Master's Thesis

TECHNISCHE UNIVERSITÄT DARMSTADT

Using Machine Learning methods for Geogenic Radon potential Mapping in Hessen

Data exploration and processing

By: Augustine M Gbondo
TropHEE

Sublogo

Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

- Data required and acquired
- VECTOR DATA
  - Radon
  - Geology
  - Uranium
- RASTER DATA
  - Soil Hydraulic Properties
  - Soil Physical Properties
  - Soil Chemical Properties
- Final Extracted Dataset *
- Outstanding work
  - Feature selection, Modelling selection and performance metrics
  - Thesis & Publishing best model result online (webmap)

Sublogo

Data Required & Acquired

TECHNISCHE UNIVERSITÄT DARMSTADT

- Field Data
- (GRP data)

- Predictor Data
- 1:300k Geological Map

- Lineament Map
→later processed

- Geological class based on the geological map of Germany, scale 1:1,000,000 (BGR, 1993). Data was re-classified based on the classification used previously for the GRP map of Germany (Bossew, 2015) and further simplified into 30 classes. Classification was mainly done by geological criteria (stratigraphy, petrography and genesis). Further, classes with similar statistical properties were merged to reduce the number of classes for computational reasons and to allow a minimum number of observations in each class. For details see Table 2 (Appendix).
- Soil hydraulic properties in 1000 m resolution (Tóth et al., 2017):
  - o saturated hydraulic conductivity
  - o saturated water content
  - o field capacity
  - o wilting point
  - o parameter α of the hydraulic conductivity curve
- Soil physical properties in 500 m resolution (Ballabio et al., 2016):
  - o clay content
  - o silt content
  - o sand content
  - o coarse fraction
  - o available water capacity
  - o bulk density
- Soil chemical properties in 500 m resolution (Ballabio et al., 2019):
  - o pH in H_2O
  - o cation exchange capacity
  - o carbon:nitrogen ratio
  - o concentration of calcium carbonate
  - o concentration of nitrogen
  - o concentration of phosphorous
  - o concentration of potassium
- Soil uranium concentration in 10 km resolution (Cinelli et al., 2019)
- SAGA wetness index derived from the digital elevation model of Germany (resolution 25 m) (BKG, 2018)
- Climate data in 1000 m resolution (DWD, 2018a, 2018b, 2018c):
  - o Temperature: annual and seasonal means 1981–2010 (DWD, 2018a)
  - o Precipitation: annual and seasonal means 1981–2010 (DWD, 2018b)
  - o Soil moisture: annual and seasonal means 1991–2010 (DWD, 2018c)

- **All above data acquired and further processed**

Sublogo

Vector Data

TECHNISCHE UNIVERSITÄT DARMSTADT

RADON DATA

- Radon Measurement data
  - Two data sets
    1. Bodenluftmessungen_Hessische_Radonstrategie[BHR] (696, 71)
    2. Bodenluftmessungen_Weitere [BW] (255, 68)

    Total of 951 measured points

RADON_RN_222_KBQM3	PERMEABILITAET_M2	PERMEABILITÄT_E15M2	RADONPOTENZIAL
131	4,06E-13	406.382461	54,79
311	1,14E-12	1139.875066	160,05
163	1,04E-12	1038.702723	82,18
92.16	1,43E-12	1429.068544	49,95
144.384	5,01E-13	501.011802	62,77

Dataset BW

- Projected CRS: ETRS89 / UTM zone 32N EPSG : 25832

This CRS was used for the whole data exploration and cleaning process

Preliminary data exploration

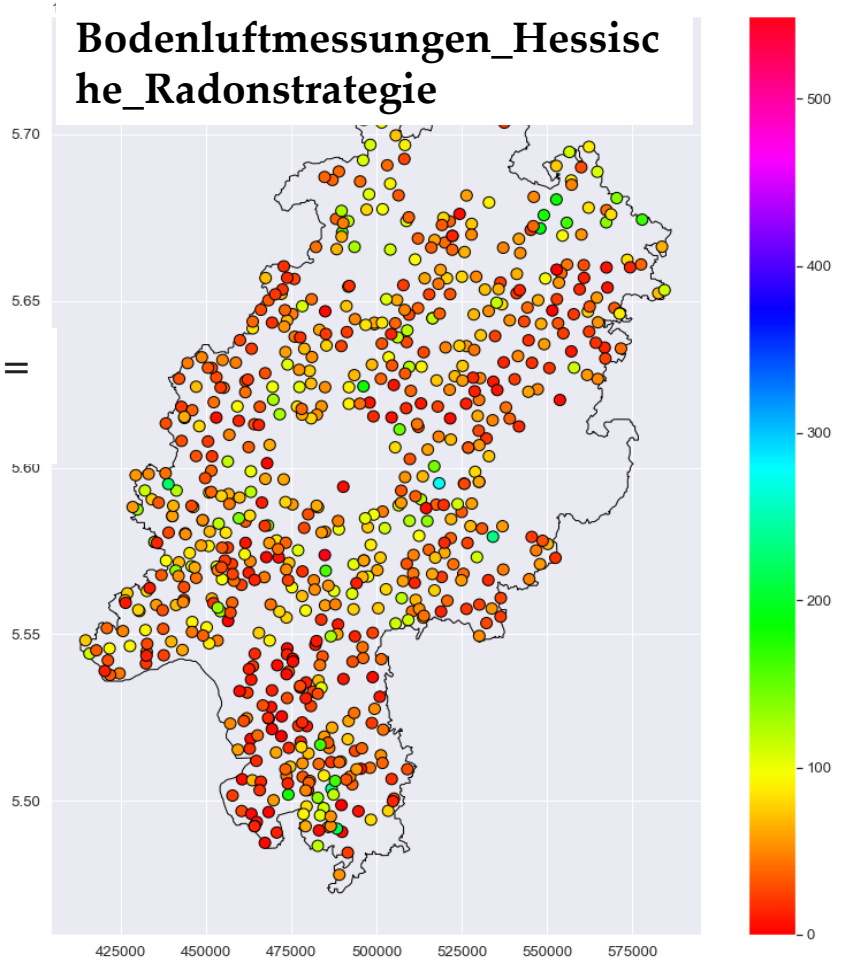
- Radon 222 (Bqm3) & (KBqm3)
- For Dataset 'BW'
  - PERMEABILITÄT_E15M2 = 59 available corresponding to the GRP
- Huge difference in permeability readings
  - BHR range = 0.02 - 225
  - BW range = 19 - 2396

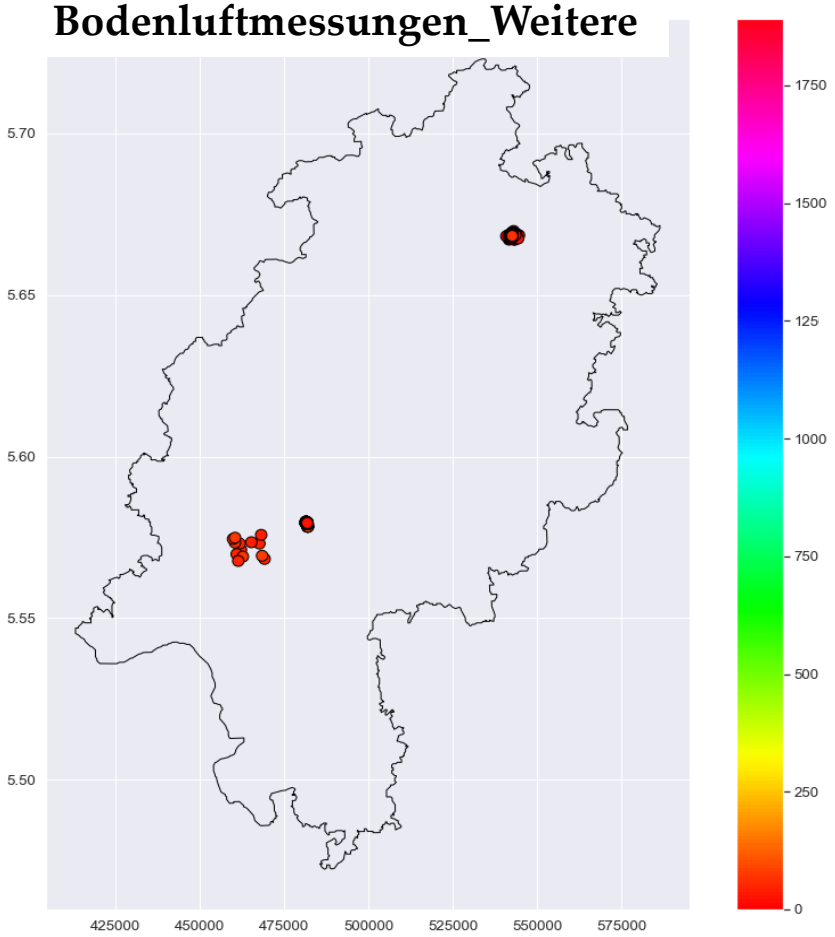
Sublogo

Vector Data

Maps
Radon Data

Sampling no. = 696





Sampling no. = 255

- Projected CRS: ETRS89 / UTM zone 32N
- EPSG : 25832

Sublogo

Vector Data

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Cleaning

- Radon data
  - Merged dataset
    951 data pts, 95 features (attributes)

