Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatially validated machine learning distinguishes texture-controlled from texture-independent metals in coastal sediments of Bohai Bay, China

Ibrahim Tabash

g202113230@kfupm.edu.sa

King Fahd University of Petroleum & Minerals

Mohammed Alzyod

King Fahd University of Petroleum & Minerals

Abdulazeez Abdulraheem

King Fahd University of Petroleum & Minerals

Research Article

Keywords: Heavy metals, Coastal sediments, Spatial cross-validation, Shapley values, Conformal prediction, Geochemical normalisation, Bohai Bay

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11695120/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spatially validated machine learning distinguishes texture-controlled from texture-independent metals in coastal sediments of Bohai Bay, China

Ibrahim Tabash[a,*], Mohammed Alzyod[b], Abdulazeez Abdulraheem[c]

[a] Department of Civil and Environmental Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

[b] Department of Electrical Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

[c] Department of Petroleum Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

* *Corresponding author:* Ibrahim Tabash, Department of Civil and Environmental Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia. E-mail: g202113230@kfupm.edu.sa

Highlights

- Eight metals in 161 coastal sediments were modelled under spatial block validation
- Test R^2 from single 80/20 splits ranged over 0.40 for Cr and 1.07 for Ni
- Texture and Al_2O_3 alone predicted Cr, Cu and Zn (R^2 0.76–0.86), not As, Cd or Hg
- Shapley values tied Cu and Zn to Al_2O_3 and As to a west–east gradient
- A texture-conditioned baseline avoided false enrichment in low-Al sands

Abstract

Machine-learning models of heavy metals in sediments are usually evaluated with a single random train–test split, which ignores spatial autocorrelation and says little about why a metal is, or is not, predictable. We re-analysed an open dataset of 161 surface sediments from the coast off Tangshan (Bohai Bay, China) to test how well Al_2O_3, grain size, water depth and location predict Cr, Cu, Zn, Ni, Pb, As, Cd and Hg. Seven learners were tuned by nested cross-validation and scored under random and spatial leave-block-out designs. Under spatial validation the best models explained 83–92% of the variance of Cr, Cu and Zn, 56% of Ni (83% when one geochemically anomalous sample was not scored), 38–62% of Pb and As, at most 32% of Cd, and none of Hg. Across 200 repetitions of a single 80/20 split, test R^2 ranged from 0.54 to 0.94 for Cr and from −0.13 to 0.94 for Ni. Predictor ablation and exact Shapley values attributed Cr, Cu, Zn and Ni mainly to Al_2O_3 and grain size and As to a west–east gradient, whereas Hg showed no spatial structure; this grouping agrees with the published receptor-model apportionment. Conformalised quantile regression brought the coverage of nominal 90% intervals to 0.87–0.93, against 0.68–0.79 without calibration. A baseline

conditioned on texture alone flagged ten stations where at least three metals exceeded their expected concentrations, most of them in the western sector, and avoided the false enrichment that linear Al normalisation produced in low-Al sands. The contrast between texture-based and location-based predictability is a simple, spatially validated complement to receptor modelling.

Keywords: Heavy metals; Coastal sediments; Spatial cross-validation; Shapley values; Conformal prediction; Geochemical normalisation; Bohai Bay

1. Introduction

Coastal sediments are the main sink for trace metals delivered by rivers, outfalls, atmospheric deposition and shipping, and they can later release these metals to the overlying water and to benthic food webs (Sharifuzzaman et al., 2016). A measured concentration is nevertheless difficult to interpret on its own, because the same Cr or Zn value may reflect natural enrichment in fine, aluminosilicate-rich mud or anthropogenic loading of coarse, metal-poor sand (Horowitz and Elrick, 1987; Loring, 1991). Normalisation to Al or to the fine fraction is therefore recommended before contamination is inferred (Loring, 1991; Kersten and Smedes, 2002). The usual procedure assumes a linear relation between a metal and a single normaliser over the whole textural range. This assumption is seldom checked, and it fails where samples lie outside the range over which the regression is well constrained.

The south coast of Tangshan (Hebei Province), on Bohai Bay, is one of the main centres of steel production, coal-fired power generation and port activity in northern China. Jiang et al. (2023) collected 161 surface sediments there in September–October 2021 and made the data openly available (Jiang, 2023). Using geochemical indices, principal component analysis and positive matrix factorisation (PMF), they found that Cr, Cu, Zn, Ni and Pb share a spatial pattern related to grain size. They assigned the largest shares of Cr, Cu, Ni and Zn (39–48%) to a natural lithogenic source, whereas Hg, Cd and As were dominated by fossil-fuel combustion (80%), steel production (82%) and agriculture (65%), respectively. Receptor modelling of this kind describes the covariance among metals. It does not show how much of the spatial variability of each metal can be predicted from sediment properties at unsampled locations, how uncertain such predictions are, or which samples carry more metal than their texture accounts for.

Machine-learning (ML) regression has been used widely to predict metals in soils and sediments from auxiliary variables (Ma et al., 2016; Tan et al., 2019; Wang et al., 2020; Chen et al., 2022; Deng et al., 2023; Hao et al., 2023; Nie et al., 2024; Lee et al., 2024; Chao et al.,

2025), including hybrid models tuned by metaheuristic optimisation for marine sediments (Yaseen et al., 2024). Three weaknesses recur in this literature. First, skill is commonly reported from one random train–test split or from random *k*-fold cross-validation. Where observations are spatially autocorrelated, held-out samples have close neighbours in the training set, so the score reflects interpolation and can be optimistic for prediction at new locations (Roberts et al., 2017; Meyer et al., 2019; Ploton et al., 2020). Second, interpretation usually rests on impurity-based variable importance, which is biased towards correlated and high-cardinality predictors (Strobl et al., 2007). The bias matters for sediment data because the clay, silt and sand fractions are compositional and sum to 100% (Aitchison, 1982). Third, most models return point predictions without calibrated uncertainty, although monitoring decisions depend on whether an observation lies outside its plausible range.

We re-analysed the Tangshan dataset with four aims: (i) to compare seven learners for all eight elements under random and spatial leave-block-out cross-validation with nested tuning, and to measure how much the score of a single random split varies; (ii) to separate the information carried by Al_2O_3, grain size, water depth and location, using predictor-set ablation and exact, group-wise Shapley values; (iii) to attach distribution-free prediction intervals and check their coverage on held-out blocks; and (iv) to derive a texture-conditioned baseline that identifies samples exceeding the concentrations expected from their composition. Spatial block validation, Shapley attribution and conformal prediction are established methods, but to our knowledge they have seldom been applied together to sediment metal data. Our working hypothesis was that metals assigned by PMF mainly to lithogenic sources are predictable from texture under spatial validation, whereas metals dominated by anthropogenic sources are not. If this holds, the contrast between texture-based and location-based predictability offers a per-metal check on source control that is independent of receptor modelling.

2. Materials and methods

2.1. Study area and dataset

The study area covers about 63 km (east–west) by 34 km (north–south) of the shallow shelf off southern Tangshan (118.26–118.99°E, 38.67–38.98°N; Fig. 1), with water depths of 12.4–36.0 m. It receives terrigenous material from several coastal rivers, and tidal and wind-driven currents rework this material into a mosaic of sandy and muddy facies (Jiang et al., 2023). The 161 surface sediments (0–2 cm) were collected on a near-regular grid in September and October 2021. Sampling, grain-size analysis and the determination of Cr, Ni, Cu, Zn, As, Cd,

Pb, Hg and Al_2O_3 are described by Jiang et al. (2023). The complete dataset was obtained from the Dryad repository (Jiang, 2023; https://doi.org/10.5061/dryad.mkkwh7150) and contains no missing values. Eight candidate predictors were available for each sample: longitude, latitude, Al_2O_3 (%), clay, silt and sand fractions (%), mean grain size (Mz, φ) and water depth (m). All eight elements were modelled as targets and are referred to as metals for brevity, although As is a metalloid. Cr, Cu, Zn and Ni, which Jiang et al. (2023) attributed mainly to lithogenic sources, are reported in most detail.

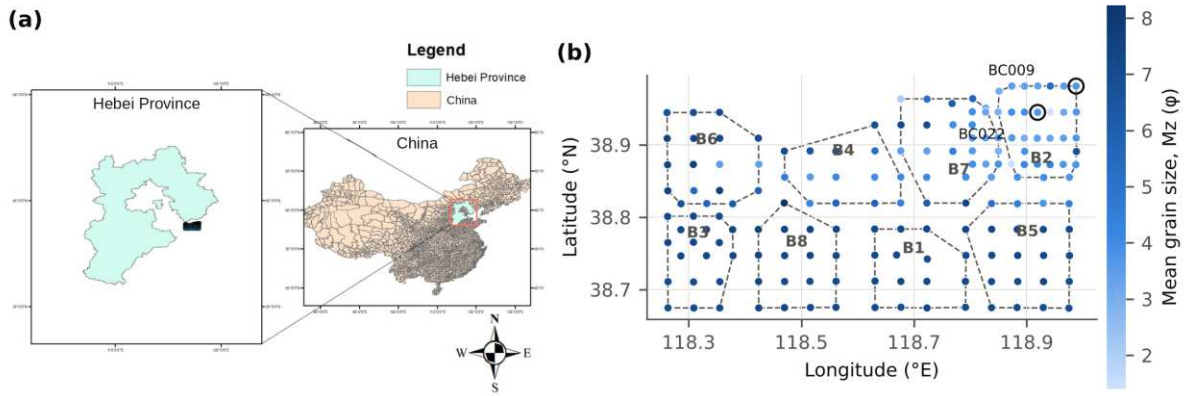


Fig. 1. (a) Study area; (b) sampling stations, mean grain size and spatial validation blocks.

