Journal of Geographic Information System, 2020, 12, 597-619
https://www.scirp.org/journal/jgis
ISSN Online: 2151-1969
ISSN Print: 2151-1950

Modeling the Spatial Distribution of Soil Heavy Metals Using Random Forest Model

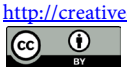
—A Case Study of Nairobi and Thirirka Rivers' Confluence

Evans Omondi, Mark Boitt

Department of Geomatic Engineering and Geospatial Information Systems, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: evan.omondi@gmail.com, mboitt@jkuat.ac.ke

How to cite this paper: Omondi, E. and Boitt, M. (2020) Modeling the Spatial Distribution of Soil Heavy Metals Using Random Forest Model. *Journal of Geographic Information System*, **12**, 597-619.
https://doi.org/10.4236/jgis.2020.126035

Received: October 3, 2020
Accepted: November 20, 2020
Published: November 23, 2020



Abstract

Modeling the spatial distribution of soil heavy metals is important in determining the safety of contaminated soils for agricultural use. This study utilized 60 topsoil samples (0 - 30 cm), multispectral images (Sentinel-2), spectral indices, and ancillary data to model the spatial distribution of heavy metals in the soils along the Nairobi River. The model was generated using the Random Forest package in R. Using R^2 to assess the prediction accuracy, the Random Forest model generated satisfactory results for all the elements. It also ranked the variables in order of their importance in the overall prediction. Spectral indices were the most important variables within the rankings. From the predicted topsoil maps, there were high concentrations of Cadmium on the easterly end of the river. Cadmium is an impurity in detergents, and this section is in close proximity to the Nairobi water sewerage plant, which could be a direct source of Cadmium. Some farms had Zinc levels which were above the World Health Organization recommended limit. The Random Forest model performed satisfactorily. However, the predictions can be improved further if the spatial resolutions of the various variables are increased and through the addition of more predictor variables.

Keywords

Random Forest, Sentinel 2, Heavy Metals, Spectral Indices, Spatial Modeling

1. Introduction

Efficient waste management is essential for improving the quality of living and sustainability of a country. However, this remains a challenge to developing countries as it's often expensive to execute. Managing an efficient waste man-

agement system requires efficient integrated systems that are sustainable and socially supported [1] [2].

In Kenya, waste disposal poses a great challenge, especially to the urban centers, including the country's capital city Nairobi, the coastal city of Mombasa, and Kisumu. Poor waste disposal in these metropolitan areas has been attributed to urbanization, rapid population growth due to rural-urban migration, sprawling of slum areas, lack of proper dumpsite areas, and a long-term lapse in the enforcement of urban planning and environmental laws [3]. Nairobi records the highest tonnage of garbage generated among these cities, approximately 2977 tons daily, of which 774 t/day (26%) remains uncollected [4]. Some of the waste generated are significant sources of heavy metal contamination to the environment, particularly industrial and electronic wastes (17,350 tons annually) [1].

Heavy metals like Manganese (Mn), Zinc (Zn), Copper (Cu), and Iron (Fe) are essential micronutrients for the growth and development of plants and the human body. While other elements like Cadmium (Cd), Lead (Pb), and Chromium (Cr) have no known benefits to both human and plant physiological processes. These toxic metals have been linked to several health problems in humans, such as hallucinations, diarrhea with blood, abdominal pain, dermatitis, liver and kidney failure, lung disease, hepatic damage, mutagenic, teratogenic, and carcinogenic effects [5].

Studies have shown the importance of remote sensing in detecting heavy metal stress [6] [7]. Heavy metals have adverse effects on plants. They inhibit physiological and metabolic processes like photosynthesis by reducing the canopy chlorophyll content, thus affecting growth and productivity [8]. Therefore, chlorophyll content acts as an important bio-indicator of a plants' health status [9] [10]. Changes in chlorophyll content can alter the spectral reflectance of both the near-infrared and visible portions of the electromagnetic spectrum. Therefore, the red-edge region is closely associated with chlorophyll content in plants [11] [12]. Additionally, some studies have shown that it can be an important indicator of heavy metal stress levels in plants [13] [14]. This can be exploited together with other variables to model heavy metal contamination in soils.

Machine learning approaches like cubist, Principle Component Analysis, and Support Vector machine have been used to map heavy metal contamination in soil, and they have performed reasonably well [15] [16] [17]. However, researchers are always looking for machine learning algorithms, additional variables, and sensors that can provide higher prediction accuracies [15] [18].

The random forest classifier is a combination of multiple decision trees, where each tree is built from a random vector independently sampled from the input vector, and each decision tree casts a vote to find out the most popular class to assign the input vector [19]. In recent years it has been used for many different applications including, image classification [20] [21] [22], vegetation mapping [23] [24], however very few studies have been done focusing on the use of Random forest for the spectral analysis of soil, and specifically on heavy metals in

soil [25] [26] [27] [28]. Nonetheless, the studies that focused on heavy metals did not consider variables at a higher spatial resolution and additional dependent variables such as HMSSI, SAVI, and WDVI.

In Kenya, the rivers in Nairobi County are too polluted with heavy metals for human use [29]. Incidentally, recent google earth imagery shows the establishment of peri-urban farms on the banks of the polluted river. And to our knowledge, there is no information in existence regarding the spatial distribution of soil heavy metals in the banks along the Nairobi River. Therefore, it is paramount to establish the pollution levels of heavy metals in the soils used to grow these crops.

This study's novelty is to utilize the red-edge and optical bands from multi-temporal Sentinel 2 satellite imageries, with a temporal resolution of 10 days, Random Forest, and ancillary data to model the distribution of heavy metals in the soils used to cultivate the peri-urban farms. Therefore, this research aims to map the distribution of toxic heavy metals (Cd, Pb, and Zn) in soils irrigated using water from the polluted Nairobi River. This is achievable by first conducting a random soil sampling and laboratory analysis for heavy metal contamination along the Nairobi River riparian, followed by determining the performance of the environmental parameters and spectral indices in the predictions. Lastly, modeling and validating the distribution of heavy metal contamination in the soils along the river.

2. Materials and Methods

2.1. Study Area

The study area shown in **Figure 1** is located at the confluence of the Nairobi and Thiririka rivers (1˚11'57.36"S - 37˚07'09.23"E). It shares a border between Ruiru sub-county in Kiambu County and Kasarani sub-county in Nairobi County. It covers a total area of approximately 17 km^2.

The climate is described as warm and temperate. It lies at an altitude of 1544 m above sea level and receives annual average precipitation of 752 mm.

According to the 2019 Census, Kasarani sub-county had a population of 780,656, and Ruiru sub-county had a population of 490,120 [30]. **Figure 2** shows the decadal increase in population around our study area.

The main socio-economic activity in the area is farming, and over the past two decades, there has been a growth of peri-urban agriculture along the river channel. Therefore, we chose an area along the river channel with a continuous and a high concentration of peri-urban farms as our study area.

2.2. Data and Methodology

Figure 3 gives a graphical illustration of the methodology used in this study.

2.2.1. Soil Sampling and Laboratory Analysis

60 soil samples were collected from 30 cm deep holes (A horizon) along the

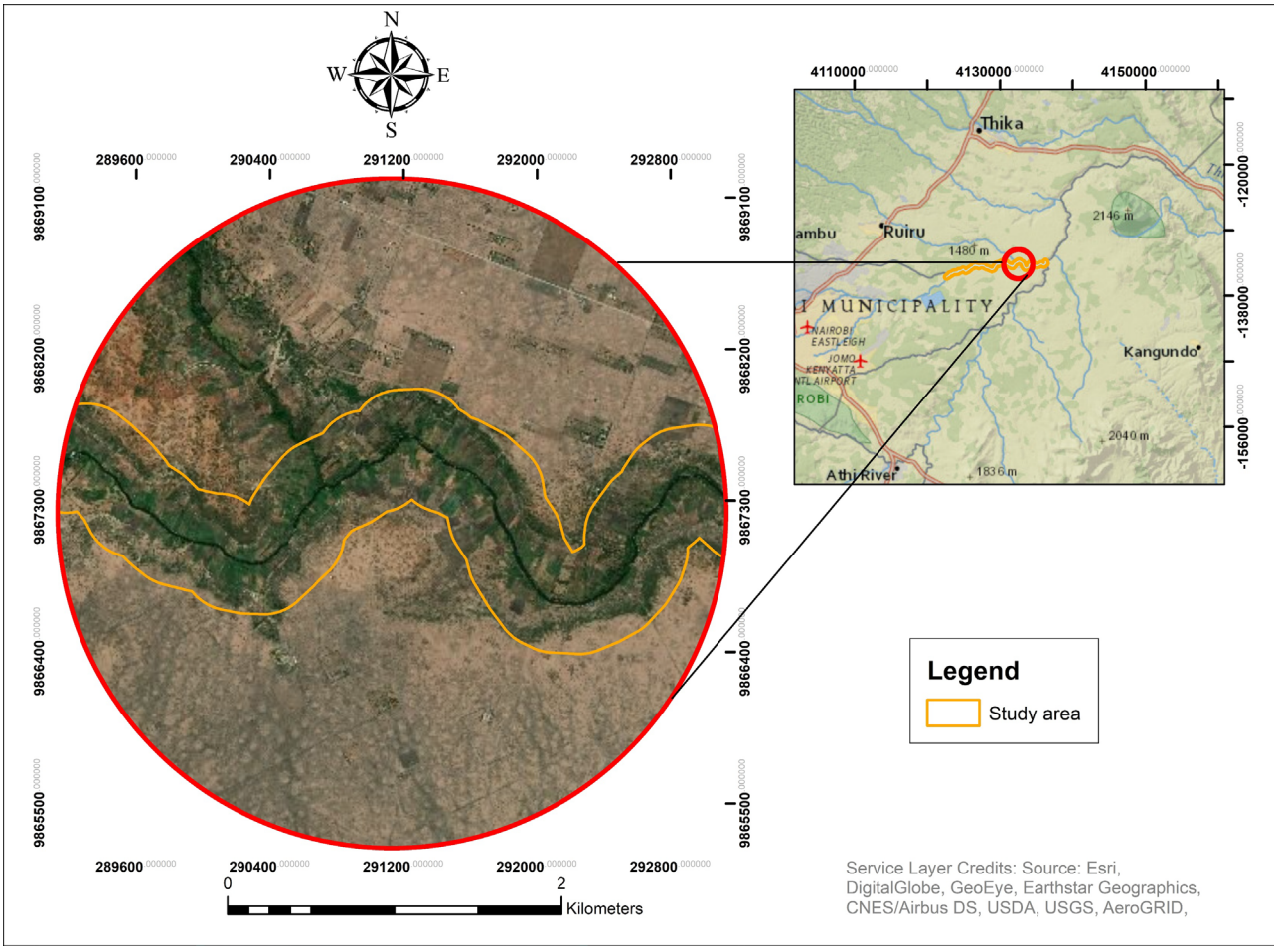


Figure 1. Study area.

