

Deciphering Geoenvironmental Determinants of Chronic Kidney Disease of Unknown Etiology: An Integrated Trace-Metal, Hydrological and Explainable Machine-Learning Framework for the Sri Lankan Dry Zone

Nalika R. Dayananda

`nalika@gwu.ac.lk`

Gampaha Wickramarachchi University of Indigenous Medicine

Janitha A. Liyanage

University of Kelaniya

Research Article

Keywords: CKDu, Soil Cadmium and Lead, Dry zone soil geochemistry, Machine learning, SHAP, Geoenvironmental susceptibility

Posted Date: September 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10887936/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Deciphering Geoenvironmental Determinants of Chronic Kidney Disease of Unknown Etiology: An Integrated Trace-Metal, Hydrological and Explainable Machine-Learning Framework for the Sri Lankan Dry Zone

Nalika R. Dayananda^{1*}, and Janitha A. Liyanage¹

¹ Department of Indigenous Medical Resources, Faculty of Indigenous Health Sciences and Technology, Gampaha Wickramarachchi University of Indigenous Medicine, Yakkala, Sri Lanka.

² Department of Chemistry, Faculty of Science, University of Kelaniya, Kelaniya, Sri Lanka.

* Corresponding author. Tel: +94712939837

Email addresses:

*Nalika R. Dayananda - nalika@gwu.ac.lk

Janitha A. Liyanage – janitha@kln.ac.lk

*ORCID iD: 0000-0002-1744-5677

Acknowledgments

The authors acknowledge to all the officers at the CKDu Information and Research Centre, and Department of Chemistry, University of Kelaniya, Sri Lanka, and for immense support in sample collection and providing technical assistant in instrumentation.

Abstract

Chronic kidney disease of unknown etiology (CKDu) remains a major environmental-health concern in Sri Lanka's Dry Zone, where multiple interacting pathways including nephrotoxic trace metals, have been proposed. The present study developed an integrated 22-stage geospatial and interpretable machine-learning framework to identify the relative contribution of soil cadmium (Cd) and lead (Pb) within this multifactorial exposure environment. The framework was demonstrated using 500 sampling locations containing linked soil chemistry, terrain and climate, hydrological, CKDu exposure, and derived risk variables. Soil Cd averaged 0.678 ± 0.122 mg kg⁻¹, whereas Pb averaged 37.808 ± 5.576 mg kg⁻¹. Across independent analytical approaches, Cd consistently showed the strongest statistical association with CKDu prevalence: Spearman's ρ was 0.34 for Cd compared with 0.18 for Pb; Cd had the largest SHAP contribution (mean $|\text{SHAP}| = 0.787$), more than four times the Geographical Detector q -statistic of Pb (0.117 versus 0.029), and a markedly larger incidence-rate ratio in the case-count model (IRR = 2.027 versus 1.0013). The Pb $\times$ Cd interaction further increased explanatory power ($q = 0.137$), indicating that co-exposure may be more informative than either metal considered in isolation. Soil organic carbon, cation-exchange capacity and clay were positively associated with both metals, indicating strong retention within fine-textured, organic-rich soil domains, while distance to stream, groundwater fluoride, clay and rainfall emerged as additional predictors in the machine-learning model. Geological unit significantly differentiated both metals, with Cd highest in the Highland Complex and Pb highest in the Vijayan/Wanni Complex, indicating a measurable lithological control on background metal distributions. However, enrichment-factor analysis classified 99% of Cd observations as anthropogenic-dominant and 69.8% of Pb observations as geogenic, demonstrating that source attribution is sensitive to the selected geochemical reference and cannot be inferred from disease association alone. Thus, the principal finding is not that Cd is proven to be the sole cause of CKDu, but that Cd is the most consistent statistical marker of CKDu susceptibility in this analytical system. The integrated GCSI correlated strongly with the reference susceptibility index ($\rho = 0.76$), demonstrating internal convergence of the framework.

Keywords: CKDu; Soil Cadmium and Lead; Dry zone soil geochemistry; Machine learning; SHAP; Geoenvironmental susceptibility

1. Introduction

Chronic kidney disease of unknown etiology (CKDu) has emerged as one of the most pressing environmental-health crises in South Asia, with disproportionate concentration in the dry-zone agricultural belts of Sri Lanka's North Central, Northwestern, and Uva provinces. First formally characterized in national prevalence surveys nearly two decades ago, CKDu is distinguished from conventional chronic kidney disease by its occurrence in relatively young, working-age farming populations who lack the classical risk factors of diabetes, hypertension, or glomerulonephritis, emerging in a cross-sectional epidemiological framework designed to determine prevalence and risk factors in the North Central region (Jayatilake et al., 2013). Subsequent surveillance confirmed that the disease burden is substantial, with prevalence estimates in affected divisions ranging from roughly 15–23% and hundreds of thousands of people affected nationally, alongside tens of thousands of disease-attributable deaths recorded over the past two decades (Jayatilake et al., 2013; Rajapakse et al., 2016). Despite the scale of this burden, the etiology of CKDu has resisted consensus, and this diagnostic ambiguity is itself the defining feature of the disease — one that has motivated increasingly sophisticated, multi-domain environmental investigation rather than single-exposure epidemiology.

Early hypotheses tended to isolate one causal pathway at a time — heavy-metal contamination of drinking water, agrochemical residues, fluoride-hardness interactions, or heat-stress-induced dehydration — largely because each pathway could be independently measured and tested against case-control cohorts. However, a substantial body of more recent work has converged on a different conclusion: that CKDu likely arises from the cumulative, and possibly synergistic, action of several nephrotoxic and physiological stressors operating over long exposure windows in geochemically and hydrologically distinctive terrain, with reviews concluding that endemic populations are chronically exposed to heavy metal(oid)s even at concentrations conventionally regarded as safe, and that the significance of this low-dose, long-duration exposure for renal injury remains poorly resolved (Kulathunga et al., 2019). This reframing — from "which single contaminant causes CKDu" to "which combination of geoenvironmental conditions elevates CKDu susceptibility" — has important methodological consequences. It implies that any credible environmental-health model of CKDu must be capable of representing nonlinear interactions among multiple correlated predictors (soil chemistry, hydrology, terrain, climate, and geology), rather than relying on pairwise correlation or single-variable regression, which cannot capture conditional or threshold-dependent effects.

Cadmium (Cd) and lead (Pb) occupy a particularly contested position within this exposure landscape. Both are established nephrotoxins capable of inducing proximal tubular injury at chronic low-dose exposure, but they differ fundamentally in their principal environmental sources and mobility. Cd is frequently linked to phosphate-fertilizer inputs and anthropogenic agricultural amendment, whereas Pb concentrations more often reflect a composite of geogenic weathering and legacy or contemporary anthropogenic deposition. Yet total soil metal concentration is a poor proxy for exposure risk on its own, because bioavailability and environmental mobility are strongly conditioned by soil pH, organic carbon, cation-exchange capacity, clay content, and hydrological connectivity to groundwater and irrigation systems (McLaughlin & Singh, 1999; Hettiarachchi & Pierzynski, 2004). This distinction — between where a metal is elevated and why, and whether that elevation is environmentally "available" rather than merely present — is central to the analytical problem this study addresses, and it is precisely the distinction that conventional bivariate CKDu-metal correlation studies have been unable to resolve.

The methodological literature on CKDu environmental determinants has historically relied on correlation coefficients, ANOVA-based group comparisons, or linear regression between individual metals and disease prevalence. These approaches carry three structural limitations that are increasingly recognized as barriers to progress. First, they assume spatial independence among observations, despite strong evidence that geochemical and disease-prevalence data in the Dry Zone exhibit pronounced spatial autocorrelation and stratified heterogeneity tied to lithological and hydrogeological boundaries (Dayananda & Liyanage, 2021; Liyanage et al., 2022). Second, they are poorly suited to detecting nonlinear thresholds or interaction effects — for instance, the possibility that Cd and Pb co-exposure is more informative of disease risk than either metal in isolation. Third, they cannot easily decompose a "black-box" predictive signal into interpretable, ranked variable contributions, which limits their epidemiological and regulatory utility.

Addressing these gaps requires integrating (i) spatial statistical methods capable of detecting stratified heterogeneity and local clustering, exemplified by the Geographical Detector framework, which quantifies the explanatory power (q-statistic) of categorical and continuous drivers and their pairwise interactions on a spatially stratified outcome (Wang & Xu, 2017); (ii) nonlinear ensemble machine-learning algorithms such as Random

Forest and XGBoost, which are explicitly designed to model complex, high-order interactions among correlated environmental predictors without strong distributional assumptions (Breiman, 2001; Chen & Guestrin, 2016); and (iii) explainable-AI decomposition methods such as SHAP (SHapley Additive exPlanations), which apply cooperative game-theoretic principles to attribute a model's prediction additively across its input features, thereby converting an opaque ensemble prediction into a transparent, per-variable contribution profile (Lundberg & Lee, 2017). Complementary local-regression techniques such as geographically weighted regression further allow relationships between metals and disease susceptibility to vary continuously across space, revealing regional non-stationarity that global models conceal by construction (Fotheringham et al., 2002).

