Determination of radon priority areas – a classification problem

Bossew P.

German Federal Office for Radiation Protection (BfS), Berlin

Bundesamt für Strahlenschutz

v.7.11.17

IWEANR 2017

2nd International Workshop on the European Atlas of Natural Radiation

Verbania, Italy, 6 – 9 Nov 2017

Content

- Idea of RPA and Euratom-BSS
- Concept → definition → estimation → validation
- Uncertainties & errors
- Examples of procedures

The idea of radon priority area

A very good recent article about idea and concept of “reference level“ and “radon priority area”, which in my understanding addresses very well the “spirit” behind these concepts:

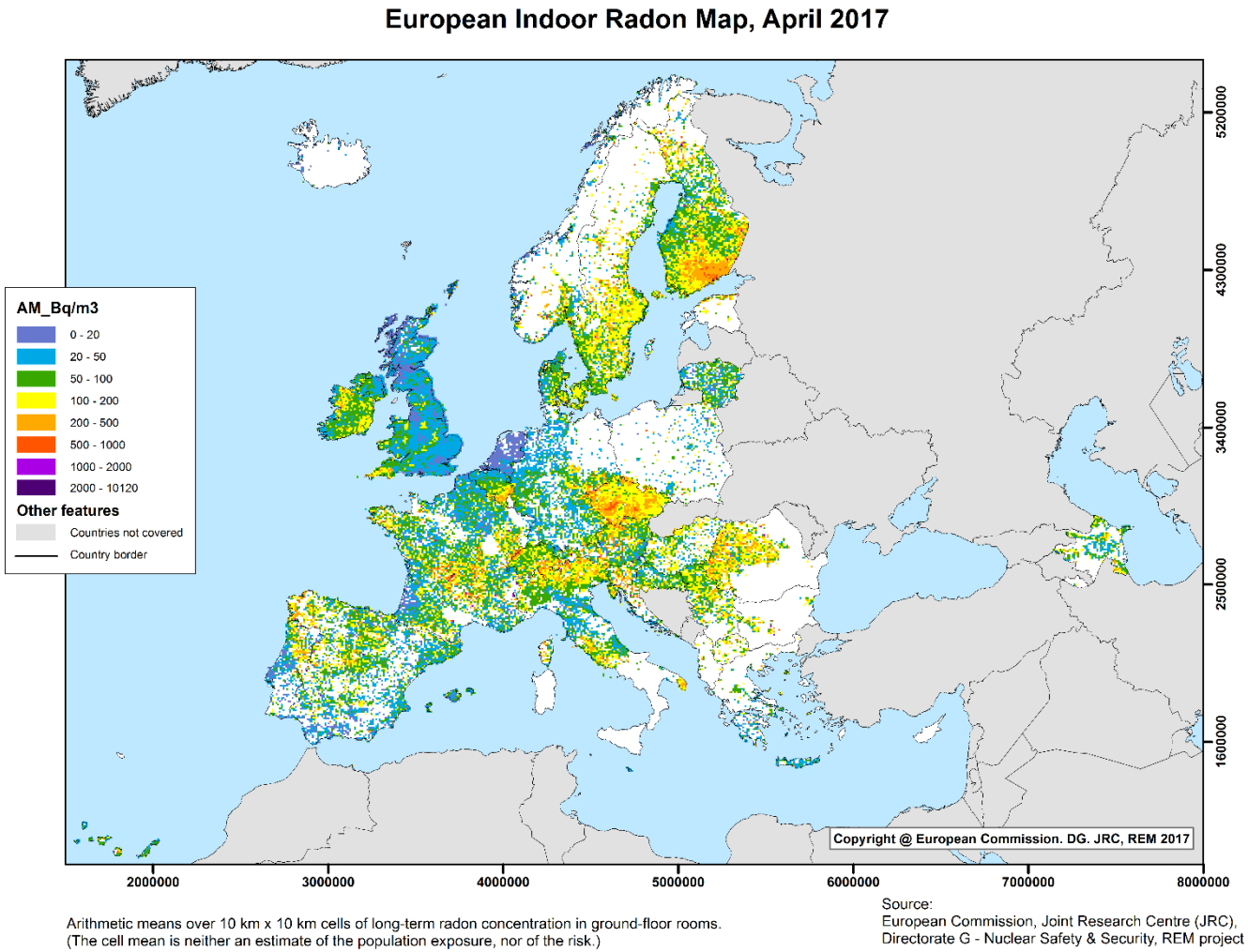
F. Bochicchio, G. Venoso, S. Antignani and C. Carpentieri:

Radon reference levels and priority areas considering optimisation and avertable lung cancers.

