Estimation of Radon Priority Areas – sources of error and uncertainty

Peter Bossew

German Federal Office for Radiation Protection (BfS), Berlin

14th INTERNATIONAL WORKSHOP
GARRM
(on the GEOLOGICAL ASPECTS OF RADON RISK MAPPING)

September 18th – 21th, 2018
Prague, Czech Republic

v. 14.9.18

Content

- Motivation:
  Why is this so important?
- Sources of uncertainty
- Consequences & how to deal with?

Indoor radon - essentials

odourless, tasteless, colourless...

let's skip this part...

[LARUC Uni Cantabria]

Reminder: RPA definitions

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., $AM_B(C)\times$population) is among the upper 10%

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.

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).

Definition → Estimation

No matter how *defined*...

RPAs have to be *estimated* ...

... done with available data ...

... using a certain method.

Data, estimation method and actual performing estimation are prone to **errors** and **uncertainty**.

Data

Data are always
dirty, noisy, incorrect, erroneous,
incomplete, ill-defined, uncertain.

Data as observations:

- measurement uncertainty (not only counting uncertainty! Sampling and measurement procedures include uncertainty, sometimes this is the most important part, but difficult to quantify)
- "semantic" uncertainty
  (Value reported ground-floor measurement, in fact first floor,...)
- wrong (e.g., geology wrongly classified)
- sloppiness errors (manual copying of data, wrong insertion into table, Excel misreads decimal point, x and y coordinates confused,...)

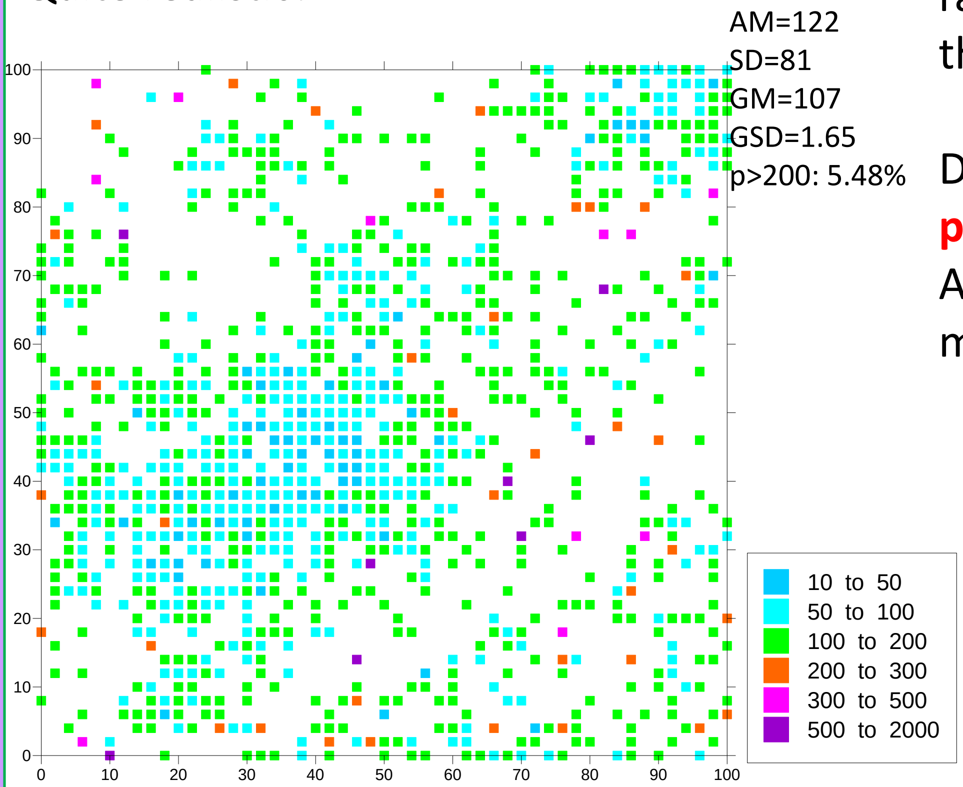
Data as samples from a population:

- not representative
  (relevant if the target is a statement about the population!)
- finite / limited sample size ⇒ estimation uncertainty

Sample size effect:
A numerical experiment, 1

The municipality Gigritzpatschen (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.