- Features of Interest
  - GK300_GEOL
  - RADON_RN_222_KBQM3
  - RADON_RN_222_BQL
  - PERMEABILITÄT_E15M2
  - RADONPOTENZIAL
  - RADONKONZE

RADONPOTENZIAL

Extremely high values

RADON_RN_222_KBQM3	PERMEABILITAET_M2	PERMEABILITAET_E...	RADONPOTENZIAL ▼
68.22256	0	95.234932	3217.480077
78.32012	0	69.254313	490.871595
38.8309	0	83.252526	487.809554

removed

Negative values

RADON_RN_222_KBQM3	PERMEABILITAET_M2	PERMEABILITAET_E...	RADONPOTENZIAL ▲
75.3664	0	169.258542	-329.758055
53.88105	0	165.173085	-247.22964
15.27105	0	115.65987	-241.696703
34.7245	0	225.444893	-98.358431
5.80988	0	124.878789	-60.213075

removed

Column Formatting

RADON_RN_222_KBQM3	PERMEABILITAET_M2	PERMEABILITÄT_E15M2	RADONPOTENZIAL
131	4,06E-13	406.382461	54,79
311	1,14E-12	1139.875066	160,05
163	1,04E-12	1038.702723	82,18
92.16	1,43E-12	1429.068544	49,95
144.384	5,01E-13	501.011802	62,77

-- Reformat column to float

-- Replace (“,” with “.”

Sublogo

Vector Data

Filling N/A Data

1. Data is available

2. Using GRP/CRn eq.

   GRP = CRn / (−log10 (k) − 10)

   (Method ineligible because)

   - K map values are negative values

3. Correlation and regression of 'BHR Map'

   (Method ineligible because)

   - Many factors affect CRn
   - Parameter 'k' is not always continue in space

Correlation and regression of 'BHR Map'

--Strong correlation between [Observed and Modelled RADONPOTENIAL for IDW interpolation

-- Regression equation shows strong promise

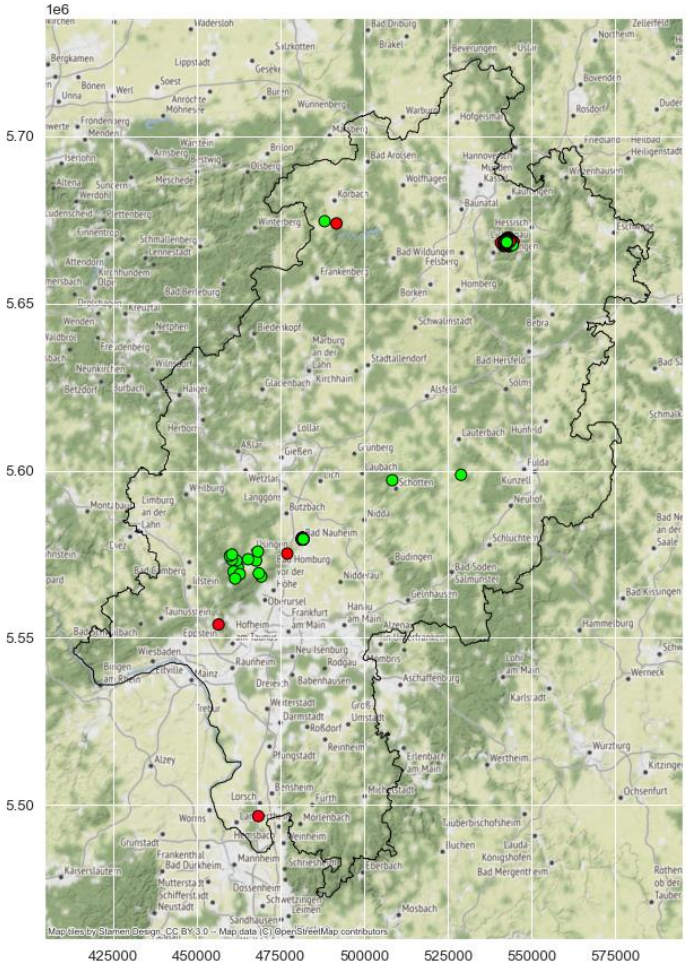
Sublogo

Vector Data

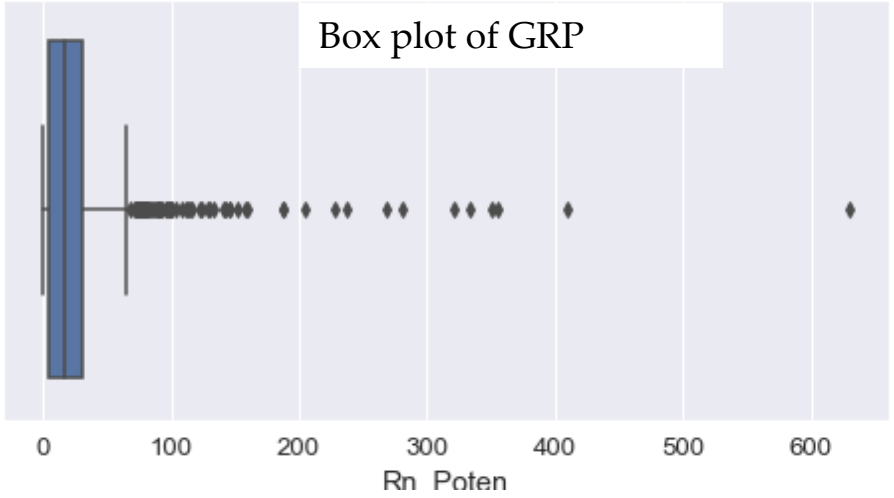
3. Correlation and regression of 'BHR Map'