Radiation Protection Dosimetry xxx, 2017.

doi:10.1093/rpd/ncx130

Motivation



arithmetic means (AM) over 10 km x 10 km grid cells of annual indoor radon concentration in ground-floor rooms of dwellings.

Not a measure of exposure or of risk!

https://remon.jrc.ec.europa.eu/About/Atlas-of-Natural-Radiation

The EURATOM - BSS

presentation of Stefan Mundigl, Monday!

ISSN 1977-0677

Official Journal of the European Union

L 13

English edition

Legislation

Volume 57

17 January 2014

Contents

II *Non-legislative acts*

DIRECTIVES

★ 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, and repealing Directives 89/618/Euratom, 90/641/Euratom, 96/29/Euratom, 97/43/Euratom and 2003/122/Euratom 1

Text in all EU languages: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013L0059

Rn priority area (RPA) - 1

- The term “Radon priority areas” does not appear in the Eur. BSS; only a qualitative definition (Art. 103(3)):
  “… areas where the radon concentration (as an annual average) in a significant number of buildings is expected to exceed the relevant national reference level.”
- In these areas, Rn measurement is required in **workplaces** (Art. 54(2)) (in ground floor and basement rooms). Regarding **dwellings**, acc. Ann. XVIII (6),
  *“Strategy for reducing radon exposure in dwellings and for giving **priority** to addressing the situations identified under point 2 [about defining and estimating these areas]”* shall be established.
- Term RPA has been adopted in Europe to emphasize that the reason for this obligation is that in these areas, taking action has priority.
- Implicitly, this implies that Rn exposure should be reduced everywhere, if possibly with lower priority (given usually limited resources);
  after all, Ann. XVIII (13) states as part of the Rn action plan:
  [Establish] “long-term goals in terms of reducing lung cancer risk attributable to radon exposure”

This slide and others taken from: Bossew et al. (2017): Indoor radon: Implementation of the new Basic Safety Standards in Europe. Presentation, Radio-2017, Goiânia, Brazil, 25 – 29 Sept 2017. Partly modified and updated.

Rn priority area (RPA) - 2

- BSS “definition” is vague (politically motivated and to allow flexibility); needs to be translated into an operable definition.
- Once one has an operable definition, one must select a method how to estimate the RPA, given data.
- It may turn out that data still have to be acquired, i.e. surveys performed, and statistical methodology developed (BSS Ann. XVIII (2)).
- $\Rightarrow$ Definitions different between countries! … see later
  May create problems of harmonization, communication and credibility.

MetroRn
WP 4.4

<u>Workflow:</u>

Concept $\rightarrow$ Definition $\rightarrow$ Estimation $\rightarrow$ Validation

Role of MetroRadon

The EMPIR – MetroRadon project has the general purpose of providing QA support to the “supply chain”:

- primary standards
- calibration
- measurement (low concentrations, Tn interference)
- RPA definition & estimation
- Inconsistencies across borders

Some topics of this presentation are closely related to MetroRadon!

see presentation Valeria Gruber et al., Monday

MetroRn
WP 4.1.1

Definitions

- The “fuzzy” or “conceptual” definition of the BSS has to be translated into an operable definition.

BSS: “… *areas where the radon concentration (as an annual average) in a significant number of buildings is expected to exceed the relevant national reference level.”*

- Examples for operable definitions:
  - A municipality is labelled RPA, if AM(C)>RL
  - A grid cell is labelled RPA, if prob(C>RL_C)>RL_p. (E.g., prob(C>300)>10%)
  - A municipality is labelled RPA, if its dominant geology is one with GM(C in this geology)>RL.
- Next step: find a method to <u>estimate</u> the areas according to the definition.

