An Outlier-Aware Hybrid Framework for Spatial Prediction and Risk Assessment of Agricultural Soil Heavy Metal Pollution

Maseeha Banu
Adarsh Ramangouda Malawade, Amrit Kumar Kamble, Mithun P Y, Pavan M
Department of CSE (AI & ML)
Malnad College of Engineering, Hassan, Karnataka, India
Email: mb@mcehassan.ac.in

***Abstract*—**Agricultural soils can contain heavy metals from natural materials and human activities, but their concentrations rarely follow a uniform pattern across a landscape. Direct sampling and laboratory analysis provide reliable measurements at selected locations; they do not, by themselves, describe the many unsampled places between them. Prediction is further complicated when a small number of observations differ sharply from their neighbours. Such observations may be measurement errors, but they may also mark genuine local pollution. This paper reviews spatial prediction and risk-assessment methods for agricultural soil heavy metals, with particular attention to the complementary ideas in two published hybrid studies: one separates normal and outlier-influenced areas before prediction, while the other combines semi-supervised concentration prediction, pollution classification, and geographic information system analysis. Building on these studies and related reviews, the paper presents a conceptual outlier-aware workflow that links data checks, spatial outlier investigation, region appropriate prediction, uncertainty assessment, risk interpretation, and mapping. The workflow is a proposed synthesis, not an experimentally validated model. The review argues that reliable pollution maps should preserve meaningful hotspot signals, use spatially appropriate validation, and communicate uncertainty alongside predicted concentrations and risk classes. ***Index Terms*—** Heavy metals, machine learning, risk assessment, spatial outliers, spatial prediction.

I. INTRODUCTION

Agricultural soil is more than a medium in which crops grow. It stores water and nutrients, supports soil organisms, and forms part of the environmental system that connects land use with food production. When potentially toxic elements accumulate in soil, the concern extends beyond the field itself. Depending on the element, its chemical form, soil conditions, and crop type, contamination may affect soil functions and create pathways for transfer into plants and the wider food chain [2], [3].

The elements commonly examined in agricultural soil studies include arsenic (As), cadmium (Cd), chromium (Cr), mercury (Hg), lead (Pb), zinc (Zn), copper (Cu), and nickel (Ni). Some occur naturally in parent materials, while additional inputs can come from mining, industrial activity, waste handling, irrigation water, atmospheric deposition, and agricultural materials. Their distribution is rarely uniform. A field near an industrial discharge point, for example, may show a different concentration pattern from a field with similar soil properties but no nearby source [1], [2].

A soil survey gives direct evidence at the locations where samples are collected. However, laboratory analysis is time-consuming and costly, so a regional survey usually contains a limited number of sampling points. The practical task is therefore to estimate concentrations at locations that were not sampled. This is the purpose of predictive soil-pollution mapping. Spatial interpolation and regression methods have long been used for this task, while machine-learning models now make it possible to include a wider range of environmental predictors [3], [4].

One difficulty deserves special attention: not every observation follows the dominant spatial pattern. A sample with a concentration very different from nearby samples may indicate a laboratory or recording problem. It may also identify a real hotspot associated with a local pollution source. Treating every unusual value as noise can erase the very locations that a pollution assessment should identify. Yang et al. addressed this issue by detecting spatial outliers, dividing the study area into normal and outlier regions, and applying different prediction methods to the two parts [1].

A related but distinct challenge is the shortage of labelled samples. Chen et al. addressed this by using Bayesian co-training with semi-supervised Bayesian

 ©IJENGT: All Rights are Reserved

regression to estimate heavy-metal concentrations at unsampled points. Their framework then used an improved multiple-kernel support vector machine for pollution-status evaluation and geographic information system (GIS) analysis for spatial representation [2]. These two studies offer complementary ideas: one focuses on how unusual spatial observations should be handled, and the other links concentration prediction with pollution-risk classification.