2.2. Exploratory and spatial-structure analysis

Monotonic associations between metals and predictors were measured with Spearman's rank correlation. Spatial autocorrelation was assessed with global Moran's *I* (Moran, 1950) on log-transformed concentrations, using binary distance-band weights of 4.5, 10 and 20 km on coordinates projected to a local equirectangular grid (km). Standardised empirical semivariograms, $\gamma(h)/\sigma^2$, were computed in 4-km lag classes to estimate the distance over which samples remain correlated. The clay, silt and sand fractions are closed compositions (they sum to 100% in every sample) and are therefore perfectly collinear. Sand was omitted from the linear models (MLR and PLSR), whereas the tree and kernel models received all eight predictors.

2.3. Learning algorithms and hyperparameter tuning

Seven regression algorithms from the linear, kernel and tree-ensemble families were compared (Table 1): multiple linear regression (MLR); partial least squares regression (PLSR; Wold et al., 2001); support vector regression with a radial-basis kernel (SVR; Smola and Schölkopf, 2004); Gaussian process regression with an anisotropic squared-exponential kernel plus white noise (GPR; Rasmussen and Williams, 2006); random forest (RF; Breiman, 2001); extremely

randomised trees (ERT; Geurts et al., 2006); and gradient-boosted regression trees (GBRT; Friedman, 2001). Predictors and target were standardised for MLR, PLSR, SVR and GPR, and predictions were back-transformed to concentration units. Hyperparameters were tuned by grid search inside each outer training set (nested cross-validation) with three inner folds and RMSE as the selection criterion; the GPR kernel hyperparameters were estimated by maximising the marginal likelihood. All analyses were run in Python 3.13 with scikit-learn 1.9.1 (Pedregosa et al., 2011), NumPy 2.5.3, SciPy 1.18.1 and Matplotlib 3.11.2.

Table 1. Learning algorithms, predictor sets, pre-processing and hyperparameter search spaces.

Model	Predictors (n)	Pre-processing	Tuned hyperparameters (grid)
MLR	7 (sand omitted)	Standardised X, y	—
PLSR	7 (sand omitted)	Standardised X, y	Latent components: 1–5
SVR (RBF)	8	Standardised X, y	$C \in \{1, 10, 100\}$; $\gamma \in \{1/(p \cdot \mathrm{var}\, X), 0.05\}$; $\varepsilon = 0.1$
GPR	8	Standardised X, y	Constant × RBF (8 length scales) + white noise; marginal-likelihood optimisation, 2 restarts
RF	8	None	$m_{\mathrm{try}} \in \{2, 4, 8\}$; minimum leaf size $\in \{1, 3\}$; 200 trees
ERT	8	None	$m_{\mathrm{try}} \in \{2, 4, 8\}$; minimum leaf size $\in \{1, 3\}$; 200 trees
GBRT	8	None	Trees $\in \{200, 500\}$; depth $\in \{2, 3\}$; learning rate 0.05; subsample 0.8

m_{try}, number of predictors sampled at each split; RBF, radial-basis function; p, number of predictors. The final RF models (Section 2.6) used 500 trees and the most frequently selected settings.

2.4. Validation design

Two outer cross-validation schemes were applied to every combination of metal and model. In *random cross-validation*, five-fold cross-validation was repeated three times with different random partitions. In *spatial leave-block-out cross-validation*, the stations were grouped into eight contiguous blocks by k-means clustering of the projected coordinates (on average 20 stations per block, with a mean maximum extent of 18.4 km; Fig. 1b), and each block was held out in turn (Roberts et al., 2017; Valavi et al., 2019). Three different clusterings served as repetitions, and the inner tuning folds were also grouped by block. With eight blocks, the median distance from a held-out station to its nearest training station increased from about 3.9 km (the sampling interval) to 5.7 km, with a maximum of 15.9 km. The scheme therefore tests prediction into partly unsampled sub-areas, not interpolation between neighbours. Out-of-fold predictions were pooled over the 161 stations in each repetition, and metrics are reported as the mean ± standard deviation (SD) across repetitions.

Three further experiments used fixed, untuned models to limit computation. (i) *Single-split variability.* An RF configured as in a typical single-split study (50 trees, two predictors per split, minimum leaf size 1, eight predictors) was trained on a random 80% of the data and tested

on the remaining 20%. This was repeated for 200 random seeds, and the training-set, out-of-bag and test-set R^2 were recorded. (ii) *Block-size sensitivity.* RF (200 trees, four predictors per split) and MLR were evaluated with random five-fold cross-validation and with 32, 16, 8 and 4 spatial blocks (mean block extents of 7.2, 11.6, 18.4 and 30.1 km), with three repetitions each. (iii) *Predictor-set ablation.* RF (200 trees, half of the predictors per split) was evaluated under eight-block spatial cross-validation with six predictor sets: Al_2O_3 only; texture only (clay, silt, sand and Mz); Al_2O_3 and texture; Al_2O_3, texture and depth; coordinates only; and all eight predictors.

2.5. Performance metrics

Agreement between observed (y_i) and out-of-fold predicted ($\hat{y}_i$) concentrations was measured by the coefficient of determination (equivalent to the Nash–Sutcliffe efficiency for out-of-sample predictions), the root-mean-square error (RMSE), the mean absolute error (MAE) and the mean bias ($\hat{y} - y$):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ , \quad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

where $\bar{y}$ is the mean observed concentration and n the number of stations. The normalised RMSE (nRMSE) is the RMSE expressed as a percentage of the mean concentration. R^2 and RMSE are sensitive to single extreme values. Spearman's rank correlation between observations and predictions was therefore also examined, and the influence of individual extreme samples was assessed by recomputing R^2 without them while keeping them in model training.

2.6. Exact Shapley attribution

Final RF models were trained on all 161 stations with 500 trees and, for each metal, the hyperparameters selected most often in the spatial cross-validation. Predictions were decomposed into additive predictor contributions with Shapley values (Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017). With only $M = 8$ predictors, the interventional Shapley values could be computed exactly by enumerating all $2^8 = 256$ coalitions:

$$\varphi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(M-|S|-1)!}{M!}\left[v(S \cup \{j\}) - v(S)\right]\ , \quad v(S) = \frac{1}{n_b}\sum_{b=1}^{n_b} f\left(x_S, x_S^{(b)}\right) \quad (3)$$

where F is the set of predictors, f is the fitted model, and the background set $\mathbf{x}^{(b)}$ comprises all $n_b = 161$ stations. The decomposition satisfies the efficiency property (the contributions and the mean prediction sum to each prediction) with a maximum numerical error below 10^{-10}. To avoid dividing credit arbitrarily among the collinear compositional fractions, Shapley values were also computed with four predictor groups as players: location (longitude and latitude), Al_2O_3, texture (clay, silt, sand and Mz) and depth. Global importance was summarised as the share of the mean $|\varphi|$ per predictor or group, and Shapley dependence plots were used to examine functional forms. Shapley values describe how a fitted model uses its predictors and are not estimates of causal effects.

2.7. Conformalised prediction intervals

Ninety-per-cent prediction intervals ($\alpha = 0.10$) were derived by conformalised quantile regression (CQR; Romano et al., 2019), which gives finite-sample, distribution-free marginal coverage under exchangeability (Lei et al., 2018). Within each outer spatial fold, two of the seven training blocks were reserved for calibration, and GBRT quantile models (300 trees, depth 2, learning rate 0.05) for the 5th and 95th percentiles were fitted to the remaining five. The conformity scores on the calibration blocks, $E_i = \max\{\hat{q}_{lo}(\mathbf{x}_i) - y_i, y_i - \hat{q}_{hi}(\mathbf{x}_i)\}$, were used to widen the quantile band by Q, their $\lceil (n_c + 1)(1 - \alpha) \rceil$-th smallest value, where n_c is the number of calibration stations:

$$C(x) = \left[\hat{q}_{lo}(x) - Q,\ \hat{q}_{hi}(x) + Q\right] \tag{4}$$

Calibrating on whole blocks keeps the calibration and test sets spatially separated in the same way. The prediction-interval coverage probability (PICP) and the mean prediction-interval width (MPIW) of CQR on the held-out blocks were compared with those of uncalibrated quantile models trained on all seven training blocks.