MetroRn
WP 4.1.1,
4.3.1

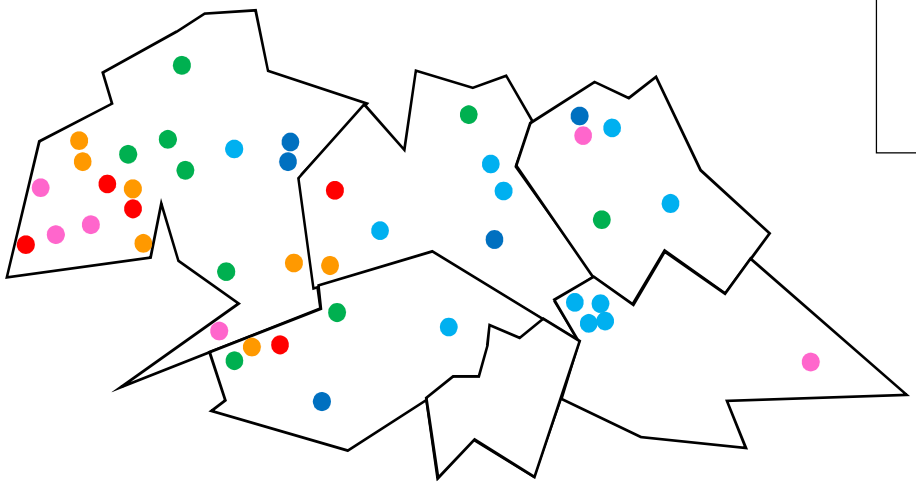
Estimation

- Once a definition is given:
- Based on data, the RPAs have to be **estimated** conforming the definition.
- Estimation is a **statistical procedure**! ..... **classification**
- It results in “random objects”, which are subject to **uncertainty**!
- **Data:**
  - observations (measurements) of the same quantity which defines the classification categories
    e.g.: measure indoor Rn; categories based on indoor Rn, e.g. RL=300
  - observations of different quantity
    e.g.: measure geogenic RP; categories based on indoor Rn RL
  - auxiliary data which define a trend
    e.g. geological units

On classification, 1

Wikipedia: “In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs…” ---- better: **object**

Given: Municipalities with scattered measurements
Task: Municipalities are the objects to be classified
Question: To which category does each municipality belong?



Among possible categories:

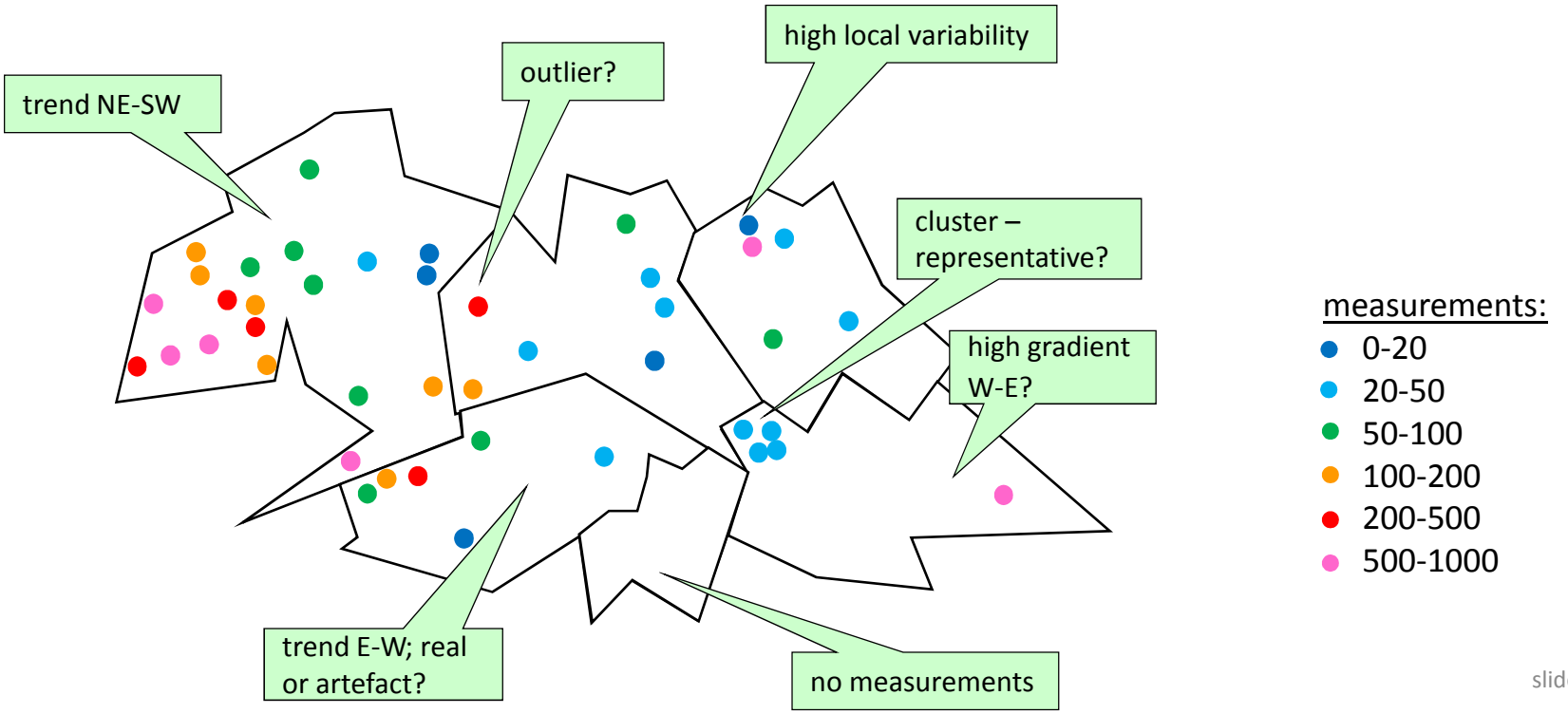
I: AM<100	A: prob(>300)<1%
II: AM=100-300	B: 1%≤prob(>300)<10%
III: AM>300	C: prob(>300)≥10%

