HIROSAKI UNIVERSITY

Radiation Environment and Medicine 2019 Vol.8, No.2 94–104

REM
Radiat Environ Med

Regular Article

Radon Priority Areas and Radon Extremes – Initial Statistical Considerations

Peter Bossew*

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

Received 11 February 2019; revised 25 May 2019; accepted 13 June 2019

Radon is acknowledged an important health hazard. Indoor radon is believed to be the second cause of lung cancer after smoking. Therefore, indoor Rn has increasingly been subject to regulation for the last years; in Europe, the EU directive on basic safety standards for protection against ionizing radiation. Among other, it requires delineation of radon priority areas, i.e. areas in which action related to prevention and remediation of high indoor radon concentrations should be taken with priority. Whatever the exact definition of these areas, also in those not declared priority areas, high radon concentrations can occur, if with lower frequency. Arguing that also scattered instances of high radon, which are too scarce to justify declaring an area radon priority, deserve mitigation attention, a particular label might be given to these areas, indicating the presence of extremes within an otherwise non-priority area. First statistical considerations about radon extremes and anomalies are presented in this paper.

Key words: Radon priority area, anomaly, extreme

1. Introduction

As radon (Rn) is acknowledged an important health hazard[1)], indoor Rn has increasingly been subject to regulation for the last years. In Europe, the relevant document is the EURATOM directive on basic safety standards for protection against ionizing radiation (BSS)[2)]. Among other, it requires delineation of Rn priority areas (RPAs), i.e. areas in which action related to prevention and remediation of high indoor Rn concentrations should be taken with priority.

The BSS definition states that an RPA is an area where it is expected that in a significant number of houses the long-term mean Rn concentration exceeds the national reference level. For practical implementation, this vague definition has to be translated into an operable one. It is based on a Rn measure, for example the mean over an area or a geographical unit (grid cell, municipality etc.) or the probability that within the unit indoor Rn exceeds a reference level.

Once defined, the RPAs have to be estimated from data, either directly from indoor Rn or covariates such as uranium content in topsoil. Effectively, the spatial domain (a country) is classified into two subsets, one consisting of units labelled RPA, the other one of units labelled non-RPA. (Also the partition into several classes of "priorityness" can be chosen.) Apart from classification uncertainty which has to be expected since the respective areas result from a statistical estimation process, it must be expected that some individual houses do not conform to the RPA definition. For example, a house located in an

*Peter Bossew : German Federal Office for Radiation Protection, Köpenicker Allee 120-130, Berlin, Germany
E-mail: pbossew@bfs.de

area labelled non-RPA, or a cluster or sub-area within the non-RPA, can still have Rn concentration exceeding the reference level. The obvious reason is the high spatial variability of Rn concentrations, resulting in a possibly long "right tail" of the frequency distribution of Rn concentrations.

The physical reason for such phenomenon may be the presence of geographically "small" features which generate high Rn concentrations, well within an otherwise low-radon area. Such features can be tectonic faults or local (secondary) uranium mineralization or highly permeable or karst rock formations. Also anthropogenic reasons can lead to locally high radon burden, due to mineral processing residues or mining activity. Being small in extension, compared to the areas where they are not present, these features contribute little to the mean, but may still pose a radiation problem for that small area. The problem occurs if the RPA definition relies on the mean (or another central measure) of the Rn distribution only, while occurrence of extremes is measured by other statistics such as high quantiles or dispersion measures.

One may therefore think on integrating such additional measures into the criterion which defines RPA or non-RPA, i.e. respecting "small" or "minor" phenomena although they contribute little to the overall picture, dominated by the "background".

In this paper, the matter of "anomaly" and statistical procedures of anomaly detection will be discussed. As an illustrative example, the dataset of the European Indoor Radon Map (EIRM)[3] will be used. In this initial stage of discussing the idea of labelling the occurrence of anomalies, no proposal for practical implementation will be given. Physical causes will not be addressed in this paper and examples will not be related to possible physical causes, i.e. not be physically interpreted. Of course, this must be done in a further stage of development.

2. Statistical concepts

2.1. Anomalies, extremes, outliers

Qualitatively speaking, an anomaly within a metric dataset is an instance which is significantly different from its neighbours. (This is a dataset in which the data have locations, such as positions in time series or geographical locations.) A metric is required to define what a neighbour or a neighbourhood is; in spatial settings, this is usually (but not necessarily) the Euclidean distance. The concept of anomaly cannot be separated from the one of background (BG), represented by the "normal" neighbourhood of the anomaly. If one succeeds to model the BG, an anomaly may be defined as a statistically significant residual from the model.

Outliers are values which seem not to belong to a population. Evidently, an assumption about the population is required to decide this. Reasons of outliers can be: observation error or uncertainty; an accidentally isolated extreme of the BG population; an instance which belongs to a different population. Importantly, a multivariate outlier is not necessarily an outlier of any individual univariate (marginal) distributions involved.

An *extreme* is simply the highest or lowest value of a set. It does not say anything about its nature.

The term *hot spot* seems to be mostly used for points, or cluster of points, or small regions, where the variable takes anomalously high values. Reasons can be a region in which the background process takes high levels or a region which is the domain of a separate process.

Anomaly and hot spot: mostly seem to denote "true" effects, i.e. not related to observation; "outlier" seems to be neutral in this respect, i.e. can also denote observation effects. A summary on outlier detection and problems involved can be found in[4].

2.2. Estimating the background from data

Background estimation consists in fitting a model to the observations, possibly including previous knowledge in Bayesian reasoning. The problem is, how to distinguish between data which belong to the BG, and which do not. If no criterion is given for labelling certain values as not belonging to the BG, using all values may lead to distortion of the estimated BG away from the "true" one by anomalies. So-called masking and swamping effects may be the consequence. The following definitions are adapted from[4].