2.8. Texture-conditioned baseline and excess ratio

To identify samples whose concentrations exceed what their composition accounts for, a texture-conditioned baseline, C_{base}, was predicted for every station from Al_2O_3, clay, silt, sand and Mz only, without location or depth. An RF (300 trees, two predictors per split, minimum leaf size 3) was used in eight-block spatial cross-validation, and the three repetitions were averaged. Because the training data themselves contain enriched samples, the baseline was made resistant to them by iterative exclusion, in analogy with the iterative 2σ technique for geochemical background (Matschullat et al., 2000; Reimann et al., 2005). Within each training

set, samples whose positive out-of-bag residuals exceeded $2\sigma_{MAD}$ were removed and the model was refitted, for up to five iterations. Here $\sigma_{MAD} = 1.4826 \times \text{median}|r_i - \text{median}(r)|$ is a robust estimate of the scale of the residuals r (Rousseeuw and Croux, 1993). The excess ratio (ER) and the criterion used to flag a station are:

$$\text{ER} = \frac{C_{\text{obs}}}{C_{\text{base}}} \; ; \quad \text{station flagged if } C_{\text{obs}} - C_{\text{base}} > 1.645\, \sigma_{\text{MAD}} \tag{5}$$

ER generalises the enrichment factor: its reference is a non-linear function of the whole textural composition, not a single normaliser. If the residuals were normally distributed, about 5% of stations (roughly eight of 161) would exceed the one-sided threshold of $1.645\sigma_{MAD}$ in the absence of any enrichment. Flags for a single metal are therefore treated as screening results, and more weight is given to stations flagged for several metals. For comparison, the same procedure was applied with an ordinary least-squares regression on Al_2O_3 alone, which corresponds to conventional linear Al normalisation (Loring, 1991; Kersten and Smedes, 2002).

3. Results

3.1. Sediment characteristics, correlations and spatial structure

Descriptive statistics are given in Table 2. Concentrations of Cr, Cu, Zn and Ni varied about ten-fold (coefficients of variation 36–45%). The distributions of Cr, Cu and Zn were left-skewed. That of Ni was right-skewed because of a single sample (BC022, 147 mg kg^{-1}) with more than twice the next-highest concentration (69.2 mg kg^{-1}). The sediments ranged from sand (Mz as low as 1.4 φ; sand fraction up to 98.6%) to clayey silt (Mz up to 8.2 φ). Their texture was bimodal, as shown by the large difference between the mean (28.2%) and median (0.9%) sand content.

Table 2. Descriptive statistics of metal concentrations and sediment properties (n = 161).

Variable	Unit	Mean	SD	Min	Median	Max	CV (%)	Skewness
Cr	mg kg^{-1}	84.34	32.06	11.33	94.64	135.54	38.0	−0.52
Cu	mg kg^{-1}	30.94	12.06	3.51	35.96	49.88	39.0	−0.44
Zn	mg kg^{-1}	83.91	30.64	18.35	97.42	126.58	36.5	−0.42
Ni	mg kg^{-1}	37.92	17.23	5.18	44.40	147.00	45.4	1.30
Pb	mg kg^{-1}	27.32	4.70	14.16	28.05	35.31	17.2	−0.58
As	mg kg^{-1}	13.85	4.00	4.81	13.91	25.45	28.9	0.31
Cd	mg kg^{-1}	0.227	0.092	0.040	0.221	0.796	40.5	1.50
Hg	mg kg^{-1}	0.128	0.049	0.014	0.138	0.199	38.6	−0.54
Al_2O_3	%	11.77	1.25	6.25	12.23	13.37	10.6	−1.91
Clay	%	21.45	10.15	0.52	26.29	51.11	47.3	−0.17
Silt	%	50.38	23.34	0.87	66.34	79.05	46.3	−0.47
Sand	%	28.17	32.85	0.00	0.94	98.61	116.6	0.50
Mean grain size (Mz)	φ	5.73	1.75	1.40	6.99	8.23	30.6	−0.50
Water depth	m	25.83	3.73	12.41	25.99	36.04	14.4	−0.38

Data from Jiang (2023). SD, standard deviation; CV, coefficient of variation; Mz, mean grain size.

Cr, Cu, Zn and Ni were strongly and positively correlated with Al_2O_3 (ρ = 0.79–0.86), clay (ρ = 0.74–0.79) and Mz (ρ = 0.77–0.81), and negatively correlated with sand (ρ = −0.80 to −0.83). Water depth was essentially unrelated to any metal (|ρ| ≤ 0.10; Fig. 2a). The associations with Al_2O_3 and the texture variables were weaker for Pb (|ρ| = 0.55–0.66), As (0.43–0.54) and Cd (0.60–0.73), and absent for Hg (|ρ| ≤ 0.11). Moran's I at the 4.5-km band was high for Zn (0.90), Cu (0.85), Cr (0.74) and Ni (0.67), moderate for Pb, Cd and As (0.59–0.67), and close to zero for Hg (−0.03). The semivariogram of Cu rose from 0.34σ² at the shortest lag to the sill at about 20–24 km. That of Hg stayed at the sill from the first lag onwards (pure nugget), which indicates no spatial structure at the 4-km sampling scale (Fig. 2b).

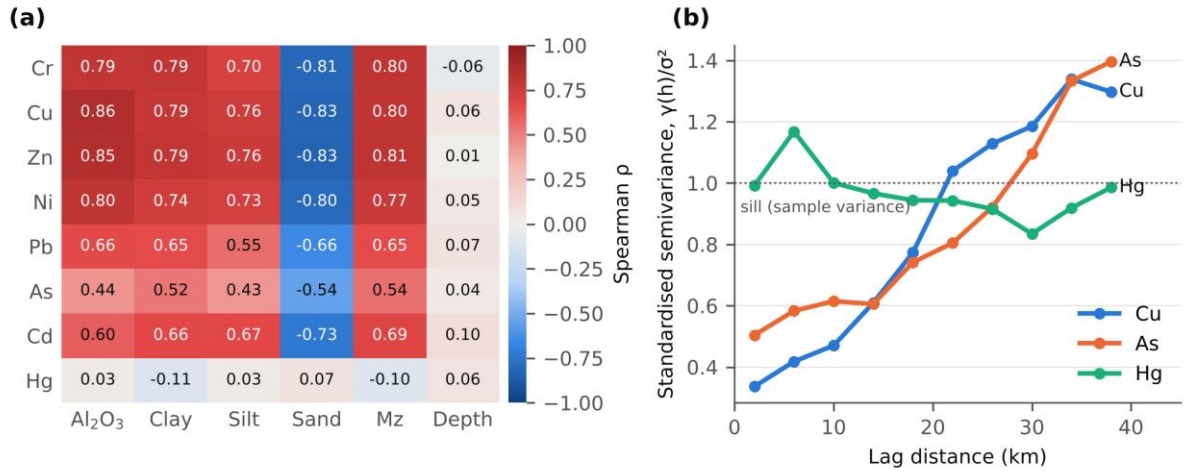


Fig. 2. (a) Metal–sediment Spearman correlations; (b) standardised semivariograms of Cu, As and Hg.

3.2. Model comparison under random and spatial validation

Under spatial leave-block-out validation, predictive skill differed far more between metals than between algorithms (Fig. 3; Table 3; Table S1). The best models explained 83–92% of the variance of Cr (RF, $R^2 = 0.833$), Cu (GBRT, 0.897) and Zn (GPR, 0.915), and the nRMSE of RF for these metals was 11–16%. Pb (best $R^2 = 0.622$, ERT) and Ni (0.559, RF) were intermediate. As (0.379, RF) and Cd (0.324, PLSR) were poorly predicted, and no model predicted Hg appreciably better than its mean (best $R^2 = 0.020$). For the well-predicted metals, the best non-linear model exceeded MLR by 0.06–0.07 in R^2 (for example, 0.915 against 0.846 for Zn), which corresponds to a 14–26% lower RMSE. RF ranked first, or within 0.025 of the best model, for six of the eight metals, and its scores varied little between repetitions (mean SD of $R^2 = 0.015$). GPR failed for Ni under spatial validation ($R^2 = -0.11$), most probably because the extreme sample BC022 distorted the fitted kernel.

Table 3. Cross-validated performance of seven learners for Cr, Cu, Zn and Ni.

Metal	Model	R^2_{sp}	R^2_{rd}	ΔR^2	$RMSE_{sp}$ (mg kg^{-1})	MAE_{sp} (mg kg^{-1})	$Bias_{sp}$ (mg kg^{-1})
Cr	MLR	0.776 ± 0.020	0.816 ± 0.007	+0.040	15.10	10.97	−0.56
	PLSR	0.794 ± 0.007	0.824 ± 0.007	+0.030	14.50	10.46	−0.49
	SVR	0.805 ± 0.012	0.827 ± 0.006	+0.022	14.12	11.09	−0.67
	GPR	0.802 ± 0.011	0.817 ± 0.013	+0.015	14.23	11.42	+0.85
	RF	**0.833 ± 0.005**	0.832 ± 0.014	−0.001	13.04	10.03	+0.66
	ERT	0.817 ± 0.015	0.828 ± 0.001	+0.011	13.68	10.46	+0.98
	GBRT	0.824 ± 0.012	0.833 ± 0.013	+0.009	13.39	10.44	+0.81
Cu	MLR	0.829 ± 0.023	0.873 ± 0.006	+0.044	4.97	3.54	−0.26
	PLSR	0.843 ± 0.008	0.864 ± 0.011	+0.021	4.76	3.38	−0.32
	SVR	0.857 ± 0.003	0.886 ± 0.015	+0.029	4.55	3.35	−0.07
	GPR	0.893 ± 0.004	0.913 ± 0.003	+0.020	3.92	3.11	+0.17
	RF	0.891 ± 0.009	0.901 ± 0.014	+0.010	3.97	2.91	−0.16
	ERT	0.885 ± 0.007	0.907 ± 0.003	+0.022	4.07	2.96	+0.09
	GBRT	**0.897 ± 0.002**	0.911 ± 0.005	+0.014	3.85	2.95	−0.04
Zn	MLR	0.846 ± 0.021	0.883 ± 0.007	+0.037	11.97	8.65	−0.70
	PLSR	0.860 ± 0.014	0.877 ± 0.009	+0.017	11.43	8.27	−0.79
	SVR	0.876 ± 0.020	0.904 ± 0.006	+0.028	10.71	7.79	+0.63
	GPR	**0.915 ± 0.003**	0.927 ± 0.002	+0.012	8.92	6.83	−0.11
	RF	0.905 ± 0.006	0.909 ± 0.008	+0.004	9.43	7.12	−0.29
	ERT	0.896 ± 0.005	0.916 ± 0.001	+0.020	9.85	7.28	+0.02
	GBRT	0.909 ± 0.006	0.919 ± 0.007	+0.010	9.22	7.08	−0.36
Ni	MLR	0.515 ± 0.010	0.536 ± 0.001	+0.021	11.96	5.63	−0.51
	PLSR	0.523 ± 0.008	0.537 ± 0.005	+0.014	11.86	5.47	−0.64
	SVR	0.523 ± 0.009	0.554 ± 0.006	+0.031	11.86	4.94	−1.30
	GPR	−0.109 ± 0.245	0.273 ± 0.033	+0.382	18.00	9.20	+0.30
	RF	**0.559 ± 0.015**	0.573 ± 0.010	+0.014	11.40	5.02	−0.19
	ERT	0.503 ± 0.059	0.542 ± 0.014	+0.039	12.09	5.64	+0.09
	GBRT	0.510 ± 0.024	0.413 ± 0.038	−0.097	12.01	5.62	+0.10

