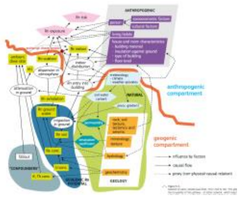
Federal Office for
Radiation Protection

How to use machine learning for (radon) mapping?

Eric Petermann, Peter Bossew

Workshop Geological Aspects of Radon Risk Mapping

22 Sep 2021, Prague

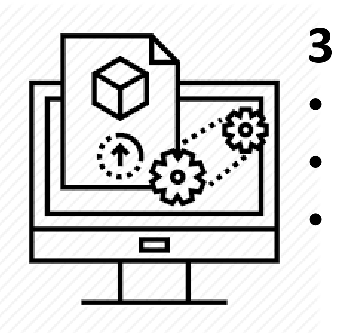


1. Motivation

- Complex system!
- Data everywhere

2. Machine Learning

- What is it?
- How does it work?

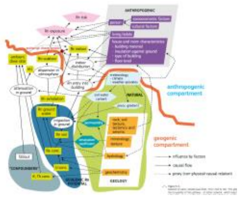


3. ML model building

- Predictor selection
- Tuning
- Interpretation

4. Conclusion & outlook

Federal Office for
Radiation Protection



1. Motivation

- Complex system!
- Data everywhere

2. Machine Learning

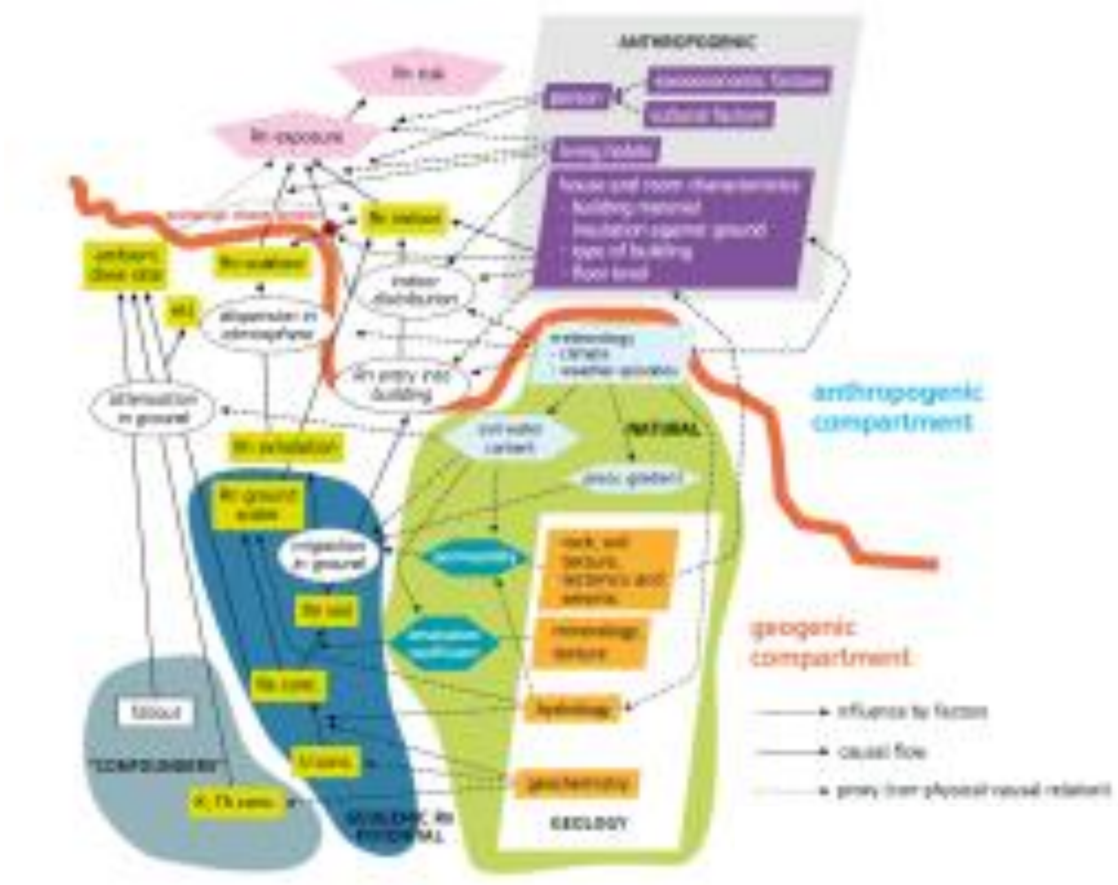
- What is it?
- How does it work?