measurements:
- 0-20
- 20-50
- 50-100
- 100-200
- 200-500
- 500-1000

On classification, 2

Step 1: Objects (municipalities) have to be assigned a quantity Z, based on measurements; e.g. Z=AM(x), GM(x), Med(x), Q90(x), SpatM(x), prob(x>RL),… (x=measurements)
Step 2: Classify these quantities into categories.

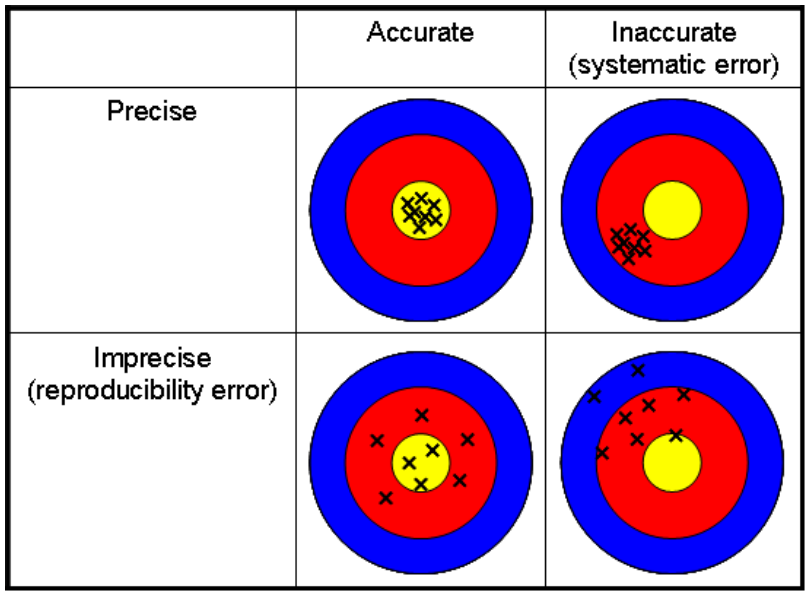
Problems:
- Z have uncertainty: precision depends on number of measurements, true dispersion; accuracy depends on representativeness
- true variability: trend or local variability ⇒ high chance of local misclassification



MetroRn
WP 4.3.1

Errors & uncertainties

- **precision / accuracy**
  accurate: low bias = representative
  precise: low random uncertainty

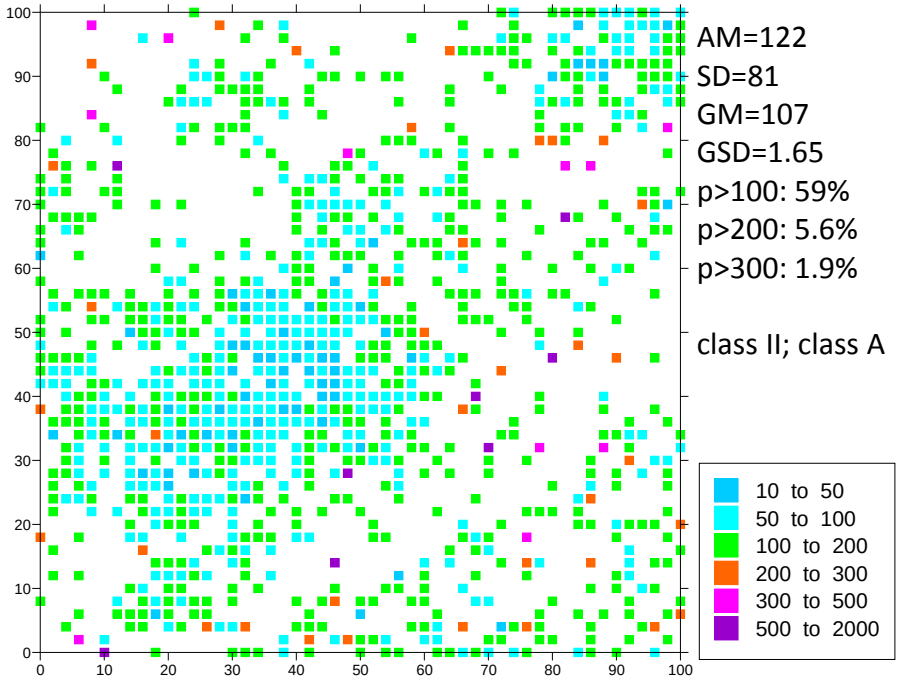


- **classification errors**
  - 1.kind error: effect detected, although not existing in reality; "false alarm"
  - 2.kind error: effect which exists in reality, but has not been detected; "false non-alarm"