Subscripts sp and rd denote spatial leave-block-out cross-validation (eight blocks) and repeated random five-fold cross-validation, respectively. R^2 values are the mean ± SD of three repetitions; $\Delta R^2 = R^2_{rd} - R^2_{sp}$. The highest R^2_{sp} for each metal is in bold. Bias is the mean of predicted minus observed concentrations. MLR, multiple linear regression; PLSR, partial least squares regression; SVR, support vector regression; GPR, Gaussian process regression; RF, random forest; ERT, extremely randomised trees; GBRT, gradient-boosted regression trees. Results for all eight metals are summarised in Table S1.

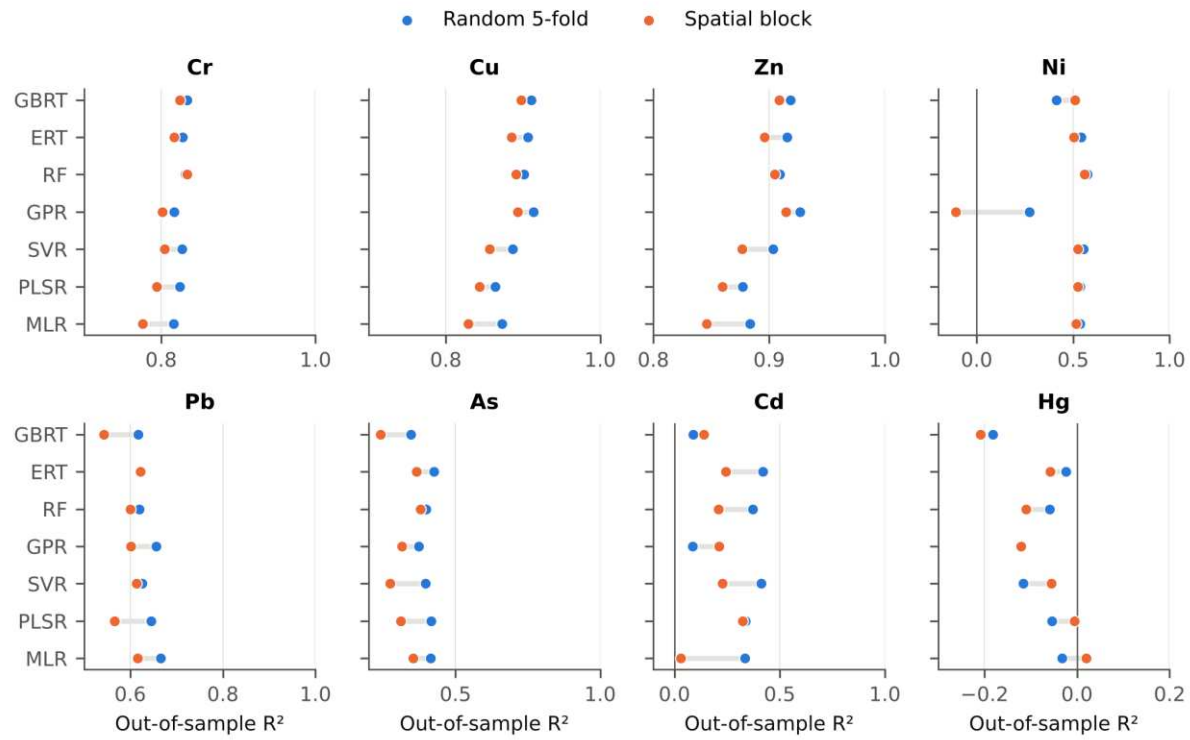


Fig. 3. Out-of-sample R^2 of seven learners under random and spatial block cross-validation.

For Cr, Cu and Zn, R^2 under spatial validation was only 0.018–0.023 lower than under random cross-validation when averaged over the seven learners, and at most 0.01 lower for RF. The difference was larger for the weakly structured metals: on average 0.041 for Pb, 0.076 for As and 0.095 for Cd, and as much as 0.30 for individual learners (MLR for Cd) and 0.38 (GPR for Ni).

The low scores for Ni are due almost entirely to BC022. This sandy station (Mz 3.8 φ) contained 147 mg kg^{-1} Ni but only 46 mg kg^{-1} Cr and 52 mg kg^{-1} Zn, so that its Ni/Cr ratio was 3.2, against a median of 0.43 for the other samples. A model driven by texture cannot be expected to reproduce such a value, and RF predicted 23–26 mg kg^{-1} at this station in every repetition (Fig. 4). When the same out-of-fold predictions were scored without BC022, the spatial R^2 of Ni rose from 0.559 to 0.832 for RF and from 0.515 to 0.778 for MLR. Similarly, scoring without the sandy Cd outlier BC009 (0.796 mg kg^{-1}) raised the spatial R^2 of Cd from 0.209 to 0.384 for RF. Both samples were kept in training and in all reported metrics, and they are considered further in Section 3.6.

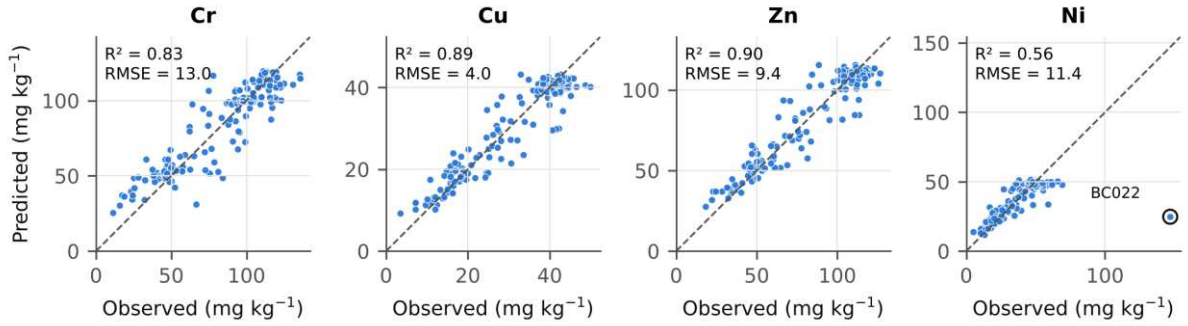


Fig. 4. Observed versus spatially cross-validated random forest predictions of Cr, Cu, Zn and Ni.

3.3. Variability of a single random split

Repeating a conventional 80/20 random split 200 times with an identical RF configuration gave widely different test scores (Fig. 5; Table 4). The 5th–95th percentile range of the test R^2 was 0.67–0.90 for Cr, 0.78–0.94 for Cu and 0.80–0.94 for Zn. For Ni it was 0.09–0.89, and the extremes ran from −0.13 to 0.94 depending on whether BC022 fell in the test set. The median test R^2 of the single splits (0.82–0.90 for Cr, Cu and Zn) was close to the spatial cross-validation estimates (0.83–0.91), but the best single split exceeded these estimates by 0.06–0.11. A test R^2 of 0.95 or more was reached in only 8 of the 800 combinations of metal and split (six for Zn and two for Cu). The training-set R^2 (0.94–0.99) reflects the in-sample fit of fully grown trees and is not a measure of predictive skill. In this dataset, therefore, the score of one split can overstate or understate model skill by an amount comparable to the differences between algorithms.

Table 4. Single-split test R^2 statistics (200 random 80/20 splits) against spatial cross-validation.