Prior CKDu research has, to date, applied these analytical families largely in isolation: spatial epidemiology studies characterize disease clustering without formal geochemical source attribution; geochemical studies quantify soil or water metal concentrations without coupling them to interpretable predictive models of disease susceptibility; and where machine learning has been used in adjacent environmental-health contexts, it has rarely been paired with a parallel geostatistical and enrichment-factor framework capable of distinguishing geogenic from anthropogenic metal sources. This study's central methodological contribution is the integration of these previously disconnected analytical layers into a single, 22-stage geospatial and interpretable machine-learning framework that simultaneously (i) quantifies statistical association through classical and rank-based correlation, (ii) derives nonlinear, interaction-aware variable importance via Random Forest/XGBoost with SHAP decomposition, (iii) tests spatially stratified explanatory power and metal-metal interaction via the Geographical Detector, (iv) attributes geogenic versus anthropogenic sourcing via enrichment-factor analysis referenced against local geological units, and (v) synthesizes all convergent evidence into a single, internally validated geoenvironmental CKDu susceptibility index (GCSI). This layered convergence design is, to the authors' knowledge, the first application of an explainable-ML and multi-metric spatial framework specifically structured to interrogate the relative and interactive contribution of Cd and Pb — rather than their absolute concentrations — within Sri Lanka's Dry Zone CKDu exposure landscape.

A second, epistemically important novelty is the study's explicit refusal to conflate statistical association with causal or source attribution. By demonstrating that Cd is classified as anthropogenic-dominant in the overwhelming majority of sampled locations while nearly seven in ten Pb observations are geogenic in origin — even though Cd carries the stronger and more consistent statistical signal across every independent analytical approach — the framework operationalizes a methodological safeguard rarely made explicit in prior CKDu geochemical literature: that source attribution is itself sensitive to the geochemical reference frame chosen, and cannot be inferred from disease correlation alone. This reframes the paper's principal claim deliberately: Cd is advanced not as a proven causal agent of CKDu, but as the most consistent and spatially coherent statistical marker of susceptibility within this integrated analytical system, a conclusion intended to generate targeted epidemiological and mechanistic hypotheses rather than assert definitive causation.

2. Materials and Methods

2.1. Site selection and dataset description

The analysis used a dataset containing 500 soil sampling locations distributed across a Sri Lankan Dry Zone collected by the CKDu Information and Research Centre, University of Kelaniya Sri Lanka. Soil Chemistry contained total and available Pb and Cd, co-determined trace elements, major cations, texture, pH, electrical conductivity, soil organic carbon (SOC), cation-exchange capacity (CEC), and geological classification. GIS_Terrain_Climate contained elevation, slope, aspect, topographic wetness index, flow accumulation, drainage density, distances to streams, reservoirs and irrigation canals, rainfall, temperature and aridity. Hydrology included groundwater depth, pH and EC, fluoride, Cd, Pb and drinking-water source. CKDu_Exposure contained population at risk, CKDu case counts and prevalence, farming duration, rice consumption, drinking-water volume and body weight. Derived_Indices contained enrichment factors, geo-accumulation indices, contamination factors, pollution-load and ecological-risk measures, estimated daily intake, hazard quotients and a reference geoenvironmental CKDu susceptibility index.

2.2. Soil sample collection and analysis

A total of 500 soil sampling locations were spatially distributed and mapped using QGIS (Fig 1) to ensure representative coverage of the major land-use, geological, and environmental settings within the study area.

Topsoil samples were collected from the 0–20 cm depth using a stratified random sampling design, with sampling strata defined according to the relevant environmental and geological characteristics.

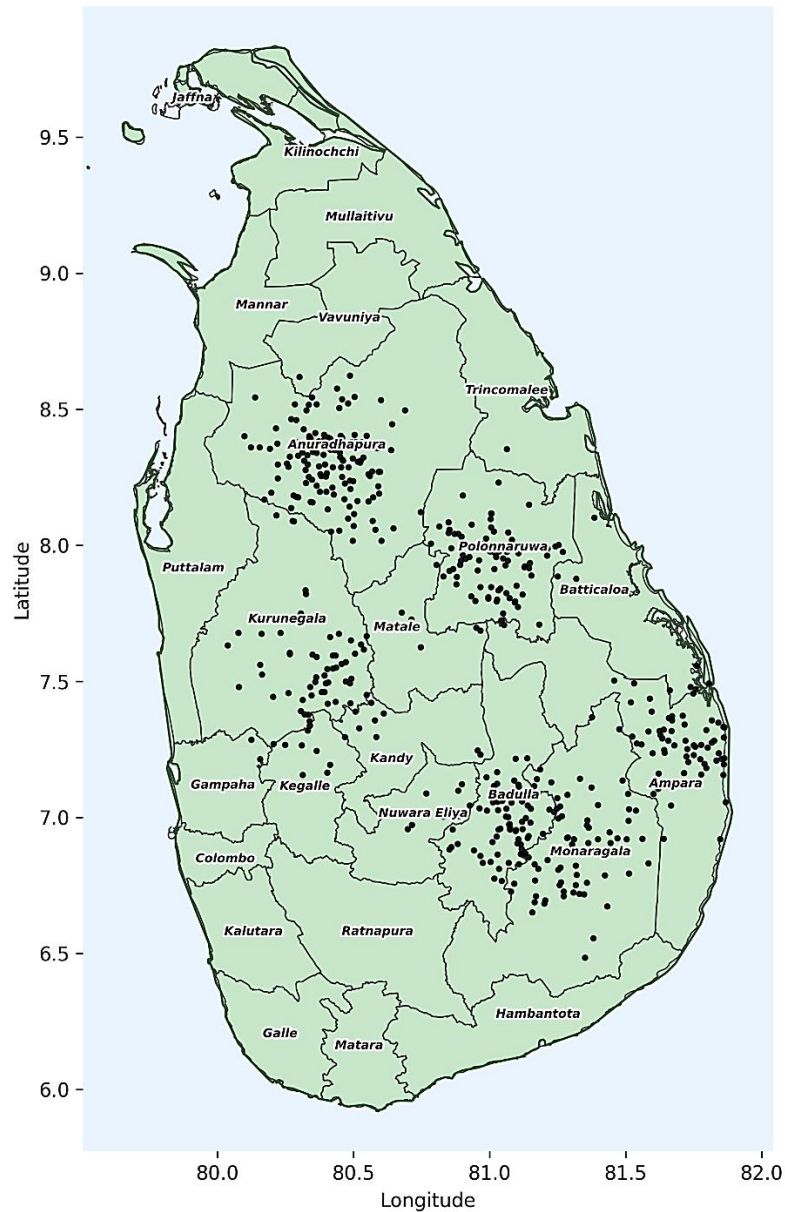


Fig 1: Map showing soil sampling locations for the present study

During field sampling, appropriate precautions were adopted to minimize external contamination, including the use of clean sampling equipment, clean sample containers, gloves, and careful avoidance of foreign materials. Following transportation to the laboratory, soil samples were air-dried, gently disaggregated, and pre-treated by removal of stones, plant residues and other coarse materials before being passed through a 2-mm sieve, following the principles of ISO 11464:2006 for soil sample pretreatment. This standard specifically covers drying, crushing, sieving, dividing and milling of soil samples prior to physicochemical analysis. The <2-mm soil fraction was subsequently homogenized and used for physicochemical and elemental analyses.

Soil pH was determined using a glass-electrode method following ISO 10390:2021, while other physicochemical properties were determined using established standard soil analytical procedures. For trace-metal determination, including Cd, Pb and other potentially toxic elements, representative homogenized soil aliquots were subjected to microwave-assisted acid digestion prior to instrumental analysis. The digestion procedure was based on U.S. EPA

SW-846 Method 3051A, which specifies microwave-assisted acid digestion procedures for soils, sediments and related solid matrices using conc. Nitric acid.