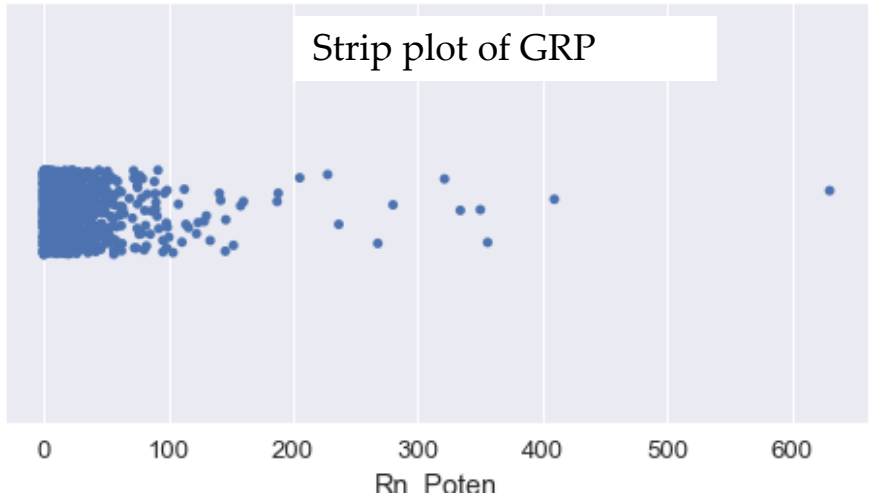
TECHNISCHE UNIVERSITÄT DARMSTADT

Dealing with Outliers



-- **Bad Nauheim** and **Kerenbach** have the highest number of null values





Box plot stats
Mean = 26.1
iqr = 25.1
q1 = 4.8
median = 17.1
q3 = 29.9

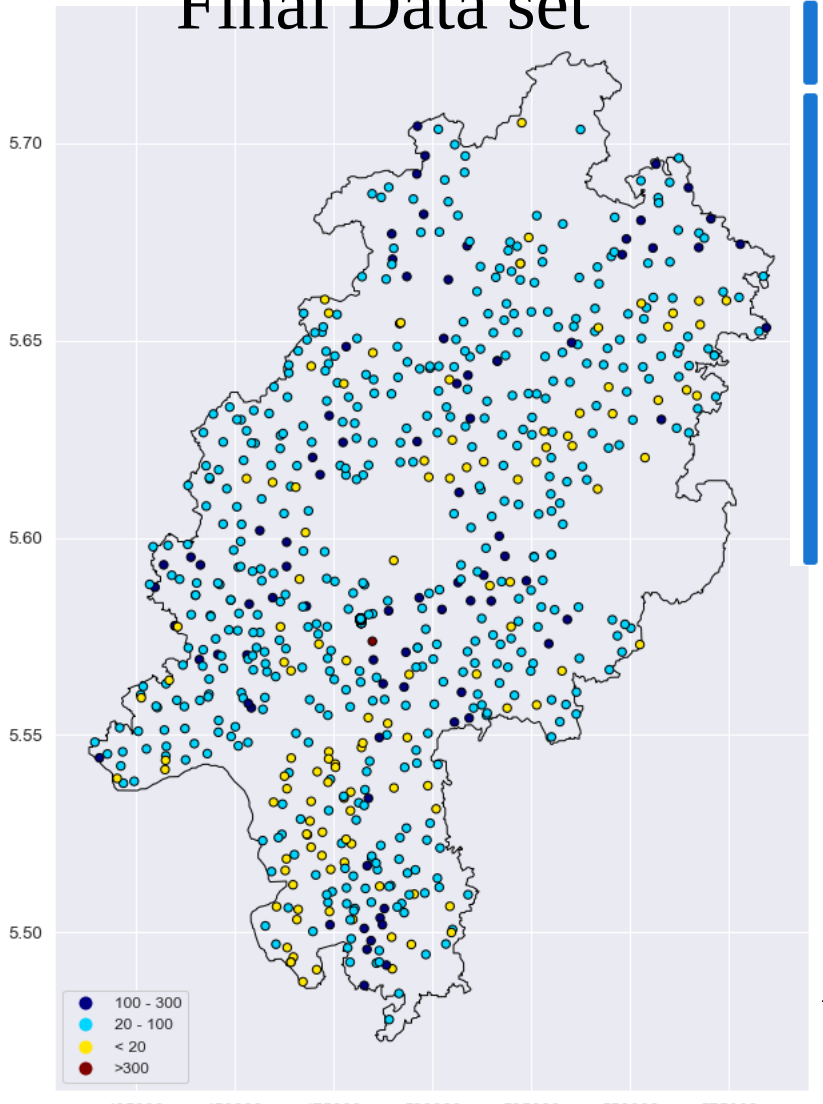
-- Most values are below 100 KBqm3

Sublogo

Vector Data

TECHNISCHE UNIVERSITÄT DARMSTADT

Final Data set




```
[183]: print(df4.shape)
       df4.head(3)
```

```
(744, 17)
```

[183]:

MESS_ID	OSTWERT	NORDWERT	Rn_Poten	PERMEABILI	PERMEABI_1	RADON_RN_2	RADON_FEHL	RADON_RN_3	CO2_PPM	LANDK
62.0	481853.162	5578395.410	64.28	1,28E-13	127.752997	186000.0	20518.284529	186.000	0.0	Wett
63.0	481881.152	5578355.426	92.14	1,08E-13	108.021901	273333.0	16072.751268	273.333	0.0	Wett
64.0	481914.138	5578338.433	40.93	4,50E-14	44.997924	137000.0	11135.528726	137.000	0.0	Wett

-- Final Measure points of 744

- Features of interest
  - Strata (hlnug)
  - Rn222 conc. (Bqlm3)
  - Rn222 conc. (KBqlm3)
  - Rn222 conc. (Intervals)
  - Permeability (e-15m2)
  - Cordinates & Geometry(for python interface)
  - Landkreis

Sublogo

Vector Data

TECHNISCHE UNIVERSITÄT DARMSTADT

GEOLOGY DATA

- Geology data
  - Two data sets
    1. GUEK300_GEOLOGIE (176, 11)
    2. GUEK300_englisch_Laura_Eck_2021 (176, 13)

    Total of 176 polygon geometries in each dataset

- Projected CRS: ETRS89 / UTM zone 32N EPSG : 25832

```
df300k = pd.DataFrame(geo_x())
df300k
```

[66]:

	unique values	null count
FORMATION	65	0
PETROGRAPH	76	0
SUBGRUPPE	9	94
GRUPPE	12	68
SUBSERIE	2	119
SERIE	17	27

-- GUEK300_GEOLOGIE

- 8 different Geology classification

- Only FORMATION, PETROGRAPH and SYSTEM has complete dataset

```
    c2.columns = ['unique values','null count']
    return(c2)

dfGeo = pd.DataFrame(geoclass_Geo())
dfGeo
```

[5]:

	unique values	null count
STRATIGRAP	65	25
Strat	26	0
strata_condense	12	0
hlnug	19	0

-- GUEK300_Laura_Eck

- 4 different Geology classification

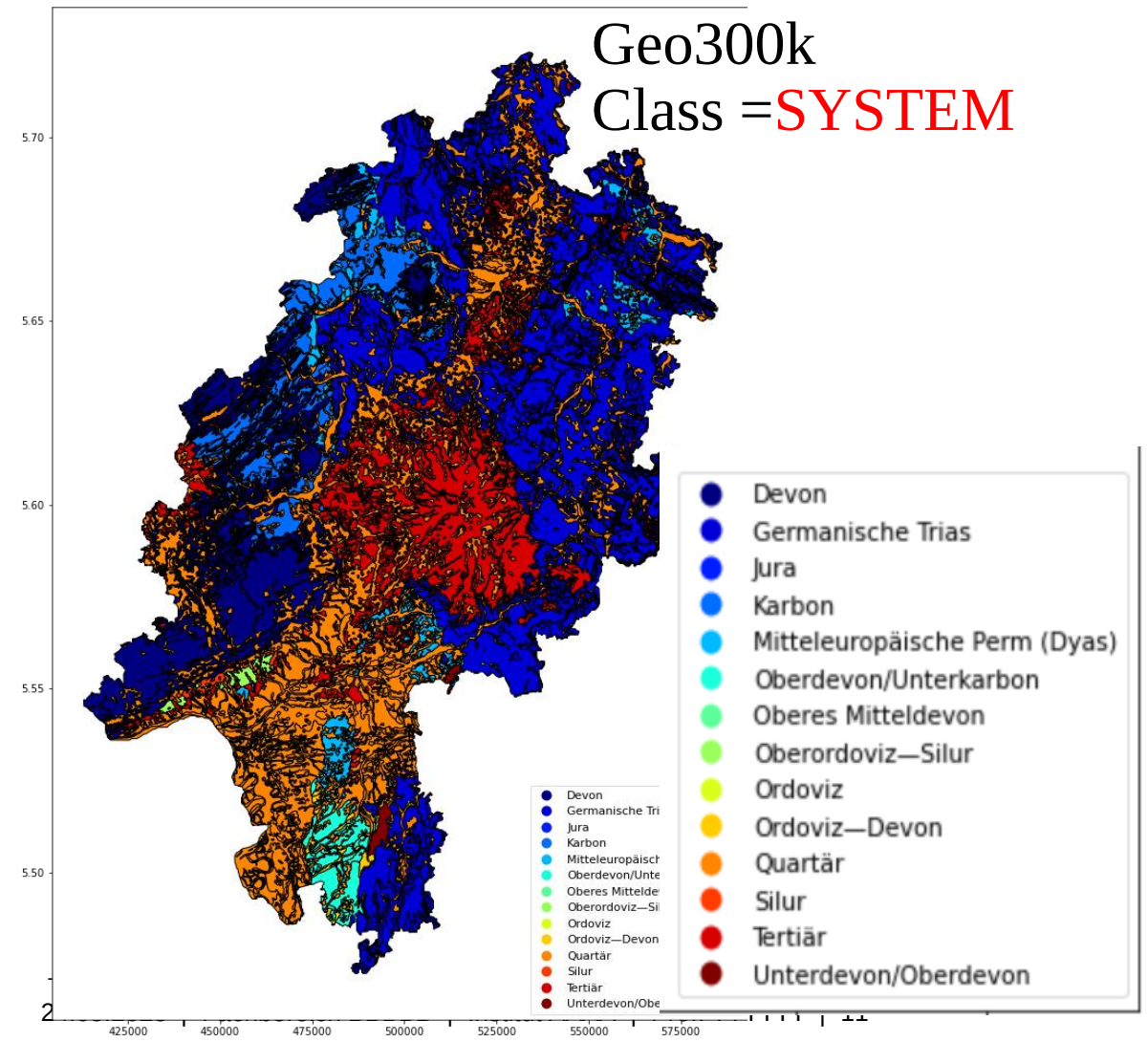
- Only STRAT, ‘strat_condense’ and ‘hlnug’ has complete dataset

