# ANR HOUSES

Harmonized Operation of Uncertainties in Spatialized Environmental Systems

Definition of the synthetic cases to discuss challenges in uncertainty assessment

**December 2023**

# Programme

13h30 - 13h45 : news du projet, rappel objectifs

13h45 - 14h45 : description cas synthétiques + premiers tests

14h45 - 15h30 : discussion, organisation

15h30 – 15h45 : pause

15h45  - 16h45 : application d’une nouvelle approche par Stéphane aux cas synthétiques (projet ISLANDER)

16h45 - 17h15 : résumé et suite

# Quelques nouvelles du projet

## Nouveau recrutement

- Priscillia Labourg (**bienvenue!**) – thèse WP3 CNRS/UTC (Sébastien + Benjamin) avec IRIT (Romain + Hélène) + BRGM (Stéphane et moi)

## Administratif

- Accord de consortium: **OK** (sur site ANR, version papier à venir)
- Accord de reversement thèse de Priscillia: **vient très vite** (lenteur administrative interne)
- Plan de gestion des données: **OK** (sur site ANR)

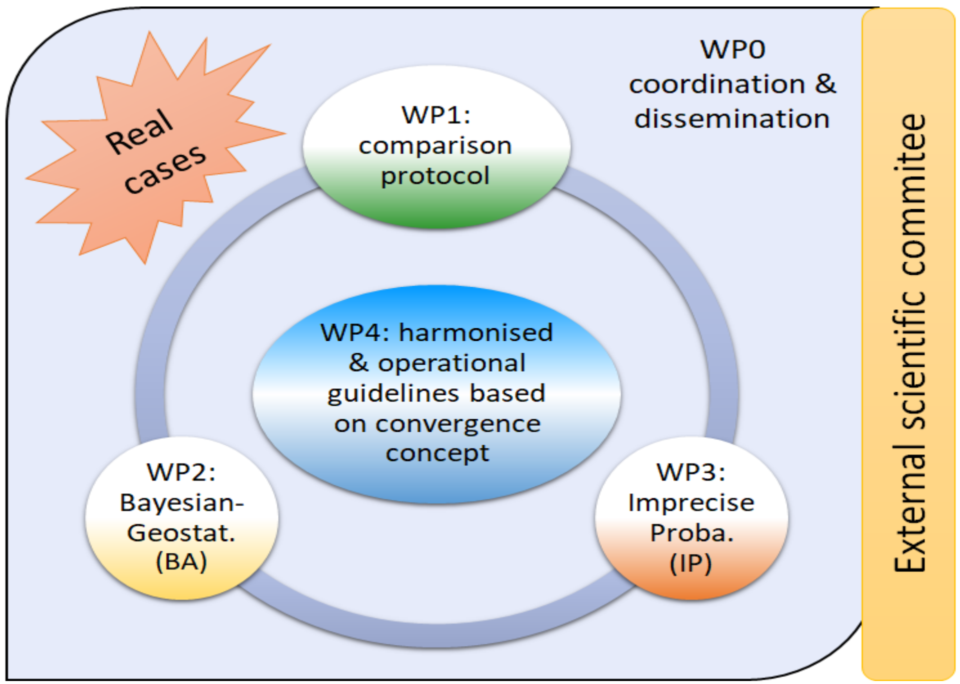
## Communication https://anrhouses.github.io/talks/

- **Article :** Belbèze, S., Rohmer, J., Négrel, P., & Guyonnet, D. (2023). Defining urban soil geochemical backgrounds: A review for application to the French context. JGE, 107298.
- **Journees de la Geostatistique :** Spatial prediction with spatially clustered data based on transfer learning,
- **GDR MADICS:** Explicability in Machine learning for Geoscience processes
- Faites moi remonter vos actions!
- A venir ?
  - EGU, geostats2024, GeoEnv 2024
  - IPMU2024
  - Autres…

# …Motivation for HOUSES…

## Objective:

Define a **harmonized framework** to exhaustively and transparently reflect **all uncertainties** along the **modelling chain** of **spatial data** while keeping **track of their origins** (knowledge imperfection and/or random variability)




**Budget** (ANR grant): 582 keuros; (total): 1.23 Meuros

**Duration:** 42 months (Starting date **3 April 2023**)

## Early career scientists:

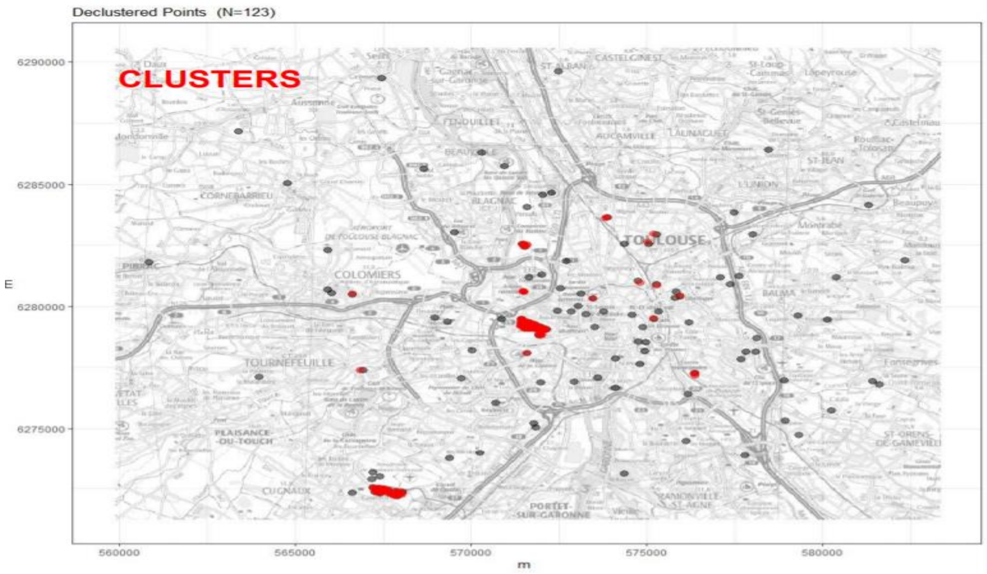
- 1 18-month post-doc (WP2)
- 1 12-month post-doc (WP4)
- 1 Phd (WP3, 1/2 salary)
- 1 research engineer (WP3)

