International Journal of
Current Trends in Research
https://www.ioro.org/ijctr

Validation-Aware AI for Soil Contamination Assessment and Remediation

Rajesh Kumar Mishra[1], Divyansh Mishra[2], Rekha Agarwal[3]

[1]*ICFRE–Tropical Forest Research Institute*

(Ministry of Environment, Forests & Climate Change, Govt. of India)

P.O. RFRC, Mandla Road, Jabalpur, MP-482021, India

[2]*XLRI – Xavier School of Management, Jamshedpur, Jharkhand 831001, India*

[3] *Department of Physics, Government Science College, Jabalpur, MP, India–482 0012*

E-mail: rajeshkmishra20@gmail.com, rajesh.mishra0701@gov.in, divyanshspps@gmail.com,rekhasciencecollege@gmail.com

Abstract

Machine learning models for mapping heavy metal concentrations in soil are increasingly popular for guiding decisions about which areas to remediate. However, the reported accuracies of these models are typically based upon the same set of clustered samples that were used to create the model. In this study, we use a controlled simulation to assess the reliability of reported accuracies and their implications for remediation decisions. Twenty synthetic lead-contaminated sites (each 100 x 100 grid) were created with known concentrations in each cell. Each site was then sampled using preferential clustered sampling methods to obtain 300 laboratory samples. Eight different model configurations were evaluated using both random and spatial-block cross-validation against a complete census of all unsampled cells. Random cross-validation overestimated the census R2 for 19 out of 20 sites (mean 0.86 vs. 0.74), while spatial-block cross-validation was pessimistic (0.66). Prediction intervals calibrated using random residuals provided 82% coverage at a nominal 90%, and 64% outside of the area of applicability; spatial-block calibration provided 94% and 84%. By treating each prediction as a probability and applying a cost-ratio rule, we were able to reduce missed exceedances from 31% to 5% at a miss-to-false-alarm cost ratio of 10. Additional space-filling sampling resulted in the lowest map error, while preferential and uncertainty-driven sampling resulted in the highest recall of contaminated cells. Models developed solely from spectral proxies did not transfer well to new sites. These results suggest that reporting accuracy under multiple validation schemes, utilising uncertainty-aware decision rules, and clearly defining the limits of applicability will improve decision-making. All findings are simulation-based and will need to be verified with field data.

Keywords: *Spatial cross-validation; Conformal prediction; Area of applicability; Heavy-metal mapping; Risk-based decision-making*

1. Introduction

Soils that have high levels of lead, cadmium, arsenic, and other potentially harmful substances are one of the longest-lasting effects of mining, smelting, industrial processes, and waste disposal [1]. The first step in managing these types of lands is creating a map, a representation of which areas have concentrations above a certain screening level and therefore require some type of action. Since laboratory testing is costly, the majority of maps are created using a small number of samples along with additional information such as reflectance spectra, terrain characteristics, and distance to known sources. Currently, machine learning (ML) models are the primary method used to supplement the sample data, following the growth of digital soil

© 2026 The Author(s). Published by IORO Publications. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0) (https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided the original author and source are credited, a link to the license is provided, and any changes are indicated.

mapping [2]–[4] and soil spectroscopy [5]. The author of this study previously reviewed the use of AI in assessing and remediating contaminated soils [6] and the broader role of digital soil mapping and artificial intelligence in soil science [7].

Additionally, a separate simulation study based on Streeter-Phelps kinetics [8] was conducted to compare ML models on data generated from a known process [9]. This type of design is beneficial for the same reason here: when the data-generating process is known, the true state of each location is known and claims about map accuracy can be evaluated against it rather than against the sample that was used to develop the model. In this study, we apply this logic to contaminated soil and to a question that directly relates to remediation practice: how much confidence should a decision-maker have in the accuracy reported with an ML contamination map? This is not an idle question. At contaminated sites, samples are rarely taken randomly; instead, they tend to cluster around suspected sources and accessible areas, which is a form of preferential sampling that can negatively impact both model fitting and validation [10]. Random k-fold cross-validation is known to provide overly optimistic accuracy estimates in spatially structured data when samples are collected preferentially [11–13]. Spatial cross-validation has been suggested as a solution to this problem; however, it has also been criticised as being a biased estimator of map accuracy in its own right, with design-based validation on probability samples being offered as the only valid alternative [14].

Therefore, practitioners are left with two imperfect internal estimates and typically no independent reference. A simulation can provide that reference. This study will address five research questions. RQ1: How far do random and spatial-block cross-validation deviate from true map accuracy, for which models, and how does this vary with sample size? RQ2: Do prediction intervals calibrated with each scheme achieve their nominal coverage, and does an area-of-applicability (AOA) mask [15] identify where they fail? RQ3: Does converting predictive uncertainty into a risk-based decision rule reduce the cost of missing contaminated ground compared to thresholding the mapped mean? RQ4: Which strategy for collecting additional laboratory samples most improves the map and the detection of exceedances? RQ5: Can spectral-proxy models be transferred between sites? The contributions of this study are fourfold: (i) a reproducible multi-site simulator of lead contamination with site-specific proxy relationships; (ii) a quantitative comparison of three accuracy estimators against a census of unsampled cells; (iii) an evaluation of split-conformal intervals and AOA masks, and of Bayes-optimal decisions derived from them, under clustered sampling; and (iv) a comparison of five sampling strategies for follow-up investigation. All code and result tables are provided as supplementary material.

2. Literature Review

2.1. *Machine learning for soil contaminant mapping*

Digital soil mapping is an approach that uses environmental variables (covariates) to predict measured soil characteristics at each location within a study area [2]. Due to their ability to model non-linear relationships and use many different types of covariate data without requiring extensive parameter tuning, machine learning-based approaches such as Random Forests [16] and Gradient Boosting [17] have become the most common methods used in digital soil mapping. The advantages and disadvantages of using these types of models were reviewed by Wadoux et al. [3]. Random Forests have also been suggested as a general framework for spatial prediction where the distance from each sample point to all other sample points are included as additional covariates [4]. Spectroscopic proxies are particularly useful for predicting contamination levels in soils due to their low cost and high density.

In addition, there are currently large spectral libraries available [5]. However, spectroscopic proxies do not directly measure the amount of a specific trace metal present in the soil; rather, they measure the presence of iron oxides, clay, and organic matter, which are related to the concentration of the trace metal. Therefore, the relationship between spectroscopic proxies and trace metals will vary among different sites. Classical geostatistical methods such as inverse-distance and kriging remain the standard against which all other mapping techniques should be evaluated [18].