Masking effect: One "strong" anomaly "masks" a second, "weaker" one, if the latter emerges as anomaly only after removal of the first one from the dataset. This happens if the anomalies "draw" or "distort" the empirical distribution, as estimates of the anticipated BG distribution (in particular mean and standard deviation) such that the second anomaly does not appear as one, although it is.

Swamping effect: An anomaly may distort the empirical distribution in such way that another observation appears as anomaly although it is not. Only after removal of the anomaly it is recognized as a normal observation.

Robust BG estimation includes leave-one-out procedures which may let identify estimates distorted or "contaminated" by isolated anomalies. Also iterative procedures may mitigate the problem, by excluding values identified as anomalies in the subsequent run, until the result has become stable. If in a spatial setting, the BG is established by interpolation, anomalies can contaminate the estimated covariance or variogram, which can distort the local expectation.

In radon science, lognormal (LN) distributions have been found to be rather universal univariate BG models, e.g.[5]. High dispersion, measured e.g. by the geometrical

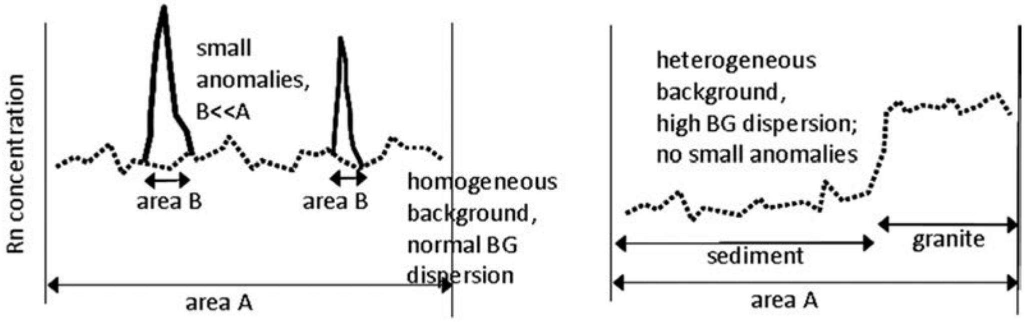

Fig. 1. Two types of anomalous areas. Left: due to presence of small anomalies; right: due to background gradient.

standard deviation (GSD), may point to contamination of the BG by anomalies. If in an area, dispersion is higher than "usual", i.e. than in comparable areas, this raises suspicion that it does not represent the BG process alone, but that an anomaly process is present in that area which inflates dispersion. Also a BG process not stationary within the area (i.e., has a trend) would lead to this effect.

The parameters of the LN distribution are μ and σ, which can be estimated in two ways; AM and SD denote arithmetical mean and standard deviation. $GM=\exp(\mu)$ and $GSD=\exp(\sigma)$ are geometrical mean and geometrical standard deviation, respectively.

(1) $\mu'=AML=AM(\ln Z)$; $\sigma'=SDL=SD(\ln Z)$

(2) Establish the normal QQ plot with empirical quantiles of ln z and determine μ' and σ' by linear regression. This also allows checking whether the LN hypothesis is approximately valid. (Here and in the following the prime denotes the estimate.)

2.3. The chance to sample a rare anomalous value in an area

Assume an area of size A within which a sub-area B is anomalous. It therefore occupies a fraction $p=B/A$. The chance that one random draw (i.e. one physical sample) from A would be taken from B, is thus p.

The chance that of n random draws *none* is taken from B, equals

$$prob(k=0|p,n)=Bin(0|p,n)=(1-p)^n,$$

in binomial logic, where n is the number of "trials" (i.e. sample size) and k, the number of "successes" (i.e., finding B). Hence, the probability, that at least one draw out of n comes from B, equals

$$q=prob(k\geq 1|p,n)=1-(1-p)^n.$$

From this, one can calculate the minimum number of random draws (i.e., minimum sample size) necessary to find B by random sampling with given probability q. By rearranging,

$$n_{min}=\log(1-q) / \log(1-p).$$

For example, let B be one tenth of A, $p=B/A=0.1$. For having the chance of $q=50\%$ to find B, one has to take at least 7 random draws (i.e. a sample of size 7). For a chance $q=95\%$, one needs $n_{min}=28$. If B is smaller, say one hundredth of A, $p=0.01$, for chances $q=50\%$ and 95%, 11 and as much as 298 random draws are necessary, respectively.

The simple calculation shows that finding small sized anomalies in an area by random sampling requires very dense sampling which is usually little realistic. Reversely, if one recognizes an area as anomalous – the following sections are devoted to this task – , one may ask whether the presence of anomalous values that represent a small anomalous sub-area inside the area is responsible, or rather another property of that area.

A typical property of an area which may lead to labelling it as anomalous, without containing anomalous values in the sense that they represent a small-scale locally separate process (B<<A or p<<1 in the above example), is the following. Imagine an area whose geological base is granite at one side and some sediment on the other. Certainly its geogenic Rn potential (GRP[1]) (and consequently indoor Rn concentration) will not be uniform over the area, but show an increasing trend from sediment to granite. The mean GRP over the area may not be representative for either part. Dispersion will be high, also not representative for neither granitic nor sedimentary geological units alone. Both parts of the area represent the BG, but not the area as a whole, without any local anomalies necessarily present. RPA status may be wrongly identified. Speaking statistically, in such area there will be gradient or trend of the response quantity

[1] The GRP quantifies the availability of geogenic Rn for exhalation from the ground surface and for infiltration into buildings.

(GRP or indoor Rn concentration). the BG process will not be stationary and individual values cannot be regarded as random draws from one population. The situations are schematically shown in Figure 1.

In this paper, we are mainly interested in cells which are anomalous, either because they contain local anomalies, or because they have a gradient inside.

2.4. Measures of spatial anomaly

In this section, several statistics will be presented that may serve for identifying isolated anomalous values, or anomalous areas. Distinguish between identifying objects (individual data or data aggregates such as cells) as anomalous with respect to an area, neighbourhood or domain, and identifying aggregates as containing anomalies. In the following (a) and (b) belong to the first, (c) and (d) to the second type.