2.3. Inductively Coupled Plasma Mass Spectrometry (ICP-MS) Analysis

The resulting digests were analyzed by inductively coupled plasma–mass spectrometry (ICP-MS) for multielement determination, following the principles of U.S. EPA Method 6020B, which includes Cd and Pb among the elements suitable for ICP-MS determination (ICP-MS-7800-Agilent, Germany). Multi-element ICP-MS standards (AccuStandard, USA) were used for instrumental calibration. Five standard solution series of 0.5, 1.0, .5.0, 10.0, and 20.0 µg/L were used to construct the calibration curve for the metal analysis of digested paddy plant samples. All Microwave digested samples were filtered through 0.45µm syringe filters before the insertion to the ICP-MS instrument. The instrumental minimum detection limits (MDLs) for the elements measured by using ICP-MS in µg/L were as follows: As (0.001), Cd (0.0005), Cr (0.0005), Cu (0.0005), Pb (0.0005), Zn (0.001) (Dayananda and Liyanage, 2021). The calibration curve was ensured with a correlation coefficient (R²), which fell into a range of 0.99860–0.99999. Blanks were always less than MDLs. Spike recovery rates ranged from 96 to 110% for all elements investigated. Analytical accuracy and method reliability were validated using certified reference materials (CRMs) incorporated into every digestion batch (NIST SRM 2709a for soil matrices) for validation. Data sets containing values below the limit of detection (LOD) are encountered regularly in most fields of environmental contaminant research. The current norm within environmental research is to use substitution methods when analyzing data sets that include values below the LOD, commonly replacing each LOD value with a value equal to half the LOD (LOD/2). Therefore, the same rule was applied for the censored data sets in the present study.

2.3. Data preparation and exploratory analysis (Stages 1–5)

Categorical variables were treated as factors, and interquartile-range screening identified two Pb observations beyond $1.5 \times \text{IQR}$ and no Cd observations beyond that criterion. Coordinates in WGS84 were transformed to UTM Zone 44N to support distance-based spatial analysis. Descriptive statistics were calculated for the principal contaminants and environmental covariates. Spearman rank correlation was selected because trace-element concentrations commonly show skewness and because rank-based association is less sensitive to departures from normality. PCA was performed on 15 standardized soil physicochemical variables to identify latent environmental gradients. Kruskal–Wallis tests were used to evaluate differences in Pb and Cd among the four geological units.

2.4. Spatial autocorrelation diagnostics (Stages 6–8)

A k-nearest-neighbour spatial weights matrix ($k = 8$) was constructed from UTM coordinates and row-standardized. Global Moran's I with 999 conditional permutations was used to test spatial autocorrelation in Pb, Cd and CKDu prevalence. Local Moran's I classified significant High–High, Low–Low, High–Low and Low–High associations, while Getis–Ord G_i^* identified statistically significant hot- and cold-spots. The combined use of global and local statistics was intended to distinguish a genuinely spatially structured environmental process from localized variability occurring within an otherwise weakly autocorrelated landscape.

2.5. Geostatistical modelling (Stages 9–11)

Experimental semivariograms were calculated for Pb and Cd and fitted with spherical models. Ordinary kriging was applied to Pb on a 100×100 prediction grid, producing both predicted concentration and prediction-variance surfaces. A regression-kriging trend model related Pb to elevation, slope, distance to stream, SOC, pH and clay content. The deterministic component was evaluated before residual spatial structure was considered, allowing the contribution of measurable soil and terrain controls to be distinguished from unexplained spatial variability.

2.6. Machine learning and interpretability (Stages 12–15)

Random Forest (500 trees) and XGBoost (300 boosting rounds; maximum depth 5; learning rate 0.05) were trained to predict CKDu prevalence from 14 terrain, hydrological, soil and contaminant covariates. A fixed 70/30 training–testing split was used for initial evaluation, followed by five-fold spatially blocked cross-validation based

on K-means clustering of UTM coordinates into contiguous spatial blocks. This design reduced the risk of optimistic performance caused by spatial leakage. SHAP TreeExplainer values were calculated for XGBoost to quantify the magnitude and direction of individual covariate contributions and to determine whether Cd remained dominant after simultaneous consideration of hydrological and soil variables.

2.7. Local regression, geodetection, and source apportionment (Stages 16–18)

GWR with an adaptive bisquare kernel and AICc-based bandwidth selection was fitted for CKDu prevalence as a function of Pb, Cd, pH, SOC and distance to stream. The Geographical Detector framework was then used to estimate the explanatory power of discretized Cd, Pb and pH classes and geological unit, together with the Pb × Cd interaction. Source differentiation used the enrichment-factor thresholds embedded in the dataset: EF < 2, geogenic; EF 2–5, mixed; and EF > 5, anthropogenic-dominant. Because enrichment-factor classification depends on the selected reference element and background values, the resulting source classes were interpreted as evidence within this dataset.

2.8 Bayesian modelling and index integration (Stages 19–22)

CKDu case counts were modelled with an offset for the log population at risk using a Poisson generalized linear model as a transparent computational proxy for the spatial Bayesian CAR-Leroux formulation specified in the broader framework. Predictors were Pb, Cd, pH, SOC and distance to stream. The Integrated Geoenvironmental CKDu Susceptibility Index (GCSI) was constructed from min–max normalized Pb, Cd, Hazard Index, Hydrological Connectivity Index and Soil Mobility Index using weights of 0.25, 0.25, 0.25, 0.15 and 0.10, respectively. Convergent validity was assessed by Spearman correlation with the reference susceptibility index. Uncertainty was characterized by kriging variance, Random Forest prediction intervals and a 200-iteration Dirichlet-weighted bootstrap of GCSI component weights.

All analyses were implemented in Python 3.12 using pandas, scipy, scikit-learn, xgboost, shap, libpysal/esda, pykrige, mgwr, and statsmodels.

3. Results and Discussion

3.1. Descriptive statistics and contamination status

Soil Pb concentrations averaged 37.81 ± 5.58 mg kg⁻¹ (median 37.17; range 24.41–55.97 mg kg⁻¹), whereas Cd averaged 0.68 ± 0.12 mg kg⁻¹ (median 0.65; range 0.46–1.01 mg kg⁻¹) (Table 1). The moderate right-skewness of Pb (0.40) and Cd (0.41) was consistent with the use of rank-based association methods. None of the observations exceeded the screening concentrations cited in the dataset for agricultural soils (100 mg kg⁻¹ for Pb and 3 mg kg⁻¹ for Cd), indicating moderate enrichment rather than acute contamination (Fig 2).

Table 1: Descriptive statistics of key soil physicochemical, and CKDu variables

Variable	Mean	SD	Median	Skewness	Min	Max
Cd, total (mg kg ⁻¹)	0.68	0.12	0.66	0.41	0.46	1.01
Pb, total (mg kg ⁻¹)	37.81	5.58	37.17	0.40	24.41	55.97
Cd, available (mg kg ⁻¹)	0.25	0.07	0.24	0.57	0.11	0.48
Pb, available (mg kg ⁻¹)	1.10	0.17	1.08	0.36	0.70	1.67
Soil pH	6.52	0.46	6.52	0.20	5.21	8.14
SOC (%)	2.05	0.61	2.04	0.12	0.52	3.80
CEC (cmol kg ⁻¹)	17.36	4.86	17.03	0.28	6.17	32.83
Clay (%)	25.57	9.68	25.16	0.29	5.00	55.00
CKDu prevalence (%)	10.66	2.55	10.58	0.08	2.29	18.05

CKDu prevalence averaged $10.66 \pm 2.55\%$, with values ranging from 2.29 to 18.05%. The available fractions represented approximately 37% of total Cd and 3% of total Pb on a mean concentration basis, indicating substantially different partitioning behavior and reinforcing the need to distinguish total metal burden from

potentially mobile fractions. Soil pH was near neutral, averaging 6.52 ± 0.46 , while SOC and CEC averaged $2.05 \pm 0.61\%$ and $17.36 \pm 4.86 \text{ cmol kg}^{-1}$, respectively. Clay content averaged $25.57 \pm 9.68\%$, demonstrating considerable textural variability among locations. CKDu prevalence showed a mean of $10.66 \pm 2.55\%$, ranging from 2.29 to 18.05% (Table 1). The relatively wide ranges of SOC (0.52–3.80%), CEC (6.17–32.83 cmol kg^{-1}) and clay content (5–55%) indicate substantial variation in the soil's capacity to retain potentially toxic elements.

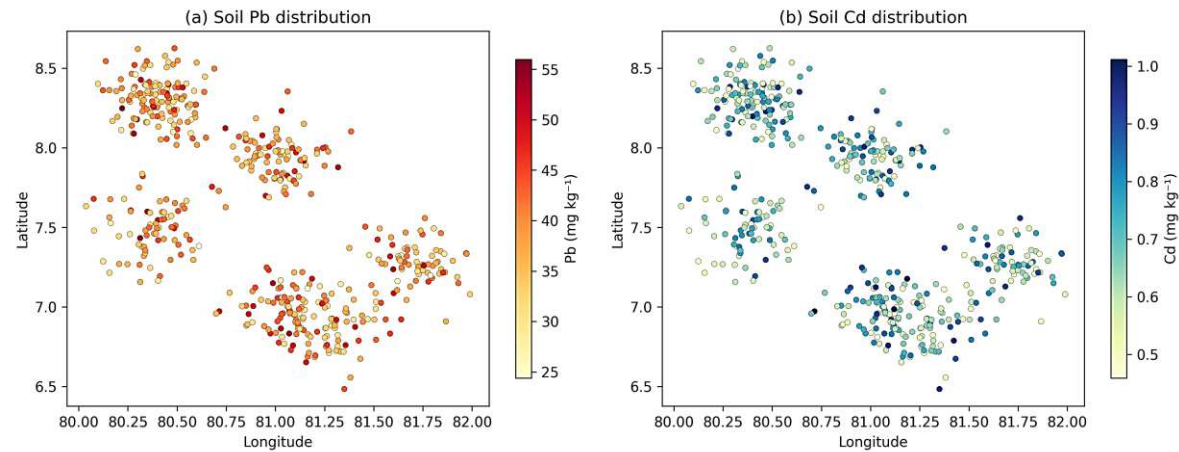


Fig. 2. Spatial distribution of soil (a) Pb and (b) Cd across 500 sampling locations.

