

Radon priority areas as random objects

Peter Bossew

German Federal Office for Radiation Protection (BfS), Berlin

IAMG2018

2-8 September, Olomouc, Czech Republic



v. 4.9.18

Content

- Rationale:
 - Importance of indoor radon
 - Legal background
- Definitions of Radon priority areas
- Classification and classification uncertainty
- Concept of random object
- Sources of uncertainties
- Examples

Indoor radon - essentials

Indoor radon – most important contribution to dose!

Second most important cause of lung cancer after smoking!

In Europe estimated about 62,000 lung cancer fatalities per year attributed to Rn.

(Gaskin et al., Envir. Health Perspectives 125, 5 (2018); incl. RU, TR; missing: BiH, LV, MD, MK, MT, RS, UA)

(figure appears a bit overestimated to me)

Sources of indoor Rn:

1. Geogenic Rn (most important in most cases)

2. Building materials

3. Tap water, natural gas

Concentrations of indoor Rn controlled by

Geogenic factors:

Geology, soil type, U concentration in topsoil, permeability, granulometry,...

Anthropogenic factors:

Construction type (tightness of structures in contact with the ground),
life or usage patterns (ventilation)

Very high local and temporal variability → makes prediction very difficult.

Legal background



Basic Safety Standards (BSS)

Council Directive 2013/59/Euratom laying down basic safety standards for protection against the dangers arising from exposure to ionizing radiation

<http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2014:013:TOC> (OJ L, 17.01.2014)

Art. 103,3; RPA:

“Member States shall identify areas where the radon concentration (as an annual average) in a significant number of buildings is expected to exceed the relevant national reference level.”

Conceptual definition, which has to be translated into an **operable** definition.

Art. 54, 74, annex XVIII; Radon Action Plan:

In areas according Art.103,3: Buildings with public access and workplaces must be measured and if above RL, remediated. New buildings: particular Rn prevention. Strategy to reduce Rn in dwellings.



Reference level (RL): must be $\leq 300 \text{ Bq/m}^3$ (BSS Art 54,1 & 74,1). Most countries chose 300, Ireland and others: 200

These areas are called Radon Priority Areas (RPA), to indicate priority in taking action.
Formerly also “Radon Prone Areas”

RPA definitions, 1

Some examples of operable RPA definitions, based on different Rn measures:

- An area B (grid cell, municipality...), in which the mean population-weighted indoor concentration C exceeds the reference level (RL); $AM_B(C) > RL$; measure = AM_B
- same, but indoor concentration in *dwellings on ground floor*
- An area B, in which the probability that C exceeds the RL, is greater than p (typically 10%); $prob_B(C > RL) > p$; measure = $prob_B$
- The areas B which represent the upper 10% of $AM_B(C)$; measure = percentile
- An area, in which the collective exposure (e.g., $AM_B(C) \times \text{population}$) is among the upper 10%

Important:

There is no “natural” definition of RPA! Therefore, also no “true” RPA!

RPAs always depend on definition and to some extent, on estimation method.

This is partly a political decision, partly a pragmatic one (i.e., availability of data).



Consequence:

RPAs may, in general, not be comparable across borders. This may create communication and credibility problems. Discussing this and proposing solutions is another subject of the Metro Radon project. One way may be a map of the Rn hazard index (RHI – currently under development) as “universal” (but still to an extent deliberate) measure of Rn “priorityness”.