3. ML model building

- Predictor selection
- Tuning
- Interpretation

4. Conclusion & outlook

Why ML - Radon is a complex issue!



Cinelli et al (2019): European atlas of natural radiation

<u>Challenges in Rn mapping</u>

- Very complex system
- Interplay of a variety of factors
- Observations does not necessarily reflect long-term mean due to temporal variability, e.g. effect of weather on short-term measurements of Rn -> noise
- Individual extreme values (caused by weather, issues during sampling etc.) can have a significant effect on predictions for a large area
- ML is able to consider many driving factors (or proxies) as predictors
- ML suitable for modelling complex non-linear processes

Why ML - Data everywhere

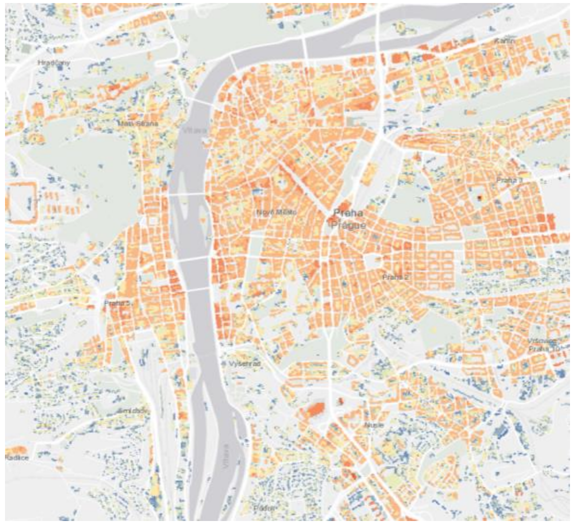
https://www.esa.int/Space_in_Member_States/Germany/Die_Copernicus-Dienste

https://esdac.jrc.ec.europa.eu/resource-type/european-soil-database-soil-properties

- More and more data available
- Satellite missions (NASA, ESA -> copernicus)
- Open access to data sets and maps on national/continental/global scale
- Citizen science

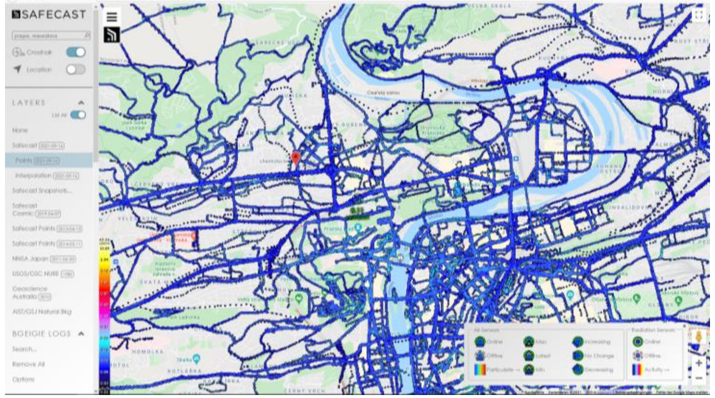
→ **A lot of suitable data for explaining geogenic and indoor Rn variability!**