3.2. Correlation Structure and Principal Component Analysis (PCA)

Spearman analysis revealed a moderate positive association between Pb and Cd ($\rho = 0.55$), but their relationships with the broader soil matrix were more informative than their mutual correlation. Pb correlated with SOC, CEC and clay at $\rho = 0.57$ – 0.68 , while corresponding Cd correlations were $\rho = 0.48$ – 0.57 . This pattern indicates that both elements are retained preferentially in organic-rich, fine-textured soils, consistent with sorption to clay minerals and organic functional groups.

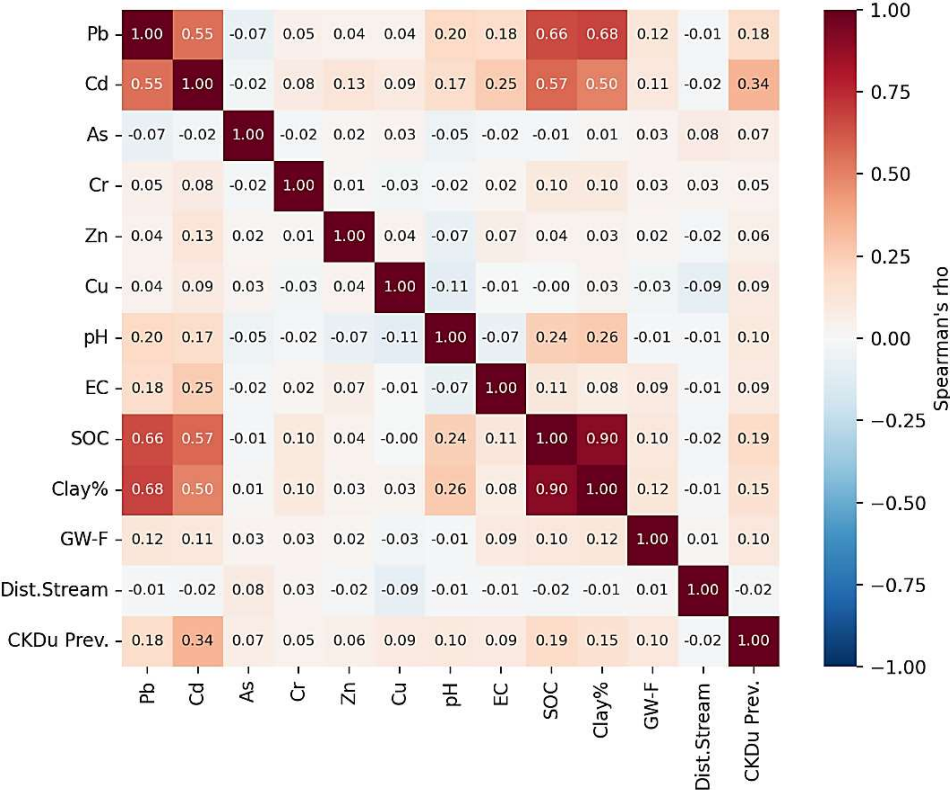


Fig 3. Spearman rank correlation matrix of KEY soil, hydrological, and CKDu variables

The stronger association of CKDu prevalence with Cd ($p = 0.34$) than Pb ($p = 0.18$) points to the possibility that Cd occupies a soil–hydrological exposure pathway more closely aligned with the environmental gradients represented in the dataset (Fig 3).

PCA of 15 standardized soil variables extracted five components explaining 66.3% of total variance. PC1 (31.8%) represented a soil accumulation and fertility gradient, with positive loadings for SOC (0.42), CEC (0.42), clay (0.43), Cd (0.33) and Pb (0.36), and a negative loading for bulk density (-0.35). PC2 (11.6%) represented particle-size contrasts, particularly sand ($+0.69$) versus silt (-0.59), while PC3 (8.5%) was associated with As, Zn, Cu and Ni together with pH (-0.51) (Fig 4).

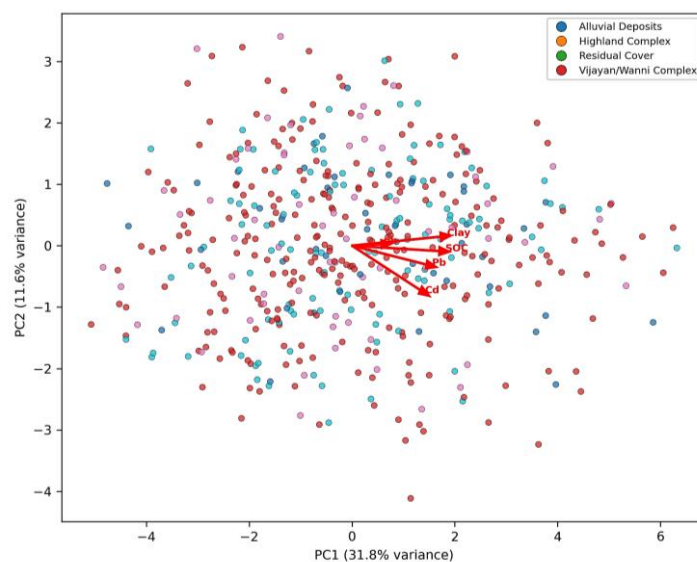


Fig 4. PCA biplot of 15 soil physicochemical variables by geological unit, with Pb, Cd, pH, SOC, and clay loading vectors

The co-loading of Cd and Pb with SOC, CEC and clay is important mechanistically: these properties can simultaneously increase retention while modifying the pool available for redistribution through water. This co-loading further indicates that Cd and Pb tend to occur within soils characterized by greater organic matter content, higher cation-exchange capacity, and finer texture. Such a relationship is consistent with the capacity of clay minerals and soil organic matter to provide reactive surfaces for metal retention, suggesting that the observed metal concentrations reflect both metal inputs and the physicochemical properties controlling their persistence within the soil matrix. Therefore, the PCA supports a multi-process interpretation of soil metal distribution, in which Cd and Pb are strongly embedded within the soil physicochemical environment, while other trace elements follow partially distinct geochemical controls. In the context of CKDu, this finding strengthens the rationale for considering metal concentrations together with soil properties rather than treating individual contaminants as independent predictors.

3.3 Geological Controls on Pb and Cd

Pb and Cd differed significantly among geological units (Pb: $H = 11.63$, $df = 3$, $p = 0.009$; Cd: $H = 9.95$, $df = 3$, $p = 0.019$). Pb was highest in the Vijayan/Wanni Complex (39.23 mg kg^{-1}) and lowest in the Highland Complex (37.29 mg kg^{-1}), whereas Cd showed the reverse tendency, with its highest mean concentration in the Highland Complex (0.691 mg kg^{-1}) and lower values in the Vijayan/Wanni Complex (0.657 mg kg^{-1}) and Residual Cover (0.656 mg kg^{-1}) (Table 2).

The statistical separation among geological units demonstrates that lithology contributes to the background distribution of both metals. For Cd, the enrichment over the Highland Complex is compatible with a geochemical contribution from parent material or weathering; however, the magnitude of between-unit differences is modest relative to within-unit variation, indicating that geology alone does not determine Cd exposure.