RPA definitions, 2

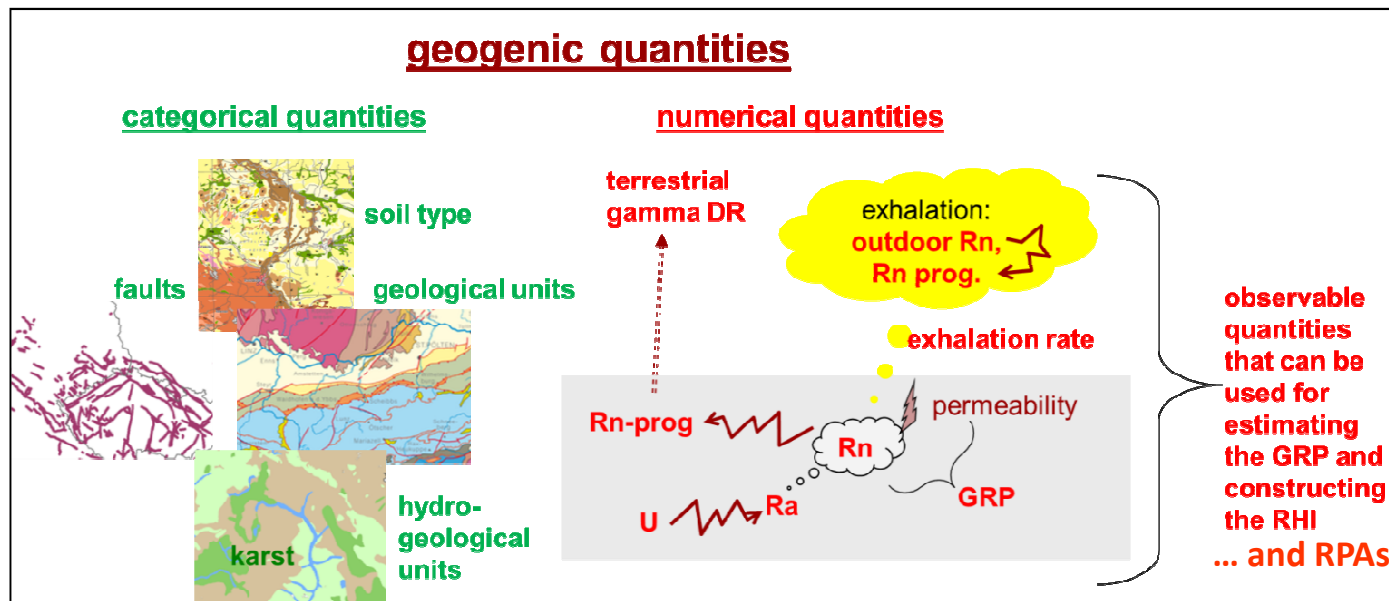
Multinomial:

Instead of 2 classes (RPA / non-RPA), several classes of “Rn-priorityness”; approach chosen by some countries.

Multivariate:

Although the BSS definition relies on indoor Rn concentration, one may chose to base estimation on other Rn-related variables instead or additionally.

Examples: geogenic Rn potential, U concentration in the ground, terrestrial gamma dose rate, geological unit, tectonic features etc.



GRP = geogenic Rn potential
RHI = geogenic Rn hazard index

RPA estimation – a classification problem

Decision about whether a geographical unit shall be labelled RPA or not (in the case of multinomial definition, which grade of “Rn priorityness” it should be assigned): a **classification problem**.

If estimated from secondary quantities: conditional and cross classification.

Existing solutions are pragmatic in the sense that they have to rely on available data and on external “political” parameters such as reference levels and tolerable uncertainty.

Classification uncertainty

Whereas the uncertainties of the estimated actual levels of the Rn measure are commonly quantified by confidence intervals, the ones of classes are given by **first and second kind classification error** probabilities.

The complication consists in the large spatial variability of indoor Rn, also in small scale (~high nugget). Whether estimated from indoor Rn directly or from secondary quantities, this may lead to large classification uncertainty.






In particular:

for geographical units whose Rn measure is close to the class limits.

“Random object”

- A quantity which is an outcome of a statistical estimation procedure is a random quantity.
- An area is a spatial object.
- The label of a grid cell -- 0/1, or a class 1...N --, is a random variable as estimation result.
- \Rightarrow the object (grid cell, or union of contiguous cells) to which the label is attached, is a “random object”.
- Understood as realizations of a stochastic process, all realizations of RPA maps look differently.
- Task: define and quantify uncertainty of such object, i.e. whether it is present or not (= cell labelled 1 or 0), or to which class it belongs with which probability.
- **Relevance: Whether an area is labelled RPA or not could make huge economical difference!**