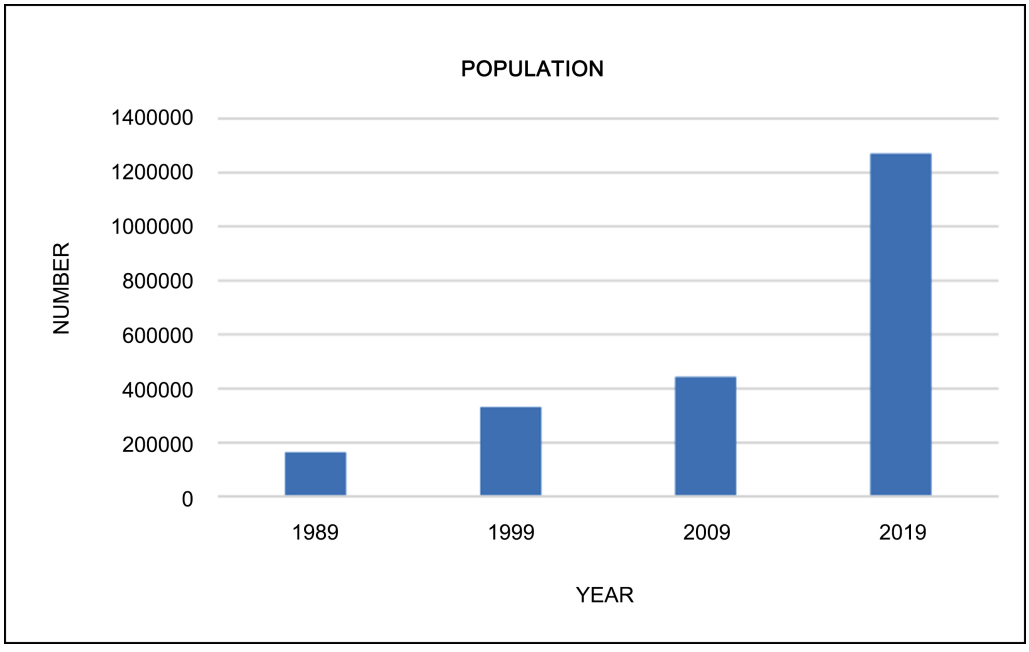


Figure 2. Study area population increase against time.

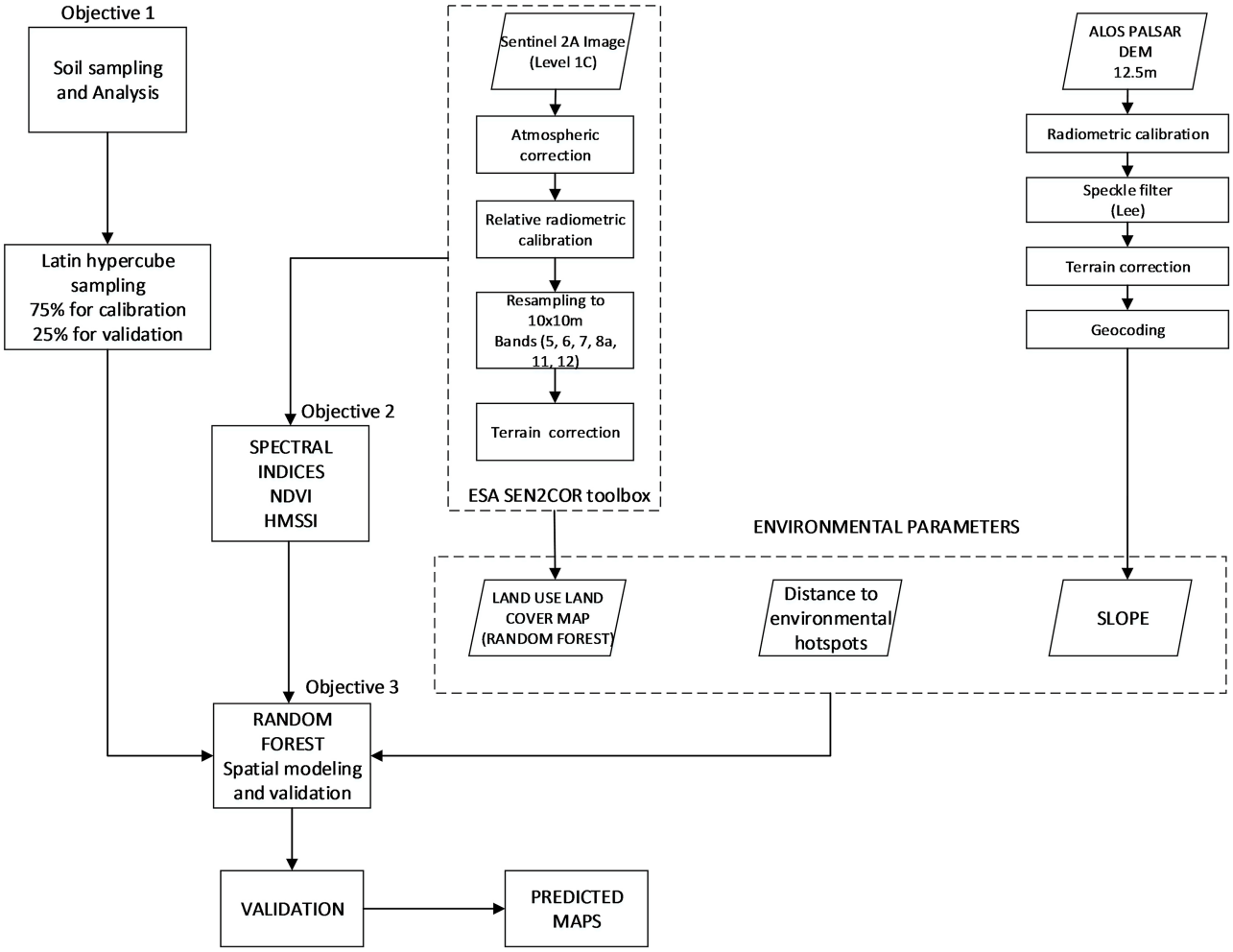


Figure 3. Methods flowchart.

Nairobi River riparian at Ruai, per the European Soil sampling guidelines for pollution studies [31]. Sampling locations covered all the affected peri-urban farms within our study area. The collection points' geographic coordinates were recorded using a 5 m accuracy hand-held Garmin GPS, as shown in **Figure 4**.

The samples were collected using an auger and taken to the Jomo Kenyatta University of Agriculture and Technology for analysis. They were air-dried for three days under room temperature. A 2 mm polyethylene sieve was used to sieve the soil. They were later analyzed for the concentrations of Lead, Zinc, and Cadmium.

0.5 g of each sample was added into a pre-cleaned Pyrex test-tube. 8ml of concentrated hydrochloric acid and 3 ml of concentrated perchloric acid were added. The mixture was heated in an aluminum block at 200˚C for a period of 3 hrs until it was dry. After the test-tubes cooled down, 5% HNO3 was added and then heated at 70˚C for 1 hr. with occasional mixing.

After cooling down, the mixture was decanted into a polyethylene tube and centrifuged at 3500 rpm for 10 min. All of the elements' concentrations were determined using an inductively-coupled plasma-atomic emission spectrometry.

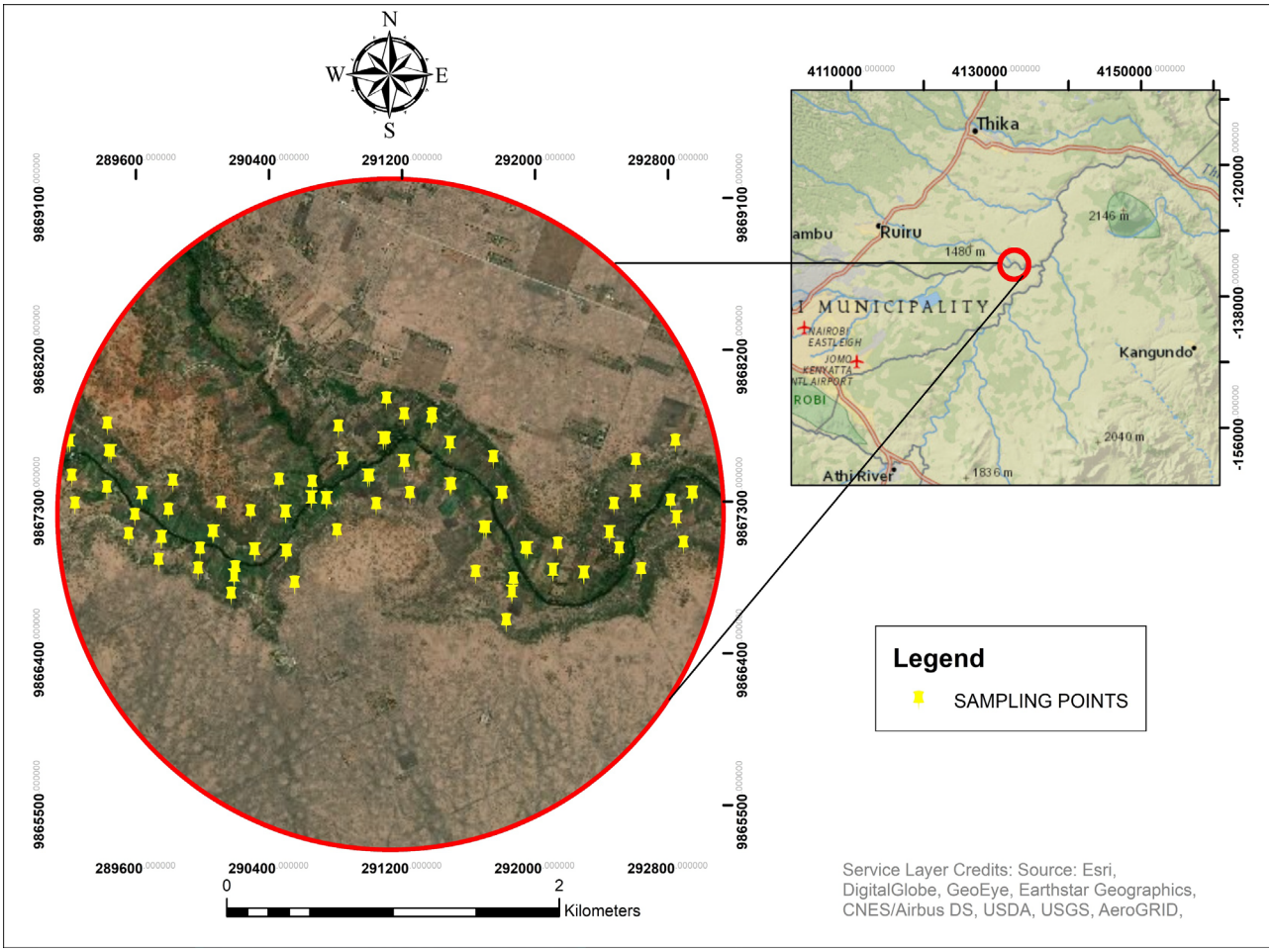


Figure 4. Sampling point locations.

2.2.2. Sentinel 2

Multi-temporal sentinel 2A (Level 1C) imagery from January 2019 to December 2019, with a spatial resolution of 10 m and 20 m, was acquired from the European Space Agency website (ESA https://earthexplorer.usgs.gov/).

All the products were radiometrically and geometrically corrected using the sent2cor tool in SNAP and projected to the WGS 1984/UTM zone 37˚S map projection.

In R version 3.6.1, the images were used to extract predictor variables (spectral indices, Land-use land cover map, and spectral bands) needed in the Random Forest model. Table 1 describes the covariate obtained from Sentinel 2.

2.2.3. ALOS PALSAR Pre-Processing

Radiometric calibration is the first important step for Alos Palsar pre-processing. It converts the signal number values to backscatter in sigma naught. Using the Lee filtering method, Speckle filtering was done to reduce the salt and pepper noise caused by speckle noise.