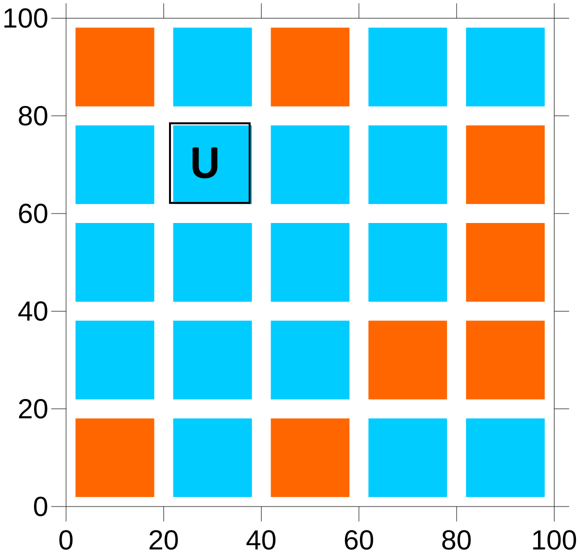


AM=122
SD=81
GM=107
GSD=1.65
p>200: 5.48%

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

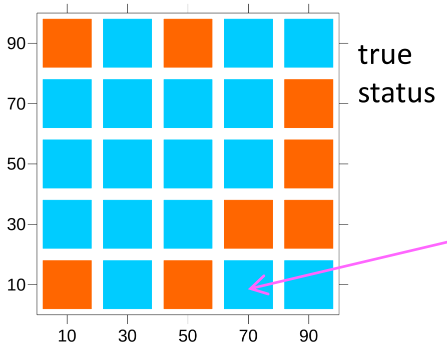
orange = RPA

blue = non-RPA

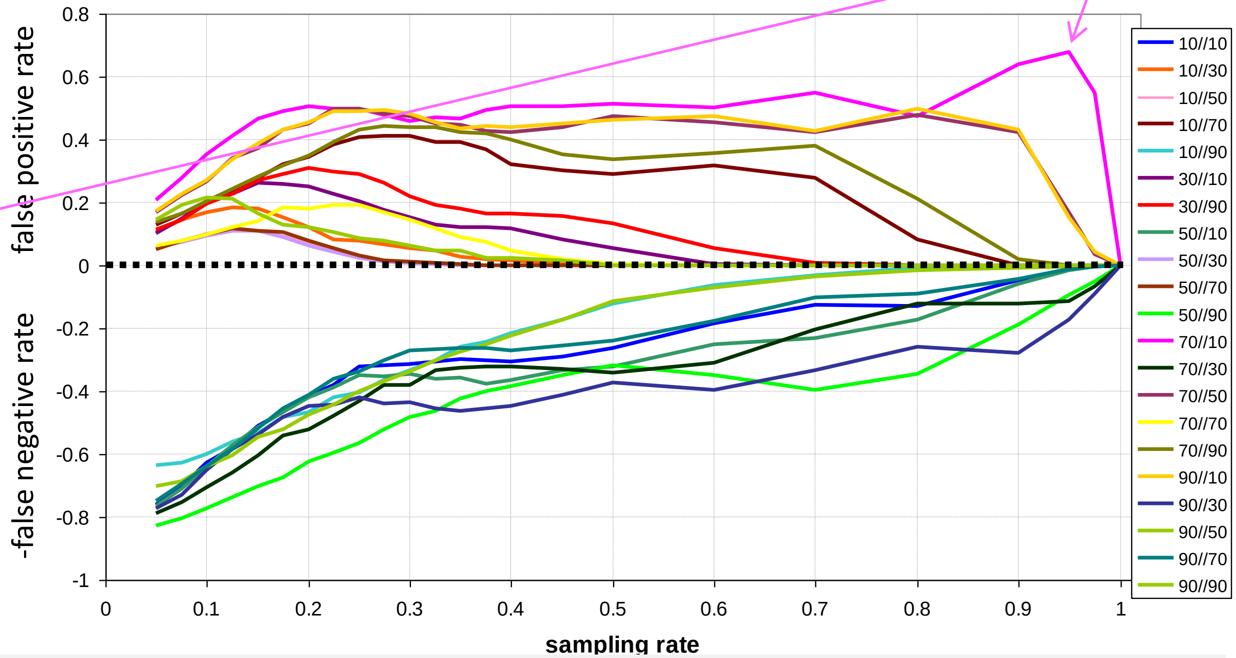
Sample size effect:
A numerical experiment, 2

Finite population! (Statistically → sample without replacement)

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



in this cell, p(true)=0.098≈0.1!