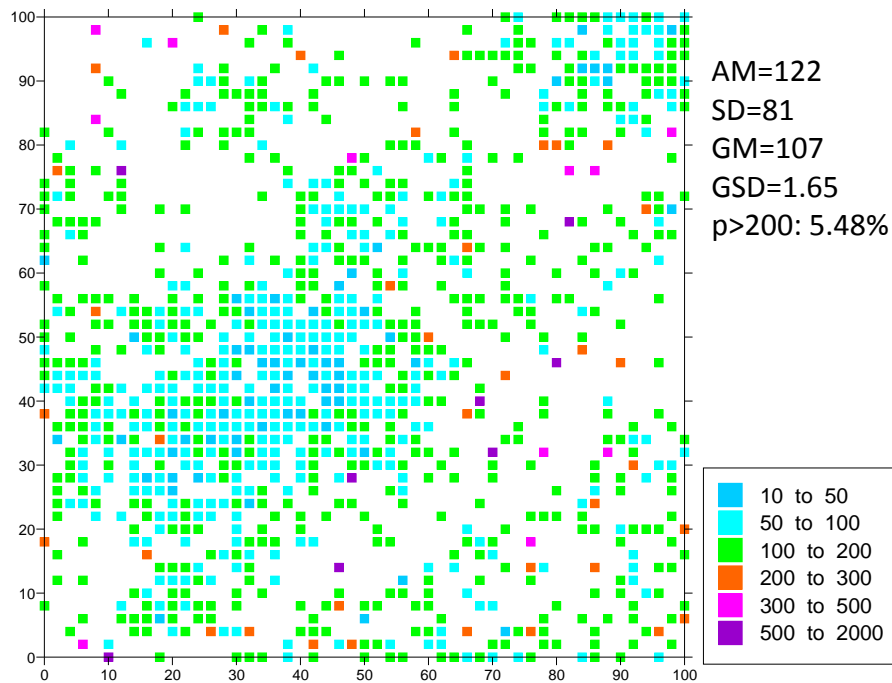
Sources of variability & uncertainty

- Intrinsic data uncertainty – Data as observations
 - Radiometric data: counting uncertainty;
 - “Semantic” data uncertainty: possibly erroneous attribute of measurement, e.g. room recorded to be in ground but in reality first floor
- Data as sample of a population
 - Sample size   **Ex. I**
 - (in particular tricky for true finite populations!)
 - Uncertainty about representativeness of sample used for inference.
- Model uncertainty
 - “minimal model”: sample statistics, e.g. mean or exceedance probability from raw data → sampling statistic → uncertainty (SD of the mean, bias of SD).
Due to the complicated structure of natural controls, Rn is variable on all scales.
 - Data uncertainty inflates dispersion → bias e.g. of exceedance probability  **Ex. II**
 - Structural uncertainty: choice of model
 - Estimation uncertainty, e.g. of regression parameters   **Ex. III**
 - ⇒ Prediction uncertainty

 ... Examples discussed here

Example I: Sample size effect: A numerical experiment, 1

The municipality Gigrizpatschen (AT),
Rn concentrations in all N=1004 houses.
Quite realistic!

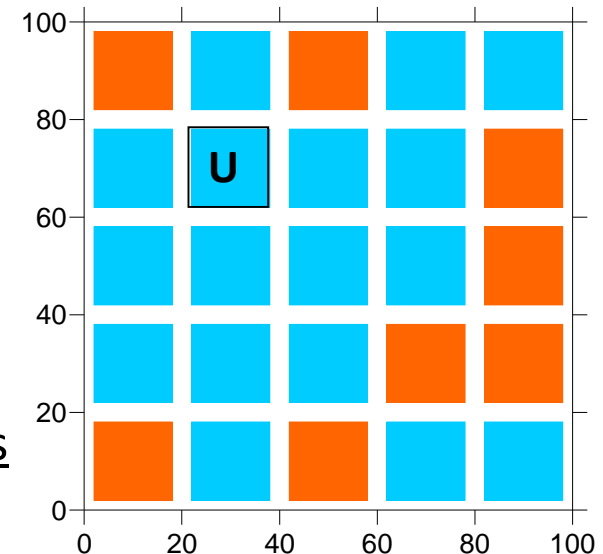


In a survey, we cannot measure all houses, but a number n , selected randomly. I.e., a representative sample in the best case.

