Towards multivariate modelling of geogenic radon

(parts 1 & 2 together)

P. Bossew[1], W. Meyer[1], M. Bleher[2]

Bundesamt für Strahlenschutz / German Federal Office for Radiation Protection,
[1]Berlin, [2]Munich

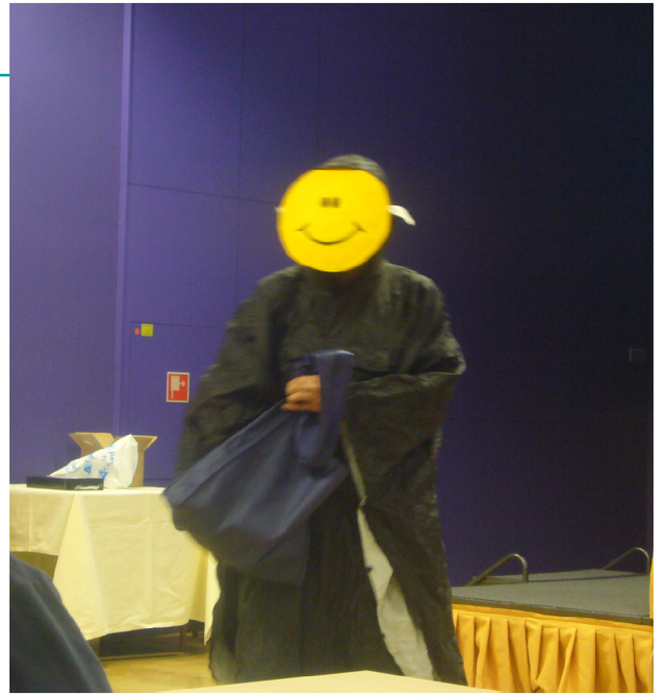
10th INTERNATIONAL WORKSHOP
on the
GEOLOGICAL ASPECTS OF RADON RISK MAPPING
22 – 25 Sept 2010, Prague, Czech Republic

v. 21 Sept 2010

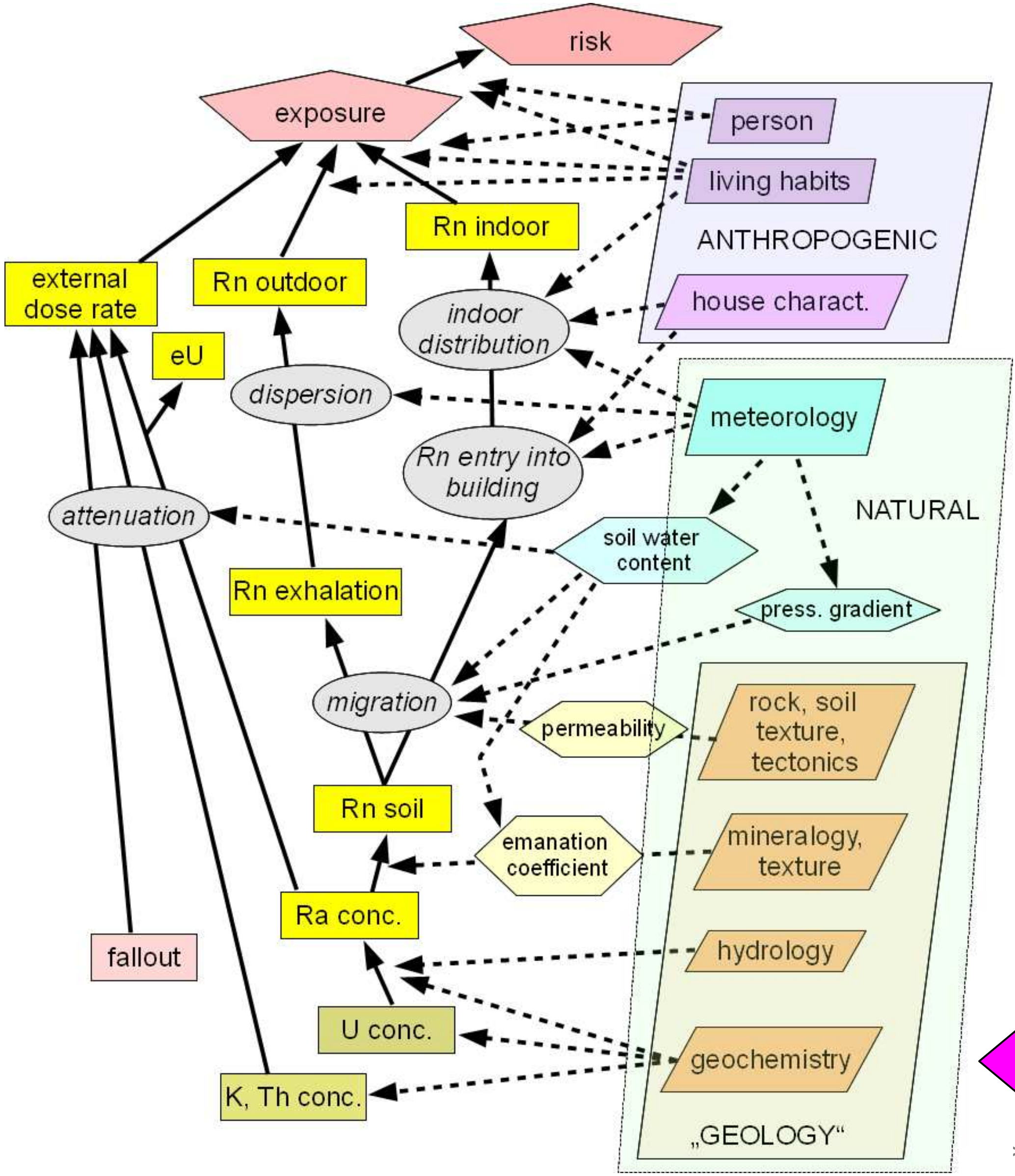
Bundesamt für Strahlenschutz

Content

- Reminder: physical complexity
- Concepts
  - types of variables
  - construction of target variable
- Regression modelling
- Correlation between variables
- Geostatistical modelling of residua
- Target variables
- Examples



Physical complexity



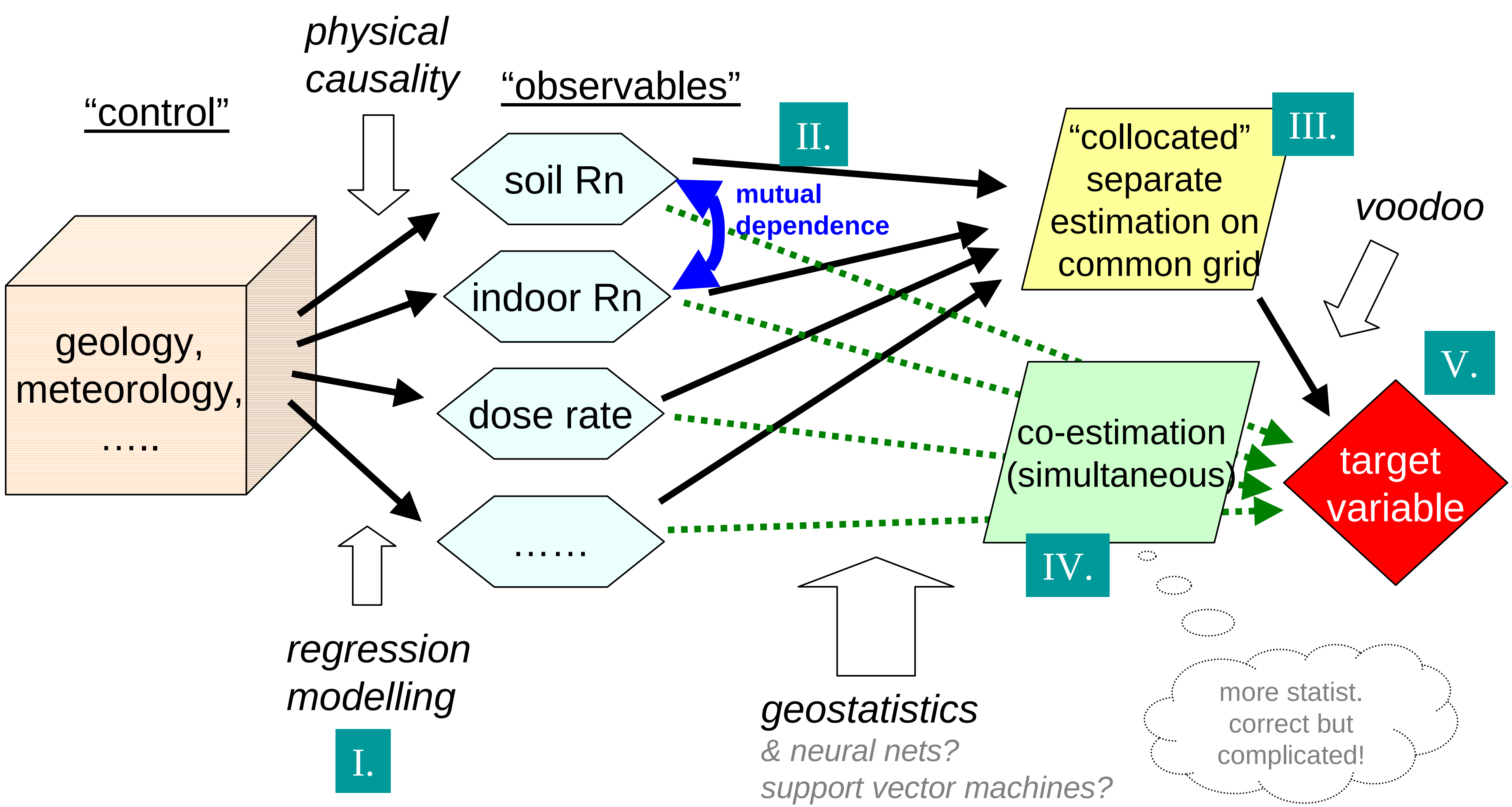
- complex pathway from Rn source – migration – exposure and risk
- some variables not trivial to define accurately
- spatially complicated structure:
  - high nugget
  - hot spots
  - geological predictors not easy to define

