



**HAL**  
open science

# The benefits of using a reference sampling for mitigating the impact of legacy soil data errors on Digital Soil Mapping outputs.

Philippe Lagacherie, Mairer Arregui, David Fages

## ► To cite this version:

Philippe Lagacherie, Mairer Arregui, David Fages. The benefits of using a reference sampling for mitigating the impact of legacy soil data errors on Digital Soil Mapping outputs.. *pedometrics'24*, Feb 2024, Las Cruces, New Mexico, United States. hal-04466401

**HAL Id: hal-04466401**

**<https://hal.inrae.fr/hal-04466401>**

Submitted on 19 Feb 2024

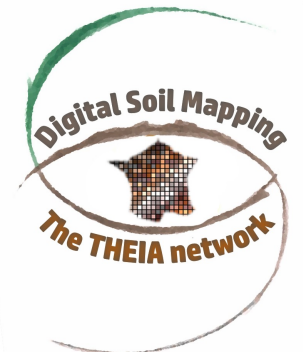
**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

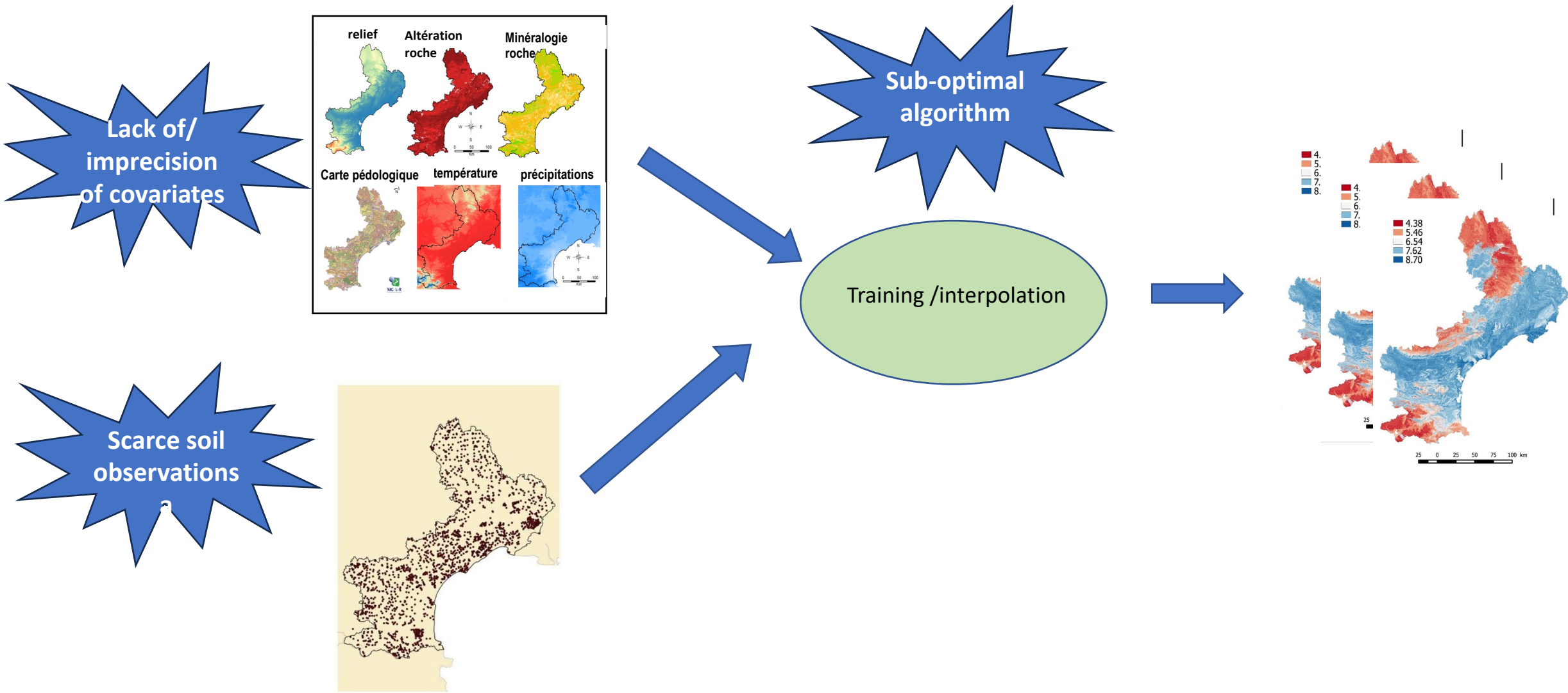
# The benefits of using a reference sampling for mitigating the impact of legacy soil data errors on Digital Soil Mapping outputs.

Philippe Lagacherie<sup>1</sup>, Maider Arregui<sup>2</sup> and David Fages<sup>1</sup>