Declare an area (U) RPA, if in U:

$$\text{prob}_U(z > 200) > 0.1$$

Areas U: quadratic fractions of the municipality.



True RPA status

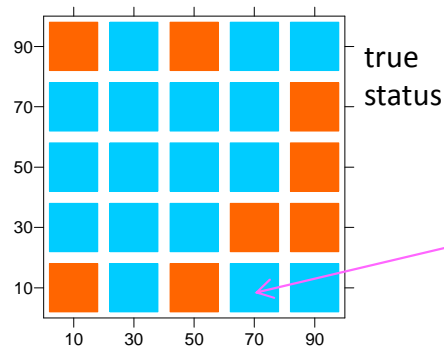
orange = RPA

blue = non-RPA

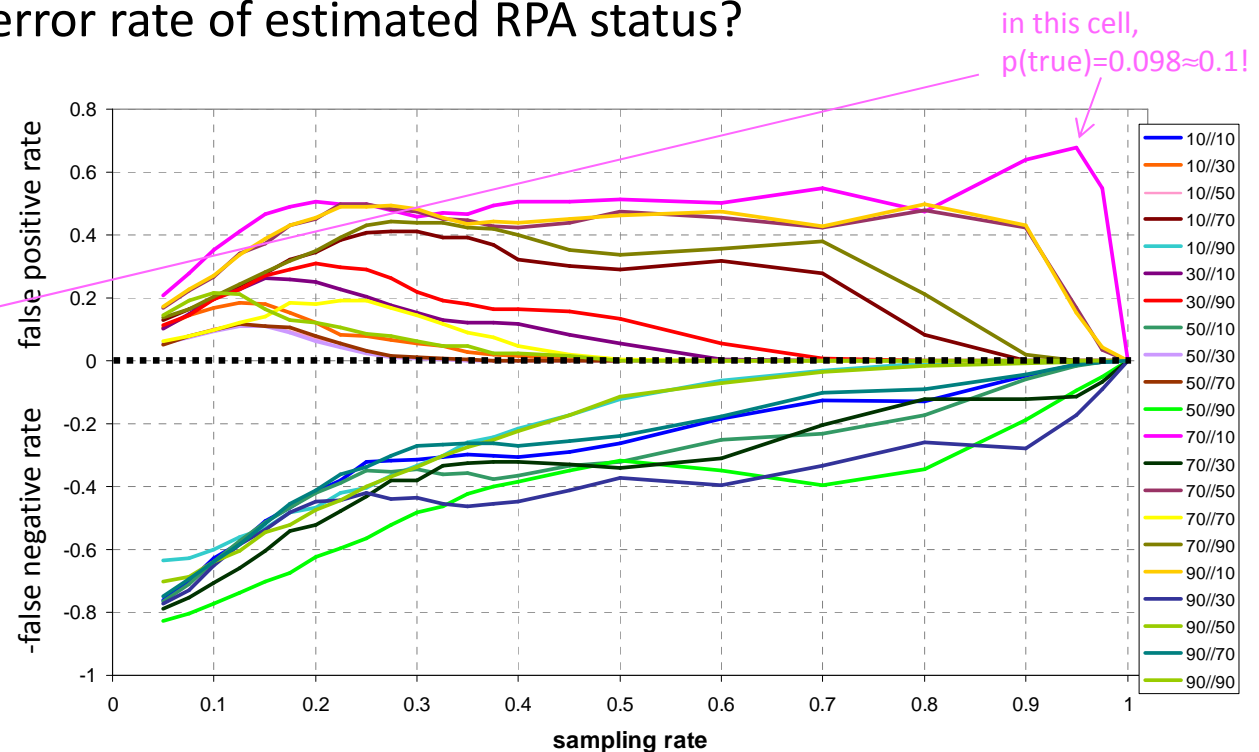
Sample size effect: A numerical experiment, 2

Finite population! (Statistically \rightarrow sample without replacement)

Question: For given sampling rate, assuming representative (random) sampling, which is the error rate of estimated RPA status?



method: many virtual “sampling campaigns” (2000-5000 realizations), calculate FP and FN rates of estimated RPA status



Even for high sampling rate, error chance can be high!