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.

Variability and uncertainty

A quantity Z can be truly variable in space or time... (typical for Rn quantities, as for most environmental quantities!)

Take a sample of the quantity from a given space or time interval, $\{z_1,...,z_n\}$

True mean of Z ... M, true SD .... S

sample mean AM(z_i) .... m, sample SD(z_i) ... s, estimates of M and S.

Uncertainty of the sample mean .... unc(m) = $s/\sqrt{n}$ = consequence of variability (s) (sample size, QA)

Variability = natural, irreducible!

Uncertainty of estimate: can be reduced (sample size, QA)

Methods – 1: complexity

Methods are almost always

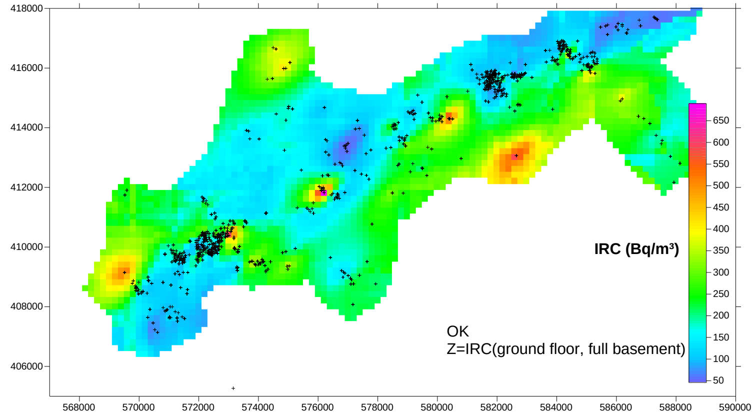
simplistic, idealization of a situation, make possibly unrealistic assumptions.

simplistic: not all relevant controls considered

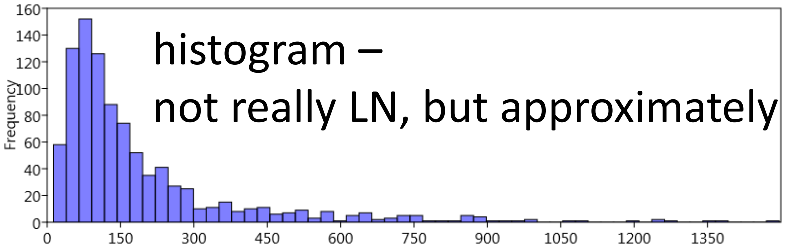
some typical unrealistic / idealistic assumptions:

- normal distribution (also LN often idealization!),
- homoscedasticity,
- statistical independence (to be able to use CLT),
- infinite sample,
- sample with replacement from finite population while it should be without,
- uncertainty of predictor ignored (regression! – leads to biased estimates of regression coefficients)

