Radon priority areas and radon extremes

- an initial study -

Peter Bossew

German Federal Office for Radiation Protection (BfS), Berlin

ICHLERA 2018

9th International Conference on High Level Environmental Radiation Areas

September 24-27, 2018, Hirosaki University, Aomori, Japan

v. 26.9.18

Content

- Rationale:
  - Importance of indoor radon
  - Legal background
- Definitions of Radon priority areas
- Extremes, anomalies, outliers:
  Causes and statistics
- How to detect and to model anomalies?
- Examples

Very short introduction

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.
(incl. RU, TR; missing: BiH, LV, MD, MK, MT, RS, UA) *(figure appears a bit overestimated to me)*
Japan: estimated ca. 3,100 (Gaskin et al., Envir. Health Perspectives 125, 5 (2018))

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 / EU

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)


Official Journal of the European Union

L 13

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

Price: EUR 4

EN

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 Bq/m³ (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., $AM_B(C)\times$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".

Metro RADON

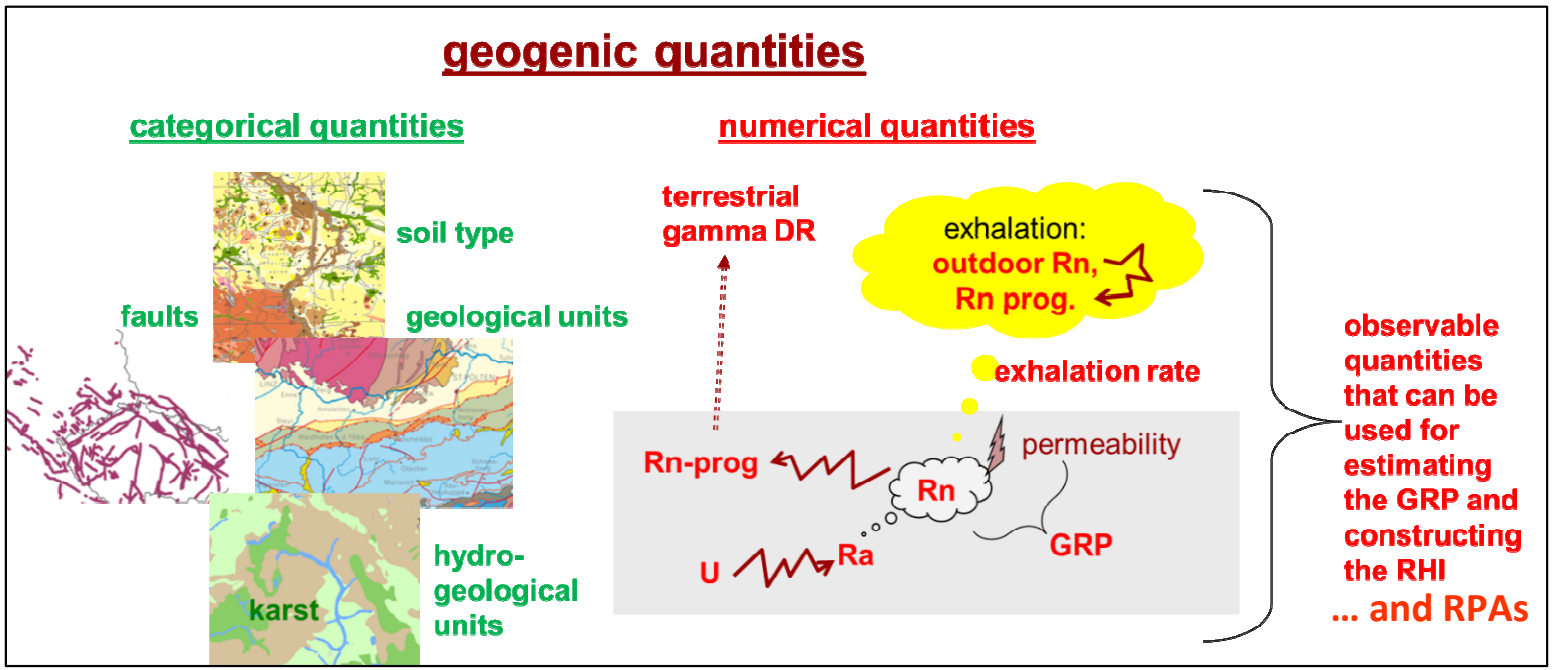
RPA definitions, 2

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

The problem

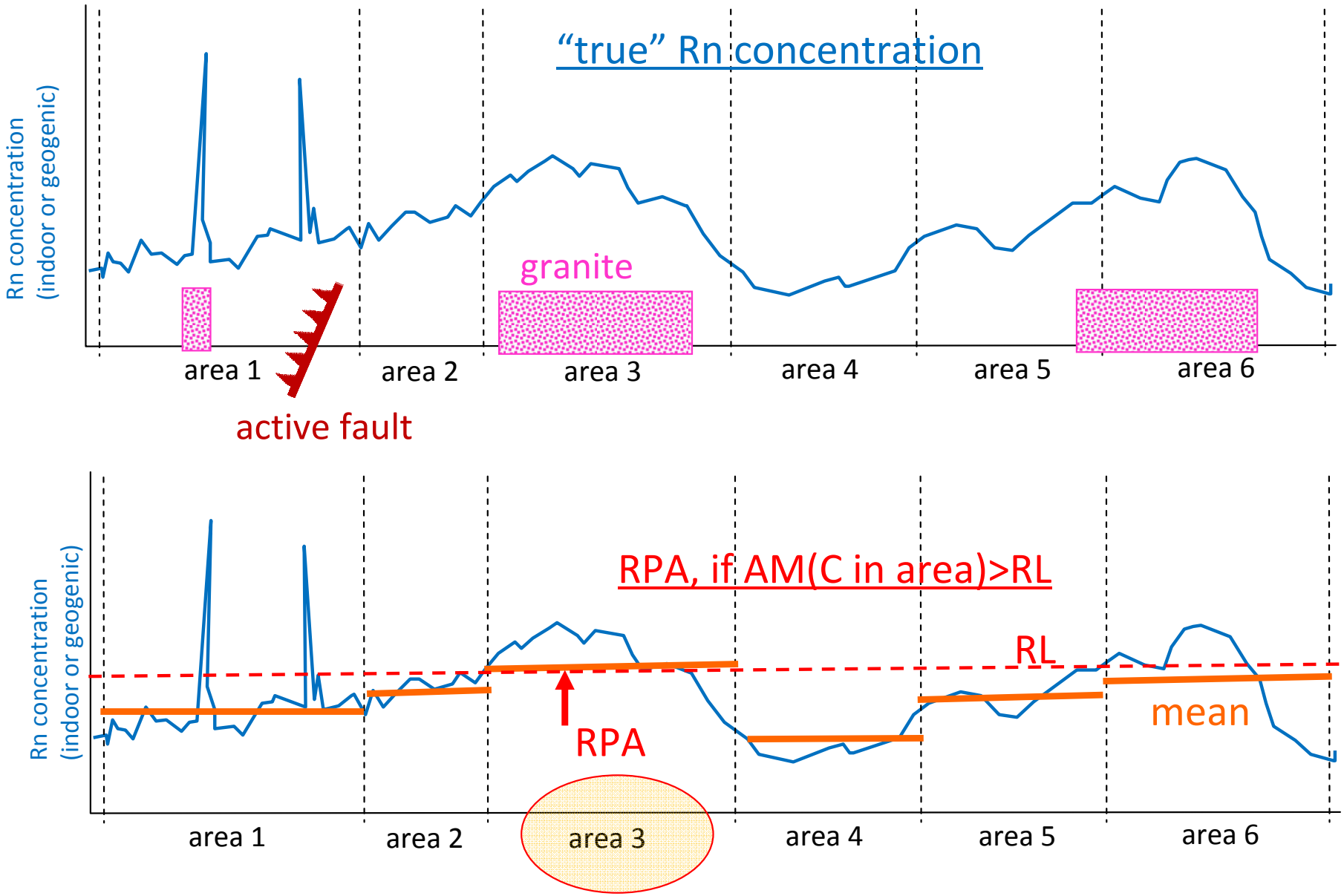
- Rn concentrations are spatially very variable. Reason: the variability of the physical factors which cause Rn concentration.
- Therefore, also in an area labelled non-RPA, i.e. “harmless”, small sub-areas can still have high Rn concentrations.
- Occurrence probability of small extremal areas should be quantified – possibly a complementary RPA criterion.
- Usual RPA criteria (last 2 slides) cover “most” of an area, but also houses in small “anomalous” areas deserve attention!