This is the case,

if a cells contains few houses / if true variability is high / if true p is close to class limit.

Sample size effect: Theory

In a cell U: population N, true “successes” K (i.e. $z > 200$).
True success probability: $p = K/N$.

Sample:

size n (sampling rate n/N), successes k.

Hypergeometric distribution:

$$\text{prob}(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

CDF, for estimating whether success rate $(k/n) = p' > 0.1$ (e.g.):

$$\text{prob}(X \leq k) = 1 - \frac{\binom{n}{k+1} \binom{N-n}{K-k-1}}{\binom{N}{K}} {}_3F_2 \left[\begin{matrix} 1, k+1-K, k+1-n \\ k+2, N+k+2-K-n \end{matrix}; 1 \right]$$

generalized hypergeometric function

Unpleasant! Little chance for practical use!!

If $N, K \gg n$, i.e. low sampling rate, but n still “large”, and $p \gg 0, \ll 1$:

$$\text{prob}(X \leq k) \approx \Phi \left(\frac{k - np}{\sqrt{np(1-p)}} \right)$$

- conditions often not fulfilled;
- true p not known → replace by p'
→ replace $\Phi \rightarrow t_{n-1}$??

Example II: Overdispersion

Observed Rn concentration: Y, true: Z

pdf of Y: $h(y) = \int g(y|z) f(z) dz$... compounded distribution, g =error distribution due to measurement uncertainty, f =true natural distribution.

$g(y|z)$ propagates into exceedance probability $\text{prob}(Y > y_0) = 1 - H_Y(y_0)$

For $\text{unc} \sim Z$, $\text{Var } Y = \text{Var } Z + \text{AM}(Z * \text{unc})^2$

Example: data as in (I)

$\text{prob}(Y > 200)$ calculated numerically for different relative measurement uncertainties, $Y \sim N(Z, Z * \text{unc})$.

True $\text{prob} = 5.48\%$ considerably inflated!

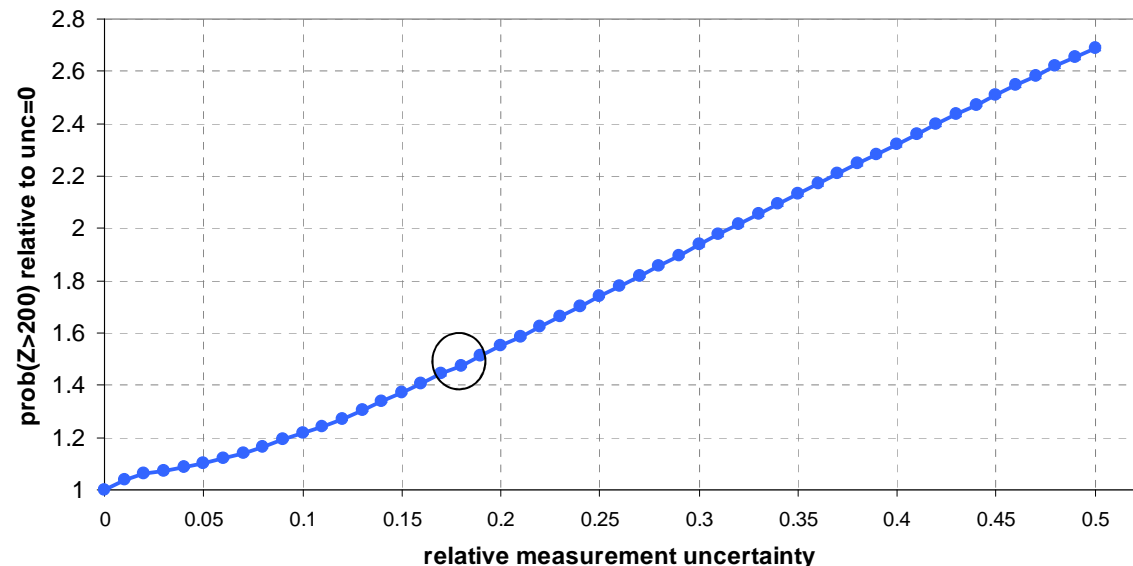
More realistic error model:

$\text{rel.unc} \sim \exp(-Z)$:

$\text{unc} = 30\%/10\%$ for $Z = 10/1000 \text{ Bq/m}^3$:

$\text{prob}(Y > 200) = (8.1 \pm 0.6)\%$,

i.e. 1.48 times true value... ○



Conclusion: RPA status probably systematically overestimated (=false positives) due to dispersion inflation caused by measurement uncertainty.

Question: How to “de-compound” or invert, to retrieve true exceedance prob.?

Example III: Estimation uncertainty

Model: Binomial cross-classification, one secondary predictor variable. Idea:

- RPA definition -> threshold for primary variable (Z = indoor Rn conc. exceed. prob.);
- Find threshold of secondary variable (Y = geogenic Rn potential) on which the RPA map is then based
- Construct truth table & ROC graph
- Perform statistic in ROC space, according constraints, e.g. tolerated 1. and 2. kind errors or optimized classification strength.

Procedure is easy and robust; drawbacks: ignores actual levels of the variables (similar to indicator kriging); ignores spatial correlation. Advantage: easy control over classification error probabilities.

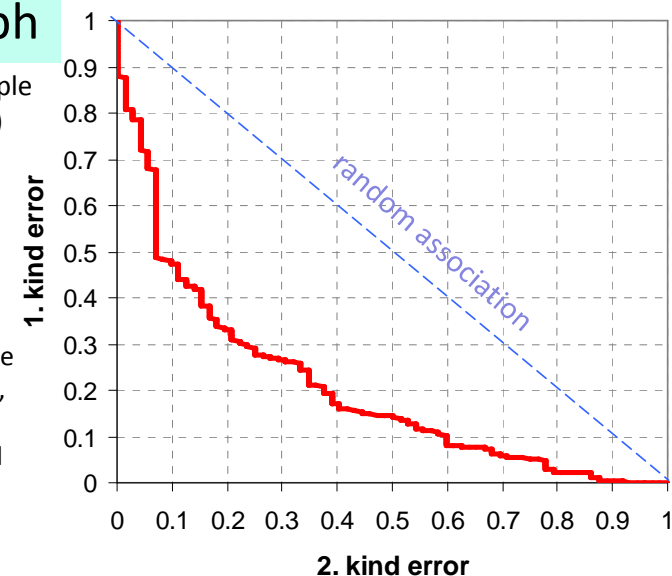
truth table

		primary variable	
		no effect	effect
secondary variable	estimated no effect	correct estimate true negative (TN)	wrong estimate false negative (FN) second kind error
	estimated effect	wrong estimate false positive (FP) first kind error	correct estimate true positive (TP)