A common statistic not discussed here is deviation of data from a fitted BG surface by analysing the residuals.

(a) Fractal criteria

Finding isolated anomalies, or small clusters in a spatial setting may be facilitated by applying a method which does not assume a local distribution or covariance model. It relies on the idea that, with A the area around point, |A| its area and z(A) the mean over A, an anomaly behaves as

$$z(A) \sim |A|^{-\beta}, \ \ \beta = -\lim_{|A|\to 0} \log z(A)/\log|A| \qquad (1)$$

(E.g.)[6] z can relate to the values of individual data, or to aggregates (areas or cells). If the exponent β lies in the critical region of exponents determined from many data, the point may be labelled anomaly. However, the estimated distribution will again be contaminated by the anomalies. The reliability of β estimated from data depends on the precision of the z(A). If the objects to be classified are areas, the precision of z, which is the mean over the area in this case, depends on the number of data within.

(b) Bivariate outliers

Consider regularly spaced data, such as on grid nodes or aggregates of grid cells, like in the European Indoor Radon Map (EIRM, see below).

Build a "lagged" bivariate dataset, $\underline{Z} := (Z, L_h Z)$ with lag operator L_h or $\underline{z} := (z(x), z(x+h))$, For each point, compute the Mahalanobis distance d_M and filter those in a p-critical region, say p=0.99. These would be called anomalous cells.

$$d_M(\underline{z})^2 := (\underline{z}-\underline{m})^T C^{-1} (\underline{z}-\underline{m}) \qquad (2)$$

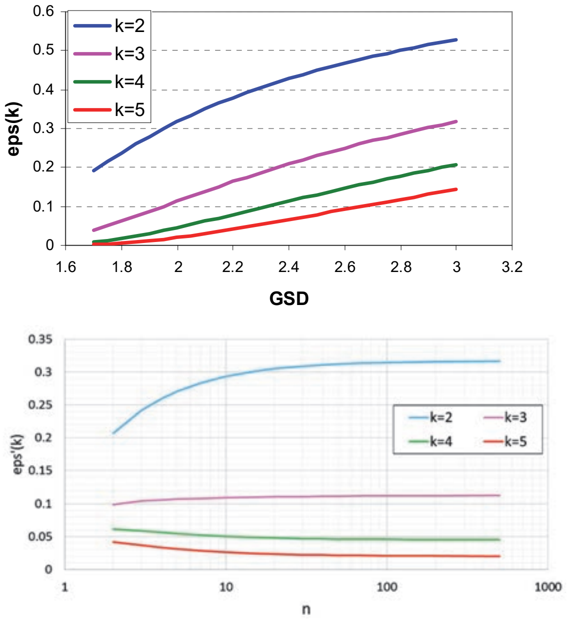

Fig. 2. Upper graph: ε_k in dependence of GSD, for several k; Lower graph: ε'_k for GSD=2.

The vector $\underline{m}$ equals (AM(Z),AM(Z)), C – the covariance matrix. For stationary fields, the off-diagonal element C_{12} equals $C_{12}=C(h)= \sigma^2-\gamma(h)$, σ^2 and $\gamma(h)$ - variance and semivariance for lag h, respectively. Alternatively, it can be estimated from the Pearson correlation coefficient r between Z and $L_h Z$, $\sigma^2\ r=cov=\sigma^2-\gamma(h)$ for stationary fields. If γ is estimated from the entire dataset, the contamination problem occurs again.

For irregularly spaced data, z(x+h) could be defined as values at locations within distance h ± tolerance from x, as usually done for variogram estimation.

For the cell-aggregated data of the EIRM, one would choose lag h=1 grid=10 km and Z=AML per cell. In practice, z(x+h) could be the mean of the nodes or cells surrounding node or cell x, for h=1 grid size.

(c) Univariate cell statistics

High quantiles or exceedance probabilities are useful to estimate the upper tails of a univariate BG model, but do not per se indicate anomalies in the above sense. However, for a contaminated BG model, these extremes may represent anomalies. Q_p denotes the p-quantile of data in an area (i.e. 100·p% of values are below Q_p), P(z) the probability that in that area, the level z is exceeded. For a LN BG, these statistics are defined

$$Q_p=\exp(\Phi^{-1}(p|GM,GSD)) \text{ and} \qquad (3a)$$
$$P(z) := prob(Z>z)=\Phi((\ln GM - \ln z)/\ln GSD), \qquad (4a)$$

respectively, Φ - standard normal, GM and GSD – geometrical mean and standard deviation within the considered area. Estimators are

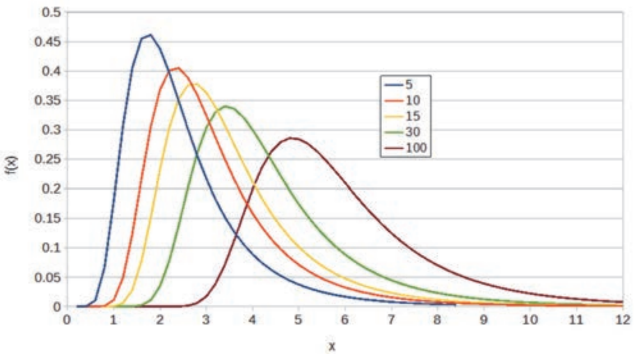

Fig. 3. Probability density functions of the maxima of X~LN(0,ln(2)), for sample size given in the legend.

$$Q_p' = \exp(t_{p;n-1}\, SDL + AML) \text{ and} \tag{3b}$$

$$P'(z) = t_{n-1}(u\ \zeta), \tag{4b}$$

With u:=√(n/(n+1)), ζ:=(ln z-AML)/SDL, n – sample size[7].

P is often used as a criterion for defining RPAs. As discussions over the last years have shown, a common choice in Europe is z=300 Bq/m^3 and P(z)=0.1, i.e. an area is labelled RPA if the probability that indoor Rn concentration exceeds 300 Bq/m^3, exceeds 10%.