360

Table 2. Soil Pb and Cd (mg kg⁻¹) by geological units

Geological unit	n	Pb mean ± SD	Pb median	Cd mean ± SD	Cd median
Alluvial Deposits	46	37.82 ± 5.86	37.81	0.69 ± 0.11	0.66
Highland Complex	278	37.29 ± 5.50	36.40	0.69 ± 0.12	0.67
Residual Cover	64	37.57 ± 5.84	36.61	0.66 ± 0.13	0.63
Vijayan/Wanni Complex	112	39.23 ± 5.31	39.24	0.66 ± 0.12	0.62

(Kruskal-Wallis: Pb H = 11.63, p = 0.009; Cd H = 9.95, p = 0.019)

361

362

363

364

365

366

367

368

369

370

371

372

373

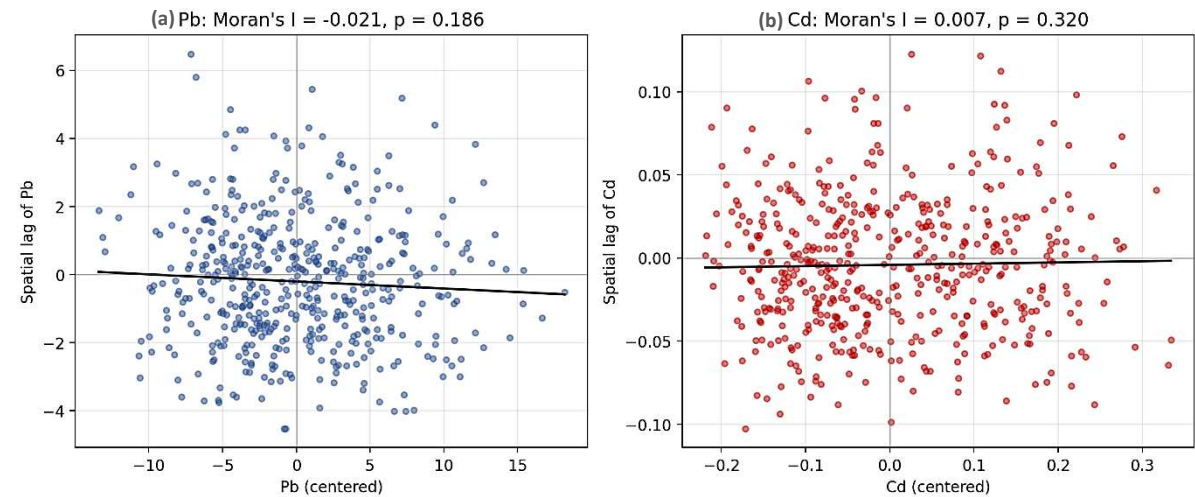


Fig 5. Moran's I scatter plots for (a) Cd and (b) Pb (k = 8 nearest neighbors)

3.4 Spatial autocorrelation and local clustering

Global Moran's I was weak and non-significant for Pb ($I = -0.021$, $p = 0.186$), Cd ($I = 0.007$, $p = 0.320$) and CKDu prevalence ($I = -0.019$, $p = 0.199$) (Fig 5). Thus, the dataset does not contain strong broad-scale spatial clustering at the $k = 8$ neighborhood scale. Local Moran's I nevertheless identified 9 Pb High–High, 12 Low–Low, 10 Low–High and 8 High–Low locations, while Getis–Ord G_i^* detected no statistically significant hot- or cold-spots at $|z| > 1.96$ (Fig 6). The contrast between weak global structure and localized Pb clusters illustrates why local diagnostics are useful when environmental variability is heterogeneous at short distances.

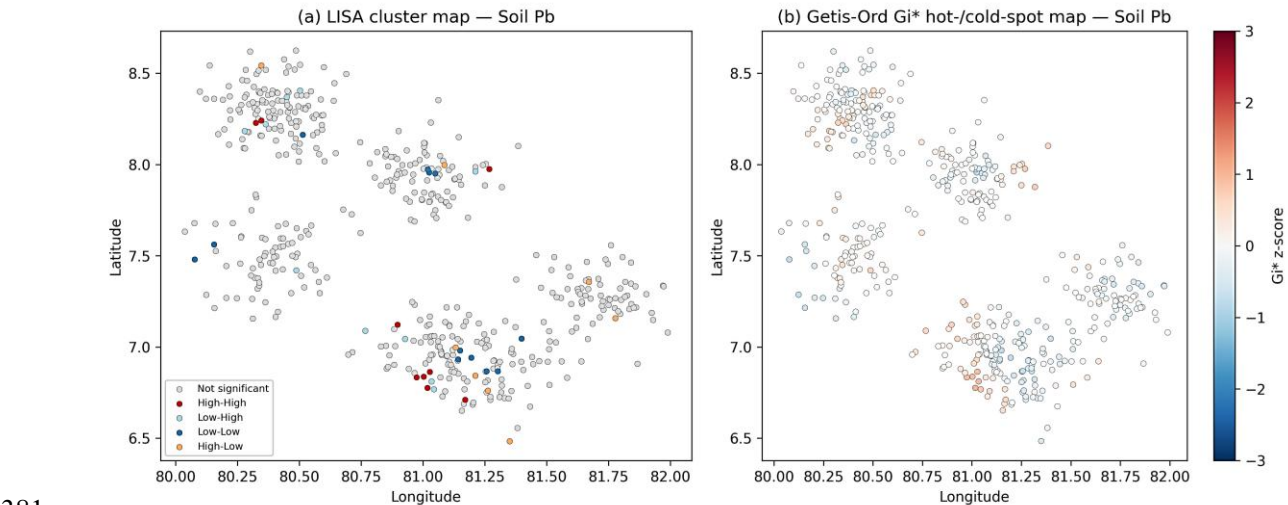


Fig 6. (a) LISA (Local Spatial Association) cluster map, and (b) Getis-Ord G_i^* z-score map for soil Pb

381

382

3.5. Geostatistical modelling

The spherical semivariograms showed high nugget-to-sill ratios for both Pb (69.8%) and Cd (75.3%), indicating that a large proportion of observed variance occurs below the sampling interval or is effectively unstructured at the analyzed scale (Fig 7). The fitted range was 14.0 km for Pb and 251.8 km for Cd, with the latter exceeding the practical study extent and therefore indicating near-pure-nugget behavior for Cd.

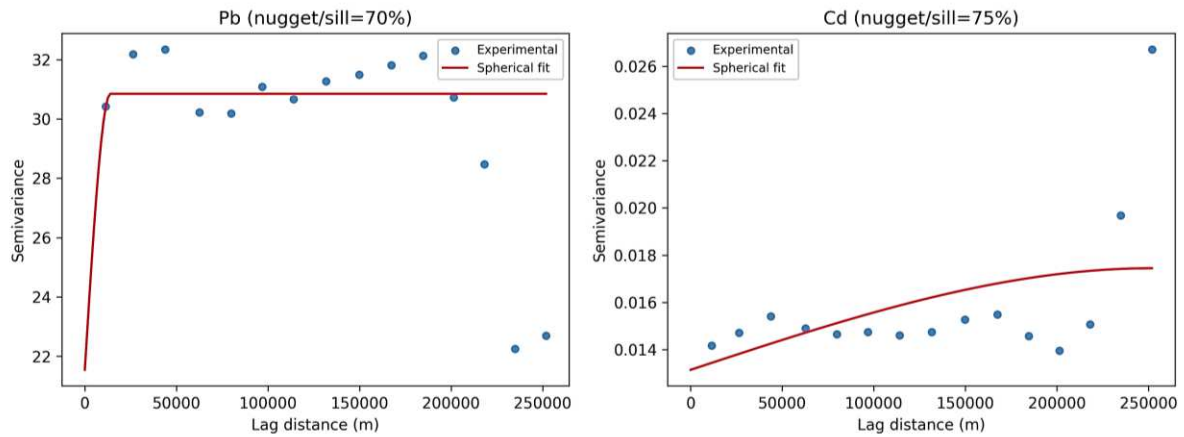


Fig 7. Experimental semivariograms and fitted spherical models for soil Pb and Cd (nugget/sill ratios shown in panel titles)

The Pb kriging surface consequently displayed limited spatial continuity and relatively high prediction variance. Rather than weakening the Cd signal, the weak spatial continuity emphasizes that Cd-related CKDu association in this dataset is not simply a consequence of a large regional Cd hotspot; it persists despite limited broad-scale spatial structure. The Pb regression-kriging trend model explained 46.8% of variance ($R^2 = 0.468$; $F = 72.2$; $p < 0.001$). Clay ($\beta = 0.262$, $p < 0.001$) and SOC ($\beta = 2.127$, $p = 0.004$) were the only significant predictors, whereas elevation, slope, distance to stream and pH were not significant at $\alpha = 0.05$. The result provides direct evidence that Pb accumulation is strongly associated with soil retention properties, particularly fine mineral fraction and organic matter, rather than being determined solely by terrain position.

