Estimation of radon prone areas by bivariate classification using ROC analysis

P. Bossew

German Federal Office for Radiation Protection, Berlin, Germany

12th INTERNATIONAL WORKSHOP GARRM

(on the GEOLOGICAL ASPECTS OF RADON RISK MAPPING)

Prague, Czech Republic, 16-19 Sept 2014

Bundesamt für Strahlenschutz

Rationale & background

- Indoor Rn: considered a hazardous pollutant already at low concentrations.
- Therefore: regulation. In Europe, 2014: Basic Safety Standards (BSS)
- Among other topics:
  - maximal reference value (RV): 300 Bq/m^3 for residences and workplaces;
  - action plan including delineation of **radon prone areas** (not using this term)
- Concept not precisely defined
- However defined: how to estimate RPA?

Content

- Concepts of radon prone area (RPA)
- Estimation of RPA, defined through geogenic **radon potential** (RP), by binary classification
- Examples
- Radon prone geologies

Concepts & definitions - 1

- Geogenic radon potential (RP):

  "What Earth delivers" in terms of Rn.

  Saxony Ministry of Environment: *"The radon potential is the property of soil to release Rn into soil air, and together with it make it available at ground surface."*

- Definitions: various in literature.

  Here: "Neznal RP": $RP = C(soil)/(-\log_{10}(k)-10)$

  C(soil): Soil radon, kBq/m^3

  k: permeability, m^2

  observed according a defined and QAed protocol.

Concepts & definitions - 2

RPA:

- **(a) through indoor Rn**

  - area in which mean C(in)>threshold;

  - area in which prob[C(in)>threshold]>threshold' (e.g. prob(C>300)>2%)

  - original draft of BSS (2011): "*Radon-prone area means a geographic area or administrative region defined on the basis of surveys indicating that the percentage of dwellings expected to exceed the national reference level is significantly higher than in other parts of the country";*

  Formally (B=country, RV=reference value):

  $U \subset B$ is RPA if $prob[C(x)>RV; x \in U] > \alpha\, prob[C(x)>RV; x \in B]$

- **(b) geogenic RPA: through RP:**

  area in which mean or other statistic of the RP>threshold; formally:

  U is RPA if $stat[RP(x): x \in U] > RP_0$

  Instead of RP: other geogenic quantity (proxy of RP) can be used

Estimation of RPA

- Here: geogenic concept of RPA adopted, i.e. based on RP or proxy Z
- Task: Define a threshold Z_0 or RP_0 of the Z
- RP is by itself no radiologically relevant quantity $\Rightarrow$ link RP_0 to one which is.
  Best candidate of course: indoor Rn
- Hence: find RP_0 which corresponds to a given threshold of indoor Rn....
  "Calibration" of the predictor RP

Method: binary classification by ROC, 1

- Domain B covered with grid;
  cells e.g. 10 km × 10 km
- Estimate mean Z (or other stat) in each cell.
- Given: indoor Rn (C) data; estimate mean C (or other stat) in each cell
- Classify cells twofold:
  stat(C in cell)>C_0 yes/no… criterion CRIT
  stat(Z in cell)>Z_0 yes/no

Method: binary classification by ROC, 2

- Create a truth table:

		<u>*observed*</u> classification, according criterion CRIT	
		pos	neg
<u>*predicted*</u> classification, according value of rp.	pos	TP	FP
	neg	FN	TN