A measure that exploits the GSD only, is the relative excess probability, high values of which point to high dispersion,

$$\varepsilon_k := \text{prob}(Z > k\ \text{med})/\text{prob}(Z > \text{med}) \in (0,1] \tag{5a}$$

The denominator equals 0.5. Under LN hypothesis, this is

$$\varepsilon_k = 2[1-\Phi(\ln k/SDL)] \tag{5b}$$

The considered area could be a grid cell within which data are aggregated. ε_k in dependence of GSD for different k is shown in Figure 2. For example, with GSD=2.4 we find that the probability to exceed the median 3 times (k=3, pink graph) equals about 20%, whereas for GSD=1.9 (a typical value for 10 km × 10 km cells), ε_3=about 10%. The sampling distribution of ε_k is not available. By simulation, the lower graph, Figure 2, is found for GSD=2. For k=3, ε_k is almost independent of sample size n. (Method: generate many (10,000) variates $s^2 \sim \sigma^2\ \chi^2_{n-1}/(n-1)$, calculate ε_k according formula (5b) with GSD=exp(√s^2) and take the average of the realizations.)

(d) Extremes

Within an area (e.g. a grid cell), the process of which the data are a sample, is characterized by its univariate distribution F_Z. The maximum of a sample of size n, $z_{max} \equiv y_n := \max\{z_1,...,z_n\}$ is exactly distributed $Y_n \sim F_Z(y)^n$ for statistically independent z_i. Probability density functions of the maxima of Z~LN(0,ln(2)) for different n are shown in Figure 3.

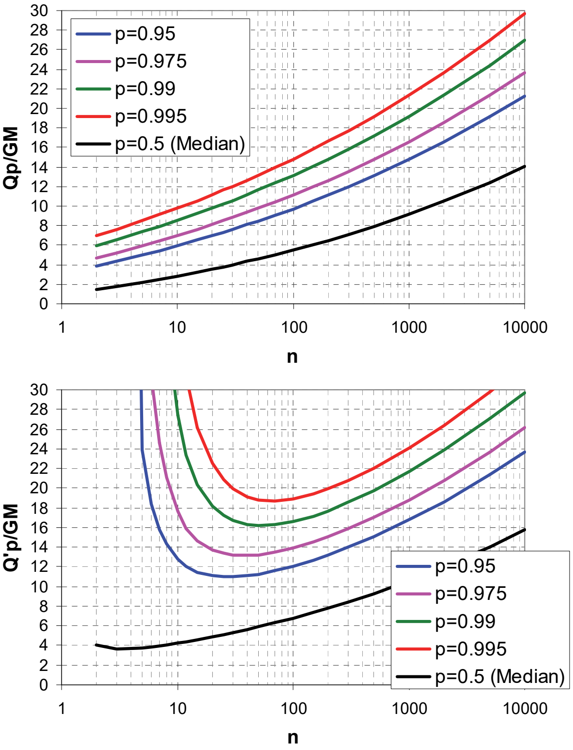

Fig. 4. Upper graph: Expected quantiles of the maximum of a LN sample of size n, relative to the geometrical mean GM. Lower graph: same, for unknown true GM. Both for GSD=2.

From extreme value theory (more specifically the Fisher-Tippett-Gnedenko theorem, which is an analogue to the CLT for maxima) it is known that Y_n is asymptotically distributed ~exp(-exp(-(y_n-μ_n)/σ_n)), the so-called Gumbel or EV type 1 distribution. Its mean EY_n and standard deviation √Var Y_n are μ_n+EM·σ_n and $\sigma_n\pi/6$, respectively; EM – the Euler-Mascheroni constant, 0.57722···; however, μ_n and σ_n are complicated functions of the parameters of F_Z. Convergence is very slow, therefore it is preferable in practice to simulate the distributions functions and approximate them by power functions, which yield good fits for the Z~LN examples. The matter shall not be discussed further here. Literature and textbooks about extreme value theory are abundant. The p-quantile of the random variable Y_n is

$$Q_p(Y_n) = F_Z^{-1}(p^{1/n}) \tag{6}$$

Calculating the expectation EY_n is much more complicated. An observed value $z^\#$ may be called anomaly if it significantly exceeds the predicted extreme, e.g. $z^\# > Q_{95}(Y_n)$. The BG (represented by F_Z) contamination problem occurs also here. Under the assumption of

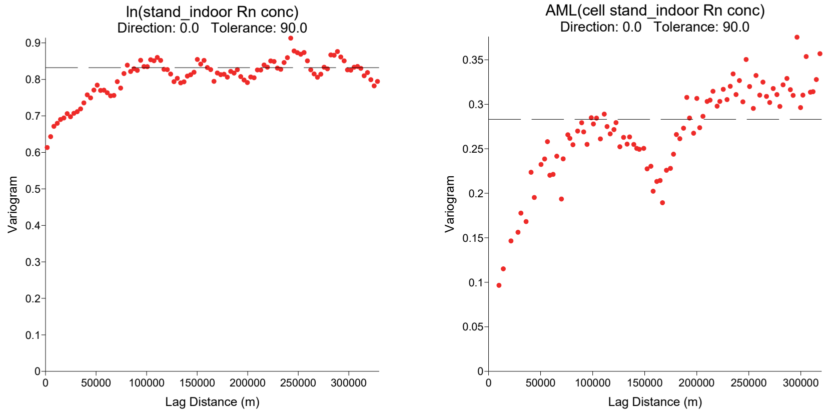

Fig. 5. Empirical variograms: Left graph: ln(indoor Rn concentration); Right graph: Same data, aggregated into 10 km × 10 km cells. Data from Germany. Dashed lines: sample variances.