The task

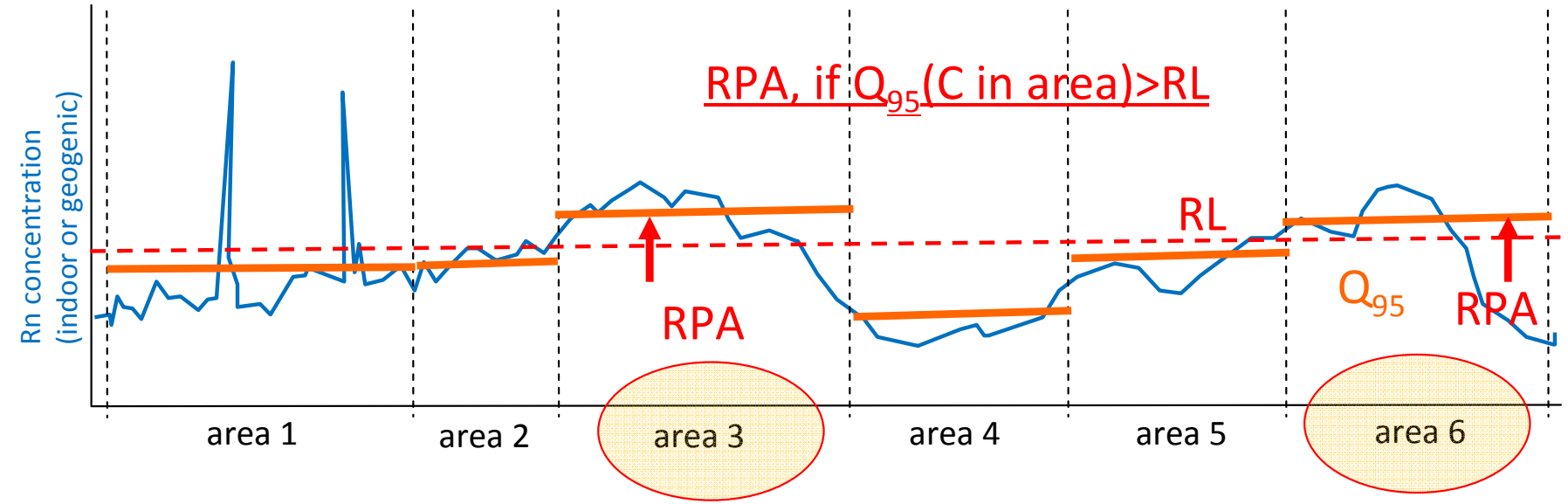
Find indicators of the presence of spatially small extremes
- *from data*
- *via models*
- *from physical predictors*