(1) Géosciences pour une Terre durable brgm (2) ARMINES (3) heudiasyc (4) IRIT (5) HESUS

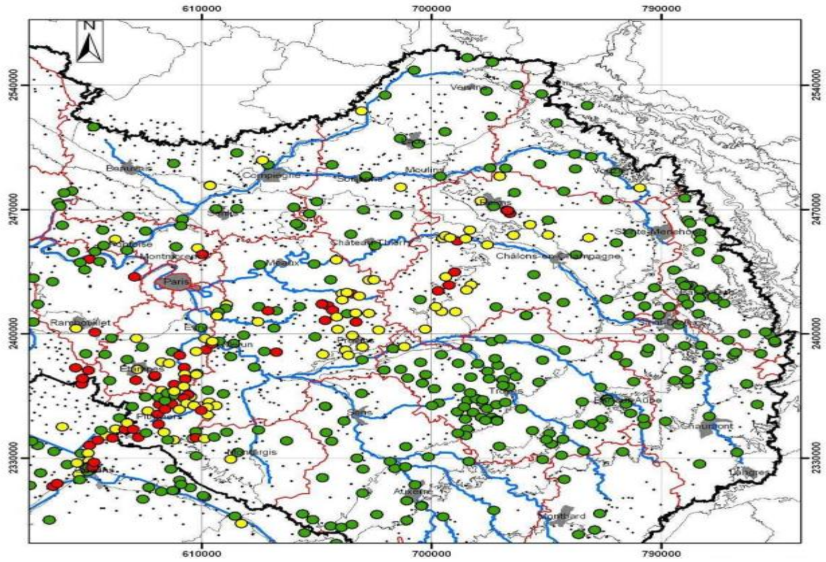
| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Statistics for environments | | | | | |
| Geostatistics | | | | | |
| Bayesian analysis | | | | | |
| Imprecise probability | | | | | |
| Decison making under uncertainty | | | | | |
| Operational use | | | | | |

# WP1 - Test cases for developments

Experiments based on **real** **S**parse **I**mprecise **C**lustered cases




*Sparse and clustered data for geochemical background mapping in Toulouse city* ***[Belbèze et al. 2019]***




*Clustered data for Trace elements' concentrations over a very large area in Paris basin* ***[Gourcy et al. 2011]***

# WP1 - Test cases for developments

## Random experiments based on large datasets (1/2)

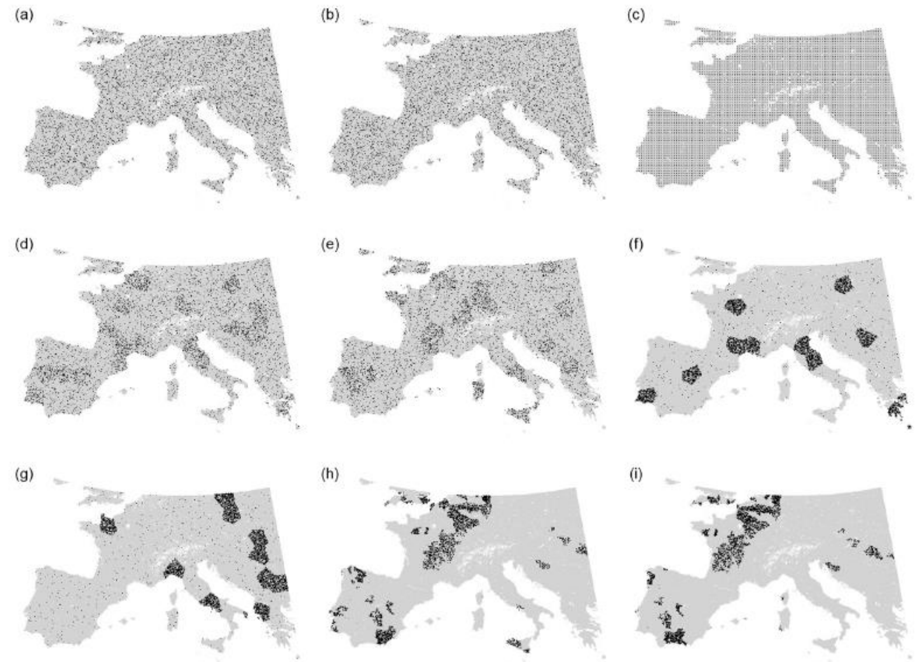



Fig. 2. Examples of studied spatial samples. (a-b) simple random samples; (c) systematic random sample; (d-e) moderately clustered samples; (f-g) strongly clustered samples; (h-i) strongly clustered, gapped samples. Except for the systematic sample (c), the sample size always amounted to 5000. The systematic sample had an expected size of 5000 but realized samples varied in size between 4998 and 5056.

