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

Jeremy Rohmer[1], Stephane Belbeze[1], Dominique Guyonnet[1]

[1]BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2, France

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

Abstract. Machine learning (ML) models have become key ingredients for digital soil mapping. To improve the interpretability of their predictions, diagnostic tools such as the widely used local attribution approach known as SHapley Additive exPlanation (SHAP) have been developed. However, the analysis of ML model predictions is only one part of the problem, and there is an interest in obtaining deeper insights into the drivers of the prediction uncertainty as well, i.e., to explain why an ML model is confident, given the set of chosen covariate values, in addition to why the ML model delivered some particular results. In this study, we show how to apply SHAP to local prediction uncertainty estimates for a case of urban soil pollution, namely, the presence of petroleum hydrocarbons in soil in Toulouse (France), which pose a health risk via vapour intrusion into buildings, direct soil ingestion and groundwater contamination. Our results show that the drivers of the prediction best estimates are not necessarily the drivers of confidence in these predictions, and we identify those leading to a reduction in uncertainty. Our study suggests that decisions regarding data collection and covariate characterisation, as well as communication of the results, should be made accordingly.

1 Introduction

Maps of soil physical properties such as cation exchange capacity, pH, soil organic content, and hydraulic properties, as well as chemical properties such as concentrations of heavy metals (arsenic, mercury, etc.) and radionuclides (Caesium 137, Plutonium-239+240) and target-oriented indicators such as erodibility and soil compaction (see, e.g., Panagos et al., 2022), are essential for multiple types of studies, such as pollution assessment, urban planning, and construction design. In recent years, these maps have led to many advances in improving spatial prediction in the domain of digital soil mapping, denoted DSM (McBratney et al., 2003), with the development of methods and approaches based on either the geostatistical framework (Chilès and Desassis, 2018) or machine learning (denoted ML) techniques (Wadoux et al. 2020).

Beyond spatial prediction, the question of uncertainty in spatial prediction has emerged as a key challenge (Heuvelink and Webster, 2022: Sect. 4). Historically, this question has been addressed with kriging (see, e.g., Veronesi & Schillaci, 2019 for a discussion of DSM). Techniques based on ML have increasingly been used or adapted for this purpose. For instance, the popular quantile random forest model (e.g., Vaysse & Lagacherie, 2017) has been used to produce soil information worldwide in the SoilGrids 2.0 database together with uncertain information (Poggio et al., 2021). Along these lines, improvements in

validation procedures have been proposed (Schmidinger & Heuvelink 2023) together with tools for assessing prediction error and transferability (Ludwig et al., 2023).

However, quantifying the uncertainty is only one part of the problem, and there is an interest in gaining deeper insights into the influence of each covariate on the overall prediction uncertainty. This is the objective of global sensitivity analysis (Saltelli et al., 2008), which can be conducted within two settings: either "factor fixing" to identify noninfluential covariates or "factor prioritisation" to rank the covariates in terms of importance. The expected results can be of different types: the former setting provides justifications for simplifying the spatial predictive model by removing the noninfluential covariates, whereas the latter setting provides justifications for prioritising future characterisation efforts by focusing on the most important variables. In DSM, this question has been addressed with the tools of variance-based global sensitivity analysis (i.e., the Sobol' indices, as implemented by Varella et al., 2010) or with variable importance scores together with potentially recursive feature elimination procedures (as implemented by Poggio et al., 2021, and Meyer et al., 2018). Both approaches provide a "global" answer to the problem of sensitivity analysis, i.e., by exploring the influence over the whole range of variation of the covariates. However, these methods do not enable us to measure the influence of the covariates locally, i.e. for a prediction at a certain spatial location.

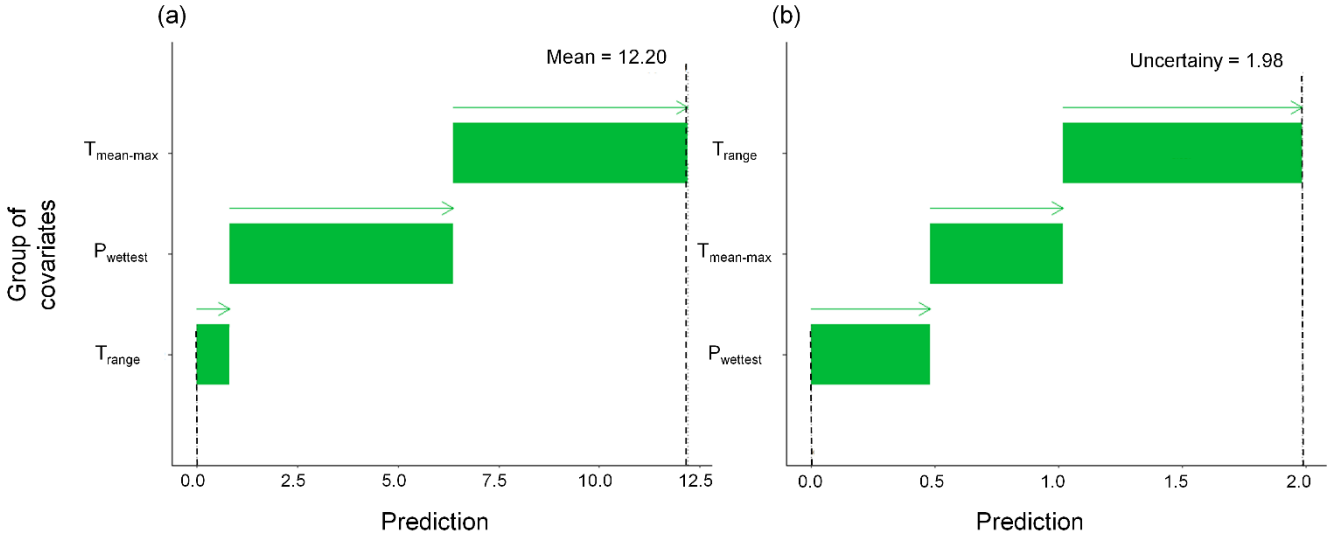


Figure 1: Example of SHAP-based decomposition of the prediction best estimate (modelled by the conditional mean of a random forest model, Panel a) and of the prediction uncertainty (modelled by the interquartile width estimated via a quantile random forest model, Panel b) for the variable of interest in the synthetic test case (fully described in Sect. 2.1) at a certain location of the study area. Each horizontal bar represents the contribution to the prediction (indicated by the vertical dotted line) of the considered covariates (indicated on the vertical axis) that correspond to the mean diurnal range T_{range}, to the group of covariates including the maximum and mean temperatures of the warmest quarter of the year $T_{mean\text{-}max}$, and to the precipitation of the wettest month $P_{wettest}$. Note the differences in the ordering of the groups, which indicate that the contributors to the mean and to the uncertainty estimate differ.

Recently, an alternative local approach was proposed by relying on a popular method from the domain of interpretable machine learning (Molnar, 2022) based on Shapley values (Shapley, 1953). This method has shown promising results in attributing the contributions of each covariate to any spatial prediction (Padarian et al., 2020; Wadoux et al., 2023; Wadoux and Molnar, 2022).

To date, the application of Shapley values to DSM has focused mainly on the prediction best estimate, and little information has been provided on the local prediction uncertainty. Motivated by a case of pollution concentration mapping in the city of Toulouse, France (Belbeze et al., 2019), we aim to investigate how to use Shapley values to decompose the local uncertainty, measured either by an interquantile width or by a variance estimate. Our objective is to explore the relationship between the drivers of the prediction best estimate and the drivers of confidence in the predictions. Providing evidence of differences in the dominant drivers is expected to have implications in terms of data collection and covariate characterisation. Communication of the results is also expected to be adapted accordingly. Figure 1 illustrates the type of result that can be derived with the approach. In this example (based on the synthetic test case fully detailed in Sect. 2.1), the mean prediction (best estimate, left panel) of the variable of interest at a certain location does not have the same contributors as the local uncertainty measured by the interquartile width (right panel). The group of covariates including the maximum and mean temperatures of the warmest quarter of the year (named $T_{mean\text{-}max}$) was identified as the first and second most important contributors, respectively. The identification of the least influential group of covariates also differed across both cases; this is illustrated by the mean diurnal range $T_{range,}$ which has little impact on the prediction result but strongly influences the confidence in the result.

The remainder of the paper is organised as follows. We first describe two application cases that motivated this study. In Sect. 3, we provide further details on the statistical methods used to estimate the local contributions to the prediction uncertainty. In Sect. 4, we apply the methods and provide an in-depth analysis of the differences in the drivers of prediction best estimates and uncertainty. In Sect. 5, we discuss the practical implications of the proposed procedure and its transferability to global-scale projects.

2 Case study

2.1 Synthetic test case

The first test case is synthetic. It aims to predict a virtual species suitability surface, denoted by *y*, over central Europe (Figure 2). This surface is calculated based on six bioclimatic covariates, defined in Table 1 (with prior normalisation between 0 and 1), which were extracted from the WorldClim dataset (available at www.worldclim.org/data/bioclim.html), as follows:

$$y(s) = 10 \times T_{\text{range}}(s) + 5 \times T_{\text{max}}(s) + 5 \times T_{\text{mean}}(s) + 5 \times P_{\text{wettest}}(s) + 10^{-4}.P_{\text{driest}}(s) + 10^{-4}.P_{\text{coldest}}(s), \quad (1)$$

By construction, this model has two characteristics that are used here to validate the proposed methods described in Sect. 3.4: (1) the last two covariates have negligible influence and (2) the covariates T_{max} and T_{mean} are strongly dependent. The dataset is based on the vignette of the R package *CAST*, which is available at https://hannameyer.github.io/CAST/articles/cast02-AOA-tutorial.html. A series of 25 "virtual" soil samples were randomly extracted (highlighted by square-like markers in Fig. 2) across the study area.

Table 1. Description of the covariates for the synthetic test case.

Covariate	Unit	Description	Identifier in the WorldClim dataset
T_{range}	°C	Mean diurnal range, i.e., mean of the monthly (max temperature - min temperature).	*Bio2*
T_{max}	°C	Max temperature of the warmest month	*Bio5*
T_{mean}	°C	Mean temperature of the warmest quarter	*Bio10*
$P_{wettest}$	mm	Precipitation of the wettest month	*Bio13*
P_{driest}	mm	Precipitation of the driest month	*Bio14*
$P_{coldest}$	mm	Precipitation of the coldest quarter	*Bio19*

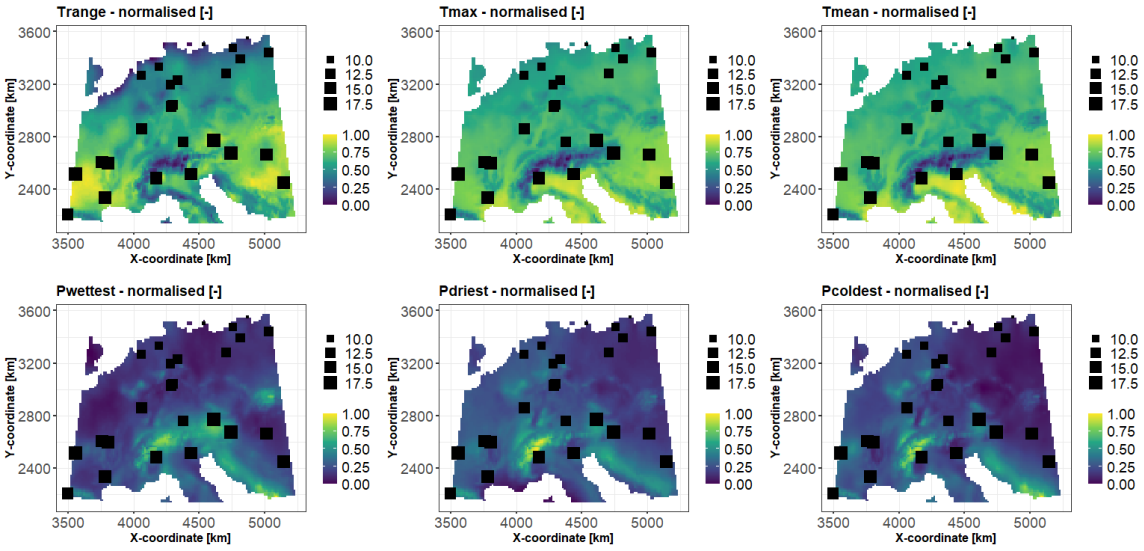


Figure 2: Covariates (with prior normalisation between 0 and 1) used in the synthetic test case (see Table 1 for a detailed description). The spatial distributions of the 25 soil samples are indicated by square-shaped markers. The size of each square is proportional to the synthetic variable calculated from the covariates based on Eq. (1).