- **simplified!**
- **yellow:** some observed Rn variables

Wanted

- spatial prediction of a variable which quantifies Rn hazard
  - its levels
  - its support (“Rn prone areas”)
- <u>hazard</u>:
  - define “hazard” variable Y
  - estimate / predict from *observable quantities*
  indoor concentration, conc. in soil air, external dose rate,…
  - and from observable *physical controls*
  geology, soil properties, geographic location…
- **¿¿** $E[Y(x)](x \in U)$, $prob[Y(x)>T](x \in U)$ **??**
  U = prediction support (point, cell, admin. unit,…)

Concept & road map I. to V.



I. regression modelling

General linear model (GLM)

$$Z(U) = \mu_k(U) + \alpha_{kl}(U) + \beta_{klm}(U) + \ldots + f(x) + \varepsilon(x)$$

- areal mean e.g. over geological unit U_k
- first categorical effect, e.g. floor level (l)
- second categorical effect, e.g. age of house (m)
- deterministic continuous effect, e.g. air pressure
- (correlated) random fluctuation

factors may or may not be contingent / correlated
ex.: contingent: age of construction ~ building material

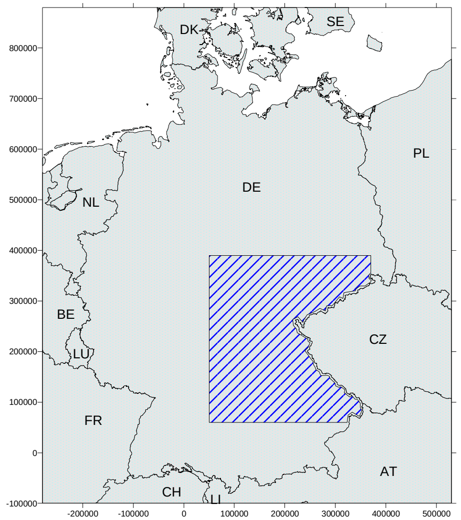
separate estimation: e.g. kriging with external drift
simultaneous estimation: e.g. regression kriging

division deterministic // stochastic part: non-trivial question!

Z = physical observable, or derived: e.g. log(indoor concentration)
U = spatial unit, $x \in U$. (e.g. $U=\{x\}$)

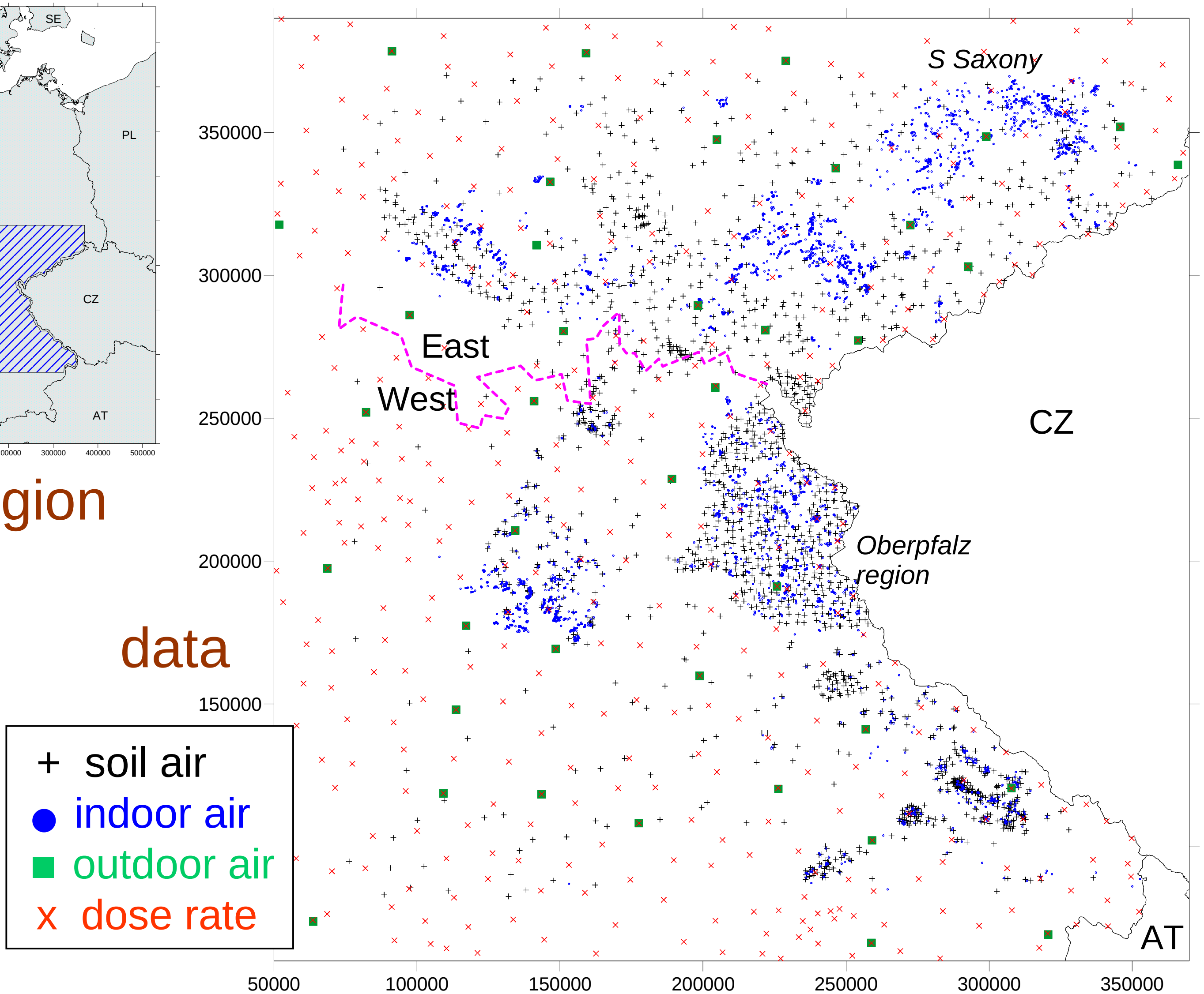
because factors are multiplicative!

example (1/4)