ELSEVIER

Ecological Informatics
Volume 69, July 2022, 101665

### Dealing with clustered samples for assessing map accuracy by cross-validation

Sytze de Bruin [a], Dick J. Brus [b], Gerard B.M. Heuvelink [c], Tom van Ebbenhorst Tengbergen [a], Alexandre M.J-C. Wadoux [d]

**Response:**
Soil organic carbon stock (OCS) in the 0–30 cm layer (Soilgrids)

**Covariates:**

- Seven terrain properties derived from the digital elevation model EU-DEM version 1.1 (Copernicus Land Monitoring Service - EU-DEM — European Environment Agency (europa.eu));
- GEDI forest height (Potapov et al., 2021);
- Seven CHELSA V2.1 climate variables (Karger et al., 2020);
- Seven generalized land cover classes derived from the 2017 Copernicus land cover map (Buchhorn et al., 2020);
- Three soil properties from SoilGrids (only used for predicting AGB);
- Two spatial coordinates (x, y) and distance from the coast, the latter computed using a land mask of the study area that was derived from the other covariates.

https://zenodo.org/record/6513429#.ZBnDunbMJPY

# Objectives

## Compare / discuss methods for uncertainty management

- Not a hackathon
- But ‘challenge problems’ as a basis for discussion

ELSEVIER
Reliability Engineering & System Safety
Volume 85, Issues 1–3, July–September 2004, Pages 11-19

Challenge problems: uncertainty in system response given uncertain parameters

William L. Oberkampf [a], Jon C. Helton [b], Cliff A. Joslyn [c], Steven F. Wojtkiewicz [d],

## Six experiments

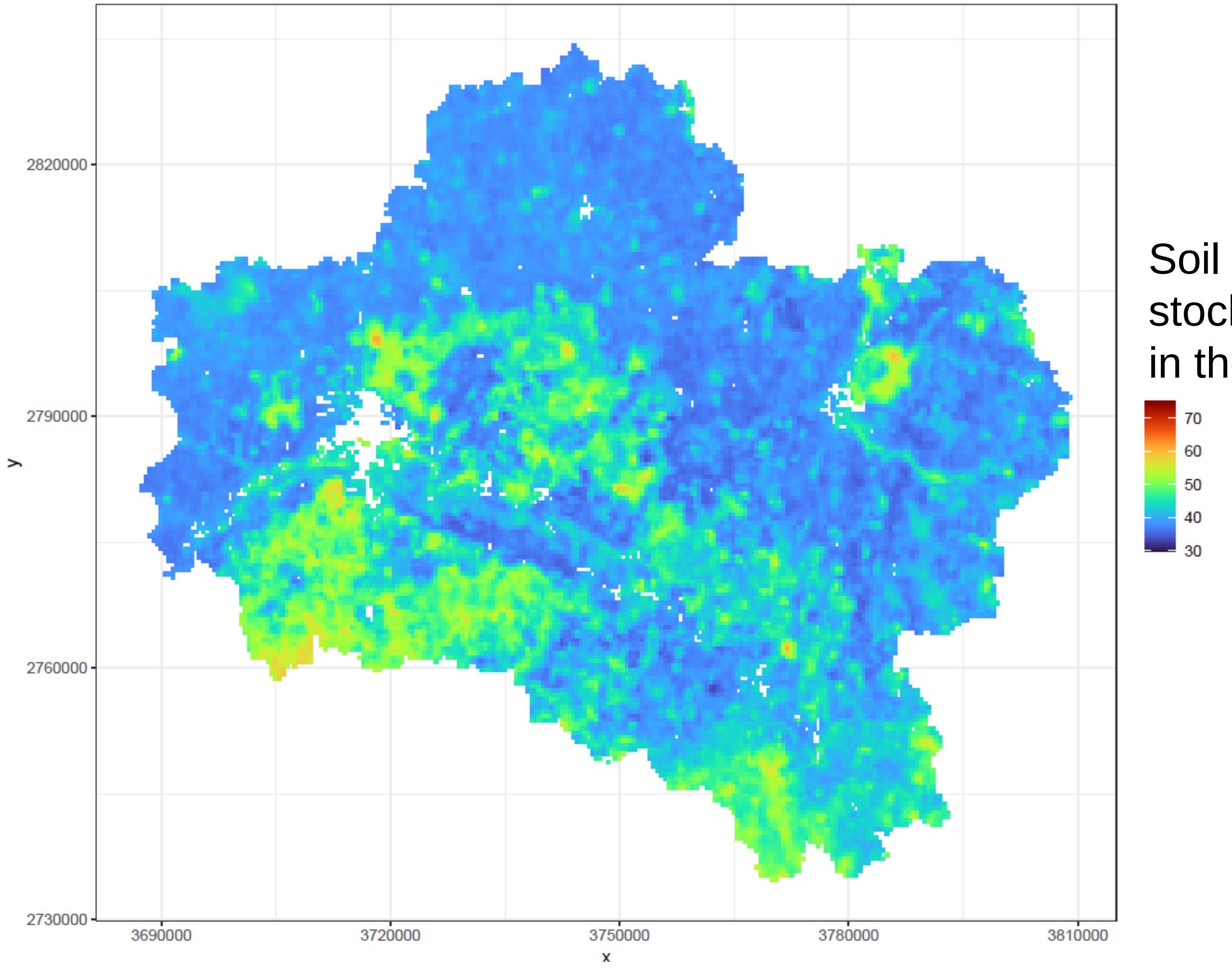
- **Reference clustered case ‘C’:** moderate number of points (200)