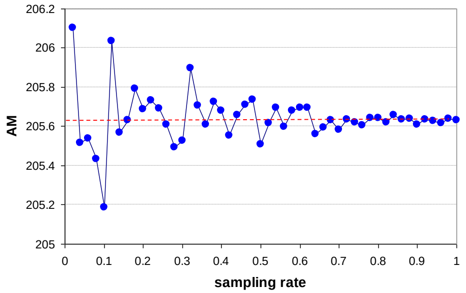
Ex.: simplification & sample size



Real region,
from Metro Rn WP4 mapping exercise

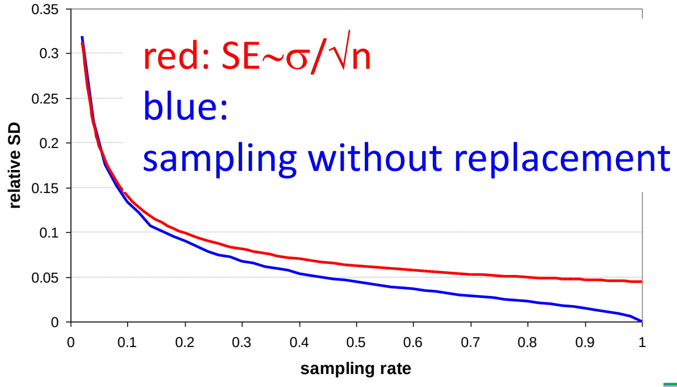


histogram –
not really LN, but approximately

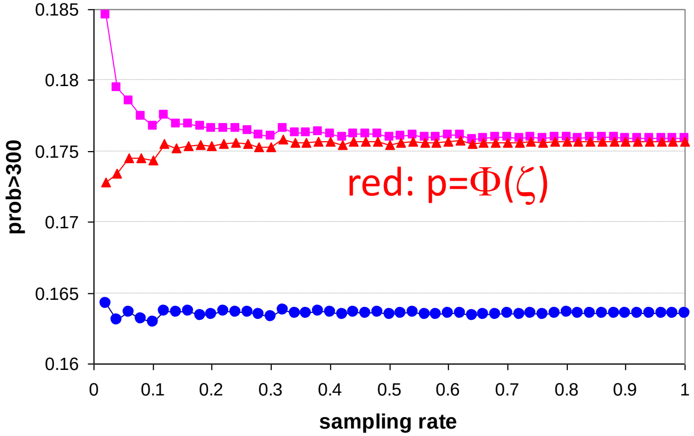


well known… AM is unbiased (=accurate), but precision depends on sample size

pink: $p' := prob_B(Z>z) = t_{n-1}\left(\zeta\sqrt{\frac{n}{n+1}}\right);\ \zeta := \frac{\ln z - AM_B(\ln Z)}{SD_B(\ln Z)}$



red: SE~σ/√n
blue:
sampling without replacement



red: p=Φ(ζ)

blue: simulation, counting instances

Methods – 2: diversity

Normally, “the true” method does not exist. In many cases, different methods or models are conceivable, which are all “more or less” correct. Results may be different.

Naturally, one would look for the “best” model, to be chosen by some validation procedure.

But:

- Different validation criteria may lead to preference of different models. (Accuracy, precision, 1st / 2nd kind error rates, cross-validation correlation, RMSE, other metrics?)
- The best available model may not be the best possible (which is not known).
- To avoid overfitting, one may remove predictors
- For practical reasons, one may opt for a compromise between model complexity and correctness.

Choice of model → structural uncertainty

Tentative Taxonomy of estimation approaches

certainly not complete!

	spatial correlation not considered	spatial correlation considered
univariate	• sample stat, stat(z) • enhanced by assuming univar. distribution, e.g. LN, stat’(z) → cut-off, I(stat(z))	• geostat. model & cut-off, I(Z*) • indicator kriging (hard/soft), I*(z)
multivariate	• ANOVA type • logistic-type regression, logi(z)=g(y) • geographically weighted, local regression etc. → cut-off, I(f(y)) • full bivariate through copula • cross-classification	• co-kriging et al. & cut-off, I(Z*;y) • regression kriging & cut-off, I(f*(y)) • indicator regression kriging, I*(f(y)) • indicator co-kriging, I*(z;y) • machine learning

I – indicator according RPA class definition; more complicated for multinomial class → I_1, I_2,..
* - interpolation, e.g. kriging type
z, y – primary and secondary variables

Example : European indoor Rn data, 1

Enhanced empirical exceedance probability in cell B by LN modelling, given n data Z in B:

p:=prob(Z>300 Bq/m³); RPA criterion: p>0.1 estimated from cell statistics:

$$p' := \mathrm{prob}_B(Z > z) = t_{n-1}\left(\zeta\sqrt{\frac{n}{n+1}}\right); \quad \zeta := \frac{\ln z - AM_B(\ln Z)}{SD_B(\ln Z)}$$