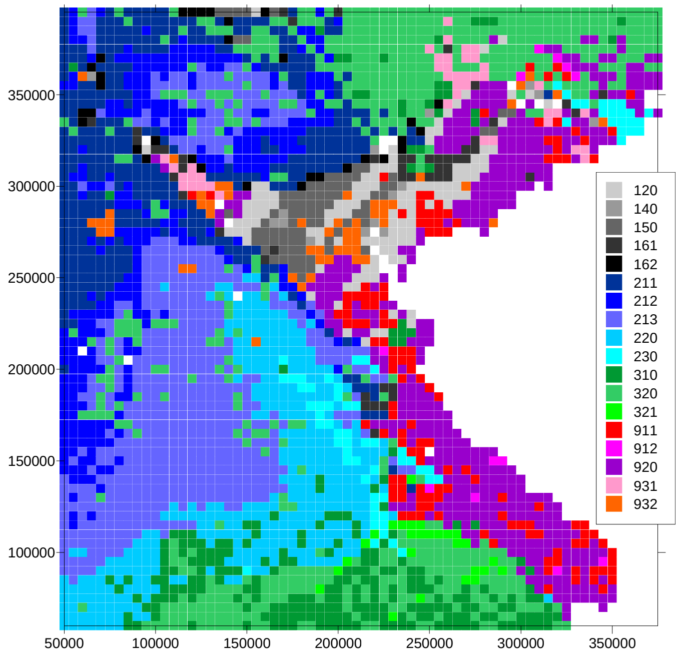
study region

data

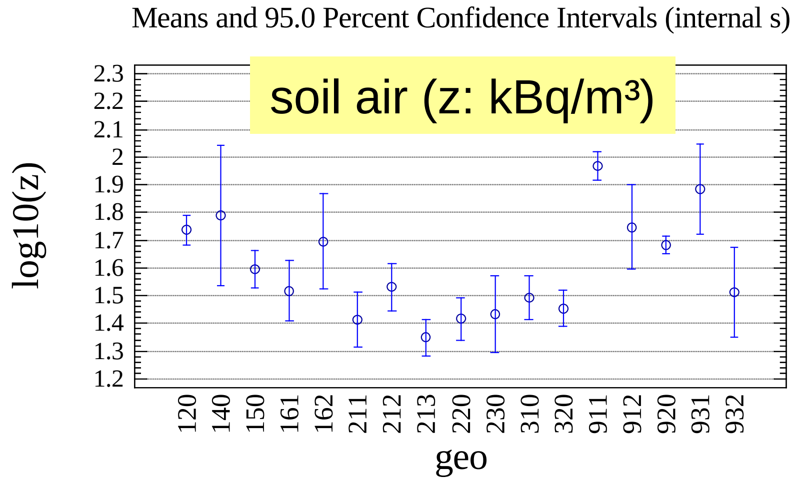


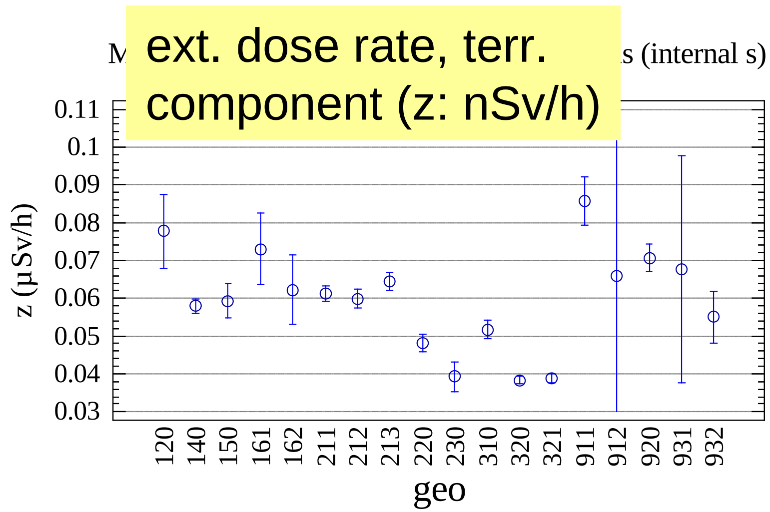
example (2/4)

geo units:



simplified, from German geol. map 1:1M,
coding following Klingel & Kemski

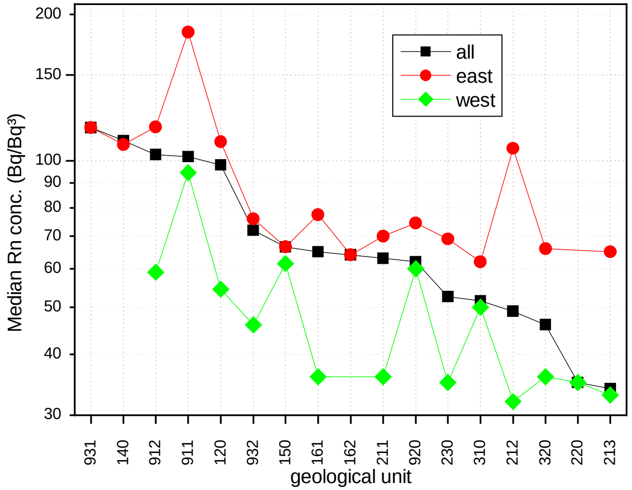


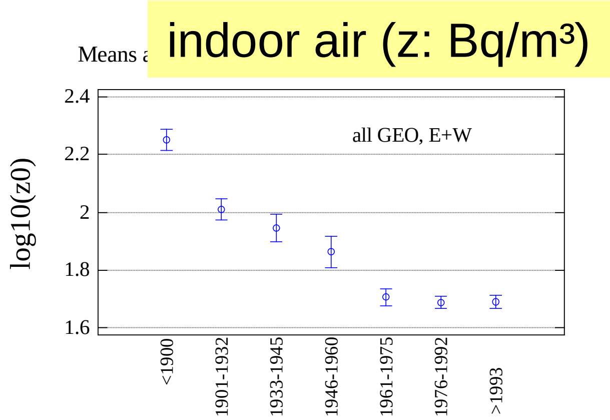


example (3/4)

indoor Rn:

dependencies of factors:
- geology
- former East- / West-Germany
- year of construction

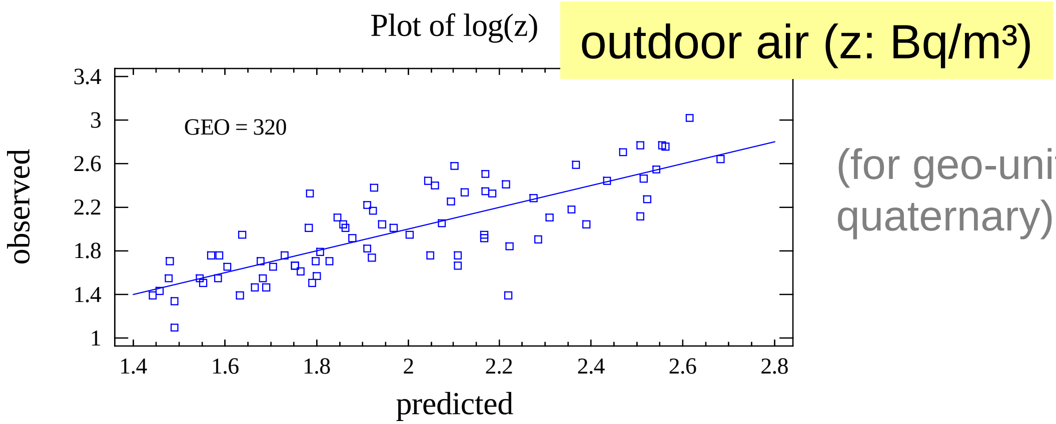




---

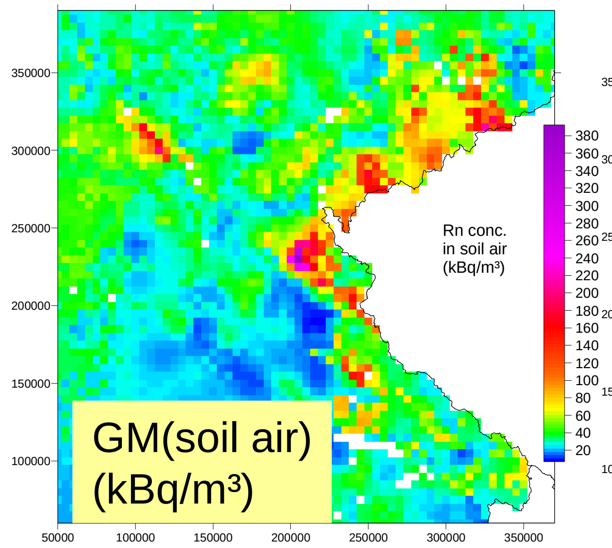
outdoor radon:

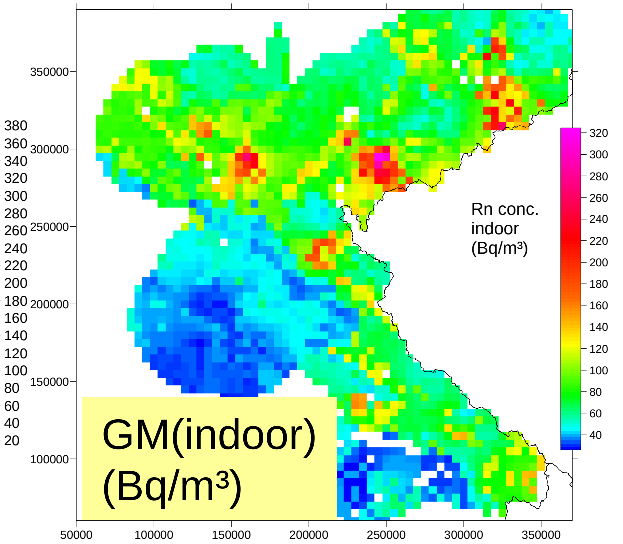
factors:
- geology
- distance from N sea