- **assessing uncertainty**
  - parametric
  - non-parametric: MC, bootstrap

A numerical experiment, 1

The municipality Gigritzpatschen (AT),
Rn concentrations in all 1004 houses.



In a survey, we cannot measure all of them, but a number k, selected randomly. I.e., a representative sample.

Then we classify the municipality according 2 schemes:

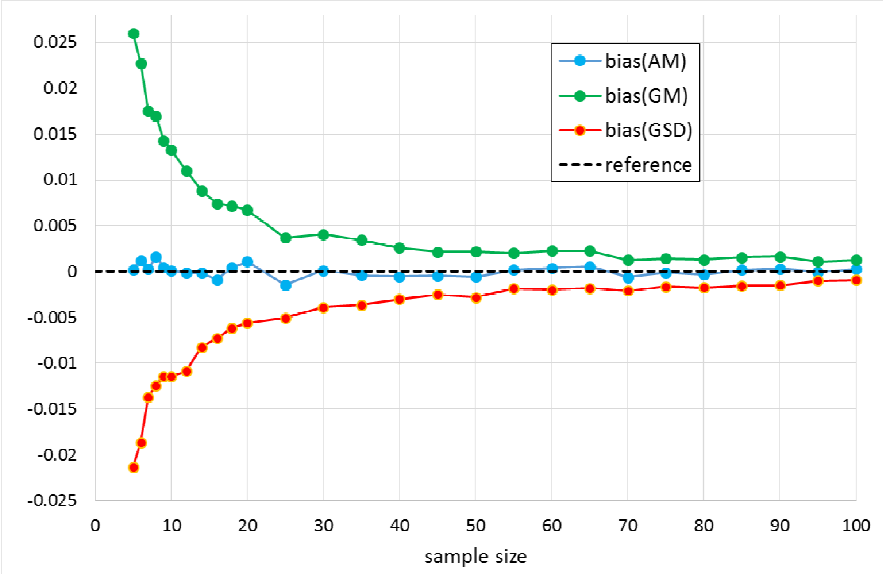
scheme 1:
if AM<100: class I; if >100: class II
scheme 2:
if p(>300)<2%: class A; else class B.

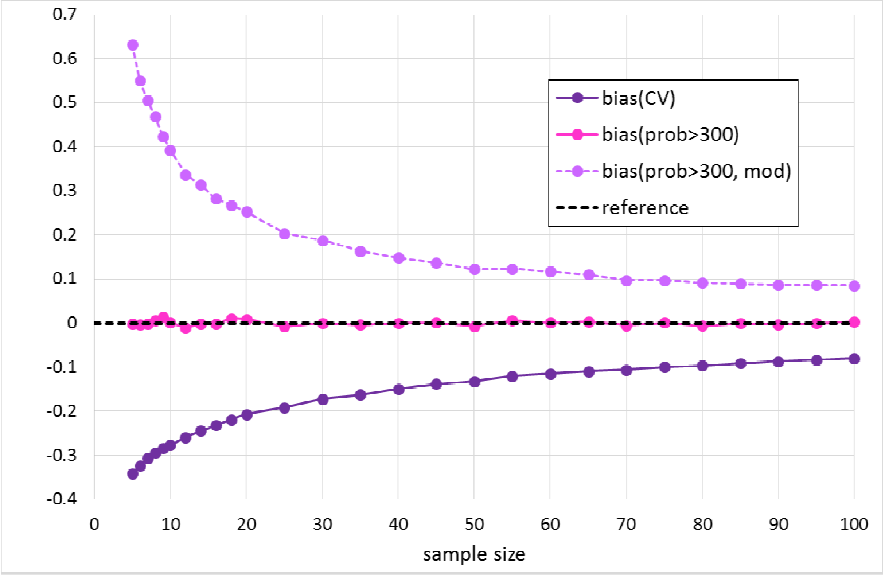
Question:
With which probability will we misclassify?

Method:
Take k random samples many times, compute statistics over realisations.