2.2. *Validating spatial predictions*

Roberts et al. demonstrated that random cross-validation results in overly optimistic estimates of error when data is structured spatially, temporally or hierarchically; therefore, they suggested using blocked cross-validation [11]. Ploton et al. demonstrated that large-scale ecological maps which were reported as having high levels of accuracy did not perform well when validated spatially [12]. Meyer & Pebesma indicated that many of the global ML maps are evaluated in a manner that does not represent how they will be used [13]. Wadoux et al. argued that spatial cross-validation can be biased (typically pessimistic) since the held-out block(s) are further from the training data than the prediction location(s) of a map; thus, probability-sample validation should be considered the standard [14]. Table 1 provides an overview of how this literature, along with research on uncertainty and sampling, relates to the current study.

Table 1. Positioning of the present study against closely related work.

Source	Focus	Relevance to this study
Roberts et al. [11]	Cross-validation for structured data	Motivates spatial-block validation; tested here against a census
Ploton et al. [12]	Spatial validation of large-scale ecological maps	Evidence that random CV can mislead; extended here to contaminated soil
Wadoux et al. [14]	Bias of spatial CV as a map-accuracy estimator	Counterpoint; both CV types compared with a census
Meyer and Pebesma [15]	Dissimilarity index and area of applicability	AOA used to partition coverage and error
Lei et al. [19]	Split-conformal regression intervals	Interval construction; calibration data varied
Wadoux et al. [20]	Sampling design for random-forest mapping	Context for the comparison of follow-up strategies

2.3. *Uncertainty quantification and applicability*

Conformal prediction generates intervals that have finite sample marginal coverage guarantees when there is exchangeability [21], [22]; the split version used in this study calibrates absolute residuals from the held-out data [19]. Exchangeability is violated by both clustered sampling and spatial dependence between the calibration and mapping locations; therefore, the guarantee cannot be assumed for maps. Meyer and Pebesma's dissimilarity index and AOA can be used to identify prediction locations where the combination of covariates lies outside of those represented in the training data [15].

2.4. *Sampling design for follow-up investigation*

Design-based and model-based sampling are used for different types of inference [23]. Spatial coverage sampling and conditioned Latin hypercube sampling in covariate space are two common approaches to sampling for model-based mapping [24], [25]; and sampling designs have been developed specifically for random forest mapping [20]. Active learning is a method of selecting additional data points based on the areas of greatest uncertainty in the current model [26]. Closed-loop, adaptive experimentation has been discussed in a larger context as well, such as in autonomous laboratories [27], of which the sampling loop in Section 3.8 is a simple field analogue.

However, whether these strategies will be effective in a contaminated-site setting where the goal is to find a rare exceedance rather than achieve average accuracy has received less attention, and is evaluated in Section 4.4.

2.5. *From maps to remediation decisions*

The remediation decision is asymmetric; leaving contaminated land un-remediated will generally cost more than remediating non-contaminated land. Using a predictive distribution, the Bayes-optimal decision is to treat a location if the probability of being above the screening value is greater than a threshold determined by the cost ratio (a well-known result from Bayesian decision theory). Thresholding the mapped mean does not take into account either the asymmetry or the uncertainty. The practical question addressed in Section 4.5 is whether the quality of the uncertainty estimate used to make the decision is important.

3. Methodology

3.1. *Experimental design*

Six experiments (E1–E6) were run on synthetic sites. Each site is a 100 × 100 grid of cells (10 m cells, 1 km^2 in total) on which the true natural-log lead concentration is known everywhere. A preferential, clustered sample of laboratory measurements is drawn, models are trained on it, and predictions are compared with the truth in all unsampled cells (the "census"). The census provides a population-level reference that no field study can offer, which is the reason for using simulation. Unless stated, results are means ± standard deviations across independent sites. Fig. 1 shows one representative site.

3.2. *Synthetic site generator*

The log concentration at each cell is the sum of a background concentration (30 mg/kg), the concentrations of three point sources, a wetness-related accumulation term and a spatially correlated residual. Each source produces an exponentially decaying plume with a source-specific strength and decay length, stretched along a random "wind" direction to produce anisotropy (Table 2). Seven spectral-type covariates were derived from latent soil properties: an iron-oxide index, a clay index, an organic-matter index, a vegetation index, a moisture index and two non-linear interaction terms. The iron-oxide, clay and organic-matter indices respond to contamination and lithology, with many of the coefficients drawn anew for each site, so that the proxy-metal relationship varies among sites, as it does in practice. Vegetation cover attenuates the spectral signal, and independent noise is added. Four terrain-type covariates complete the full feature set: elevation, a topographic wetness index, slope and distance to the nearest source, the last representing a known point of release. A site was generated for each random seed; across the 20 sites used in E1, E3 and E5, the mean share of cells above 120 mg/kg was 9.4% (range 4.2-19.0%) and the maximum cell concentration averaged 1,048 mg/kg (range 452-2,192 mg/kg).