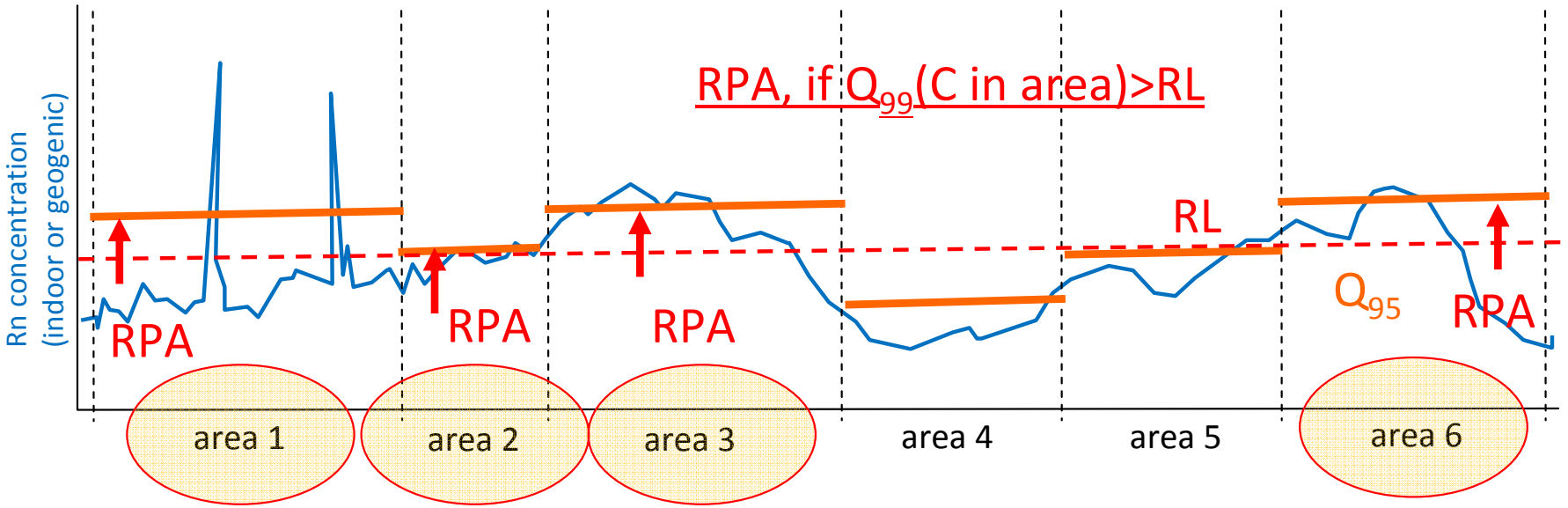
An artificial example

background and anomalies, 1



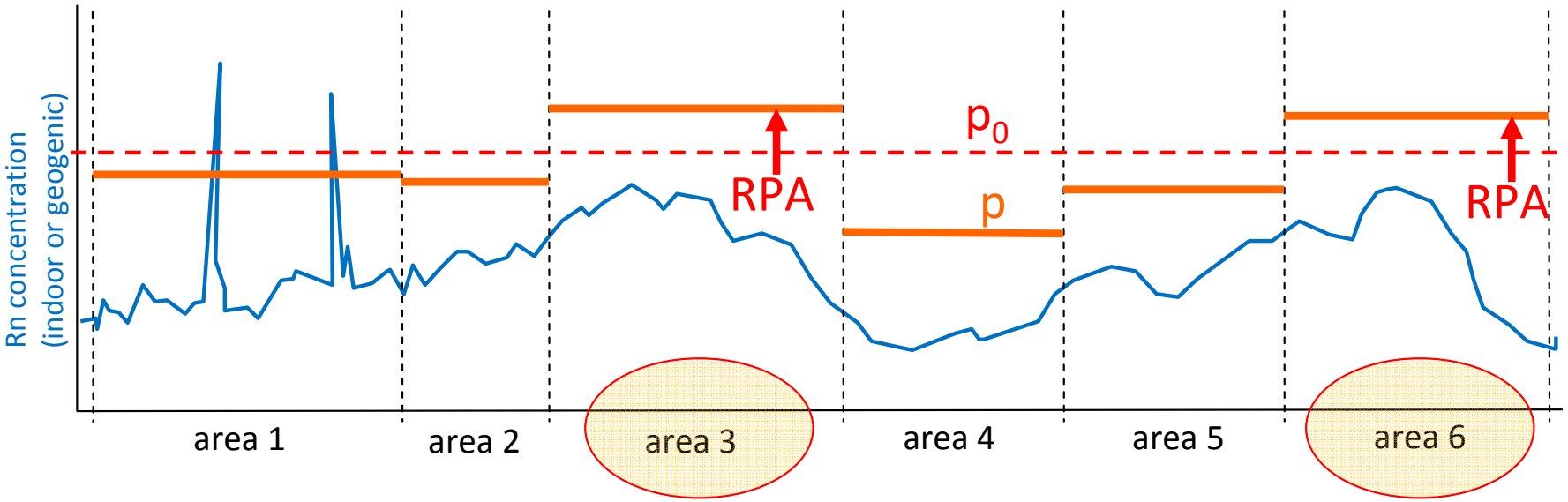
background and anomalies, 2



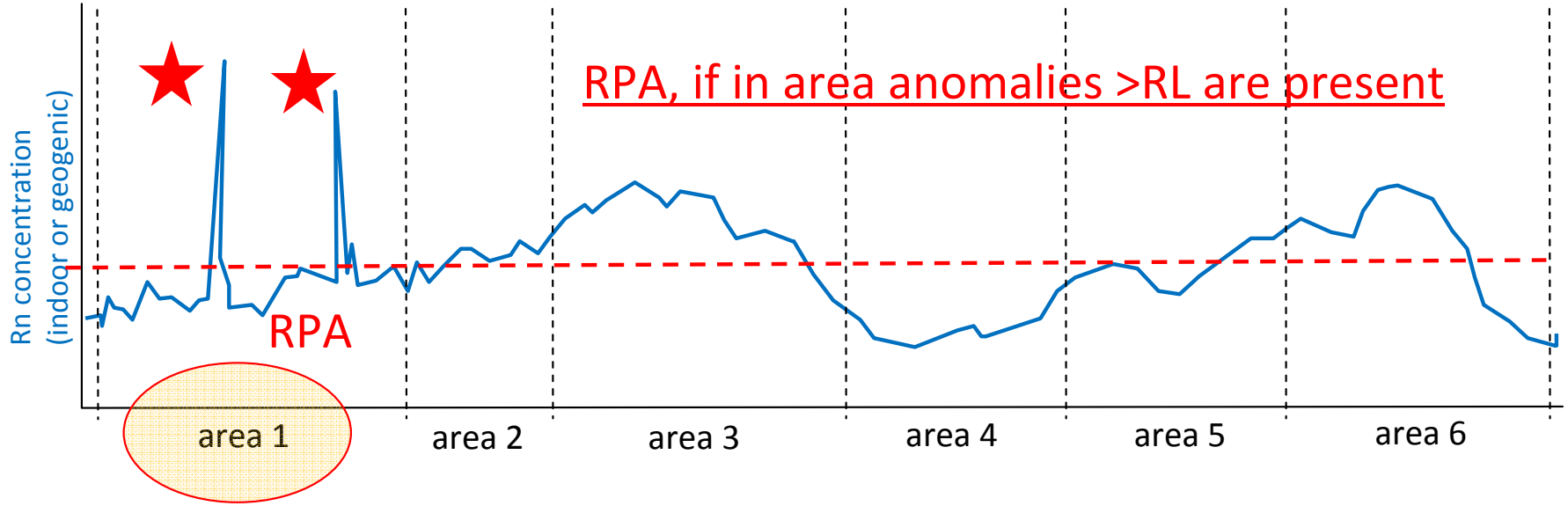


background and anomalies, 3

RPA, if p=prob(C>RL in area)>p_0



RPA, if in area anomalies >RL are present



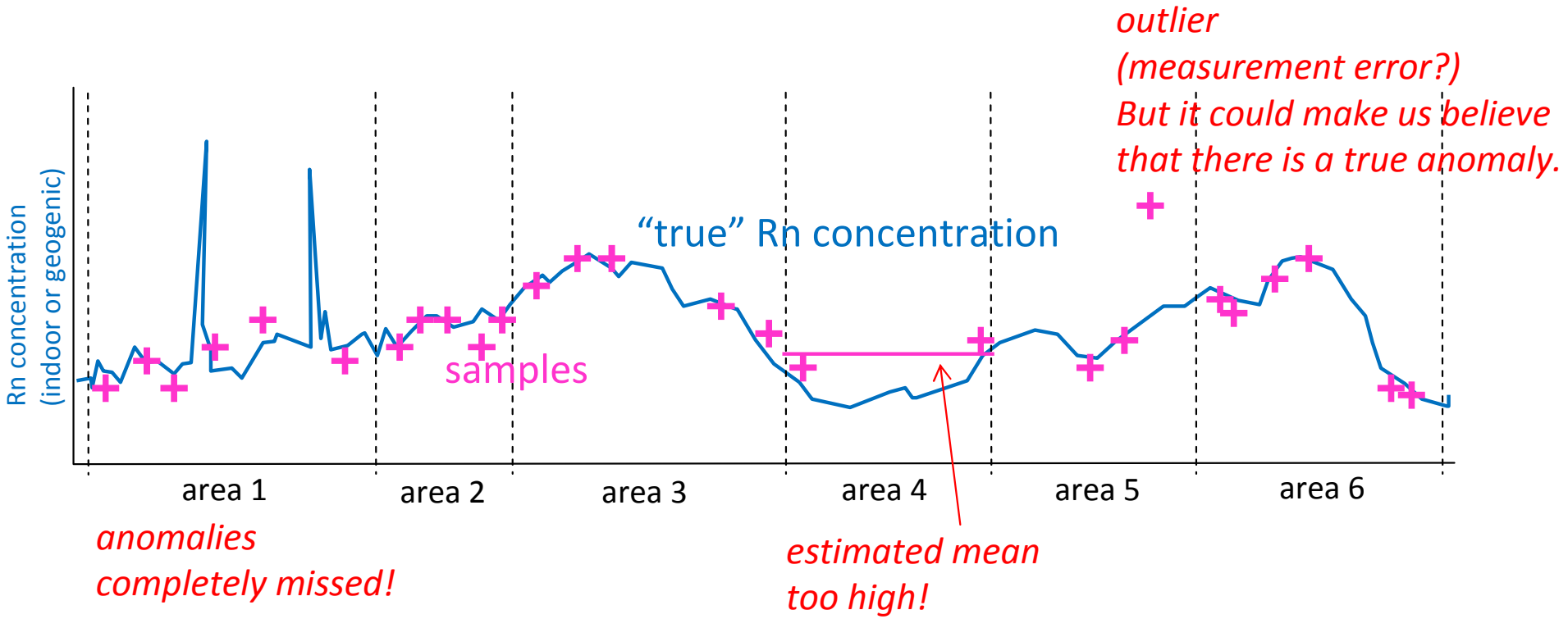
Some concepts

& little bit physics

Physical causes of small anomalies

- **Geological** anomalies, e.g. small granite outcrops
- **Secondary U mineralization**:
  hydrothermal activity, contact metamorphism,…
- Active **faults**
- **Karst** formations (high macro-permeability)
- Anthropogenic: **mining** residues,…

truth and sample



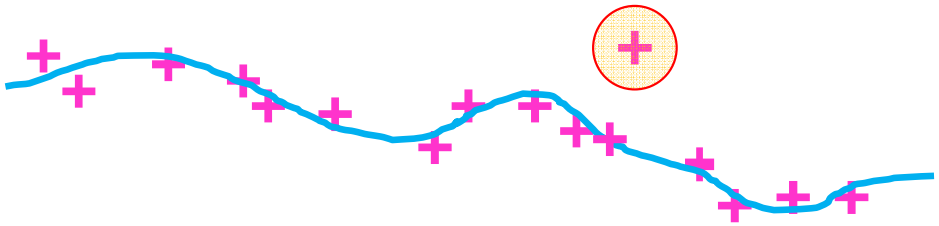
Questions:

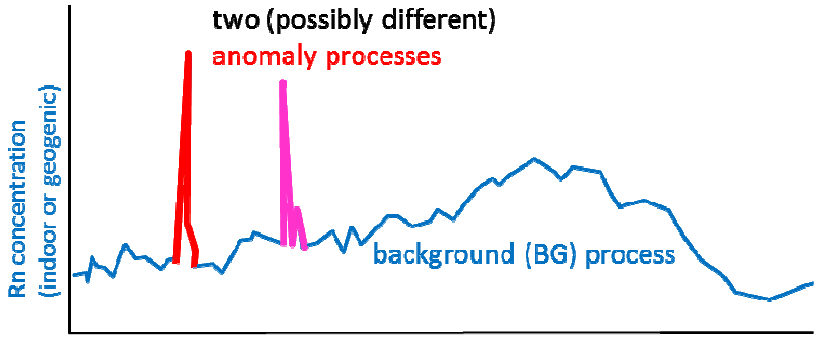
How can one estimate statistics from limited and uncertain sample values?

- AM is known to be sensitive against outliers
- High quantiles or exceedance probabilities of high RL are difficult to estimate from data
- "Rare" events are likely be missed
- Opposite: Outliers can represent "spurious" anomalies
- After all, what is an "anomaly" ?

What is an anomaly?

- Concepts of “anomaly” and “background” cannot be separated!
  Definition of anomaly depends on definition of BG and vice versa.
- Qualitatively:
  Anomaly is a value which is significantly different from its vicinity.
- Identifying anomalies:
  - residuals from interpolated smooth surface





  - high difference from neighbours
  - assuming a distribution within a vicinity → outliers of the distribution
  - high Hölder exponent α

    from fractal theory: $z(A) \sim |A|^{\alpha}$ ($z(A)$ = mean over A, $|A|$=area of A) → $\alpha = \lim_{|A| \to 0} \log z(A) / \log |A|$

Do not confuse an anomaly with an extreme of the background!