Sublogo

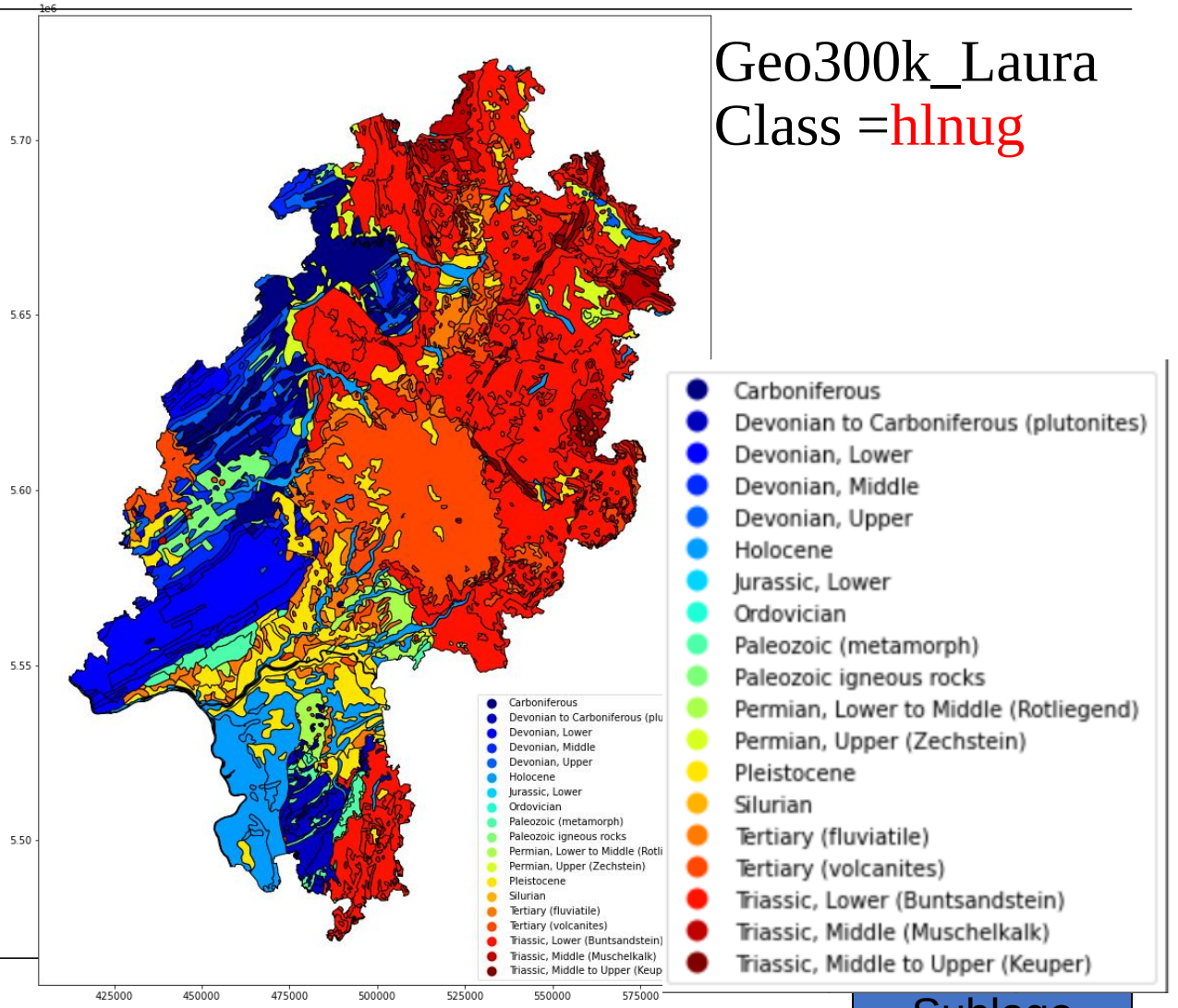
Comparing Geology classification 1.