Building height ->
number of floor levels



https://land.copernicus.eu/local/urban-atlas/building-height-2012?tab=mapview

Safecast-> radioactivity monitoring



https://map.safecast.org/?y=50.0963&x=14.4014&z=14&l=1&m=0

1. Motivation
- Complex system!
- Data everywhere

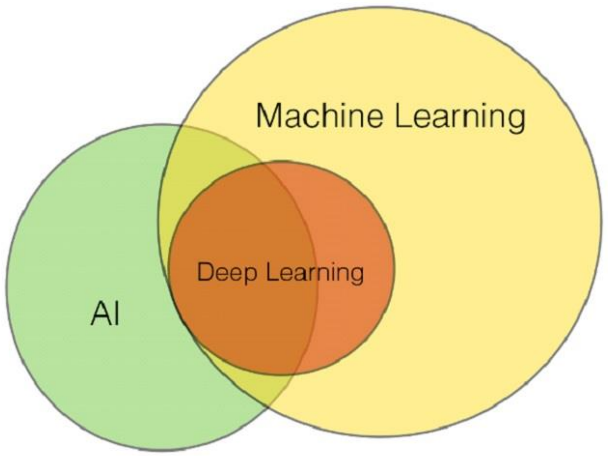
2. Machine Learning
- What is it?
- How does it work?

3. ML model building
- Predictor selection
- Tuning
- Interpretation

4. Conclusion & outlook

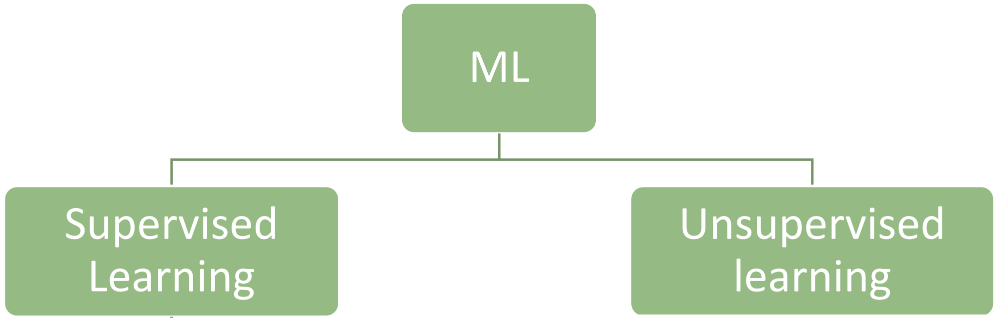
What is Machine learning?

AI vs. Machine learning



www.en.wikipedia.org

- Component of AI -> extract knowledge from large data sets
- Non-parametric
- Data-driven
- Supervised vs. unsupervised ML



-> Rn concentration

-> Exceedance probability

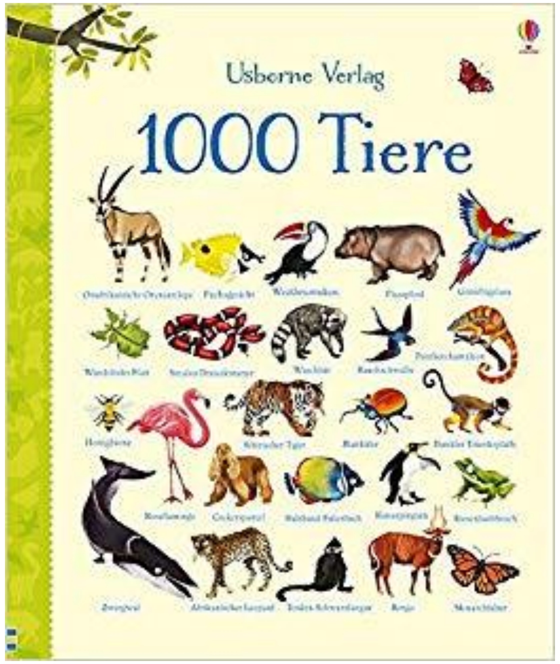
Examples:

- Image and speech recognition
- predictive marketing (Amazon, Google)
- autonomous driving

How does ML work?

Training

Test



© Usborne; https://www.amazon.de/1000-Tiere-Jessica-Greenwell/dp/1782321179

Example: How toddlers learn distinguishing animals (-> classification)