The speckle filtering was followed by terrain correction. Finally, geocoding of the image was done using ground control points obtained from 1:50,000 topo-

graphic maps from the Ministry of Lands and Physical Planning. This was to ensure that the image was properly georeferenced.

2.2.4. Environmental Parameters

Environmental parameters have proven useful as ancillary data in improving the accuracy of predicting the distribution of pollutants and other soil attributes [15] [32] [33]. In this study, we utilized selected environmental parameters (Anthropogenic parameters like distance to environmental hotspots, geomorphology data like slope, and a land-use land cover map) to predict the distribution of heavy metal contamination in the soils along Nairobi River (Table 2).

The choice of these parameters was informed by our field observations and their use in other predictive models performed in almost similar environmental conditions.

1) Digital Elevation Model

An ALOS PALSAR Digital Elevation Model (D.E.M.) with a spatial resolution of 12.5 m was acquired from the Japanese Space Agency website (https://www.eorc.jaxa.jp/ALOS/en/about/palsar.htm). The tiles were merged using the Mosaic tool in ArcMap version 10.6. A map of the variations in elevation is shown in Figure 5. The DEM was used later used to generate the slope variable used to predict the model.

2) Distance to environmental hotspots

The distance to the closest environmental hotspots (*i.e.*, Industries, roads, mines, dumpsites, and water treatment plants) was generated using the Euclidean distance tool and Extract multi values to points tool in ArcMap 10.6. A map showing the distance to hotspots is shown in Figure 6.

3) Land cover map

Land use and a land cover map (Figure 7) was generated from a 10 m spatial resolution Sentinel 2 A image, with a <5% cloud cover. The image classification was done using the Random Forest package in R version 3.6.1 software.

Table 1. Sentinel 2 data.

Covariate	Representative digital data	Source of data
Spectral Indices	NDVI: Bands (8 − 4)/(8 + 4) HMSSI: Bands [(7/5) − 1]/[(4 − 2)/6]	Sentinel 2 imagery

Table 2. Environmental parameters used to predict the distribution of soil heavy metals along the Nairobi River.

Environmental covariates	Representative digital data	Source of data
Digital Elevation Model (m)	Slope	12.5 m ALOS PALSAR
Distance to environmental hotspots (m)	Hotspots (Sewer treatment plants, dumpsites, Industries, and roads)	Google earth pro
Land-use land-cover map	Bands 4, 3 and 2	10 m Sentinel 2

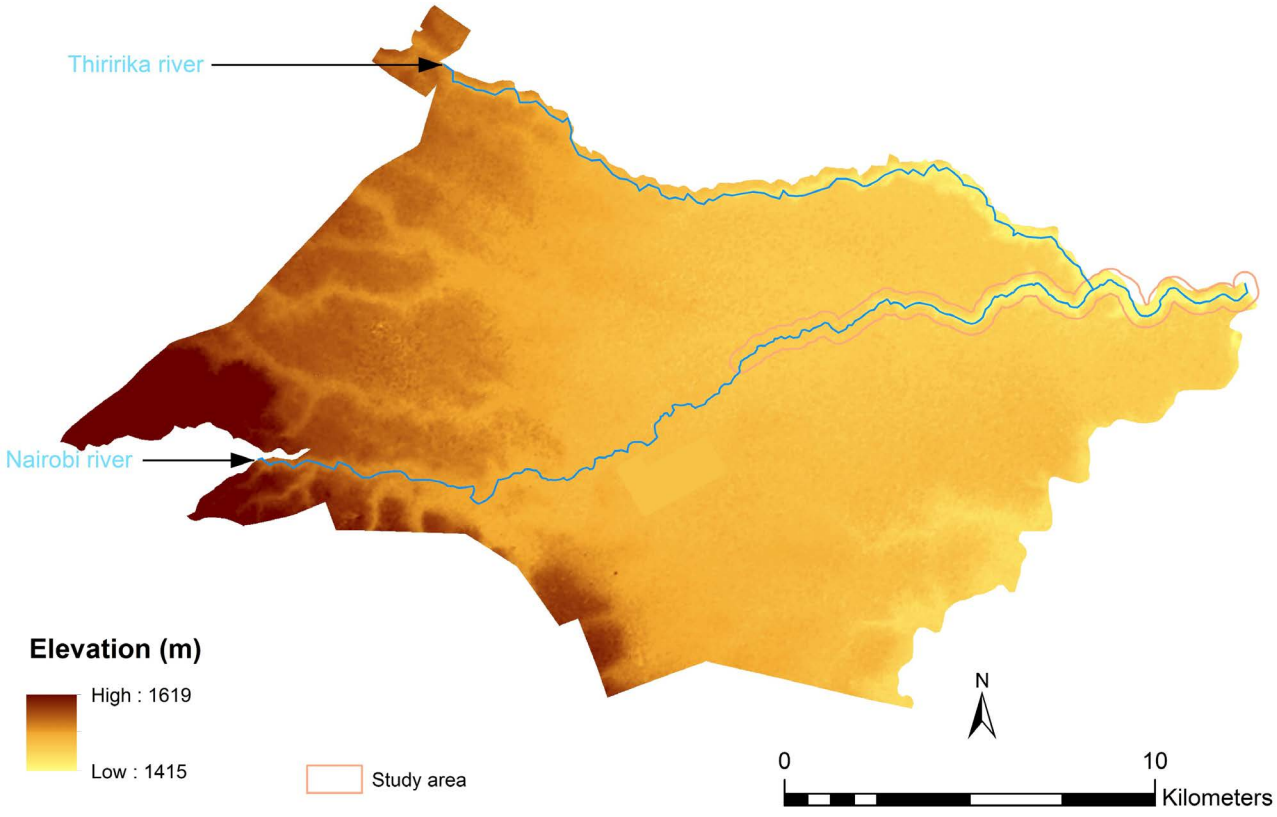


Figure 5. A 12.5 m ALOS PALSAR DEM showing the topography of the study area.

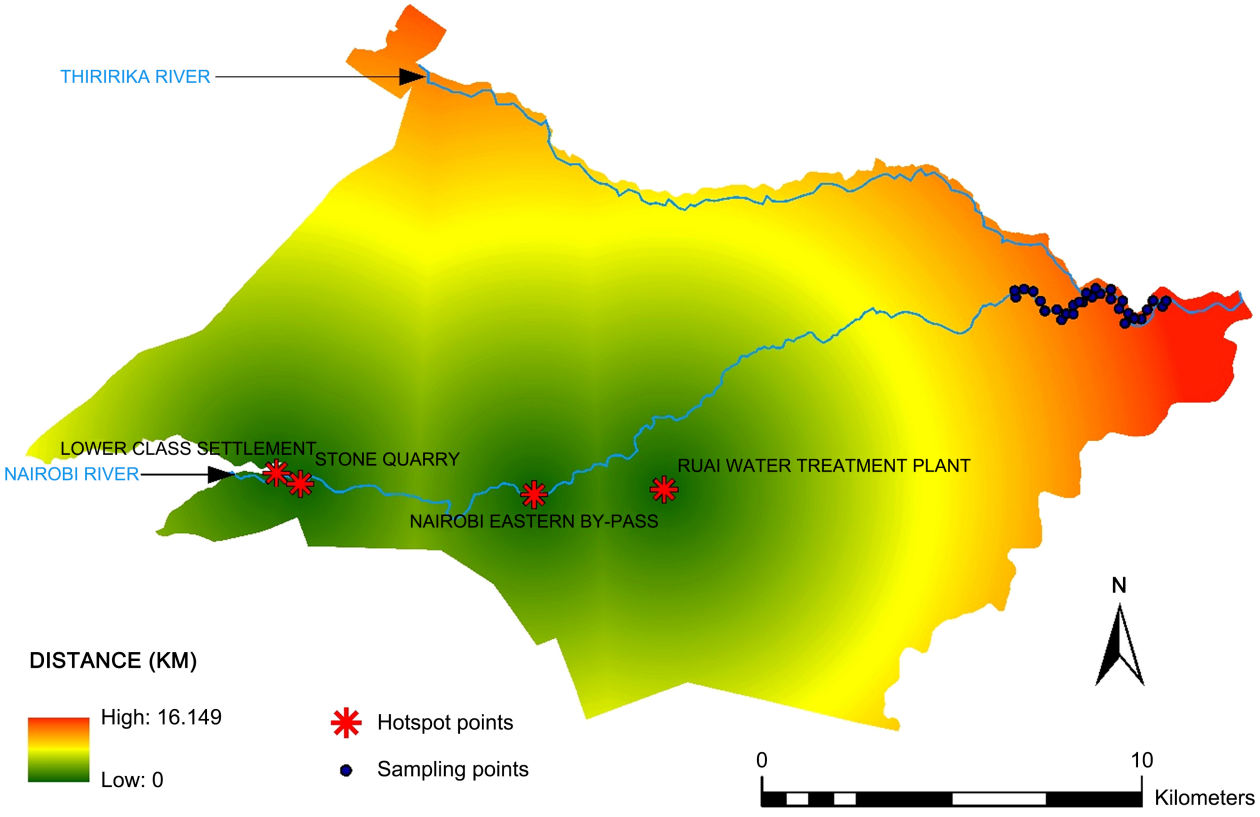
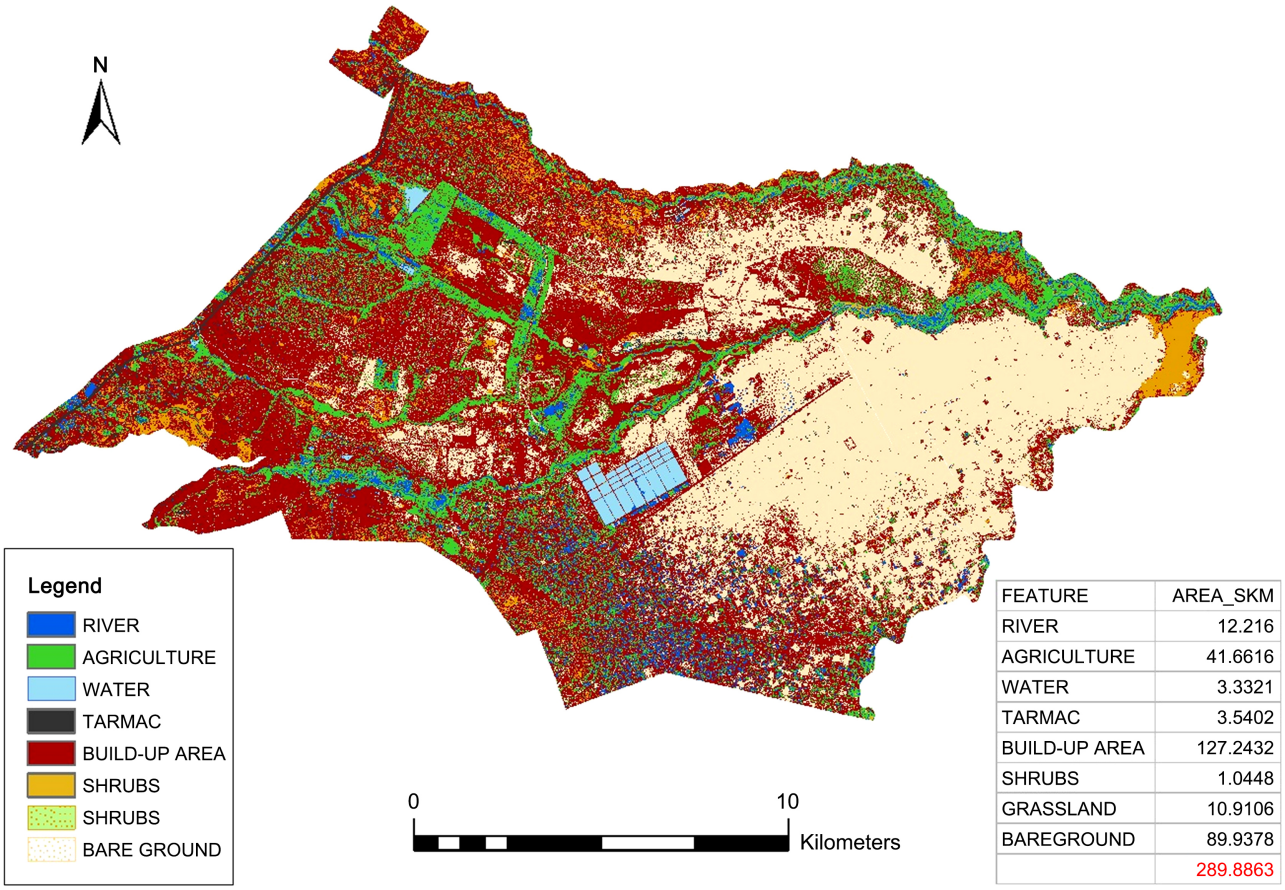


Figure 6. Distance to environmental hotspots.