	Single 80/20 split (200 repetitions)					Spatial CV
Metal	**Training R^2 (mean)**	**OOB R^2 (mean)**	**Test R^2 (median)**	**Test R^2 (5th–95th pct.)**	**Test R^2 (min–max)**	**R^2 (tuned RF)**
Cr	0.974	0.811	0.823	0.674–0.899	0.540–0.941	0.833
Cu	0.983	0.877	0.887	0.778–0.940	0.738–0.967	0.891
Zn	0.985	0.889	0.895	0.798–0.945	0.742–0.960	0.905
Ni	0.937	0.554	0.796	0.086–0.893	−0.127–0.941	0.559

Single-split protocol: RF with 50 trees, two predictors per split and minimum leaf size 1. Spatial CV: tuned RF under spatial leave-block-out cross-validation. OOB, out-of-bag; CV, cross-validation. The training R^2 is an in-sample value and is given only to show the gap between apparent and out-of-sample performance.

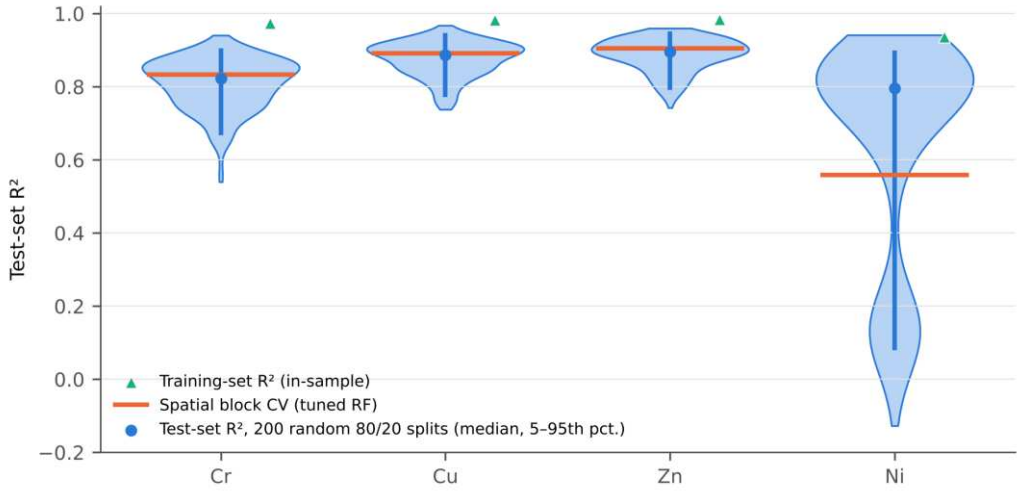


Fig. 5. Distribution of test R^2 across 200 random 80/20 splits versus spatial cross-validation.

3.4. Sensitivity to block size and contribution of predictor groups

Predictive skill was stable from random cross-validation down to eight blocks (mean extent 18.4 km). It declined when only four blocks were used (about 30 km, beyond the semivariogram range of 20–24 km), that is, when each test block lay largely outside the spatial and textural range of the training data (Fig. 6, where R denotes random five-fold cross-validation). For Cu and Zn the decline was much smaller for RF than for MLR (Cu: 0.904 to 0.783 against 0.873 to 0.650; Zn: 0.913 to 0.798 against 0.883 to 0.635). For Pb, As and Cr at four blocks, MLR matched or exceeded RF, which is in keeping with the limited ability of tree ensembles to extrapolate beyond the range of the training predictors.

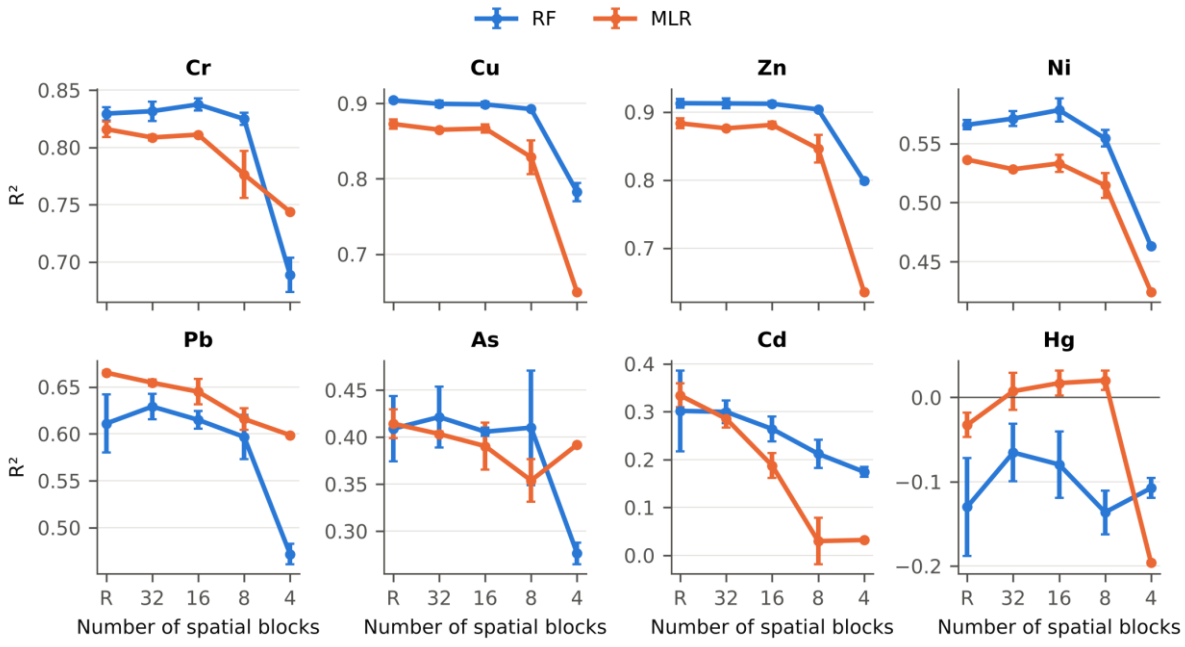


Fig. 6. Sensitivity of out-of-sample R^2 to spatial block size for RF and MLR.

In the predictor-set ablation (Fig. 7a), Al_2O_3 alone gave a spatially validated R^2 of 0.65–0.78 for Cr, Cu and Zn. Adding the grain-size variables raised it to 0.76–0.86, and depth and coordinates added a further 0.02–0.07. For As the pattern was reversed: Al_2O_3 alone performed worse than the mean ($R^2 = -0.54$), Al_2O_3 with texture reached only 0.18, and coordinates alone gave 0.45. Cd behaved in the same way at lower skill (0.19 from Al_2O_3 with texture and 0.29 from coordinates). No predictor set gave positive skill for Hg. When the skill obtained from composition is plotted against the skill obtained from coordinates (Fig. 7b), the metals fall into three groups: texture-controlled (Cu, Zn and Cr, and to a lesser degree Ni and Pb), spatially structured but not texture-controlled (As and Cd), and spatially unstructured (Hg).

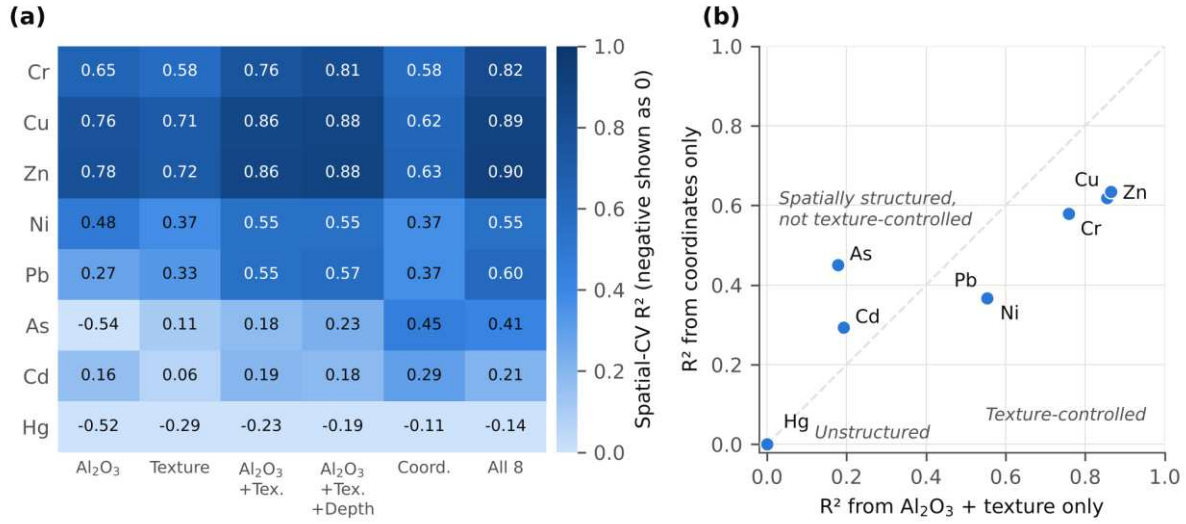


Fig. 7. (a) Spatial cross-validation R^2 by predictor set; (b) texture-based versus coordinate-based predictability.

3.5. Shapley attribution

The grouped Shapley values (Fig. 8a) were consistent with the ablation. Al_2O_3 accounted for 69% of the mean absolute contribution for Zn and 64% for Cu, and Al_2O_3 and texture together accounted for 76–83% for Cr, Cu, Zn and Ni. Location dominated As (75%), was substantial for Cd (33%) and Pb (25%), and contributed only 12–16% for Cr, Cu, Zn and Ni. Among individual predictors (Fig. 8b), Al_2O_3 ranked first for Cr, Cu, Zn and Ni. It was followed by longitude for Cu and Zn, by Mz and longitude for Cr, and by sand, Mz and silt for Ni.