- **Sparse case 1 ‘SC1’:** clustered with 100 points
- **Sparse case 2 ‘SC2’:** clustered with 50 points

- **Imprecise cases:**
  - **‘CIo’** With outliers: 10 outliers outside the clusters
  - **‘CIc’** With left-censored data: <median OCS (outside the clusters)
  - **‘CIi’** With intervals: relative error of 30% (outside the clusters), of 10% (within the clusters)

- **With or without** covariates ‘**XX-cov**’ (14)

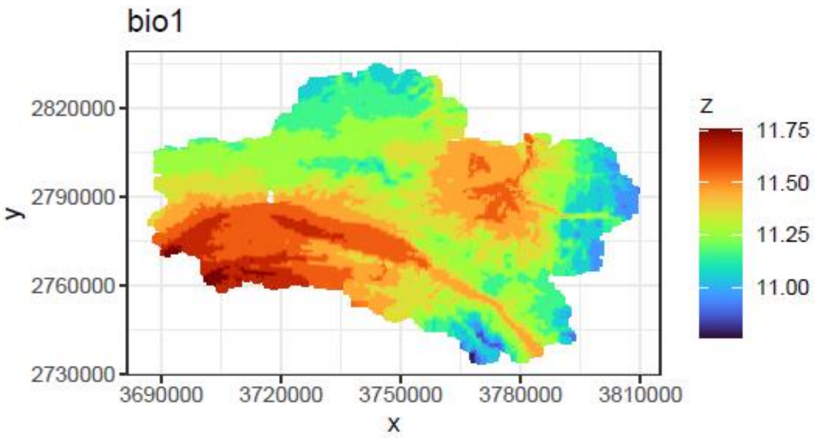
# Variable of interest over Loiret department