The aim of this paper is to bring these ideas together in a practical conceptual framework for agricultural soil assessment. The paper reviews the role of spatial outliers, compares common prediction approaches, examines how predicted concentrations can inform risk assessment, and proposes a workflow that carries uncertainty through to the final maps. No new field dataset is analysed here, and the proposed integrated workflow is not presented as a model whose accuracy has already been demonstrated.

II. REVIEW SCOPE AND METHOD

This paper uses a focused narrative-review approach. The two IEEE Access studies supplied as the starting point were treated as anchor papers because they directly address spatial outlier-aware soil mapping and hybrid agricultural-soil risk assessment [1], [2]. They were supplemented with peer-reviewed review and research articles on machine-learning based soil-contamination prediction, spatially informed mapping, and risk assessment [3]–[7].

The literature was examined around five questions: (1) why limited sampling makes soil-pollution mapping difficult; (2) how spatial outliers can be distinguished from errors; (3) what geostatistical and machine-learning methods contribute to prediction; (4) how predicted concentrations can be interpreted as pollution or risk categories; and (5) what is needed to make the resulting maps reliable for decision-making. Search concepts included “soil heavy-metal contamination,” “spatial prediction,” “spatial outlier detection,” “machine learning,” “geo statistics,” “risk assessment,” “GIS,” and “uncertainty.”

This is not a PRISMA systematic review or a meta analysis. It does not claim that every relevant publication has been identified, and it does not pool model-performance values across studies. The discussion distinguishes results reported by the source papers from recommendations inferred for the proposed framework. That distinction matters because accuracy values from different regions, datasets, target metals, and validation designs cannot be compared as though they came from one controlled experiment.

III. THE SPATIAL NATURE OF SOIL HEAVYMETAL CONTAMINATION

A. Sources and Environmental Context

Heavy metals in soil may reflect the mineral composition of the parent material or inputs associated with human activity. Mining and metal-processing sites, industrial emissions, wastewater, waste disposal, and some agricultural inputs are examples of possible anthropogenic sources. The importance of each source depends on local land use, geology, hydrology, soil characteristics, and the history of the site [2], [3]. For this reason, a model developed in one region should not be assumed to work equally well in another without checking the local conditions.

B. Why Concentrations Vary Across Space

Soil properties and pollution sources change over distance. Nearby points may share similar parent material or land use, but a local source can create a steep concentration gradient over a short distance. Spatial heterogeneity means that relationships between a contaminant and its explanatory variables can change from one part of a study area to another. Spatial dependence also matters: samples close to each other may be more alike than samples far apart. Both features influence the choice of prediction model and the design of model validation [1], [3].

C. A High Value Is Not Automatically a Bad Value

Outlier review should begin with data quality. Units, coordinates, sample labels, laboratory records, and detection limits should be checked before an unusual result is interpreted. If the result remains plausible, the analyst should consider whether a nearby source or a distinct local process could explain it. The 2019 hybrid mapping study was motivated by precisely this issue: the mechanism producing an outlier dataset may differ from the process represented by the majority of observations [1].

IV. METHODS USED IN SPATIAL PREDICTION AND RISK ASSESSMENT

A. Geostatistical and Interpolation Methods

Inverse distance weighting estimates an unsampled value from nearby observations, assigning greater influence to closer points. Kriging uses a model of spatial dependence, commonly represented by a variogram, to produce a spatial estimate and, under its assumptions, an estimation variance. These methods

 ©IJENGT: All Rights are Reserved

are useful baselines because their assumptions and spatial behaviour can be examined. Their limitations become important when sampling is sparse, the spatial pattern is strongly non-stationary, or a small hotspot is surrounded by much lower values [1], [3].

B. Machine-Learning Methods

Machine-learning models can represent nonlinear relationships between heavy-metal concentrations and environmental covariates. Random forests, support vector regression, gradient-boosting methods, and neural networks are among the approaches discussed in soil-contamination prediction literature [3], [4]. Their usefulness depends on having training data that represent the range of conditions in the target area. A complex algorithm cannot compensate for missing source information, poor sampling design, or inconsistent laboratory measurements.