2.2 Real test case

The real case focuses on DSM to predict the total petroleum hydrocarbon (C10-C4) concentration over the city of Toulouse (located in southwestern France), as part of the definition of urban soil geochemical backgrounds (see the comprehensive review by Belbeze et al. (2023)). In this case, petroleum hydrocarbons have multiple sources, such as road and air traffic, industrial emissions, and residential heating. The presence of this pollutant may inhibit several soil functions and hence prevent the delivery of associated ecosystem services (Adhikari and Hartemink, 2016). A primary soil function that may be jeopardised is the ability of the soil to provide a platform for human activities in a risk-free environment. Petroleum hydrocarbons in soil pose a risk to human health via several pathways, namely, direct soil ingestion, which is a particularly sensitive pathway for young children; exposure through respiration via vapour intrusion into buildings; and contamination of groundwater used for drinking water purposes. Notably, our study uses the data of this case to illustrate and discuss the applicability of the proposed approach and is not meant to supplement the results of Belbeze et al. (2019).

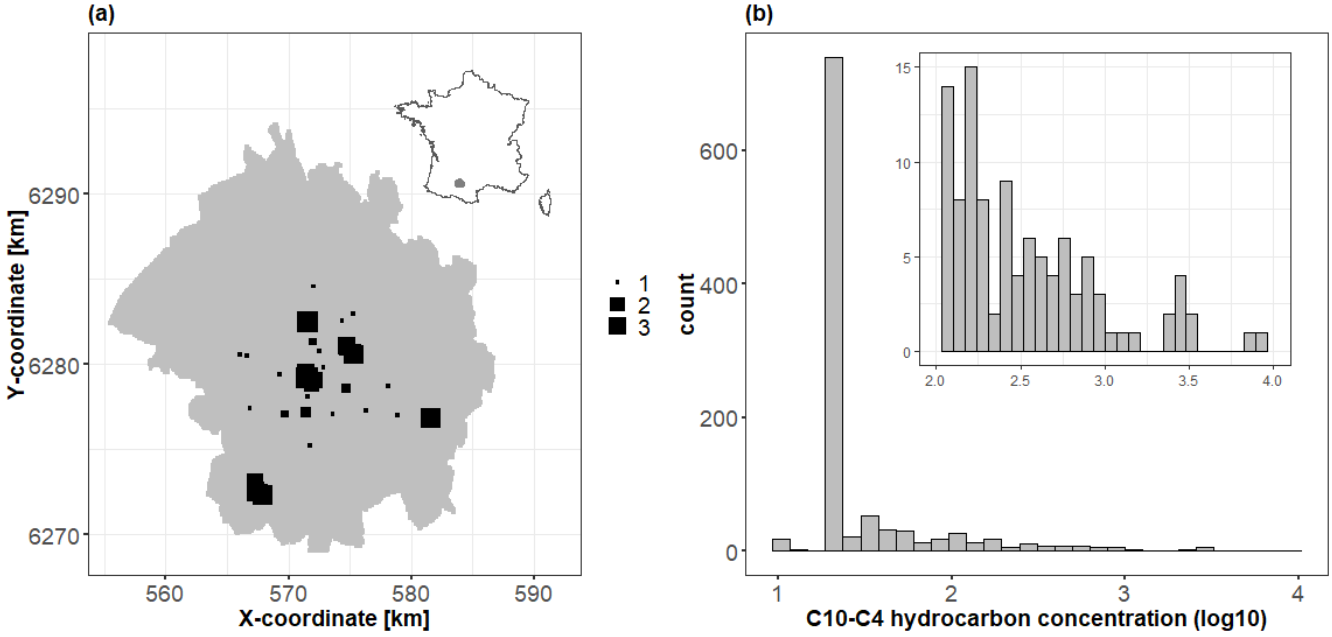


Figure 3: (a) Spatial locations of the 1,043 soil samples (square-like markers) across the city of Toulouse, which is located in southwestern France (see the location in the top right inset map). The size of each square is proportional to the logarithm (base 10) of the C10-C4 hydrocarbon concentration (expressed in mg/kg). (b) Histogram of the logarithm (base 10) of the C10-C4 hydrocarbon concentration (expressed in mg/kg) with a magnified view of the interval 2.0-4.0 (top right inset panel).

We use 1,043 soil samples collected over a depth interval [0, 2 m] to analyse the logarithm (base 10) of the C10-C4 hydrocarbon concentration, expressed in mg/kg (Fig. 3). We aim to predict the concentration over the whole city of Toulouse using a fine grid of spatial locations (one point every 100 m, i.e., >45,000 grid points) with the covariates described in Table 2. Figures 4 and 5 depict the spatial distributions of the considered covariates of continuous and categorical types, respectively.

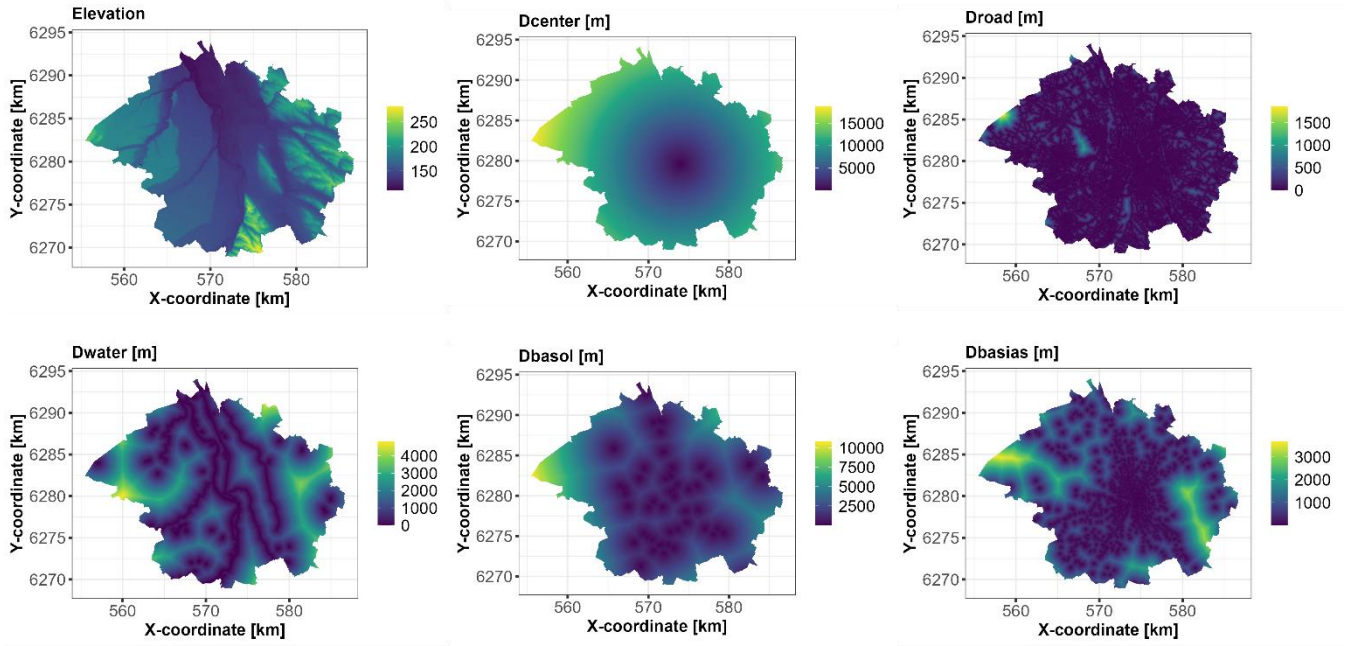
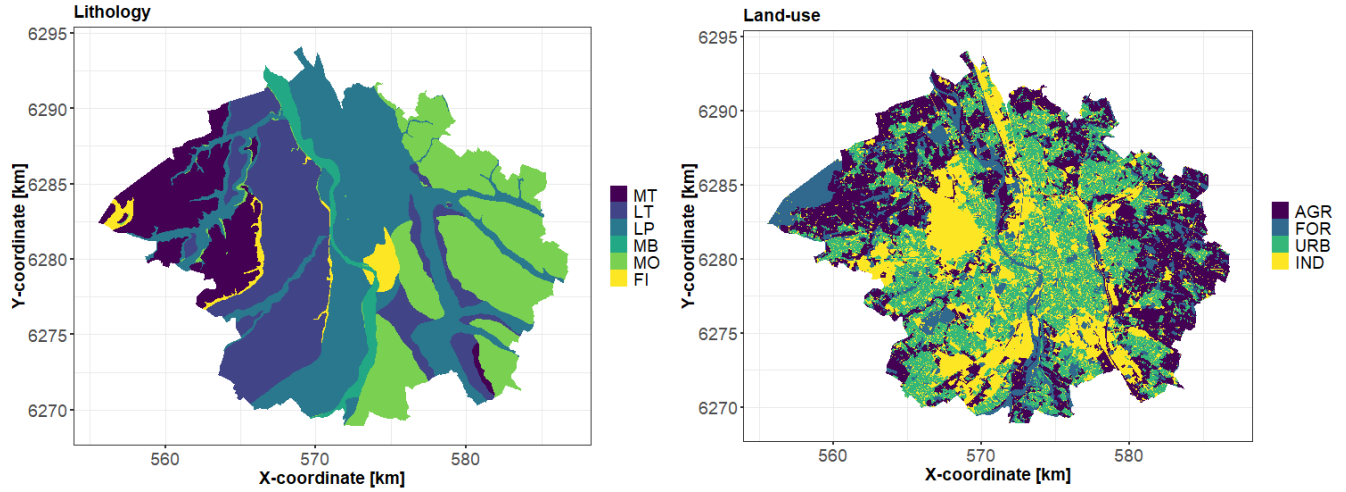


Figure 4: Covariates of continuous type over the city of Toulouse (see Table 2 for a detailed description).



120

Figure 5: Covariates of categorical type over the city of Toulouse. Left: Lithology (MT: medium terrace alluviums, LT: low terrace alluviums, LP: low plain alluviums, MB: major riverbed alluviums, MO: molasses, and FI: fill materials). Right: Land use (AGR: agriculture, FOR: forests and grasslands, and IND: industrial and commercial economic activities); see Table 2 for a detailed description.

125

In addition to these covariates, we follow the approach proposed by Behrens et al. (2018) to better account for spatial dependence: we also consider seven additional covariates, namely, the two geographical coordinates, X and Y, and five geographical covariates that correspond to the distances to the southeastern, northeastern, southwestern and northwestern corners of a rectangle around the city (denoted by D_{se}, D_{ne}, D_{sw}, and D_{nw}) and the distance to the centre location of this rectangle (denoted by D_{middle}). A total of 15 covariates are considered.

Table 2. Description of the covariates of the real test case.