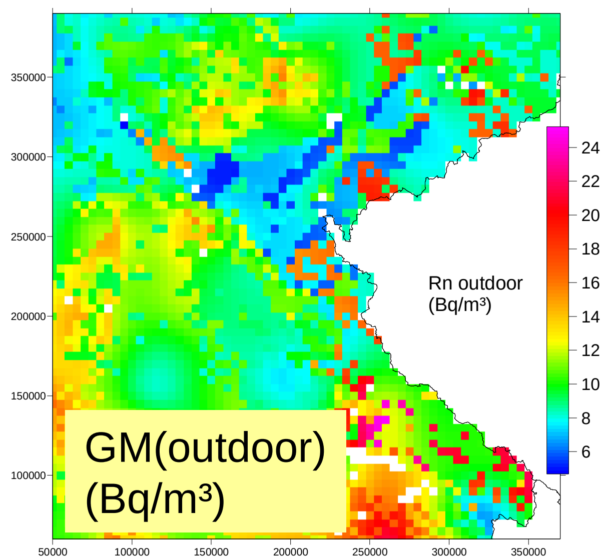


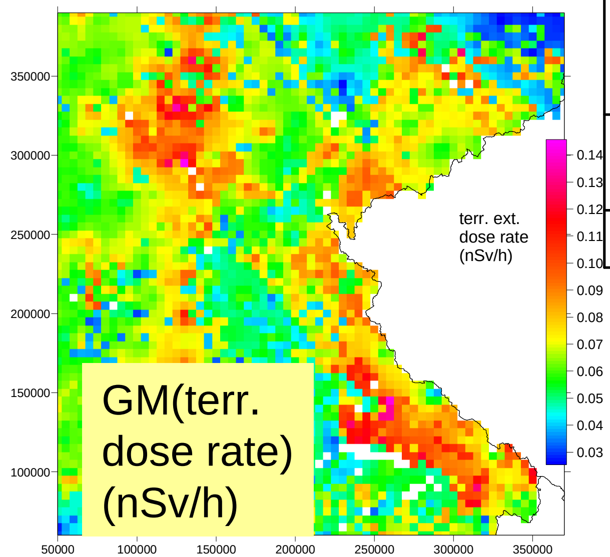
(for geo-unit 320 = quaternary)

example (4/4)









result, separate estimation

regression models:

variable	**factors**
soil Rn, $\ln(Z_s)$	geo-units
indoor Rn, $\ln(Z_{in})$	▪ geo-units ▪ construction year ▪ “East-West”
outdoor Rn, $\ln(Z_{ou})$	▪ geo-units ▪ distance from sea
dose rate, $\ln(Z_d)$	geo-units

sharp edges along geological borders !

II. correlation between variables

symbolically:

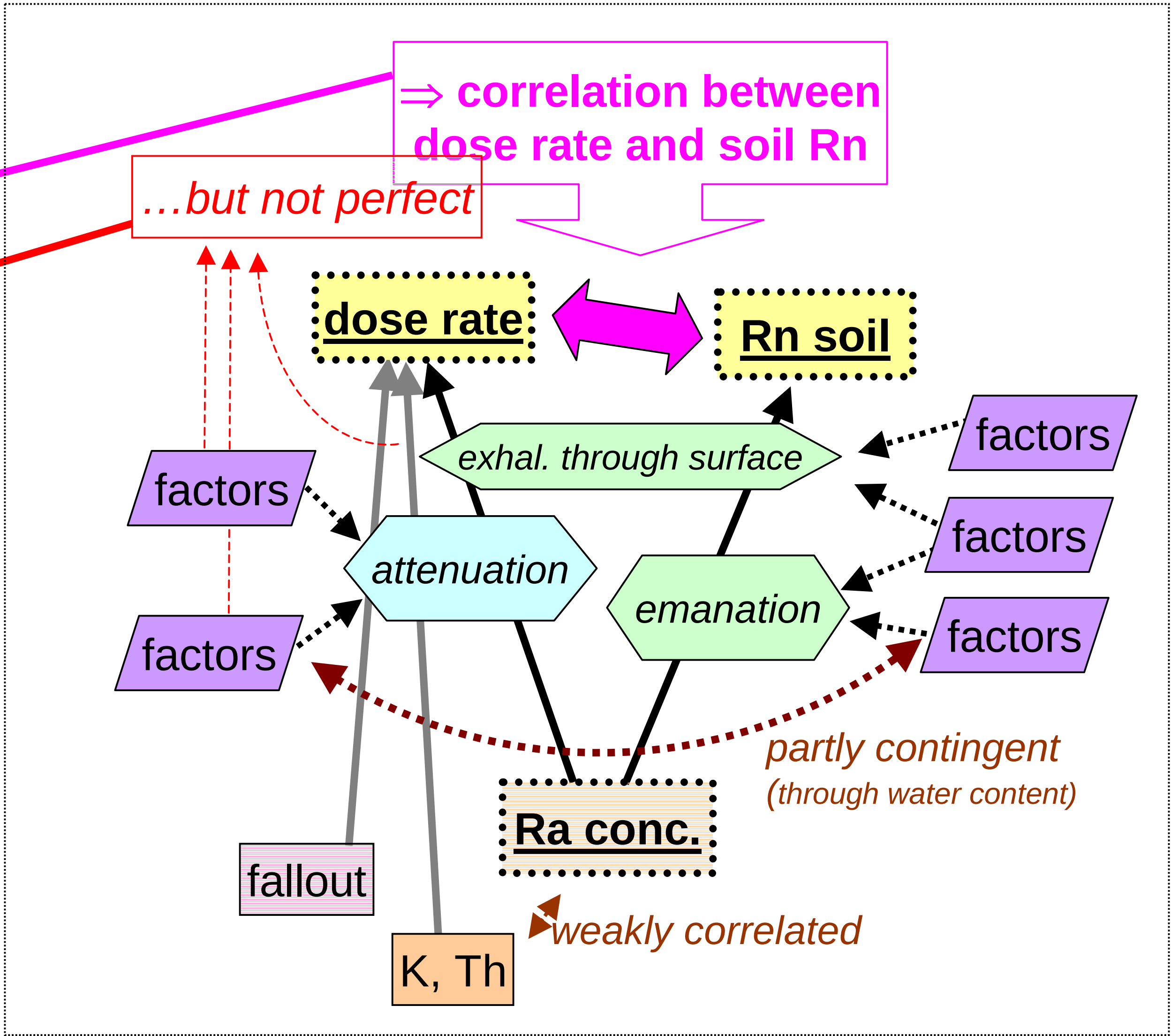
(Rn)(x) =
(dose rate)(x)
+ ε(x)

statistically:

cov(Z^1, Z^2),
r^2(Z^1, Z^2),
τ(Z^1,Z^2) (rank corr.)

Z may be derived from original variables, like log, some g(Z),…

variables indexed by upper indices.
Lower indices: sample index



examples for mutual dependencies

- "$Z^1(U) = f_\theta(Z^2(U))$"
- Example 1:

  $$C_{indoor} = B + T * C_{soil}$$

  = physical model, valid in steady state
  B = building material, T = transfer "factor"

- Example 2:

  $$D_{(ext.\ dose\ rate)} = D_0 + \delta\, C_{soil}$$

  D_0… influence of K, T; δ ~ emanation, attenuation,…

- Example 3:

  $$C_{outdoor} = C_0 + \beta\, C_{soil} + \gamma * (\text{dist. from sea})$$

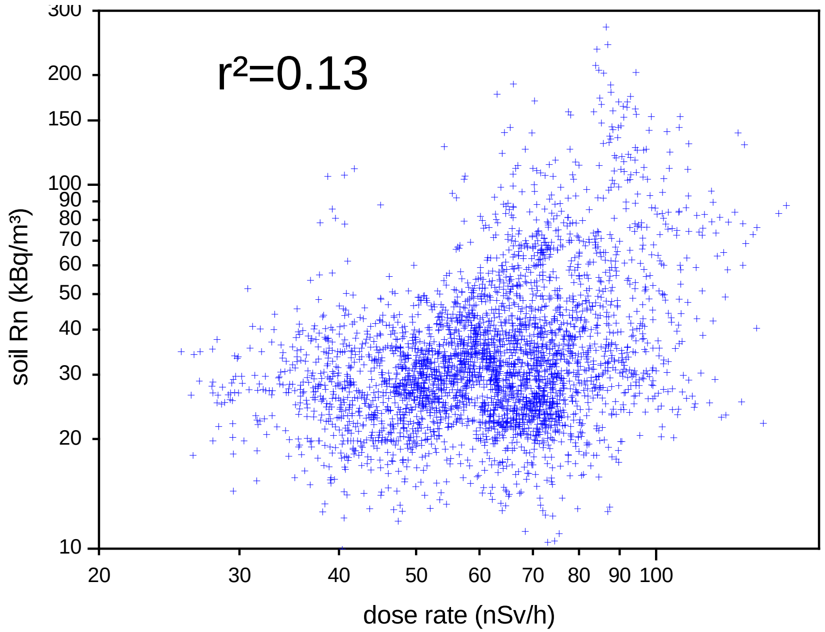
  C_0: contribution from distant locations,
  β: exhalation (~pressure diff., snow, humidity,..)
  γ: proxy for dilution by "clean" sea air