Anomalies and BG extremes represent different (physical and statistical) processes!

Terminology

Outlier: a value which one suspects not belonging to an assumed population.

Reasons can be:
- observation error or uncertainty
- an accidentally isolated extreme of the population
- an instance which belongs to a different population

Extreme: The highest or lowest value of a set. Does not say anything about its nature.

Anomaly: a point or a region, at which the variable behaves differently from most of the data domain or of its neighbourhood, or simply being "too large" (or too small) than is considered typical for that part of the field.

An anomaly may be an outlier, but does not need to be one necessarily.

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
- A region which is the domain of a separate process.

Distinguishing these can be quite complicated!

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... but terminology is not completely unambiguous!

Some mathematics

.... modelling

Often data are not sufficient for estimating occurrence of anomalous events.

On the other hand, establishing models from which statistics are derived, can also be sensitive to uncertainty.

Popular model for Rn: Lognormal distribution (LN) – Requires either

- estimating geometrical mean (GM) and standard deviation (GSD) from data, GM=exp(AM(ln Z)), GSD=exp(SD(ln Z)); or

- establish QQ-plot with empirical quantiles and determine GM and GSD by regression.
  This also allows checking whether LN is (approximately) valid.

Then, quantile $Q_p=\exp(\Phi^{-1}(p|GM,GSD))$

exceedance probability: $\text{prob}(Z>z)=\Phi((\ln GM - \ln z)/\ln GSD)$

Φ=standard normal, more precisely to be replaced by t_{n-1}, n=sample size.

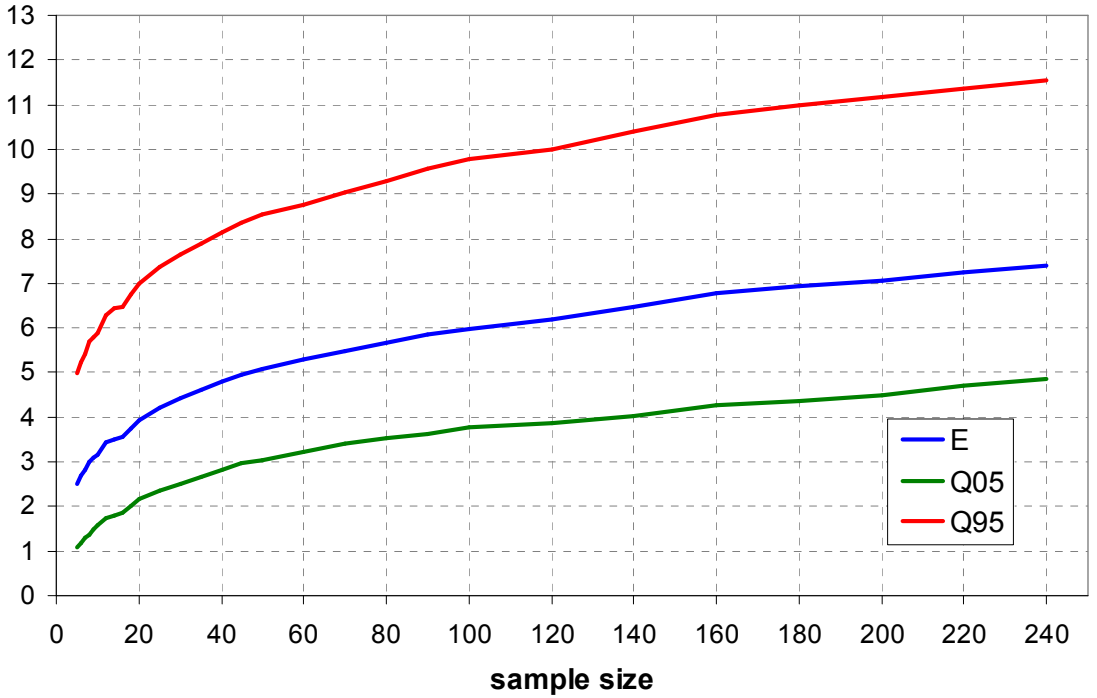
Big problem: One cannot assume LN for anomalies on top of background!