Covariate	Unit	Description	Source
Elevation	[m]	Digital elevation model postprocessed from LIDAR data and gridded at a 10 m x 10 m resolution	Based on the processing detailed in (Belbeze et al., 2022)
Lithology	-	8 categories based on the grouping of Belbeze et al. (2019)	Based on the CHARM database available at https://www.data.gouv.fr/fr/datasets/cartes-geologiques-departementales-a-1-50-000-bd-charm-50/
Land-Use	-	4 categories based on the grouping of Belbeze et al. (2019)	Based on Copernicus (2012), available at https://doi.org/10.2909/debc1869-a4a2-4611-ae95-daeefce23490
D_{basias}	[m]	Distance to industrial sites (abandoned or active) potentially at the origin of pollution	BASIAS database (Leprond (2013)), available at https://www.data.gouv.fr/en/datasets/inventaire-des-sites-pollues/
D_{basol}	[m]	Distance to (potentially) polluted sites	BASOL database, available at https://www.data.gouv.fr/fr/datasets/base-des-sols-pollues/
D_{road}	[m]	Distance to the closest roads	Based on the processing detailed in (Belbeze et al., 2019)
D_{water}	[m]	Distance to the closest rivers	Based on the processing detailed in (Belbeze et al., 2019)
D_{center}	[m]	Distance to the city centre	Based on the processing detailed in (Belbeze et al., 2019)

X and Y coordinate	[m]	Geographical coordinate in the coordinate reference system of France, Lambert 93	-
D_{se}, D_{ne}, D_{sw}, D_{nw}, D_{middle}	[m]	Distances to the southeastern, northeastern, southwestern and northwestern corners of a rectangle around the city and the distance to the centre location of this rectangle	Based on the approach of Behrens et al. (2018)

3 Methods

3.1 Local attribution framework

We first consider that the value of the variable of interest $y(s)$ (e.g., a soil property or pollutant concentration) at a given spatial location *s* is related to *d* covariates $\mathbf{x}(s) = \{x_1(s), x_2(s), \dots, x_d(s)\}$. The mathematical relationship is modelled by an ML model (denoted by *f*(.)), where $f(\mathbf{x}(s))$ is assumed to resemble $y(s)$ as closely as possible; that is, $y(s) \approx f(\mathbf{x}(s))$. The ML model also provides a measure of the uncertainty (denoted by *u*(s)) of this prediction that is related to $\mathbf{x}(s)$ through the function *g*(.), which may differ from *f*(.). In this study, we focus on the random forest model (denoted by RF) used for regression (Breiman, 2001) and on its quantile regression variant, denoted by qRF (Meinshausen 2006), because this ML model has proven to be very efficient in multiple studies for DSM, as outlined in the introduction. Further details are provided in Appendix B. To reflect the magnitude of the RF prediction uncertainty $u(s)$ at spatial location s, we use the interquantile half-width (denoted by IQW_α) at a given level α defined as follows:

$$u(s) = g(\mathbf{x}(s)) = IQW_\alpha = q^{\frac{1+\alpha}{2}}(y|\mathbf{x}(s)) - q^{\frac{1-\alpha}{2}}(y|\mathbf{x}(s)), \quad (2)$$

where $q^\tau(y|\mathbf{x}(s))$ is the conditional quantile at level τ. In particular, the interquartile width corresponds to $IQW_{\alpha=0.50}$.

Our objective is to decompose *u*(s) at a given spatial location *s* as a sum of the $\mu_{i=1,..,d}(s)$ specific to the values of the covariates $\mathbf{x}(s)$ within the setting of the additive "feature attribution approach" (as defined by Lundberg and Lee (2017): Sect. 2) as follows:

$$u(s) = g(\mathbf{x}(s)) = \mu_0 + \sum_{j=1}^{p} \mu_j(s), \quad (3)$$

where μ_0 (named the *base value*) is a constant value (see the definition in Sect. 3.2). This decomposition can also be applied to *f*(.) as described in previous studies, as indicated in the introduction.

Importantly, Eq. (3) does not aim to linearise *g*(.) but to compute the contribution of each covariate to the particular prediction uncertainty value $g(\mathbf{x}(s))$. This means that the decomposition provides insights into the influence of the particular instance of

the covariates $\mathbf{x}(s)$ relative to $g(\mathbf{x}(s))$: (1) the absolute value of $\mu(s)$ informs on the magnitude of the influence at location s, directly expressed in physical units, which facilitates interpretation, and (2) the sign of $\mu(s)$ indicates the direction of the contribution, i.e., whether the considered covariate influences the prediction upwards or downwards in relation to the base value μ_0. Both aspects are outlined in Fig. 1; the width of the horizontal bar and the arrow are indicators of (1) and (2), respectively. To quantify $\mu(s)$ in Eq. 3, we rely on the SHapley Additive exPlanation (SHAP) approach, which was developed by Lundberg and Lee (2017) based on the Shapley values described in Sect. 3.2.

3.2 Shapley additive explanation

The SHAP approach relies on the Shapley value (Shapley, 1953), which is used in game theory to evaluate the "fair share" of a player in a cooperative game, i.e., it is used to fairly distribute the total gains to multiple players working cooperatively. It is a "fair" distribution in the sense that it is the only distribution satisfying some desirable properties (efficiency, symmetry, linearity, and 'dummy player'). See the proofs by Shapley, 1953, and Aas et al., 2021: Appendix A for a comprehensive interpretation of these properties from an ML model perspective.

Formally, we consider a cooperative game with d players and let $D_S \subseteq D = \{1, \dots, d\}$ be a subset of $|D_S|$ players. We define a real-valued function that maps a subset D_S to its corresponding value val: $2^{D_S} \to \mathbb{R}$ and measures the total expected sum of the payoffs that the members of D_S can obtain by cooperation. The gain that the ith player obtains is defined by the Shapley value with respect to val(.):

$$\mu_i(s) = \frac{1}{d}\sum_{D_S \subseteq D\setminus\{i\}} \binom{d-1}{|D_S|}^{-1} (\mathrm{val}(D_S \cup \{i\}) - \mathrm{val}(D_S)), \quad (4)$$

Equation 4 can be interpreted as the weighted mean over the contribution function differences for all subsets D_S of players not containing player i. This approach can be translated for ML-based prediction by viewing each covariate as a player and by defining the value function val(.) as the expected output of the ML model conditional on $\mathbf{x}_S^*$, i.e., when we only know the values of the subset D_S of inputs (Lundberg and Lee, 2017). This approach is flexible with respect to the output of the ML model and can be applied to the conditional mean of the RF model, as well as to the uncertainty measure computed with the qRF model (Eq. 2).

Formally,

$$\mathrm{val}(D_S) = \mathrm{E}\left(h(\mathbf{x})\middle|\mathbf{x}_{D_S} = \mathbf{x}_{D_S}^*\right), \quad (5)$$

where h(.) can correspond to either the conditional mean, denoted by f(.), or the uncertainty estimate, denoted by g(.).

In this setting, the Shapley value can then be interpreted as the contribution of the considered covariate to the difference between the prediction $h(\mathbf{x}^*)$ and the base value μ_0. The latter can be defined as the value that would be predicted if we did not know any covariates (Lundberg and Lee, 2017). In the application case, we are interested in pollution prediction; in this context, we choose $\mu_0 = 0$, which means that if we do not know any covariates, no pollution is expected (and there is no uncertainty). In this way, μ_i in Eq. 3 corresponds to the change in the expected model prediction if f(.) is used (or in the

uncertainty if $g(.)$ is used) when conditioning on that covariate and explains how to depart from 0. By construction, $\mu_i = 0$ indicates the absence of influence for the i[th] covariate related to the 'dummy player' property of the method. In addition, the sum of the inputs' contributions is guaranteed to be exactly $h(\mathbf{x}^*(s)) - \mu_0$, which is related to the 'efficiency' property of the method.

In this study, we aim to calculate the Shapley values for both the prediction best estimates modelled by $f(.)$ and the uncertainty modelled by $g(.)$. To facilitate comparison across the study area between these different prediction objectives, we use a scaled version of the Shapley absolute value, i.e., $\mu(s)/(f(\mathbf{x}^*(s)))$, expressed in %. This means that the contributions, regardless of the prediction objective, are analysed in Sect. 4 with a common scale, which is chosen here as the value of the prediction best estimate for the given considered instance.

In practice, the computation of the Shapley value may be demanding because Eq. (4) implies covering all subsets D_S (the number of which grows exponentially with the number of covariates d, i.e., 2^d) and Eq. (5) requires solving integrals of dimension 1 to $d-1$. When the SHAP approach is applied to a large number of spatial locations (in our case, >45,000), the computational complexity is high. To alleviate the computational burden, a possible option is to rely on the group-based approach proposed by Jullum et al. (2021), which can be used to adapt Eq. 4 for a group of covariates. Considering a partition $\boldsymbol{G} = \{G_1, G_2, \dots, G_g\}$ of the covariate set D, the Shapley value for the i[th] group of covariates G_i is as follows:

$$\mu_{G_i}(s) = \frac{1}{g}\sum_{T \subseteq \boldsymbol{G}\backslash\{G_i\}} \binom{g-1}{|T|_g}^{-1} (\text{val}(T \cup \{G_i\}) - \text{val}(T)), \quad (6)$$

where the summation index T runs over the groups in the sets of groups $\boldsymbol{G}\backslash\{G_i\}$ and $|T|_g$ is the number of groups in T. This means that this group Shapley value is the game theoretic Shapley value framework applied to groups of covariates instead of individual covariates. The group Shapley values possess all the Shapley value properties. The practical advantage of working with groups is that computing Eq. 6 has a relative computational cost reduction of 2^{d-g}, hence making possible the use of an exact method by considering all combinations of covariates for computing the Shapley values.

This definition raises the practical question of how to define groups. As explained by Jullum et al. (2021), this definition can be based on knowledge/expertise, i.e., on information that makes sense in relation to the problem at hand. The main advantage is that this facilitates the interpretation of the Shapley values. The second grouping option, which is complementary to the one based on expertise, consists of identifying covariates that provide redundant information because they share a strong dependency. The groups of dependent covariates can be identified with a clustering algorithm (see Hastie et al., 2009: Chapter 14) by taking as input the matrix of pairwise similarities. This approach does not, however, ensure that the effect of dependence among the covariates is completely removed, which may influence the SHAP results, as was extensively investigated by Aass et al. (2021). To account for this, we rely on the method proposed by Redelmeier et al. (2020) using conditional inference trees (Hothorn et al., 2006) to model the dependence structure of the covariates.

3.3 Overall procedure

The proposed approach, named group-based SHAP (see the implementation details in Appendix A), comprises three steps.

Step 1 aims to build and train the RF models based on the dataset of soil samples together with the covariate values, which is named the training dataset. The RF hyperparameters correspond to the number of variables at which to possibly split in each node (denoted by m_{try}) and the minimal node size at which to split (denoted by n_s). Their values are tuned via a repeated 10-fold cross-validation process (Hastie et al., 2009: Chapter 7). Although the RF model is efficient in taking into account many covariates, performing a screening analysis prior to the SHAP application within the cross-validation procedure is useful for facilitating its implementation. By reducing the number of covariates directly during the construction of the RF, the SHAP computational burden can be largely alleviated, as discussed in Sect. 3.2, and the applicability to global-scale projects where hundreds of covariates are present can be improved (see the discussion in Sect. 5.2).

To eliminate the covariates of negligible influence, different options are available in the literature (see, e.g., the procedure based on recursive feature elimination described by Poggio et al., 2021). Here, we propose relying on a popular method in the ML community, namely, the dependence measure based on the Hilbert-Schmidt Independence Criterion *HSIC* of Gretton et al. (2005). This generic measure has the advantages of being applicable (1) to any type of dependence, i.e., linear, monotonic, or nonlinear (see the discussion by Song et al., 2022); (2) to mixed variables, i.e., continuous or categorical (as in our case described in Sect. 2); and (3) without the use of RF importance measures, for which limitations have been identified in the literature, as has been extensively discussed (see, e.g., Benard et al., 2022 and references therein). In addition, the combination of *HSIC* with a hypothesis testing procedure (El Amri and Marrel, 2021) provides a rigorous setting for quantifying the significance of the considered covariate through the computation of p values. Further details are provided in Appendix C.

Step 2 is optional and aims to identify groups of covariates. The objective is twofold. First, by grouping covariates suitably with respect to the problem at hand, the Shapley values can be easily interpreted. This can be done based on expert knowledge or/and by identifying covariates that share a strong dependence to reduce the redundancy of information, for instance, using the *HSIC* pairwise dependence measure (Appendix C). The second practical implication is a reduction in the computational burden of the Shapley value estimation, as discussed in Sect. 3.2.