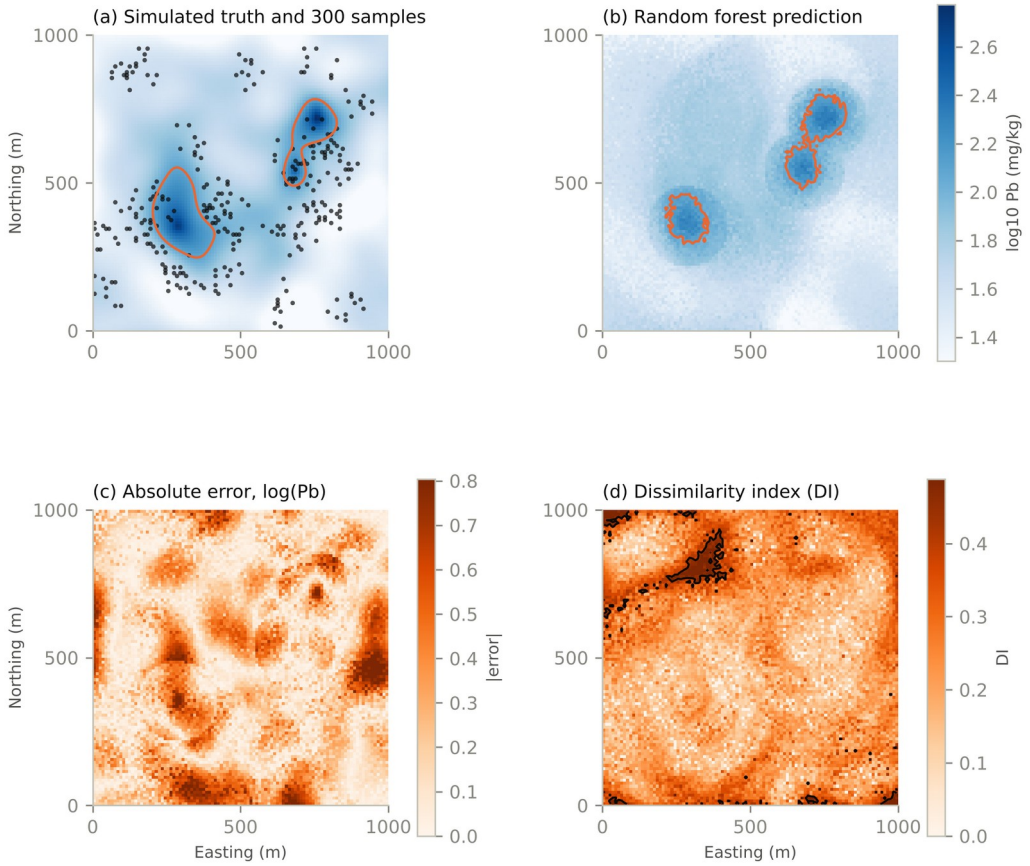


Fig. 1. Representative simulated site. (a) True log10 Pb with the 300 preferential samples (dots) and the 120 mg/kg contour (orange). (b) Random-forest prediction from the full covariate set, with the same contour. (c) Absolute prediction error on the natural-log scale. (d) Dissimilarity index; dark cells to the upper left lie beyond the area-of-applicability limit. Panels (a) and (b) share one colour scale.

Table 2. Settings of the synthetic site generator and experiments.

Item	Setting
Grid	100 × 100 cells of 10 m (1 km^2)
Background Pb	30 mg/kg
Sources	3 per site; log-strength increment U(2.3, 3.1); decay length U(6, 10) cells; plume stretch 1.8 along a random direction
Other log-Pb terms	Wetness accumulation, weight 0.25 (standardised); correlated residual, smoothing scale 6 cells, weight 0.30
Exceedance threshold	120 mg/kg (illustrative, not regulatory)
Covariates	7 spectral-type (iron-oxide, clay, organic-matter, vegetation, moisture, 2 interactions); 4 terrain-type (elevation, wetness, slope, distance to source)
Proxy coefficients	Redrawn for each site (iron-oxide on contamination U(0.2, 1.0); clay on lithology U(0.3, 0.9); others as coded)
Sampling	Clustered preferential; clusters of up to 10, spread 3.5 cells; laboratory error SD 0.08 (log)
Sample sizes	300 (E1, E3, E5, E6); 60–800 (E2); 60 to 300 by 20 (E4)
Replicates	20 sites (E1, E3, E5); 10 sites (E2, E4); 20 site pairs (E6)

The 120 mg/kg exceedance threshold is used as an example of a screening value that will provide a reasonable number of exceeding cells; however, this is not a regulatory limit. Since the simulator was developed by the researcher, the way the simulator works (smooth plumes, stationary residuals, and linear plus interaction proxies) may have influenced how well the models performed to an unknown degree.

Therefore, the results presented in this study should be interpreted as characteristics of the methods used given the stated conditions and should not be applied to any specific real-world location.

3.3. *Sampling and laboratory measurement*

Preferential sampling was imitated by drawing cluster centres with probability proportional to 0.15 + exp(−d/20), where d is the distance in cells to the nearest source, and placing up to ten samples around each centre with a Gaussian spread of 3.5 cells. This produces dense clusters in suspected hot spots and sparse coverage elsewhere, as in Fig. 1a. Laboratory values were the true log concentration plus Gaussian error with a standard deviation of 0.08 (log units).

3.4. *Models*

Eight models were evaluated (Table 3), including: Inverse Distance Weighting on Coordinates (using k=8 for Distance-Weighted Nearest Neighbors); Partial Least Squares Regression (using 4 components) on Spectral Covariates and on All Features; Support Vector Regression with Radial Kernel; Random Forest (using 200 trees, 3 minimum leaf size, and 50% of features at each split) on Spectral Features and on All Features [16]; Gradient Boosting (using 200 trees, 3 depth, 0.05 learning rate, and 0.8 subsample) [17]; and Two-Hidden-Layer Neural Network (using 64 and 32 units, early stopping). All models were built using Scikit-Learn 1.9.1 [28] and used the hyperparameters listed above; however, these hyperparameters were set prior to running the models and were not optimised for individual sites, so the results represent reasonable default performance rather than the optimal performance of each model.

3.5. *Accuracy estimators*

Three estimates of coefficient of determination (R^2) and root-mean-square error (RMSE) on the log scale were compared. Random 10-fold cross-validation ignores location. Spatial-block 5-fold cross-validation partitions the site into twenty-five 20 × 20 cell blocks that are assigned at random to five folds. The census R^2 is computed on all unsampled cells and is the reference. Because the census weights every unsampled cell equally, it corresponds to a design-based target population of the whole site, which is the quantity a map user cares about [14]. "Gap" denotes a cross-validation estimate minus the census value.

3.6. *Prediction intervals and area of applicability*

Using all available variables, we calculated split-conformal intervals [19], [21] in the format of mean +/- q; q is the ceil((n+1)(1-alpha))/n empirical quantile of absolute out-of-fold residuals at alpha = 0.10 (nominal coverage 90%). We evaluated two types of calibration: residuals from random 10-fold and from spatial-block cross-validation. The Area of Applicability (AOA) was determined as per Meyer and Pebesma [15]: we standardised and weighted each variable based upon random-forest importance; we then calculated the dissimilarity between each training point and its nearest training point outside its own spatial fold relative to the average pairwise training distance; finally, we established a threshold as the largest training dissimilarity that did not exceed the third quartile plus 1.5 interquartile ranges. All cells with a dissimilarity greater than this threshold are considered outside the AOA. We reported both coverage and RMSE inside and outside the AOA.

3.7. *Risk-based decisions*

Let L be the cost of a false alarm (treating clean ground) and rL the cost of a miss (leaving exceeding ground untreated). With a Gaussian predictive distribution centred on the model prediction and with standard deviation σ taken as the root-mean-square out-of-fold residual, the Bayes rule treats a cell when the probability of exceeding 120 mg/kg is greater than $p^* = 1/(1 + r)$. Four policies were compared for $r \in \{1, 2, 5,$

10, 20}: thresholding the predicted mean; the risk-based rule with σ from random-CV residuals; the same rule with σ from spatial-block residuals; and a conservative rule that treats a cell when the mean plus 1.645σ exceeds the threshold. The cost per exceeding cell is (false alarms + r × misses) divided by the number of truly exceeding unsampled cells. Paired differences between policies across the 20 sites were tested with the Wilcoxon signed-rank test [29].