representativeness of the z in the investigated area, the frequency of $z^{\#}$ may be an indicator of its anomality.
Under LN hypothesis, it can be written

$$\text{SDL } \Phi^{-1}(p^{1/n}|0,1) = Q_p(\ln Y_n)\text{-AML, or}$$
$$\exp(\text{SDL } \Phi^{-1}(p^{1/n}|0,1) = Q_p(Y_n)/GM =: \psi \qquad (7)$$

GM = exp(AML) – the geometrical mean of Z. This function is shown in Figure 4 for several probabilities p in dependence of sample size n, for GSD = 2. For example, if we have n = 20 values, the 99% quantile (green line, Figure 4) of the maximum is about 10 times the GM. This is to be understood as repeating the experiment with n = 20 many times, in 99% of cases the maximum will be below 10 times the GM. If the true GM is not known, Φ^{-1} is replaced by t^{1} with n-1 degrees of freedom, which leads to the lower graph of Figure 4.
As indicators of anomaly one may use

$$z_{max}(\text{observed}) \,/\, Q_{95}(Y_n) =: \psi_1 \qquad (8a)$$
$$\text{prob}(Z > Q_{95}(Y_n)) = \#\{z_i > Q_{95}(Y_n)\}/n =: \psi_2, \qquad (8b)$$

The frequency of exceeding the predicted 95-quantile of the maximum. ψ_2 requires knowing the individual values within an area, ψ_1 only the distribution parameters.
For these quantities, $\psi_1 > 1 \Leftrightarrow \psi_2 > 0$ holds.

2.5. *Autocorrelated fields*

The considerations in (c) and (d) hold for statistically independent data. In most real cases, the data are autocorrelated, $\text{cor}(Z, L_h Z) \neq 0$, i.e. the z_i are not independent but have an autocorrelation or autocovariance structure often quantified by the variogram in spatial settings. As an example, the structure is shown in Figure 5 for German indoor Rn concentration data (long-term means, ground floor of living rooms in buildings with basement): logarithms of individual data (n = 21,953; left graph) and aggregates in 10 km × 10 km cells (n = 510; right graph), with minimal 5 data per cell. We recognize a correlation length of about 100 km. The nugget effects (variogram for lag = 0) equal 72% and 32% of the variances, respectively. This means that autocorrelation of data within 10 km cells (left graph) is present. For h = 10 km it amounts to r = 1-variogram(10000)/variance = 0.16, or r^2 = 2.5%, which is very low. Therefore, the error committed by treating the data as independent is probably low. On the other hand, correlation between neighbouring cells AMLs (right graph) equals 0.68, or r^2 = 46% which is quite high. If regarding cells as samples from a large domain (country), neglecting the autocorrelation structure may therefore lead to considerable error. In the future, this issue must be investigated more profoundly.

Correlated random fields typically have "patches" of high values opposed to uncorrelated fields in which extremes occur randomly. Adler (1981)(Theorem 6.3.1; p.133)[8] has shown that for a 2-dim Gaussian random field (0,σ) the expected density of local maxima above z, for high z, is proportional to $(z/\sigma^2)\exp(-z^2/(2\sigma^2))$. Much literature has been devoted to modelling extremes of autocorrelated spatial random fields.

Autocorrelation of the field is also the reason, why the mean variance with an area depends on the size of the area. For the example underlying Figure 5, total sample variance equals $\sigma^2 = 0.832$ (left graph), corresponding GSD = $\exp\sqrt{\sigma^2} = 2.49$. The empirical mean GSD in 10 km × 10 km cells equals 2.13. Theoretically (calculation not shown here), one finds 2.22 by evaluating the variogram model found by fitting the empirical variogram, Figure 5 (left graph).

If accounting for autocorrelations, calculations become much more complicated, but this is beyond the scope of

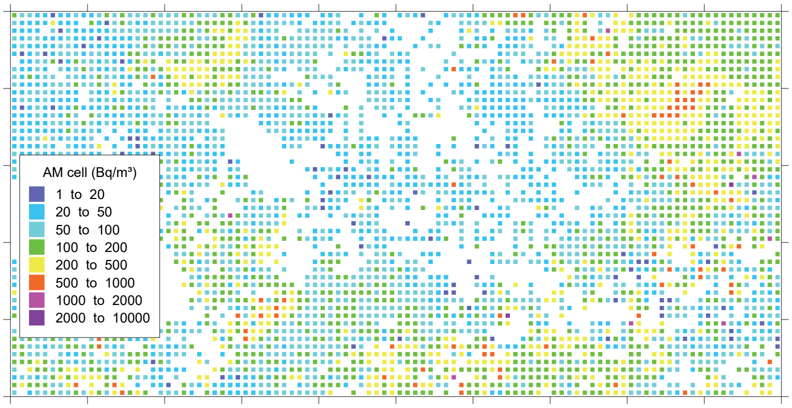

Fig. 6. Arithmetic mean indoor Rn concentration in 10 km × 10 km cells. 1 axis tick = 100 km. Empty cells: no data.

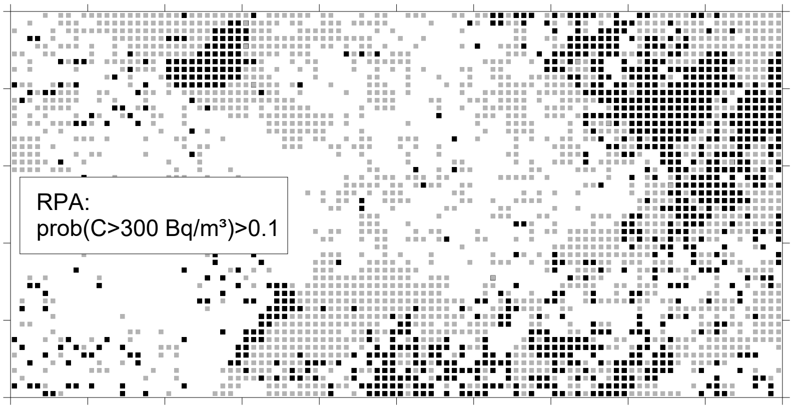

Fig. 7. Black cells: Defined as radon priority areas according the criterion given in the legend.

the simple considerations presented in this article.