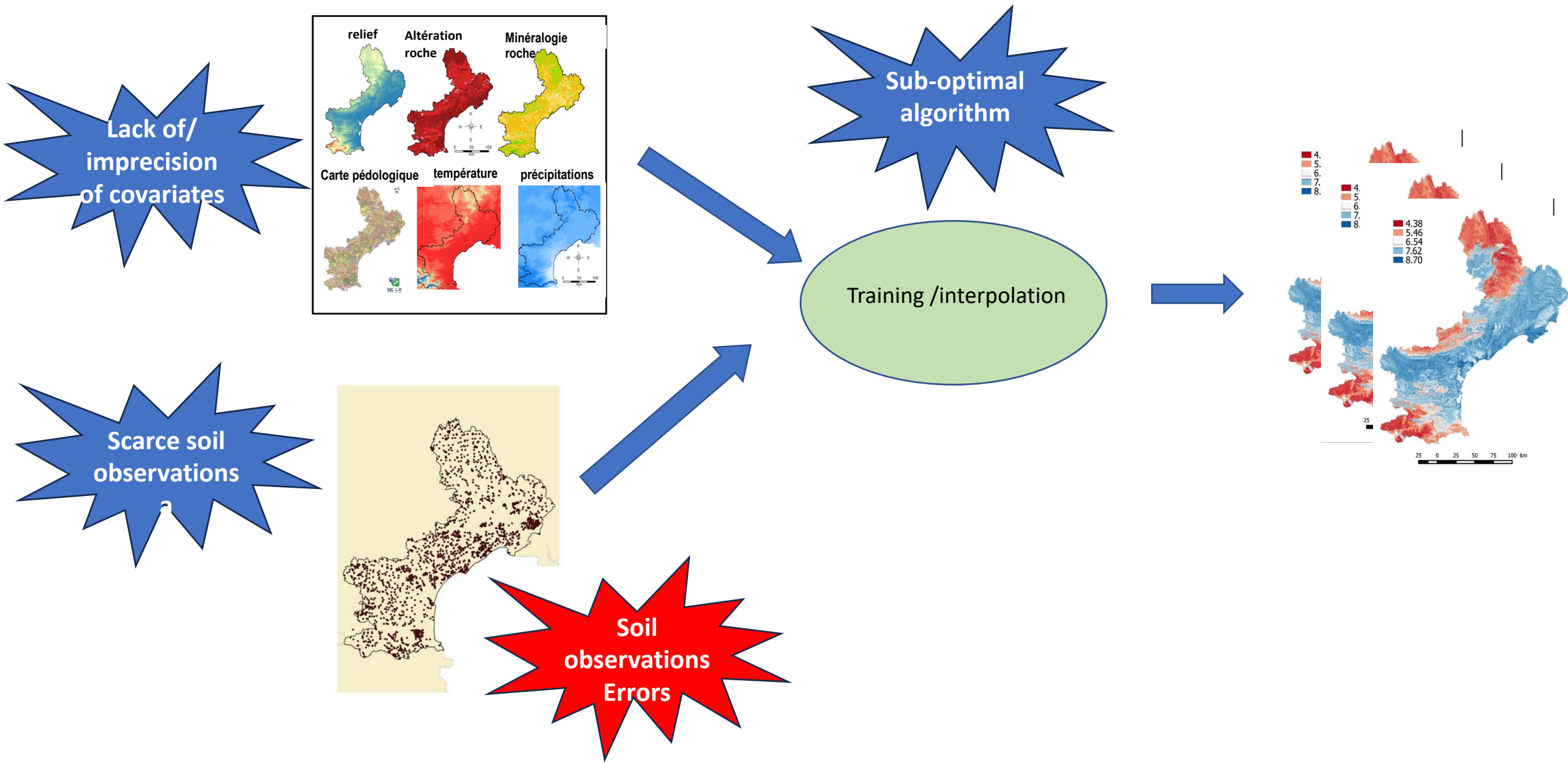
1. *LISAH, INRAE, Montpellier, France*
2. *BRL Exploitation, Nimes, France*

The logo for INRAE, featuring the letters 'INRAE' in a bold, teal, sans-serif font.The logo for LISAH, with the letters 'LISAH' in a bold, multi-colored font (brown, green, blue).The logo for BRL Groupe, featuring the letters 'BRL' in a bold, blue, sans-serif font with a green leaf-like shape to the right, and the word 'Groupe' in a smaller font below.The logo for 'L'EUROPE S'ENGAGE L'OCCITANIE AGIT', with the text in a bold, blue and red font.The logo for the European Union, featuring a blue rectangle with twelve yellow stars arranged in a circle, and the text 'UNION EUROPÉENNE' below.The logo for the Occitanie region, featuring a red rectangle with a yellow sun-like symbol and the text 'La Région Occitanie Pyrénées - Méditerranée'.The logo for Terra OccitanIA, featuring a circular emblem with a landscape scene and the text 'Terra OccitanIA'.The logo for Digital Soil Mapping The THEIA network, featuring a circular emblem with a globe and the text 'Digital Soil Mapping The THEIA network'.

# Source of errors in Digital Soil Mapping



# Source of errors in Digital Soil Mapping



# Sources of errors on legacy soil information

---

- Analysis protocols
- Geo-referencing (mostly in the pre-GPS era)
- Sample processing before soil analysis (sieving, grinding, decarbonatation, ...)
- Soil database entry errors
- Age of observations (for time-variant soil properties)



**Lagacherie, P., Arregui, M., Fages D., Evaluating the quality of soil legacy data used as input of Digital Soil Mapping models. Accepted in EJSS**