TECHNISCHE UNIVERSITÄT DARMSTADT

Geo300k
Class =SYSTEM



Geo300k_Laura
Class =hlnug

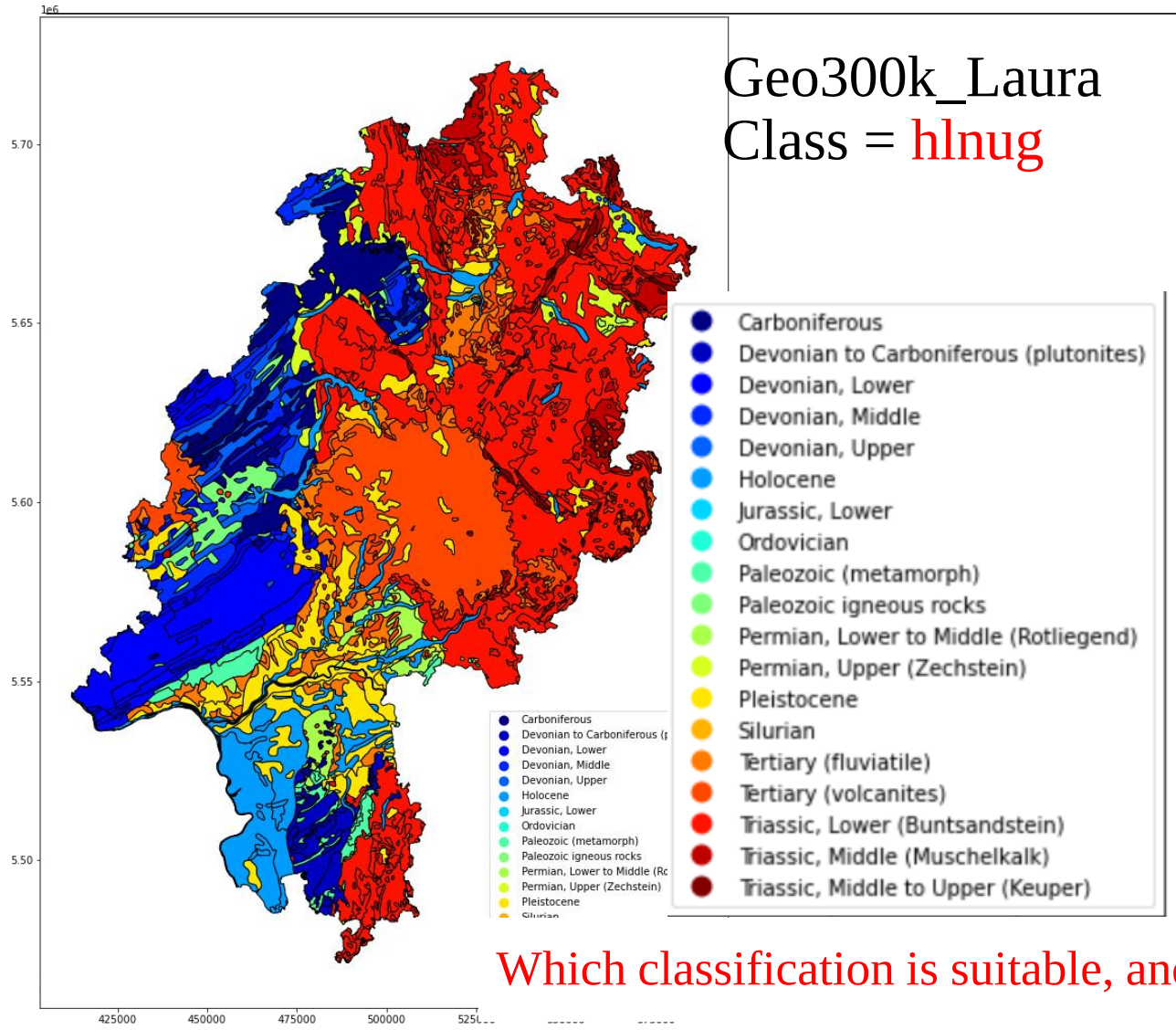


Sublogo

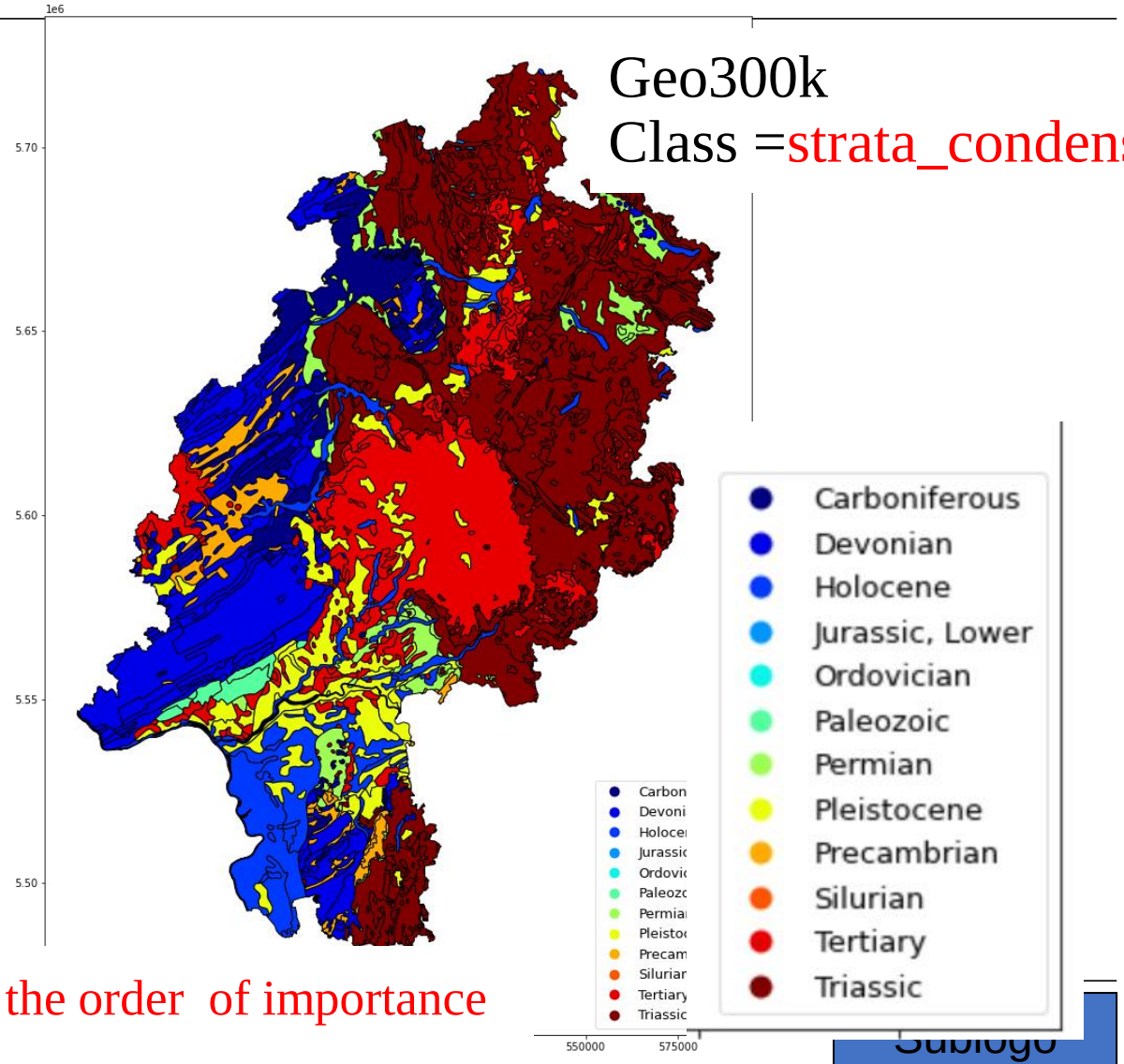
Comparing Geology classification 2.

