Supplement of SOIL, 10, 679–697, 2024
https://doi.org/10.5194/soil-10-679-2024-supplement

Supplement of

Insights into the prediction uncertainty of machine-learning-based digital soil mapping through a local attribution approach

Jeremy Rohmer et al.

Correspondence to: Jeremy Rohmer (j.rohmer@brgm.fr)

S1. Synthetic test case

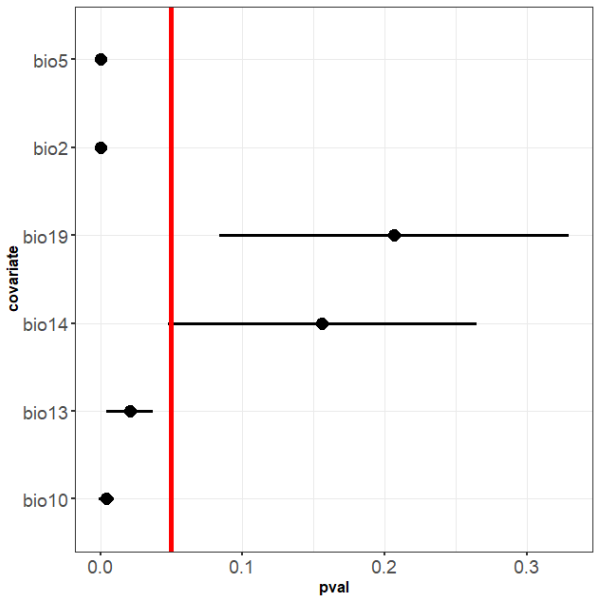

Figure S1. Screening analysis showing the p-values of the *HSIC*-based test of independence for the synthetic case. The dots indicate the mean value estimated over the replicates of a 10-fold cross validation (repeated 25 times). The lower and upper bounds of the error-bar are defined at +/- one standard deviation. When the dot merges with the error-bar, this indicates that the value of the standard deviation is low. The vertical red line indicates the significance threshold at 5%. When the p-value is below 5%, it indicates that the null hypothesis should be rejected, i.e., the considered covariate has a significant influence on the variable of interest, and is retained in the RF construction.

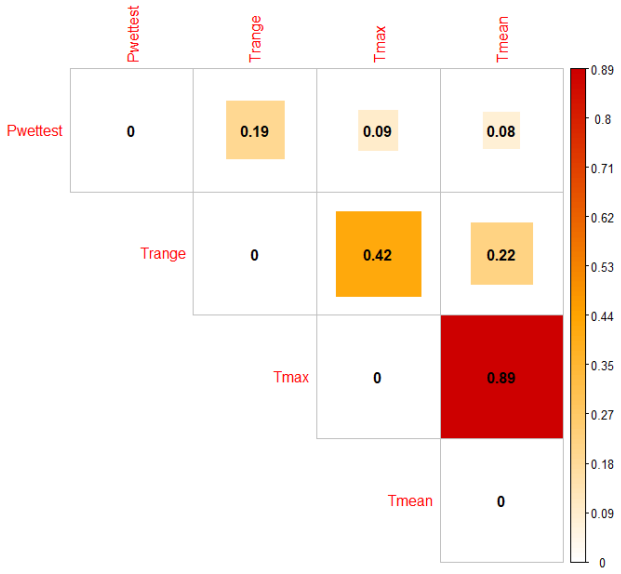

Figure S2: Matrix of pairwise *HSIC* dependence measure considering the covariates retained after applying the screening analysis for the synthetic test case. The warmer the colour, the higher the dependence.