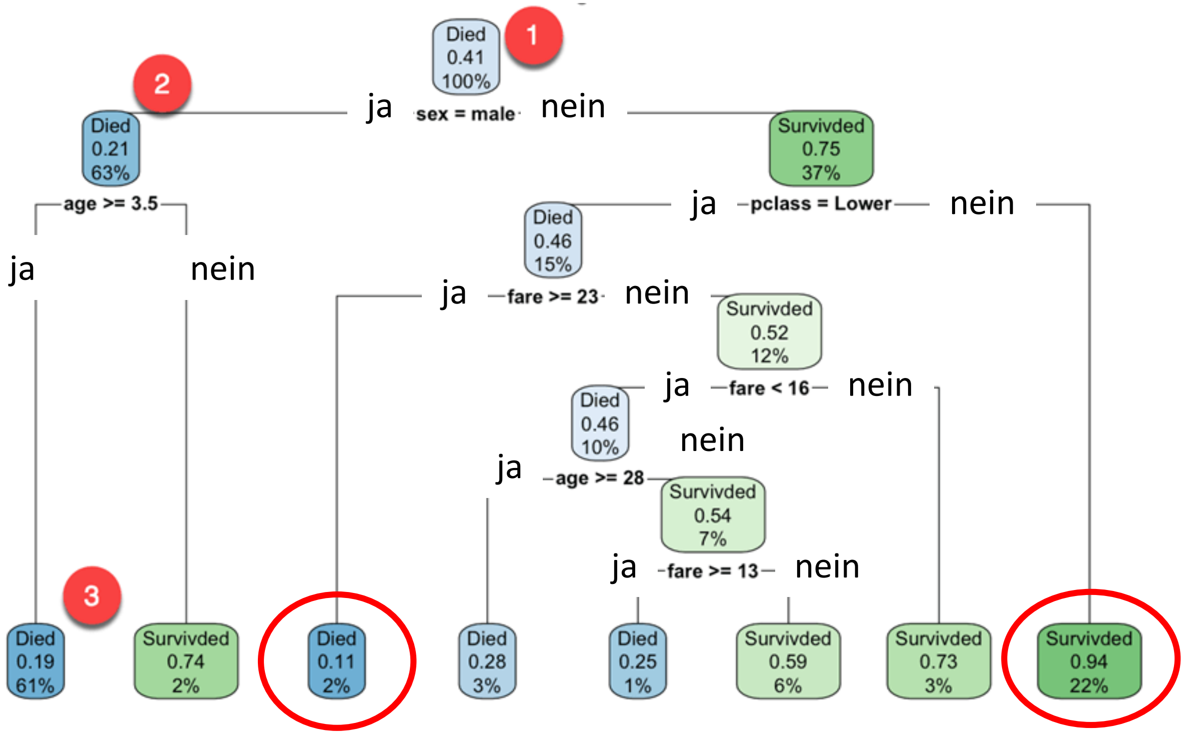
- Assign attributes/properties (colour, size, shape etc.) to terms/label
- Relationship between attributes and terms created
- → test data required that was not used for training to test generalizability
- Algorithms highly flexible -> risk of overfitting (learning of patterns in training samples)
- Reducing the risk of misinterpretation of relationships: direction of view, relative position

Algorithms: Example Random Forest

Regression tree:
Example survival probability sinking of Titanic


https://de.wikipedia.org/wiki/RMS_Titanic

Target variable: died/ survived
4 properties/predictors: sex, age, passenger class, fare



https://www.guru99.com/r-decision-trees.html

- Combination of many (e.g. n=500) decorrelated decision trees
- Every decision tree is build only with a fraction of available data (e.g. 80 %)
- At every split only a subset of available predictors (e.g. 3 out of 9) is evaluated and used for splitting the data
- Optimization criteria: reduction of prediction error