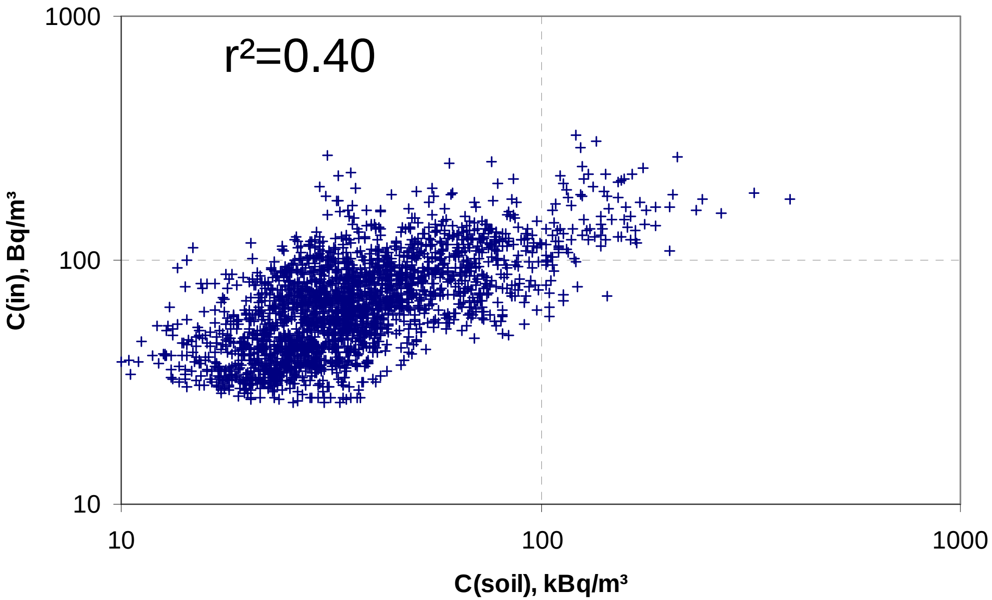
all model parameters θ are themselves random variables $\theta(x)$

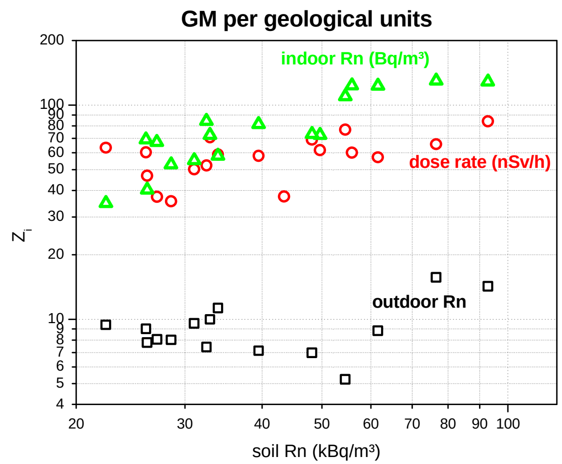
hopefully: spatial variabilities of $\theta(x)$ are low.

not yet examined !

examples: soil Rn ~ dose rate, indoor Rn







so far bad to moderate correlations!

likely reason:

regression models are insufficient,

in particular geological classification

still questionable

III. collocating

- Problem: how to
  (a) construct target variable of different Z^i
  (b) estimate correlation parameters θ,
  if the Z^i not sampled at the same locations ?
- (1) first separate estimation on same set of locations ("sampling set"), e.g. **grid** or locations of one chosen variable. ("collocated estimation")
  (2) model afterwards:
  - $Z^i = f^{ij}_{\theta ij}(Z^j)$ transfer models
  - $Y=Y(Z^i)$ target variable

  } *later slides!*

excursion: sampling sets

- variable Z^i: samples $\{z^i_j\} \equiv \{z^i(x_j)\}$, $j=1...n^i$
  $\xi^i := \{x_j\}$ = sampling set of variable Z^i.
- in general: $\xi^i \neq \xi^k$!
  (locations of indoor Rn and soil Rn samples are different)
- joint sampling set: $\xi := \cup\xi^i$
  (=all sampling locations of all variables)
- grid: Ξ = set of grid nodes or cells
  $\xi \rightarrow \Xi$ requires interpolation

IV. co-estimation

- estimating variable Z^1, exploiting information contained in $Z^2,\ldots,Z^m$
- based on cross-covariances
  $C^{ij}(h) = cov(Z^i(x), Z^j(x+h))$

  a) co-kriging
  b) co-simulation
- problems:
  - estimation of co-variograms
  - technical implementation for m>2
- different approach: ¿¿¿ neural networks and support vector machines ???
- problem: implementation? theoretically more complicated, but seems to have big potential!
- in my view: most statistically satisfying methods!
- not done --- for now!

V. target variable

now starts the more complicated part !

- <u>method 1: "transfer"</u>
  first estimate one common Z^1 out of available Z^i using transfer models from collocated estimates $\rightarrow Z^{1(i)*}$, then $Y=f(Z^{1(i)*})$

inspired by H. Friedmann's proposal (JRC geogenic expert group)

- <u>method 2: "multi-variate proper"</u>
  $Y=f(Z^1,\ldots,Z^m)$. (Some Z^j may be missing.)
  "global" and "local" versions

method 1: cooking recipe

I.

1. estimate all m variables Z^i separately on common grid $\Xi \Rightarrow Z^{1*},\ldots,Z^{m*}$

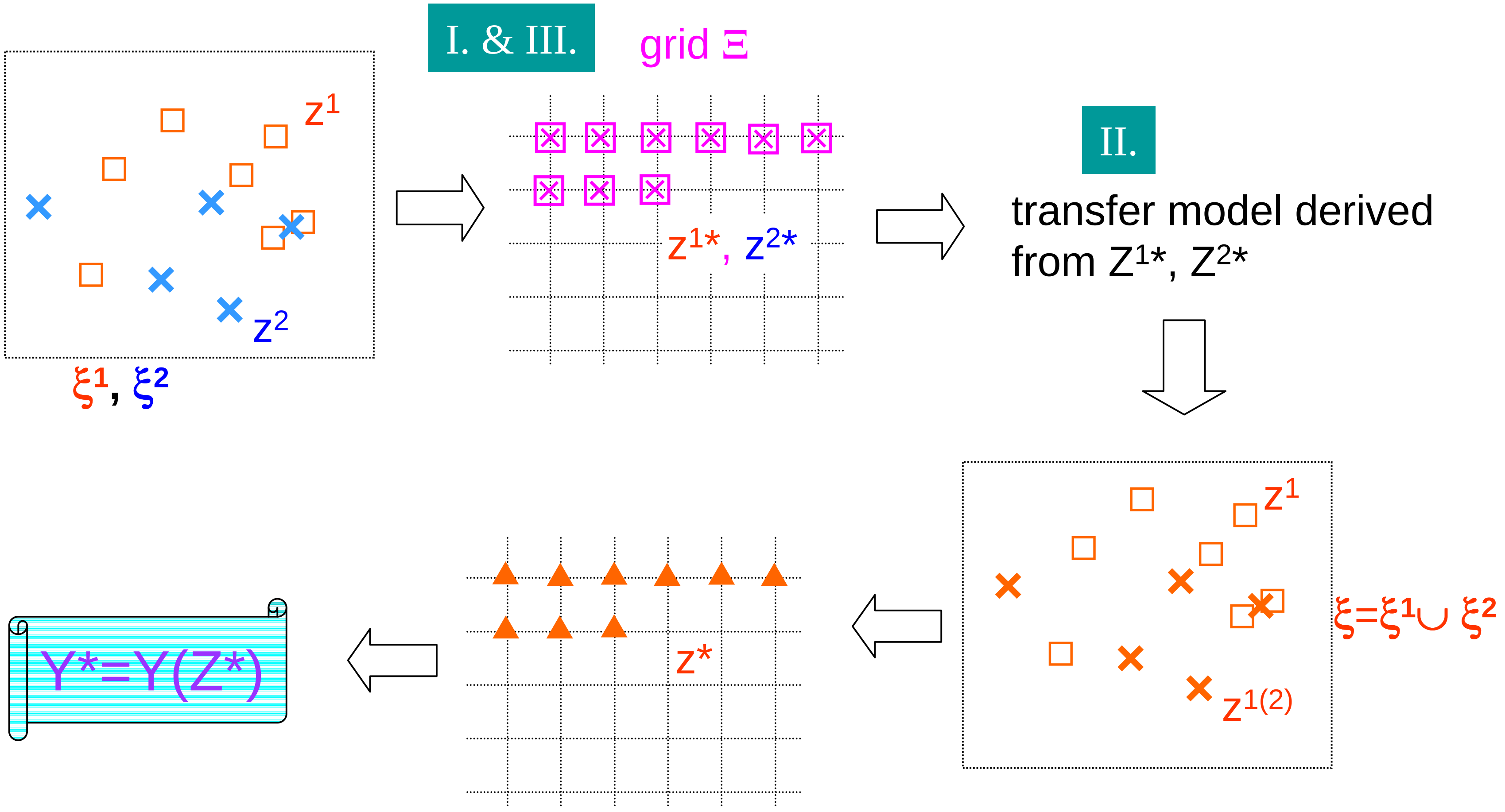
III.