Mapping of extremes

Geostatistical treatment of extremes of a process or their occurrence probability relies on modelling joint distributions of maxima. Since modelling of extremal events has obvious practical importance in may disciplines of natural and social science, this matter has been studied extensively for a few decades, for example[9, 10] Mathematical treatment is challenging and shall not be discussed in this introductory paper.

Moreover, we are less interested in the dependence structure of the extremes of the BG field, but in the anomalous processes sitting on top of the BG. Their spatial dependence is the one of their physical causes. Therefore, we think that studying the latter is a prerequisite of understanding the former.

At this point, anomalies identified through the statistics described above will just be labelled and plotted in post-map style. Of course, this prevents inference on anomalies in non-sampled areas (empty cells), which would require interpolation based on spatial extreme value statistics.

2.6. Classification

For each value of a spatial dataset, one can calculate statistics such as the ones presented in section 2.3. As a next step, the data have to be classified after some criterion, whether they shall be called anomalous or not. The straightforward approach consists in comparing a value with the critical limits of the null hypothesis (H_0: "no anomaly"). The distribution of the null hypothesis has to be inferred from the data in the considered domain. This leads again to the problem of contaminated BG, because the distribution is estimated prior to anomaly detection, including possible anomalies. An iterative procedure may be envisaged to resolve this: in step (n +1), the null distribution is estimated from the data excluding the ones identified as anomalous in step (n). However, this may involve considerable computational effort.

Three approaches for defining the null hypothesis may be considered for labelling an area i.e. an aggregate of individual data as anomalous. In the following, call summarily ϑ the statistical indicators (β, d_M, ε_k, $\psi_{1,2}$) or any other.

1. Calculate ϑ for each area or cell of the domain.

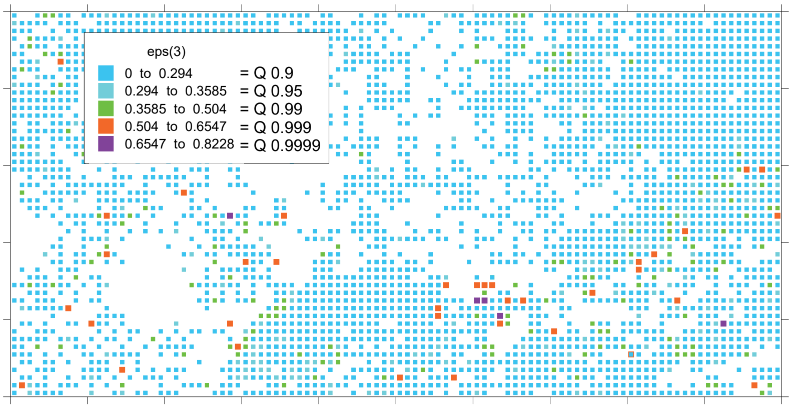

Fig. 8. Exceedance statistic ε_3. Colour codes classified according percentiles of the ε_3.

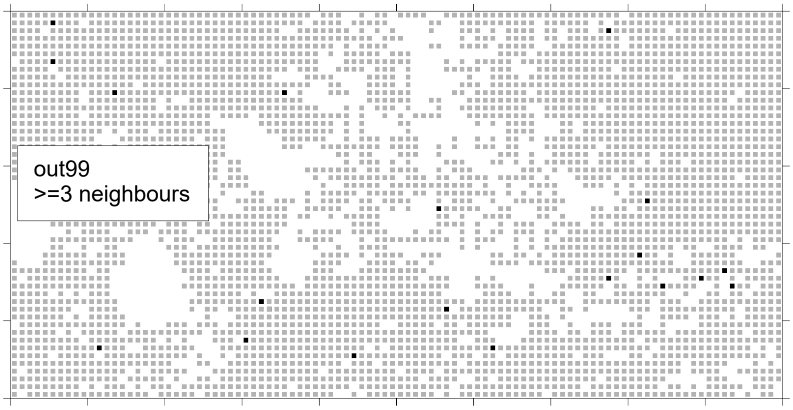

Fig. 9. Bivariate outliers according Mahalanobis statistic. The 99% quantile has been chosen for labelling cells purple. Only cells with at least 3 neighbours included in the calculation.

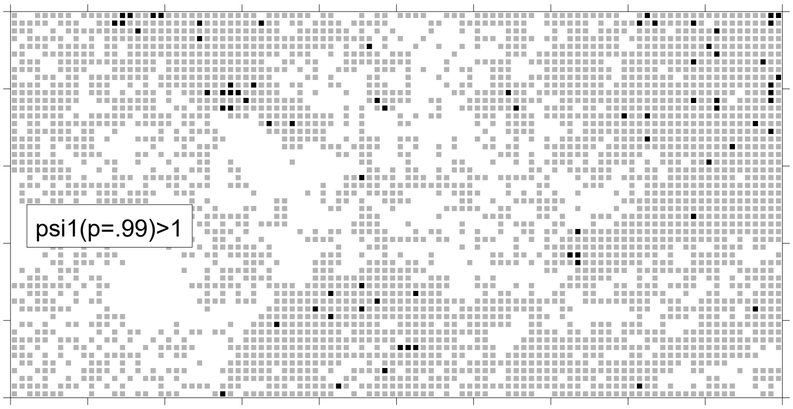

Fig. 10. Extreme value statistic ψ_1 with $P=0.99$.

Build the distribution of ϑ in the domain and calculate the critical limits, i.e. 1-$\alpha/2$ quantiles for a given α. Decide whether $\vartheta \in$ critical region. If this is the case, the cell is called anomalous. The problem is that empirical ϑ derived from different sample sizes n (number of data within the area) are mixed together. This affects ε_k and ψ, but not β and d_M.