The dependence of Zn on Al_2O_3 was strongly non-linear (Fig. 8c). Contributions rose gradually up to about 11.5% Al_2O_3, increased by about 40 mg kg^{-1} between 11.5% and 12.5%, and levelled off above 12.5%. Contributions to Ni increased with Mz, most steeply between about 5 and 7 φ, and BC022 was the only large positive contribution in coarse sediment. Water depth lowered the predicted Cr by up to 13 mg kg^{-1} at stations shallower than about 24 m, although

its marginal correlation with Cr was negligible ($\rho = -0.06$). For As, the contribution of longitude changed sign near 118.6°E, from +1.5 to +4 mg kg^{-1} in the west to −0.7 to −3.3 mg kg^{-1} in the east. Shapley values were not interpreted for Hg, for which the model had no predictive skill.

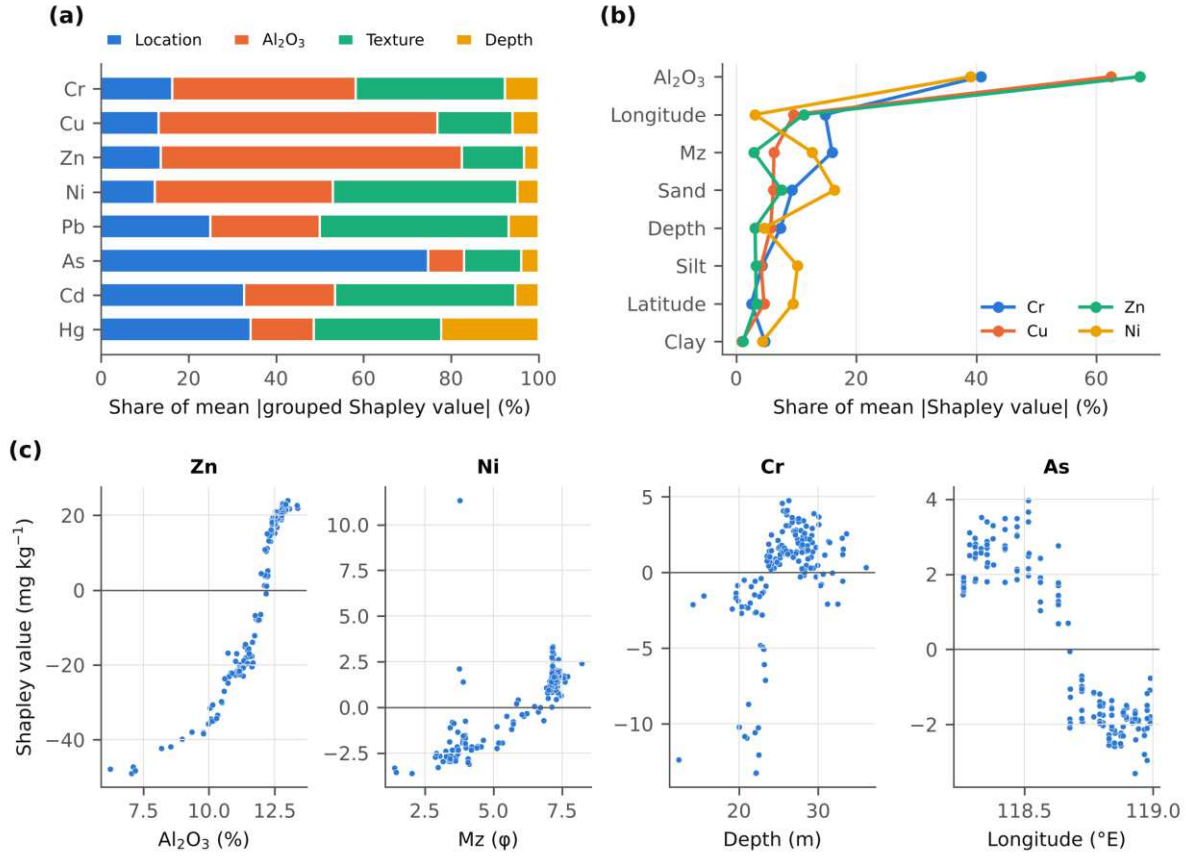


Fig. 8. Exact Shapley attribution: (a) predictor groups, (b) individual predictors, (c) dependence plots.

3.6. Prediction intervals and texture-conditioned excess

Uncalibrated quantile models under-covered consistently under spatial validation: the empirical coverage of the nominal 90% intervals was only 0.68–0.79 (Table 5). Conformal calibration brought the coverage to 0.87–0.93 for all eight metals. The calibrated intervals were wide, however. Their mean width was 78–94% of the mean concentration for Cr, Cu, Zn and Ni and 46% for Pb, which reflects the small calibration sets (two blocks, about 40 stations) and the heterogeneity between blocks.

Table 5. Coverage and mean width of 90% prediction intervals before and after conformal calibration.

	Uncalibrated QR		CQR		
Metal	**PICP**	**MPIW (mg kg^{-1})**	**PICP**	**MPIW (mg kg^{-1})**	**MPIW / mean (%)**
Cr	0.718	42.9	0.865	74.6	88
Cu	0.718	14.8	0.903	25.6	83
Zn	0.696	41.7	0.894	65.7	78
Ni	0.733	27.4	0.899	35.6	94
Pb	0.725	7.5	0.928	12.6	46
As	0.785	7.8	0.909	12.4	89
Cd	0.677	0.135	0.903	0.236	104
Hg	0.760	0.124	0.888	0.165	129

Values are the mean of three repetitions under spatial leave-block-out cross-validation. QR, gradient-boosted quantile regression without calibration; CQR, conformalised quantile regression; PICP, prediction-interval coverage probability; MPIW, mean prediction-interval width.

The texture-conditioned RF baseline reproduced the concentrations of Cu and Zn with a spatially validated R^2 of 0.84–0.85 and those of Cr with 0.73, against 0.52–0.58 for linear Al normalisation (Table 6). The linear baseline gave negative expected concentrations at seven of the eight stations with Al_2O_3 below 9.5% for Cu, Zn and Ni (six for Cr). It consequently flagged six to eight of these low-Al sandy stations as enriched in Cr, Cu, Zn and Ni, with excess ratios as high as 33 (Cu) and 37 (Ni). The RF baseline flagged none of them for Cu, Zn or Ni and one for Cr.

Over the whole area, the RF baseline flagged 15–18 stations for each of Cr, Cu, Zn and Ni (Fig. 9), about twice the number expected by chance (Section 2.8). Ten stations were flagged for at least three of the five metals Cr, Cu, Zn, Ni and Pb. Eight of them (BC058, BC059, BC070, BC081, BC087, BC105, BC106 and BC153) lie between 118.31 and 118.52°E in the western sector, and the other two lie farther east (BC050 at 118.72°E and BC094 at 118.84°E). Within the western group, BC058, BC059 and BC070 are relatively coarse (Mz 3.4–3.6 φ) but contain 78–116 mg kg^{-1} Cr, which is close to the concentrations in fine muds elsewhere. The Cd flags formed a separate cluster in the south-west (118.26–118.47°E) and a short row near 38.93°N between 118.63 and 118.72°E, and they included the north-eastern station BC009. The Ni flags included BC022.

Table 6. Texture-conditioned random forest baseline versus linear Al normalisation for identifying metal excess.

	Texture-conditioned RF baseline			Linear Al normalisation			
Metal	**R^2**	**Flagged (*n*)**	**Flagged, low-Al (*n*)**	**R^2**	**Flagged (*n*)**	**Flagged, low-Al (*n*)**	**$C_{base} \leq 0$ (*n*)**
Cr	0.725	17	1	0.550	11	6	6
Cu	0.837	15	0	0.515	11	8	7
Zn	0.852	16	0	0.579	10	8	7
Ni	0.501	18	0	0.341	15	8	7
Pb	0.512	11	0	0.494	13	3	0
As	0.125	19	2	0.008	19	3	0
Cd	0.252	16	0	0.231	19	0	1
Hg	−0.206	4	0	−0.042	1	1	0

Flagged, stations with $C_{obs} - C_{base} > 1.645\sigma_{MAD}$; low-Al, the eight stations with $Al_2O_3 < 9.5\%$; $C_{base} \leq 0$, stations for which the baseline is zero or negative. All values are from spatial leave-block-out cross-validation.