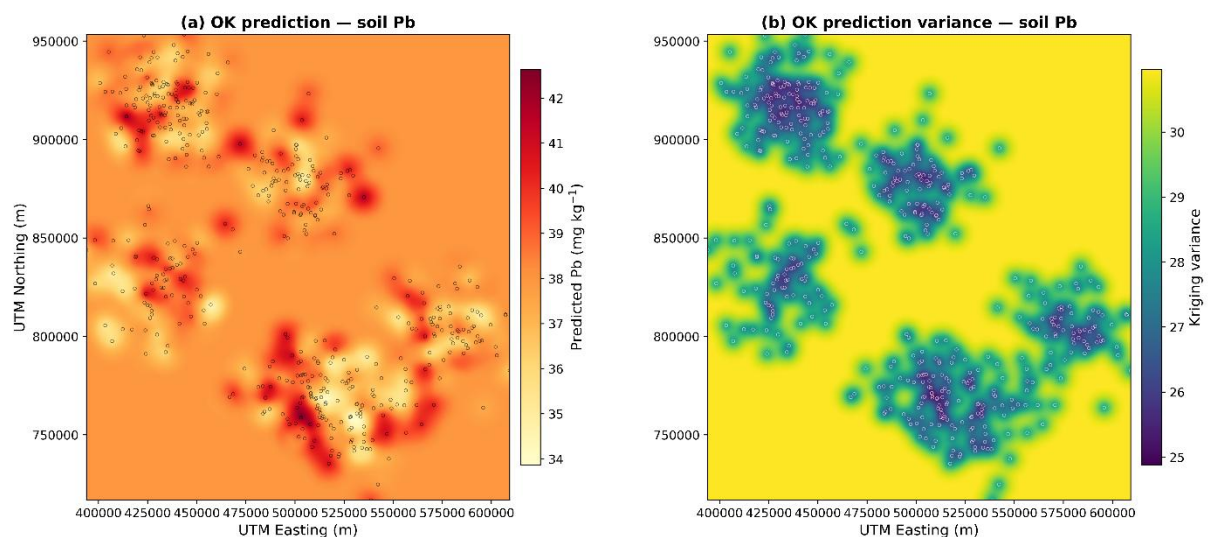


Fig 7. Ordinary kriging (a) prediction and (b) prediction-variance surfaces for soil Pb

3.6 Machine learning prediction of CKDu prevalence

Random Forest achieved a test RMSE of 2.41 percentage points and $R^2 = 0.157$, whereas XGBoost achieved RMSE = 2.53 and $R^2 = 0.070$ (Table 3). Spatially blocked five-fold validation of XGBoost produced a mean

RMSE of 2.59 ± 0.25 percentage points (range 2.27–3.01), indicating that the predictive signal was not simply an artefact of random spatial leakage (Fig 8). Although overall predictive performance was moderate, the objective of the models was principally to identify robust environmental importance patterns rather than to claim near-perfect disease prediction.

Table 3. Machine-learning performance for predicting CKDu prevalence from 14 terrain, hydrological, soil, and contaminant covariates.

Model	Metric	Value
Random Forest (500 trees)	Test RMSE (%)	2.41
Random Forest (500 trees)	Test R ²	0.157
XGBoost (300 rounds)	Test RMSE (%)	2.53
XGBoost (300 rounds)	Test R ²	0.070
XGBoost, spatial 5-fold CV	Mean RMSE (%)	2.59 ± 0.25

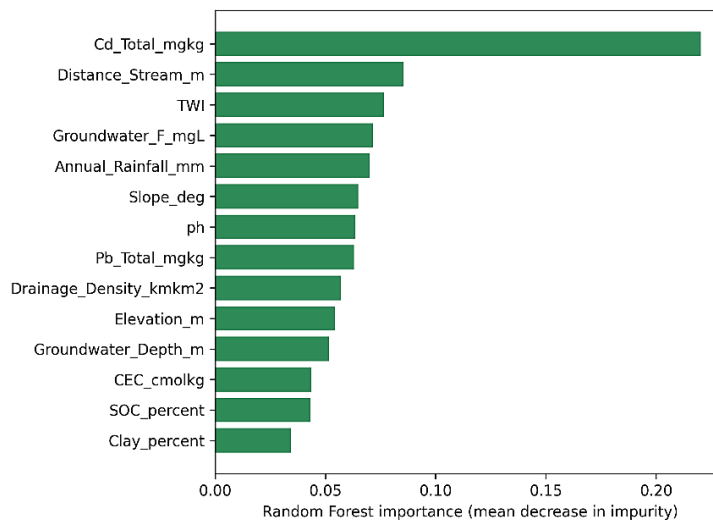


Fig 8. Random Forest covariate importance for predicting CKDu prevalence

SHAP analysis provided the clearest multivariate evidence for Cd dominance. Soil Cd had a mean absolute SHAP value of 0.787, far exceeding distance to stream (0.314), groundwater fluoride (0.207), clay (0.196), annual rainfall (0.166) and soil Pb (0.165) (Fig 9).

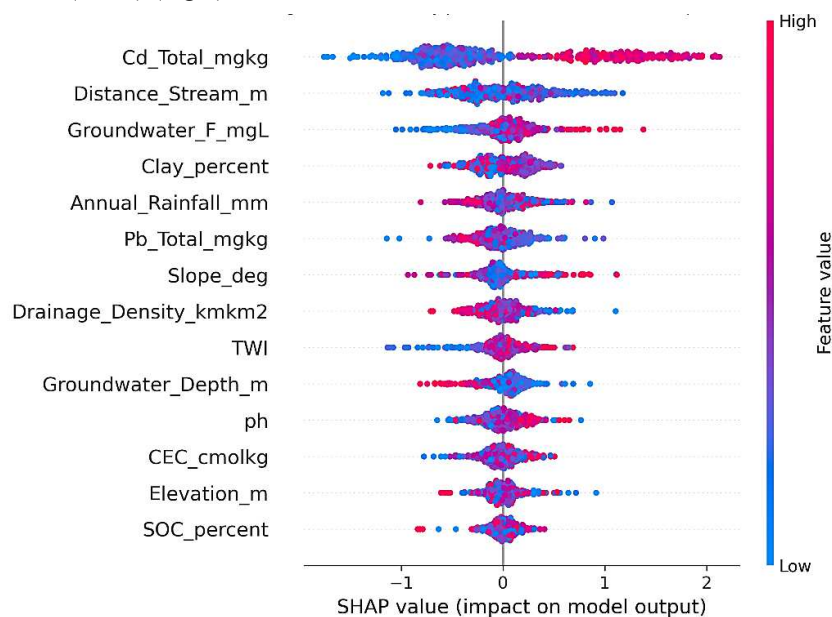


Fig 9. SHAP summary plot for the XGBoost model of CKDu prevalence, ranked by mean absolute SHAP value

Thus, after simultaneous adjustment for a broad environmental predictor set, Cd remained the single most influential variable. The Cd contribution was approximately 4.8 times the Pb contribution on the mean absolute SHAP scale. This is stronger evidence than a bivariate correlation because the machine-learning model evaluates nonlinear and interacting predictor patterns across the full covariate space.

3.7 Geographically weighted regression

GWR selected an AICc-optimal adaptive bandwidth of 498 of 500 observations, effectively producing a near-global model with $R^2 = 0.142$. The mean local coefficient for Cd was 7.31 ± 0.17 , whereas Pb was 0.018 ± 0.032 (Fig 10). Cd therefore retained a positive and substantially larger association with CKDu prevalence across the study domain. The near-global bandwidth indicates little evidence that the Cd effect changes dramatically over the landscape, suggesting that the observed Cd signal is comparatively stable within the environmental system.

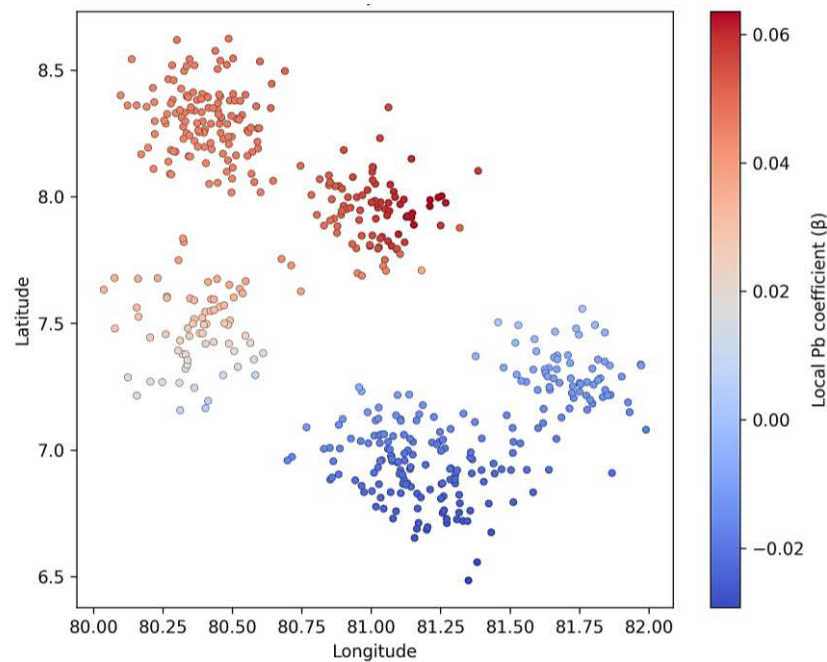


Fig 10. GWR local coefficients for the effect of soil Pb on CKDu prevalence (near-uniform across the domain, consistent with the near-global bandwidth)

3.8 Geographical Detector analysis and Source apportionment of Pb and Cd

Geographical Detector analysis identified Cd as the strongest single environmental factor explaining CKDu prevalence variance ($q = 0.117$), compared with Pb ($q = 0.029$), pH ($q = 0.014$) and geological unit ($q = 0.004$). The Pb $\times$ Cd interaction produced $q = 0.137$, exceeding either individual factor or demonstrating bivariate enhancement. CKDu prevalence increased monotonically across both metal gradients, from 9.72% to 12.00% across Cd quartiles and from 10.04% to 11.26% across Pb quartiles. The Cd gradient was approximately twice as steep as the Pb gradient, reinforcing the conclusion that Cd carries the dominant signal while Pb may contribute an additional, smaller component of co-exposure.