Step 3 is to compute the Shapley values associated with each covariate or group identified in *Step 2* to decompose the prediction uncertainty provided by the qRF model (trained in *Step 1*).

4 Results

4.1 Application to the synthetic test case

Using the 25 randomly selected soil samples described in Sect. 2.1, we construct an RF model to estimate the conditional mean, which is used as the best estimate of the prediction, and a qRF model to estimate the interquartile width (*IQW)*, which is used as the uncertainty estimate. To select the RF parameters *ns* and m_{try}, we repeat a 10-fold cross-validation exercise 25 times (Hastie et al., 2009: Chapter 7) by varying *ns* from 5 to 10 and m_{try} from 1 to 4. The number of random trees is fixed at 1,000; preliminary tests have shown that this parameter has little influence provided that it is large enough. This tuning procedure selects the pairs (*ns*, m_{try}) for which the average relative absolute error is minimised, identifying (*ns*=5, m_{try}=3) as

the combination that results in the lowest error of 6.8% (averaged over 25 replicates of the 10-fold cross-validation) with a frequency of 72% (i.e., 18 replicates out of 25). A screening analysis is performed within the cross-validation procedure using *HSIC* measures combined with a hypothesis testing procedure using the sequential random permutation-based algorithm developed by El Amri and Marrel (2021) with up to 5,000 random replicates. Averaged over the replicates of the 10-fold cross validation (repeated 25 times), the p values for the first four covariates reach a maximum value of 2% (Supplementary Materials A). The p values of the last two covariates are 16% (with a standard deviation of 10%) and 20% (with a standard deviation of 12%). Using a significance threshold of 5%, this means that the last two covariates are not statistically significant; hence, the number of covariates can reasonably be reduced from 6 to 4. This result is consistent with the construction of the synthetic case described by Eq. (1).

Using the trained RF model, Figure 6 shows the best estimate of the true value of the synthetic variable of interest (Panel a) and the prediction best estimate (Panel b) together with the uncertainty measure (Panel c) at 10,000 grid points across the European study area (with a spatial resolution of ≈13.5 x 13.5 km²). Overall, the RF predictions relatively well reproduce the true spatial distribution (comparing Panels a and b in Fig. 6), with an average relative absolute error of approximately 5%. The uncertainty indicator reaches the highest values (highlighted in yellow in Fig. 6c) where observations are sparsely distributed, particularly in the Alps (zone Z_1) and in northern Germany (zone Z_2). It is important to note that our objective here goes beyond improving the predictive capability of the RF model: given this level of prediction uncertainty (Fig. 6c), we aim, in the following, to investigate the main drivers of this uncertainty and whether they differ from the drivers of the best estimate of the prediction (Fig. 6b).

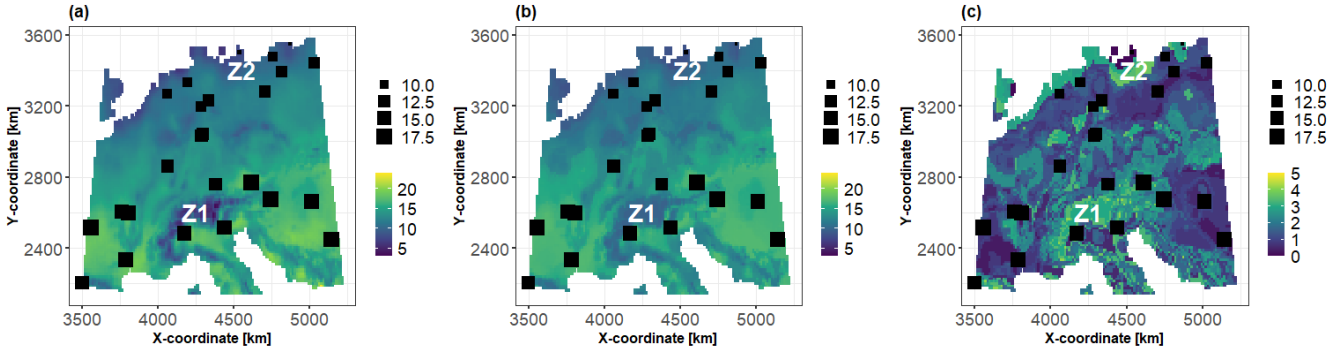


Figure 6: (a) Spatial distribution of the synthetic variable of interest modelled by Eq. 1; (b) Prediction best estimate using the conditional mean of the RF model. (c) Interquartile width *IQW* computed using the 25th and 75th quantiles of the qRF model. The 25 soil samples are indicated by squares whose sizes are proportional to the value of the synthetic variable of interest. The results are more specifically discussed in Sect. 4.1 for the zones indicated by $Z_{1\text{-}2}$.

To ease the interpretation, we group the temperature variables T_{max} and T_{mean} because they are, by nature, physically related. To further support this choice, we apply a partitioning around medoids clustering algorithm (Rdusseeun & Kaufman 1987) using the matrix of the pairwise *HSIC* dependence measures (provided in Supplementary Materials A). The average silhouette

width reaches 0.32 and 0.41 for two and three groups, respectively, which justifies the use of three groups, namely, $T_{mean\text{-}max}$, T_{range}, and $P_{wettest}$.

Using the trained RF model and the selected groups of covariates, we apply the group-based SHAP approach to decompose the data at the 10,000 grid points of the study area. An example of this analysis at the grid point of coordinates (4,206,729 m, 2,149,423 m) is provided in Fig. 1. To facilitate comparison across the study area, we plot the scaled Shapley values (as defined in Sect. 3.2) and use them to map the contributions to the prediction best estimate (i.e., the conditional mean, Fig. 7a) and to the corresponding uncertainty (i.e., *IQW*, Fig. 7b). With regard to the prediction best estimate, the upper panels of Fig. 7 show that the three groups of covariates have Shapley values in the range [25, 50%] over ≈75% of the whole study area. Figure 7 shows that the major contributors to the prediction best estimate and to the uncertainty differ. This is exemplified by the two zones where the uncertainty is the highest (see Fig. 6c). In the central zone around the Alps (zone Z_1 in Fig. 7), the major contributors to the best estimate and to the uncertainty are $P_{wettest}$ (with contributions in the range of 50-75%), the group $T_{mean\text{-}max}$ (with contributions in the range of 10-25%) and T_{range} over a more limited spatial extent. On the other hand, in northern Germany (zone Z_2 in Fig. 7), $P_{wettest}$ contributes the most to the uncertainty, with contributions in the range of 10-25%, whereas the group $T_{mean\text{-}max}$ contributes the most to the prediction mean, with contributions in the range of 50-75%. These differences in importance in both zones may also be related to these covariates having uniquely high and low values, respectively, which are not represented in the training dataset (see Fig. 2). This means that in these two zones, the decisions regarding the characterisation of the covariates are different; this is discussed in more detail in Sect. 5.1.

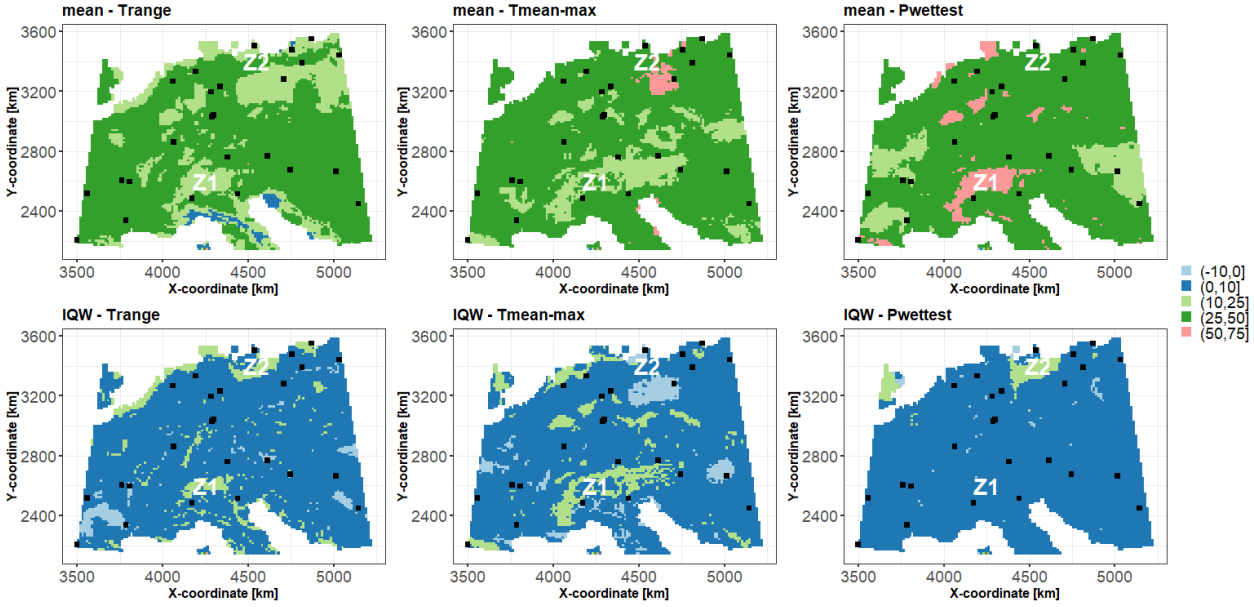


Figure 7: Scaled Shapley values (in %) for the synthetic test case considering the prediction best estimate modelled by the RF conditional mean (top) and the prediction uncertainty modelled by the qRF interquartile width *IQW* (bottom). The black squares indicate the spatial locations of the 25 soil samples. The results are more specifically discussed in Sect. 4.1 for the zones indicated by $Z_{1\text{-}2}$.

4.2 Application to the real case

We construct an RF model using the 1,043 soil samples described in Sect. 2.2 to estimate the conditional mean, which is used as the best estimate of the prediction, and the interquartile width *IQW*, which is used as the uncertainty estimate, with the 25th and 75th quantiles computed using a qRF model. One additional difficulty is related to the clustering of the observations, as shown in Fig. 3a. To alleviate this problem, we follow an approach similar to that of Bel et al. (2009). First, we use an inverse sampling intensity weighting to assign more weight to the observations in sparsely sampled zones and less weight to the observations in densely sampled zones. Second, to estimate the sampling intensity, we use a two-dimensional normal kernel density estimation with bandwidth values estimated based on the rule of thumb of Venables and Ripley (2002). Finally, the inverse sampling intensity weights, with prior normalisation to between 0 and 1, are used for RF training during bootstrap sampling (see Appendix B) to create individual trees with different probability weights by following a method similar to that of Xu et al. (2016).

To select the RF parameters *ns* and m_{try}, we use a cross-validation exercise similar to that used for the synthetic case. This tuning procedure selects the pairs (*ns*, m_{try}) for which the average relative absolute error is minimised, yielding (*ns*=5, m_{try}=2) as the combination that results in the lowest error of ≈12% (averaged over 25 replicates of the 10-fold cross-validation) with a frequency of 76% (i.e., 19 replicates out of 25).