-> grouping of sample data into smaller statistically more similiar subsets

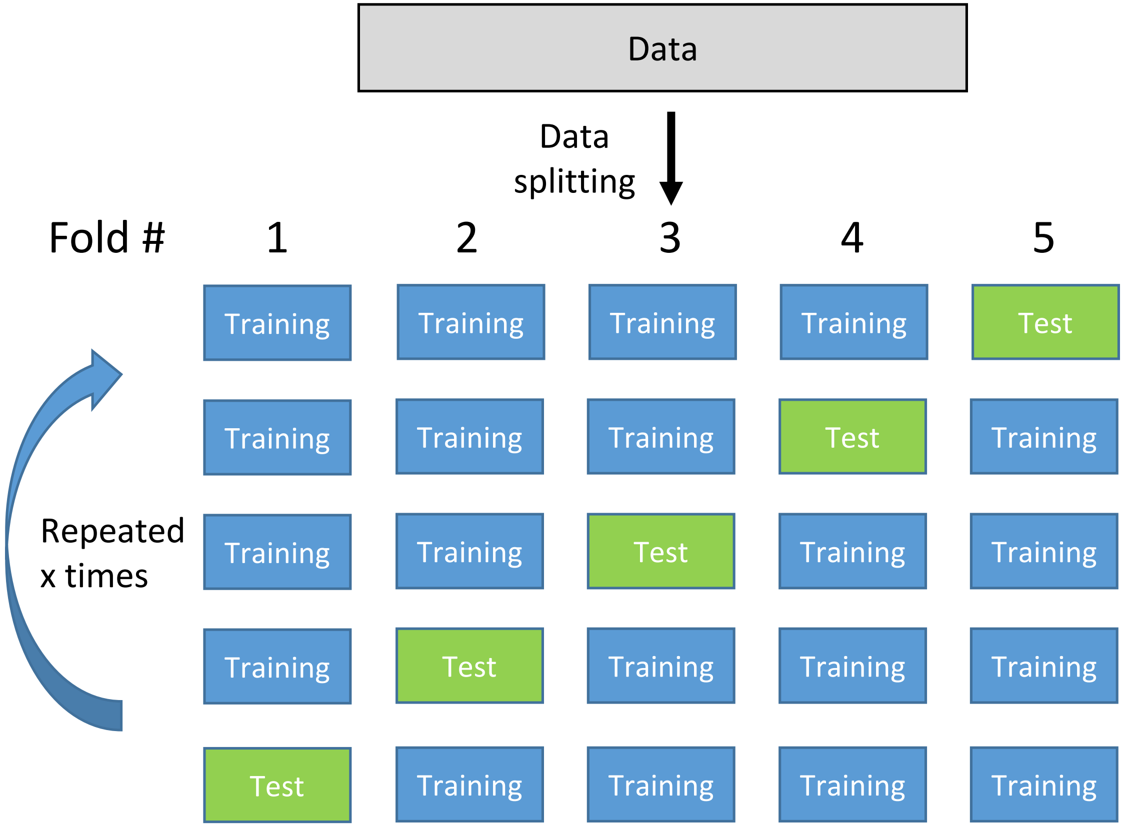
1. Motivation
- Complex system!
- Data everywhere

2. Machine Learning
- What is it?
- How does it work?

3. ML model building
- Predictor selection
- Tuning
- Interpretation

4. Conclusion & outlook

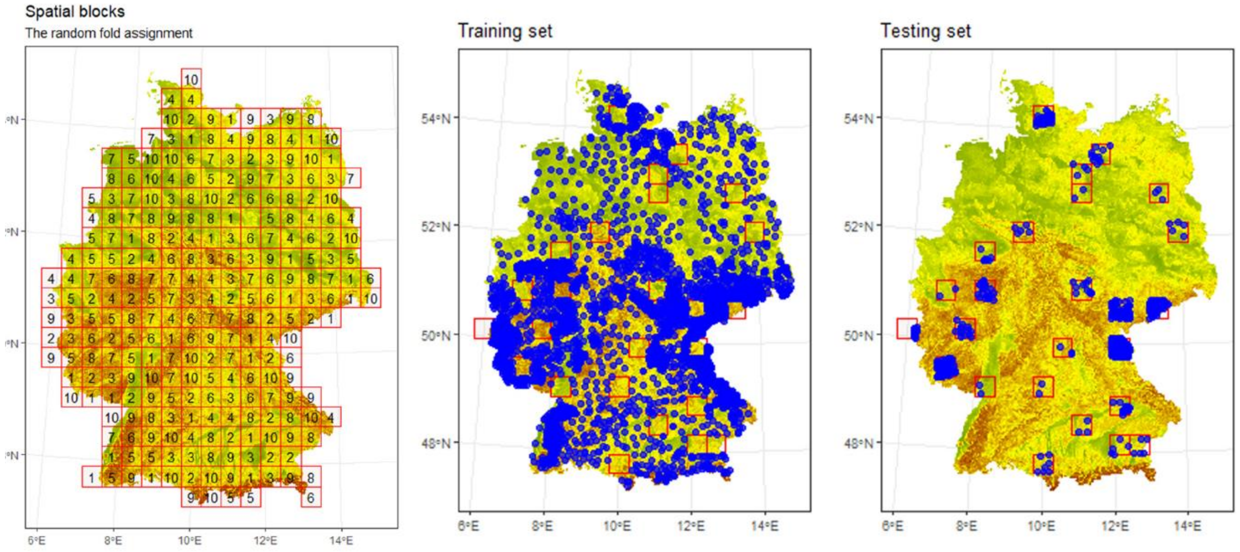
Model building – training and testing via cross-validation