2. estimate transfer models between variables, $Z^i = f_\theta(Z^j)$, data = $(Z^{i*},Z^{j*})(\Xi)$

II.

3. apply the models to the original data $\{z^i\}$ on ξ^i.
4. $\Rightarrow$ e.g., $Z^{1(2)}=f^{12,}{}_{\theta 12}(Z^2)$, $Z^{1(3)}=f^{13}{}_{\theta 13}(Z^3)$
   new dataset $\{z^{1,new}\} := \{z^1\}\cup\{z^{1(2)}\}\cup\{z^{1(3)}\}$ on $\xi=\xi^1\cup\xi^2\cup\xi^3$
   (Z^1=soil Rn, Z^2=indoor Rn, Z^3=dose rate; $Z^{1,new}$ = information of all)
5. use $\{z^{1,new}\}$ for modelling $Z^{1,new}$ on grid Ξ.
   (Contains information of original Z^1, Z^2, Z^3.)
6. target variable $Y=Y(Z^{1,new})$ (could be $Z^{1,new}$ itself)

method 1


I. & III.
grid Ξ
z1
z2
ξ1, ξ2
z1*, z2*
II.
transfer model derived from Z1*, Z2*
z1
z1(2)
ξ=ξ1∪ ξ2
z*
Y*=Y(Z*)

method 2

1. estimate all available or wanted variables Z^i separately on common grid $\Xi \Rightarrow Z^{1*},\ldots,Z^{m*}$ (= method 1)
   wanted but not available: use regression, I.
2. multivariate target variable $Y=Y(Z^1,\ldots,Z^m)$
3. How to?....
   - multivariate classification (CZ, DE, USA, others)
   - continuous
   -- (a) estimate *local* $F_Y(x)$; ($\rightarrow$ 3rd next slide)
   -- (b) *global* distribution $G_{Z^1,\ldots,Z^m}$ for a region B. ($\rightarrow$ 4th next slide)
   in both cases, spatial association enters through the input variables Z^i.
   If |B| large $\Rightarrow$ G independent of B.

proven and viable concept, could be implemented relatively easily

missing input

in general, some Z^j missing!

consistency:

strong version:

at each point x, $Y(Z^1,\ldots,\blacksquare^i,\ldots,\blacksquare^j,\ldots,Z^m)$ must be independent of which Z^i are missing (up to statistics).

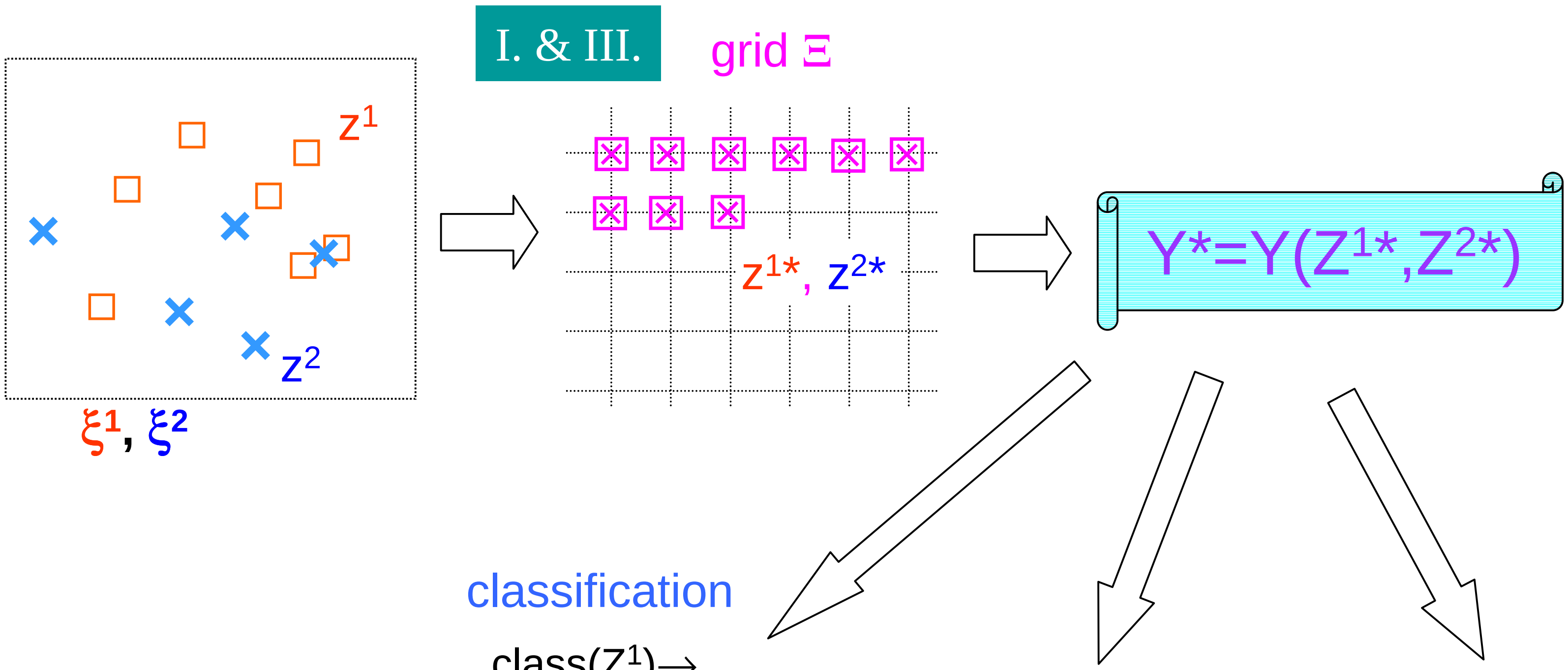
E.g.: $Y(Z^1,Z^2,Z^3) \cong^! Y(Z^1,Z^2,\blacksquare)$ … maybe not realistic

weak version: “conservative”

$Y(Z^1,Z^2,\blacksquare) \geq^! Y(Z^1,Z^2,Z^3)$ … $\geq$ means “higher hazard”

Serious constraint on admissible Y !

method 2



classification

class(Z^1)→

class(Z^2)→

	level 1	level 2
level 1	Y(1;1)	Y(2;1)
level 2	Y(2;1)	Y(2;2)

(standard; not further discussed here)

local $[F_Y(y)](x)$

next slide

global $[G_{\underline{Z}}(\underline{z})](x)$

2nd next slide

example for Y from local $F_Y(x)$ (a)

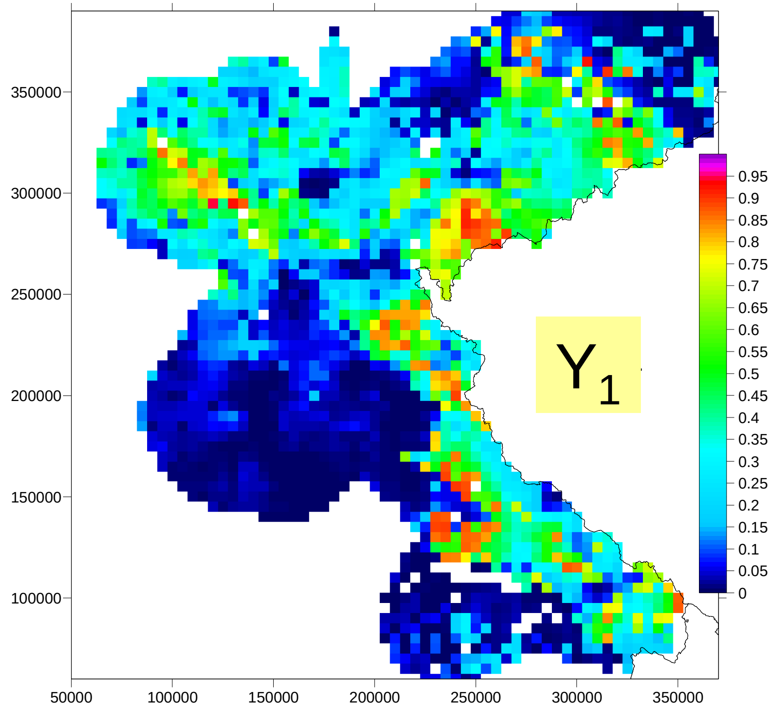
- at x, generate many realizations $(z^1,\ldots,z^m)^{(k)} =: \underline{z}^{(k)}$ using global $\mathrm{cov}(Z^1,\ldots,Z^m)$ from II.
- for each $\underline{z}^{(k)} \rightarrow Y^{(k)}$
- statistics of Y over (k) (e.g. $AM\{Y^{(k)}\}$) $\rightarrow$ Y