Establishing a BG model is relatively (!) easy. This helps us determining extremes of the BG process, but says nothing about the anomaly process.

distribution of maxima, 1

Let Z a random variable ~ LN with GM=1 and GSD=2.
Take n independent random samples $\{z_1,\ldots,z_n\}$
Which is the maximum value of these?

Graph (based on 100,000 simulations):
Expected maximum $E(z_{max})$,
90% confidence interval
(Q05 – Q95)

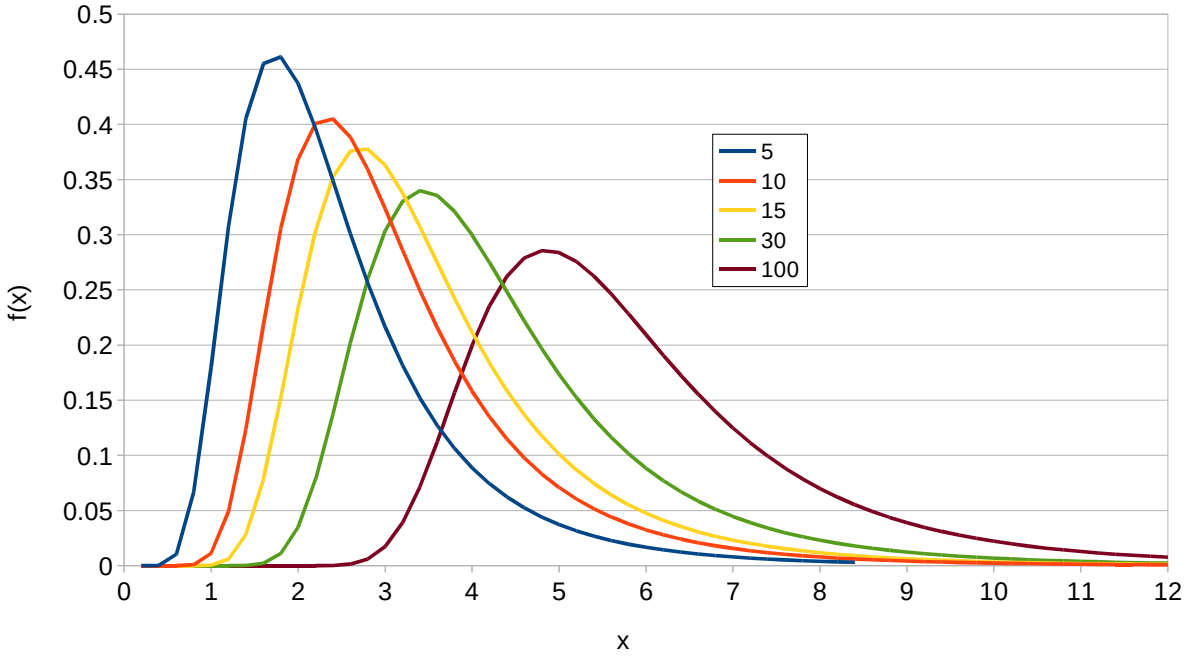


If we have a sample of size n and use a model for the BG (such as LN), and if we find a value $z^{\#}$ which is above the expected maximum,
$z^{\#} > E(z_{max})$ or $> Q95(z_{max})$ $\Rightarrow$ **candidate for anomaly!**

distribution of maxima, 2

The maximum $z_{max} = y_n := \max\{z_1,...,z_n\}$ is exactly distributed $\sim F_Z(y_n)^n$ for $Z \sim F_Z$; but difficult to handle

From extreme value theory it is known that for F=LN or similar,
y_n is asymptotically distributed
$\sim \exp(-\exp(-(y_n-\mu_n)/\sigma_n))$
(Gumbel or EV type1 distribution … Fisher–Tippett–Gnedenko theorem),
mean $=\mu_n+\gamma\sigma_n$, SD$=\sigma_n\pi/\sqrt{6}$,
γ=Euler-Mascheroni constant = 0.57722...;
μ_n and σ_n are (complicated) functions of parameters of F_Z and sample size n.
However, converges slowly; better: fit simulated by power function
(in the above example $r^2>0.995$)



pdf of maxima of LN(0,ln(2)) for different sample sizes

Q_p easy to calculate: $Q_p(Y_n)=F_Z^{-1}(p^{1/n})$

Dilemma!

If we estimate directly from data, i.e. without model assumptions (such as LN):
perhaps we miss important but spatially small phenomena because of insufficient amount of data;

If we use a model, derived from data and possibly additional knowledge (in Bayesian statistics: prior), and define anomalies as deviations from the model:
- the choice of model assumptions may be wrong and
- its parameterization may be inaccurate.

Possible pragmatic way out ?