TECHNISCHE UNIVERSITÄT DARMSTADT

Geo300k_Laura
Class = hlnug



Geo300k
Class =strata_condense



Which classification is suitable, and the order of importance

Vector Data

Vector points to raster

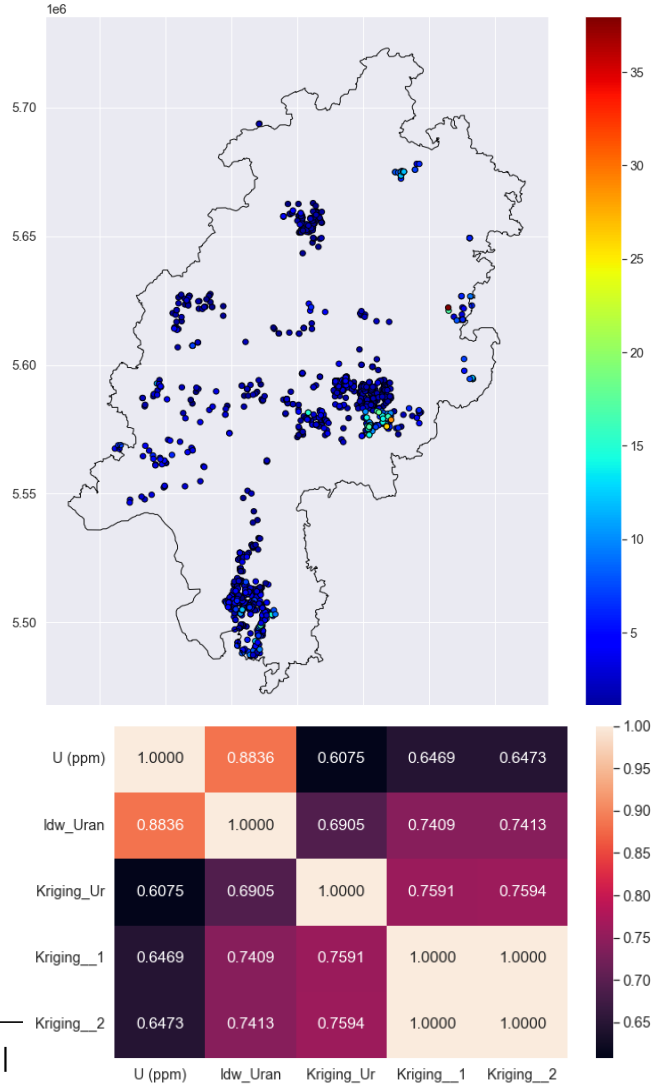
TECHNISCHE UNIVERSITÄT DARMSTADT

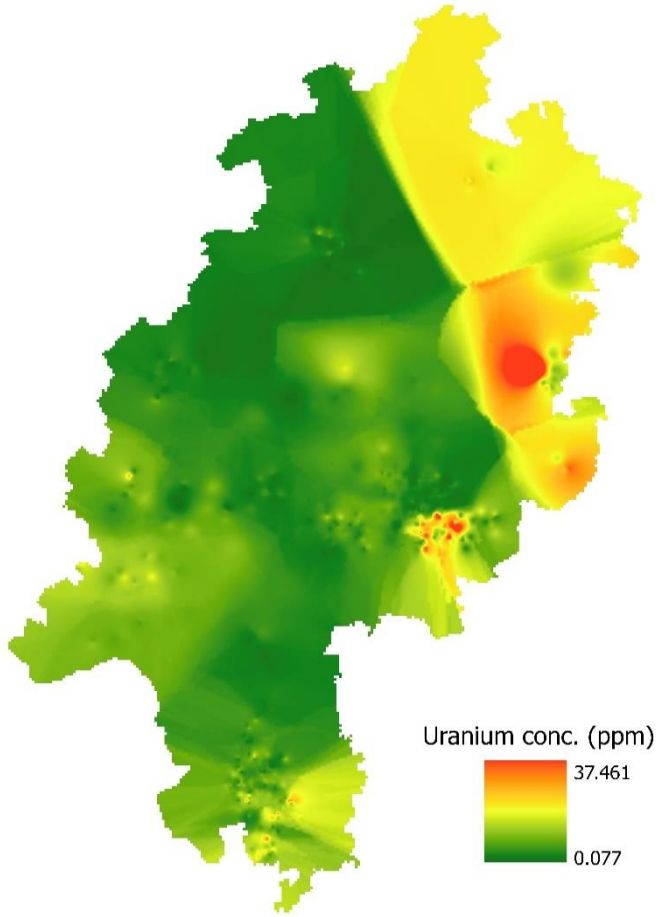
Uranium Data

- 1 data sets
  - Uran_ppm_Rohda
    One pt with 238 ppm
    was removed

- Projected CRS: ETRS89 / UTM zone 32N EPSG : 25832

- Four interpolation techniques were used. 1 IDW and 3 Kriging.
  -IDW showed the best result





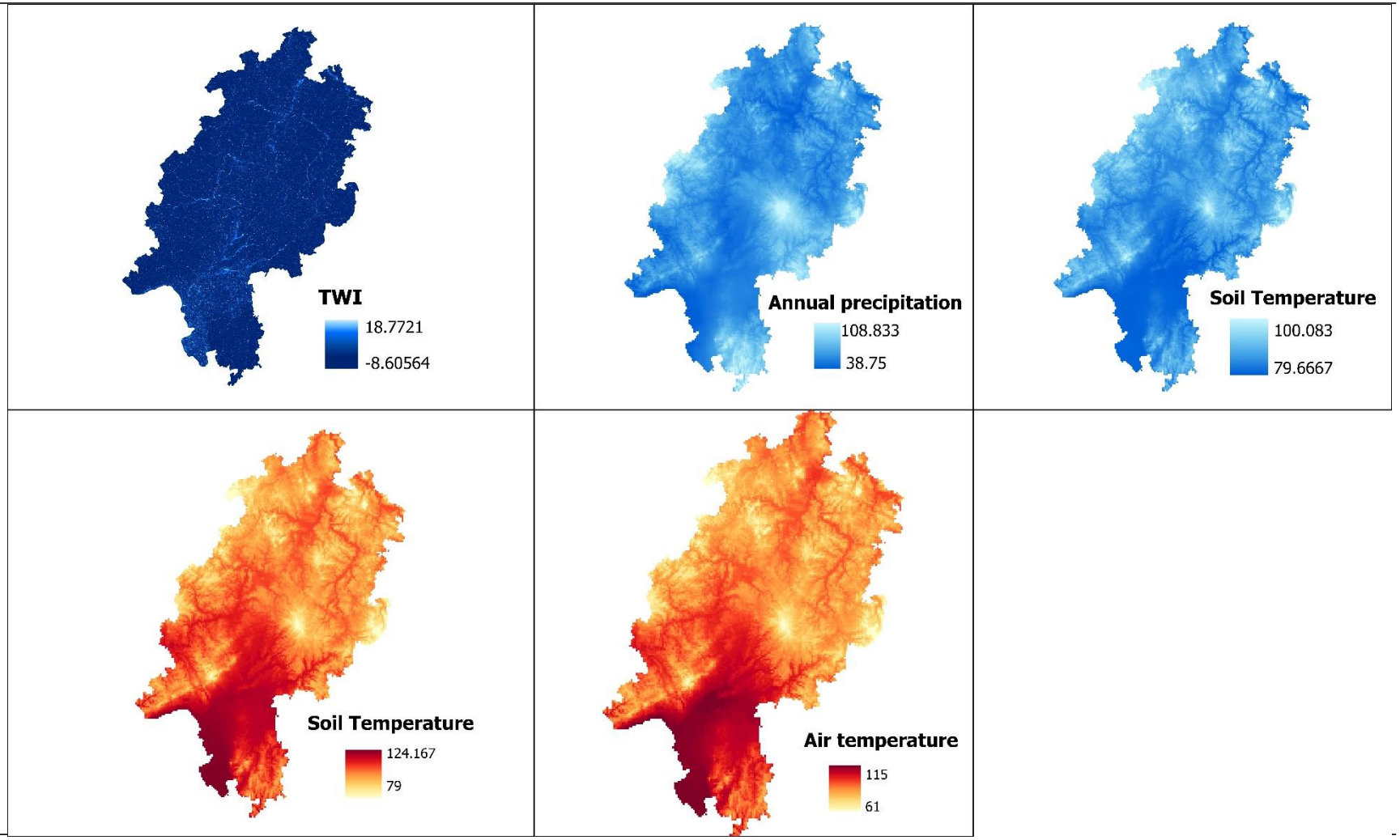
Sublogo

Raster Data

Climate Data

- Climate data (Annual average from 1991-2020)
  - Three data sets
    1. Precipitation
    2. Air Temperature
    3. Soil Moisture
    4. Soil Temperature
    5. SAGA wetness Index

Projected from: DHDN-3degGK-Zone3
To
Projected CRS: ETRS89 / UTM zone 32N EPSG : 25832



https://opendata.dwd.de/climate_environment/CDC/grids_germany/seasonal/precipitation/

Raster Data

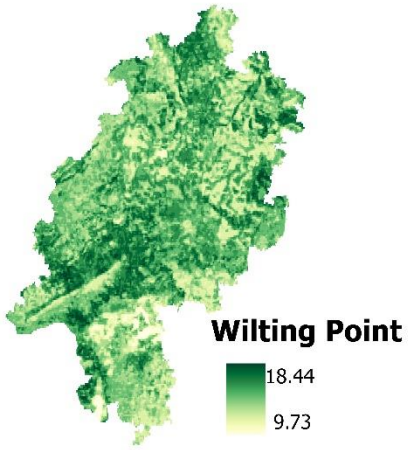
TECHNISCHE UNIVERSITÄT DARMSTADT

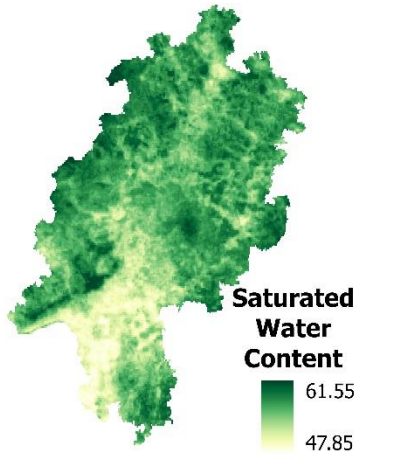
SOIL PROPERTIES

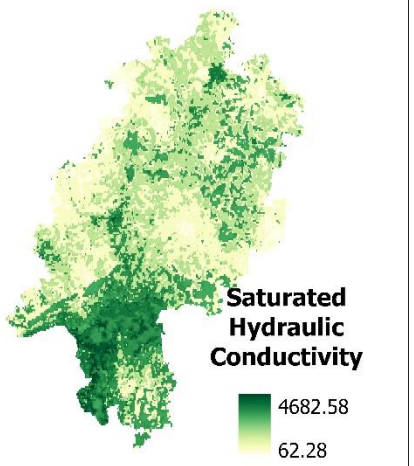
- Soil Hydraulic properties (weighted mean of 7 layers)
  - 4 data sets
    1. Saturated Water Content (250km)
    2. Saturated hydraulic conductivity (KS)
    3. Field Capacity
    4. Wilting point