Environmental predictors may include soil pH, organic matter, texture, elevation, vegetation indicators, land use, geology, climate, and distance to potential pollution sources. Ma et al. showed how spatial regionalization indices can be added to machine learning to improve the mapping of heavy metals in agricultural soils [5]. This supports the broader point that spatial context should be represented explicitly rather than left entirely to the model to infer from coordinates.

C. Hybrid and Semi-Supervised Approaches

A hybrid model assigns different tasks to methods that have complementary strengths. Yang et al. combined geographically weighted extreme learning machine (GWELM) for the normal region with distance-decay-function interpolation for the outlier region, after identifying spatial outliers using a fieldbased theory spatial outlier detection method (FTSOD) [1]. The method was evaluated using cadmium and arsenic measurements from Huizhou, China. The authors reported 103 sampling points and identified 16 cadmium and 13 arsenic outlier points in that dataset [1].

Chen et al. focused on the difficulty of assessing agricultural soil with limited labelled samples. Their hybrid intelligent evaluation model used Bayesian cotraining with semi-supervised Bayesian regression to predict concentrations at unsampled points. An improved multiple-kernel support vector machine was then used to assess pollution status, with an artificial fish swarm algorithm used for parameter selection and ArcGIS/Kriging used for spatial analysis [2]. The study considered eight elements—As, Cd, Cr, Hg, Pb, Zn, Cu, and Ni—in agricultural areas of Huangpi and Xinzhou, Wuhan, China. The authors used 364 sampling points in Huangpi and 155 in Xinzhou, together with additional dense sampling locations used as prediction targets. They reported an average singlefactor assessment accuracy of 96.66% and an overall evaluation accuracy of 97.42% for their study [2]. These are study-specific assessment results, not directly comparable regression scores or guarantees of performance in other regions.

D. From Predicted Concentration to Risk Information

A concentration map and a risk map answer different questions. The first estimates where a metal may be present and at what level. The second interprets that estimate using an assessment method, exposure context, and applicable criteria. Single-factor indices evaluate individual elements, while comprehensive indices combine information across elements. Ecological and human-health assessments may require additional information about bioavailability, exposure pathways, land use, and receptor characteristics [2], [5], [7].

For agricultural land, threshold selection must match the jurisdiction, soil-use category, pH conditions, and applicable regulatory document. A threshold taken from a different country or land-use category should not be transferred without justification. This review therefore does not insert universal numerical limits. Any India-specific threshold used in a later case study must be checked against the current official standard and the relevant land-use classification.

TABLE I. COMPARISON OF THE TWO ANCHOR STUDIES

Study	Method and evidence relevant to this review
Yang et al. (2019) [1]	Huizhou soil dataset; Cd and As; 103 sampling points. FTSOD identifies spatial outliers; GWELM predicts in normal areas; distance-decay interpolation estimates values in outlier-influenced areas. Demonstrates why unusual observations may need separate treatment.
Chen et al. (2023) [2]	Agricultural areas of Huangpi and Xinzhou, Wuhan; eight metals. Bayesian co-training / Semi-BR predicts unsampled concentrations; improved MKSVM assesses pollution status; GIS and Kriging support spatial representation. Demonstrates a prediction-to-assessment workflow under limited labelled sampling.

TABLE II. METHOD FAMILIES AND THEIR MAIN CONSIDERATIONS

Method family	Useful role	Main caution
IDW / kriging	Spatial interpolation and baseline maps.	Sensitive to sampling pattern and spatial assumptions.
Random forest / boosting	Nonlinear prediction with environmental covariates.	Can overfit; performance depends on representative training data.
SVR / neural models	Flexible regression when relationships are nonlinear.	Requires tuning and careful validation.