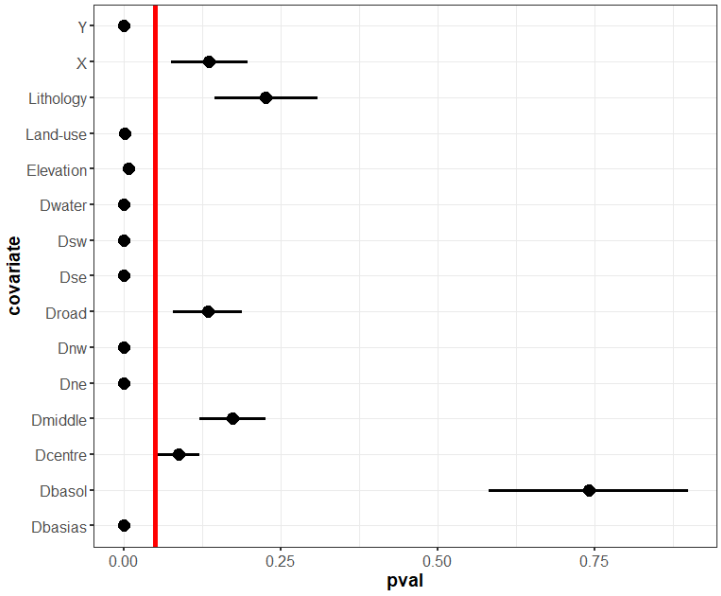


Figure 8: Screening analysis showing the p values of the *HSIC*-based test of independence (described in Appendix C) for the Toulouse case. The dots indicate the mean values estimated over the replicates of a 10-fold cross-validation (repeated 25 times). The lower and upper bounds of the error bars are defined as +/- one standard deviation. When the dot merges with the error bar, the value of the standard deviation is low. The vertical red line indicates the significance threshold at 5%. When the p value is less than 5%, the null hypothesis should be rejected, i.e., the considered covariate has a significant influence on the hydrocarbon concentration and is retained in the RF construction.

A screening analysis is performed within the cross-validation procedure using *HSIC* measures combined with a hypothesis testing procedure. Figure 10 shows the statistics of the p values calculated over the replicates of the 10-fold cross-validation (repeated 25 times). Several observations can be made:

- The distance to potentially polluted sites, D_{basos}, has a minor influence, contrary to the distance to industrial sites D_{basias} (abandoned or active). This is due to the high dependence of D_{basos} (whose *HSIC* is on the order of 0.93-0.95; Supplementary Materials B) with elevation and geographical coordinates. This is also supported by Fig. 3, which suggests that polluted sites tend to be located in relatively low-lying areas in the vicinity of the city centre. In other words, the inclusion of D_{basos} in the analysis is redundant with respect to the information provided by the covariates on which it is dependent;
- Land use has a strong impact, whereas lithology appears to have little impact, with a p value on the order of 25%, i.e., larger than the significance threshold. This is interpreted as being related to the hydrocarbon nature of the pollution, which is less strongly related to geological processes than heavy metal pollution, for instance;
- The distance to roads was not included even though its relation to hydrocarbon concentration was expected. This is due less to its dependence on the other covariates, whose *HSIC* values are as high as 0.14 (Supplementary Materials B), than to its very dense spatial distribution: the value of this covariate varies very little over a large area, as indicated by the almost homogeneous colour in Fig. 3, i.e., very few zones are discriminated by this covariate in this case.
- Out of all the cross-validation replicates, nine covariates have a statistically significant influence on hydrocarbon concentration considering a significance threshold of 5%. These covariates are retained in the construction of the RF model.

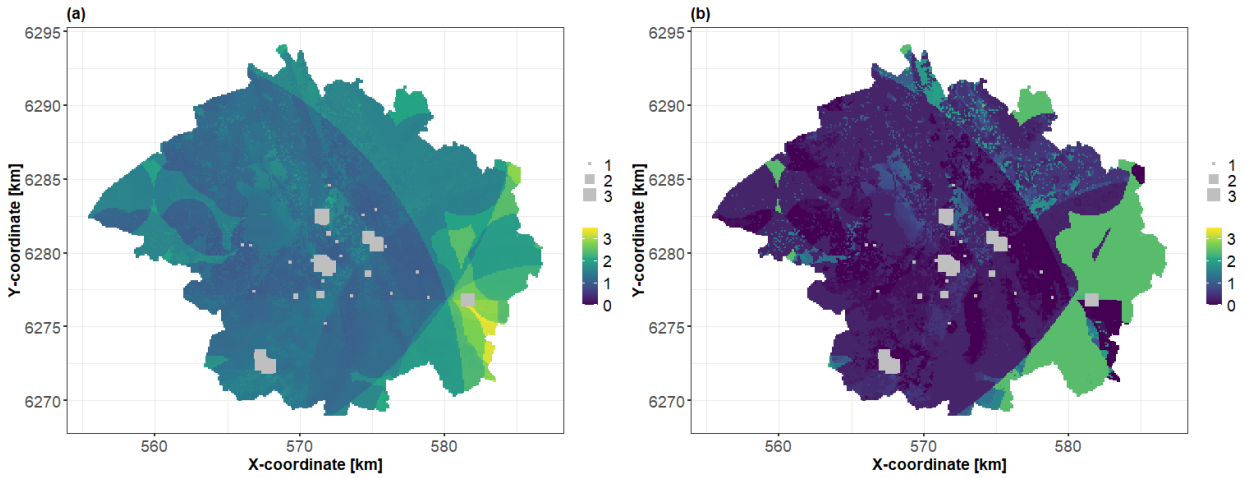


Figure 9: (a) Prediction best estimate of the hydrocarbon concentration (log_{10} scale) for the city of Toulouse using the conditional mean of the RF model. The soil samples are indicated by squares whose sizes are proportional to the log_{10} of the hydrocarbon concentration. (b) Interquartile width *IQW* computed using the 25th and 75th quantiles of the qRF model (expressed on the same scale as the hydrocarbon concentration).

Figure 9 shows the prediction (Panel a) of the hydrocarbon concentration together with the uncertainty measure (Panel b) at the grid points across the city with a spatial resolution of 100 × 100 m². Notably, a large proportion of the city has a predicted concentration varying between 1 and 2 (on a $\log_{10}$ scale), with the exception of the southeastern part, where the concentration is predicted to be >3. In this zone, the uncertainty of the prediction is the highest, with values ranging up to ≈2.5. Outside this zone, a large proportion of the study area has uncertainty estimates of <1.0, with some zones having uncertainties of <0.01, particularly in the vicinity of the observations.

To ease the interpretation of the Shapley values, we define two groups of covariates whose contributions are considered together with the land use and the elevation:

- The group $D_{basias\text{-}water}$ includes D_{basias} and D_{water}. The analysis of the joint influence is meaningful because this group reflects the general tendency of industrial sites to be located close to a water supply;
- The *geographical* group includes D_{ne}, D_{se}, D_{nw}, D_{sw}, and the Y-coordinate. This group of covariates was introduced to improve the predictive capability of the RF (see Sect. 2.2). Interpreting the respective influence of each of these individual covariates is often difficult in practice, and grouping them makes sense in this regard.

To support these choices, we analyse the pairwise dependence measure (Supplementary Materials B), which confirms the moderate-to-high pairwise dependence of all variables in this group. This analysis also indicates that land use has a low dependence with the other variables.

Using the trained RF model, we apply the group-based SHAP approach to decompose the data at >45,000 grid points in the study area. As for the synthetic test case, we compute the scaled Shapley values and use them to map the contribution to the prediction best estimate (i.e., the conditional mean; Fig. 10, left) and to the corresponding uncertainty (i.e., *IQW*; Fig. 10, right). With regard to the prediction best estimate, the left panels of Fig. 10 show that the four covariate groups have contributions, to some extent, of approximately the same order of magnitude. The different groups have high influence, with scaled Shapley values within the range [25, 50%] but in different areas of the city, namely, in the central part of the city for *elevation* and *land use*, i.e., in low-lying industrial areas (Fig. 4); in the western part for $D_{basias\text{-}water}$; and in the eastern part for the *geographical* group. The *geographical* group has the largest contribution in the southeastern part of the city where the predicted values are the largest, as outlined by the dark green zone in Fig. 10, left. This is also confirmed by the analysis of the boxplots provided in Supplementary Materials B. Regarding the prediction uncertainty, the scaled Shapley's values are mainly in the range [0, 25%], but with some particular areas where the determining factors for either the best prediction estimate or uncertainty, or both, are not necessarily the same.

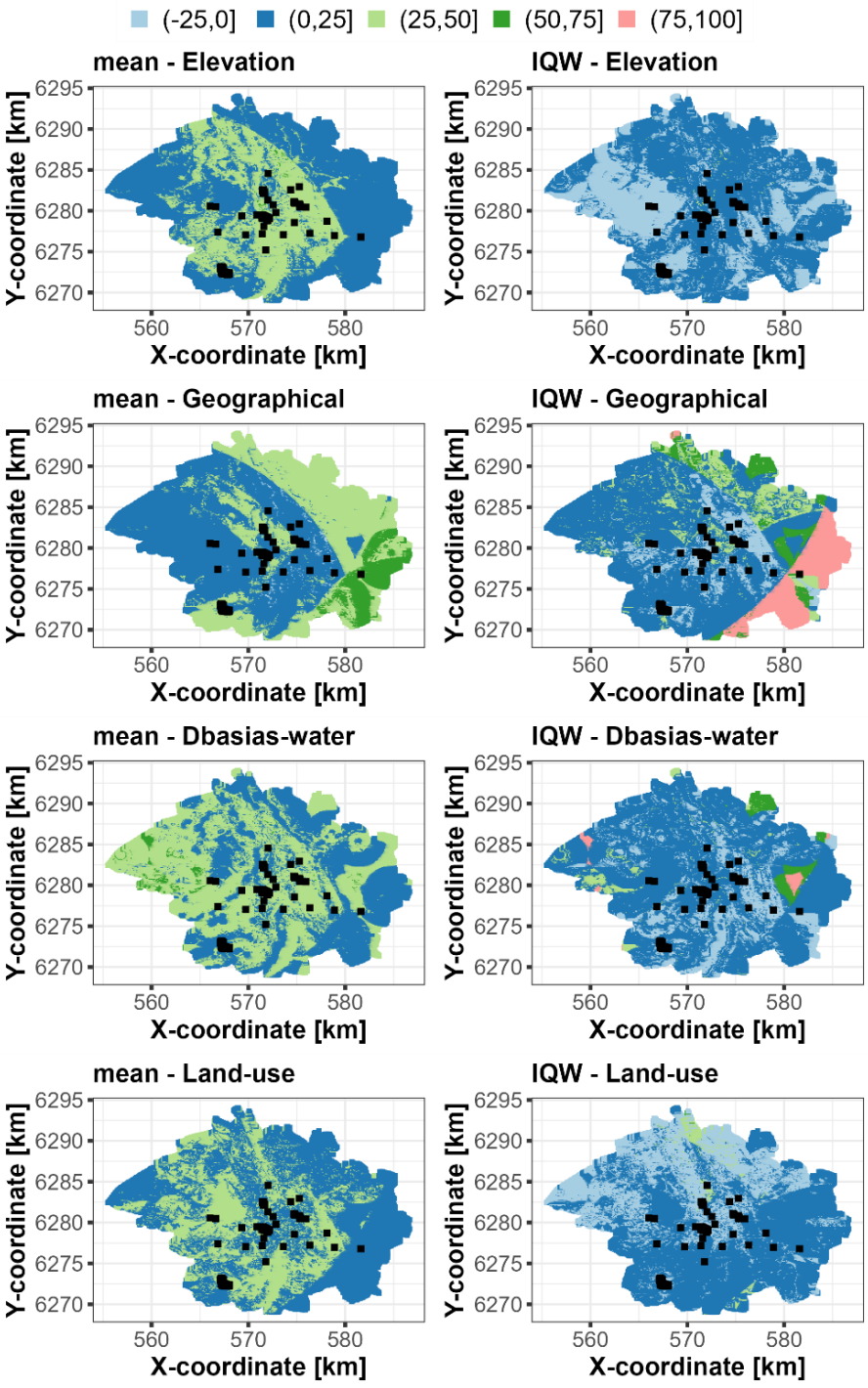


Figure 10: Scaled Shapley values (in %) for each group of covariates of the Toulouse test case considering the prediction best estimate using the RF conditional mean (left) and the prediction uncertainty using the qRF interquartile width (*IQW*) (right). The black squares indicate the locations of the soil samples used for RF training.