A numerical experiment, 2

bias = AM_{sim}(estimated quantity)/(true quantity)-1



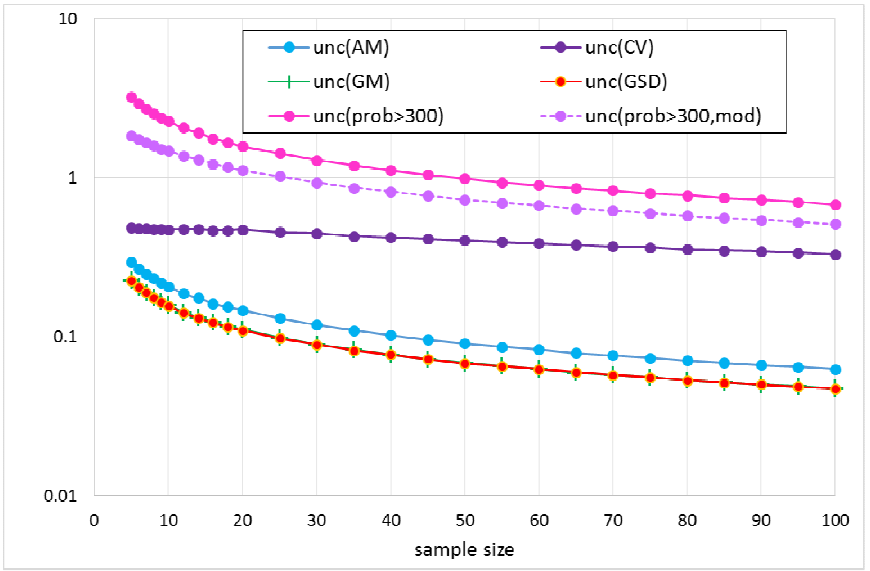


<u>bias:</u> measure of **accuracy**
<u>rnd. uncertainty:</u> measure of **precision**

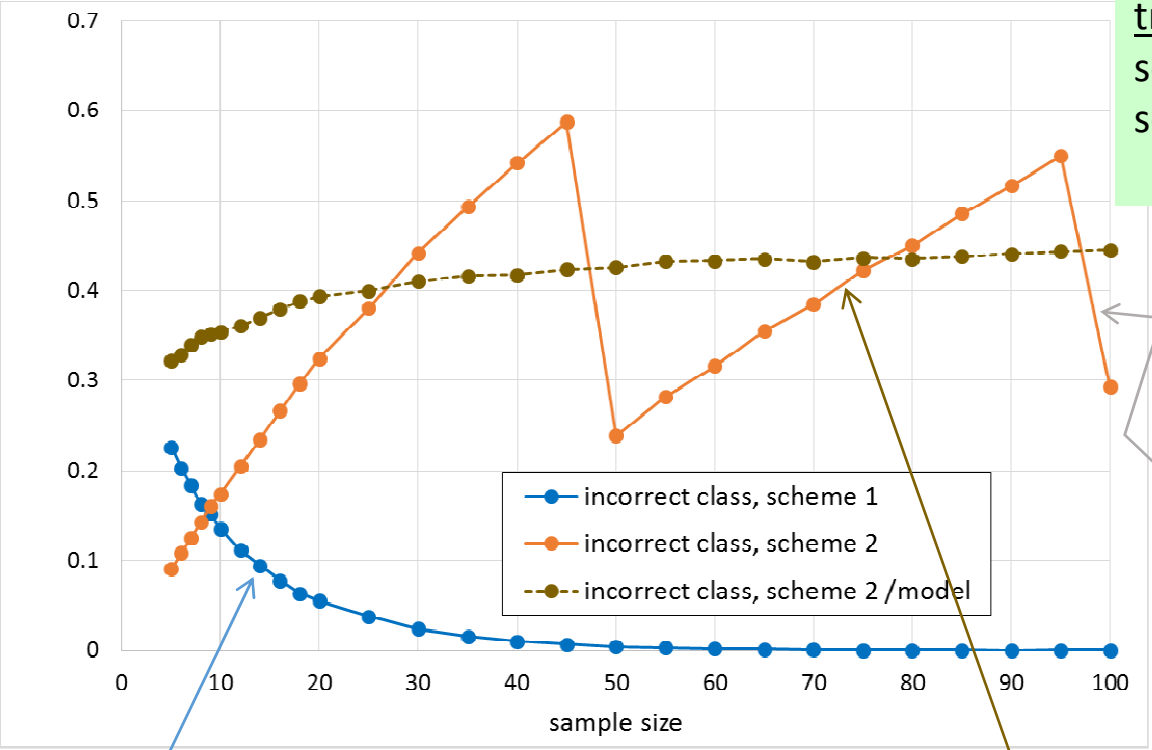
<u>empirical prob>300</u> = (number of observations>300) / (total number = k)
<u>modelled prob>300:</u> assume the k observations LN distributed, estimate GM, GSD and prob=1-Φ[(ln(300)-ln(GM))/ln(GSD)]

AM_{sim}, CV_{sim}: statistics over simulations (here: 50,000)

unc=CV_{sim}(estimated quantity)



A numerical experiment, 3



<u>true:</u>
scheme 1: AM>100 ⇒ class II (=RPA);
scheme 2: (prob>300)<0.02 ⇒ class A (=non-RPA)

Where does this funny curve come from ??