Z = long-term Rn concentration in ground floor dwellings;

z = 300 Bq/m³,

B = 10 km × 10 km cells

* … OK modelling

# … classified p< / >0.1

<u>univariate</u>

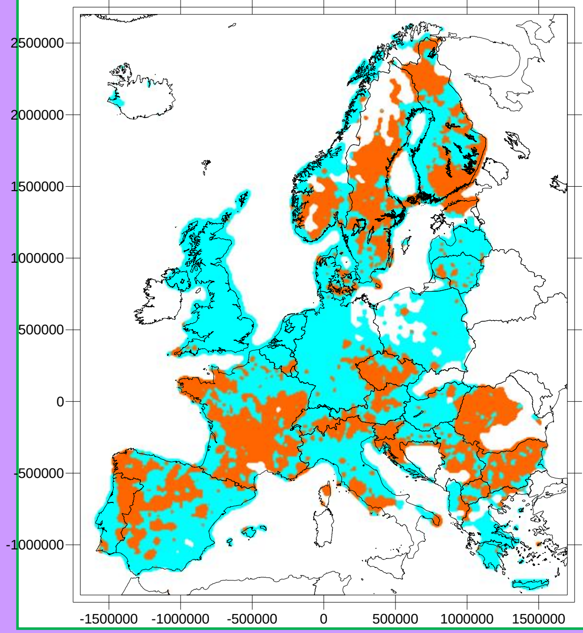
<u>bivariate,</u> Y= U conc.