FEATURE	AREA_SKM
RIVER	12.216
AGRICULTURE	41.6616
WATER	3.3321
TARMAC	3.5402
BUILD-UP AREA	127.2432
SHRUBS	1.0448
GRASSLAND	10.9106
BAREGROUND	89.9378
	289.8863

Figure 7. Classified image of the study area.

2.3. Spectral Indices and Remote Sensing Images

The health of vegetation can provide important ancillary information when modeling heavy metals' spatial distribution in soil. Multi-temporal Sentinel 2 images of Nairobi County were acquired from the USGS Earth Explorer for a period ranging from January 2019 to December 2019. This study used a higher spatial resolution optical sensor, Sentinel 2 (10 m) to improve the prediction accuracy.

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + \mathrm{R}} \tag{1}$$

$$\mathrm{HMSSI} = \frac{\mathrm{CIred\text{-}edge}}{\mathrm{PSRI}} \tag{2}$$

$$\mathrm{PSRI} = \frac{\mathrm{R}_{680} - \mathrm{R}_{500}}{\mathrm{R}_{750}} \tag{3}$$

$$\mathrm{CIred\text{-}edge} = \left(\frac{\mathrm{R}_{783}}{\mathrm{R}_{705}}\right) - 1 \tag{4}$$

$$\mathrm{SAVI} = \frac{(\mathrm{B8} - \mathrm{B4})}{\mathrm{B8} + \mathrm{B4} + \mathrm{L}} \times (1 + \mathrm{L}) \tag{5}$$

$$\mathrm{WDVI} = \mathrm{B8} \times \mathrm{B4} \tag{6}$$

where L = 1.

A study by [15] has shown that the health of vegetation can be an important indicator of heavy metal contamination in soil, and it could also improve the accuracy of a soil heavy metal prediction model. The indices used in this study were derived from Sentinel-2, and they include Normalized Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), Weighted Difference Vegetation Index (WDVI) [34] and the novel Heavy metal stress-sensitive index (HMSSI) by [35]. HMSSI was developed to improve the accuracy of detecting heavy metal stress in Chinese rice fields using multi-temporal sentinel-2 images.

In calculating HMSSI, two red-edge spectral indices, namely plant senescence reflectance Index (PSRI) and Red-edge chlorophyll Index (CIred-edge), were used. Low (CIred-edge) index values indicate low chlorophyll and high stress in vegetation, while an increase in PSRI indicates an increase in canopy stress [35].

2.4. Spatial Modeling and Validation

This study created a prediction model for each element (Cd, Pb, and Zn) using the Random Forest package (version 4.6 - 14) in R version 3.6.1 [36]. Random forest operates by constructing multiple decision trees. Each tree is built from a random vector independently sampled from the input vector, and each decision tree casts a vote to find out the most popular class to assign the input vector [19].

The Random forest has some advantages over other classification and regression tree algorithms. In addition to eliminating bias, it reduces the variance in the predictions usually associated with tree-based approaches by growing the trees further, then averaging their predictions [37]. Another important advantage of Random Forest over other machine learning algorithms is when the training data is small; it has the ability to capture complex and non-linear relationships between predictors and the outcome [37]. It's also important to note that with random forest, the accuracy of prediction increases with an increased number of predictor variables [38].

In this study, the various model outputs were validated using the "out of bag" (OOB) testing. The OOB samples are the observations not included in the model, and since they are not used to predict the model, they are used to test it.

To test prediction quality, 75% of the predictor variables were used for calibration, while the remaining 25% was used for validation. The datasets were chosen using a Latin hypercube sampling to ensure that both the validation and calibration datasets were appropriately represented.

After selecting the training and test sets, we fitted the random forest model using default parameters [36]. The number of trees (ntree) was set at 500. The model was then fine-tuned by changing the number of variables randomly sampled at each stage (mtry) to 13. Evaluation of the model was done using Root Mean Square Error (RMSE), Bias, and coefficient of determination (R^2), which were calculated using the Out of Bag Error Estimation. In this, 25% of predictor

variables were used to validate the trees.

$$\text{RMSError} = \sqrt{1 - r^2}\text{SDy} \tag{7}$$

where SDy is the standard deviation of y

$$\text{Bias} = E(H) - \theta \tag{8}$$

where H is the expected values of the estimator less the values θ being estimated

$$R^2 = \text{MSS}/\text{TSS} = (\text{TSS} - \text{RSS})/\text{TSS} \tag{9}$$

where MSS is the model sum of squares, and TSS is the total sum of squares associated with the outcome variable.

3. Results and Discussion

3.1. Soil Sampling and Analysis

A total of 60 samples were collected from selected points within the study area and later analyzed for heavy metals in a lab. The metals' different concentrations are shown in **Figure 8**, **Figure 9**, and **Figure 10**. **Table 3** shows the safety thresholds for heavy metals recommended by WHO, FAO, and USEPA.

The results indicate that for Zinc heavy metal, 17 out of the 60 soil samples collected exceeded the WHO/FAO permissible limits. According to [41], Zinc concentrations in the study area can be presented by anthropogenic activities like waste combustion at Dandora dumpsite, Steel processing activities at Nairobi Industrial area, and stone quarrying.

Additionally, 12/60 Lead samples exceeded the WHO/FAO/USEPA permissible safety limits. Lead is mainly used in the manufacture of lead storage batteries. In this case, it's highly likely that its presence results from leachates and run-offs from electronic waste components at the Dandora dumpsite. Lead poisoning occurs when there is direct ingestion of Lead contaminated soil. Vegetables produced in soils with less than 300 ppm of Pb contamination are considered safe for consumption. The risk increases with an increase in the concentration in soil.

Table 3. Heavy metals safety limits. World Health Organization (WHO), Food and Agriculture Organization (FAO), United States Environmental Protection Agency (USEPA) [39] [40].

Samples	Standards	Zn	Pb	Cd
Water (mg/L)	WHO/FAO (2007)	2.0	5.0	0.01
	USEPA, 2010, USEPA., 2010	2.00	0.015	0.005
Soil (mg/kg)	WHO/FAO (2007)	300 - 600	250 - 500	3.0 - 6.0
	USEPA, 2010, USEPA., 2010	200	300	3.0
Plant (mg/kg)	WHO/FAO (2007)	60.0	5.0	0.2
	USEPA, 2010, USEPA., 2010	–	–	–

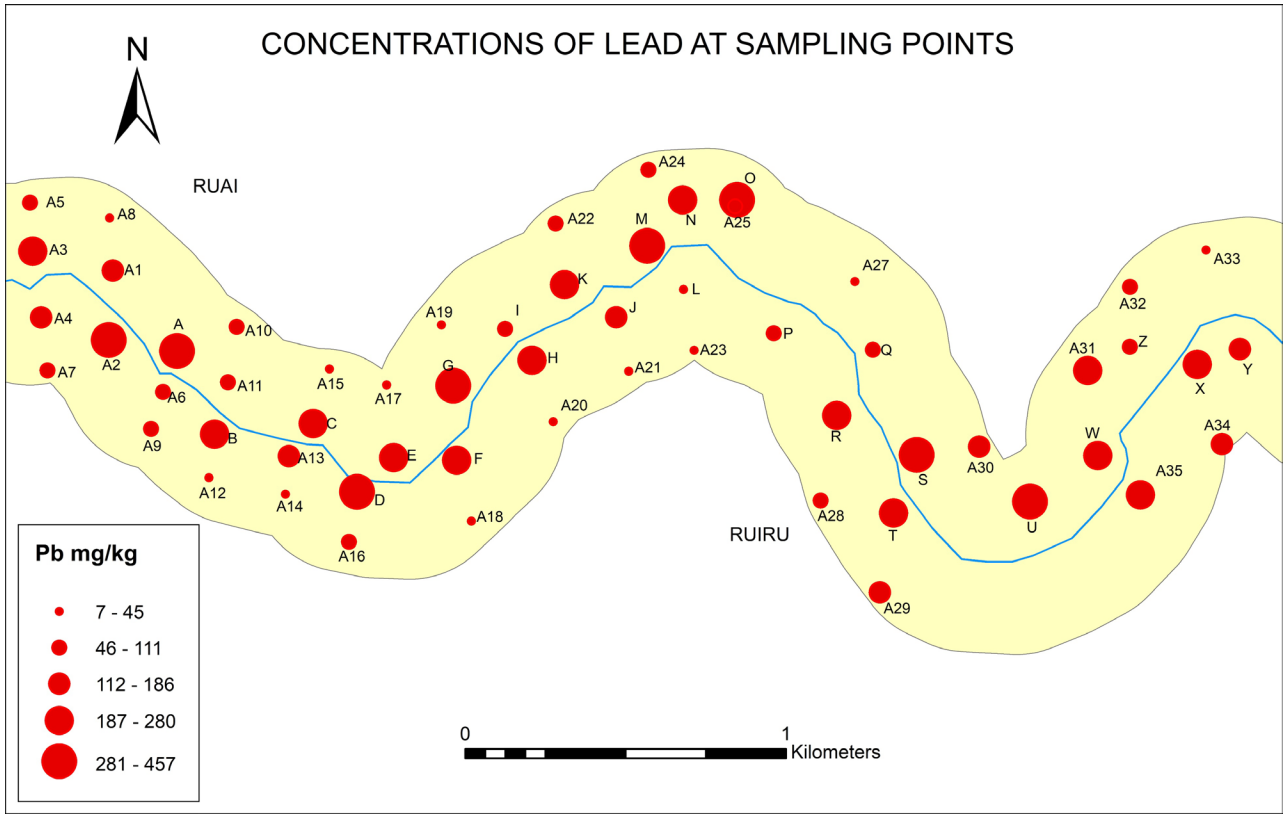


Figure 8. Lead concentrations at sampling points.