# Objectives of the day

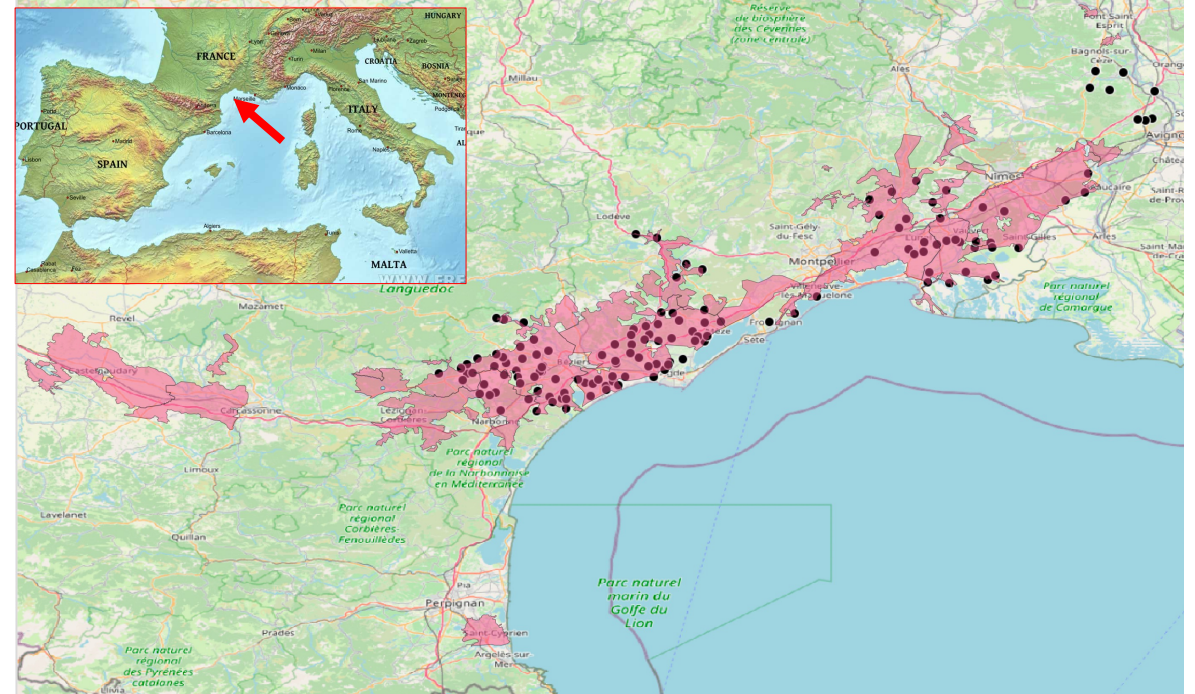
---

- To evaluate the impact of the legacy soil observation errors on DSM prediction performances
- To evaluate the benefits on DSM predictions of prior corrections of input soil data errors by using a control sampling.

# The data

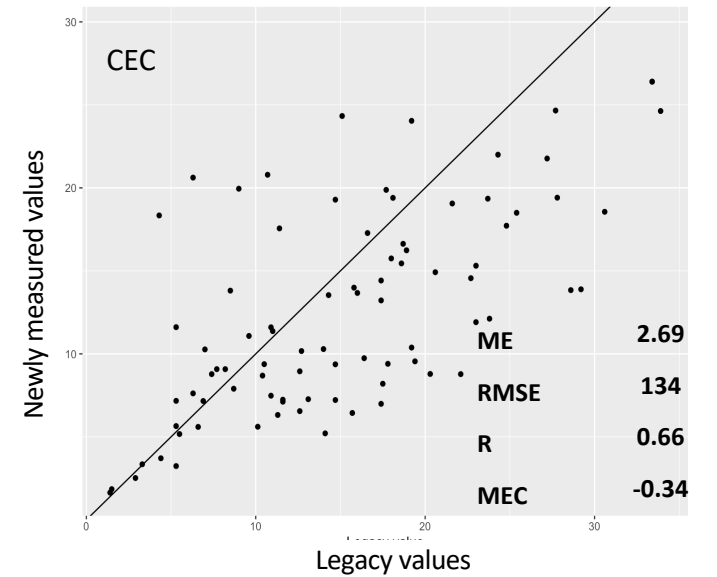
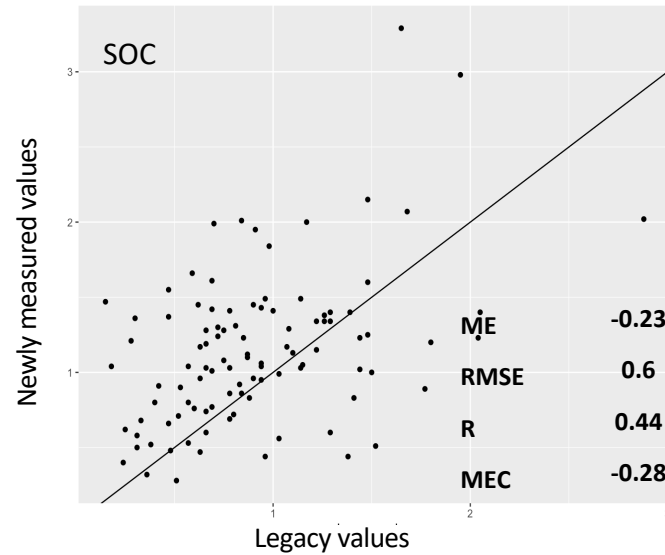
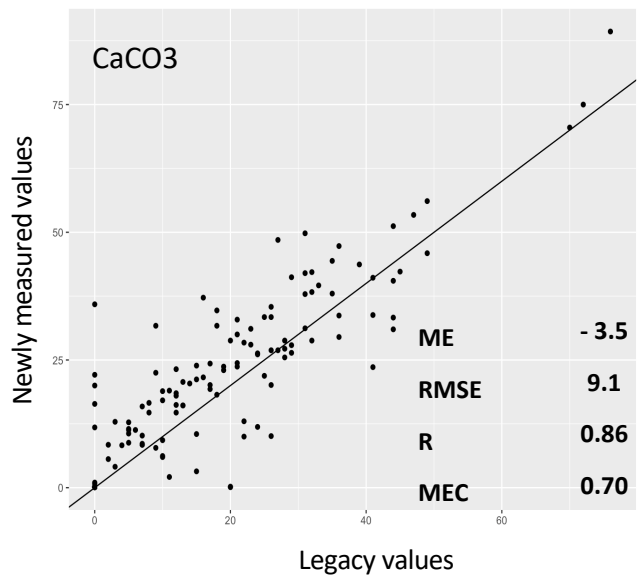
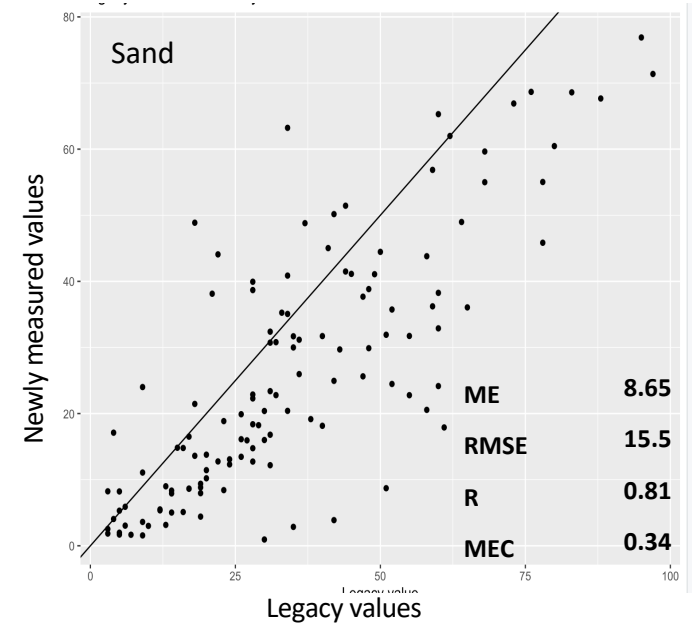
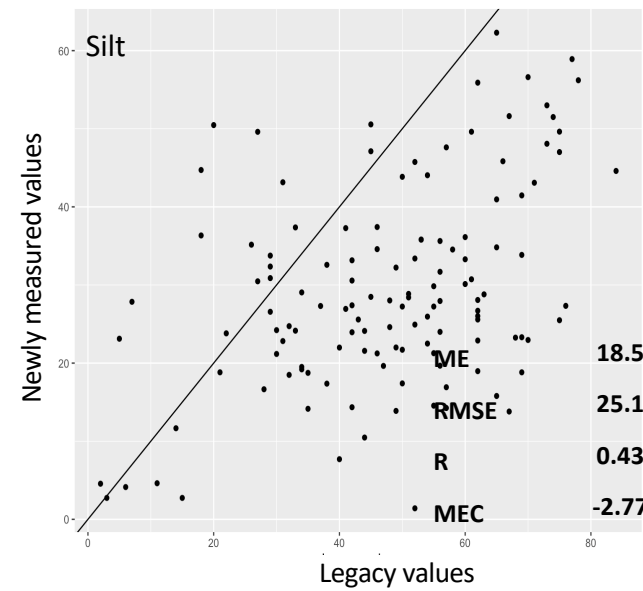
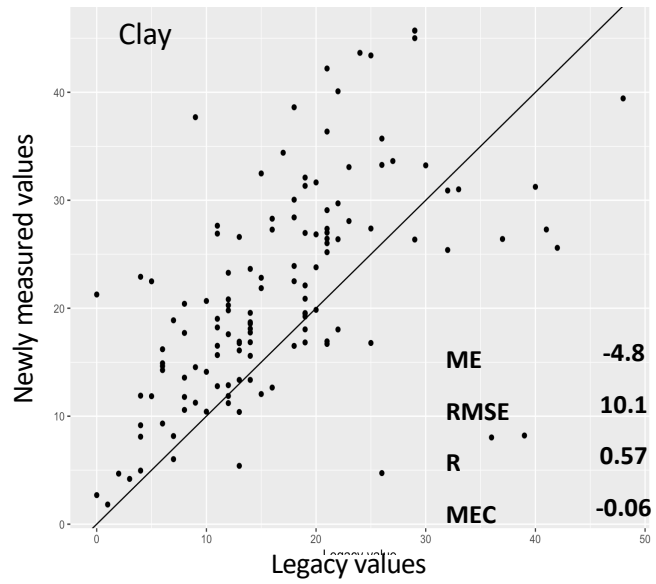
- Study area : BRL irrigation perimeter ( 6 636 km<sup>2</sup> in the Coastal plain of Occitanie, southern France)