3.8. *Additional sampling strategies*

Starting from 60 preferential samples, 20 further samples were added in each round up to 300, and a random forest was refitted each round. Five strategies were compared on 10 sites: continued preferential clustered sampling (the status quo); simple random sampling; maximin space-filling sampling, which repeatedly selects the cell farthest from all existing samples [25]; feature-space k-means, which selects the cell nearest each cluster centre in standardised covariate space, in the spirit of covariate-space designs [24]; and uncertainty-driven active learning [26], which selects cells with the greatest variance across trees subject to a minimum separation of 4 cells. Map RMSE and recall of exceedance cells (share of truly exceeding unsampled cells predicted to exceed) were recorded.

3.9. *Cross-site transfer*

For 20 pairs of independently generated sites, models were trained on 300 samples from site A and applied to all cells of site B. Random forest on full features, random forest on spectral features only, and PLSR on spectral features only were evaluated with spatial-block cross-validation at site A, census R^2 at site A and census R^2 at site B. Because proxy relationships were redrawn for each site, the experiment measures sensitivity to proxy shift, not general transferability.

3.10. *Learning-curve experiment and reproducibility*

E2 repeated the three accuracy estimates for random forest with 60, 100, 200, 400 and 800 samples on 10 sites. Every site, sample and fold was generated from a fixed seed, and the complete code and result tables are provided as supplementary files (see Data Availability).

4. Results

4.1. *Validation optimism and pessimism (E1)*

Table 3 and Fig. 2 report the three accuracy estimates for each model. Gradient boosting on the full features had the highest census R^2 (0.76 ± 0.05), followed closely by random forest (0.74 ± 0.05), support vector regression (0.71 ± 0.08) and inverse-distance weighting on coordinates alone (0.70 ± 0.10). Models restricted to spectral covariates were much weaker (census R^2 0.28 for random forest and 0.18 for PLSR).

Table 3. Accuracy of eight model configurations under three estimators (R^2 and RMSE on log (Pb); mean ± SD over 20 sites). Gap = CV estimate − census R^2.

Model (features)	R^2 random CV	R^2 spatial-block CV	R^2 census	RMSE census (log)	Gap (random - census)	Gap (spatial - census)
GBM (full)	0.88 ± 0.04	0.68 ± 0.10	0.76 ± 0.05	0.32 ± 0.04	+0.12	-0.08
RF (full)	0.86 ± 0.05	0.66 ± 0.10	0.74 ± 0.05	0.34 ± 0.04	+0.12	-0.08
SVR (full)	0.85 ± 0.05	0.64 ± 0.12	0.71 ± 0.08	0.35 ± 0.04	+0.14	-0.08
IDW (coordinates)	0.93 ± 0.02	0.50 ± 0.19	0.70 ± 0.10	0.36 ± 0.04	+0.24	-0.19
ANN (full)	0.70 ± 0.12	0.09 ± 0.50	0.55 ± 0.12	0.45 ± 0.06	+0.15	-0.46
PLSR (full)	0.71 ± 0.08	0.54 ± 0.16	0.52 ± 0.14	0.47 ± 0.09	+0.20	+0.02
RF (spectral)	0.51 ± 0.18	0.29 ± 0.25	0.28 ± 0.18	0.57 ± 0.11	+0.23	+0.01
PLSR (spectral)	0.43 ± 0.20	0.26 ± 0.28	0.18 ± 0.23	0.61 ± 0.13	+0.26	+0.08

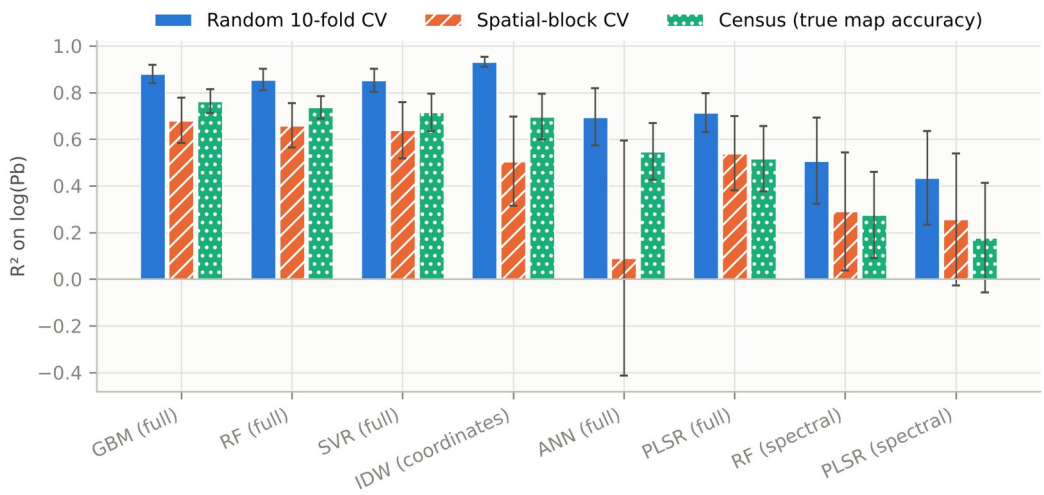


Fig. 2. Random 10-fold cross-validation, spatial-block cross-validation and census R^2 on log (Pb) for eight model configurations, 20 sites each (bars: mean; whiskers: ±1 SD).

Random cross-validation overstated census accuracy for every model by between 0.12 and 0.26 R^2 units. For random forest, it exceeded the census value at 19 of 20 sites (95%), with a mean gap of +0.12 (0.86 against 0.74). The inverse-distance interpolator showed the largest optimism, with a random-CV R^2 of 0.93 but a census R^2 of 0.70 (gap +0.24): clustered samples sit close to one another, so leaving one out costs little, whereas the map must also be right far from any sample. Spatial-block cross-validation moved in the opposite direction. For random forest it understated census accuracy (0.66 against 0.74, gap −0.08) at 15 of 20 sites (75%), and for the neural network it was unstable (0.09 ± 0.50), with census R^2 of 0.55 ± 0.12. Spatial-block estimates matched the census only for the smoother or low-signal models, PLSR on full features (gap +0.02) and random forest on spectral features (+0.01). These results agree in direction with the argument that spatial cross-validation can be pessimistic for map accuracy [14], while showing that random cross-validation is the more misleading of the two under clustered sampling in this setting.

4.2. *Effect of sample size (E2)*

The optimism of random cross-validation was largest when data were scarcest (Fig. 3, Table 4). With 60 samples, the random-CV R^2 of random forest was 0.72, whereas the census R^2 was 0.19 ± 0.42; with 100 samples, the corresponding values were 0.74 and 0.51. By 800 samples, the gap had narrowed to 0.06 (0.88 against 0.82). Spatial-block R^2 was near the census value at 200 samples (0.65 against 0.67) but plateaued at about 0.68 thereafter, while the census R^2 kept rising, which is the pessimism noted above. Large site-to-site variability at small n (SD 0.42 for census R^2 at n = 60) means that a single investigation of this size cannot reliably report its own accuracy under either scheme.