S2. Toulouse real case

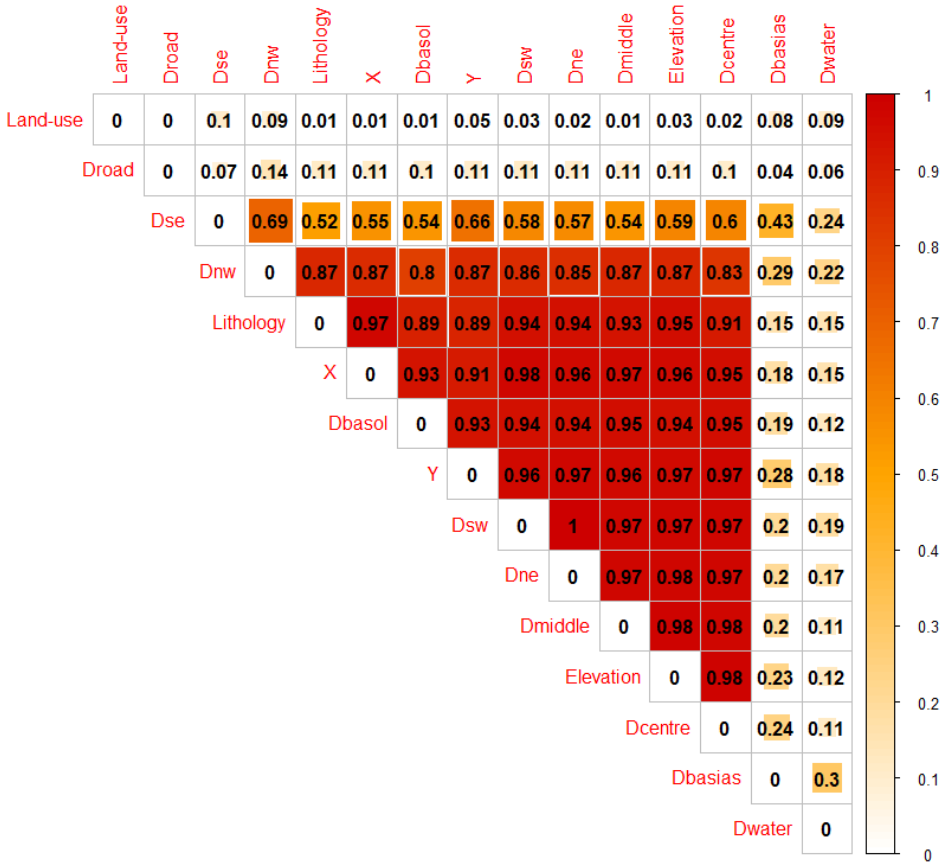

Figure S3 Pairwise dependence matrix for all covariates.

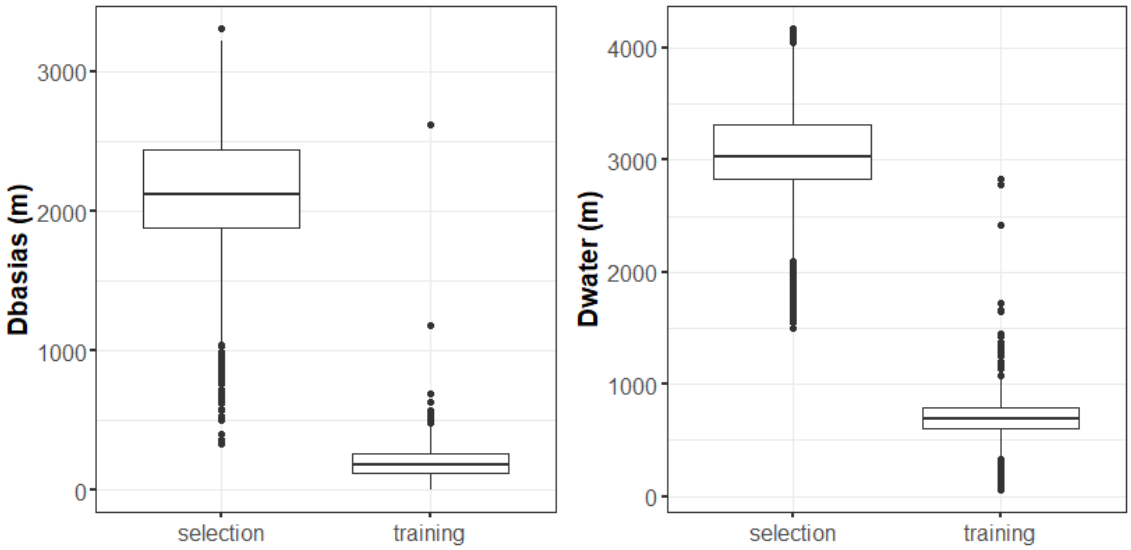

Figure S4 Boxplots of the distance values for the training dataset and for the locations (named "selection") where the corresponding group of covariates contributes significantly to the uncertainty, with a scaled Shapley value exceeding that of the prediction best estimate by more than 25%. The bottom right-hand panel compares the proportion of land used categories (AGR: Agriculture, FOR: Forests and grasslands, IND: Industrial and commercial economic activities) for the selection and the training dataset.