2. A variant consists in calculating the distribution of GSD in the domain and deriving the ones of ε_k and $\psi = Q_p(Y_n)/GM$, which depend only on GSD, followed by calculating the critical region. Both also depend on sample size n, but for ε_k no sampling

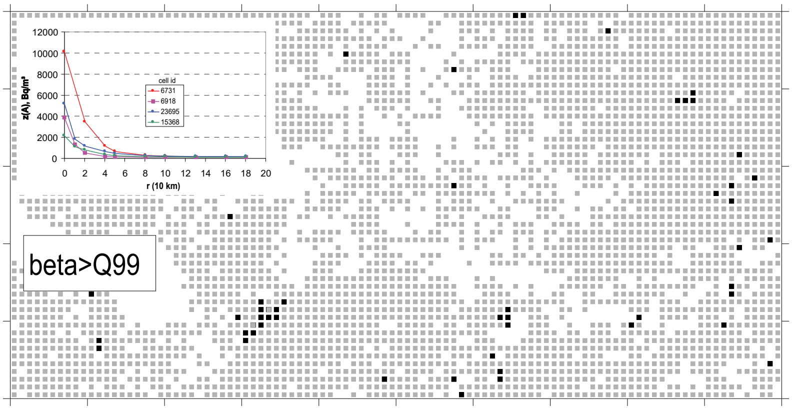

Fig. 11. Fractal β exponent; black cells: $\beta > Q99 = 0.194$. Inset: Dependence z(A) against r(A) for four data points.

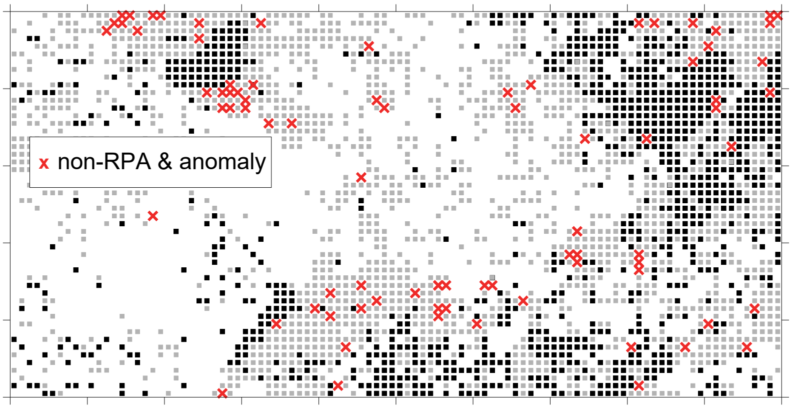

Fig.12. RPA map as Figure 7 with anomalies marked (x) which are located in non-RPA.

distribution seems available.

3. Without reference to the entire set: A cell is called anomalous, if $\psi_1>1$ or equivalently, $\psi_2>0$.

In the example below, β, d_M and ε_3 (the latter despite its n–dependence, at this stage) will be treated according (1), while ψ_1 according (3). Variant (2) has not been evaluated yet.

Remember that the initial question was to identify cells which are not being labelled RPA, but should still considered anomalous because high values can be expected within, be it due to small anomalies or heterogeneous BG. Resorting to the RPA definition: area is RPA if $P(300)>0.1$, we seek to identify the cells with $P(300) < 0.1$ & $\vartheta \in$ crit.

3. Example: the European indoor radon map

The European Indoor Radon Map (EIRM)[3] consists of 10 km × 10 km cells in which indoor Rn data (long-term measurements, ground floor living rooms) have been aggregated by institutions of participating countries. These cells are the areas discussed before. So far (late 2018), about 1.1 million individual data from 34 countries are included, which however remain with the participants for reasons of data protection. Only the cell statistics (n, AM, SD, AML, SDL, median, min, max) are communicated to the Joint Research Centre (JRC) of the European Commission, which collects the data and generates the map. These are the empirical data of the example in the following. In the figures, a 1000 km × 500 km rectangle from Central Europe has been selected. Figure 6 shows the original AM values from the database.

For Figure 7, the probability to exceed 300 Bq/m^3 has been calculated according formula (4b) and the cells classified whether P>10% (black) or not (gray). Since the calculation requires that the cell contains at least 2 data, the map contains more void areas (cells with n=1) than Figure 6.

In Figure 8, the ε_3 are shown, calculated according formula (5b). Quantiles belonging to the 90%, 95%, 99%, 99.9% and 99.99% percentiles of the entire European dataset have been used for classification.

For identifying bivariate outliers according Mahalanobis distance, formula (2), the following practical procedure has been chosen. For each cell, the left and the upper neighbour was identified. A bivariate dataset $\{(\underline{x}_1, \underline{x}_1(\text{left})), (\underline{x}_1, \underline{x}_1(\text{up})), (\underline{x}_2, \underline{x}_2(\text{left})), (\underline{x}_2, \underline{x}_2(\text{up})),\cdots\}$ was created, i.e. twice as large as the original one. From this, the covariance C(h = 1 cell) was determined by calculating the Pearson correlation coefficient, r = 0.622. (An alternative would be calculating the variogram γ(10 km).) Then, d_M was calculated for all cells. These were filtered according (a) retaining those cells which have $d_M > Q_{99}$, but whose neighbours have $d_M \leq Q99$, and (b), which have at least 3 neighbours. Q_{99} is the 99% quantile of all d_M. Neighbours are defined according the 4-neighbour rule, i.e., only cells which share a side are considered as neighbours. (a) serves to find isolated anomalies. The result is shown in Figure 9.

Figure 10 shows the extreme value statistic ψ_1 for $P = 0.99$, formula (8a). The black points are the ones in which the recorded maximum exceeds the theoretical one with 99% probability. The estimate may be strongly biased for cells with low occupancy (n), as the ψ_1 were calculated according Figure 4 (upper). Using the probably more correct version substituting the normal be the t-distribution in formula (7) would factually exclude cells with n < 30 from consideration, see Figure 4 (lower). This issue has to be investigated further.