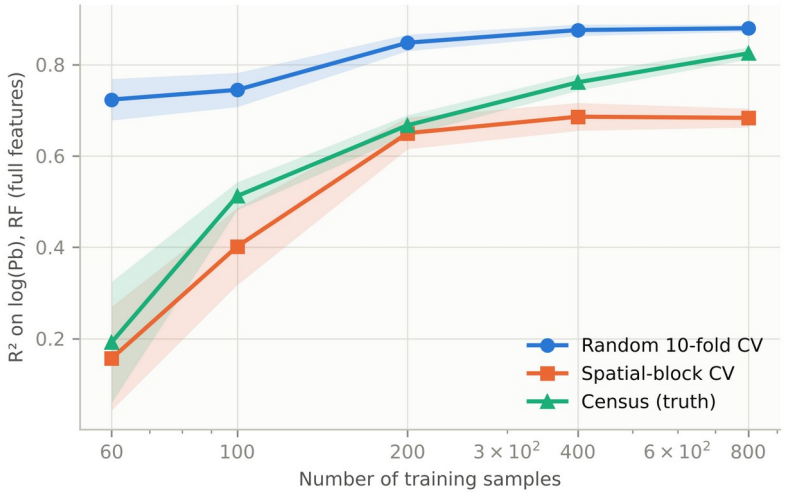


Fig. 3. Random forest R^2 on log (Pb) as a function of the number of preferential training samples under random 10-fold CV, spatial-block CV and census (10 sites; shading ±1 SD). The horizontal axis is logarithmic.

Table 4. Random forest R^2 under random CV, spatial-block CV and census by training-sample size (mean ± SD over 10 sites).

n	Random CV	Spatial-block CV	Census
60	0.72 ± 0.14	0.16 ± 0.36	0.19 ± 0.42
100	0.74 ± 0.12	0.40 ± 0.27	0.51 ± 0.10
200	0.85 ± 0.06	0.65 ± 0.11	0.67 ± 0.07
400	0.88 ± 0.04	0.69 ± 0.10	0.76 ± 0.06
800	0.88 ± 0.03	0.68 ± 0.07	0.82 ± 0.04

4.3. *Prediction-interval coverage and applicability (E3)*

Nominal 90% intervals calibrated on random cross-validation residuals covered only 82.0% ± 5.7% of unsampled cells (Table 5, Fig. 4a). The half-width q was 0.449 ± 0.068 log units, corresponding to a multiplicative factor of about 1.6. Calibrating on spatial-block residuals widened the half-width to 0.695 ± 0.119 (factor about 2.0) and gave 93.8% ± 3.7% coverage, slightly conservative, which mirrors the pessimism of spatial-block R^2. The two calibrations bracket the nominal level.

Table 5. Empirical coverage of nominal 90% conformal intervals (random forest, full features; mean ± SD over 20 sites).

Region	Random-CV calibration	Spatial-block calibration
Overall	0.820 ± 0.057	0.938 ± 0.037
Inside AOA	0.828 ± 0.055	0.942 ± 0.034
Outside AOA	0.638 ± 0.233	0.838 ± 0.177

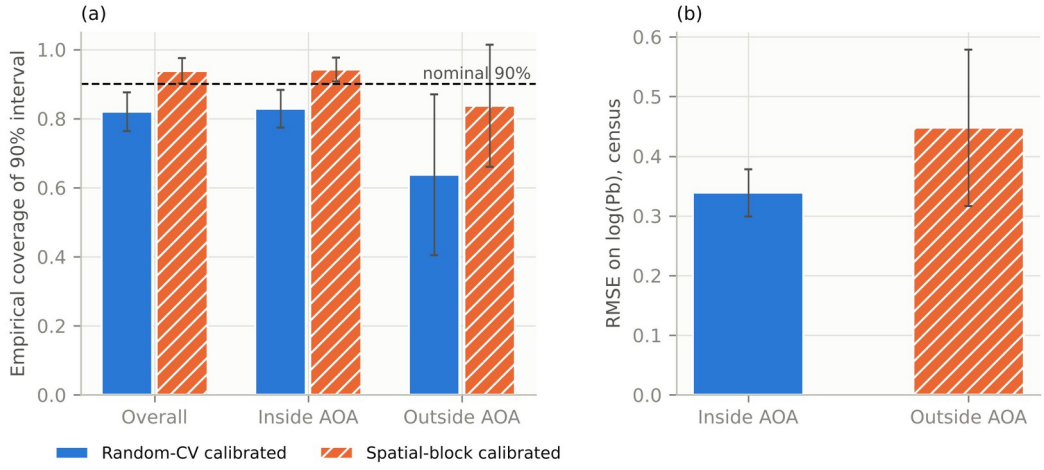


Fig. 4. (a) Empirical coverage of nominal 90% conformal intervals for random-CV and spatial-block calibration, overall and inside and outside the area of applicability (AOA). (b) Census RMSE on log (Pb) inside and outside the AOA. Bars: mean over 20 sites; whiskers: ±1 SD.

On average, 4.9% ± 3.5% of cells lay outside the AOA. Prediction error was higher there (RMSE 0.448 ± 0.131 against 0.339 ± 0.040 inside; Fig. 4b), and coverage fell to 63.8% ± 23.3% with random-CV calibration and 83.8% ± 17.7% with spatial-block calibration. The AOA therefore identified cells where the model was less accurate and where intervals were less trustworthy, but the share of such cells was small and the variation between sites large, so the mask is a flag for caution and not a guarantee. In the representative site, 2.9% of cells fell outside the AOA (Fig. 1d).

4.4. *Additional sampling strategies (E4)*

Fig. 5 and Table 6 compare the strategies. For overall map accuracy, space-filling maximin sampling was best at every sample size after the first round: at n = 140 its RMSE was 0.338 ± 0.020 against 0.414 ± 0.092 for continued preferential sampling, and at n = 300 it was 0.280 ± 0.022 against 0.323 ± 0.031. Random sampling was second (0.295 at n = 300). Continued preferential sampling gave the highest RMSE and the largest between-site variation, since clusters add little new information about unrepresented regions.