Measure k houses; let i=number with Rn>300. Can be understood as Bernoulli trial: k=tries, i=successes.

The true prob>300 equals p=0.019. Empir. probability i/k>0.02, i.e. that we have class B, means, i>0.02*k.

Then, prob[more than i=k*0.02 observations above 300, with k tries and true (prob>300)=0.019] equals 1-CBin(int(0.02*k),k, (prob>300)).

CBin= cumulative Binomial distribution.
int: because the number of observations is always an integer number. At each increment of k equalling 1/0.02=50, this number jumps by +1 ⇒ this causes the form of the curve.

probability that falsely classified as I instead of II, i.e. falsely classified as non-RPA. This error shall be below β (2.-kind error). For β=0.1 ⇒ k≥13 houses to be measured

probability that falsely classified as B instead of A, i.e. falsely classified as RPA. This error shall be below α (1.-kind error). For α=0.1 ⇒ **not achievable !!**
In this example, there will always be a high risk of "false alarm"! Even increasing with sample size!

MetroRn
WP 4.3.1

Consequence for sampling design

- If maximum tolerable classification error rates (α, β) are given as external constraint ("political parameter"):
- How to design a sampling scheme which guarantees classification which meets that condition?
- Difficult statistical question!
  Possibly not always achievable?
  (see Gigritzpatschen example)

- Still working on this problem $\rightarrow$ MetroRn WP4

Cross-classification, 1

Statement of the problem

- Objects (such as municipalities) shall be classified whether RPA or not, or into which RPA class they belong.
- *Classifier* = statistic on indoor Rn concentration (AM, prob>RL, etc.). This follows from BSS.
- No or not sufficient indoor Rn data available.
- Therefore: Use different available variables (GRP, U conc. in soil, ADR, occurrence rate of vampires, geological units, etc.)
- Derive *secondary classifiers* for these variables, e.g. U-conc. < or > 1 ppm, etc., or for combinations of such variables → geogenic Rn hazard index GRHI.

Cross-classification, 2

Example 1

definition:
municipality has AM(Rn)
< 100 …… class I
100-200 …. class II
> 200 …… class III

shall correspond as well as possible!

available data:
U concentration in soil, ppm
< x1 …... class I
x1-x2 …. class II
> x2 …... class III

task: find optimal x1, x2

Example 2

definition:
municipality has AM(Rn)
< 100 …… class I
100-200 …. class II
> 200 …… class III

shall correspond as well as possible!

available data:
geological map=set of geol. units
subset1 …... class I
subset2 …... class II
subset3 …... class III

task: find optimal subsets 1,2,3 of geological units

Cross-classification, 3

What does "optimally" mean?

- Classification error rates as low as possible?
  (But they cannot be minimized independently, in general)
- Conforming to pre-set tolerable error rates?
  (… done this way in DE)
- More general: minimizing a loss function?
- Now 2 sources of classification errors:
  - Variability within municipality
    (*has been discussed in the Gigritzpatschen example)*
  - Imperfect relationship between the predictor
    (e.g. U conc.) and the primary classifier (indoor Rn)

Done rigidly → quite complicated !