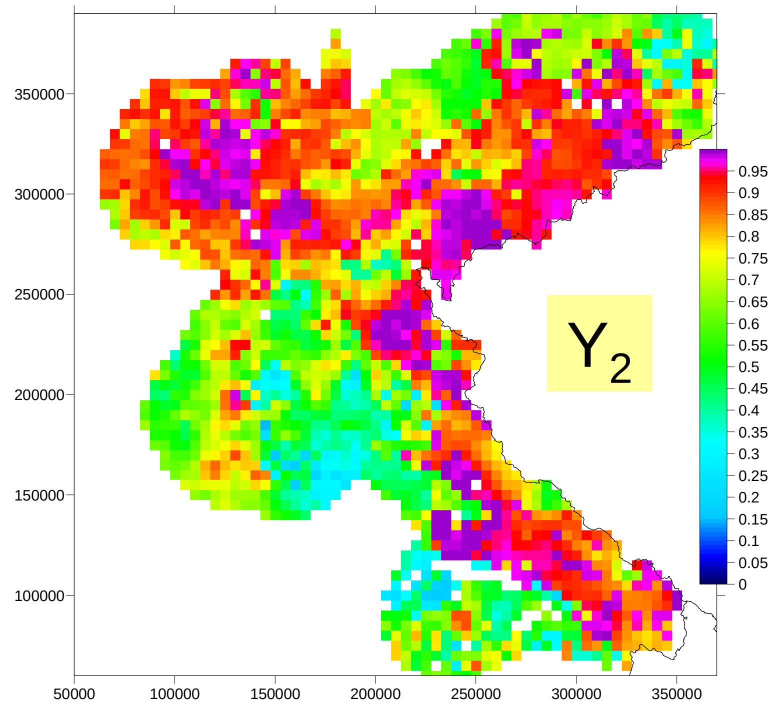
$\forall$ x

- technically: e.g. Cholesky method: A, such that AA^T=cov.; generate $u^i \sim N(0,1)$ indep.; $\Rightarrow \underline{z}=\underline{\mu}+A\,\underline{u} \sim N(\underline{\mu}, \mathrm{cov})$
- needs to be demonstrated yet !

example for Y from global G (b); 1/4

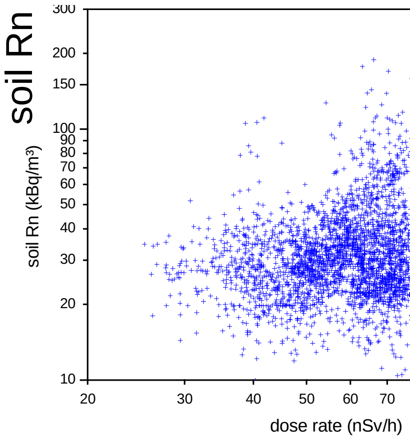
- $Y_1(x) := \text{prob}(z^1 < \zeta^1, \ldots, z^m < \zeta^m) \equiv F_{\underline{z}}(\underline{\zeta})$ ($\underline{z} := (z^1, \ldots, z^m)$)
- $Y_2(x) := 1 - \text{prob}(z^1 > \zeta^1, \ldots, z^m > \zeta^m)$
- 3-variate scoring (m=3): indoor conc., soil conc., dose rate
- use (for now) empirical distributions F_{emp}





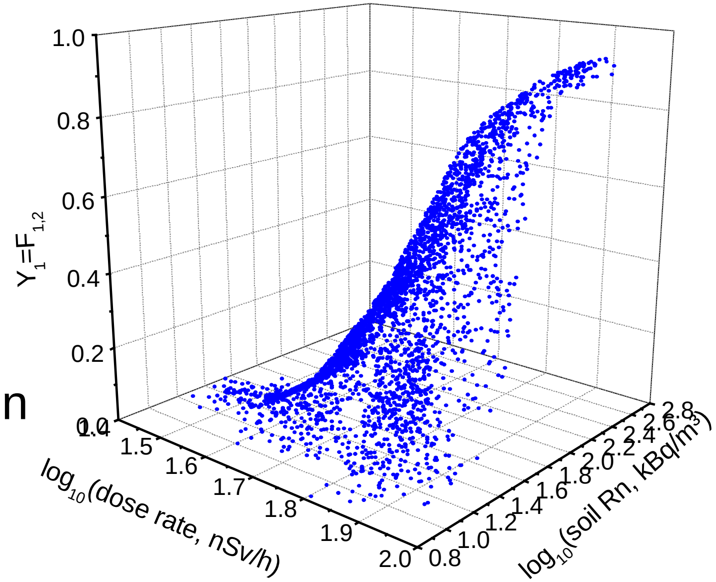
missing: e.g., $Y_1(\zeta^1, \zeta^2, \blacksquare)(x) = F_{\underline{Z}}(\zeta^1, \zeta^2, \infty)$ (not yet implemented)