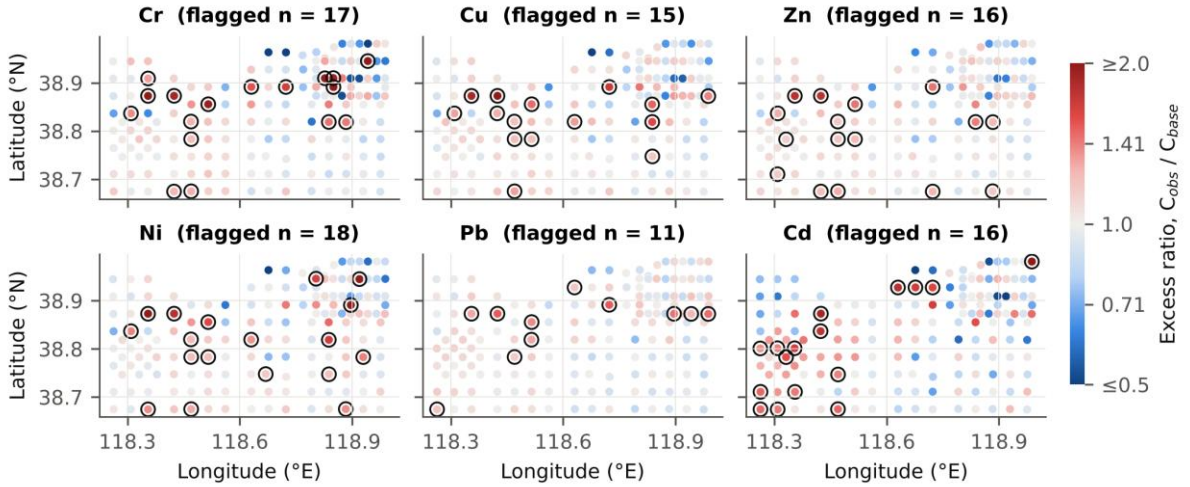


Fig. 9. Excess ratios relative to the texture-conditioned baseline; circles mark flagged stations.

4. Discussion

4.1. Predictability as a check on source control

Spatially validated predictability from sediment composition separates the metals into the same groups as the PMF apportionment of Jiang et al. (2023), which was obtained independently. Cr, Cu, Zn and Ni, assigned by PMF mainly to a lithogenic source, were predicted from Al_2O_3 and grain size alone with R^2 of 0.55–0.86 (0.76–0.86 for Cr, Cu and Zn), and 76–83% of their Shapley attribution fell on these compositional variables. Hg, assigned mainly to fossil-fuel combustion, was not predictable by any learner or predictor set and had a pure-nugget semivariogram. Diffuse atmospheric deposition combined with short-range variability, possibly including analytical noise, would produce such a pattern, but the present data cannot separate these contributions at a 4-km sampling interval. Cd and As, assigned to steel

production and agriculture, were intermediate: texture explained neither, but both were partly predictable from location. For As, the steady west-to-east decline in the Shapley contributions of longitude points to a regional gradient in supply and not to grain-size sorting.

The agreement holds at the level of groups and not of individual values. Within the lithogenic group, the PMF natural share is largest for Ni, whereas predictability is highest for Zn and Cu. The two quantities also measure different things. PMF decomposes the covariance among the eight metals into source profiles but does not indicate whether a profile is spatially predictable. The comparison in Fig. 7b is computed for each metal separately from out-of-sample predictions, requires no assumption about the number of sources, and can be obtained for any dataset that includes texture.

Texture-based predictability should not be read as proof of a natural origin. An anthropogenic metal that is delivered with, or sorbed onto, fine particles is also predictable from texture. The moderate natural shares of Cr, Cu, Zn and Ni in the PMF solution (39–48%), set against their high predictability, suggest that part of their anthropogenic load is distributed in proportion to the fine fraction. For the same reason, the texture-conditioned excess of Section 4.3 is a conservative indicator of contamination.

4.2. Validation design and the claims a model can support

ML studies of metals in soils and sediments often report high test R^2 from a single random split (e.g. Tan et al., 2019; Wang et al., 2020; Chen et al., 2022). With about 160 samples, we found that the outcome of one split spans 0.2–0.4 in R^2 for well-behaved metals and more than 1.0 when a single extreme value can fall in either the training or the test set (Table 4). An R^2 of 0.95 for Cu or Zn requires an RMSE below about one-fifth of the between-sample standard deviation; out of sample this was almost never achieved and was approached only by training-set fits. For datasets of this size we therefore suggest four reporting practices: (i) repeated or nested cross-validation with a measure of dispersion in place of a single split; (ii) spatial block cross-validation with blocks at least as large as the autocorrelation range (Roberts et al., 2017; Ploton et al., 2020); (iii) a linear reference model, without which the gain from a complex learner cannot be judged; and (iv) a statement of how influential samples were handled.

Random cross-validation was only slightly optimistic for the texture-controlled metals (difference in R^2 of about 0.04 or less), although these metals were strongly autocorrelated. Their spatial pattern is largely inherited from that of the predictors (Moran's *I* of Mz was 0.96 at 4.5 km), so the model learns a relation between composition and metal that transfers between

blocks and does not depend on location as such. The optimism was larger for As and Cd, whose spatial structure is not carried by the predictors and has to be learned from the coordinates (Hengl et al., 2018; Meyer et al., 2019). Even for Cu and Zn, however, R^2 fell by 0.11–0.12 for RF and by 0.22–0.25 for MLR once the held-out area exceeded the autocorrelation range. These figures, and not random-split scores, indicate what to expect when a model is applied to unsampled parts of Bohai Bay. Transfer to other coasts was not tested.

4.3. From linear normalisation to a texture-conditioned baseline

Geochemical normalisation is the standard means of separating natural from anthropogenic variability (Loring, 1991; Kersten and Smedes, 2002), but its linear form is poorly suited to strongly bimodal sediments such as those off Tangshan. Extrapolating the Al regression to the low-Al sands gave negative baselines and hence apparent enrichment at almost all of these stations. The RF baseline avoided this artefact because it reproduces the steep, non-linear dependence of the metals on Al_2O_3 and grain size (Fig. 8c) and cannot predict outside the range of observed concentrations. It also explained 0.16–0.32 more of the variance of Cr, Cu, Zn and Ni than the linear baseline. The steep rise between 11.5% and 12.5% Al_2O_3 probably reflects the bimodal texture of the area, in which Al_2O_3 separates sands from muds, and it should not be taken as a general geochemical threshold. Similarly, the lower Cr predicted at stations shallower than about 24 m (Section 3.5) is a conditional relation that marginal correlation does not show. Winnowing of fines in shallower, more energetic water is one possible explanation, but it cannot be confirmed without hydrodynamic or organic-carbon data.

The flagged stations form a multi-metal group in the western sector, where eight of the ten stations exceeding the texture-expected level for at least three of Cr, Cu, Zn, Ni and Pb are located. This group coincides with the north-western and south-western zones of elevated Cr, Cu, Zn, Ni and Pb described by Jiang et al. (2023). Our analysis adds that part of the elevation is not explained by finer texture, particularly at the coarser stations BC058, BC059 and BC070. The south-western Cd cluster and the isolated Ni (BC022) and Cd (BC009) maxima in coarse sediment may reflect particulate inputs that are not subject to grain-size control, but sampling or analytical artefacts at single stations cannot be excluded with the data available. These interpretations can be tested by resampling the flagged stations with duplicate analysis, by sequential extraction to separate residual from labile fractions, and by Pb isotope or metal-ratio analysis to discriminate between geogenic, steel-related and fossil-fuel sources.

4.4. Implications for monitoring

For Cr, Cu and Zn, measurements of texture and Al_2O_3, which are inexpensive compared with trace-metal analysis, predicted concentrations in held-out sub-areas with an RMSE of 4.0 (Cu), 9.4 (Zn) and 13.0 (Cr) mg kg^{-1} for RF, or 11–16% of the mean. One practical design is to measure texture and Al_2O_3 at high spatial density, analyse metals at a subset of stations, and use the texture-conditioned baseline together with the calibrated intervals to decide where further sampling is needed. Stations whose concentrations exceed the upper conformal bound or the baseline threshold would be resampled first. This approach does not replace chemical analysis of Hg, Cd and As, whose spatial variability is not tied to texture. For these metals, denser sampling or source-specific covariates are needed, such as distance to outfalls, river mouths or industrial frontage, or modelled atmospheric deposition. The width of the calibrated intervals (Table 5) also means that point predictions should not be used to assess compliance with sediment quality guidelines at individual stations.

4.5. Limitations

The study rests on a single survey of one coastal sector with 161 samples, and on the analytical quality reported by the original authors. Conclusions on transferability are limited to spatial extrapolation within the surveyed area, and validation against independent surveys of Bohai Bay is still needed. Total organic carbon, Fe and Mn oxides, sulphides and carbonate also control metal partitioning; they were not available and may account for part of the unexplained variance. The coverage guarantee of conformal prediction is marginal and assumes exchangeability between calibration and test blocks, which holds only approximately for spatial blocks. The flags in Fig. 9 are screening results: the threshold admits a chance exceedance rate of about 5% per metal, and the excess is measured relative to a texture-conditioned expectation within the surveyed population. Where contamination is itself correlated with the fine fraction, the baseline absorbs part of it, so the excess is conservative and should not be read as an absolute anthropogenic load.

5. Conclusions

The re-analysis of 161 surface sediments from the Tangshan coast supports four conclusions.

(1) Under spatial block validation, Cr, Cu and Zn were well predicted from Al_2O_3, grain size, water depth and location (R^2 = 0.83–0.92). Ni was predicted equally well once a single

anomalous sample was left out of scoring (R^2 = 0.83). Pb, As and Cd were weakly predictable, and Hg was not predictable.

(2) With the same model and data, the test R^2 of a single random 80/20 split varied by 0.2–0.4 for Cr, Cu and Zn and by more than 1.0 for Ni. Repeated, spatially blocked validation with a linear reference model is a more defensible basis for reporting skill in datasets of this size.