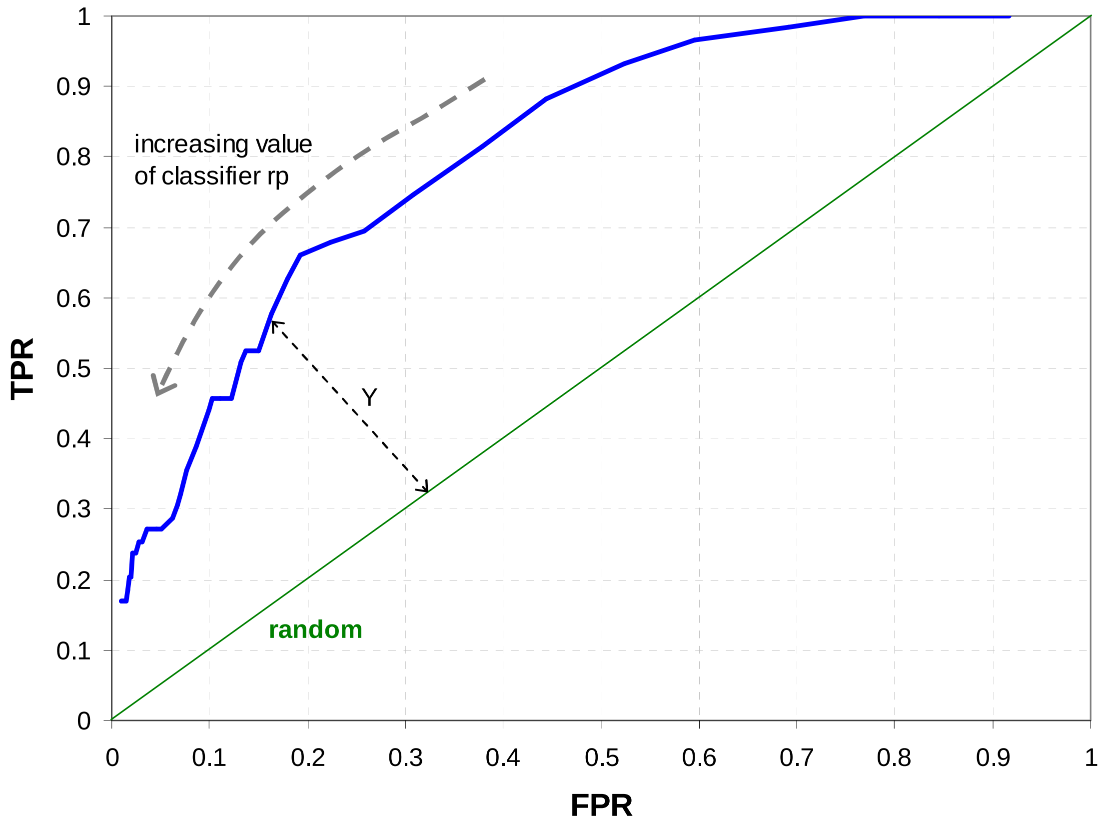
TP=number of true positives
FP=False positive
FN=False negative
TN=True negative

- Calculate statistics:

  True positive rate
  TPR = TP/(observed positives)=TP/(TP+FN)
  False positive rate
  FPR = FP/(observed negatives)=FP/(FP+TN)

Method: binary classification by ROC, 3

Plot a graph
TPR vs. FPR = "ROC" graph
(receiver operating characteristic)



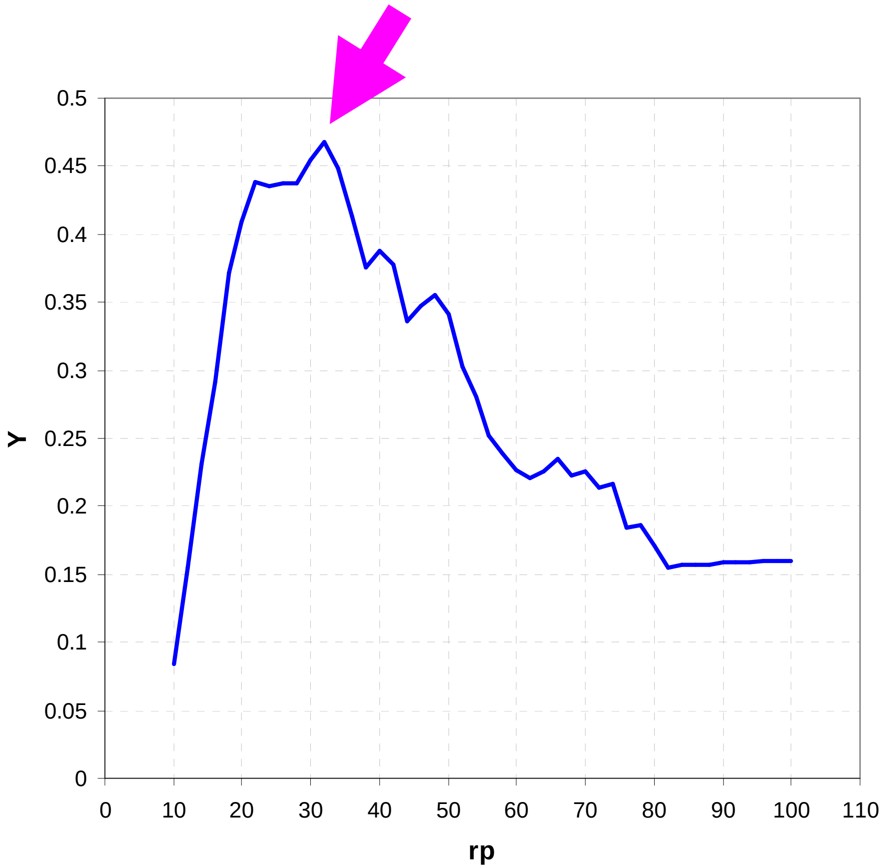
Define a score (loss function, target function) to find the optimal point of the ROC graph; e.g.