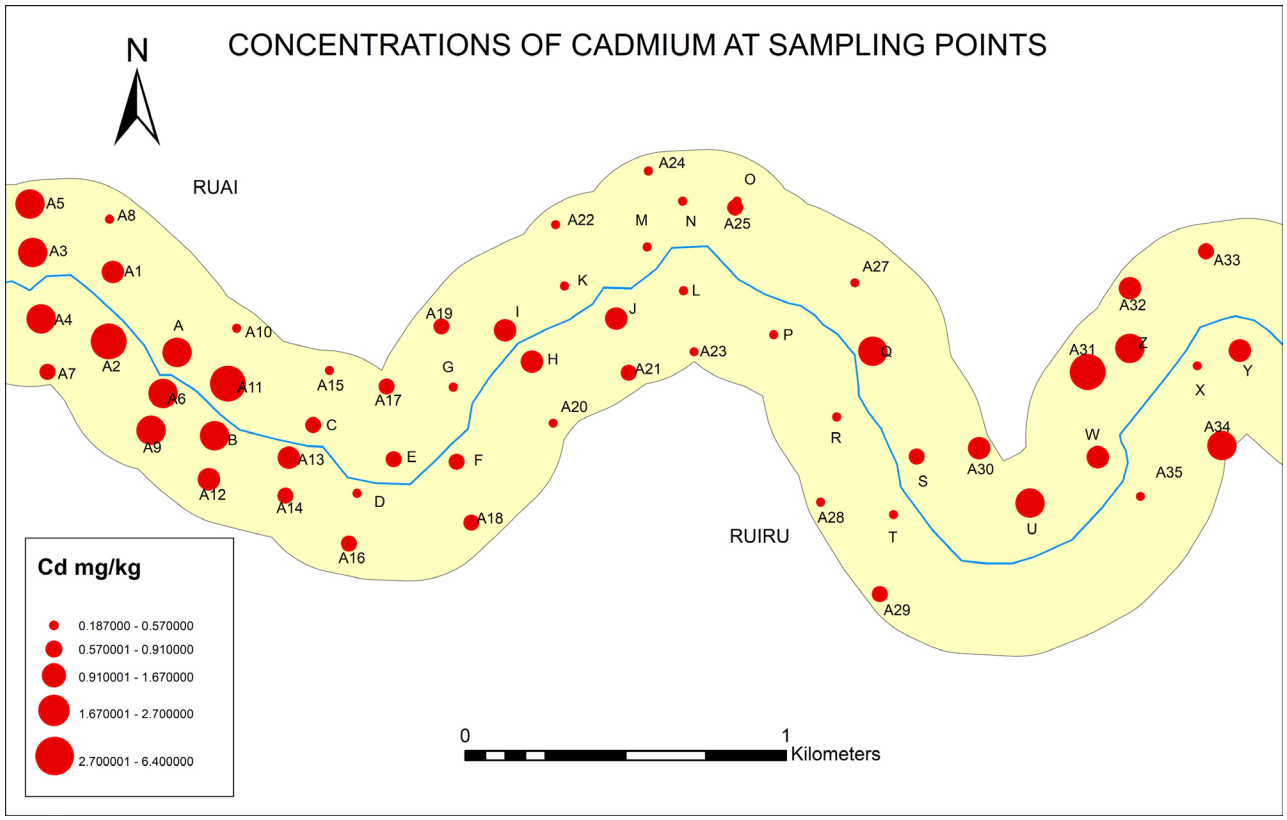


Figure 9. Cadmium concentrations at sampling points.

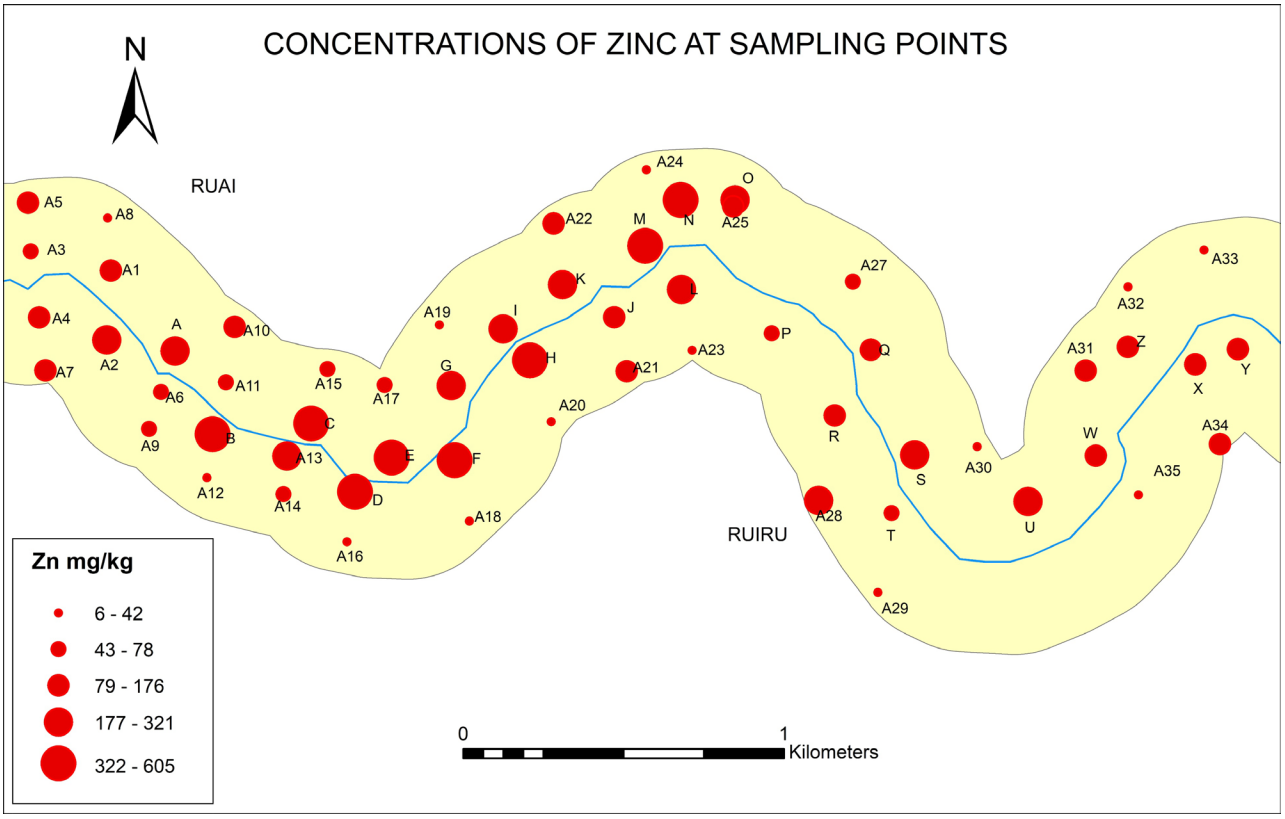


Figure 10. Zinc concentrations at sampling points.

High Cadmium levels were found. However, non-exceeded the WHO/FAO/USEPA permissible limits. Cadmium pollution is linked to industrial effluents and water treatment sludge. Agricultural inputs like pesticides and fertilizers also increase their total concentration in soils [42].

According to [41], the high Cadmium concentrations can be attributed to the fact that the Nairobi river passes through several environmental hotspots, like the Dandora dumpsite, Industrial area, Nairobi sewerage treatment plant, and high traffic networks like the busy Nairobi Eastern bypass.

The distinct variation in the heavy metals results from a combination of different anthropological activities like the application of agricultural inputs, discharge of effluents from Industries and a sewerage treatment plant, leachates, and run-offs from Dandora dumpsite into the river, and mining activities.

3.2. Statistical Analysis of Heavy Metal Data

One of the benefits of RF and other classification and regression tree algorithms is that they do not require the input data to be normalized. Therefore, no box plots and histograms were required for this study.

The general statistics for (75%) Calibration and (25%) Validation are shown in Table 4 and Table 5, respectively.

Table 6 displays Pearson's correlation between the three soil toxic metals. There was no significant correlation between any of the metals, implying that

they possibly did not come from the same source [15] [42].

The validation results of the spatial modeling for soil heavy metals contamination are shown in Table 7. Generally, the calibration model performed well with regards to the R^2 and RMSE estimates.

Our validation results for both Lead and Zinc had a higher correlation coefficient than [15] R^2 (Zn = 0.51) and R^2 (Pb = 0.53). This improvement could be attributed to the use of finer spatial resolution (10 m) for environmental predictors and spectral images in comparison to [15] at 30 m and [16] at 1 km. Additionally, the use of a high number of evenly distributed sample points in a smaller study area (17 km^2) also meant that the accuracy of prediction in our study was improved. On the other hand, the climatic difference between Kenya, Qatar [15], and Europe [16] could be a contributing factor to the difference in

Table 4. Calibration statistics dataset. 75% of the data. (Standard Deviation) SD.

Element (mg/kg)	n	Mean	Median	SD	Variance
Pb	45	217	221	95.82	9181.7
Cd	45	0.96	0.975	0.62	0.384
Zn	45	221	261	150	22,454.76

Table 5. Validation dataset statistics 25% of the data.

Element (mg/kg)	n	Mean	Median	SD	Variance
Pb	15	229.75	213.5	64	4074
Cd	15	1.2	0.82	0.654	0.427
Zn	15	259.5	186	141	19,905

Table 6. Pearson correlation coefficients between Lead, Cadmium and Zinc (n = 60).

	Pb	Zn	Cd
Pb	1	0.39*	−0.15
Zn		1	−0.26
Cd			1

Levels of significance *p < 0.05, **p < 0.01.

Table 7. Validation results for different soil heavy metals concentration.

	R^2	RMSE	Bias
Lead	0.8335727	0.4382932	−1.076752
Zinc	0.8309954	0.5139823	−3.516596
Cadmium	0.7882876	0.2719324	0.01085672

results. Nairobi has a warm and temperate climate; Qatar is a desert while Europe is generally temperate. The soil parent rock materials and anthropogenic activities are also different. For these reasons, the choice of variables for modeling was different for the different study areas.

3.3. Variable of Importance Usage by Random Forest

The variable of importance is measured based on the out of bag samples. These are observations not included in the Random Forest model. Also, they are based on a mean square error accuracy measure. The value is averaged over all trees [36].

From the output, it's evident that the model utilized all the variables for prediction but gave more emphasis to the most important ones.

For Lead (**Figure 11**), HMSSI and SAVI were dominant within the top ten important variables. HMSSI was the best performer while the land-use and land cover map also performed considerably well in the prediction.

For Zinc, 3 WDVI, 3 HMSSI, and 2 NDVI were selected among the top ten important predictors. At the same time, distance to environmental hotspots was in the 7th position (**Figure 12**).

The top ten important variables for Cadmium's prediction (**Figure 13**) included spectral indices (3 WDVI, 3 HMSSI, and 2 NDVI). The land-use and land-cover map also performed considerably well, occupying the 5th position.