Three distinct situations are identified that are relevant from the viewpoint of uncertainty management. The first situation corresponds to the locations outlined in Fig. 11, where the groups of covariates contribute significantly to the prediction best estimate, with a scaled Shapley value exceeding that of the uncertainty by more than 25%. Interestingly, the main contributors to this situation are the three groups that are relevant to the soil prediction problem, as opposed to the *geographical* group, whose objective is to account for spatial dependence. This gives some confidence in the process underlying the RF prediction because it indicates that the best estimate is controlled mainly by the soil-relevant predictors. This also indicates that their influence on the prediction is not masked by the use of geographical covariates, i.e., the use of spatial surrogate covariates, also called pseudocovariates, as discussed by Wadoux et al. (2020): Sect. 3.3.

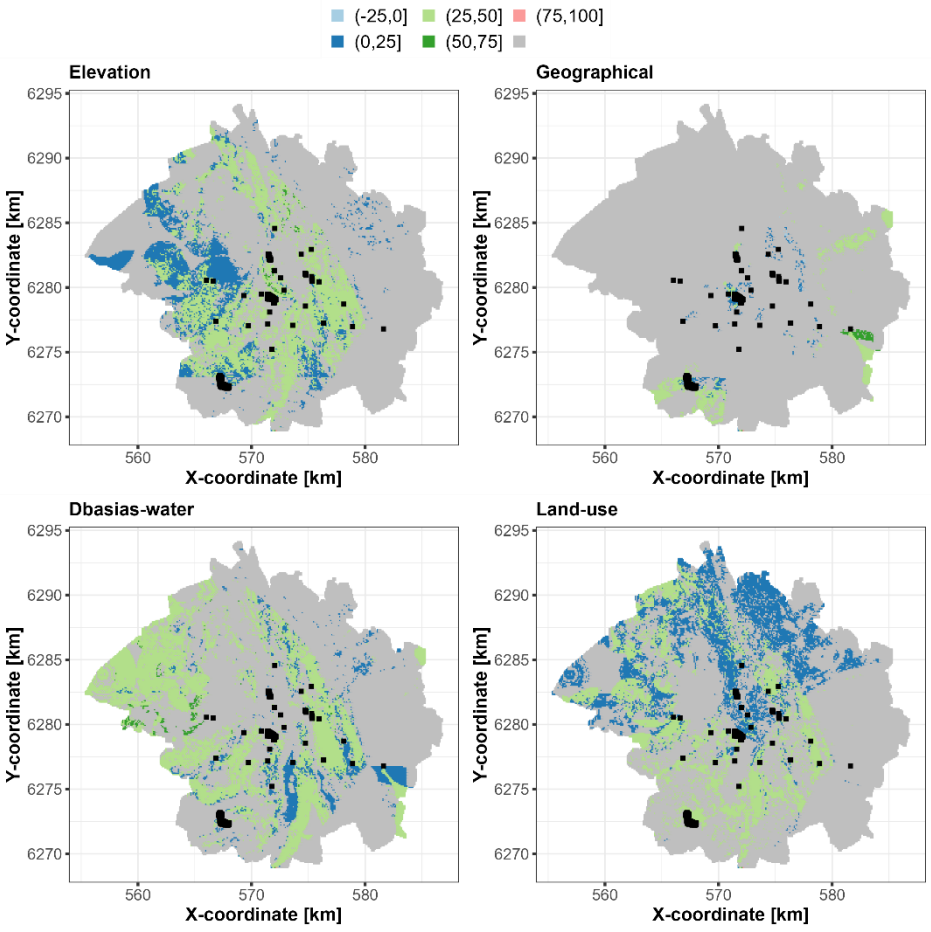


Figure 11: Locations where each corresponding group of covariates contributes significantly to the best estimate, with a scaled Shapley value exceeding that of the uncertainty by more than 25%. The colours indicate the scaled Shapley value for the best estimate. The areas in grey correspond to grid points where the condition is not met. The black squares indicate the locations of the soil samples used for RF training.

An examination of the distribution of the corresponding covariates (Fig. 12) reveals that these locations have elevation values and distances D_{basias} and D_{water} of the same order of magnitude as those in the training dataset, resulting in a prediction situation that does not rely on the extrapolation capability of the RF model. In the areas where land use contributes most to this situation, we show (Fig. 12, bottom, right-hand panel) that this is linked to agricultural areas and forests, i.e., areas where there is less chance of finding potentially polluted sites, as shown by the analysis of the training dataset.

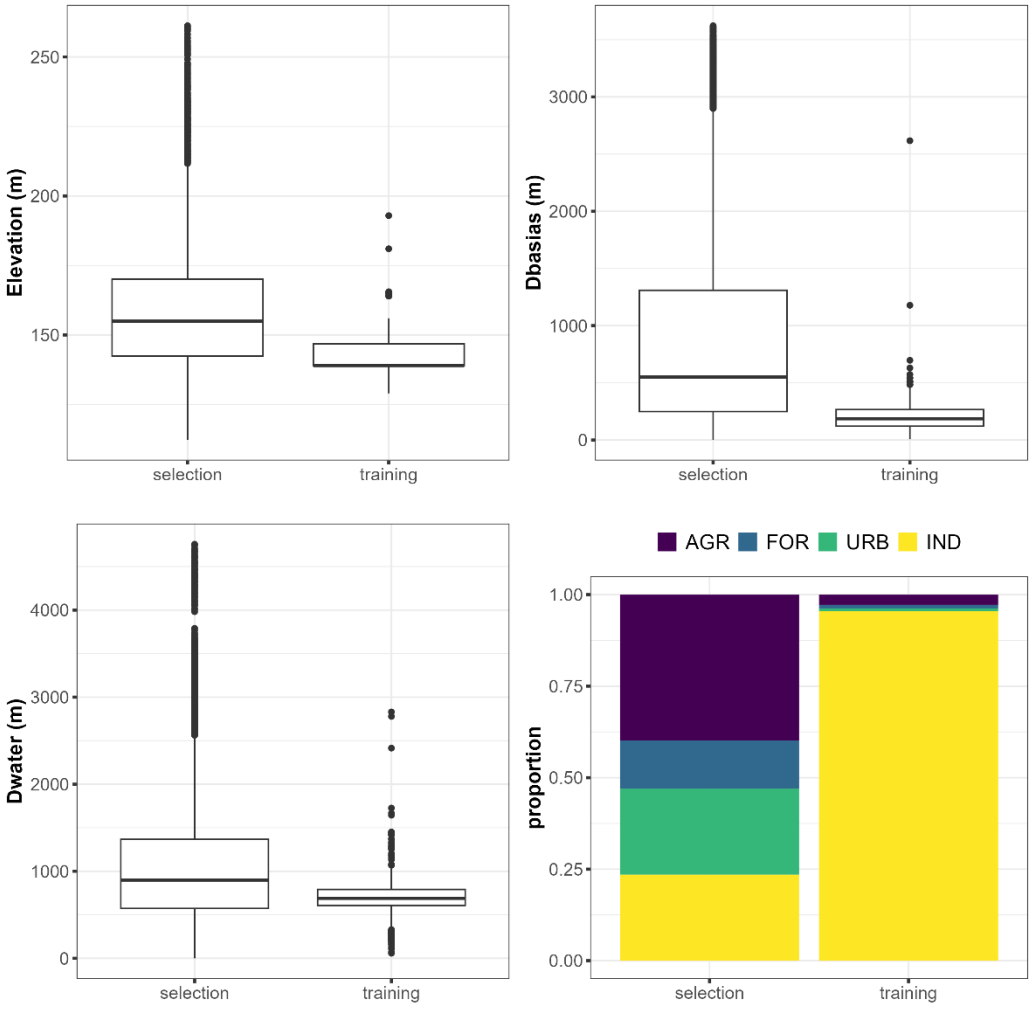


Figure 12: Boxplots of the covariate values for the training dataset and for the locations (named "selection") where the corresponding group of covariates contributes significantly to the best estimate, with a scaled Shapley value exceeding that of the uncertainty by more than 25%. The bottom right-hand panel compares the proportions of land use categories (AGR: agriculture, FOR: forests and grasslands, IND: industrial and commercial economic activities) for the selection and training datasets.

The second situation is the opposite of the first and corresponds to locations where the groups of covariates contribute significantly to the uncertainty, exceeding their contributions to the best estimate by more than 25%. These are shown in dark green and light red in Fig. 10 (right) and only concern the *geographical* and $D_{\text{basias-water}}$ groups. In these areas, the RF models predict cases that are well outside the range observed in the training dataset. This is clear for the identified area for the *geographical* group, which is located outside the spatial domain covered by the soil samples, with the exception of one soil sample located farthest east. An examination of the distributions of the distances D_{basias} and D_{water} reveals that the median values exceed those of the training dataset by factors of 10 and 4, respectively (Supplementary Materials B). This indicates that the RF model is being used here in an extrapolation mode, i.e., a mode where RF models are known to have various limitations (Takoutsing and Heuvelink, 2022).

Finally, the third situation corresponds to where the covariate groups have negative contributions (outlined in light blue in Figure 10, right), i.e., where they participate directly in reducing the prediction uncertainty. An examination of the distribution of the corresponding covariates (Supplementary material B) indicates the same result as for the first situation, with distances to the nearest rivers even smaller than those observed in the training dataset and with even more marked land use behaviour, where agriculture, forest and urban are almost the only categories identified. We also note that the *geographical* group contributes negatively to uncertainty only in the vicinity of the soil samples at distances of 1 to 2 km.

5 Discussion

5.1 Usefulness of the results

To date, the Shapley values have been used to explain individual predictions related to a certain instances of covariates by computing the contribution of each of them to the prediction. Translated for DSM, the Shapley values can be used to determine why a spatial ML model reached a certain value for the soil or chemical properties at a certain spatial location. As discussed by Wadoux and Molnar (2022), the use of Shapley values has the potential to constitute a key tool for environmental soil scientists to improve the interpretation of ML-based DSMs, particularly by providing insights into the underlying physical processes that drive soil variations.

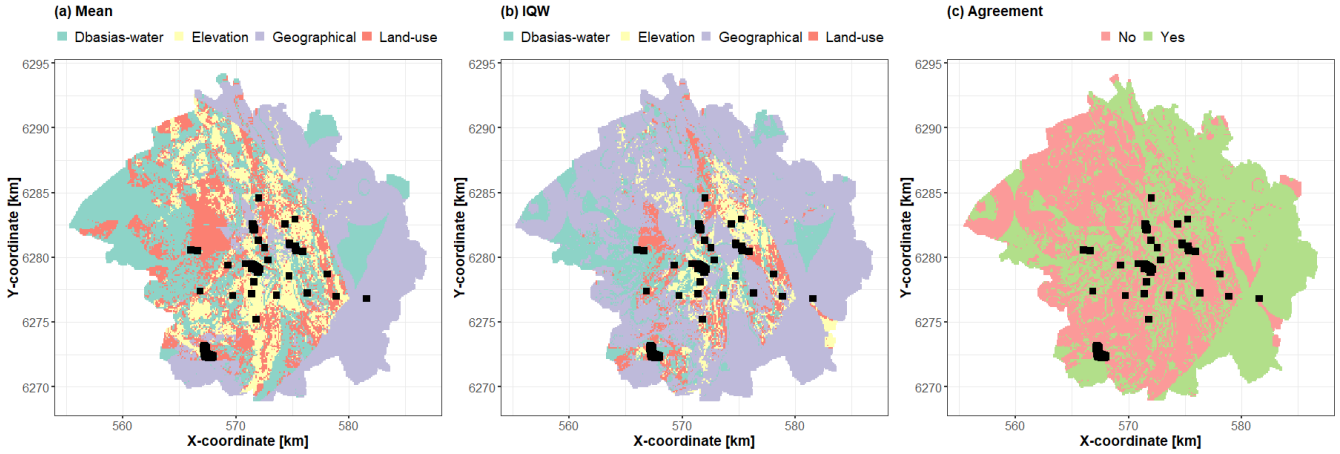


Figure 13: Spatial distributions of the most important group of covariates with respect to the scaled Shapley values for the prediction best estimate (mean, Panel a) and for the prediction uncertainty (*IQW*, Panel b) in the city of Toulouse. Panel c shows the regions where the most important group of covariates for the mean agrees with that for *IQW*. The black squares indicate the locations of the soil samples used for RF training.