Problem: Random splitting of data does not guarantee independence of training and test data (i.i.d. -> independent and identically distributed)

→ Spatial auto-correlation of samples (that´s why geostatistics can be used for mapping)

Solution: data splitting with spatial blocks larger than correlation length



More details:

- Roberts et al (2017), Ecography.
- Meyer et al. (2019), Ecological Modelling.

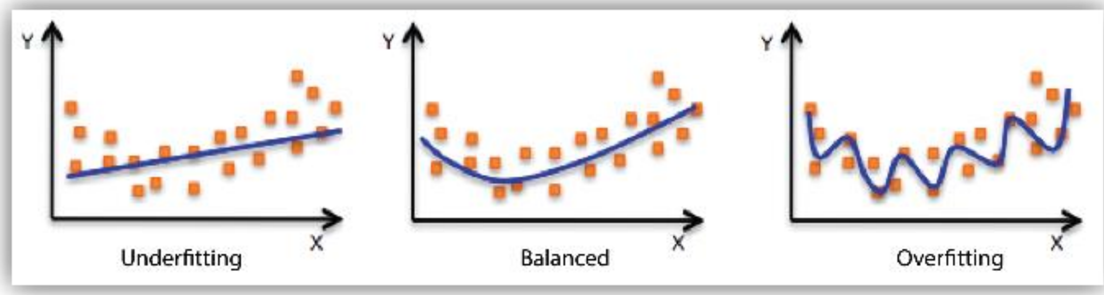
Model building – predictor selection

- Many candidate predictors, sometimes >100
- Not all of them improve model performance
- Computational expensive

→ Select only relevant predictors
  - Principle of parsimony
  - Avoid overfitting

e.g., viewing direction

- Predictor selection -> goal: finding optimal combination of predictors (criteria: test performance)



https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Different ways of predictor selection, e.g. forward selection:

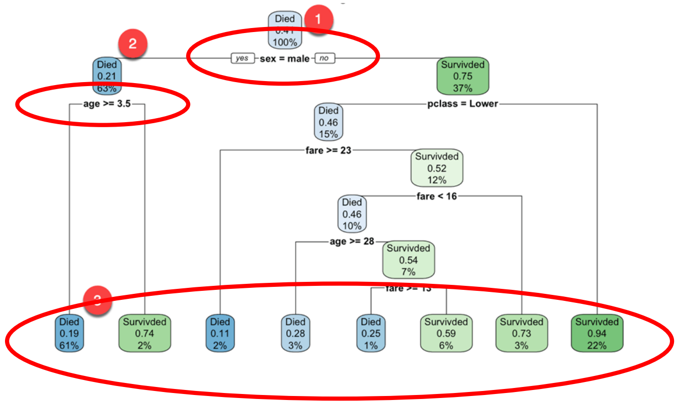
1. testing of every two predictor combination
2. Select best 2-predictor-combination
3. test all not-selected predictors as a third predictor
4. Select best 3-predictor-combination
5. Continue until adding predictors does not improve test performance

implementation for R in package CAST (Meyer, 2021)

Model building – tuning

Tuning of hyperparameters:

- Hyperparameters are e.g.: minimum number of measurements in leaves, number of predictors evaluated at every split
- Cannot be directly estimated from the data
- Importance dependent on algorithm: for random forest small impact, for deep learning large impact
- Testing different combinations of hyperparameters