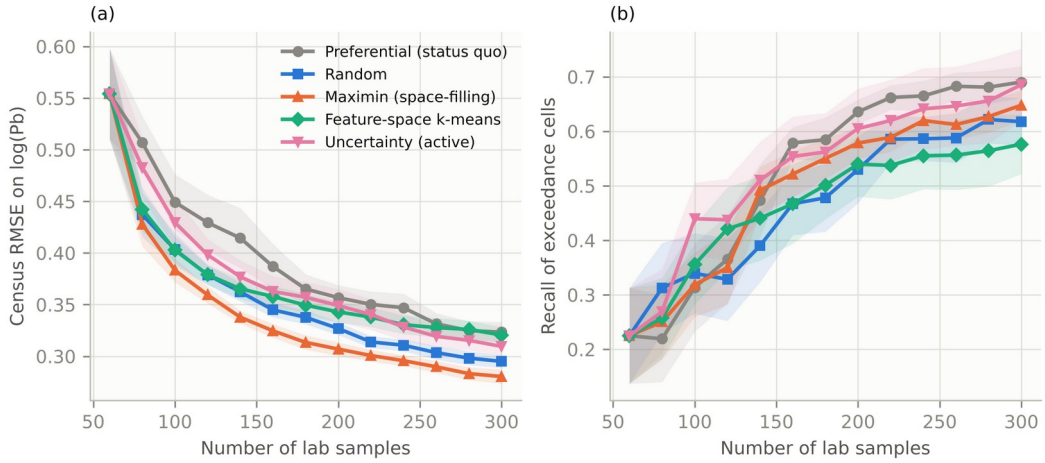


Fig. 5. Effect of five sampling strategies on (a) census RMSE on log (Pb) and (b) recall of exceedance cells, as samples are added from 60 to 300 (10 sites; shading ±1 SD).

Table 6. Map RMSE (log scale) and recall of exceedance cells for five sampling strategies (mean ± SD over 10 sites).

Strategy	RMSE n=140	Recall n=140	RMSE n=300	Recall n=300
Preferential (status quo)	0.414 ± 0.092	0.47 ± 0.21	0.323 ± 0.031	0.69 ± 0.09
Random	0.362 ± 0.029	0.39 ± 0.21	0.295 ± 0.019	0.62 ± 0.14
Maximin (space-filling)	0.338 ± 0.020	0.49 ± 0.14	0.280 ± 0.022	0.65 ± 0.09
Feature-space k-means	0.365 ± 0.031	0.44 ± 0.24	0.321 ± 0.025	0.58 ± 0.17
Uncertainty (active)	0.377 ± 0.050	0.51 ± 0.22	0.310 ± 0.027	0.69 ± 0.21

The ranking was different for finding contamination. At n = 300, recall of exceedance cells was highest for preferential sampling and uncertainty-driven sampling (both 0.69), against 0.65 for maximin, 0.62 for random and 0.58 for feature-space k-means, although the standard deviations (0.09–0.21) are large enough that most of these differences cannot be separated with ten sites. The practical reading is that preferential sampling, by concentrating on likely hot spots, retains an advantage in detecting exceedances even though it is the worst strategy for overall map accuracy, and that no strategy dominated on both criteria. Feature-space k-means, which performed comparably to random sampling on RMSE at n = 140, fell behind it later, possibly because the covariate-space designs oversample cells with rare covariate combinations that carry little information about contamination; this explanation was not tested.

4.5. *Risk-based remediation decisions (E5)*

Truly exceeding cells made up 9.2% of unsampled cells on average. Thresholding the mapped mean missed 30.7% of them regardless of cost ratio (Table 7, Fig. 6), because it ignores both the cost asymmetry and the uncertainty. Risk-based rules reduced misses as the cost ratio increased. At r = 10, the random-CV-σ rule missed 9% of exceeding cells and the spatial-block-σ rule 5%, at costs per exceeding cell of 1.58 and 1.68 against 3.17 for the mean threshold. The conservative mean-plus-1.645σ rule held misses at 3% but treated 23.6% of cells (more than twice the true exceedance share) and was costlier than the risk-based rules at every r ≤ 10, with cost 2.12 at r = 10.

Table 7. Decision cost per exceeding cell (share of exceeding cells missed in parentheses) by cost ratio r and policy (mean over 20 sites).

Cost ratio r	Map threshold (mean)	Risk-based (random-CV σ)	Risk-based (spatial-CV σ)	Conservative (mean + 1.645 σ)
1	0.41 (31% missed)	0.41 (31% missed)	0.41 (31% missed)	1.83 (3% missed)
2	0.72 (31% missed)	0.63 (21% missed)	0.64 (17% missed)	1.86 (3% missed)
5	1.64 (31% missed)	1.08 (13% missed)	1.17 (8% missed)	1.96 (3% missed)
10	3.17 (31% missed)	1.58 (9% missed)	1.68 (5% missed)	2.12 (3% missed)
20	6.24 (31% missed)	2.20 (7% missed)	2.47 (3% missed)	2.44 (3% missed)

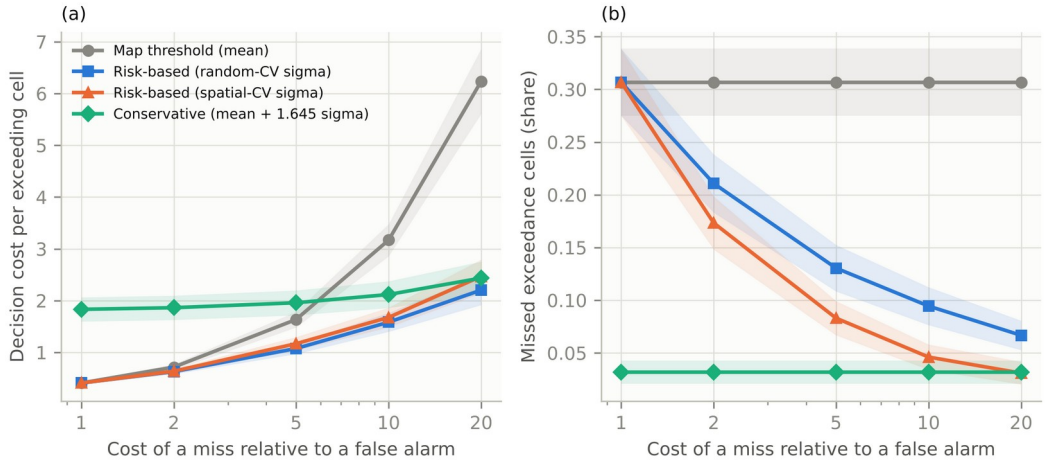


Fig. 6. (a) Decision cost per exceeding cell and (b) share of exceeding cells missed, for four decision policies as the cost of a miss relative to a false alarm increases (20 sites; shading ±1 SD).

Both risk-based rules were clearly better than the mean threshold at r = 10 (spatial-σ rule: lower cost at 19 of 20 sites, Wilcoxon $p = 1.0 \times 10^{-4}$). The choice between random-CV and spatial-block σ mattered much less for cost: the spatial-σ rule was cheaper at only 8 of 20 sites, and the paired difference was not significant (p = 0.35). Its main effect was to shift the operating point, treating a larger area (18.7% of cells against 13.7% at r = 10) to miss fewer exceedances (5% against 9%). Whether that trade is appropriate is a policy question that depends on the true cost ratio; the simulation shows only that the more honest uncertainty estimate moves the decision in the cautious direction, as it should when the map is less reliable than its internal validation suggests.