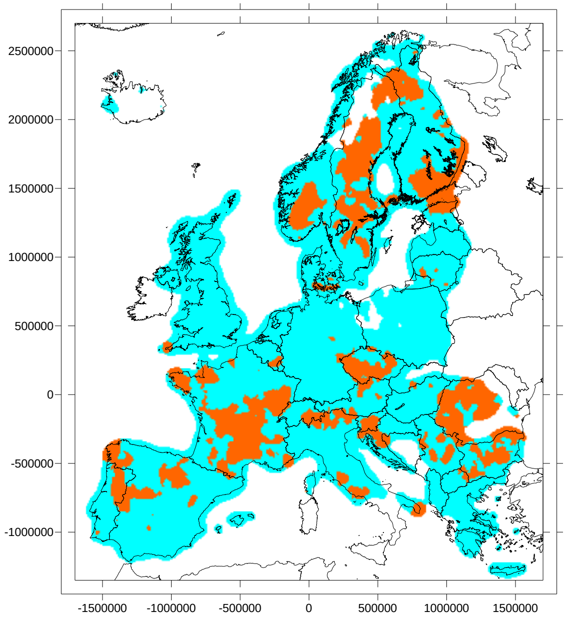
$p^{*\#}=I_{0.1}(p'^{*}(Z))$

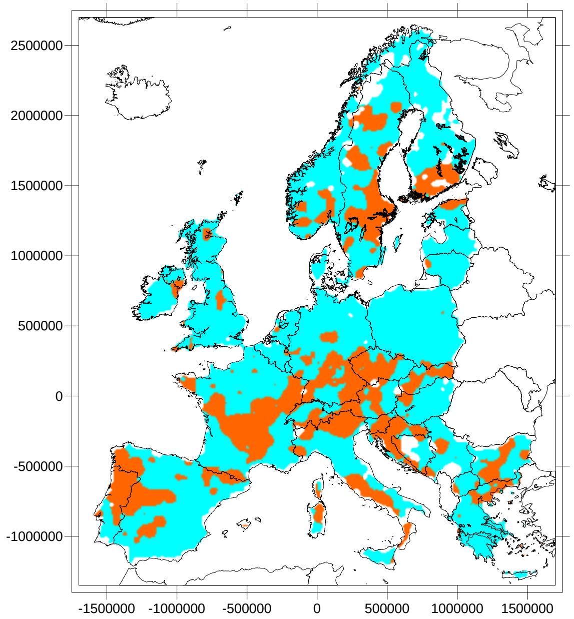
$p^{*\#}=I_{0.5}(I^{*}_{0.1}(p'(Z)))$

$p^{*\#}=I_{0.1}(p^{*}(f_{logi}(Y)))$

f_{logi} - logistic regression







Example: European indoor Rn data, 2

bivariate cross-classification p’ against U concentration in cells B

blue: prob(Z>300)<10% with 1-α confidence

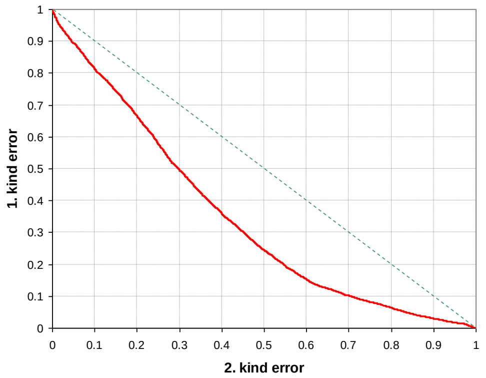
orange: >10% with 1-β confidence

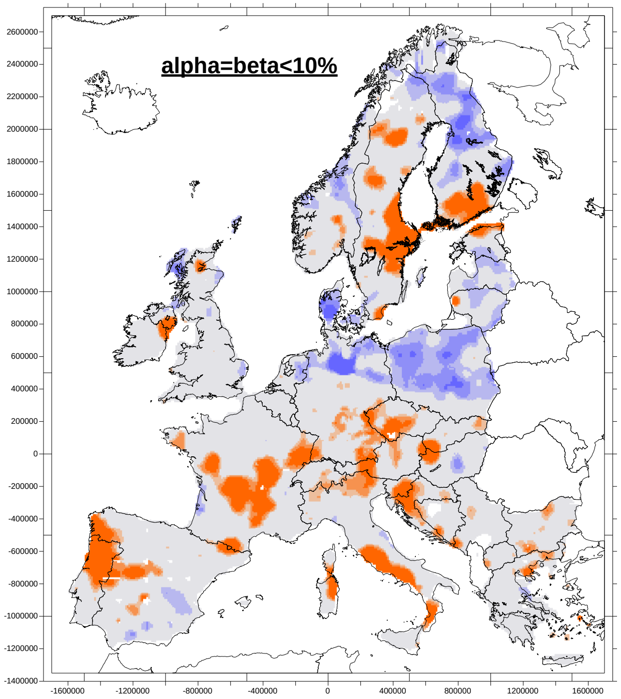
grey: undecided

… large areas because of weak association

class limits with 90%C.I. (by bootstrap)

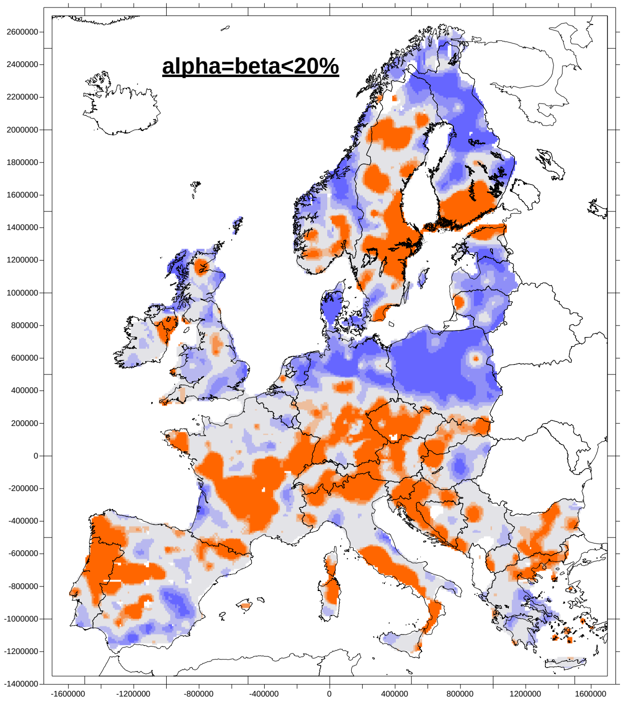
ROC curve close to diagonal: association between p’ and U not very strong!





lower: U=1.43 ppm (1.28…1.57)

upper: U=3.27 ppm (3.00…3.73)



lower: U=1.66 ppm (1.50…1.88)

upper: U=2.71 ppm (2.55…2.89)

Methods – 3: parameterization

Models must be parameterized:

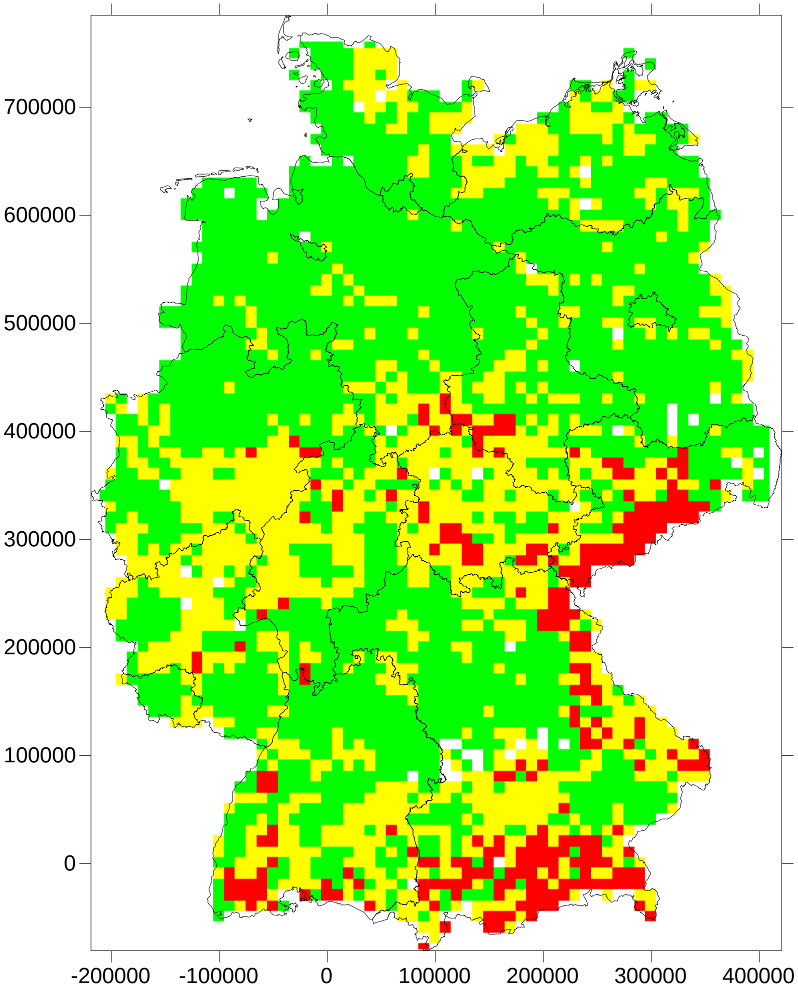
- Regression models: coefficients estimated from data
- Cross-classification: rule derived from training object

Therefore, parameters are uncertain: ⇒

Predictions are also uncertain.

Estimation uncertainty – real data!

Proposal 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:=prob$_U$(Z>300 Bq/m³) > 3 × German average ≈ 10%.

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

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

(unfortunately biased estimator)

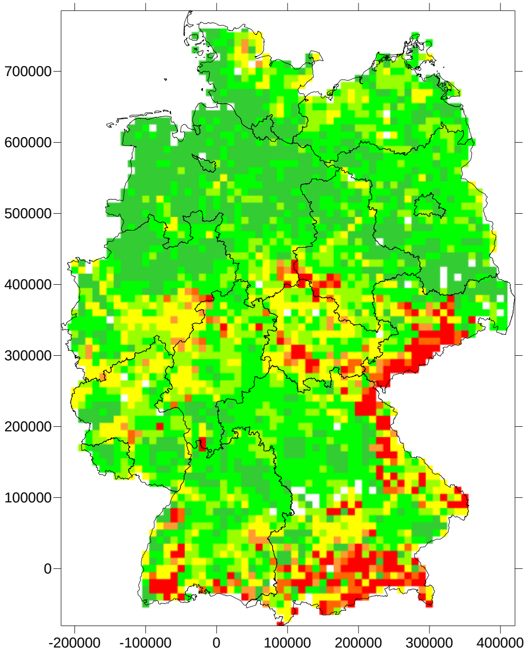
Cell U labelled RPA or non-RPA with confidence 90%, i.e. 1. and 2. kind error probability <0.1.