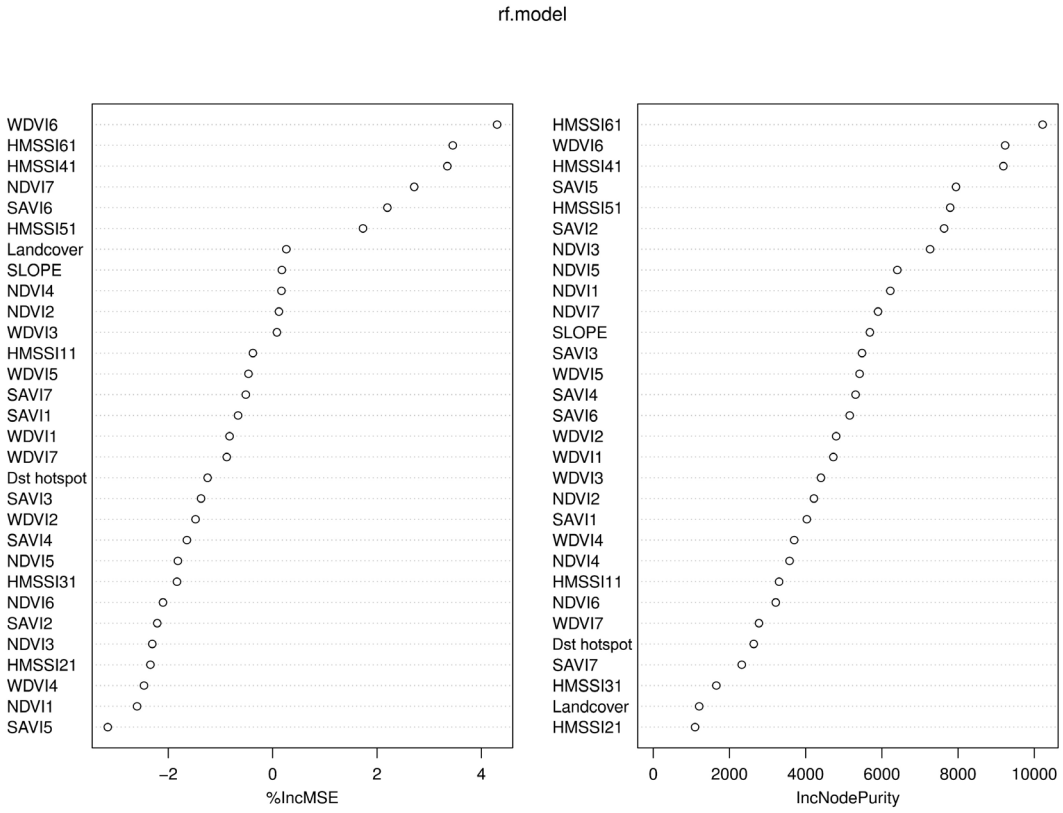


Figure 11. Ranking of predictors in Lead from Random Forest model fitting.

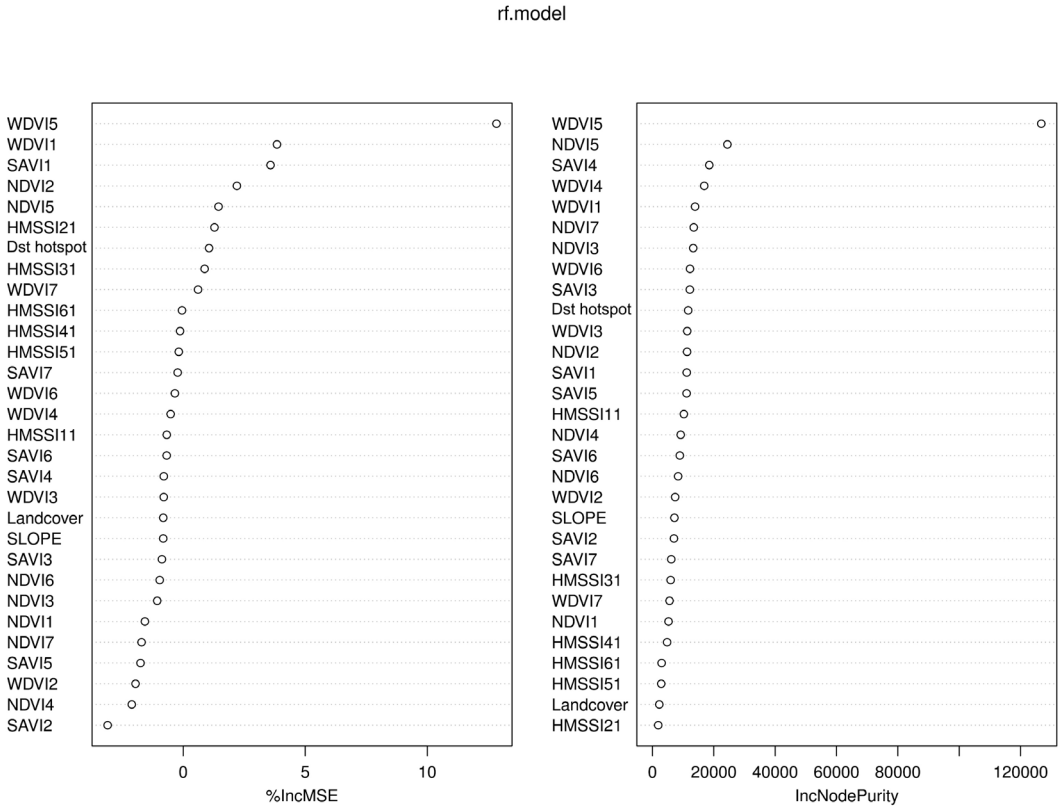


Figure 12. Ranking of predictors in Zinc from Random Forest model fitting.

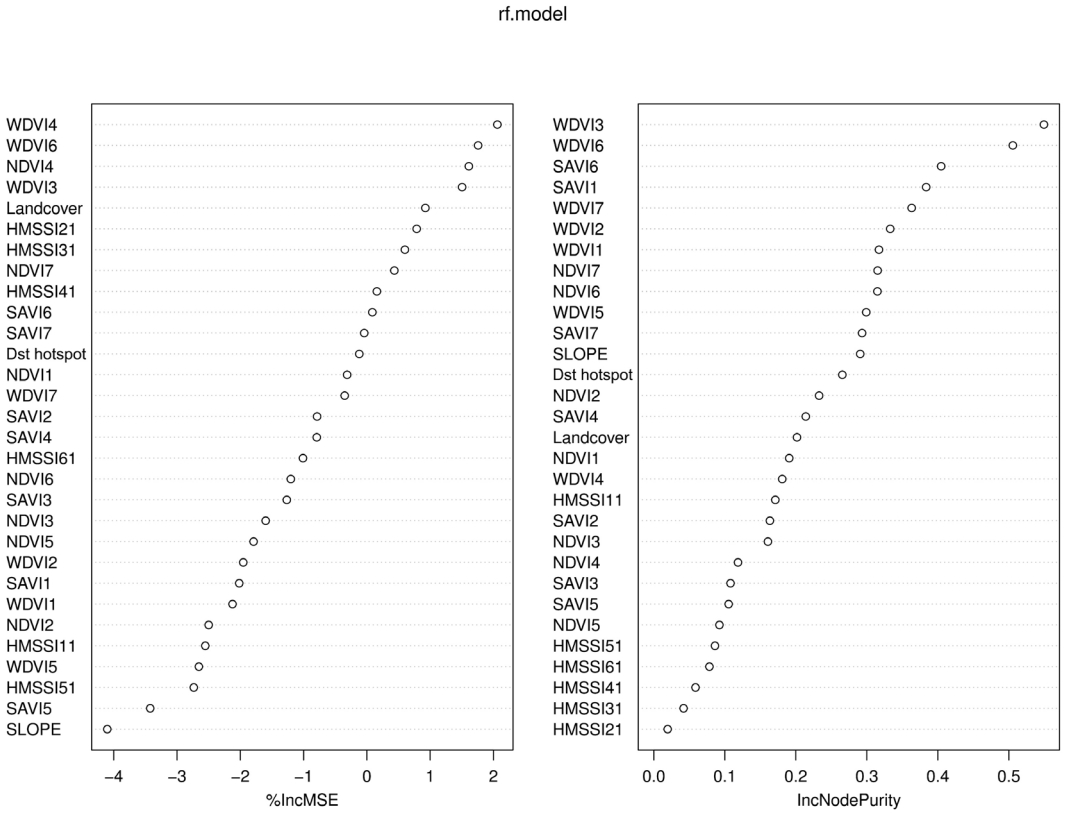


Figure 13. Ranking of predictors in Cadmium from Random Forest model fitting.

In the prediction, the random forest model included the vegetation spectral indices (NDVI, SAVI, WDVI, and HMSSI) within the top ten important variables for all the heavy metals. This indicates their importance in predicting the distribution of heavy metals, which is in line with previous studies [15] [16].

The distance to the environmental hotspots variable was high in the ranking of all the metals, implying that human activities highly influence their concentrations.

The land-use and land-cover map performed reasonably well in the predictions despite being a new additional variable absent in previous research [15] [16].

3.4. Vegetation Indices for Toxic Metals Prediction

In order to improve the predictive capability of heavy metals in soil, spectral indices were calculated for all the Sentinel 2 images. Four spectral indices HMSSI, SAVI, WDVI, and NDVI were derived.

The four variables were present within the top ten important variables in the prediction of all the heavy metals. In the prediction of Zinc, NDVI was the dominant index, followed by SAVI and HMSSI. For Lead prediction, NDVI was the most dominant index within the top 10 important variables. Of the top 10 important variables for Cadmium prediction, NDVI and HMSSI contributed three indices, each within the top important variables, followed by SAVI and WDVI.

3.5. Predicted Maps for Toxic Metals in Soil

Three predicted maps are shown in **Figure 14**, **Figure 15** and **Figure 16**. With a 300 m buffer along the rivers' riparian, we can easily tell how the three different heavy metals are distributed within the peri-urban farms.

There is a high concentration of Cadmium on the easterly end of the river. This can be linked to the study area's proximity to the Nairobi water treatment plant at Ruai. Its abundance in the water treatment plant is because, besides other sources of Cadmium, it can also occur as an impurity in detergents [41]. The water treatment plant serves a population of 4,397,073 residents, which dramatically increases the probability of high Cadmium concentrations within the wastewater. There is also a possibility that the treated wastewater being discharged into the river is still toxic.

Another possible source of Cadmium in the soil is the application of phosphate fertilizers and pesticides on the peri-urban farms. Additionally, the disposal of Industrial waste upstream, as the river passes through an industrial area, also increases the total concentration of Cadmium in soil.

Some farms have a high concentration of Lead in their soil. However, these soils are safe for agricultural production because they have not exceeded the WHO/FAO/USEPA permissible limits. Additionally, plants do not uptake Lead into their system unless the concentration levels rise above 300 ppm. Further, a study done by [41] indicates that Lead does not readily accumulate in the fruiting parts of a plant.