4.6. *Cross-site transfer (E6)*

Models trained on a single site transferred imperfectly (Table 8, Fig. 7). Random forest on full features lost about 0.2 R^2 units (30% of its value), from a census R^2 of 0.70 ± 0.08 at the training site to 0.49 ± 0.15 at a new site, and never fell below zero. Spectral-only models were much weaker and less stable: census R^2 at a new site was 0.09 ± 0.22 for random forest and 0.10 ± 0.24 for PLSR, and 30% and 40% of new sites, respectively, had negative R^2, meaning predictions worse than the site mean. Within the training site, spatial-block cross-

validation understated census accuracy for all three models, so internal validation gave no warning of the transfer failure. For random forest on spectral features, the correlation between the new-site R^2 and the absolute difference between the iron-oxide proxy coefficients of the two sites was −0.18, weak and in the expected direction, though with 20 pairs it is not distinguishable from zero.

Table 8. Transfer of site-A models to site B (R^2 on log(Pb); mean ± SD over 20 site pairs).

Model (features)	Spatial-block CV (training site)	Census (training site)	Census (new site)	Share of new sites with R^2 < 0
RF (full)	0.60 ± 0.13	0.70 ± 0.08	0.49 ± 0.15	0%
RF (spectral)	0.24 ± 0.22	0.35 ± 0.13	0.09 ± 0.22	30%
PLSR (spectral)	0.21 ± 0.27	0.31 ± 0.14	0.10 ± 0.24	40%

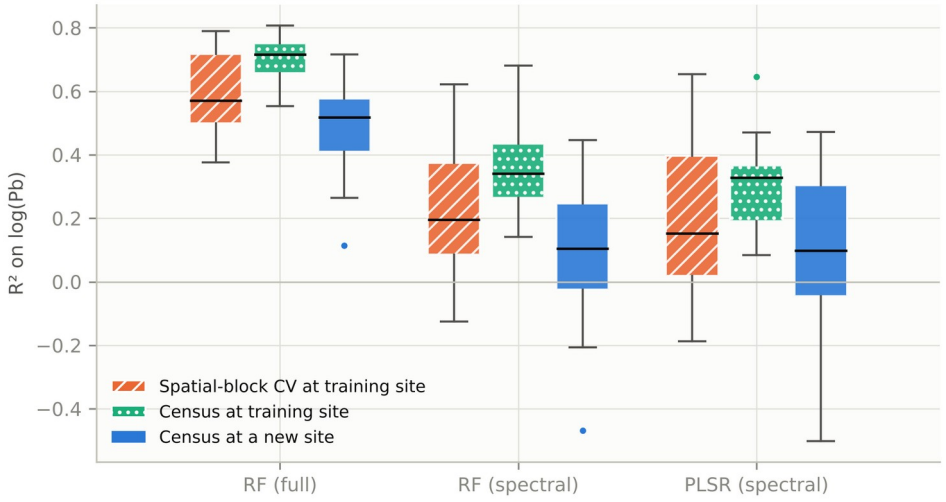


Fig. 7. R^2 on log (Pb) for three models: spatial-block CV and census at the training site, and census when the trained model is applied at a new site (20 site pairs; boxes: interquartile range; line: median; whiskers: 1.5 × IQR).

5. Discussion

The findings from this study can be summarised in three ways. First, random cross-validation consistently overestimated map accuracy across nearly all sites, and the overestimation was largest for those models and sample sizes where users are most likely to rely on a single number (i.e., interpolators and small data sets). Second, spatial-block cross-validation consistently underestimated map accuracy; therefore, neither internal estimate represents the "truth"; reporting both provides an interval within which the census value was typically found for the better-performing models (for example, for random forest, 0.66 and 0.86 bracket 0.74). Third, the implications for decision-making were greater than indicated by the differences in accuracy. A mean-threshold map identified less than one-third of exceeding cells at any cost ratio, while a simple rule based on the predictive spread reduced that miss rate by half or more at moderate cost ratios. These results are consistent with previous literature regarding optimistic random validation [11–13], and with the caution that spatial cross-validation is not an unbiased estimator of map accuracy [14]. However, these results do not resolve the debate.

In this study, the census served as a design-based reference over the entire site and cross-validation estimates were compared to it; a real investigation cannot calculate the census, which is why probability sampling for validation [23], [25] continues to be the preferred method when feasible. This simulation provides a quantitative measure of the likely direction and magnitude of errors in each internal estimate under clustered sampling. This study qualifies a common recommendation. Overall accuracy was increased most by space-filling designs, consistent with arguments for coverage sampling [25]; however, preferential

sampling continued to perform well in identifying exceedances. If a program's sole objective is to delineate contamination, then it may be reasonable to continue using some preferential effort; if a program also requires a reliable map, then it should incorporate space-filling samples. In this configuration, the uncertainty-driven strategy did not result in an improvement over random sampling for RMSE; therefore, active learning cannot be assumed to be beneficial without verification, and its minimum-separation and batch settings were not optimised in this study. Based on the assumptions made in this study, four recommendations are provided below: (i) Provide both random and spatial-block validation together and indicate which sampling design generated the data. (ii) Pass uncertainty, not just the mean, to the remediation decision and specify the cost ratio. (iii) Publish the area of applicability along with the map and consider predictions outside of that area as unsupported. (iv) Do not apply proxy-based models at a new site without local calibration samples. Each recommendation is consistent with and provides quantitative support for the cautions about AI-driven contaminated-land assessment raised in previous reviews [6], [7].

5.1. *Limitations*

The most significant limitation of this study is that all data used were generated synthetically. This simulation was conducted at a fixed location, with smooth plumes of one type of contaminant. It did not have any limitations on the sampling frame, nor did it measure the support of samples. Additionally, there were no vertical dimensions or non-point source contaminations included in the simulations. In contrast, real-world locations often have abrupt boundaries such as fill, excavation, and roads that will affect how spatially dependent the variables are, as well as how related the covariates are.

Additionally, hyperparameters were set rather than optimised; the neural network was not optimised, and both the AOA and conformal procedures were used in only one configuration. Due to the limited number of sites (10-20 per experiment), the precision of the differences reported is likely to be affected by the small sample size, especially regarding recall in E4. While conformal coverage appears to be reasonable across cells and sites, local coverage in hot spots was not evaluated and is the more relevant metric for remediation. Lastly, the Gaussian predictive distribution used in E5 is an approximation. All of these findings should be considered as hypotheses about how well methods perform based on simulations, rather than as evidence from the field. To confirm these hypotheses, they must be tested using real data, preferably with an independent probability sample.