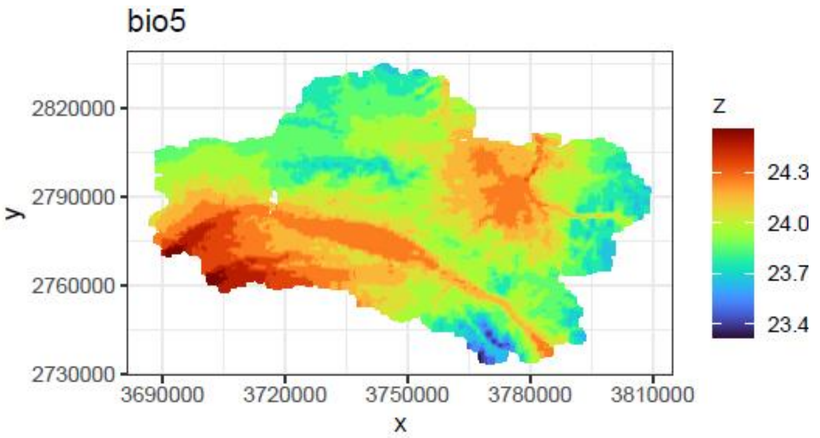


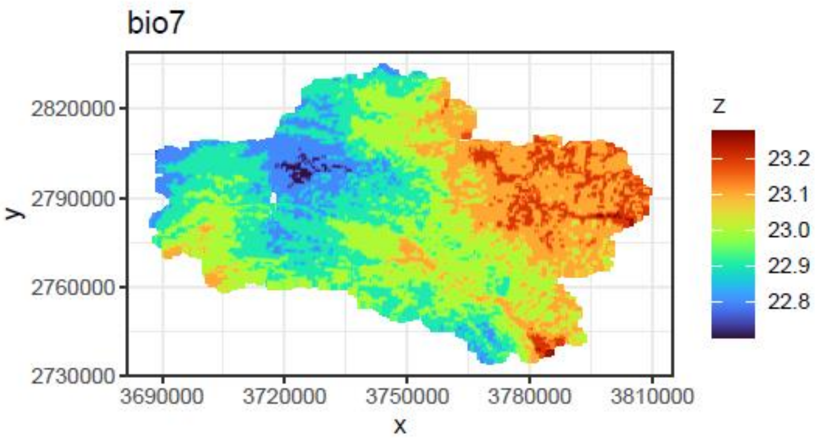

Soil organic carbon stock (OCS, Mg/ha) in the 0–30 cm layer

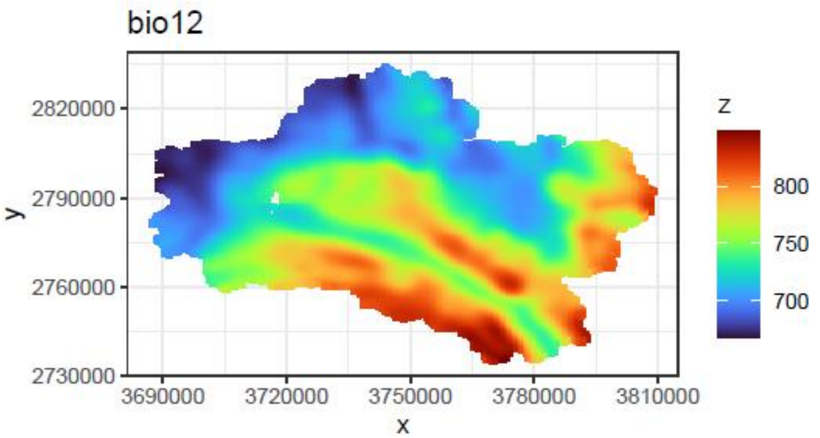
**Data source:** soilgrid.org; 27430 pts

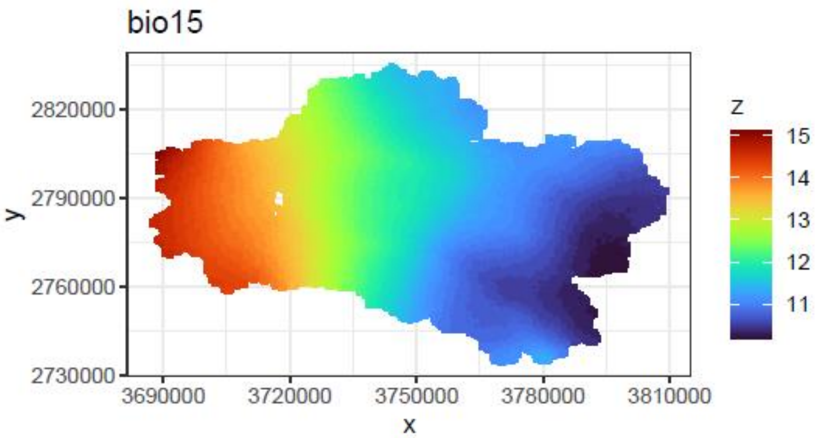
# Covariate(1/2)