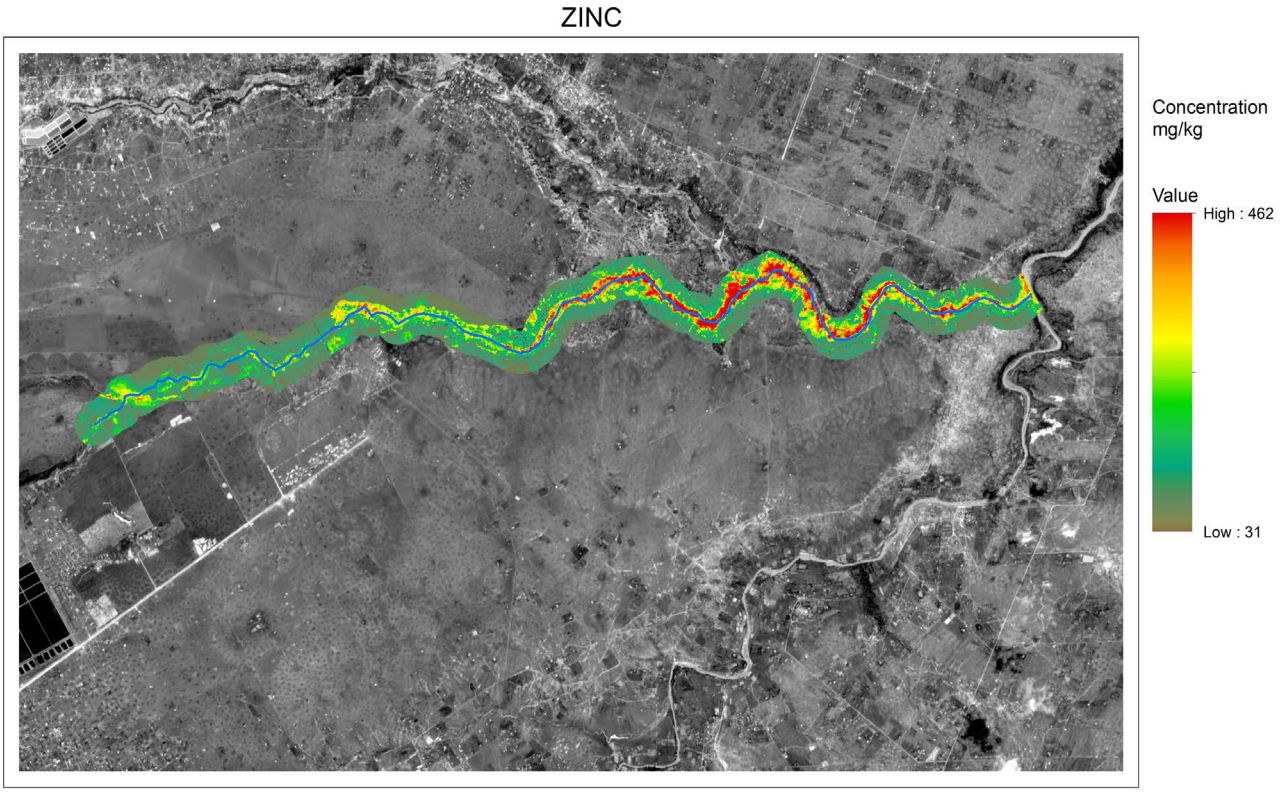


Figure 14. Predicted map (10 m resolution) of topsoil (0 - 30 cm) of Zn using the Random Forest algorithm.

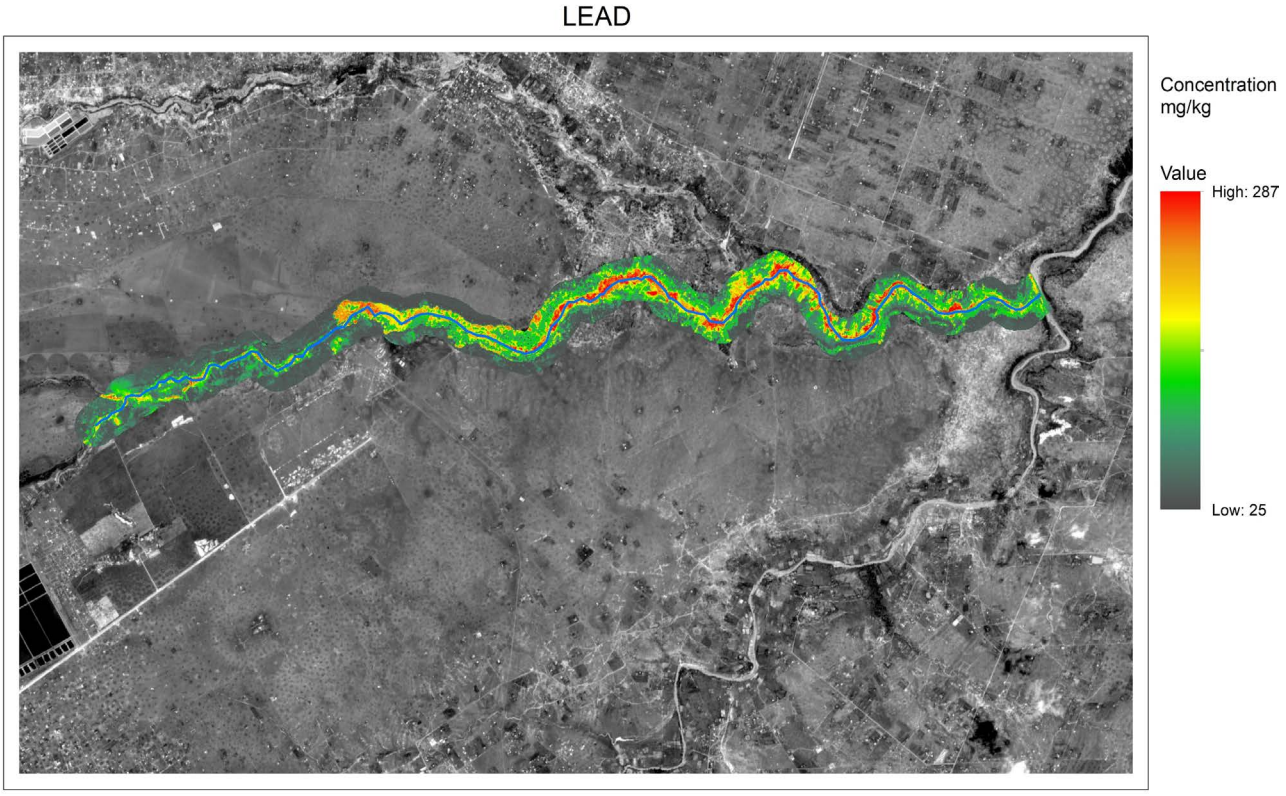


Figure 15. Predicted map (10 m resolution) of topsoil (0 - 30 cm) of Lead using the Random Forest algorithm.

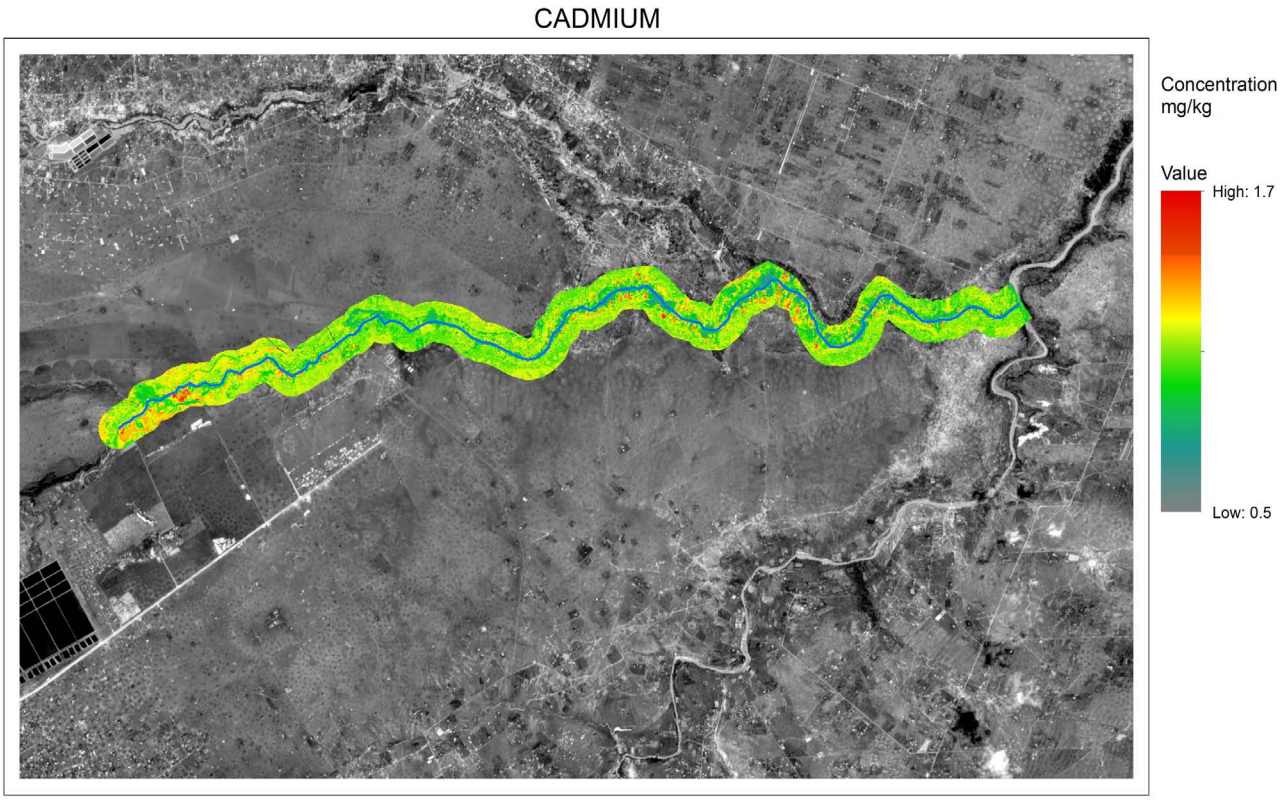


Figure 16. Predicted map (10 m resolution) of topsoil (0 - 30 cm) of Cd using the Random Forest algorithm.

Below the 300 mg/kg permissible limit, Lead poisoning can only occur through direct ingestion of contaminated soil particles.

Lead is a significant component in lead storage batteries, cable coverings, and other electronic equipment. In this regard, the probable source of Lead in the water could be leachates and run-offs from the Dandora dumpsite.

Some farms had Zinc levels, which were above the WHO/FAO/USEPA recommended limit. Zinc occurs naturally in soils in concentrations between 10 to 100 mg/kg; it's a highly toxic heavy metal at high concentrations for both plants and animals. Anthropogenic activities such as atmospheric deposition, waste combustion, mining, steel processing, and sewage sludge application continue to enrich the topsoils with Zinc. In this study, therefore, the most likely sources of Zinc are; The Dandora dumpsite where waste combustion takes place, the Nairobi Water and sewerage plant where treated wastewater is discharged into the river, and the Nairobi Industrial area where some industries carry out metal processing.

4. Conclusions

From the variable rankings, it's clear that anthropogenic activities played a significant role in the pollution levels. Additionally, the predictive maps indicate that the soils are too polluted to grow food crops, thereby posing a great risk to Nairobi's residents. Health risks notwithstanding, the Water Resources Man-

agement Authority (WARMA) doesn't license individuals to abstract water from the river because of its high pollution levels.

Legal enforcement of the existing land, health, and environmental laws should end the anthropogenic activities that pollute the river. However, In the short-term, phytoremediation of the soils can be done to manage soil toxicity.

The Random Forest model gave satisfactory results in predicting the distribution of heavy metals in soil. However, the model can be improved further if the spatial resolution of the various variables is increased and through the addition of more predictor variables. It would also be interesting to determine how other machine learning algorithms like PCA, cubist, and SVM compare with Random Forest in predicting soil heavy metals. Finally, more research needs to be done along the Nairobi River on the distribution of other potential heavy metals like mercury, arsenic, chromium, and copper.