The image shows two forms related to soil analysis. The left form is titled 'FICHE DE SOL' and contains general information such as 'C.N.A.E.R.L. D.E.M.V.', 'S.E.S. (mod. 6.0)', and '211 74 24 24 2'. It includes sections for 'CARACTÈRES GÉNÉRAUX', 'BIBLIOTHÈQUE TOPOGRAPHIQUE', 'HISTORIQUE', 'UTILISATION ACTUELLE', 'ÉTAT DE LA SURFACE', 'GÉOLOGIE', 'TYPE DE SOL', and 'CARACTÉRISTIQUES HYDRODYNAMIQUES'. The right form is titled 'CARACTÉRISTIQUES PÉDOLOGIQUES' and contains a table for 'CARACTÈRES DU SOL EN PLACE' with columns for 'Horizon', 'Couleur', 'Texture', 'Structure', 'Consistance', 'Bases', 'pH', and 'Observations'. Below this table are sections for 'ANALYSES' and 'CONCLUSIONS'.



- 6872 legacy measured soil profiles (1955- 1992) digitized by automatic text recognition (*ChemChem et al, submitted*)
- Control sampling at 129 locations with a registered legacy soil profile
- 6 topsoil properties analysed by a certified soil laboratory (AUREA): Clay%, Silt%, Sand%, CaCO<sub>3</sub>%, SOC% and CEC (meq/100g)