In this study, we complement this type of analysis by addressing the "why" question with respect to the prediction uncertainty, i.e., by explaining why the spatial ML model is confident. For this purpose, the SHAP approach for estimating the Shapley values is applied to decompose the uncertainty indicator provided by the ML spatial model. The attribution results are expected to facilitate communication between environmental soil scientists and stakeholders, which is essential for the inclusion of these new digital soil map products in current practices (see, e.g., the discussion by Arrouays et al., 2020). The SHAP results are expected to improve the framing of the prediction results together with the associated uncertainty. The following narrative follows the example of the synthetic test case (Fig. 1): "The predicted value of 12.20 is mainly attributable, by a positive factor of almost 50%, to the maximum and mean temperature of the warmest quarter of the year. The confidence in this result measured by the uncertainty indicator of 1.98 is explained by the diurnal range by almost 50%. To further increase this confidence by decreasing the uncertainty, next efforts should concentrate on the characterisation of this particular covariate".

The second implication of our study is in terms of uncertainty management. Our application to hydrocarbon concentration mapping in Toulouse, as well as to the synthetic test case, reveals that the determining contributors to the best estimate or the uncertainty may not necessarily be the same. Fig. 13 shows maps of the most important groups of covariates with respect to the scaled Shapley absolute values for the real case. These maps show that the prediction uncertainty is dominated by the *geographical* group over almost 65% of the entire study area, whereas the best estimate is influenced by this group of covariates over less than 35%; these areas are outlined in purple in Fig. 13a,b. Overall, the most important group of covariates differ for both prediction objectives over about 50% of the entire study area (outlined in red in Fig. 13c), mostly in the western and central part of the city. The dichotomy between the drivers of the best estimate and uncertainty is also illustrated for the synthetic test case in Fig. 7.

On this basis, distinct situations can be identified with different practical implications for data collection and covariate characterisation. This is illustrated in the Toulouse case in Sect. 4.2. If the primary objective of environmental soil scientists is to increase the confidence in the prediction, the characterisation efforts should be concentrated in the zones outside the spatial domain covered by the soil samples, i.e., in regions where the RF models appear to be used in an extrapolation mode. Two improvements are particularly notable: (1) the modelling of the spatial dependence in the ML model, as revealed by the high importance of the *geographical* group, and (2) the need for more samples outside the range covered by the soil samples to better characterise the pair of distances D_{basias} and D_{water}. On the other hand, if the primary objective of environmental soil scientists is to support the communication of the prediction results to end-users, this analysis provides two key elements. First, the results can be explained in the same form as the example provided above by stressing that the predicted pollutant concentration values in the central and western zones of the city are influenced mainly by the covariates that are relevant to the soil prediction problem; different parts of the city being influenced by different groups of covariates (Fig. 11). Second, the results provide evidence for confidence in how the RF model can make predictions, as discussed in Sect. 4.2 based on Fig. 12. Finally, identifying areas where groups of covariates have negative contributions is essential for prioritising actions to reduce prediction uncertainty.

5.2 Applicability to global-scale projects

The application cases analysed in this study correspond to situations with a moderate number of covariates (a few tens) and predictions at either the city scale or the national scale, with several tens of thousands of grid cells, which are representative of other real case situations described in the literature, such as those of Meyer et al. (2018), de Bruin et al. (2022), and Wadoux et al. (2023). In this section, we extend the discussion regarding the applicability to global-scale projects such as that described by Poggio et al. (2021) with hundreds of covariates and millions of grid cells. Applying the SHAP approach is challenging in this context due to its computational load, which is directly related to the number of covariates (Sect. 3.2).

As an illustration, we run SHAP for the Toulouse test case using the nine important covariates (without grouping) at 100 randomly selected grid points (on a Windows Desktop x-64 with a PC – Intel® Core™ i5-13600H, 2,800 MHz, 12-core, 16 logical processor(s) with 32 GB physical RAM), which led to an average CPU time of 2.15 seconds. Given the constraints of global-scale studies, a direct SHAP analysis would require at least 200 days of calculation on a single laptop. The first solution relies on the use of a high-performance computing architecture, as proposed by Wadoux et al. (2023). A second option involves approximating the Shapley values using, for instance, sampling algorithms (Chen et al., 2023), with some approximation errors opposite to those of the exact method used here. A third option explored in this study is the combination of screening analysis and a grouping approach. Although RF models can handle a large number of covariates, eliminating the covariates before calculating the Shapley values has a clear benefit for saving CPU time. In the real case, the SHAP computational complexity is proportional to 2^{15}=32,768. The application of screening analysis (Fig. 8) decreases the number of features from 15 to 9, resulting in a relative computational cost reduction of 2^{15-9}=64. An additional step of grouping is proposed here, with the primary objective of facilitating interpretation. Interestingly, Wadoux et al. (2023) also presented Shapley values for groups

of covariates (mean climate, climate extremes, vegetation, topography, etc.), as indicated in Figure 6 of their study. By grouping before calculating the Shapley values, an additional relative computational cost reduction can be achieved. In the Toulouse case, this implies a cost reduction of $2^{9-4}=32$, and the analysis required less than one hour for the group-based SHAP (with an average CPU time of 0.054 seconds). Given the constraints of global-scale studies, this approach would here require less than 7 days of calculation on a single laptop. With the growing concern regarding energy consumption (see, e.g., Jay et al., 2024) for scientific computing, this option provides soil scientists with efficient, energy-saving analytical tools although it requires a careful identification of the covariates of negligible influence as well as the definition of groups.

6 Summary and future work

Providing insights into the uncertainty impacting DSM is a key challenge that requires appropriate diagnostic tools. In an effort to complement the toolbox of environmental soil scientists, in this study, we assessed the feasibility of using the SHAP approach to quantify the contributions of covariates to the machine learning-based prediction uncertainty at any location in the study area. Using a real case of pollution concentration mapping in the city of Toulouse, as well as a synthetic test case, we explored the benefits of jointly analysing the contributors to the prediction best estimate and to the prediction uncertainty. Our results revealed that the drivers of the prediction best estimate are not necessarily the drivers of the confidence in the predictions: this means that decisions in terms of data collection and covariate characterisation may differ depending on the target, the prediction best estimate or the confidence/uncertainty and the way in which the results of the prediction (and their uncertainty) are communicated.

However, to integrate SHAP at a fully operational level, several lines of improvement need to be considered. First, the implementation to global scale projets still remains challenging and deserves further work to find a comprise between accuracy, efficiency and interpretability by paying a particular attention paid to estimation algorithms (Chen et al., 2023) with a potential combination with screening and grouping analysis. Second, we focused on a unique uncertainty indicator, namely, the interquartile width, but in some situations, it may not be representative of the total uncertainty, and additional developments are necessary to integrate the entire prediction probability distribution within the SHAP setting. The use of an information theoretic variant of Shapley values, as investigated by Watson et al. (2023), may be helpful here. Third, we focused on one type of machine learning model, i.e., the quantile RF model. Alternative approaches should be considered in future research: different types of machine learning models, such as deep learning techniques, which have shown promising results (see, e.g., Kirkwood et al. (2022)), and improved approaches, in particular, to address complex sample distributions such as clustering (see, e.g., de Bruin et al. (2022) and references therein). Different uncertainty measures should also be tested, for example, geostatistical methods using the kriging variance or statistical quantities derived from stochastic simulations (see, e.g., Chilès and Delfiner (2012)); Bayesian techniques (see Abdar et al., 2021 for deep learning techniques); and data-driven approaches such as cross-validation procedures (Ben Salem et al., 2017). These future studies are made possible by the model-agnostic nature of SHAP.

Author contributions

JR designed the concept. SB provided support for the data processing, for the implementation of the machine learning model, and the application to the Toulouse case. JR undertook the statistical analyses. The results were analysed by JR, SB, and DG. JR wrote the manuscript; SB and GD reviewed the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Code/Data availability

Sources of data of the covariates are listed in Table 1 and Table 2. We provide the R scripts to run the synthetic test case in the form of an R markdown on the Github repository: https://github.com/anrhouses/groupSHAP-uncertainty based on the vignette of the R package *CAST* available at: https://hannameyer.github.io/CAST/articles/cast02-AOA-tutorial.html. The data of the Toulouse test case have however restricted access restriction.

Acknowledgements

We acknowledge financial support of the French National Research Agency within the HOUSES project (grant N°ANR-22-CE56-0006). We are grateful to the "Ministere de la Transition Ecologique et Solidaire, Direction Générale de la Prévention des Risques (MTES/DGPR)" and Toulouse Metropole for letting us use the data that supported the study (Belbeze et al., 2019). We thank Sebastien da Veiga (ENSAI) for useful discussions on the use of *HSIC* measures.

Appendix A Implementation

The R package *ranger* developed by Wright & Ziegler (2017) is used to train the RF models as well for the predictions, and quantile estimates. The R package *sensitivity* (https://cran.r-project.org/web/packages/sensitivity/index.html) is used to implement the HSIC-based analysis (screening and dependence). The R package *shapr* developed by Sellereite et al. (2023) is used to implement the group-based SHAP approach. The R package *cluster* by Maechler et al. (2023) is used to implement the PAM clustering method. An R markdown based on the vignette of the R package *CAST* (available at: https://hannameyer.github.io/CAST/articles/cast02-AOA-tutorial.html) is provided on the Github repository: https://github.com/anrhouses/groupSHAP-uncertainty.

Appendix B Quantile Random Forest

The Random Forest (RF) model, as introduced by Breiman (2001), is used here for regression. It is a non-parametric technique based on a combination (ensemble named "forest") of regression trees (Breiman et al. 1984). Each tree is constructed by relying on recursive partitioning, which aims at finding an optimal partition of the covariates' domain of variation by dividing it into L disjoint sets R_1, …, R_L to have homogeneous Y_i values in each set $R_{l=1,\ldots,L}$ by minimizing a splitting criterion (e.g. based on the sum of squared errors, see Breiman et al. 1984) or when the number of observations in each partition reaches a minimal number termed nodesize (denoted *ns*). To sum up, the RF model aggregates the different regression trees as follows: (1) random bootstrap sample from the training data and randomly select m_{try} variables at each split; (2) construct n_{tree} trees $T(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}_t$ denotes the parameter vector based on which the t^{th} tree is built; (3) aggregate the results from the prediction of each single tree to estimate the conditional mean of Y as follows:

$$Y(s) = f(\mathbf{x}(s)) = E(Y|\boldsymbol{X} = \mathbf{x}(s)) = \sum_{j=1}^{n} w_j(\mathbf{x}(s))Y_j, \quad \text{(B1)}$$

where $E(.)$ is the mathematical expectation, and the weights w_j are defined as follows:

$$w_j(\mathbf{x}(s)) = \frac{\sum_{t=1}^{n_{\text{tree}}} w_t(\mathbf{x}(s), \boldsymbol{\alpha}_t)}{n_{tree}}, \text{ with } w_j(\mathbf{x}(s), \boldsymbol{\alpha}) = \frac{\mathrm{I}_{\{X_i \in \mathrm{R}_{l(x,\alpha)}\}}}{\#\{j:\, X_i \in \mathrm{R}_{l(x,\alpha)}\}}, \quad \text{(B2)}$$

where $I(A)$ is the indicator operator which equals 1 if A is true, 0 otherwise; $\mathrm{R}_{l(x,\alpha)}$ is the partition of the tree model with parameter $\boldsymbol{\alpha}$ which contains $\boldsymbol{x}$.

The RF method is very flexible and can be adapted to predict quantiles. The quantile random forest (qRF) model was originally developed by Meinshausen (2006), who proposed to estimate the conditional quantile $q_\tau(y|\boldsymbol{x})$ at level τ as

$$q_\tau(y|\boldsymbol{x}) = \inf(\boldsymbol{x}: F_{H|\mathbf{X}}(y|\boldsymbol{x}) \geq \tau), \quad \text{(B3)}$$

where

$$F_{Y|\mathbf{X}}(y|\mathbf{x}) = \sum_{j=1}^{n} w_j(\boldsymbol{x})\mathrm{I}_{\{Y_j \leq y\}}, \quad \text{(B4)}$$

where the weight are calculated in the same manner as for the regression RF model.