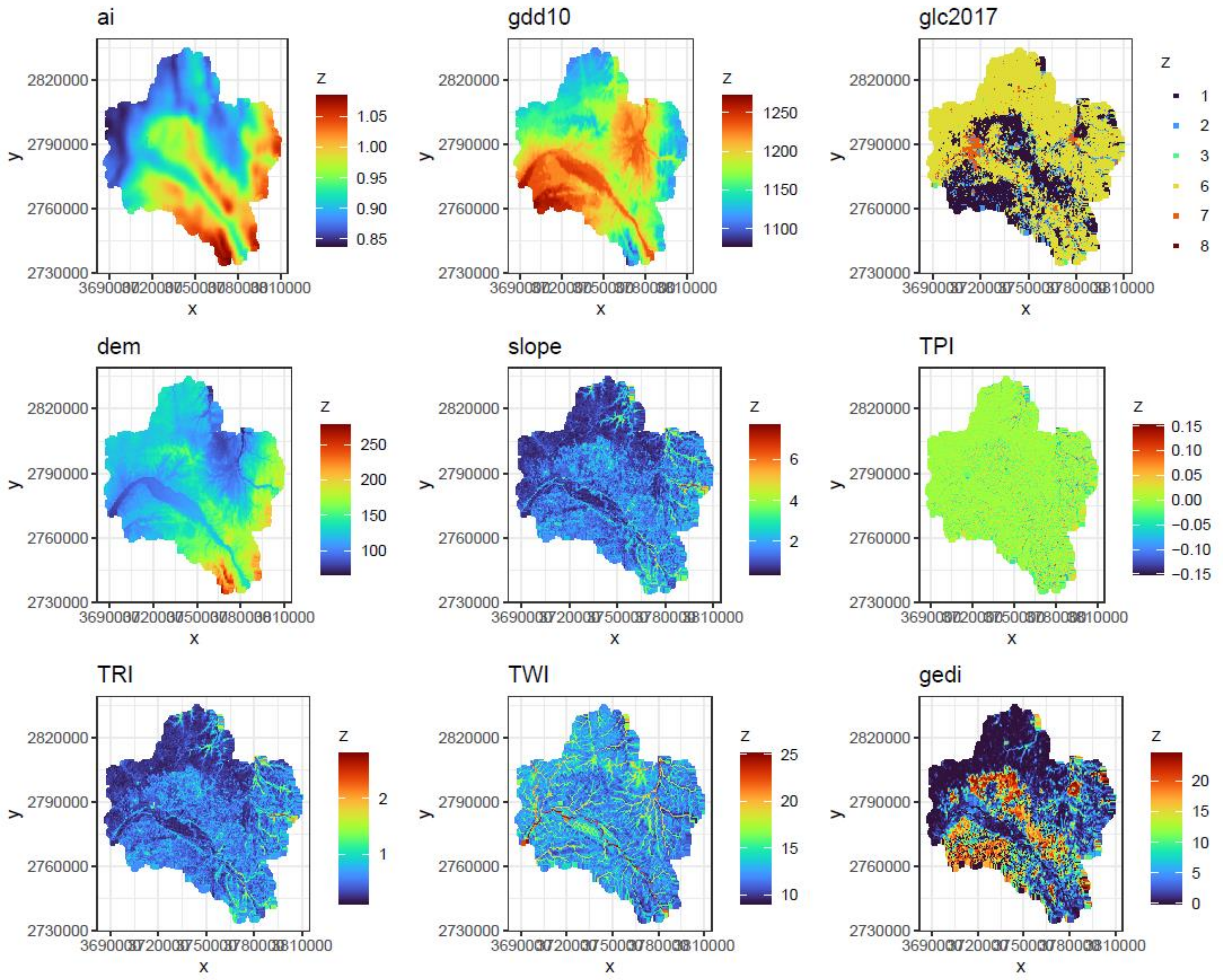



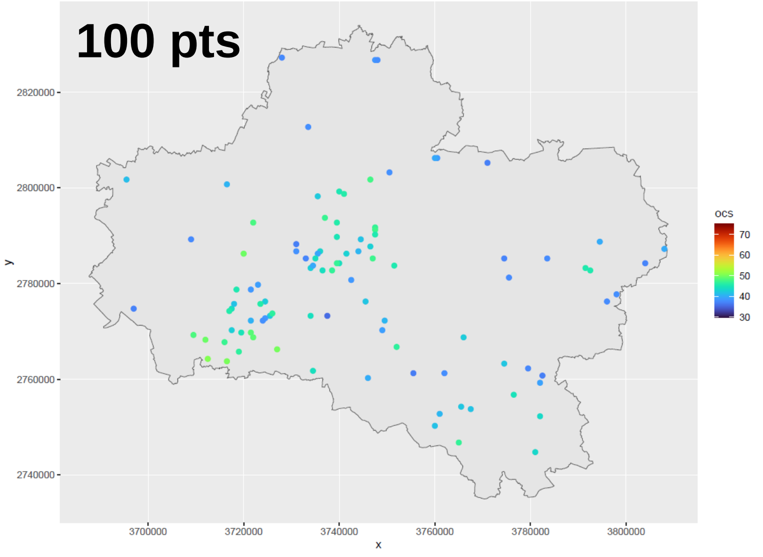

| | |
|---|---|
| bio1 | Mean annual air temperature [°C] |
| bio5 | Mean daily maximum air temperature of the warmest month [°C] |
| bio7 | Annual range of air temperature [°C] |
| bio12 | Annual precipitation [kg/m$^2$] |
| bio15 | Precipitation seasonality [kg/m$^2$] |

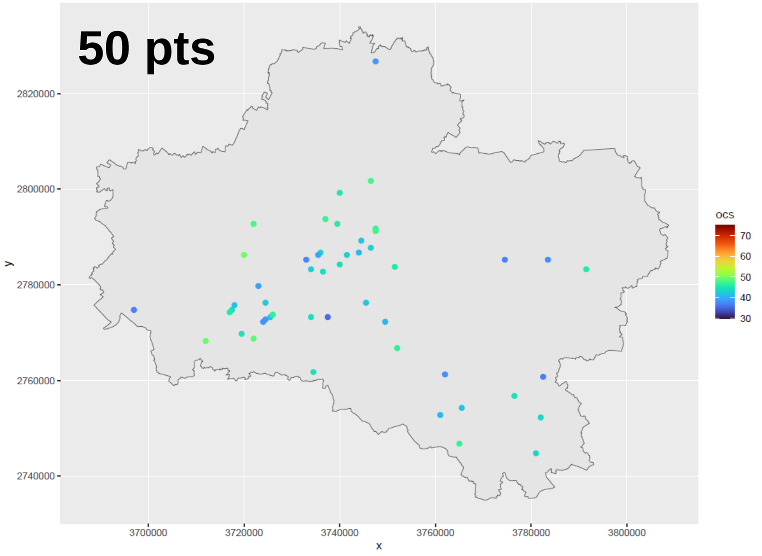
# Covariate(2/2)



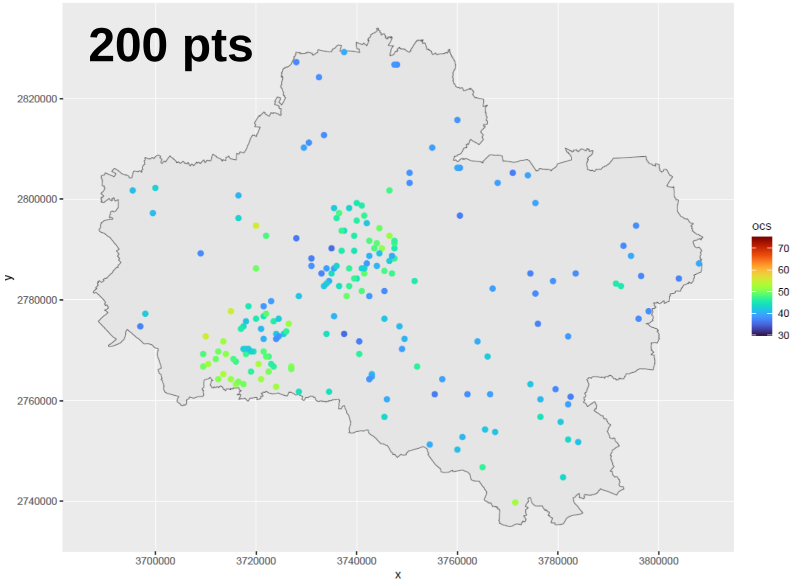

| | |
|---|---|
| ai | Aridity Index |
| gdd10 | Growing degree days heat sum above 10 °C |
| glc2017 | Landcover 2017 |
| dem | Elevation |
| slope | Slope |
| TPI | Topographic position index |
| TRI | Terrain ruggedness index |
| TWI | Topographic wetness index |
| gedi | Forest height |