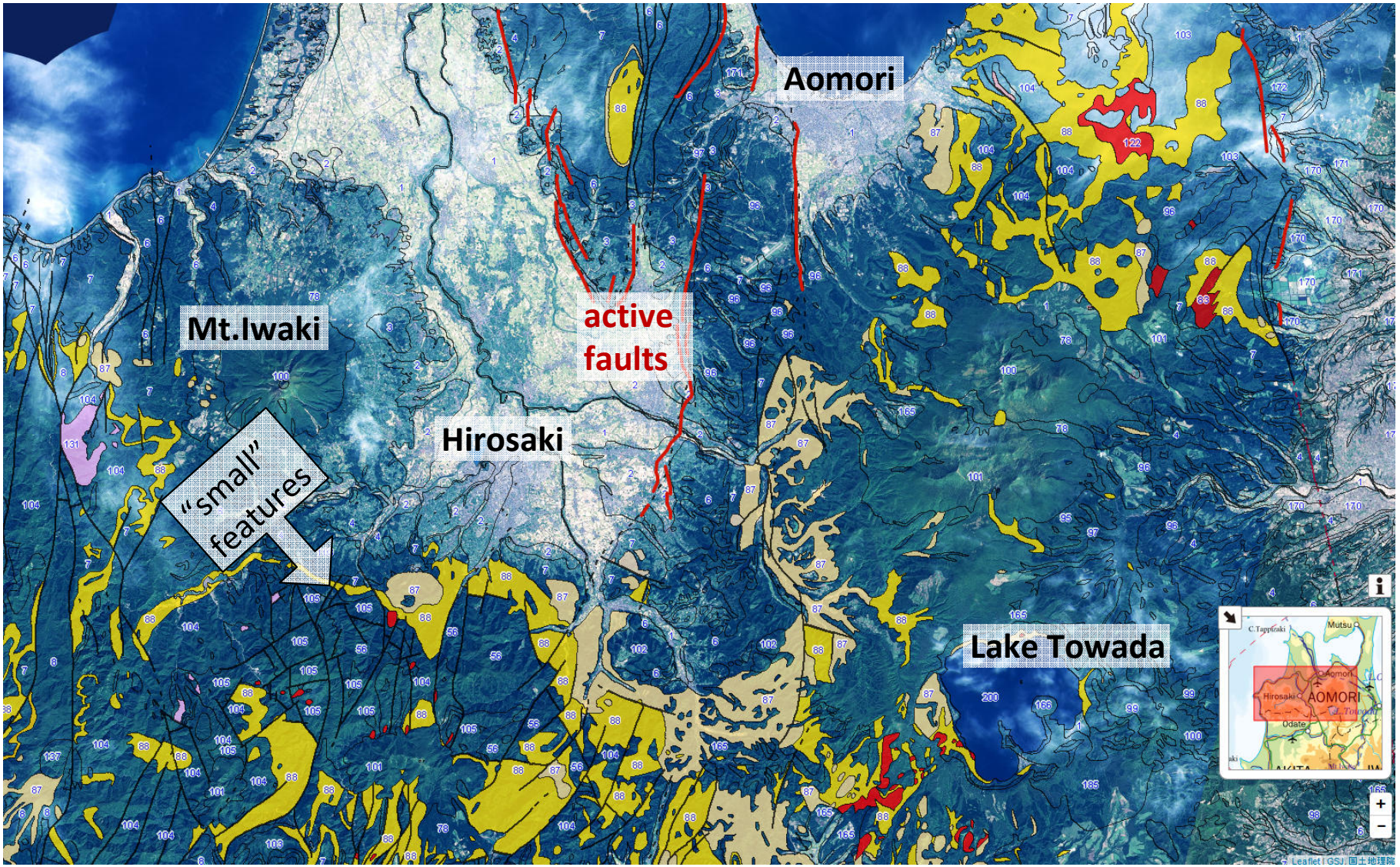
Use all available approaches together and decide from "expert knowledge":

- Inspect data and derive empirical statistics;
- Use a BG model and look for residuals and outliers;
- **Look for physical indications:** certain geological (lithology and hydrogeology) and tectonic features.

But the BG model alone will not inform us about occurrence and size of anomalies ... limit of modelling!

Real examples!

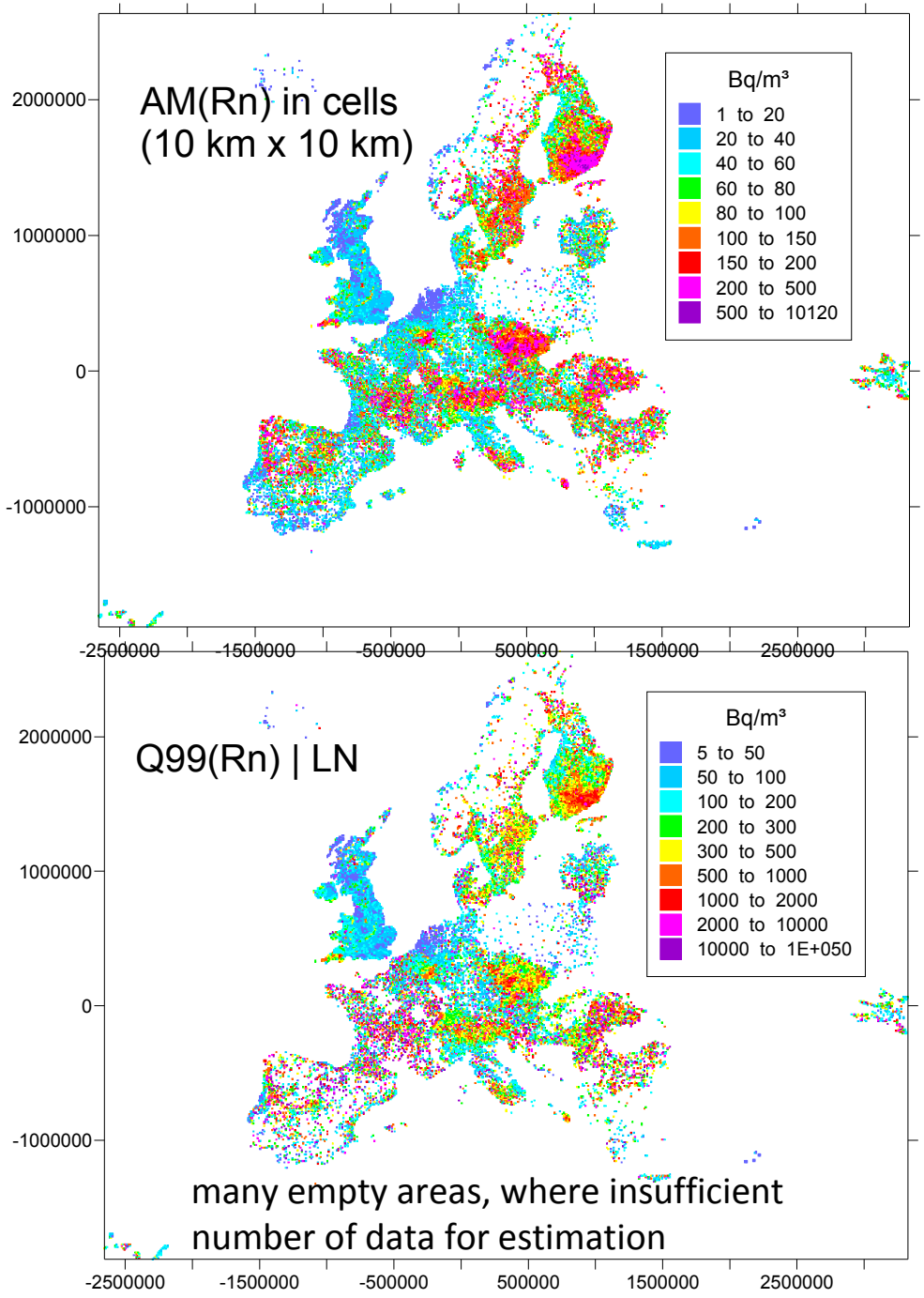
Example 1: Hirosaki region



red: felsic (acid) plutonic; yellow: felsic volcanic

Source: Geological Survey of Japan
https://gbank.gsj.jp/seamless/seamless2015/2d/index.html?lang=en

Example 2: European indoor Rn map



Empirical mean

AM(indoor Rn concentration) in 10 km x 10 km cells; ground floor rooms

ca. 28,000 non-empty cells,
ca. 1.1 mill data, 34 countries
(by Aug 2018)