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

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

Yellow: undecided

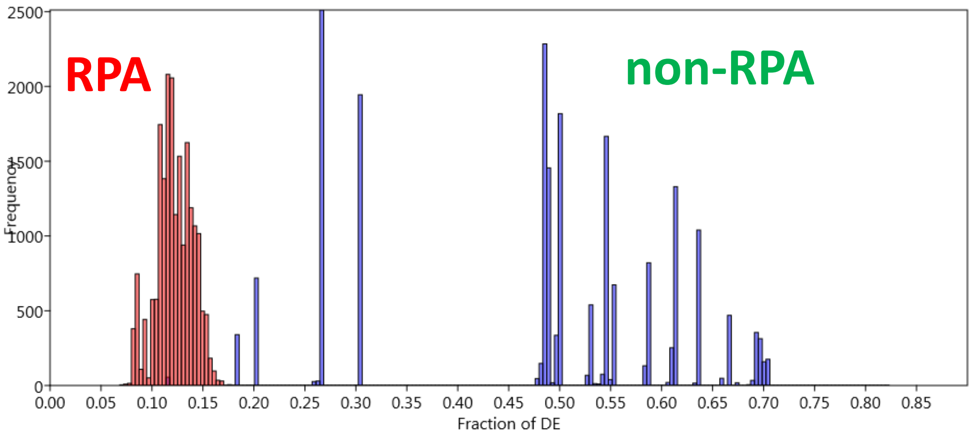
Estimation uncertainty – quantification

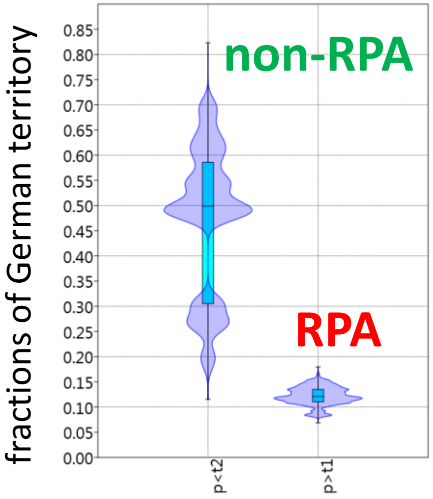


For estimation uncertainty component of model uncertainty:
Ignore unc. of input variables!
Only unc. of association Z ~ Y!

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



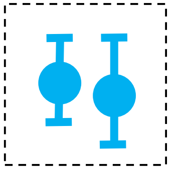


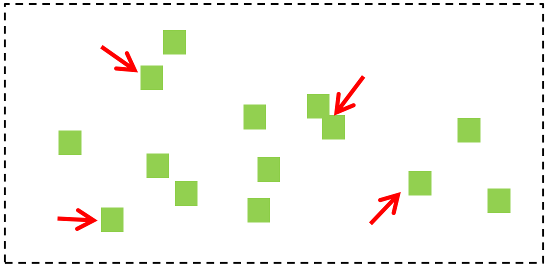
Distribution somewhat unexpected
Probably because classification is ‘very’ nonlinear transform

Summary: Sources of uncertainty

- Data uncertainty

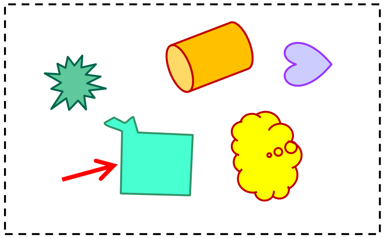
  - Intrinsic: data as observations;



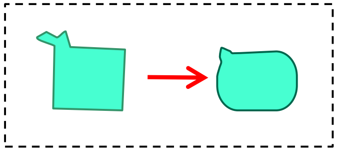


  - Data as samples

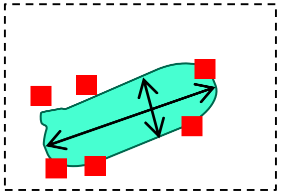
- Model uncertainty



  - Structural uncertainty: choice of model



  - Simplification uncertainty



  - Parameter estimation uncertainty

3 Levels of QA

1. **Design QA:** sampling such that the target (e.g. AM of the population in an area) can be met with given tolerance
   ⇒ sample size, representativeness
2. **Data QA:** correct measurement!!
   Classical metrology QA
3. **Evaluation QA:**
   - select proper method / model
   (easier said than done!)
   - consider, as far as feasible (because this can be complicated), model-induced uncertainty.

Aspects of 2 and 3 discussed in Metro Radon!

Finally: Why is this so important?

- Whether an area is assigned RPA or non-RPA (or a certain level of priorityness) can make a big economic difference:
  - implementation of building norms
  - measurement campaigns
  - remediation
  - property value
- Also possibly legal consequences, if RPA status is legally disputed by stakeholders!
- Administrations and decision makers want to be on the safe side – understandably.

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: Further explore uncertainty budget of RPA!**
- My impression: Unc(RPA) not taken sufficiently seriously!

- 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.

MetroRADON