# Sparse



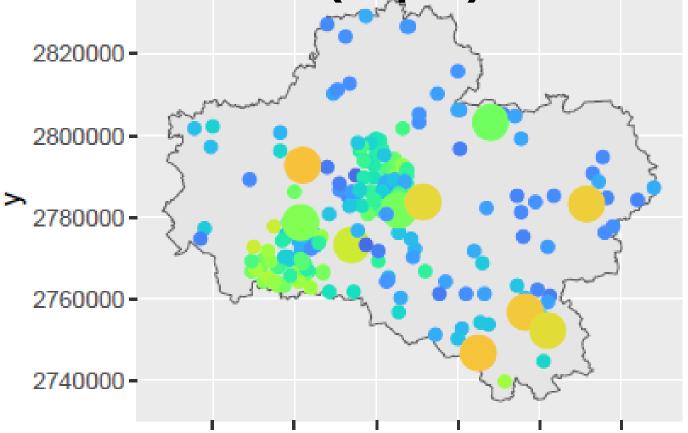




# Clustered

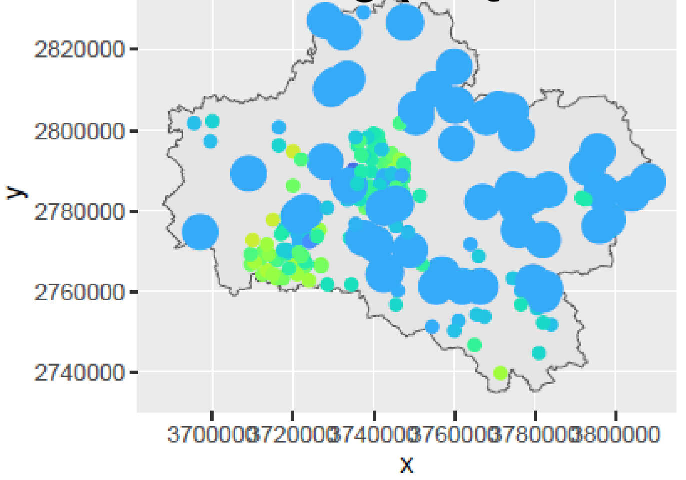



# Imprecise

**With outliers (10pts)**




**With censoring (<LQ=median)**




**Interval-valued**

Outside cluster: rela. Error of 30%

Within clusters: rela. Errror of 10%

# Data for the problems

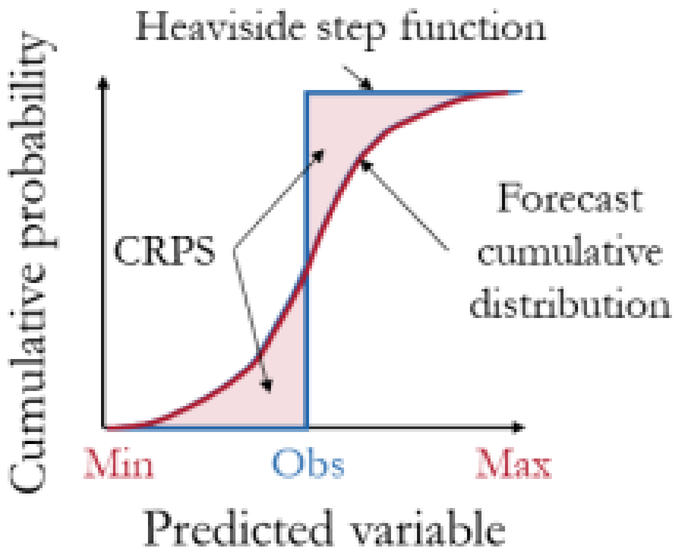
- **Training dataset**
  - **ocs at given locations for the six experiments** C, SC1, SC2, CIo, CIc, CIi
  - **Values of the 14 covariates**