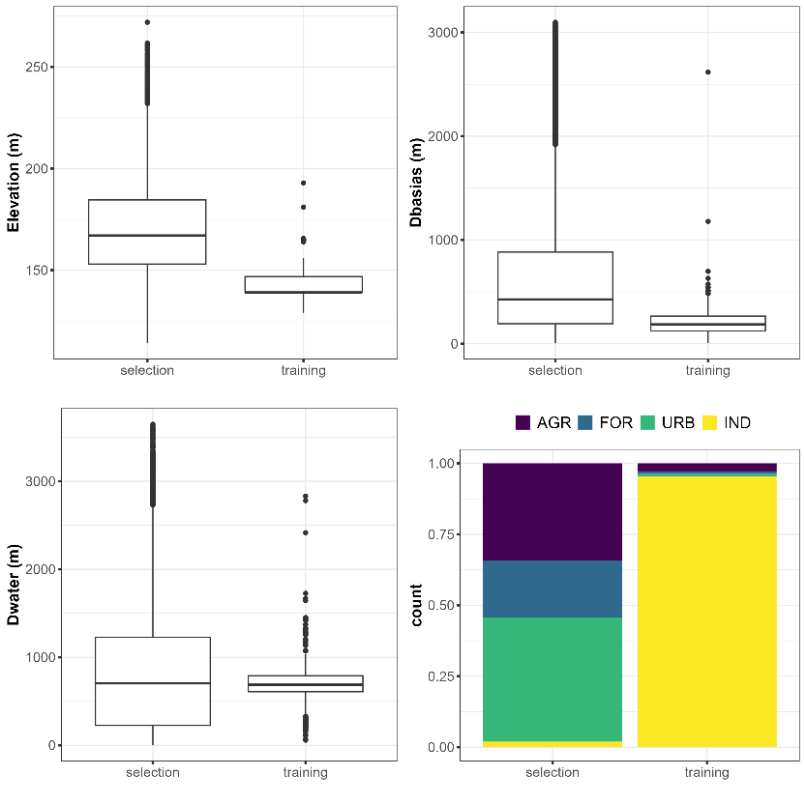

Figure S5 Boxplots of the covariates values for the training dataset and for the locations (named "selection") where the corresponding group of covariates have a negative contribution to the uncertainty. The bottom right-hand panel compares the proportion of land used categories (AGR: agriculture, FOR: forests and grasslands, IND: industrial and commercial economic activities) for the selection and the training dataset.

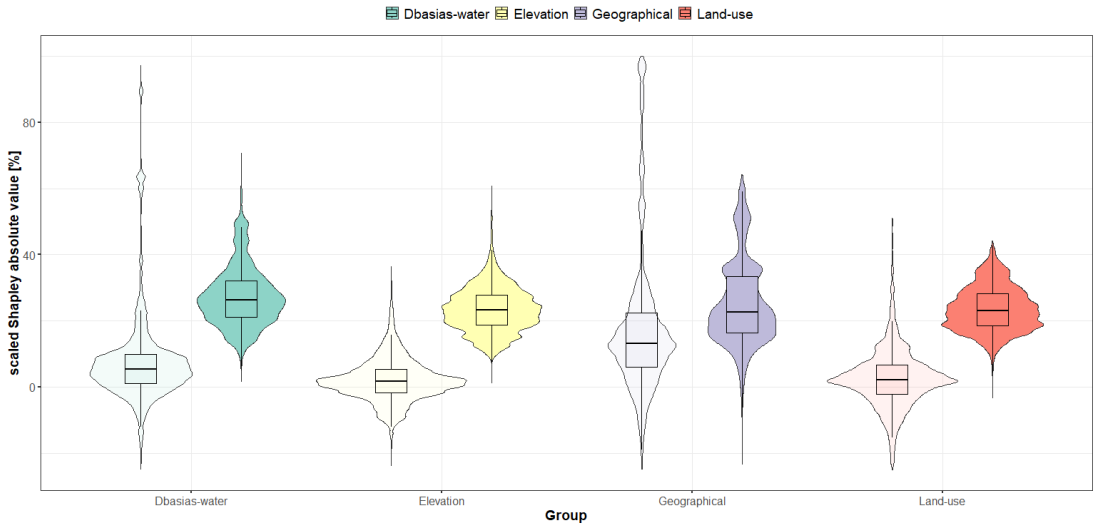

Figure S6: Statistics over the study area (>40,000 grid points) of the scaled Shapley value for the four groups of covariates using the RF conditional mean (bold colours), and the prediction uncertainty using the qRF inter-quartile width *IQW* (light colours).

S3. List of acronyms

- *IQW*: inter-quartile width
- *HSIC*: Hilbert–Schmidt Independence Criterion
- *ML*: Machine Learning
- *qRF*: quantile Random Forest
- *RF*: Random Forest
- *SHAP*: SHapley Additive exPlanation