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

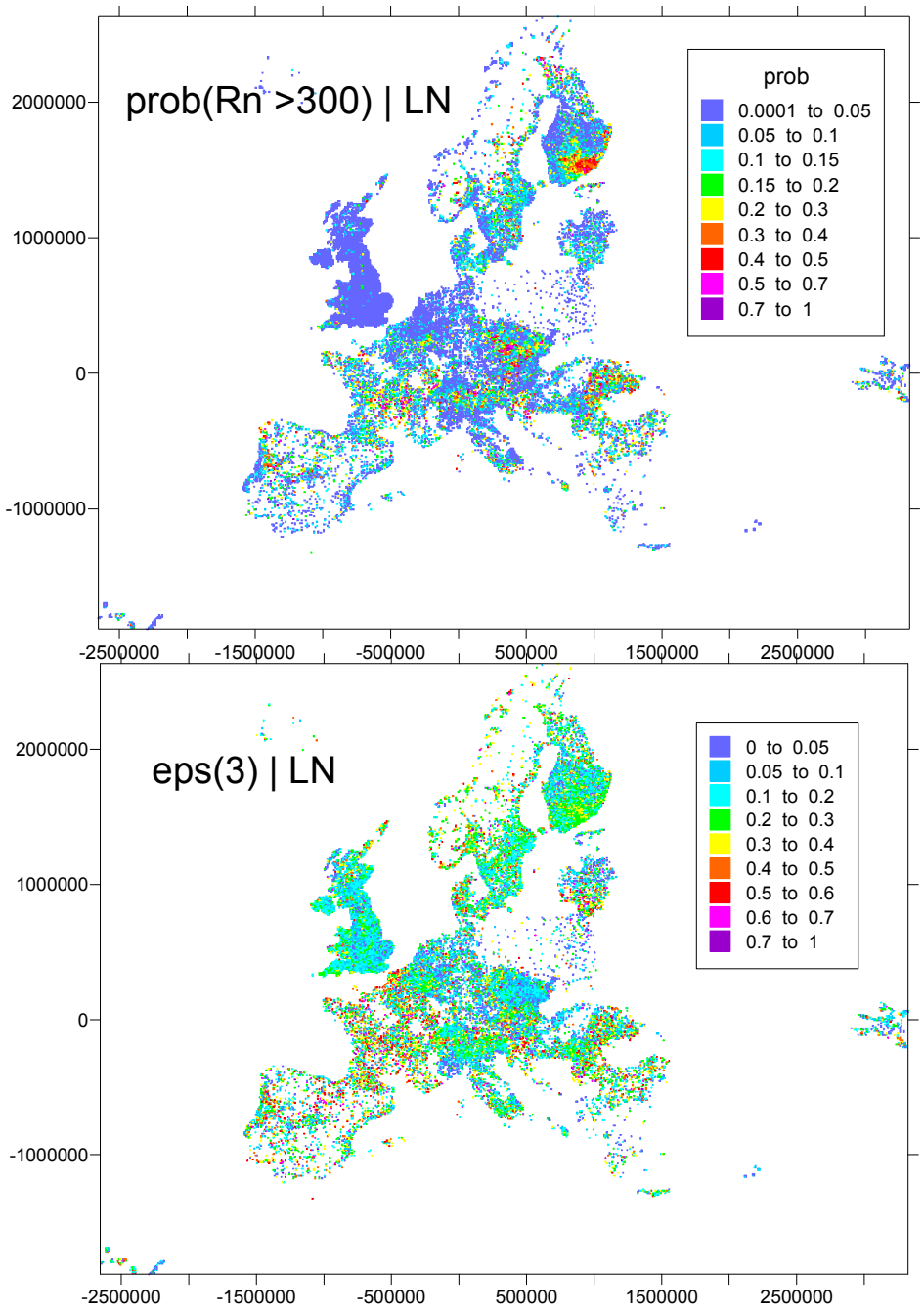
Estimated high quantile

$$\widehat{Q_p} = exp\left(t_{p;n-1}SDL + AML\right)$$

n=number of data in cell
AML:=AM(ln z)
SDL:=SD(ln z)

LN model with observed AML, SDL assumed!

Example, cont.



Estimated exceedance probability

$$p = \text{prob}_B(Z > z); \quad \hat{p} = t_{n-1}\left(\zeta \sqrt{\tfrac{n}{n+1}}\right); \quad \zeta := \frac{\ln z - \text{AM}_B(\ln Z)}{\text{SD}_B(\ln Z)}$$

LN model with observed AML, SDL assumed!

Estimated relative excess probability

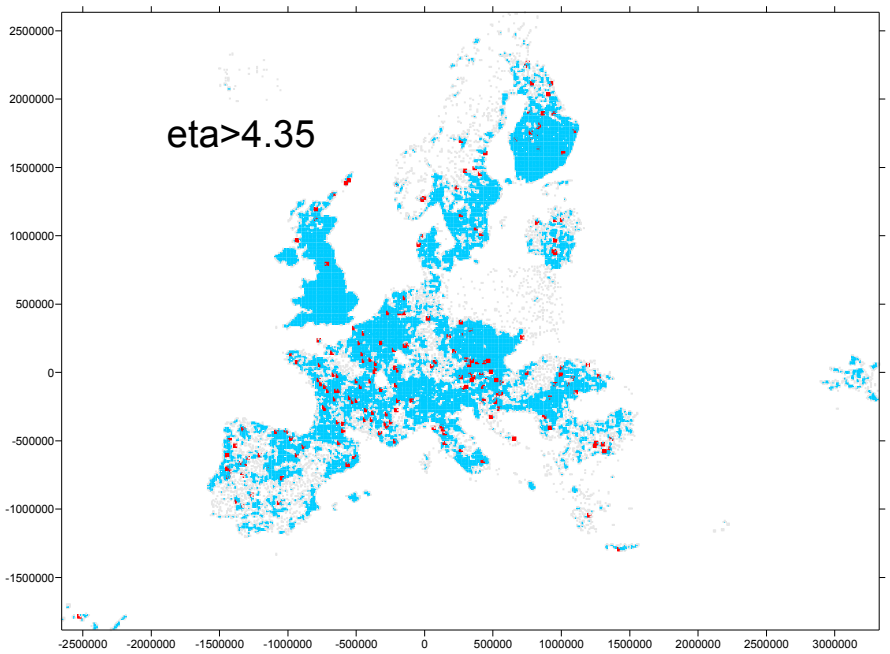
Probability that k × median is exceeded, relative to median

$$\varepsilon_k := \frac{prob(Z > k \cdot Med)}{prob(Z > Med)} \longleftarrow =0.5$$

$$\hat{\varepsilon_k} = 2\left[1 - t_{n-1}\left(\sqrt{\frac{n}{n+1}}\,\frac{\ln k}{SDL}\right)\right]$$

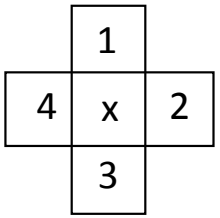
LN model with observed SDL (only) assumed!