# Soil Legacy data errors



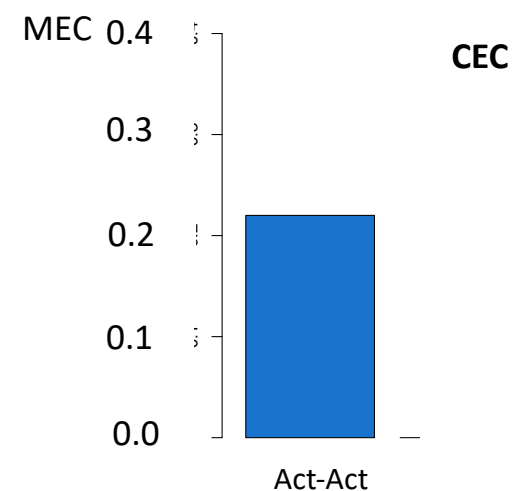
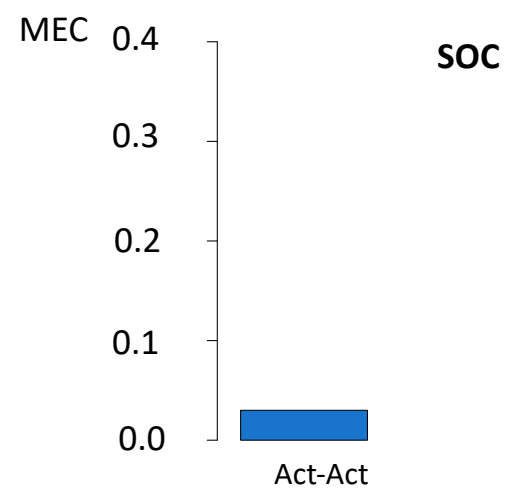
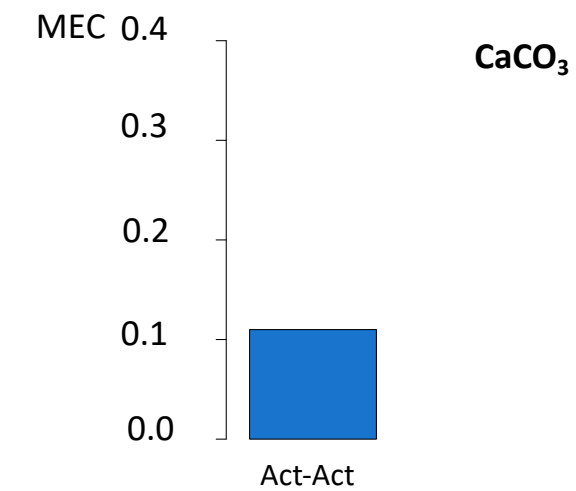
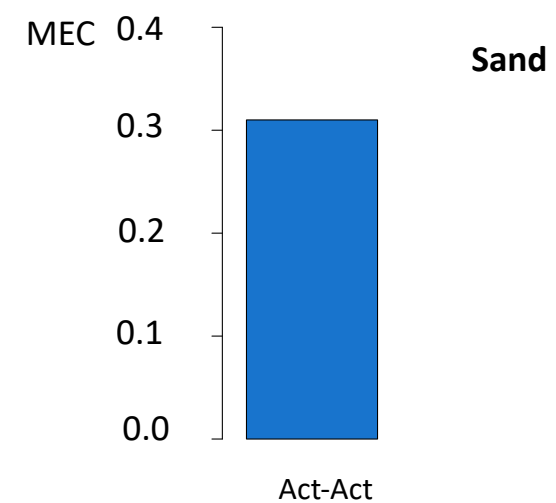
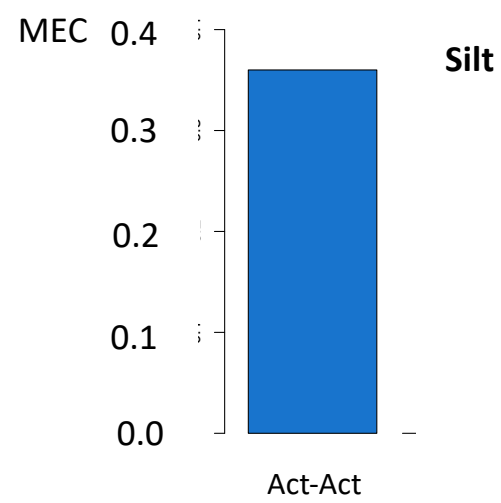
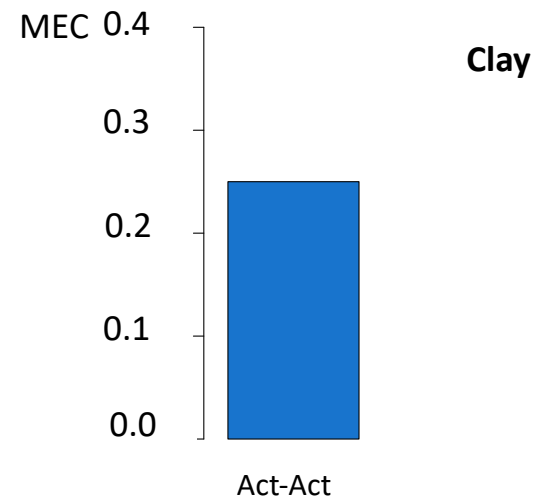