Model building – final model & mapping

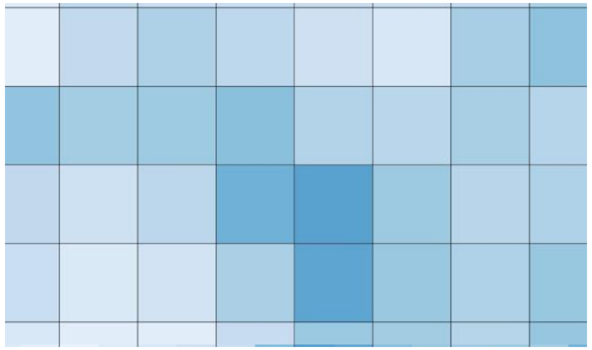
- After predictor selection and hyperparameter tuning the final model can be fitted using all available observational data
- For spatial prediction (-> mapping <-) the final model is computed for every grid cell (for random forest 1000 regression trees are computed and averaged), i.e. every grid cell needs information of all informative predictors

→ upscaling/downscaling if cell resolution of predictors is higher/ lower
→ rasterizing of polygon data (e.g. for geology): conversion of vector data (polygons) to raster data
→ A single geological unit needs to be assigned to a grid cell; dominant geology or geology at cell centre → this is a critical decision!

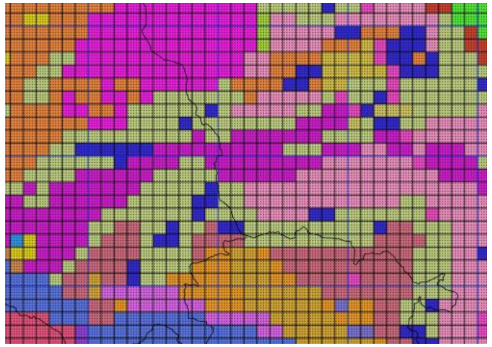
- For large-scale and/or high-resolution mapping working memory intensive

-> tiling required, i.e. dividing the mapping area in smaller units, e.g. for Germany 1km grid cells ~250 tiles
-> then, jigsaw puzzle („merging“) of tiles to the final map

Upscaling 500 m -> 1000 m



Rasterizing geology



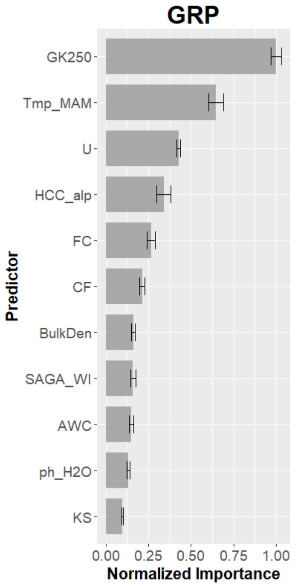
Faults -> density

Model building – interpretation

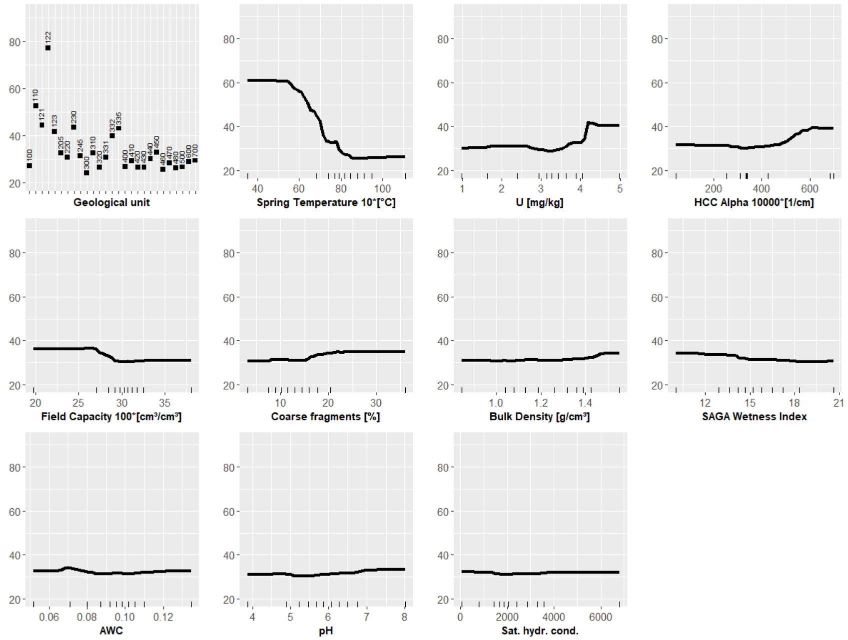
1. Variable importance: relative importance of selected predictors in the model
2. Partial dependence plots: quantitative understanding of predictor-response relationship
3. Spatial dependence plots: mapping of partial dependence