Example, cont.: fractal indicators



η:= z(cell)/AM(z, neighbouring cells)

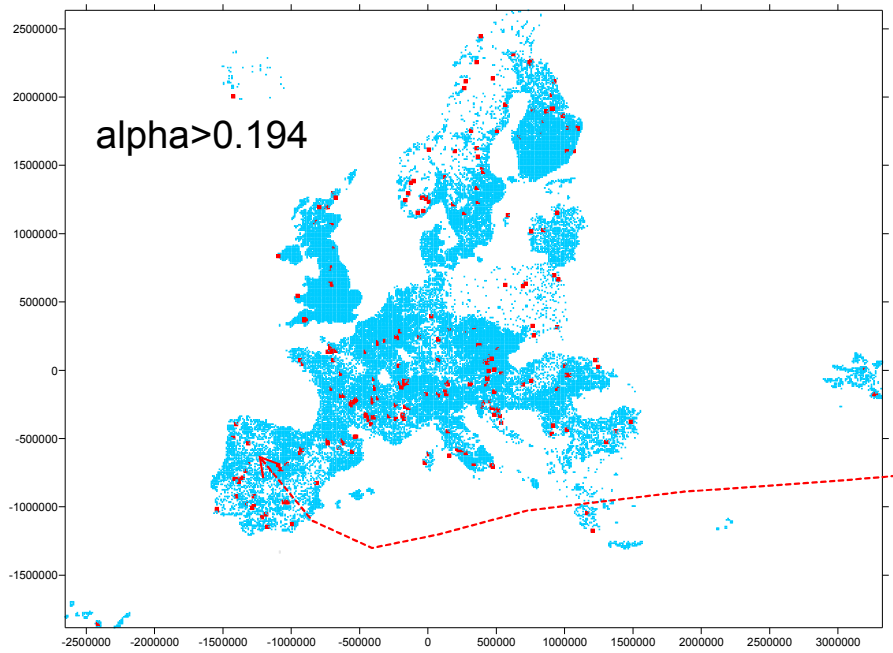
Q99(η)=4.35



4-neighbour rule applied, at least 3 non-missing neighbours

No model assumed!

Only from observed data

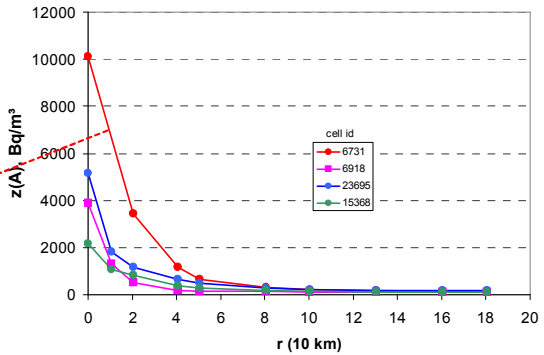


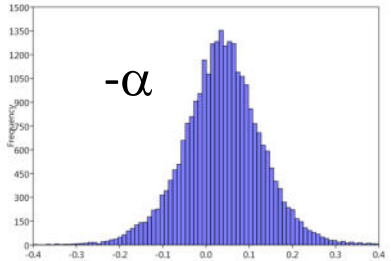
$$\alpha := -lim_{|A|\to 0} \frac{log\ z(A)}{\log |A|}$$

z(A)=mean over A,
|A|=area of A

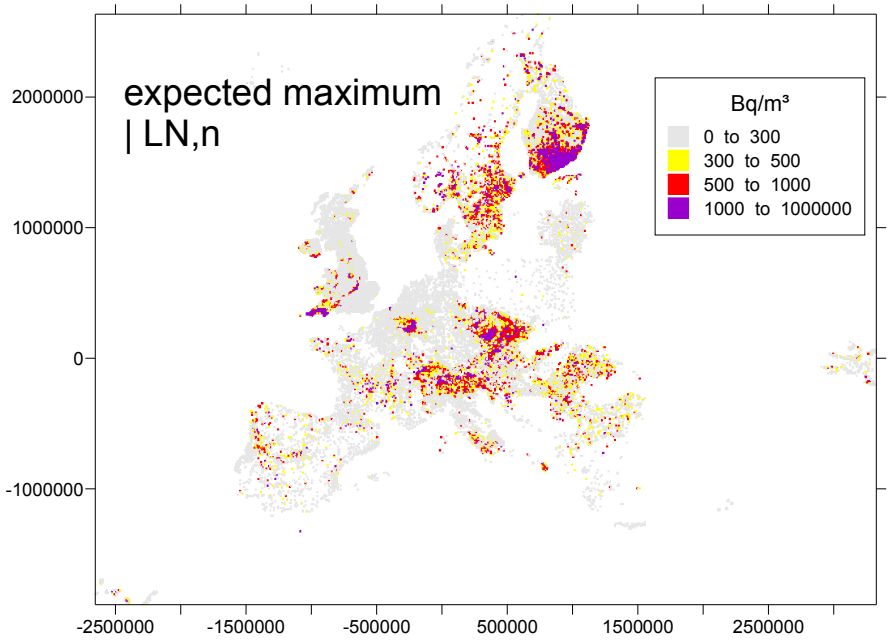
estimated by defining A as shells of different radius around a cell, then regression

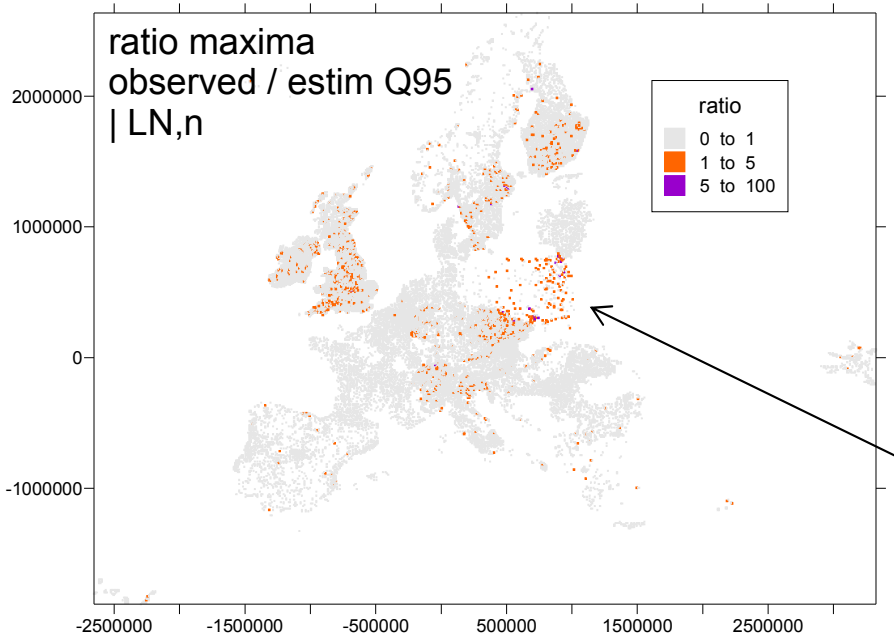
Q99(α)=0.194





Example, cont.: maxima





<u>Expected maximum</u>,
assuming LN (from data)
and given sample size n

Additional problem, not to be discussed here:
Spatial interpolation of extremes for generating "map of extremes" is mathematically demanding.

<u>ratio between observed maximum and estimated 95% quantile.</u>

If ratio > 1:
- indicator for anomaly
- or: distribution assumed wrongly

data errors?

Finally…

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!
- RPAs can be very small in space, due to small geological or tectonic features. They may not contribute much to the overall mean, but may still be locally relevant.
- Presented here: some initial ideas....
- Often not or insufficiently covered by data -
  Identification, quantification and modelling not easy mathematically!
- To do:
  - Further develop mathematical tools
  - Explore validation possibilities
  - Uncertainty in estimation of “small” but extreme phenomena
  - Once anomaly quantification achieved: define RPA criterion
  - **Develop a generic strategy to deal with Rn anomalies!**

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.

Metro RADON