- Y-score: the point farthest away from the diagonal
- d01: as close to optimal point (0,1) as possible
- force 1.kind error rate= 2.kind error rate (intersection with 2.diag.)
- force 2.kind error rate ß < 0.2 (e.g.)

Method: binary classification by ROC, 4

Optimize the score function

e.g. find maximal Y ⇒ optimal threshold $rp_{opt} = RP_0$



Result for Germany;
CRIT: E(C)>100 (see ex.1)

- Y-,d01-scores: RP_0=32
- 1.kind=2.kind error rate:
  RP_0 = ca. 28
- 2.kind error rate set to ß=0.2:
  RP_0 = ca. 23.

Extension:

Y etc. are a "metric" on ROC-space. Modify metric such as to assign different weights to x- and y-axis (⇒ different weights to 1. and 2.kind errors) … gives different RP_0.

So far, we used ordinary Y-score.

Method: binary classification by ROC, 5

Caveats

The result (i.e. the optimal RP_0) depends on two types of decisions:

Political decisions:

- the calibration criterion CRIT ↔ reference value
- relative weight of 1. and 2. kind errors

Technical decisions:

- cell size
- definition of RP
- which score for optimizing the ROC
- how to estimate CRIT and the predictor RP per cell
- …

Uncertainty:

1. and 2. kind error rates (however adjusted) relatively high.

Why? Because of not very high correlation between RP and indoor Rn, since also non-geogenic (i.e. anthropogenic) factors influence indoor Rn.

Intermezzo

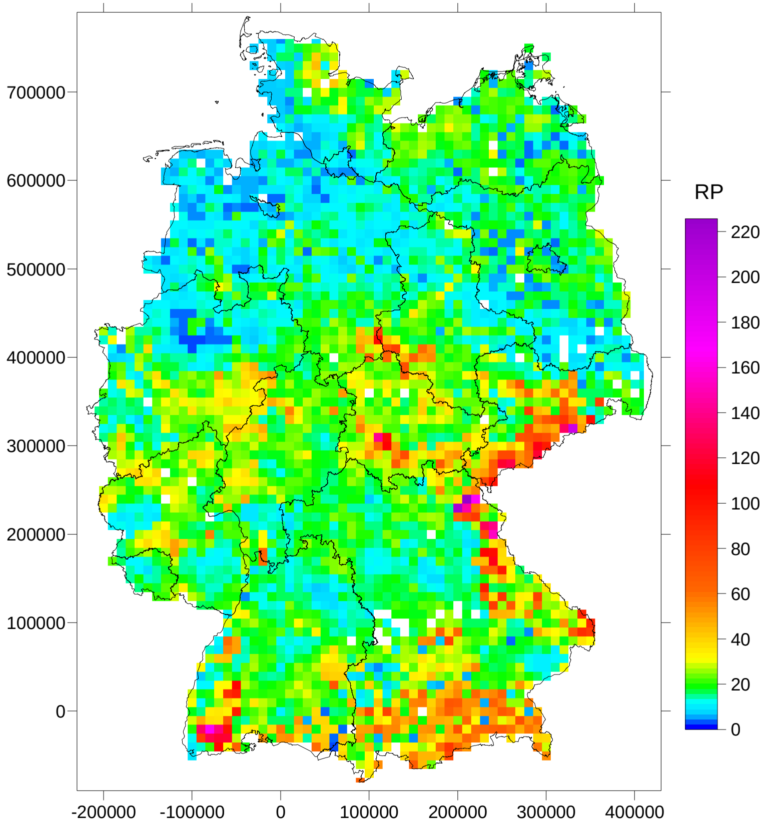
- One may ask:
  Why not use indoor Rn directly?