ROC graph

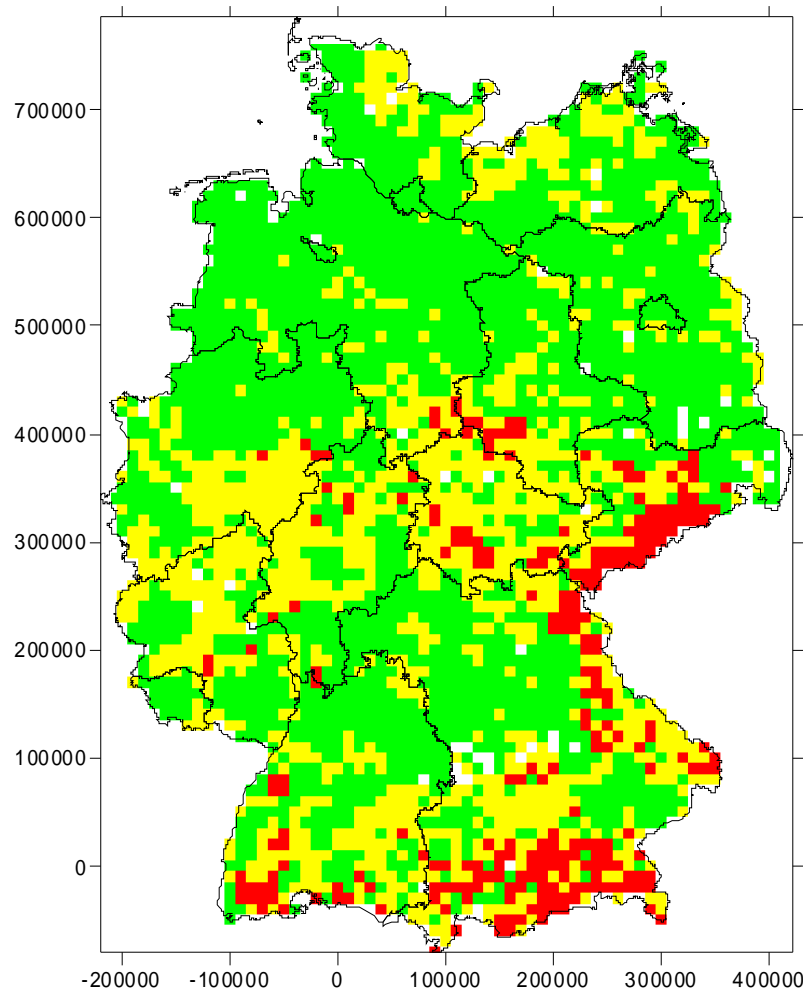
(for example next slide)

classification strength measured by statistics such as AUC (area between curve and diagonal), max. distance from diagonal or min. distance from (0,0)



Estimation uncertainty – real data!

Suggestion for RPAs, Germany, based on cross-classification method.



Primary variable: Z= indoor Rn concentration in ground floor dwellings, houses with basement;

Secondary variable: Y= Geogenic Rn potential (GRP). Modelled by SGS on U = 10 km × 10 km grid, geology as deterministic predictor.

RPA definition: grid cell U = RPA, if $p = \text{prob}_U(Z > 300 \text{ Bq/m}^3) > 3 \times \text{German average} \approx 10\%$.

p estimated by enhanced empirical exceedance prob., assuming LN within cells, $\text{GSD} = \exp(\text{SD}_U(\ln Z)) = 2$:

$$p = t_{n-1} \left[\sqrt{\frac{n}{n+1}} \frac{\ln(300) - AM(\ln Z)}{SD(\ln Z)} \right] \quad (\text{unfortunately biased estimator})$$

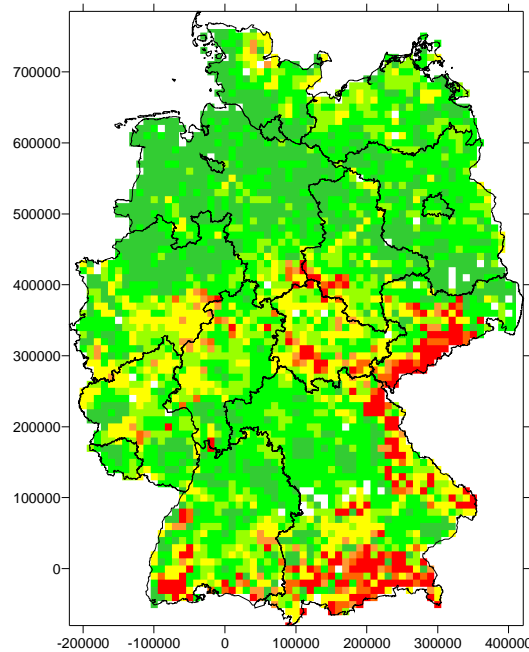
Cell U labelled RPA or non-RPA with confidence 90%, i.e. 1. and 2. kind error probability < 0.1.