6. Conclusion

The results of a controlled multi-site study demonstrate that the reported accuracy of ML maps of soil lead will be greatly influenced by the method used to estimate accuracy. Random cross-validation (RCV) overestimated census accuracy; spatial-block cross-validation (SBCV) underestimated census accuracy; and intervals based on random residuals were too narrow, especially when applied outside of the area of applicability. When predictive uncertainty was converted to a cost-aware decision rule, the number of missed exceedances was reduced much more than the differences between the two methods of uncertainty calibration. Space-filling follow-up sampling increased overall accuracy, but preferential sampling maintained an advantage in terms of recall. Models that relied solely on spectral proxies were not able to be effectively transferred between sites. Future research should apply these findings to real contaminated-site data using probability-sample validation, assess local coverage in "hot spots," extend the simulator to non-stationary and multi-contaminant environments, and assess decision rules with site-specific cost structures.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgments

We are very much thankful to the authors of different publications, as many new ideas are abstracted from them. Authors also express gratefulness to their colleagues and family members for their continuous help, inspirations, encouragement, and sacrifices without which this work could not be executed. Finally, the main target of this work will not be achieved unless it is used by research institutions, students, research scholars, and authors in their future works. The authors will remain ever grateful to Dr. Neelu Singh, Director, ICFRE Tropical Forest Research Institute, Jabalpur, Director, XLRI – Xavier School of Management, Jamshedpur & Principal, Government Science College, Jabalpur who helped by giving constructive suggestions for this work. The authors are also responsible for any possible errors and shortcomings, if any, in the paper, despite the best attempt to make it immaculate.

References

[1] B. J. Alloway, Ed., Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability, 3rd ed. Dordrecht, The Netherlands: Springer, 2013.

[2] A. B. McBratney, M. L. Mendonça Santos, and B. Minasny, "On digital soil mapping," Geoderma, vol. 117, no. 1–2, pp. 3–52, 2003.

[3] A. M. J.-C. Wadoux, B. Minasny, and A. B. McBratney, "Machine learning for digital soil mapping: Applications, challenges and suggested solutions," Earth-Sci. Rev., vol. 210, Art. no. 103359, 2020.

[4] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," PeerJ, vol. 6, Art. no. e5518, 2018.

[5] R. A. Viscarra Rossel et al., "A global spectral library to characterize the world's soil," Earth-Sci. Rev., vol. 155, pp. 198–230, 2016.

[6] Mishra, R. K., Agarwal R and Mishra D (2026) Integrated AI-Driven Assessment and Sustainable Remediation of Contaminated Soils: Innovations, Challenges and Future Directions. In Book: Contaminated Soils - Remediation Technologies for Sustainable Environment. Publisher: IntechOpen Limited 167-169 Great Portland Street, London, W1W 5PF, UNITED KINGDOM, IntechOpen. http://dx.doi.org/10.5772/intechopen.1015754, 43pp. DOI: 10.5772/intechopen.1015754.

[7] Mishra, R.K., Mishra, Divyansh, and Agarwal, R. (2025), Technological Innovations in Soil Science: Digital Soil Mapping, Artificial Intelligence and Precision Agriculture, In book: Soil Science: Global Challenges and Solutions, An International Edition, 2026, ISBN: 978-93-49938-77-9, 37-56. DOI: 10.5281/zenodo.19680845.

[8] H. W. Streeter and E. B. Phelps, "A study of the pollution and natural purification of the Ohio River, III: Factors concerned in the phenomena of oxidation and reassociation," U.S. Public Health Service, Washington, DC, USA, Public Health Bull. 146, 1925.

[9] Mishra, Divyansh, Mishra R.K., Agarwal, R., Machine learning-based prediction of river water quality: A Streeter–Phelps-grounded simulation framework and comparative model evaluation, International Journal of Civil and Environmental Engineering, Volume 1 | Issue 1 | 2026 | Article ID: 8277-1192-7665, 23-42. DOI: 10.64823/ijcee.2601003.

[10] P. J. Diggle, R. Menezes, and T. Su, "Geostatistical inference under preferential sampling," J. R. Stat. Soc. C, vol. 59, no. 2, pp. 191–232, 2010.

[11] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913–929, 2017.

[12] P. Ploton et al., "Spatial validation reveals poor predictive performance of large-scale ecological mapping models," Nat. Commun., vol. 11, Art. no. 4540, 2020.

[13] H. Meyer and E. Pebesma, "Machine learning-based global maps of ecological variables and the challenge of assessing them," Nat. Commun., vol. 13, Art. no. 2208, 2022.

[14] A. M. J.-C. Wadoux, G. B. M. Heuvelink, S. de Bruin, and D. J. Brus, "Spatial cross-validation is not the right way to evaluate map accuracy," Ecol. Model., vol. 457, Art. no. 109692, 2021.

[15] H. Meyer and E. Pebesma, "Predicting into unknown space? Estimating the area of applicability of spatial prediction models," Methods Ecol. Evol., vol. 12, no. 9, pp. 1620–1633, 2021.

[16] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.

[18] R. Webster and M. A. Oliver, Geostatistics for Environmental Scientists, 2nd ed. Chichester, U.K.: Wiley, 2007.

[19] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," J. Am. Stat. Assoc., vol. 113, no. 523, pp. 1094–1111, 2018.

[20] A. M. J.-C. Wadoux, D. J. Brus, and G. B. M. Heuvelink, "Sampling design optimization for soil mapping with random forest," Geoderma, vol. 355, Art. no. 113913, 2019.

[21] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY, USA: Springer, 2005.

[22] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv:2107.07511, 2021.

[23] D. J. Brus and J. J. de Gruijter, "Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil," Geoderma, vol. 80, no. 1–2, pp. 1–44, 1997.

[24] B. Minasny and A. B. McBratney, "A conditioned Latin hypercube method for sampling in the presence of ancillary information," Comput. Geosci., vol. 32, no. 9, pp. 1378–1388, 2006.

[25] D. J. Brus, "Sampling for digital soil mapping: A tutorial supported by R scripts," Geoderma, vol. 338, pp. 464–480, 2019.

[26] B. Settles, "Active learning literature survey," Univ. Wisconsin–Madison, Madison, WI, USA, Comput. Sci. Tech. Rep. 1648, 2009.

[27] 36. Mishra, R.K., Mishra, Divyansh, and Agarwal, R. (2026), Artificial intelligence for scientific discovery: From hypothesis generation to Autonomous laboratories, In book: Contemporary Innovations in Science, Engineering and Technology, Edition: First Edition, January 2026, ISBN: 978-93-47587-99-3, 1-20.

[28] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[29] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bull., vol. 1, no. 6, pp. 80–83, 1945.