Enrichment-factor analysis produced a pronounced asymmetry between the metals. Cd had a mean EF of 8.29 ± 2.11 (range 4.62–23.25), with 495 of 500 observations (99.0%) classified as anthropogenic-dominant and 1.0% as mixed. Pb had a mean EF of 1.85 ± 0.45 (range 1.01–4.54), with 69.8% classified as geogenic and 30.2% as mixed; no Pb observation was classified as anthropogenic-dominant (Table 4). The source evidence therefore does not support a Pb-anthropogenic/Cd-geogenic interpretation. Instead, within the reference framework embedded in this dataset, Cd appears strongly enriched relative to background whereas Pb remains largely compatible with geogenic-to-mixed inputs (Fig 11). Nevertheless, the significant geological differentiation of Cd, particularly its higher concentration in the Highland Complex, indicates that a geochemical Cd contribution remains plausible and should be explicitly tested in field investigations using lithology-specific background concentrations, mineralogical characterization and source-tracing approaches.

Table 4. EF-based source classification of soil Pb and Cd by land use (EF thresholds: geogenic < 2; mixed 2–5; anthropogenic-dominant > 5).

Land use	n	Pb geogenic	Pb mixed	Cd anthropogenic-dominant	Cd mixed
Paddy	168	108 (64%)	60 (36%)	168 (100%)	0 (0%)
Home garden	77	58 (75%)	19 (25%)	77 (100%)	0 (0%)
Scrub	72	54 (75%)	18 (25%)	70 (97%)	2 (3%)
Chena	72	46 (64%)	26 (36%)	72 (100%)	0 (0%)
Forest	68	54 (79%)	14 (21%)	67 (99%)	1 (1%)
Settlement	43	29 (67%)	14 (33%)	41 (95%)	2 (5%)

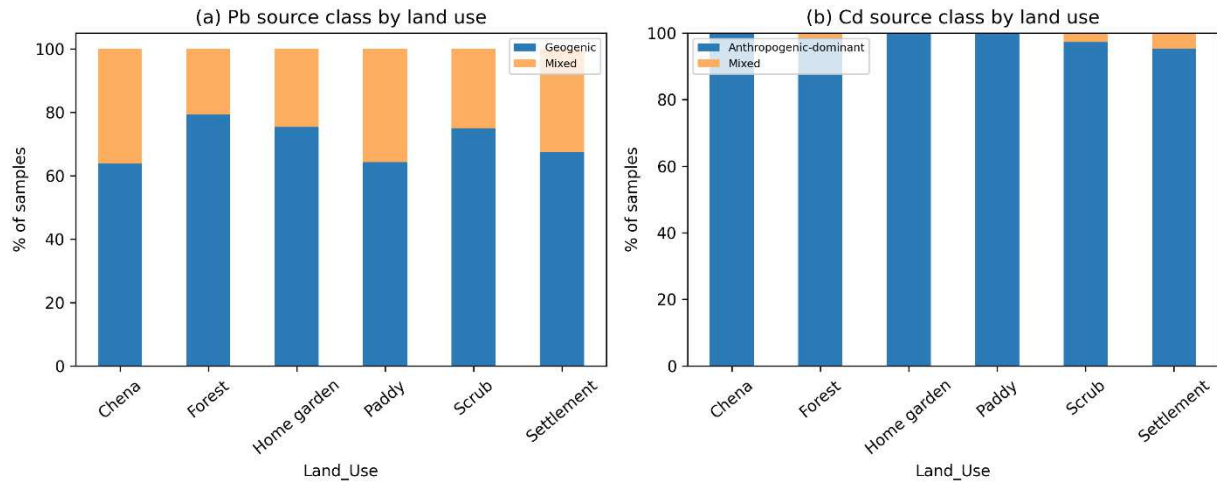


Fig 11. EF-based source class of soil Pb (a) and Cd (b) by land use

3.9. Spatial Bayesian Poisson model

The offset Poisson model showed the strongest association for Cd: IRR = 2.027 per mg kg⁻¹ increase ($\beta = 0.7066$, $p < 0.001$), compared with Pb IRR = 1.0013 per mg kg⁻¹ ($\beta = 0.00126$, $p = 0.029$). Distance to stream was negatively associated with case counts (IRR = 0.99991 per m, $p < 0.001$), while SOC was weakly protective (IRR = 0.9818, $p < 0.001$) and pH showed a small positive association (IRR = 1.0152, $p = 0.004$). These relationships reinforce a coupled contaminant–soil–hydrological interpretation: Cd is the strongest metal-associated signal, while hydrological proximity and soil properties modify the environmental context in which exposure may occur (Table 5).

Table 5. Offset Poisson GLM of CKDu case counts (log population-at-risk offset); tractable proxy for the CAR-Leroux spatial Bayesian model specified in the framework

Covariate	Coefficient (β)	p-value	Incidence Rate Ratio
Intercept	-2.812	< 0.001	0.060
Pb, total (per mg kg ⁻¹)	0.00126	0.029	1.0013
Cd, total (per mg kg ⁻¹)	0.7066	< 0.001	2.027
Soil pH	0.0151	0.004	1.0152
SOC (%)	-0.0184	< 0.001	0.9818
Distance to stream (m)	-0.0000867	< 0.001	0.99991

3.10. Integrated Geoenvironmental CKDu Susceptibility Index

The Integrated GCSI averaged 42.4 ± 12.4 on a 0–100 scale (range 9.4–78.2) and correlated strongly with the independent reference susceptibility index (Spearman $\rho = 0.76$). The 200-iteration Dirichlet-weighted bootstrap produced a mean 95% interval width of 23.3 index points, demonstrating the sensitivity of composite susceptibility estimates to component weighting (Fig 12). The index nevertheless reproduced the broad susceptibility structure of the reference formulation, supporting the internal coherence of the integrated framework.

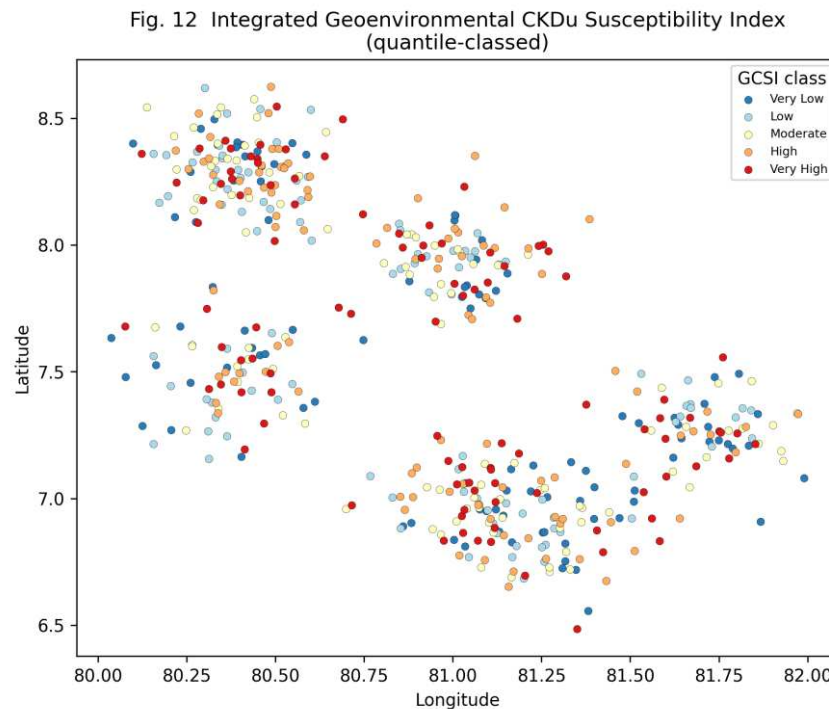


Fig. 12. Integrated Geoenvironmental CKDu Susceptibility Index (GCSI), quantile classed

The most consistent finding of this study is the predominance of soil Cd over Pb as the metal-associated environmental signal of CKDu susceptibility. Although the bivariate Cd–CKDu association was moderate (Spearman’s $\rho = 0.34$), its importance became considerably stronger when evaluated within multivariable and spatially explicit analytical frameworks. In the machine-learning analysis, Cd produced a mean absolute SHAP value of 0.787, almost five times that of Pb (0.165), indicating that variation in soil Cd contributed substantially more to the modelled CKDu prevalence signal than Pb after simultaneous consideration of terrain, hydrological, and soil variables. The GWR results further reinforce this interpretation, showing a consistently positive and substantially larger Cd coefficient across the study domain.