- Practical answer: In Germany indoor Rn data are not suitable (clustered, representativeness questionable). RP data better $\rightarrow$ used as predictor
- Theoretical answer: RPA derived from indoor C depends on state of building stock: temporal variability; geogenic RP = constant, independent of anthropogenic factors

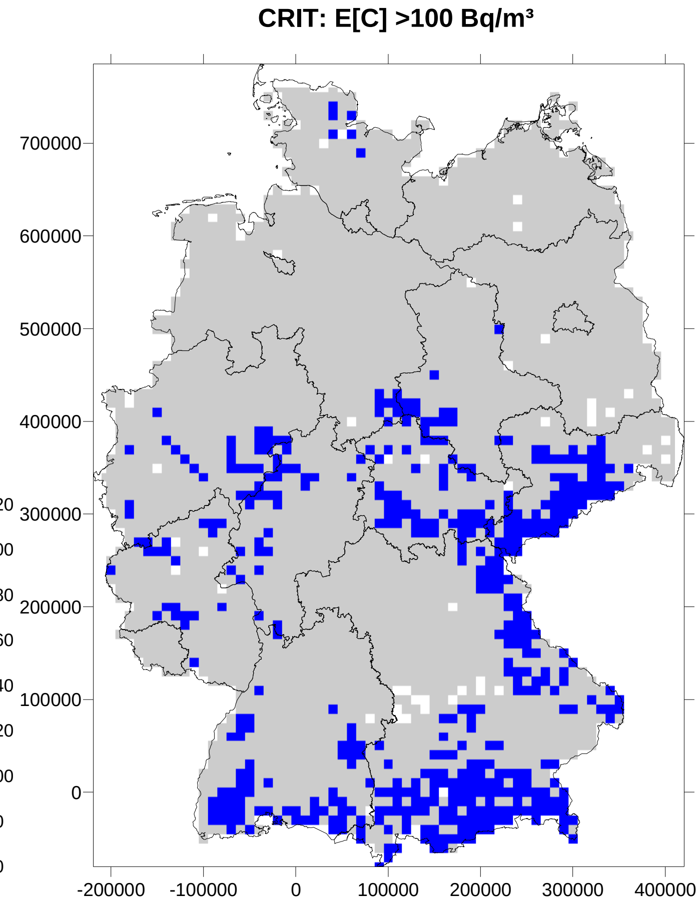
Example 1: Germany

predictor = RP,

CRIT =
E[C_{indoor} in cell]>100 Bq/m³



predictor map,
10 km × 10 km cells



Data:
3745 values RP,
15563 values indoor Rn, living rooms in ground floor, buildings with basement.