(3) The contrast between predictability from texture and from location, supported by exact Shapley values, separated texture-controlled metals (Cu, Zn and Cr, and less strongly Ni and Pb) from spatially structured but texture-independent metals (As and Cd) and from spatially unstructured Hg. This grouping agrees with the published PMF apportionment, although texture-based predictability does not by itself establish a natural origin.

(4) Conformal calibration brought the coverage of nominal 90% intervals to 0.87–0.93 under spatial validation. A texture-conditioned baseline outperformed linear Al normalisation (R^2 of 0.73–0.85 against 0.52–0.58 for Cr, Cu and Zn), avoided false enrichment in low-Al sands, and identified a western multi-metal group and south-western Cd flags as priorities for resampling and source tracing.

CRediT authorship contribution statement

Ibrahim Tabash: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Mohammed Alzyod:** Software, Validation, Writing – review & editing. **Abdulazeez Abdulraheem:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The sediment data analysed in this study are openly available from the Dryad Digital Repository at https://doi.org/10.5061/dryad.mkkwh7150 (Jiang, 2023). The Python code used for the analyses and figures will be made available in a public repository upon acceptance.

Acknowledgements

The authors thank Jiang et al. for making their sediment dataset openly available, and acknowledge the support of King Fahd University of Petroleum & Minerals (KFUPM).

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B 44, 139–177. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chao, J., Gu, H., Liao, Q., Zuo, W., Qi, C., Liu, J., Tian, C., Lin, Z., 2025. Natural factor-based spatial prediction and source apportionment of typical heavy metals in Chinese surface soil: Application of machine learning models. Environ. Pollut. 366, 125373. https://doi.org/10.1016/j.envpol.2024.125373

Chen, Y., Liu, Z., Xu, C., Zhao, X., Pang, L., Li, K., Shi, Y., 2022. Heavy metal content prediction based on Random Forest and Sparrow Search Algorithm. J. Chemom. 36, e3445. https://doi.org/10.1002/cem.3445

Deng, L., Gao, X., Xia, B., Wang, J., Dai, Q., Fan, Y., Wang, S., Li, H., Qian, X., 2023. Improving the efficiency of machine learning in simulating sedimentary heavy metal contamination by coupling preposing feature selection methods. Chemosphere 322, 138205. https://doi.org/10.1016/j.chemosphere.2023.138205

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Hao, H., Li, P., Jiao, W., Ge, D., Hu, C., Li, J., Lv, Y., Chen, W., 2023. Ensemble learning-based applied research on heavy metals prediction in a soil–rice system. Sci. Total Environ. 898, 165456. https://doi.org/10.1016/j.scitotenv.2023.165456

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518. https://doi.org/10.7717/peerj.5518

Horowitz, A.J., Elrick, K.A., 1987. The relation of stream sediment surface area, grain size and composition to trace element chemistry. Appl. Geochem. 2, 437–451. https://doi.org/10.1016/0883-2927(87)90027-8

Jiang, W., 2023. Distribution of heavy metals in coastal sediments under the influence of multiple factors: A case study from the south coast of an industrialized harbor city (Tangshan, China) [Dataset]. Dryad. https://doi.org/10.5061/dryad.mkkwh7150

Jiang, W., Chu, H., Liu, Y., Chen, B., Feng, Y., Lyu, J., Yuan, J., Wang, L., Li, J., Hou, W., 2023. Distribution of heavy metals in coastal sediments under the influence of multiple factors: A case study from the south coast of an industrialized harbor city (Tangshan, China). Sci. Total Environ. 889, 164208. https://doi.org/10.1016/j.scitotenv.2023.164208

Kersten, M., Smedes, F., 2002. Normalization procedures for sediment contaminants in spatial and temporal trend monitoring. J. Environ. Monit. 4, 109–115. https://doi.org/10.1039/b108102k

Lee, D.H., Lee, S.I., Kang, J.H., 2024. Machine learning approaches to identify spatial factors and their influential distances for heavy metal contamination in downstream sediment. Sci. Total Environ. 948, 174755. https://doi.org/10.1016/j.scitotenv.2024.174755

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L., 2018. Distribution-free predictive inference for regression. J. Am. Stat. Assoc. 113, 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Loring, D.H., 1991. Normalization of heavy-metal data from estuarine and coastal sediments. ICES J. Mar. Sci. 48, 101–115. https://doi.org/10.1093/icesjms/48.1.101

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30, 4765–4774.

Ma, W., Tan, K., Du, P., 2016. Predicting soil heavy metal based on Random Forest model. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4331–4334. https://doi.org/10.1109/IGARSS.2016.7730129

Matschullat, J., Ottenstein, R., Reimann, C., 2000. Geochemical background – can we calculate it? Environ. Geol. 39, 990–1000. https://doi.org/10.1007/s002549900084

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to

563 spatial prediction. Ecol. Model. 411, 108815.
564 https://doi.org/10.1016/j.ecolmodel.2019.108815

565 Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.
566 https://doi.org/10.2307/2332142

567 Nie, S., Chen, H., Sun, X., An, Y., 2024. Spatial distribution prediction of soil heavy metals
568 based on Random Forest model. Sustainability 16, 4358.
569 https://doi.org/10.3390/su16114358

570 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
571 Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,
572 Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in
573 Python. J. Mach. Learn. Res. 12, 2825–2830.

574 Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C.,
575 Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R.,
576 2020. Spatial validation reveals poor predictive performance of large-scale ecological
577 mapping models. Nat. Commun. 11, 4540. https://doi.org/10.1038/s41467-020-18321-y

578 Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT
579 Press, Cambridge, MA.

580 Reimann, C., Filzmoser, P., Garrett, R.G., 2005. Background and threshold: critical
581 comparison of methods of determination. Sci. Total Environ. 346, 1–16.
582 https://doi.org/10.1016/j.scitotenv.2004.11.023

583 Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S.,
584 Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F.,
585 Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial,
586 hierarchical, or phylogenetic structure. Ecography 40, 913–929.
587 https://doi.org/10.1111/ecog.02881

588 Romano, Y., Patterson, E., Candès, E., 2019. Conformalized quantile regression. Adv. Neural
589 Inf. Process. Syst. 32, 3543–3553.

590 Rousseeuw, P.J., Croux, C., 1993. Alternatives to the median absolute deviation. J. Am. Stat.
591 Assoc. 88, 1273–1283. https://doi.org/10.1080/01621459.1993.10476408

592 Sharifuzzaman, S.M., Rahman, H., Ashekuzzaman, S.M., Islam, M.M., Chowdhury, S.R.,
593 Hossain, M.S., 2016. Heavy metals accumulation in coastal sediments. In: Hasegawa,

H., Rahman, I.M.M., Rahman, M.A. (Eds.), Environmental Remediation Technologies for Metal-Contaminated Soils. Springer, Tokyo, pp. 21–42. https://doi.org/10.1007/978-4-431-55759-3_2

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25. https://doi.org/10.1186/1471-2105-8-25

Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 647–665. https://doi.org/10.1007/s10115-013-0679-x

Tan, K., Ma, W., Wu, F., Du, Q., 2019. Random forest–based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data. Environ. Monit. Assess. 191, 446. https://doi.org/10.1007/s10661-019-7510-4

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2019. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol. Evol. 10, 225–232. https://doi.org/10.1111/2041-210X.13107

Wang, H., Yilihamu, Q., Yuan, M., Bai, H., Xu, H., Wu, J., 2020. Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: A comparison of regression and random forest. Ecol. Indic. 119, 106801. https://doi.org/10.1016/j.ecolind.2020.106801

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Yaseen, Z.M., Melini Wan Mohtar, W.H., Homod, R.Z., Alawi, O.A., Abba, S.I., Oudah, A.Y., Togun, H., Goliatt, L., Ul Hassan Kazmi, S.S., Tao, H., 2024. Heavy metals prediction in coastal marine sediments using hybridized machine learning models with metaheuristic optimization algorithm. Chemosphere 352, 141329. https://doi.org/10.1016/j.chemosphere.2024.141329

Supplementary material

Table S1. Random forest performance for all eight metals and the best spatially validated learner.

	Random forest					Best learner (spatial CV)	
Metal	**R^2_{rd}**	**R^2_{sp} (mean ± SD)**	**$RMSE_{sp}$ (mg kg^{-1})**	**MAE_{sp} (mg kg^{-1})**	**$nRMSE_{sp}$ (%)**	**Model**	**R^2_{sp}**
Cr	0.832	0.833 ± 0.005	13.04	10.03	15.5	RF	0.833
Cu	0.901	0.891 ± 0.009	3.97	2.91	12.8	GBRT	0.897
Zn	0.909	0.905 ± 0.006	9.43	7.12	11.2	GPR	0.915
Ni	0.573	0.559 ± 0.015	11.40	5.02	30.1	RF	0.559
Pb	0.619	0.600 ± 0.006	2.96	2.30	10.8	ERT	0.622
As	0.399	0.379 ± 0.036	3.14	2.24	22.7	RF	0.379
Cd	0.373	0.209 ± 0.023	0.081	0.052	35.9	PLSR	0.324
Hg	−0.059	−0.111 ± 0.021	0.052	0.044	40.5	MLR	0.020

R^2_{rd}, repeated random five-fold cross-validation; R^2_{sp}, spatial leave-block-out cross-validation (eight blocks, three repetitions); nRMSE, RMSE as a percentage of the mean concentration.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- tangshanhmmlcode.zip
- GISProjectWork.mxd