# The DSM Experiment

---

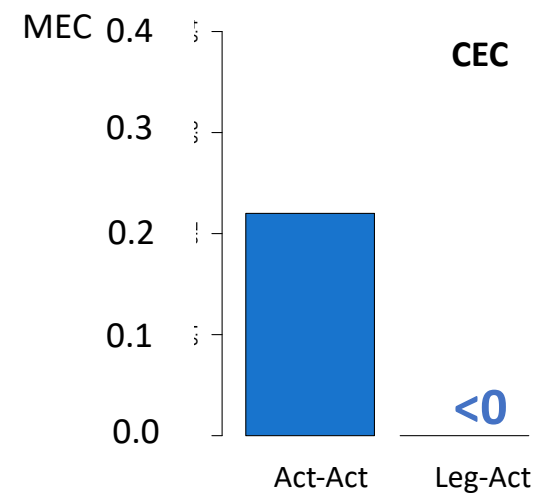
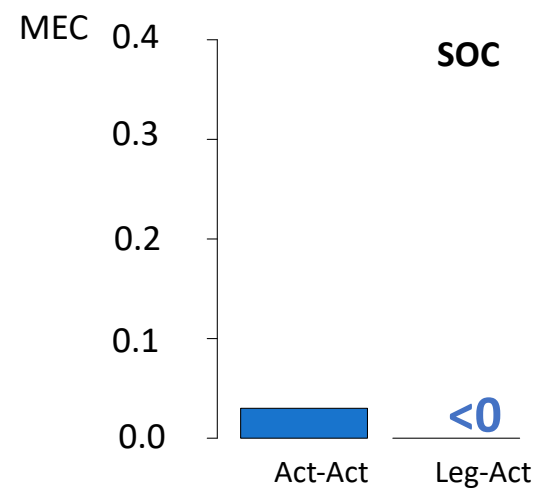
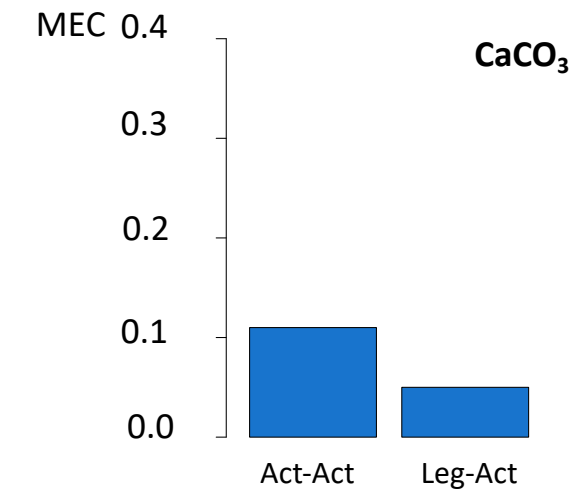
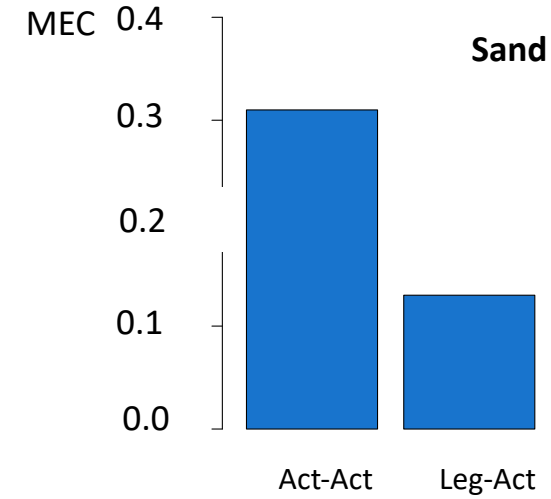
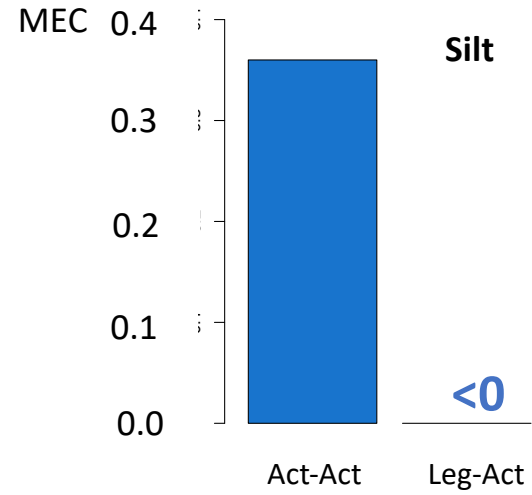
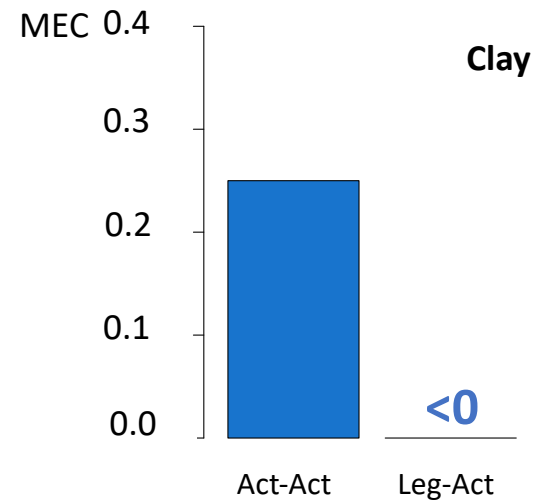
- Two alternate input datasets describing the same 129 locations: actual soil analyses (Act) and Legacy soil analyses (Leg)
- A classical DSM approach
  - Training algorithm : Random Forest
  - Soil covariates : geological maps, 1:250 000 pedological map, DEM and its derivative and a set of Remote sensing products
  - Evaluation method : 10 fold Cross Validations repeated 100 times
- 3 DSM scenario of input/evaluation data