 ©IJENGT: All Rights are Reserved

Spatial regression / GWELM	Models spatially varying relationships.	Local estimates may be unstable where samples are sparse.
Semi-supervised learning	Uses labelled and unlabelled observations together.	Unlabelled data help only when assumptions are appropriate.
Outlier-aware hybrid	Uses different strategies in normal and outlier influenced regions.	Outlier detection and region boundaries require validation.

V. PROPOSED OUTLIER-AWARE HYBRID FRAMEWORK

The framework proposed here is a synthesis of the two anchor studies rather than a direct reproduction of either one. Its central idea is to treat spatial outlier detection as a decision point in the modelling process. Unusual observations are first checked and interpreted; only then is the prediction strategy selected. The workflow has seven stages.

A. Assemble Soil and Environmental Data

The input dataset should contain sampling coordinates, measured concentrations, sampling depth and date, laboratory metadata, and relevant environmental covariates. Where available, soil pH, texture, organic matter, elevation, land use, geology, vegetation indicators, and proximity to potential sources can help explain spatial patterns [3]–[5]. Each data source should retain its unit, spatial resolution, date, and method of collection.

B. Check Data Quality Before Detecting Outliers

The initial checks should identify missing values, inconsistent units, coordinate errors, duplicate records, values below detection limits, and unusual laboratory results. A sample should not be labelled a spatial outlier simply because its concentration is high. Data quality checks help separate an avoidable measurement problem from an observation that may represent a real environmental condition.

C. Identify and Investigate Spatial Outliers

Spatial outlier detection should compare each observation with its neighbours while accounting for both concentration differences and geographic proximity. A candidate outlier should then be checked against field notes, land use, nearby facilities, topography, and other source information. If no plausible explanation can be found, the observation should remain flagged for sensitivity analysis rather than being silently deleted. This stage follows the rationale of the FTSOD-based approach described by Yang et al. [1].

D. Select Prediction Models by Region

For a normal or background-like region, a regression model using environmental covariates may be suitable. Candidate methods include random forest, support vector regression, GWELM, or regression kriging. In an outlier-influenced region, a local interpolation method or a specialised local model may better preserve a sharp hotspot. The selected method must be tested against simpler baselines; the framework does not assume that the more complicated option will always perform better.

E. Reconcile Regional Predictions

Predictions from the two regions must be combined into a continuous map. This step requires particular care near the boundary between regions. Abrupt discontinuities may be physically meaningful, but they may also be artefacts of the modelling procedure. Residual maps and cross-boundary checks should therefore be examined before the combined surface is interpreted.

F. Carry Uncertainty Into Risk Assessment

The framework should report more than a single predicted value. Where a method supports it, prediction intervals or spatial uncertainty estimates should accompany the concentration surface. Risk categories should be based on the appropriate standard and should indicate when a prediction lies close to a category boundary. Uncertainty caused by sparse samples, outlier classification, model selection, and extrapolation should be discussed explicitly.

G. Produce Decision-Oriented GIS Outputs

The final GIS product should contain at least three linked layers: predicted concentration, risk category, and prediction uncertainty. Keeping these layers separate helps users distinguish an area predicted to have high contamination from an area where the model is simply uncertain. The maps should also identify the sampled locations so that measured observations can be distinguished from modelled estimates.

 ©IJENGT: All Rights are Reserved

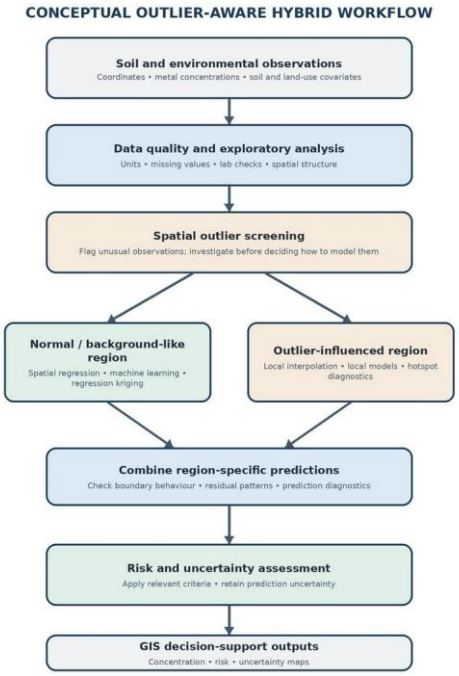


Fig. 1. Conceptual outlier-aware workflow proposed in this review. The integrated workflow requires empirical validation.