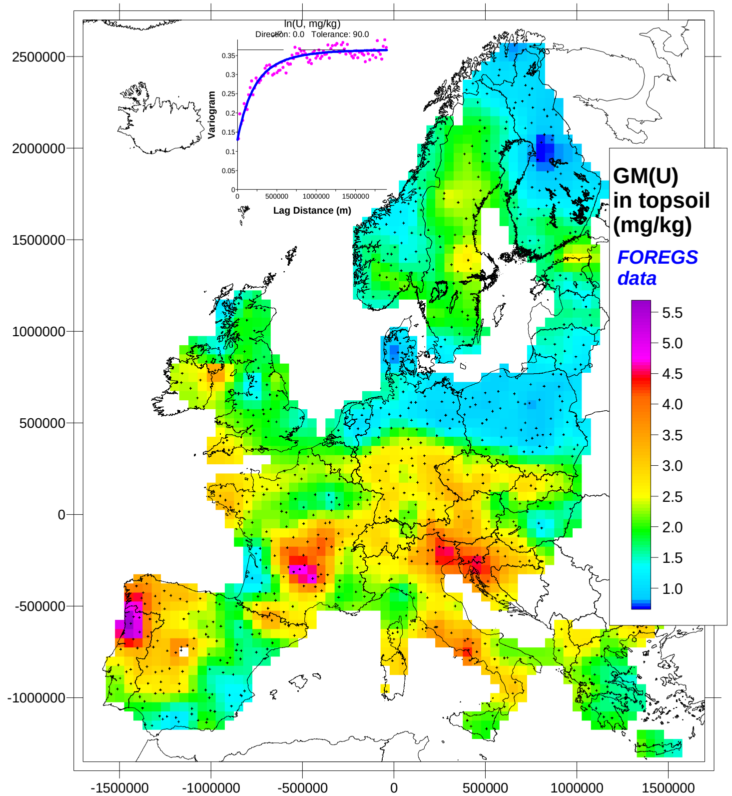
RPA map, RP_0=32

remark: E(C)>100 in cell ⇔
prob(C>100)≈38%, about 8.4 times the same probability over whole Germany;
prob(C>300)≈1.8%. about 5 times the same probability over whole Germany.

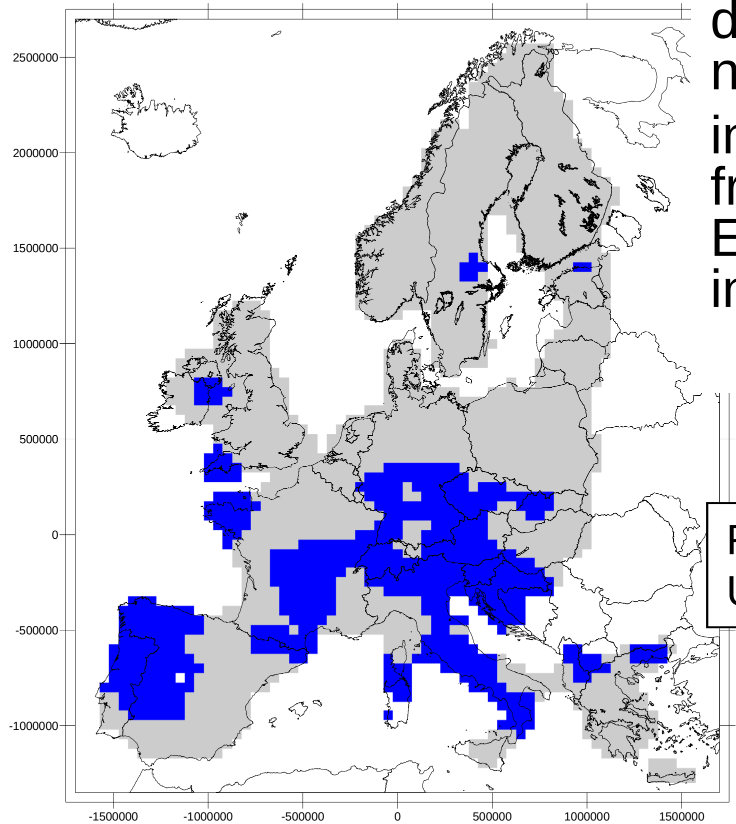
Example 2: Europe

<u>predictor</u> = U concentration in topsoil,
<u>CRIT</u> = E[C_{indoor} in cell]>100 Bq/m³,
cell=50 km × 50 km

<u>Data:</u>
U: from the FOREGS database n=843,
indoor Rn: from European indoor map