RPA: Y > 44.5 (12.0% of territory);

Non-RPA: Y < 20.2 (49.8% of territory);

Yellow: undecided

Estimation uncertainty – quantification



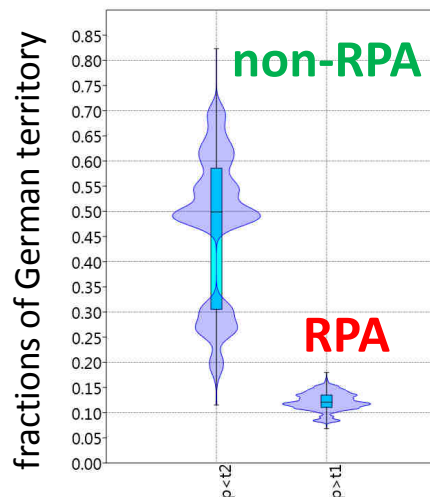
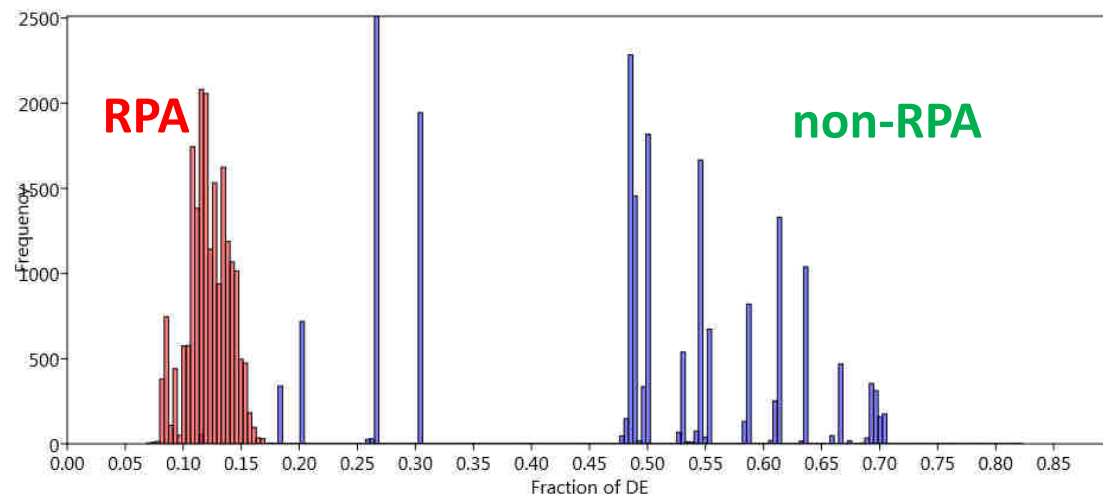
For estimation uncertainty component of model uncertainty:
Ignore unc. of input variables!
Only unc. of association $Z \sim Y$!

By bootstrap ($k=20,000$):

reddish hues: **RPA**: $CI_{90} = (38.2, 52.8)$

greenish hues: **Non-RPA**: $CI_{90} = (13.1, 26.4)$

fractions of German territory, 20,000 bootstraps



Distribution somewhat unexpected
Probably because classification is 'very' nonlinear transform

Conclusions & To-do

- RPA definition and estimation: not only academic exercise, but practically important. May have severe economic & political impact. Heavy stakeholder interest!
Therefore: QA very important!
- Uncertainty of RPA status (in terms of classification error rate, 1st/2nd kind error prob) has many sources of different types!
- Unc(RPA) can be high, in particular for spatial units close to class limits.
- To do:
 - explore uncertainty budget of RPA!
 - analytical approach for sample size effect
 - inversion of overdispersion effect
 -
- Open questions which are a big headache in practice:
 - how to communicate the fact that RPAs are “random objects”?
 - how to deal with RPA uncertainty in administrative decision-making?



RPA – a sensitive subject!

Action required in RPA can be costly → political disputes



Thank you!



Bundesamt für Strahlenschutz

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