VI. EVALUATION PLAN FOR FUTURE IMPLEMENTATION

A future implementation should be evaluated in a way that tests both prediction quality and the usefulness of the outlier-aware design. A random train-test split alone may not be sufficient for spatial data because nearby samples can share environmental conditions. Recent systematic evidence on soil heavymetal machine learning highlights weaknesses in the reporting of validation procedures and argues for more spatially meaningful evaluation [6].

A. Compare Against Transparent Baselines

At minimum, the proposed framework should be compared with ordinary kriging or another justified interpolation baseline, a conventional machinelearning model, and a hybrid model without the outlier-specific branch. This comparison would help determine whether the additional outlier-aware stage contributes measurable value. If the proposed approach is more complex but does not improve independent validation results or hotspot detection, its added complexity may not be justified.

B. Use Spatially Separated Validation

Where the study area and sample density allow it, spatial block cross-validation or geographically separated test locations should be considered. The test locations should be kept separate from the training locations during model development. The split design, block size, and any preprocessing steps should be reported clearly. A 2026 systematic review of 245 reports found that spatial block cross-validation was not used in the studies it examined, illustrating the need for stronger spatial validation practice [6].

C. Report Multiple Metrics

For concentration prediction, the study should report complementary metrics such as mean absolute error, root mean square error, and the coefficient of determination. Error maps should also be reviewed because a single average score can hide poor performance in a small but important hotspot. For risk classification, the confusion matrix, class-wise precision and recall, and the handling of imbalanced risk classes should be reported. Prediction uncertainty should be evaluated separately from point-estimate accuracy.

D. Test the Value of Outlier Handling

An ablation analysis should compare at least three settings: (1) all observations modelled by one method; (2) candidate outliers removed; and (3) candidate outliers retained and modelled through the proposed region-specific strategy. Such a comparison would show whether the outlier-aware approach preserves genuine local signals without allowing erroneous records to dominate the prediction.

VII. DISCUSSION AND RESEARCH GAPS

The literature points toward a practical shift: soilpollution mapping should be judged not only by how closely predicted values match observations, but also by whether the map preserves local pollution patterns and supports a defensible risk interpretation. The two anchor studies approach different parts of this problem. Yang et al. show how a spatially unusual subset can be treated separately, while Chen et al. demonstrate how semi-supervised prediction can be connected to pollution assessment and GIS [1], [2]. Combining these ideas is attractive, but their compatibility must be tested rather than assumed.

A key research gap is the limited integration of outlier interpretation, spatially robust validation, and uncertainty-aware risk mapping in one workflow. Recent reviews also caution against ranking algorithms solely by reported accuracy because differences in sample size, target metal, study design, covariates, and validation strategy can confound comparisons [6], [7]. These findings support contextaware model selection: the best model for a region is the one whose assumptions fit the available evidence and whose errors are acceptable for the intended decision.

TABLE III. MAIN RESEARCH GAPS AND SUGGESTED RESPONSES

Research gap	Suggested response in a future study

 ©IJENGT: All Rights are Reserved

Outliers removed without field context.	Check laboratory quality and source context; retain plausible hotspots.
Random-only validation.	Use spatially separated test locations and report the split design.
Prediction and risk kept separate.	Carry prediction uncertainty into risk maps.
Cross-study model rankings are confounded.	Use local baselines and report data, target metals, and validation conditions.