Importantly, this evidence should be interpreted as identifying Cd as the dominant metal-associated environmental variable, rather than establishing Cd as the sole causal agent of CKDu. The relatively modest predictive performance of the machine-learning models indicates that CKDu prevalence remains influenced by multiple interacting environmental determinants.

The association between Cd and CKDu becomes more meaningful when considered alongside the underlying soil geochemistry. Both Cd and Pb were positively associated with SOC, CEC, and clay content, indicating that their distribution is strongly influenced by soil retention processes. However, their geological patterns differed: Cd concentrations were highest within the Highland Complex, whereas Pb was highest within the Vijayan/Wanni Complex. Significant differences among geological units for both elements demonstrate that parent material contributes to their environmental distribution, although the relatively small between-unit differences indicate that geology alone cannot explain the observed metal variability.

This finding is particularly relevant to the question of Cd provenance. The geological differentiation of Cd, together with its integration into the principal soil-fertility and fine-texture gradient, indicates that geochemical background and lithological controls should not be excluded when interpreting Cd accumulation. Nevertheless,

the enrichment-factor results provide an important qualification: 99% of Cd observations were classified as anthropogenic-dominant, whereas Pb was predominantly geogenic-to-mixed. Thus, the present dataset does not support a simple geogenic-Cd/anthropogenic-Pb model. Rather, it suggests that Cd enrichment may reflect a combination of lithological background and external inputs, with the relative contribution requiring direct field-based source tracing.

The contrast between the geological signal and EF classification is scientifically informative rather than contradictory. Enrichment factors depend strongly on the selected background and normalizing element; consequently, apparent anthropogenic enrichment does not necessarily exclude an underlying geogenic contribution. The observed Cd pattern therefore supports a mixed conceptual model in which geological Cd availability may regulate the environmental pool, while agricultural or other external inputs can modify its magnitude and mobility. The moderate correlation between Cd and Pb ($\rho = 0.55$) indicates partial co-occurrence, but their stronger associations with SOC, CEC, and clay suggest that their accumulation is governed substantially by common soil physicochemical retention mechanisms rather than by a single source. The PCA reinforces this interpretation, with Cd, Pb, SOC, CEC, and clay loading positively on the principal component representing a fine-textured, organic-rich soil gradient. Such environments can retain trace metals through adsorption and surface-complexation processes, thereby influencing their persistence and potential transfer into soil–water and soil–plant exposure pathways.

Hydrological connectivity provides an additional dimension to this relationship. Distance to stream was the second most influential variable in the SHAP analysis, while the case-count model showed a negative association between distance to stream and CKDu occurrence. This suggests that proximity to surface-water pathways may contribute to the environmental connectivity through which soil-associated contaminants and other geochemical constituents can enter exposure pathways. The interaction detector further demonstrated that $\text{Pb} \times \text{Cd}$ jointly explained more spatial variation in CKDu prevalence ($q = 0.137$) than Cd alone ($q = 0.117$), indicating that metal exposure should not necessarily be conceptualized as independent single-contaminant processes. Within this framework, Cd represents the strongest metal-associated signal, while Pb contributes an additional, smaller component of the environmental exposure profile. The positive Pb–Cd interaction indicates that combined exposure may be more informative than either metal considered independently.

4. Conclusion

This study demonstrates that CKDu-related environmental analysis benefits from treating trace metals, soil properties and hydrological connectivity as an integrated geoenvironmental system. Across a 500-location Dry Zone-soil dataset, soil Cd consistently outperformed Pb as a statistical correlate of CKDu prevalence. Cd showed the stronger bivariate association ($\rho = 0.34$), the highest explainable machine-learning contribution (mean $|\text{SHAP}| = 0.787$), the largest Geographical Detector q -statistic (0.117), a positive near-global GWR coefficient and an IRR of 2.027 in the case-count model. The $\text{Pb} \times \text{Cd}$ interaction ($q = 0.137$) further indicates that joint exposure patterns contain information not captured by either metal alone. Cd was also embedded within a soil-retention gradient involving SOC, CEC and clay, while distance to stream, groundwater fluoride and rainfall contributed additional predictive information, supporting a multi-pathway environmental interpretation rather than a single-contaminant model.

Geological differentiation of Cd, especially its higher concentration within the Highland Complex, is compatible with a geochemical contribution and provides a testable basis for investigating parent-material weathering. However, the dataset's enrichment-factor analysis classifies 99% of Cd observations as anthropogenic-dominant. Pb, in contrast, is predominantly geogenic-to-mixed by EF and is also significantly differentiated among geological units. Therefore, the present results support Cd as the dominant statistical CKDu-associated metal but do not support the specific claim that Cd is demonstrably geogenic while Pb is demonstrably anthropogenic. Resolving that provenance question requires further field-based background determination, mineralogical characterization and source-tracing measurements.

The integrated GCSI showed strong agreement with the reference susceptibility index ($\rho = 0.76$), indicating that combining contaminant burden, hazard, hydrological connectivity and soil mobility can provide a coherent susceptibility framework. The analytical architecture is directly transferable to future field-collected data and can be used to test whether the Cd-dominant signal observed here persists under measured soil, water, exposure and clinical conditions. Such validation would provide a stronger basis for distinguishing geochemical background from anthropogenic enrichment and for identifying the environmental pathways most relevant to CKDu prevention.

5. References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Dayananda, N. R., & Liyanage, J. A. (2021). Quest to assess potentially nephrotoxic heavy metal contaminants in edible wild and commercial inland fish species and associated reservoir sediments; a study in a CKDu prevailed area, Sri Lanka. *Exposure and Health*. Advance online publication. <https://doi.org/10.1007/s12403-021-00403-x>
- Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Wiley.
- Hettiarachchi, G. M., & Pierzynski, G. M. (2004). Soil lead bioavailability and in situ remediation of lead-contaminated soils: A review. *Environmental Progress*, 23(1), 78–93. <https://doi.org/10.1002/ep.10004>
- Jayatilake, N., Mendis, S., Maheepala, P., Mehta, F. R., & CKDu National Research Project Team. (2013). Chronic kidney disease of uncertain aetiology: Prevalence and causative factors in a developing country. *BMC Nephrology*, 14, 180. <https://doi.org/10.1186/1471-2369-14-180>
- Kulathunga, M. R. D. L., Ayanka Wijayawardena, M. A., Naidu, R., & Wijeratne, A. W. (2019). Chronic kidney disease of unknown aetiology in Sri Lanka and the exposure to environmental chemicals: A review of literature. *Environmental Geochemistry and Health*, 41(5), 2329–2338. <https://doi.org/10.1007/s10653-019-00264-z>
- Liyanage, D. N. D., Diyabalanage, S., Dunuweera, S. P., Rajapakse, S., Rajapakse, R. M. G., & Chandrajith, R. (2022). Significance of Mg-hardness and fluoride in drinking water on chronic kidney disease of unknown etiology in Monaragala, Sri Lanka. *Environmental Research*, 203, 111779. <https://doi.org/10.1016/j.envres.2021.111779>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- McLaughlin, M. J., & Singh, B. R. (Eds.). (1999). *Cadmium in soils and plants*. Springer. <https://doi.org/10.1007/978-94-011-4473-5>
- Rajapakse, S., Shivanthan, M. C., & Selvarajah, M. (2016). Chronic kidney disease of unknown etiology in Sri Lanka. *International Journal of Occupational and Environmental Health*, 22(3), 259–264. <https://doi.org/10.1080/10773525.2016.1203097>
- Wang, J. F., & Xu, C. D. (2017). Geodetector: Principle and prospective. *Acta Geographica Sinica*, 72(1), 116–134. <https://doi.org/10.11821/dlxb201701010>

Data Availability: The data used to support the results of this research are shown in the manuscript and available from the corresponding author upon request.

Declarations

Conflict of Interest: The authors declare that there is not any conflict of interest regarding the publication of this manuscript. In addition, ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Author Contributions: Nalika R. Dayananda conducted Conceptualization, methodology, data analysis & investigations, and original manuscript preparation. Janitha A. Liyanage involved in Funding acquisition, resources, supervision and manuscript review.

Consent for Publication: All authors agree to the publication.

642 **Life Science Reporting:** No life science threat was practiced in this research.

643 **Funding:** Authors received no funding for the research