A map of the fractal exponent β (formula 1) is shown in Figure 11. Practically, it has been calculated by laying circles around each data point, calculating the mean within each circle and regressing the means against circle radius, The function z(A) against radius is shown for four data points in the inset. The distribution of the β (from all European data) was established and the 99% quantile (0.194) used as filter.

Finally, in Figure 12, results are summarized: Anomalies identified in the previous maps, which are located in non-RPA, are marked with crosses. For ε_3 (Figure 8), the 99% quantile was used as filter.

4. Possible implementation

In practice, one would certainly check geological, tectonic, geochemical and hydrogeological maps for presumptive physical causes of Rn anomalies, and screen data based on that knowledge. (Concerning tectonics, e.g. [11]; karst, e.g. [12]) Then one would check whether the physical causes are supported by data. However, in many cases, data density may not be sufficient to adequately reflect physical reality in the data realm (section 3.1).

In parallel, statistical investigation would be carried out along the lines indicated above, or using methods still to be developed. Areas which are recognized anomalous by statistical indicators due to presence of anomalies or heterogeneity of geological controls could be given a particular label, in addition to the label RPA / non-RPA. Of course, one would try to find *physical* causes of the identified *statistical* anomalies. Sampling density in an area can be too low to find small anomalies. Therefore, missing statistical evidence does not imply absence of anomalies; the possibility of committing a second kind error (false non-detection of an effect which is there in reality) is always present due to data sparseness. Reversely, statistical evidence is a good indication to pay attention to an area. However, the possibility of a "spurious" anomaly must not be discarded, due to data or reporting errors.

5. Conclusions

Also areas not labelled RPA can deserve attention out of radioprotection concerns. They may contain local anomalies of high Rn concentration, the RPA status may be assigned wrongly based on non-representative data or questionably if there is a BG gradient. In particular, high values may be too scarce to justify declaring the area as RPA, or a fraction which would by itself be RPA, has not enough weight against the remaining non-RPA fraction.

In this article, first statistical considerations on detecting such anomalous areas were presented. How this could be implemented practically has to be left for future discussions, as it would certainly also pose an administrative challenge. Analytical tools will have to be developed further and not least, uncertainty and classification errors addressed.

Among task to be tackled in future work are:

1. The mathematically difficult issue of statistics of autocorrelated fields must be studied further.
2. "Decontamination" of BG estimates by iterative or other procedures should be attempted.
3. Further, more powerful statistics for indication of anomalies may be found.
4. Sampling distributions of some of the statistics addressed here must be investigated.
5. Establish better suited null hypotheses to classify statistics as anomalous, in particular for ε_k.

Importantly, the matter of physical validation, i.e. investigation the association between statistical and physical phenomena will require thorough case studies.

Disclosure and acknowledgement

This work is supported by the European Metrology Programme for Innovation and Research (EMPIR), JRP-Contract 16ENV10 MetroRADON (www.euramet.org). The EMPIR initiative is co-funded by the European Union's Horizon 2020 research and innovation programme and the EMPIR Participating States.

The author declares that he has no conflict of interest.

References

1. World Health Organization 2009. WHO handbook on indoor radon - a public health perspective. http://www.who.int/ionizing_radiation/env/9789241547673/en/

2. Council Directive 2013/59/Euratom of 5 December 2013 laying down basic safety standards for protection against the dangers arising from exposure to ionising radiation etc., http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2014:013:FULL&from=EN.

3. Cinelli G, Tollefsen T, Bossew P, Gruber V, Bogucarskis K, De Felice L, De Cort M. Digital version of the European Atlas of natural radiation, J Environ Radioact. 2019;196:240–52; doi.org/10.1016/j.jenvrad.2018.02.008. Web version of the map: https://remon.jrc.ec.europa.eu/About/Atlas-of-Natural-Radiation

4. Ben-Gal I 2005. Outlier detection, In: Maimon O and Rockach L editors. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers. Available from: http://www.eng.tau.ac.il/~bengal/outlier.pdf

5. Bossew P. Radon: Exploring the Log-normal Mystery. J. Environmental Radioactivity 2010;101:10:826–834. http://dx.doi.org/10.1016/j.jenvrad.2010.05.005

6. Cheng Q. Multifractality and spatial statistics. Computers & Geosciences 1999;25(9):949–61. DOI: 10.1016/S0098-3004(99)00060-6

7. Bossew P, Tollefsen T, Cinelli G, Gruber V and De Cort M. Status of the European Atlas of Natural Radiation. Radiat Protect Dosim. 2015;167:1–3:29–36;doi:10.1093/rpd/ncv216

8. Adler R. The geometry of random fields. New York: John Wiley and Sons;1981 ISBN 0 471 27844 0.

9. Schlather M, Tawn JA. A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika. 2003;90:1:139–56; https://doi.org/10.1093/biomet/90.1.139

10. Cooley D, Naveau P, Poncet P. Variograms for spatial max-stable random fields. In: Bertail P., Soulier P., Doukhan P. editors: Dependence in Probability and Statistics. Lecture Notes in Statistics, vol 187; 2006. Springer, New York, NY; DOI https://doi.org/10.1007/0-387-36062-X_17

11. Ciotoli GC *et al*. Geogenic radon as geophysical / geochemical tracer of active faults. In Lujanienė G and Povinec PP editors. Radionuclides as Tracers of Environmental Processes. Proc. 4th Intl. Conference on Environmental Radioactivity (ENVIRA 2017), 29 May–2 June 2017, Vilnius, Lithuania; SRI Center for Physical Sciences and Technology, Vilnius 2017; ISBN 978-609-95511-4-2.

12. Dehandschutter B. The role of karst in identifying radon priority areas in Belgium. 14th Intl. Workshop on the Geological Aspects of Radon Risk Mapping- GARRM, Prague, September 2018.