The major difference with the formulation for regression RF is that the qRF model computes a weighed empirical CDF of Y for each partition instead of computing a weighed average value (as in Eqs. B2-B4).

Appendix C HSIC dependence measure

The number of covariates is 15 (Sect. 2), which is sufficiently large to pose some difficulties regarding the computational time cost of the SHAP approach (Sect. 3.2). To filter out covariates of negligible influence (screening analysis), we rely on the *HSIC* (Hilbert–Schmidt independence criterion) measure, which can capture arbitrary dependence between two random variables (potentially of mixed type, continuous or categorical). In the following, we describe the main aspects and the interested readers can refer to Gretton et al. (2005) and da Veiga (2015).

Let us associate X_i with a universal reproducing kernel Hilbert–Schmidt (RKHS) space defined by the characteristic kernel function $k_i(.,.)$. The same transformation is associated with Y by considering a RKHS space with kernel $k(.,.)$. We define the HSIC measure as follows:

$$HSIC(X_i, Y) = \mathrm{E}\big(k_i(X_i, X_i')k(Y, Y')\big) + \mathrm{E}\big(k_i(X_i, X_i')\big)\mathrm{E}\big(k(Y, Y')\big) - 2\mathrm{E}(\mathrm{E}(k_i(X_i, X_i')|X_i)\mathrm{E}(k_i(Y, Y')|Y)), \quad \text{(C1)}$$

where (X_i', Y') is an independent and identically distributed copy of (X_i, Y), and E(.) is the expectation operator.

The role of the characteristic kernel is here central because it can be defined depending on the type of the considered variables. For continuous variables, the Gaussian kernel is used, and is defined as $\exp(-\lambda\|\mathbf{x} - \mathbf{x}'\|^2)$, with λ being the bandwidth parameter chosen as the inverse of the empirical variance of the considered variable. For categorical variables, the identity function is used as a characteristic kernel.

The pairwise dependence is measured by $HSIC\big(X_i, X_j\big)_{i\neq j}$. To ease the interpretability, the scaled version of Eq. (C1) between 0 and 1 is preferably used, namely the ratio between $HSIC\big(X_i, X_j\big)$, and the square root of a normalizing constant equal to $HSIC(X_i, X_i).HSIC\big(X_j, X_j\big)$ as proposed by da Veiga (2015). From the scaled $HSIC\big(X_i, X_j\big)_{i\neq j}$, we define the similarity $S\big(X_i, X_j\big)_{i\neq j} = 1 - HSIC\big(X_i, X_j\big)_{i\neq j}$.

To perform the screening analysis, we rely on the interpretation of $HSIC(X_i, Y)$ from a sensitivity analysis perspective (da Veiga, 2015), namely its nullity indicates that X_i does not influence Y. To identify the significant X_i, the null hypothesis $'\mathrm{H}_0: HSIC(X_i, Y) = 0'$ (against the hypothesis $'\mathrm{H}_1: HSIC(X_i, Y) > 0'$) is tested, and the corresponding p value can be evaluated (El Amri and Marrel, 2021). When the p value is below a given significance threshold (typically of 5%), it indicates that the null hypothesis should be rejected, i.e., the considered covariate X_i has a significant influence on the variable of interest Y.

References

Aas, K., Jullum, M., and Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artificial Intelligence, 298, 103502, 2021.

Adhikari, K., Hartemink, A. E.: Linking soils to ecosystem services - A global review. Geoderma 262, 101–111, 2016.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion, 76, 243-297, 2021.

Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A. C., Morgan, C. L., et al.: Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Regional, 20, e00255, 2020.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., and MacMillan, R. A.: Spatial modelling with Euclidean distance fields and machine learning. European journal of soil science, 69(5), 757-770, 2018.

Bel, L., Allard, D., Laurent, J. M., Cheddadi, R., and Bar-Hen, A.: CART algorithm for spatial data: Application to environmental and ecological data. Computational Statistics & Data Analysis, 53(8), 3082-3093, 2009.

Belbeze, S., Rohmer, J., Négrel, P., and Guyonnet, D.: Defining urban soil geochemical backgrounds: A review for application to the French context. Journal of Geochemical Exploration, 107298, 2023.

Belbeze, S, Assy, Y., Le Cointe, P., and Rame, E.: CAPacité d'Infiltration des eaux pluviales du territoire de TOULouse Métropole (CAPITOUL). Technical Report BRGM/RP71904-FR (in French), 72 pp., 2022. Available at: http://infoterre.brgm.fr/rapports/RP-71904-FR.pdf

Belbeze, S., Djemil, M., Béranger, S., Stochetti, A.: Détermination de FPGA - Fonds Pédo-Géochimiques Anthropisés urbains Agglomération pilote : TOULOUSE MÉTROPOLE. Technical Report BRGM/RP-69502-FR (in French), 347 pp., 2019. Available at: http://ficheinfoterre.brgm.fr/document/RP-69502-FR

Bénard, C., Da Veiga, S., and Scornet, E.: Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. Biometrika, 109(4), 881-900, 2022.

Ben Salem, M., Roustant, O., Gamboa, F., and Tomaso, L.: Universal prediction distribution for surrogate models. SIAM/ASA Journal on Uncertainty Quantification 5(1), 1086-1109, 2017.

Breiman, L.: Random forests. Machine learning, 45, 5-32, 2001.

Chen, H., Covert, I. C., Lundberg, S. M., and Lee, S. I.: Algorithms to estimate Shapley value feature attributions. Nature Machine Intelligence, 5(6), 590-601, 2023.

Chilès J-P, and Delfiner P.: Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, New York, 2012.

Chilès, J.P., and Desassis, N.: Fifty Years of Kriging, in: Handbook of Mathematical Geosciences edited by: Daya Sagar, B., Cheng, Q., and Agterberg, F. (eds). Springer, Cham., 2018.

Da Veiga, S.: Global sensitivity analysis with dependence measures. Journal of Statistical Computation and Simulation 85(7), 1283–1305, 2015.

De Bruin, S., Brus, D. J., Heuvelink, G. B., van Ebbenhorst Tengbergen, T., and Wadoux, A. M. C.: Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics, 69, 101665, 2022.

El Amri, M. R., and Marrel, A.: Optimized HSIC-based tests for sensitivity analysis: Application to thermalhydraulic simulation of accidental scenario on nuclear reactor. Quality and Reliability Engineering International, in press, 2021.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms, in: International conference on algorithmic learning theory (pp. 63-77). Berlin, Heidelberg, Springer Berlin Heidelberg, 2005.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer: Berlin/Heidelberg, Germany, 2009.

Heuvelink, G. B., and Webster, R.: Spatial statistics and soil mapping: A blossoming partnership under pressure. Spatial statistics, 50, 100639, 2022.

Hothorn, T., Hornik, K., and Zeileis, A.: Unbiased Recursive Partitioning: A Conditional Inference Framework, Journal of Computational and Graphical Statistics, 15 (3), 651–74, 2006.

Jay, C., Yu, Y., Crawford, I., Archer-Nicholls, S., James, P., Gledson, A., et al.: Prioritize environmental sustainability in use of AI and data science methods. Nature Geoscience, 1-3, 2024.

Jothi, N. and Husain, W.: Predicting generalized anxiety disorder among women using Shapley value, J. Infect. Public Heal., 14, 103–108, 2021.

Jullum, M., Redelmeier, A., and Aas, K.: Efficient and Simple Prediction Explanations with groupShapley: A Practical Perspective, in: Proceedings of the 2nd Italian Workshop on Explainable Artificial Intelligence, 28–43. CEUR Workshop Proceedings, 2021.

Kirkwood, C., Economou, T., Pugeault, N., and Odbert, H.: Bayesian deep learning for spatial interpolation in the presence of auxiliary information. Mathematical Geosciences, 54(3), 507-531, 2022.

Leprond, H.: Annual review of the "Sensitive Buildings" project. Technical Report BRGM/RP-62878-FR (in French), 24 pp., 2013.

Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., and Meyer, H.: Assessing and improving the transferability of current global spatial prediction models. Global Ecology and Biogeography, 32(3), 356-368, 2023.

Lundberg, S. M., and Lee, S. I.: A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K.: cluster: Cluster Analysis Basics and Extensions. R package version 2.1.6, 2023.

McBratney, A. B., Santos, M. M., and Minasny, B.: On digital soil mapping. Geoderma, 117(1-2), 3-52, 2003.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software, 101, 1-9, 2018.

Meinshausen, N.: Quantile regression forests. Journal of machine learning research, 7(6), 2006.

Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.), Available at: christophm.github.io/interpretable-ml-book/ (last access 2 January 2024), 2022.

Padarian, J., McBratney, A. B., and Minasny, B.: Game theory interpretation of digital soil mapping convolutional neural networks, SOIL, 6, 389–397, https://doi.org/10.5194/soil-6-389-2020, 2020.

Panagos, P., Van Liedekerke, M., Borrelli, P., Köninger, J., Ballabio, C., Orgiazzi, A., Lugato, E., Liakos, L., Hervas, J., Jones, A. and Montanarella, L.: European Soil Data Centre 2.0: Soil data and knowledge in support of the EU policies. European Journal of Soil Science, 73(6), 2022.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil, 7(1), 217-240, 2021.

Rdusseeun, L. K. P. J., and Kaufman, P.: Clustering by means of medoids, in Proceedings of the statistical data analysis based on the L1 norm conference, Neuchatel, Switzerland, 1987.

Redelmeier, A., Jullum, M., and Aas, K.: Explaining predictive models with mixed features using Shapley values and conditional inference trees, in International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 117-137, 2020.

Sellereite, N., Jullum, M., Redelmeier, A., and Lachmann, J.: shapr: Prediction Explanation with Dependence-Aware Shapley Values. R package version 0.2.3.9100, https://github.com/NorskRegnesentral/shapr/, https://norskregnesentral.github.io/shapr/, 2023.

Song, H., Liu, H., & Wu, M. C.: A fast kernel independence test for cluster-correlated data. Scientific Reports, 12(1), 21659, 2022.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (Eds.): Global sensitivity analysis: the primer, JohnWiley & Sons, 2008.

Schmidinger, J., and Heuvelink, G. B.: Validation of uncertainty predictions in digital soil mapping. Geoderma, 437, 116585, 2023.

Takoutsing, B., and Heuvelink, G. B.: Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma, 428, 116192, 2022.

Varella, H., Guérif, M., and Buis, S.: Global sensitivity analysis measures the quality of parameter estimation: the case of soil parameters and a crop model. Environmental Modelling & Software, 25(3), 310-319, 2010.

Vaysse, K., and Lagacherie, P.: Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma, 291, 55-64, 2017.

Venables, W. N. and Ripley, B. D.: Modern Applied Statistics with S, Springer, 2002.

Veronesi, F., and Schillaci, C.: Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecological Indicators, 101, 1032-1044, 2019.

Wadoux, A. M. C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359, 2020.

Wadoux, A. M. C., & Molnar, C.: Beyond prediction: methods for interpreting complex models of soil variation. Geoderma, 422, 115953, 2022.

Wadoux, A. M. C., Saby, N., and Martin, M. P.: Shapley values reveal the drivers of soil organic carbon stock prediction. Soil, 9(1), 21-38, 2023.

Watson, D. S., O'Hara, J., Tax, N., Mudd, R., and Guy, I.: Explaining Predictive Uncertainty with Information Theoretic Shapley Values. arXiv preprint arXiv:2306.05724, 2023.

Wright, M. N., and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1–17, 2017.

Xu, R., Nettleton, D., and Nordman, D. J.: Case-specific random forests. Journal of Computational and Graphical Statistics, 25(1), 49-65, 2016.