# Reference scenario using actual analysis (Act-Act) for training and evaluating RF



\*  $MEC = 1 - MSE/Variance$

# RF trained with legacy value and evaluated with actual analyses (Leg-Act)



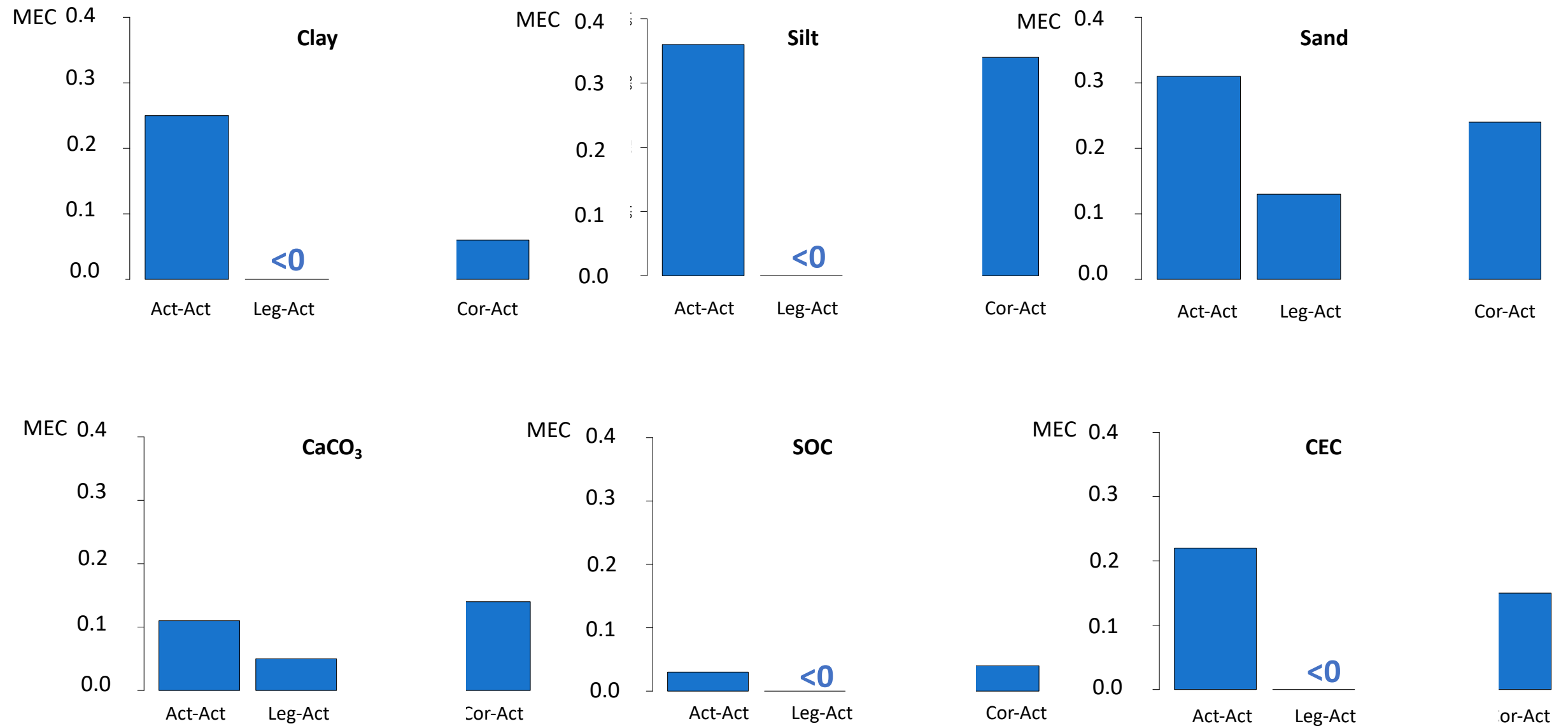
\* MEC = 1 - MSE/Variance

# Bias corrections ( $Z_{cor} = aZ_{leg} + bCaCO3_{leg} + c$ )

---

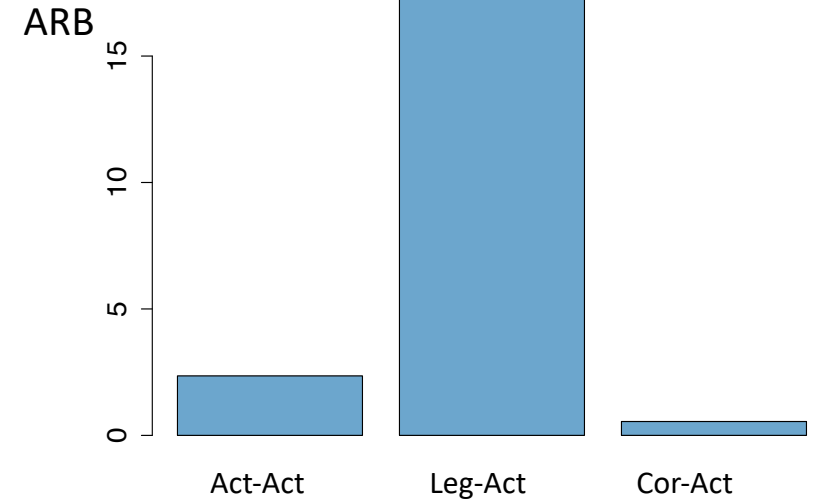
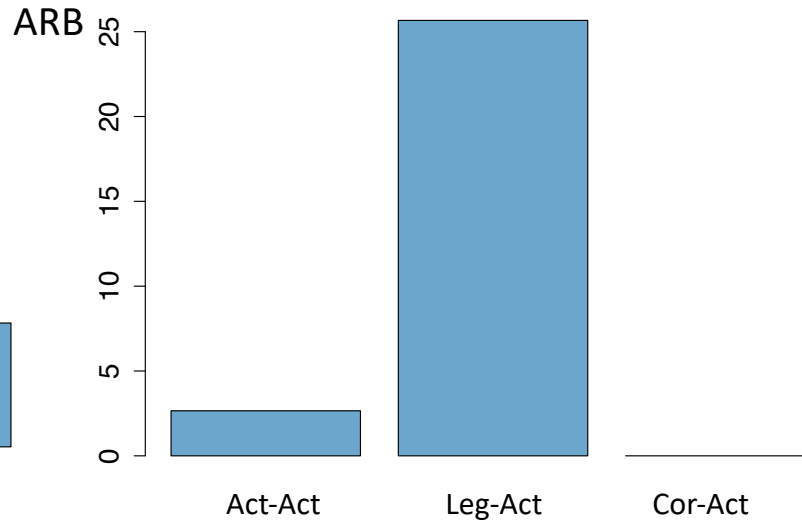
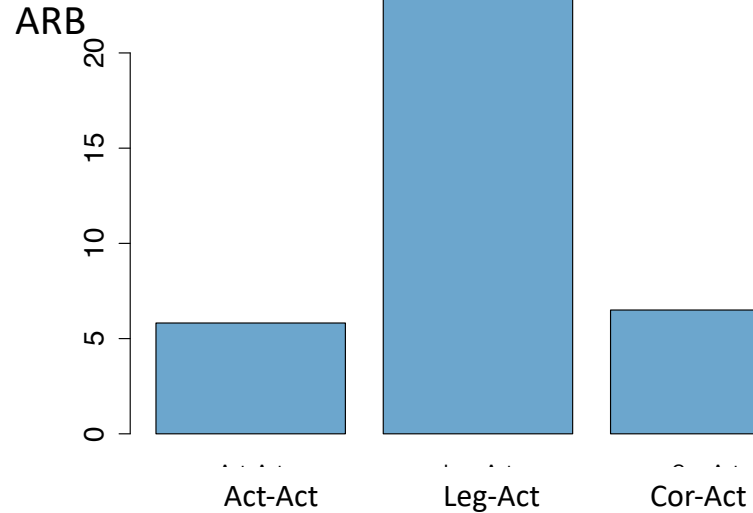
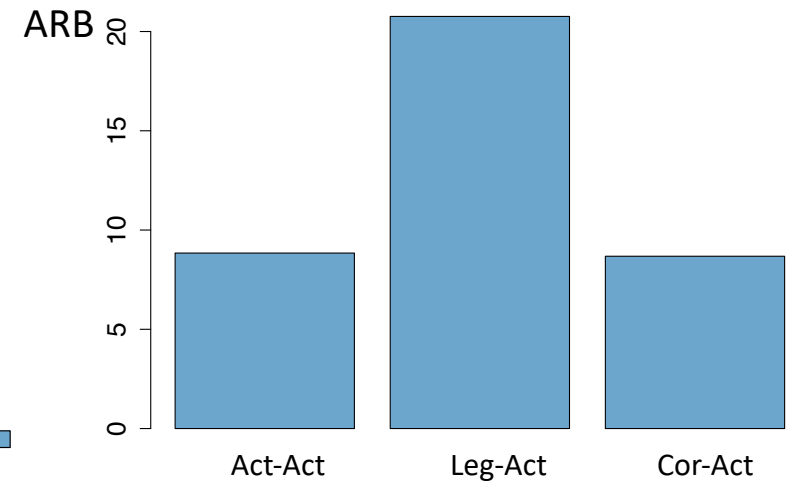
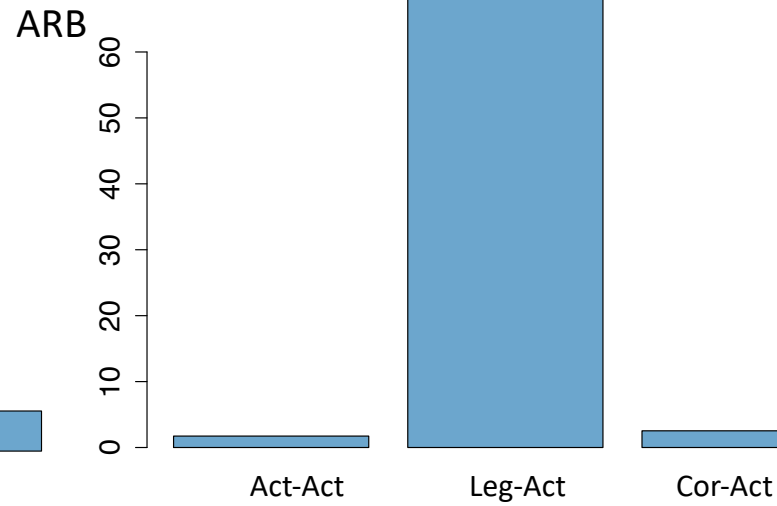
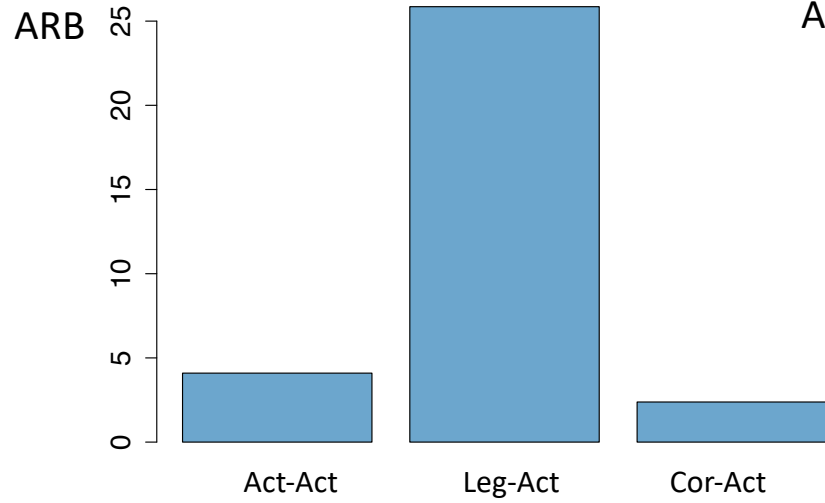
Soil property	c	a	b	R <sup>2</sup>
Clay	11.3646	0.6064	-	0.32
Silt	17.5961	0.4246	- 0.4723	0.49
Sand	1.2228	0.7133	-	0.65
CaCO3	5.2143	0.9089	-	0.75
SOC	0.6451	0.5449	-	0.20
CEC	4.577	0.522	-	0.44

# RF trained with corrected legacy values (Cor-Act)



\* MEC = 1 - MSE/Variance

# Bias removals with corrections



ARB : Absolute relative biases

# Conclusions

---

- The errors affecting legacy soil measurements of soil properties can be substantial, with possible large biases
- These errors can severely affect the performance of the DSM models trained on soil legacy data
- A control sampling using recent soil analyses performed at legacy soil profile locations can partly mitigate these effects
- Low-cost and optimized control sampling approaches should be investigated in the future