VIII. LIMITATIONS OF THIS REVIEW

This paper is a focused narrative review built around two supplied research studies and a small set of related peer-reviewed publications. It is not an exhaustive systematic review, and the search was not designed to produce a complete census of the field. No original soil samples were collected, no new model was trained, and no independent benchmark was conducted. The proposed framework should therefore be read as a structured research direction that needs to be tested with field data, not as evidence that a particular combination of algorithms will outperform existing methods in every agricultural region.

The two anchor studies were conducted in China and used different datasets, metals, methods, and assessment settings [1], [2]. Their findings are useful for method design, but transfer to an Indian agricultural context would require local sampling, current Indian standards, and validation under local soil and land-use conditions. Regulatory criteria and exposure assumptions must be selected for the specific location and research objective.

IX. CONCLUSION

Mapping heavy-metal contamination in agricultural soil is difficult because field observations are limited and pollution patterns vary across space. Spatial prediction can fill gaps between sampling points, but its results may be misleading when local hotspots are treated as ordinary noise or when the model is evaluated without accounting for spatial dependence. The reviewed studies offer complementary responses to these problems: one separates normal and outlier influenced areas before prediction, and the other links semi-supervised concentration estimation with pollution assessment and GIS [1], [2].

This review brings those ideas together in a conceptual outlier-aware workflow. It recommends checking data quality, investigating unusual observations, selecting models according to regional behaviour, validating predictions geographically, and presenting risk and uncertainty maps together. The framework is not yet an experimentally validated method. Its value must be established through independent testing against transparent baselines and through evidence that it preserves real pollution hotspots without increasing prediction error. A careful implementation could provide a more useful basis for prioritising soil monitoring and environmental management than a concentration map considered on its own.

REFERENCES

[1] W. Yang, M. Deng, X. Yang, and D. Wei, “Predictive soil pollution mapping: A hybrid approach for a dataset with outliers,” IEEE Access, vol. 7, pp. 46668–46676, 2019, doi: 10.1109/ACCESS.2019.2907198.
[2] X. Chen, C. Zhang, K. Yan, Z. Wei, and N. Cheng, “Risk assessment of agricultural soil heavy metal pollution under the hybrid intelligent evaluation model,” IEEE Access, vol. 11, pp. 106847–106858, 2023, doi: 10.1109/ACCESS.2023.3319428.
[3] Y. Zhang, M. Lei, K. Li, and T. Ju, “Spatial prediction of soil contamination based on machine learning: A review,” Frontiers of Environmental Science & Engineering, vol. 17, no. 8, Art. no. 93, 2023, doi: 10.1007/s11783-023-1693-1.
[4] W. Zhao, J. Ma, Q. Liu, L. Dou, Y. Qu, H. Shi, Y. Sun, H. Chen, Y. Tian, and F. Wu, “Accurate prediction of soil heavy metal pollution using an improved machine learning method: A case study in the Pearl River Delta, China,” Environmental Science & Technology, vol. 57, no. 46, pp. 17751–17761, 2023, doi: 10.1021/acs.est.2c07561.
[5] X. Ma, D.-X. Guan, C. Zhang, T. Yu, C. Li, Z. Wu, B. Li, W. Geng, T. Wu, and Z. Yang, “Improved mapping of heavy metals in agricultural soils using machine learning augmented with spatial regionalization indices,” Journal of Hazardous Materials, vol. 478, Art. no. 135407, 2024, doi: 10.1016/j.jhazmat.2024.135407.
[6] M. T. E. Tutul and S. D. Datta, “Machine learning and remote sensing for soil heavy metal(loid) risk assessment: A systematic review of multisource data fusion, benchmarking, and explainable AI,” Journal of Hazardous Materials, vol. 517, Art. no. 143536, 2026, doi: 10.1016/j.jhazmat.2026.143536.
[7] L. Yang, T. Yue, and M. Ma, “Soil heavy metals for sustainable risk management: A systematic review and a context-aware method selection framework,” Sustainability, vol. 18, no. 4, Art. no. 1893, 2026, doi: 10.3390/su18041893.

 ©IJENGT: All Rights are Reserved