https://www.castsoftware.com/blog/cracking-open-the-black-box-of-it-for-ceos

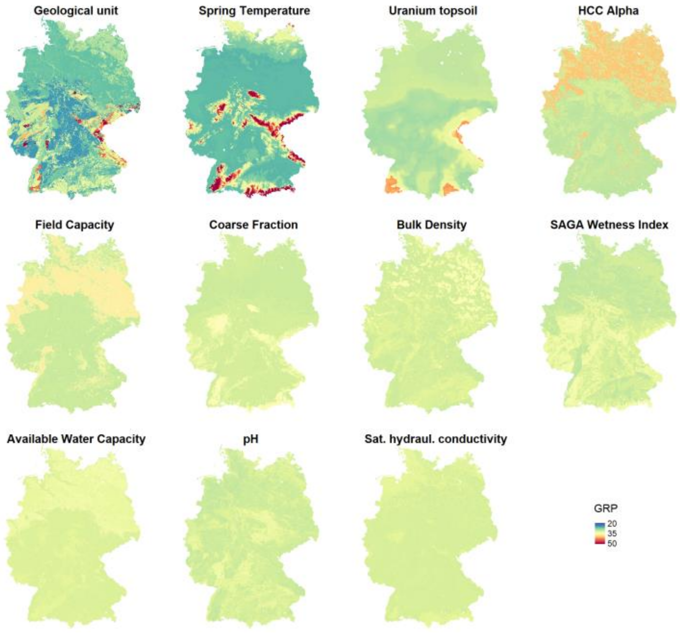
1)



2)



3)



Petermann et al (2021), Sci. Total Environ. 754; Petermann & Bossew (2021), Sci. Total Environ. 780

Model building – some practical issues

- Which algorithms are the best?
  - …it depends…
  - especially ensemble based algorithms (such as random forest) suitable for noisy data (e.g., Rn in soil)
  - deep learning used for many industry applications
- What software to use?
  - e.g., R, python, ArcGIS (?)
- How long does it take to build a model?
  - dependent on amount of observational data, predictor data, algorithm, hyperparameter setting
  - most time required for data collection and pre-processing
- What computational power is required?
  - many cores beneficial -> parallel computing
  - state-of-the-art desktop computer sufficient for (most) regional to national mapping with <10.000 observations and <50 predictors

1. **Motivation**
   - Complex system!
   - Data everywhere

2. Machine Learning
- What is it?
- How does it work?

3. ML model building
- Predictor selection
- Tuning
- Interpretation

4. Conclusion & outlook

Conclusion & outlook

- ML powerful state-of-the-art techniques for spatial mapping
- Data pre-processing and implementation requires some coding, not in a ready-to-use way included in GIS software
- Recent literature shows that ML outperforms geostatistical models in many cases-> better predictive power
- ML relies on the existence and quality of predictor data
- Prediction is solely based on the site characteristics and average observations for these set of characteristics

→ i.e. measurements nearby are not necessarily considered (contrast to geostatistics)
→ ML gives less weight to individual measurements
→ Information that is not in the predictors (i.e. outcrop of an unmapped small geological unit) won´t be in the map
→ Possible solution hybrid approaches: regression kriging, i.e.

   1) machine learning regression model
   2) Geostatistical interpolation of residuals

   - If we are lucky, the model improves, but it can also reintroduce the noise that we wanted to avoid

Thank you!

Eric Petermann
Federal Office for Radiation Protection (BfS), Berlin, Germany
epetermann@bfs.de