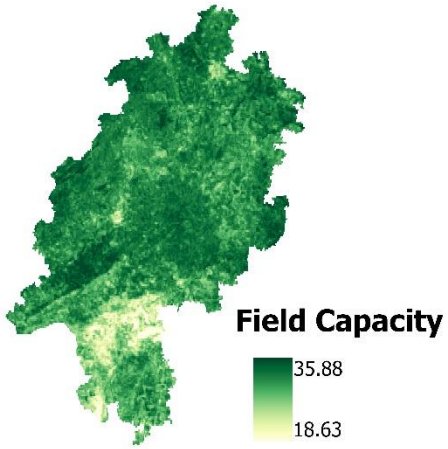
-- Projected from: GCS: WGS 1984
To
Projected CRS: ETRS89 / UTM zone 32N
EPSG : 25832

-- 1k datasets resampled to 250m









Determination of weighted sum

-- 7 sampling depth of (0cm, 5cm, 15cm, 30cm, 60cm, 100cm, 200cm

-- Hydraulic properties decrease with depth

Weighted sum = [soil depth/total soil depth]
Weights were inversed

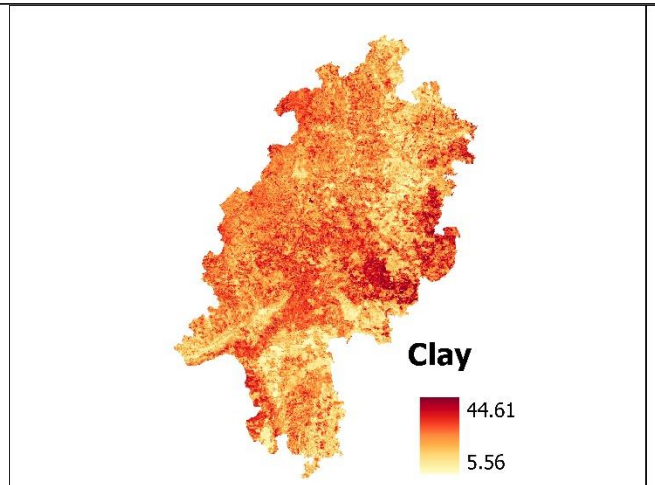
 https://esdac.jrc.ec.europa.eu/tmp_dataset_access_req_68430

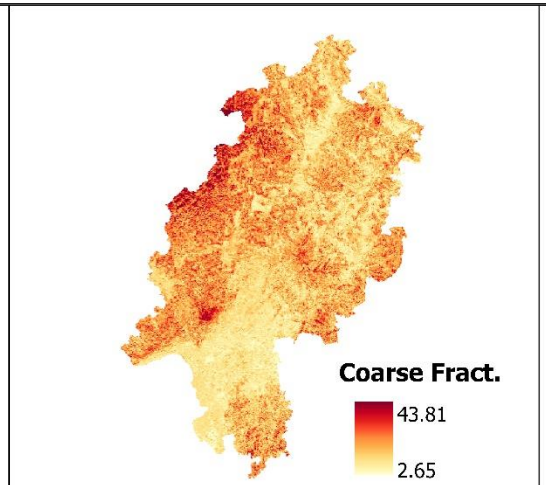
Raster Data

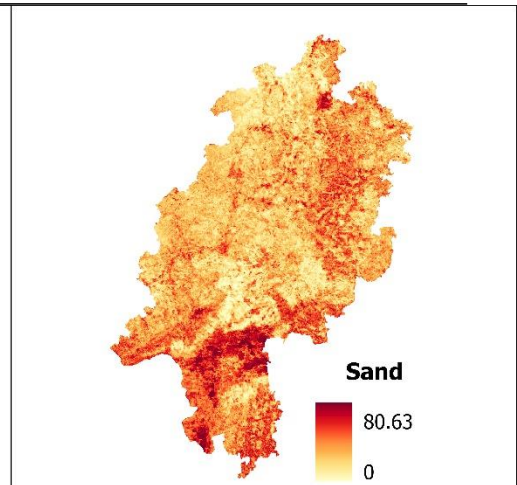
TECHNISCHE UNIVERSITÄT DARMSTADT

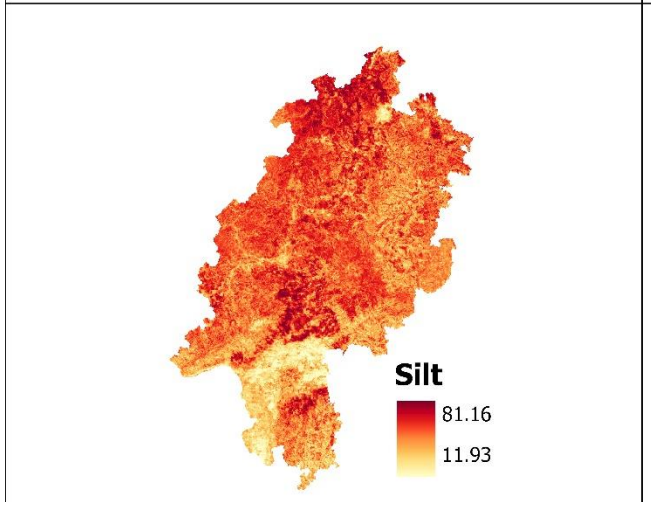
SOIL PROPERTIES

- Soil Physical properties
  - 4 data sets
    1. Clay (250m)
    2. Silt
    3. Sand
    4. Coarse fraction









-- Projected from: GCS: WGS 1984

To

Projected CRS: ETRS89 / UTM zone 32N
EPSG : 25832

-- 1k datasets resampled to 250m

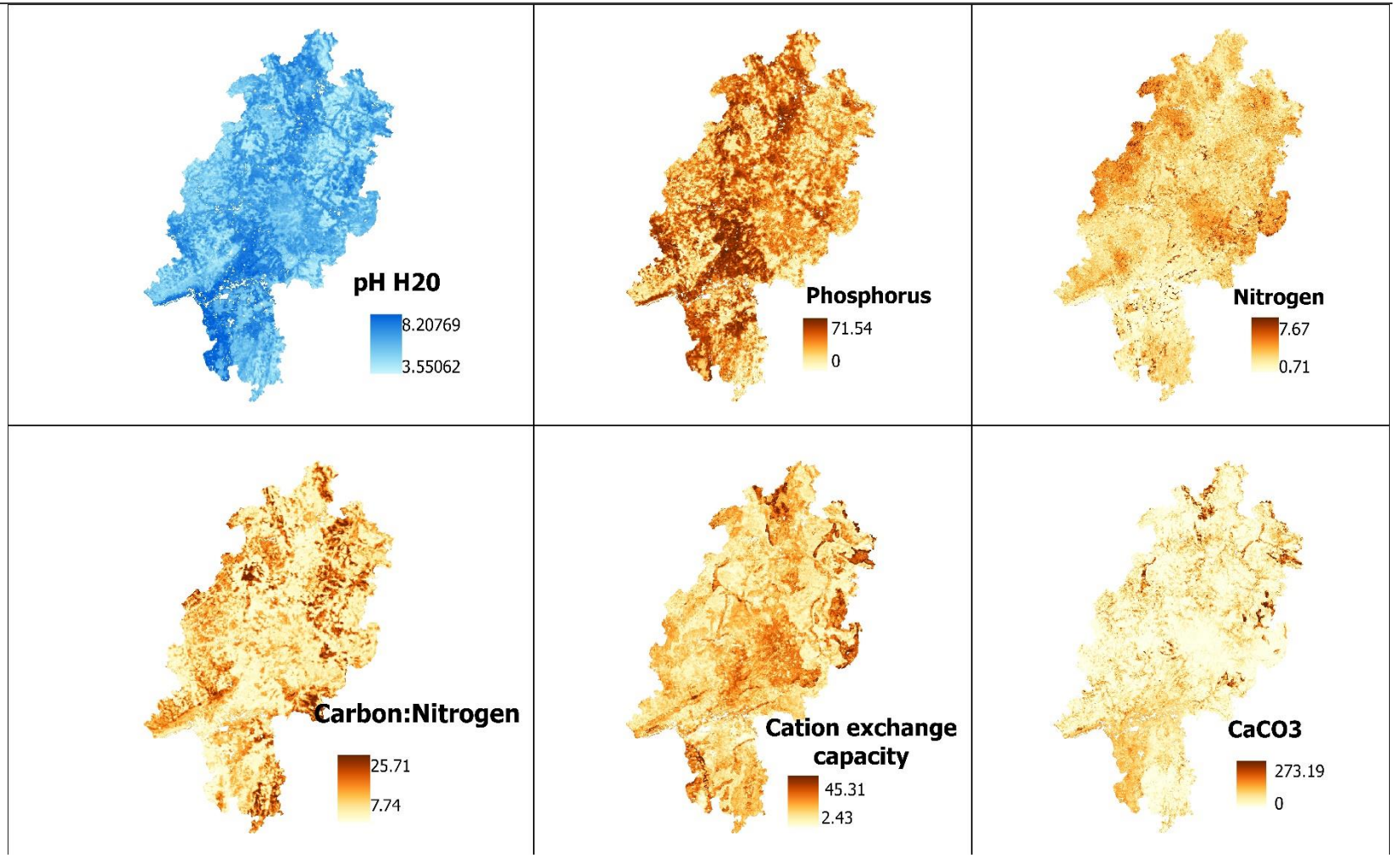
https://esdac.jrc.ec.europa.eu/tmp_dataset_access_req_68430