continued, 2/4: soil Rn & dose rate

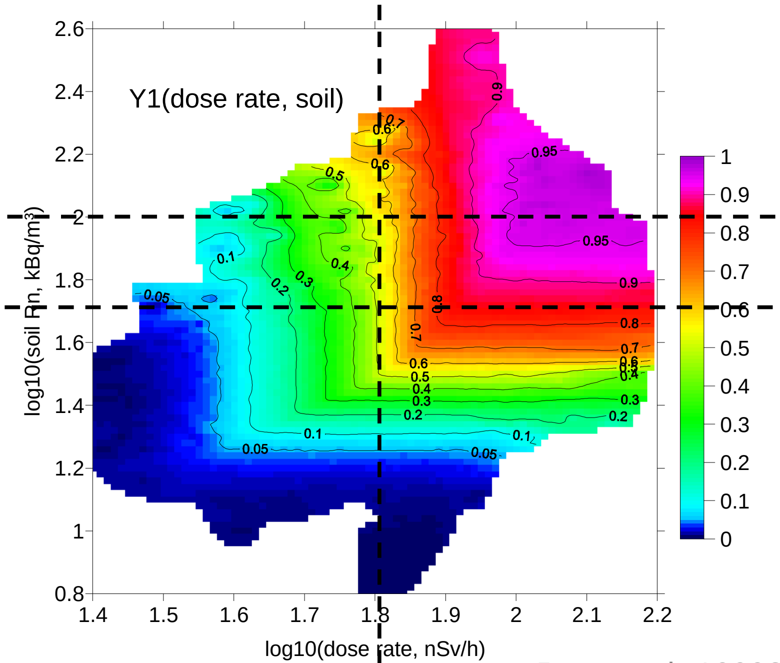


dose rate

soil Rn ~ dose rate:
bad correlation !
$r^2=0.13$



bivariate distribution
$Y_1=F_{Z^1,Z^2}$



$C_{soil}=100$

$C_{soil}=50$

AM(dose rate)=64

<u>treatment of missing variable,</u>

example: dose rate = missing

→ $Y_1(100,\infty)=1$

$Y_1(100,AM(\text{dose rate}))=0.55$

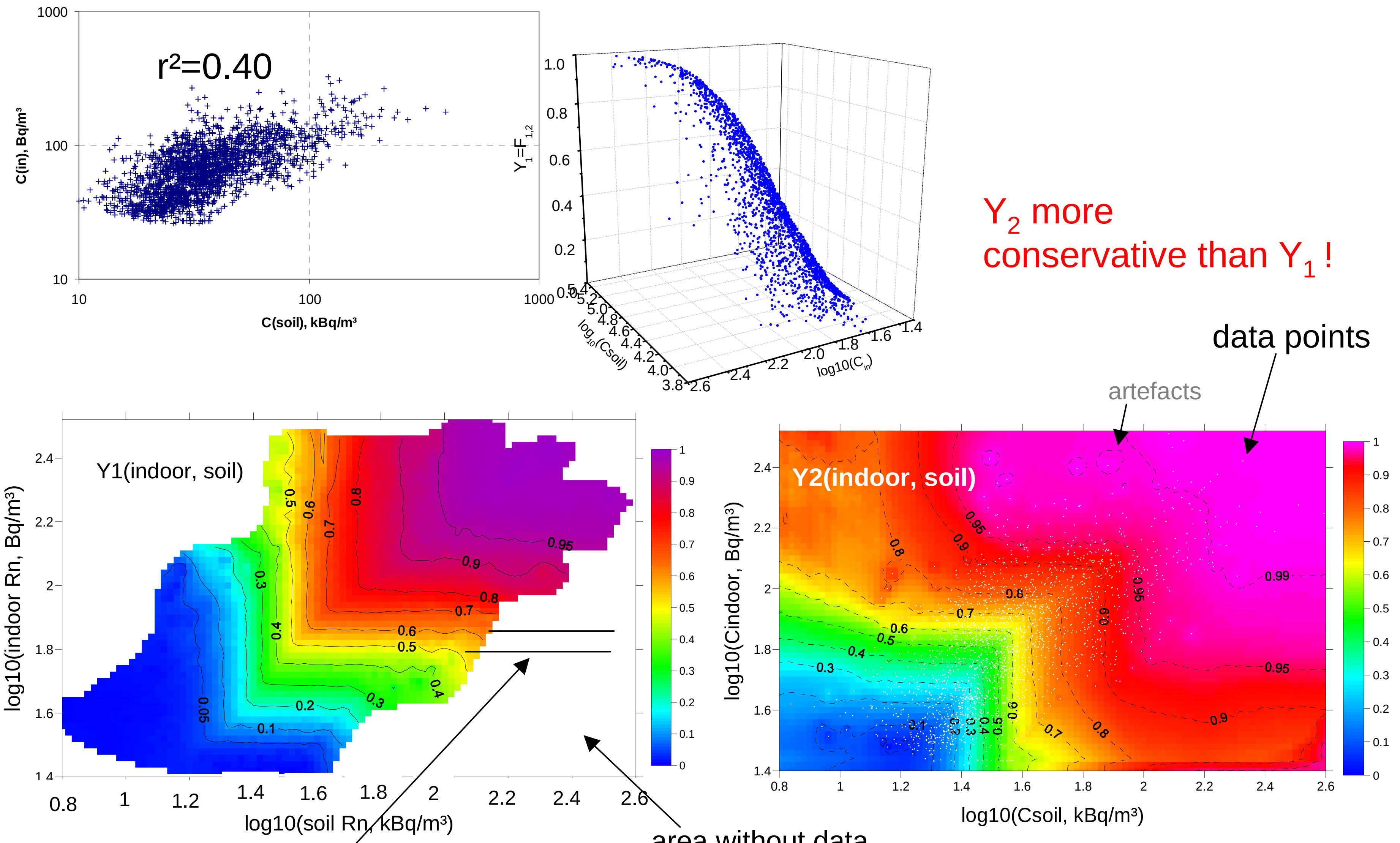
→ $Y_1(50,\infty)=0.83$

$Y_1(50,AM(\text{dose rate}))=0.55$

<u>does not require</u> a priori knowledge of $F_1(Z^1,Z^2)$!

<u>requires</u> a priori knowledge of $F_1(Z^1,Z^2)$!

continued, 3/4: soil Rn & indoor Rn



Y_2 more conservative than Y_1 !

theoretically: straight lines

area without data support blanked

improvement / generalization

on $[0,1]^n$, define a p-norm $||.||_p$, ($\rightarrow$ metric space)

$Y^p(x) := ||F_{\underline{z}}(x)||_p := (m^{-1/p})(\Sigma_i(F_{zi}(x))^p)^{1/p}$

n= number of possible covariates,

m= number of available covariates

projection onto main diagonal of sub-cube $[0,1]^m$.

i.e. Y^p is functional f^p: $f^p[\underline{Z}(x)]=||F_{\underline{Z}}(x)||_p$

p=0 ... corresponds score 1 (Y_1)

p=∞ ... corresponds score 2 (Y_2, „maximum norm“)

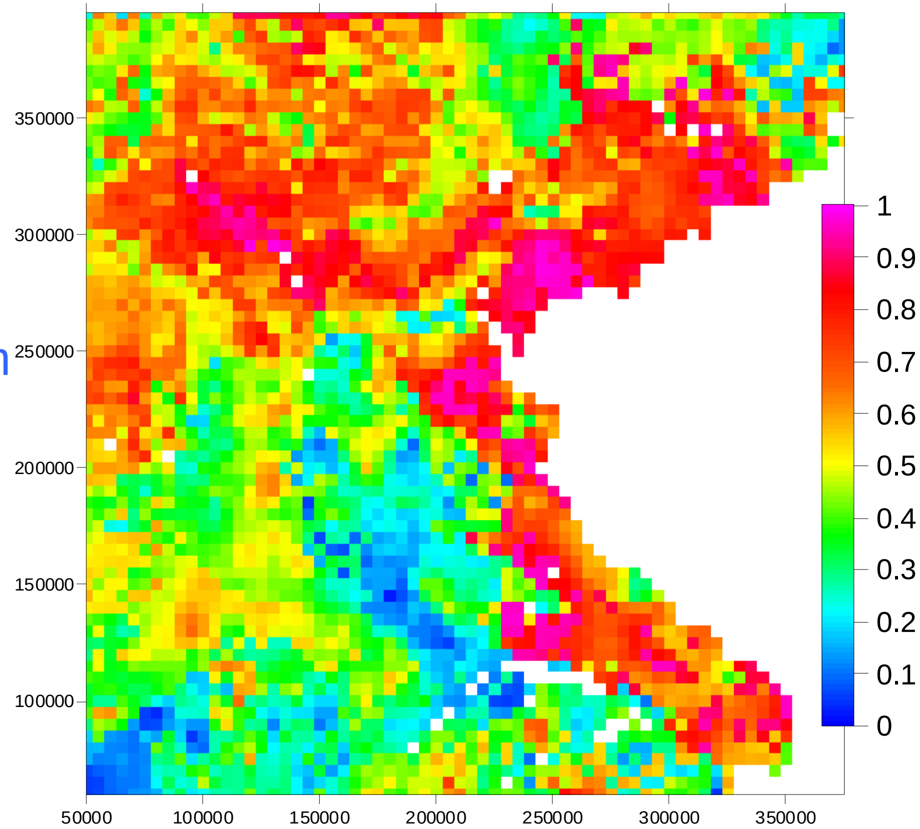
p=1 ... inner product ($F_{\underline{z}}(x) \bullet \underline{1}$)

Y^1 = joint $F_{\underline{z}}$ as if the z^i were perfectly correlated; Y^1=average of F_{zi}

advantage of concept: more adaptive and flexible !

natural treatment of missing data if m<n

example: p=2;
soil Rn, indoor Rn, dose rate



4/4: in practice

- how to estimate $Y(z^1,\dots,z^m)$?

1. given $F_{z1},\dots,F_{zm}$ from previous experience
2. nscore $z \to w \Rightarrow F_{w1},\dots,F_{wm} \sim N(\mu^i,\sigma^i)$; σ^{ij}
3. values $\{z^i_k\} \rightarrow \{w^i_k\}$ from transform table $\Rightarrow$ values = m-tuples $(w^1,\dots,w^m)_k$
4. generate N (many!) $(u^1,\dots,u^m)$ random $\sim F_{\underline{w}}$
5. $Y_1(\underline{z}) = Y_1(\underline{w})$ = prob $(w'^1<w^1,\dots, w'^m<w^m) \approx \min\{\#(u^1<w^1)/N,\dots,\#(u^m<w^m)/N\}$
   Y_2 and Y^p in analogy

- if sufficient data for F_{z^i}: easy to implement automatically !

Σ method 2: cooking recipe

1. select your input variables
   (= the ones you have)
2. regression modelling on available factors **I.**
   (geology,…) → $Z^i = factor_{1,j} + factor_{2,jk} + \ldots + f(factor_n) + \ldots$
3. estimate on common grid $\rightarrow (z^{1*},\ldots,z^{m*})(\Xi)$ **III.**
4. correlation analysis → $cov(Z^i, Z^j)$ **II.**
5. select target variable Y
   …. Z^1 // classification // F_Y // $G_{\underline{Z}}$ **V.**
6. estimate Y

1 - 4: ± straight forward;
5 - 6: still problems

problems (as usual)

1. **Geological control: proper definition of geological classes remains to be done.**
   ⇒ hopefully also better correlation between variables!
2. **Definition of input variables, in particular soil-Rn, permeability**
3. **Regression modelling: data! data! data!**
4. **target quantity: more discussion still needed.**
5. **estimation methods to be improved.**

literature review would be helpful!

here ! round table Friday!

dig deeper into geo-statistical methodology!