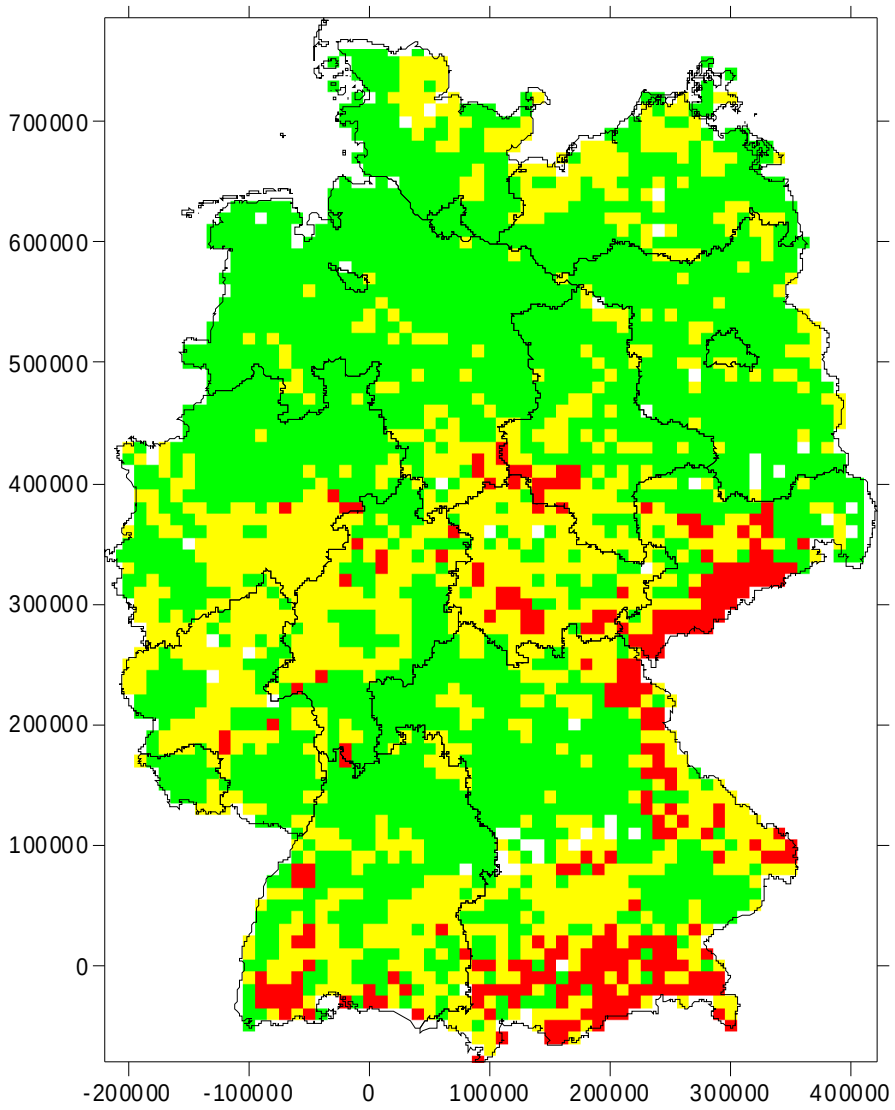
Cross-classification, 4

Open problems:

- *Multinomial* cross-classification
  (more than 2 classes: “low”, “medium”, “high”, etc.)
- *Multivariate* cross-classification
  More than 1 predictor
  possible way: “dimensional reduction” by constructing an index RHI = 1-dim (i.e. univariate) predictor… see pres. later

MetroRn
WP 4.3.4

Example, DE



- <u>Definition of RPA (proposed):</u>
  cell(10 km×10 km) in which estimated prob(C>300)=3 ×prob(mean, DE)=0.09
- $\alpha=\beta=0.1$
- classifier = indoor C exceedance probability;
- estimated from GRP → is secondary classifier
- **red:** certainly RPA (prob that it is not: ≤0.1); threshold GRP=44.5
  green: certainly non-RPA (prob that it is: ≤0.1); threshold GRP=20.2
  yellow: between
- binomial classification → trinomial through α, β concept
- GRP thresholds have uncertainty! (SD a few GRP units)

essential QA feature!!

Validation

- How to validate a classification result?
- So far no experiences in RPA classification, to my knowledge.
- May become important for legal reasons!
- Basic possibilities:
  - partition training / validation data
  - “postdiction”: develop model and then apply to instances with known true classification

Supporting research projects

MetroRadon

- QA of the chain from primary standards over measurement to RPA definition and estimation.
- Tn interference
- Inconsistencies in RPA definition across national borders

Presentation of Valeria Gruber et al., last Monday!

RESPIRE

- Rn Geo-database
- Demonstrate remediation in areas with different GRP
- Rn risk perception

Presentation of Giancarlo Ciotoli, last Monday!

Next important event in this context

GARRM 2018

14th International Workshop on the Geological Aspects of Radon Risk Mapping

Prague, 18 – 21 Sept 2018

Database of events related to Rn and Nat. Rad.: http://radoneurope.org/

ERA
European Radon Association

Improving Awareness and
Reducing Risk of
Radon Exposure Across Europe

Conclusions

- RPA definition – a sensitive topic
- RPA estimation – lot of technicalities!
  beware of statistical traps!
- Classification problems sound easier than they are,
  if performed rigidly;
- In particular 1./2. kind error considerations
  can be tricky!
- Communication of these problems to
  administrations and law makers -- ??

Thank you!

Bundesamt für Strahlenschutz