Sublogo

Raster Data

SOIL PROPERTIES

- Soil Chemical properties
  - 6 data sets
    1. PH (250m)
    2. Cation exchange
    3. Carbon:Nitrogen (C_N)
    4. CaC03-
    5. Nitrogen
    6. Phosphorus



-- Projected from: GCS: WGS 1984

To

Projected CRS: ETRS89 / UTM zone 32N
EPSG : 25832

-- 1k datasets resampled to 250m

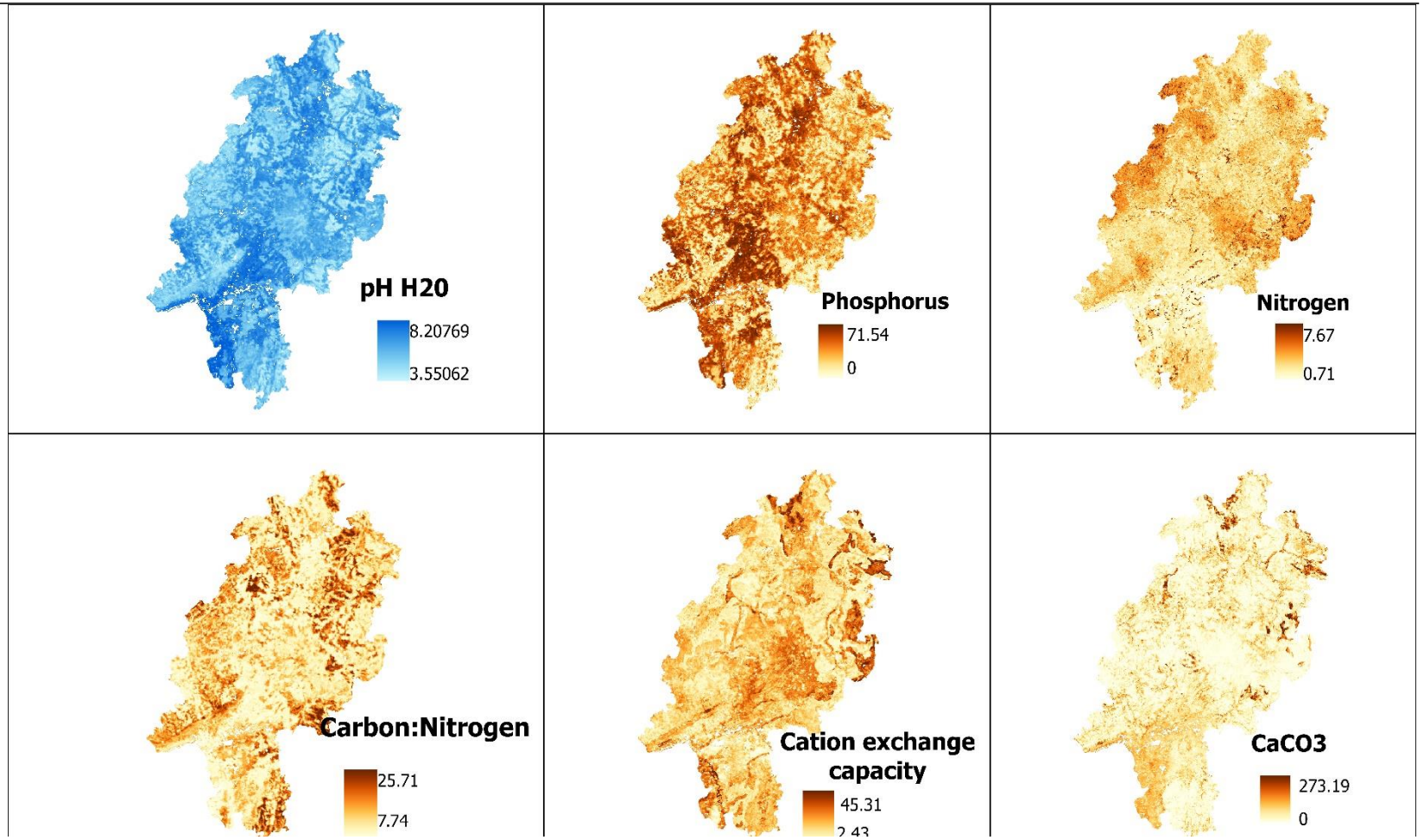
https://esdac.jrc.ec.europa.eu/tmp_dataset_access_req_68430

Sublogo

Raster Data

SOIL PROPERTIES

- Soil Chemical properties
  - 6 data sets
    1. PH (250m)
    2. Cation exchange
    3. Carbon:Nitrogen (C_N)
    4. CaC03-
    5. Nitrogen
    6. Phosphorus



-- Projected from: GCS: WGS 1984

To

Projected CRS: ETRS89 / UTM zone 32N
EPSG : 25832

-- 1k datasets resampled to 250m

https://esdac.jrc.ec.europa.eu/tmp_dataset_access_req_68430

Sublogo

Final Extracted Data *

TECHNISCHE UNIVERSITÄT DARMSTADT

[48]:
```
## Radon master data with all features
print(Rn_master.shape)
Rn_master.head(5)
```

```
(744, 41)
```

[48]:

	MESS_ID	OSTWERT	NORDWERT	Rn_Poten	PERMEABILI	PERMEABI_1	RADON_RN_2	RADON_FEHL	RADON_RN_3	CO2_PPM	...	H
0	62.0	481853.162	5578395.410	64.28	1,28E-13	127.752997	186000.0	20518.284529	186.000	0.0	...	
1	63.0	481881.152	5578355.426	92.14	1,08E-13	108.021901	273333.0	16072.751268	273.333	0.0	...	
2	64.0	481914.138	5578338.433	40.93	4,50E-14	44.997924	137000.0	11135.528726	137.000	0.0	...	
3	65.0	481939.128	5578313.443	39.72	9,72E-14	97.164279	119667.0	6506.407099	119.667	0.0	...	

Sublogo

Outstanding work

TECHNISCHE UNIVERSITÄT DARMSTADT

- Feature selection
  - Algorithm being built
- Model selection and building
  - In progress
- Model validation and performance metrics
  - Not yet started
- Publishing results as a webmap
  - Not yet started
- THESIS
  - Literature review and structure of thesis in progress

Sublogo

TECHNISCHE UNIVERSITÄT DARMSTADT

Sublogo