of the arithmetic mean according to the sample size (N)

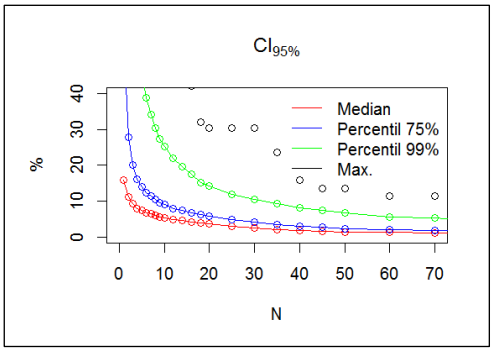


Figure 11: Grids with 30, or more, indoor radon measurements (N = 4,173; AM: Arithmetic Mean in Bq m^{-3}, RSD: Relative Standard Deviation in %)

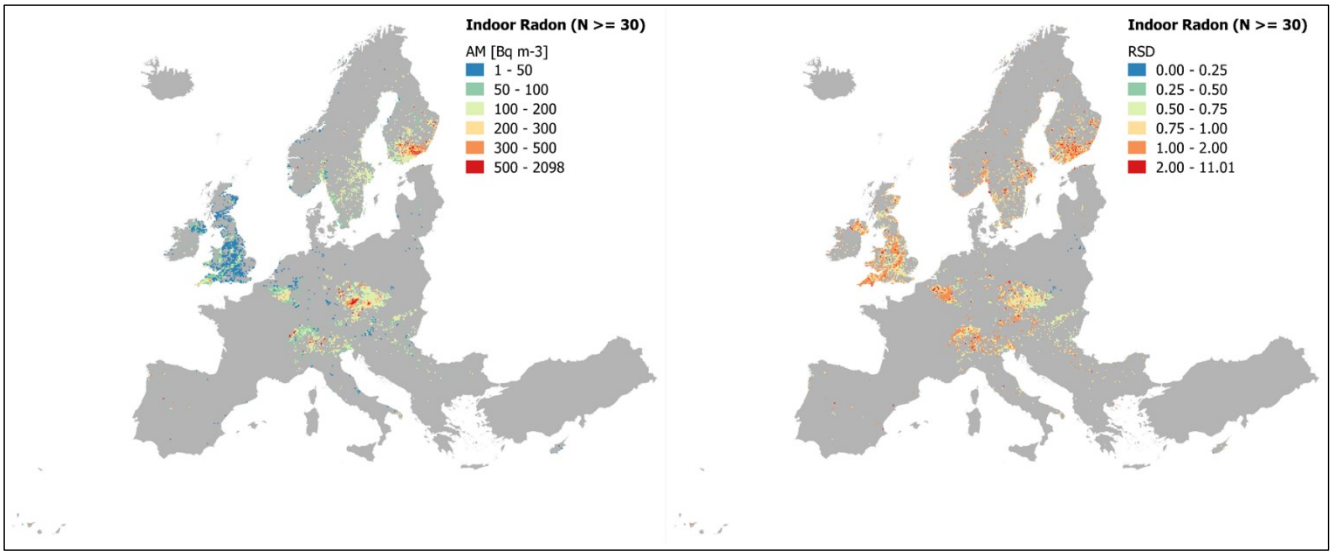
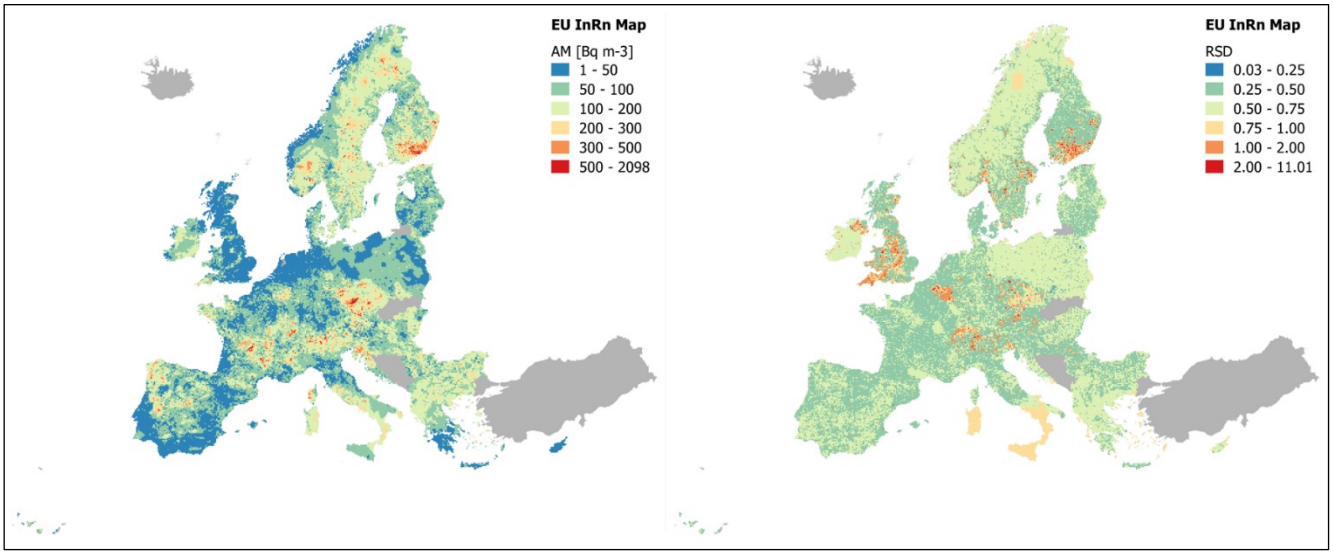


Figure 12: Final Pan-European Indoor Radon Map



5