Acknowledgements

Foremost, I would like to extend my sincere gratitude to my supervisor Dr. Mark Boitt for his guidance. Secondly, I would also like to extend my appreciation to the GEGIS department for their support and professionalism all through my study period, and to Dr. Mercy Mwaniki for assisting me in brainstorming my problem statement when I was stuck and almost giving up.

Many thanks to Patrick from the horticulture department for the support he provided during sample collection and analysis. My friend Ian Ariko, for all the reviews he made in my document and his assistance with the statistical analysis of my soil sample results, may God bless him abundantly.

Finally, I would like to appreciate my father (John Nyagara) and my mother (Rosemary Ngere) for the unwavering support and encouragement I received through my journey.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Omari, J.N., Mutwiwa, U.N. and Mailutha, J.T. (2016) The Current Status and Handling of E-Waste in Nairobi City County of Kenya. *Journal of Sustainable Research in Engineering*, **3**, 22-28.

[2] World Bank Group (2018) Municipal Solid Waste.

[3] Bello, I.A., Norshafiq, M. and Kabbashi, N.A. (2016) Solid Waste Management in Africa: A Review. *International Journal of Waste Resources*, **6**, 4-7. https://doi.org/10.4172/2252-5211.1000216

[4] Takeuchi, N. (2019) Results of the SDG 11.6.1 Data Collection Exercise. Nairobi.

[5] Adekunle, J., Oyekunle, O., Ogunfowokan, A.O., Olutona, G.O. and Durosinmi, M. (2012) Hydrology for Disaster Management Total and Exchangeable Metals in Groundwater of Ile-Ife, Southwestern Nigeria. *Special Publication of the Nigerian*

Association of Hydrological Sciences, 208-223.

[6] Lamine, S., Petropoulos, G.P., Brewer, P.A., Bachari, N.E., Srivastava, P.K., Manevski, K., Kalaitzidis, C. and Macklin, M.G. (2019) Heavy Metal Soil Contamination Detection Using Combined Geochemistry and Field Spectroradiometry in the United Kingdom. *Sensors*, **19**, 762. https://doi.org/10.3390/s19040762

[7] Sridhar, B.B.M., Vincent, R.K., Roberts, S.J. and Czajkowski, K. (2011) Remote Sensing of Soybean Stress as an Indicator of Chemical Concentration of Biosolid Amended Surface Soils. *International Journal of Applied Earth Observation and Geoinformation*, **13**, 676-681. https://doi.org/10.1016/j.jag.2011.04.005

[8] Ashfaque, F., Sahay, S., Islamia, J.M. and Iqbal, S. (2016) Influence of Heavy Metal Toxicity on Plant Growth, Metabolism and Its Alleviation by Phytoremediation—A Promising Technology. *Journal of Agriculture and Ecology Research International*, **6**, 1-19. https://doi.org/10.9734/JAERI/2016/23543

[9] Verrelst, J., *et al.* (2011) Evaluation of Sentinel-2 Red-Edge Bands for Empirical Estimation of Green LAI and Chlorophyll Content. *Sensors*, **11**, 7063-7081. https://doi.org/10.3390/s110707063

[10] Sampson, P.H., Zarco-tejada, P.J., Mohammed, G.H., Miller, J.R. and Noland, T.L. (2002) Hyperspectral Remote Sensing of Forest Condition: Estimating Chlorophyll Content in Tolerant Hardwoods. *Forest Science*, **49**, 381-391.

[11] Liu, M., Liu, X., Zhang, B. and Ding, C. (2016) Regional Heavy Metal Pollution in Crops by Integrating Physiological Function Variability with Spatio-Temporal Stability Using Multi-Temporal Thermal Remote Sensing. *International Journal of Applied Earth Observation and Geoinformation*, **51**, 91-102. https://doi.org/10.1016/j.jag.2016.05.003

[12] Del, I., Sanches, A., Roberto, C., Filho, S. and Floyd, R. (2014) Spectroscopic Remote Sensing of Plant Stress at Leaf and Canopy Levels Using the Chlorophyll 680 nm Absorption Feature with Continuum Removal. *ISPRS Journal of Photogrammetry and Remote Sensing*, **97**, 111-122. https://doi.org/10.1016/j.isprsjprs.2014.08.015

[13] Dockray, D.N.H.H.M., Barringerj, A.R. and Barber, J. (1983) Red Edge Measurements for Remotely Sensing Plant Chlorophyll Content of Educational. *Advances in Space Research*, **3**, 273-277. https://doi.org/10.1016/0273-1177(83)90130-8

[14] Horler, D.N.H., Barber, J., Darch, J.P., Ferns, D.C. and Barringer, A.R. (1983) Approaches to Detection of Geochemical Stress in Vegetation. *Advances in Space Research*, **3**, 175-179. https://doi.org/10.1016/0273-1177(83)90118-7

[15] Peng, Y., Kheir, R.B., Adhikari, K., Malinowski, R. and Greve, M.B. (2016) Digital Mapping of Toxic Metals in Qatari Soils Using Remote Sensing and Ancillary Data. *Remote Sensing*, **8**, 1-19. https://doi.org/10.3390/rs8121003

[16] Luis Rodriguez Lado, H.I.R. and Hengl, T. (2008) Heavy Metals in European Soils: A Geostatistical Analysis of the FOREGS Geochemical Database. *Geoderma*, **148**, 189-199. https://doi.org/10.1016/j.geoderma.2008.09.020

[17] Lv, J. and Yan, Z. (2014) Estimation of Pb Concentration in the Mining Tailing Areas Base on Field Spectrometry and Support Vector Machine. *The Third International Conference on Agro-Geoinformatics*, Beijing, 11-14 August 2014, 1-5. https://doi.org/10.1109/Agro-Geoinformatics.2014.6910568

[18] Grunwald, S., Vasques, G.M. and Rivero, R.G. (2014) Fusion of Soil and Remote Sensing Data to Model Soil Properties. *Advances in Agronomy*, **131**, 1-109. https://doi.org/10.1016/bs.agron.2014.12.004

[19] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

[20] Pal, M. (2005) Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing*, **26**, 37-41. https://doi.org/10.1080/01431160412331269698

[21] Xu, B. and Ye, Y. (2012) An Improved Random Forest Classifier for Image Classification. 2012 *International Conference on Information and Automation*, Shenyang, 6-8 June 2012, 795-800. https://doi.org/10.1109/ICInfA.2012.6246927

[22] Millard, K. and Richardson, M. (2015) On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sensing*, **7**, 8489-8515. https://doi.org/10.3390/rs70708489

[23] Nguyen, U., Glenn, E.P., Duc, T. and Pham, L.T.H. (2019) Mapping Vegetation Types in Semi-Arid Riparian Regions Using Random Forest and Object-Based Image Approach: A Case Study of the Colorado River Ecosystem, Grand Canyon, Arizona. *Ecological Informatics*, **50**, 43-50. https://doi.org/10.1016/j.ecoinf.2018.12.006

[24] Bailey, B.N. (2019) Mapping Aboveground Biomass of Four Typical Vegetation Types in the Poyang Lake Wetlands Based on Random Forest Modelling and Landsat Images. *Frontiers in Plant Science*, **10**, 1281. https://doi.org/10.3389/fpls.2019.01281

[25] Du Kun Tan, Q., Ma, W.B. and Wu, F.Y. (2019) Random Forest—Based Estimation of Heavy Metal Concentration in Agricultural Soils with Hyperspectral Sensor Data. *Environmental Monitoring and Assessment*, **191**, Article No. 446.

[26] Mariana Belgiu, L.D. (2016) Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, **114**, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

[27] Pouladi, N., Bjørn, A., Tabatabai, S. and Greve, M.H. (2019) Mapping Soil Organic Matter Contents at Field Level with Cubist, Random Forest and Kriging. *Geoderma*, **342**, 85-92. https://doi.org/10.1016/j.geoderma.2019.02.019

[28] Wang, H., Yilihamu, Q., Yuan, M., Bai, H., Xu, H. and Wu, J. (2020) Prediction Models of Soil Heavy Metal(loid)s Concentration for Agricultural Land in Dongli: A Comparison of Regression and Random Forest. *Ecological Indicators*, **119**, Article ID: 106801. https://doi.org/10.1016/j.ecolind.2020.106801

[29] Kibet, J. (2017) The State of Water Quality in Nairobi River, Kenya.

[30] Kenya National Bureau of Statistics (2019) Kenya Population and Housing Census Volume I: Population by County and Sub-County, I.

[31] Theocharopoulos, S.P., Wagner, G., Sprengart, J. and Mohr, M. (2001) European Soil Sampling Guidelines for Soil Pollution Studies. *Science of the Total Environment*, **264**, 51-62. https://doi.org/10.1016/S0048-9697(00)00611-2

[32] Amini, H., Schindler, C., Hosseini, V. and Yunesian, M. (2017) Land Use Regression Models for Alkylbenzenes in a Middle Eastern Megacity: Tehran Study of Exposure Prediction for Environmental Health Research (Tehran SEPEHR). https://doi.org/10.1289/isee.2017.2017-936

[33] Peng, Y., Xiong, X., Adhikari, K., Knadel, M. and Grunwald, S. (2015) Modeling Soil Organic Carbon at Regional Scale by Combining Multi-Spectral Images with Laboratory Spectra. *PLoS ONE*, **10**, e0142295. https://doi.org/10.1371/journal.pone.0142295

[34] Clevers, J.G.P.W. (1991) Application of the WDVI in Estimating LAI at the Generative Stage of Barley. *ISPRS Journal of Photogrammetry and Remote Sensing*, **46**, 37-47. https://doi.org/10.1016/0924-2716(91)90005-G

[35] Zhang, Z., Liu, M., Liu, X. and Zhou, G. (2018) A New Vegetation Index Based on

Multitemporal Sentinel-2 Images for Discriminating Heavy Metal Stress Levels in Rice. *Sensors*, **18**, 2172. https://doi.org/10.3390/s18072172

[36] Malone, B.P., Minasny, B. and McBratney, A.B. (2017) Using R for Digital Soil Mapping. Springer International Publishing, Sydney. https://doi.org/10.1007/978-3-319-44327-0

[37] Brokamp, C., Jandarov, R., Rao, M.B., Lemasters, G. and Ryan, P. (2017) Exposure Assessment Models for Elemental Components of Particulate Matter in an Urban Environment: A Comparison of Regression and Random Forest Approaches. *Atmospheric Environment*, **151**, 1-11. https://doi.org/10.1016/j.atmosenv.2016.11.066

[38] Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest.

[39] FAO and WHO (2019) Discussion Paper on the Development of a Code of Practice for the Prevention and Reduction of Cadmium Contamination in Cocoa. Joint FAO/WHO Food Standards Programme Codex Committee on Contaminants in Foods 13th Session, No. Appendix I.

[40] United States Environmental Protection Agency (2007) Framework for Metals Risk Assessment Framework for Metals Risk Assessment.

[41] Wuana, R.A. and Okieimen, F.E. (2011) Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation. *International Scholarly Research Notices*, **2011**, Article ID: 402647. https://doi.org/10.5402/2011/402647

[42] Xu, X., Zhao, Y., Zhao, X., Wang, Y. and Deng, W. (2014) Sources of Heavy Metal Pollution in Agricultural Soils of a Rapidly Industrializing Area in the Yangtze Delta of China. *Ecotoxicology and Environmental Safety*, **108**, 161-167. https://doi.org/10.1016/j.ecoenv.2014.07.001