predictor map,
50 km × 50 km cells



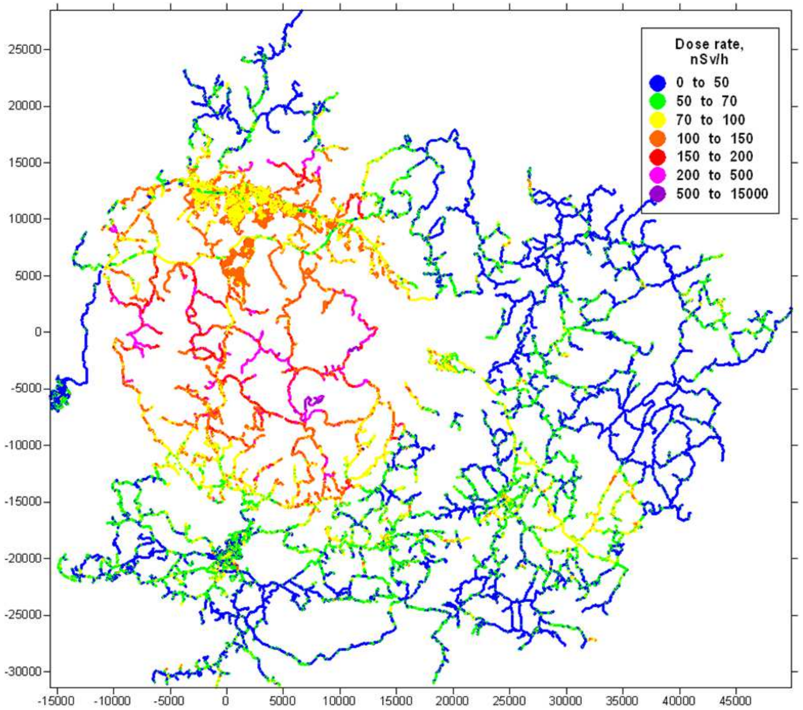
RPA map,
U_0=2.48 ppm

Example 3: Poços de Caldas, Brazil

<u>predictor</u> = ambient gamma dose rate,
<u>CRIT</u>= E[C_{indoor} in cell]>100 Bq/m³,
cell=1 km × 1 km

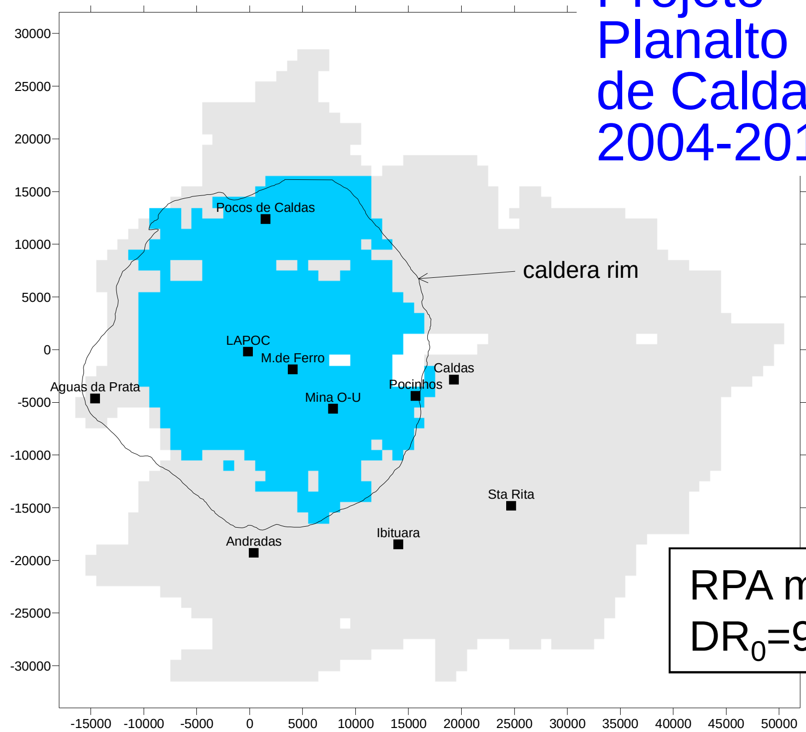
<u>Data:</u>
Doserate: 520,000 carborne measurements