- **Test dataset:** 27230 data over the whole territory

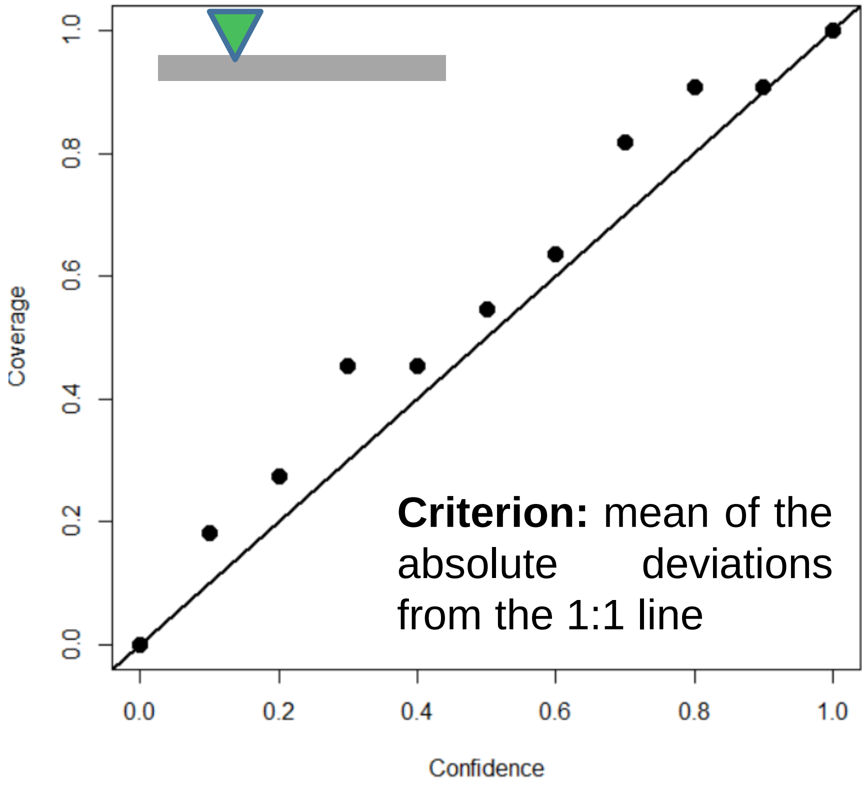
- **Objectives**
  - Assess error and uncertainty metrics
    - **Error:** RMSE, MAE, MaxAE
    - **Prediction intervals** PI:
      - width of 90-PI (w.PI),
      - coverage of 90-PI (cov.PI)
      - Mean asbolute deviation of the accuracy plot (Mcov.PI)
    - **Distribution:** continuous ranked probability score (CRPS)
    - **Any** relevant criterion!
  - Discuss evolution from experiment to experiment

- **Also available:** R scripts to generate repeated experiments

# Criteria




$$\mathrm{CRPS}(F, y) = \int_{\Omega} (F(z) - \mathbb{I}\{z \geq y\})^2 \mathrm{d}z.$$




**Criterion:** mean of the absolute deviations from the 1:1 line

# Test avec quantile Random Forest

| mae | rmse | maxe | w.PI | cov.PI | Mcov.PI | Mcrps | Case |
|---|---|---|---|---|---|---|---|
| [Mg/ha] | [Mg/ha] | [Mg/ha] | [Mg/ha] | % | % | - | |
| 2.7 | 3.5 | 21.0 | 5.0 | 80% | 5% | 1.86 | cluster |
| 2.8 | 3.6 | 21.0 | 4.7 | 80% | 4% | 1.96 | sparse1 |
| 4.1 | 4.8 | 19.0 | 5.1 | 73% | 9% | 2.93 | sparse2 |
| 3.0 | 3.9 | 20.6 | 6.8 | 85% | 3% | 2.02 | outlier |
| 3.0 | 3.7 | 19.2 | 3.9 | 44% | 26% | 2.23 | lq |

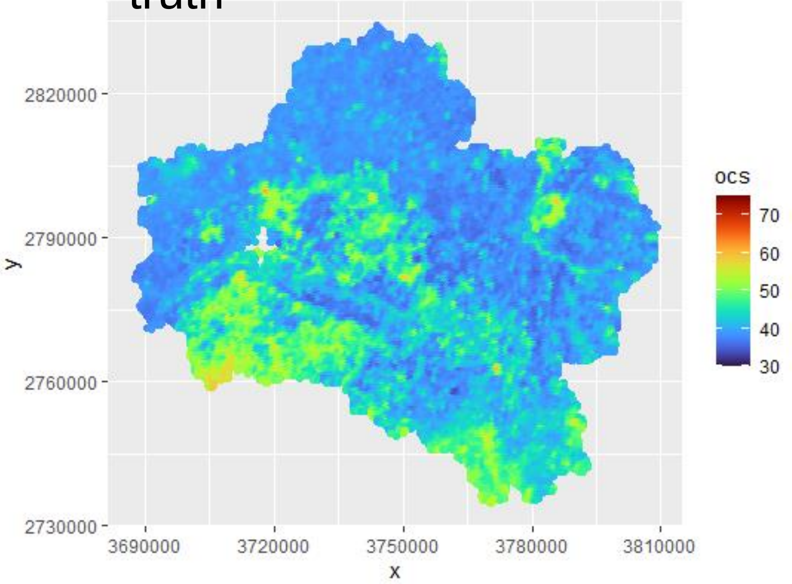
Only x,y coordinates

| mae | rmse | maxe | w.PI | cov.PI | Mcov.PI | Mcrps | Case |
|---|---|---|---|---|---|---|---|
| [Mg/ha] | [Mg/ha] | [Mg/ha] | [Mg/ha] | % | % | - | |
| 2.2 | 2.9 | 19.5 | 5.0 | 89% | 1% | 1.55 | cluster |
| 2.4 | 3.2 | 19.0 | 4.4 | 83% | 3% | 1.70 | sparse1 |
| 2.9 | 3.6 | 16.7 | 4.9 | 82% | 4% | 1.99 | sparse2 |
| 2.3 | 2.9 | 17.2 | 6.9 | 91% | 4% | 1.63 | outlier |
| 2.6 | 3.1 | 19.0 | 4.1 | 51% | 20% | 1.93 | censoring |

All 14 covariates

Example with clustered data

‘truth’




RF prediction (with all predictors)