indoor Rn: 340 indoor Rn data, living and sleeping rooms in ground floor



predictor data

Projeto Planalto Poços de Caldas, 2004-2013

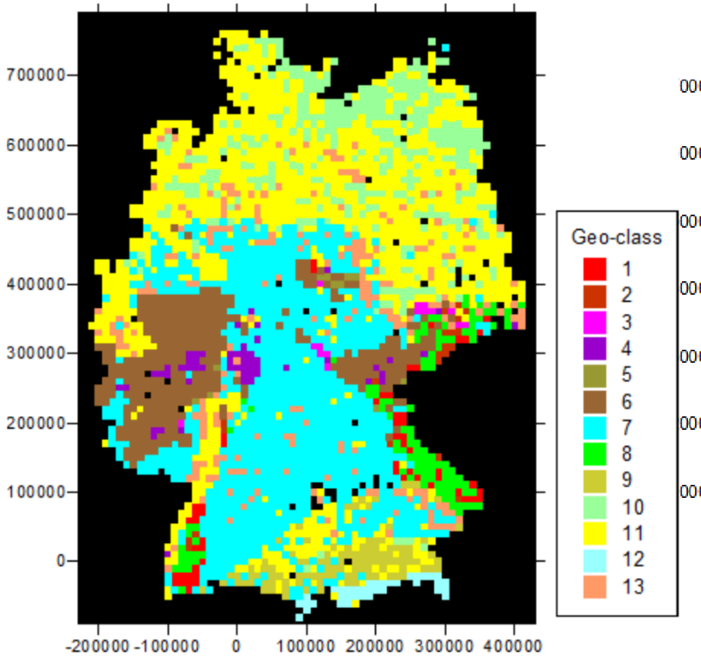


RPA map,
DR_0=92 nSv/h

Radon prone geologies

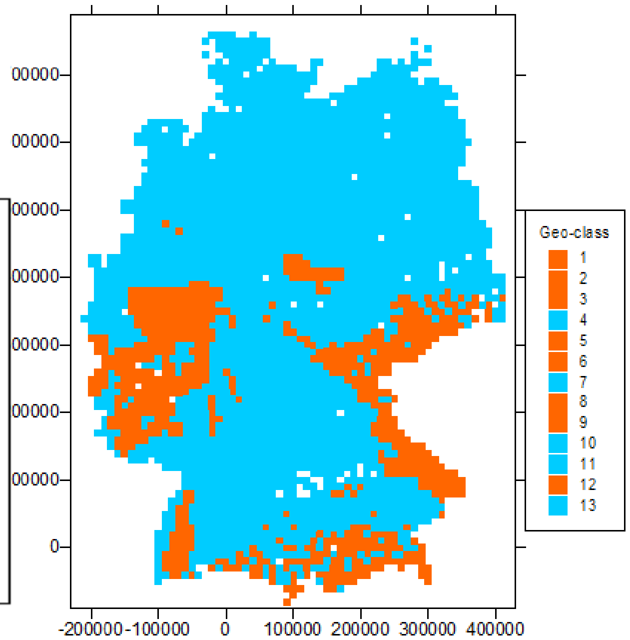
- **Extension of the method:**
  Predictor: geological units (transformed into continuous variable by combinatorial method – see article, subm. JER)
- Given: a number of geological units; separate them optimally into 2 groups to match the calibration criterion CRIT.

predictor map:
geological map,
13 generalized units

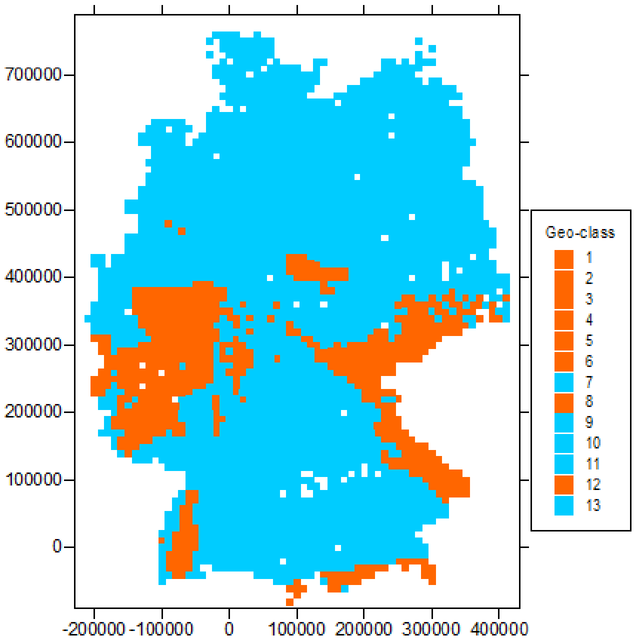


2 maps of Rn prone geologies (RPG), according 2 different calibration criteria; orange: RPG

CRIT3: GM(RP)=32; score=Y



CRIT2: prob(C>100;m0=5)>0.1; score=Y



Conclusion

- Method is computationally very simple
- But has its caveats!
- Result = threshold of the predictor (RP_0 etc.) the more reliable, the better the correlation between predictor quantity (RP) and calibration quantity (indoor Rn)
- Several “political” and “technical” decisions required
- <u>Missing:</u> uncertainty budget, significance test
- <u>Planned:</u> extension to
  - multinomial classification
    (several levels, not only RPA yes / no)
  - multivariate predictor
    (e.g. RP & U-conc. & dose rate)