JOINT INSTITUTE FOR NUCLEAR RESEARCH
Meshcheryakov Laboratory of Information Technologies

FINAL REPORT ON START PROGRAMME

Machine Learning Estimation of Heavy Metal Pollution Concentration Based on Satellite Imagery

Supervisor:
Dr. Alexander Ayriyan

Student:
Milena Aghabekyan, Armenia
Yerevan State University

Participation period:
June 30 – August 10,
Summer Session 2024

Dubna, 2024

Abstract

This study presents a Heavy Metal Contamination Model that utilizes satellite imagery and machine learning techniques to determine heavy metal concentrations in soil at unmeasured locations. The model was developed using a dataset containing satellite imagery data and corresponding heavy metal concentrations from previous research.
The Google Earth Engine platform was used to calculate indices from satellite images that represent summarized information.
Two machine learning algorithms were trained and evaluated based on their Mean Squared Error and R-squared Score: Random Forest Regressor, Gradient Boosting Regressor.
Model was then used to determine heavy metal concentrations at fifty randomly selected points in the study region, providing accurate estimates of contamination levels. The use of satellite imagery and machine learning techniques for heavy metal contamination determination offers a cost-effective and efficient alternative to physical sampling, contributing to more comprehensive and timely monitoring and management strategies.

Introduction

There are many sources of soil pollution by heavy metals such as industrial areas, paints, fertilizers, disposal of heavy metals, sewage sludge, animal manures, wastewater irrigation, pesticides, coal combustion residues, spillage of petrochemicals, and other different sources.
This study presents an approach to monitoring heavy metal contamination in soil using satellite imagery and machine learning techniques. The model aims to predict heavy metal concentrations at unmeasured locations in Egypt, which is crucial for effective management and mitigation strategies.

The datasets used in this study are the Landsat 9 and MODIS (onboard the Terra satellite) satellite imagery data, which were filtered for the specified date range and cloud cover percentage.

The study focused on predicting heavy metal concentrations at 50 randomly selected points in Egypt, using the median values of the surface reflectance bands as input features. The use of satellite imagery and machine learning techniques for heavy metal contamination identification offers a cost-

effective and efficient alternative to physical sampling, contributing to more comprehensive and timely monitoring and management strategies.

The main advantages are:

1. Cost-effective: The use of satellite imagery and machine learning techniques reduces the need of physical sampling, which can be expensive and time-consuming.
2. Efficient: The model can identify heavy metal concentrations at unmeasured locations, reducing the need for extensive field sampling and analysis.
3. Timely: The model can provide rapid results, enabling timely decision-making and response to heavy metal contamination issues.
4. Comprehensive: The model can provide information on heavy metal concentrations over a large area, enabling more comprehensive monitoring and management strategies.

The limitations of this study are:

1. Data quality: The accuracy of the model depends on the quality of the satellite imagery data and the laboratory analysis of the heavy metal concentrations.
2. Generalizability: The model may not be applicable to other regions or countries with different environmental conditions.
3. Complexity: The model may require complex machine learning algorithms and expertise in remote sensing and data analysis.

Overall, this study demonstrates the potential of using satellite imagery and machine learning techniques for estimation heavy metal contamination in soil. The result of this study can contribute to more comprehensive and timely monitoring and management strategies for heavy metal contamination in Egypt.

Materials and methods

Some Egyptian soils are polluted by heavy metals, where concentrations of Fe, Mn, and Zn are moderate to high. Industrial contaminated areas of Fe, Mn, Zn, Cu, Cd, Co, Ni and Pb were investigated. Levels of Pb, Ni, Co and Cd in soils. Soil surface samples were collected using a systematic nested gridding soil sampling design with 1 km spacing.

Hyperspectral images are a unique source for obtaining many kinds of information about the Earth's surface. Modern platforms support users to perform complex analyses with a collection of images without using any specialized software. Google Earth Engine (GEE) is a planetary-scale platform for Earth science data & analysis. Atmospheric, radiometric, and geometric corrections have been made to a number of image collections at GEE. There are over 100 satellite image collections and modeled datasets. With just a few commands, GEE enables to get a median image by specifying the collection name, date frame, and area of interest. It enables working with Earth remote sensing, utilizing satellite images to obtain snapshots of a specific area over a certain period of time, allows for various manipulations with these images, such as obtaining the median image or getting statistics for a specific area. We have an observation point, and we can receive statistical data for spectral channels within an area, conditionally one square kilometer, centered on this point.
In this research we worked with two programs:

1. MODIS
2. LANDSAT 9

Two of the most popular datasets available in GEE are MODIS (Moderate Resolution Imaging Spectroradiometer) and Landsat 9.

MODIS dataset

The MODIS dataset is a collection of high-resolution, multi-spectral imagery collected by NASA’s Terra and Aqua satellite between 2000 and 2020. The MOIDS instrument captures images of the Earth surface at a spatial resolution of 250-500 meters, making it suitable for studying a wide range of environmental phenomena.

GEE provides a wide range of MODIS products, including:

1. Surface reflectance: The reflected solar radiation from the Earth's surface, measured in 36 spectral bands.

2. Atmosphere corrected reflectance: The reflected solar radiation from the Earth's surface, corrected for atmospheric effects.

3. Land surface temperature: The temperature of the Earth's surface, measured in Kelvin.

4. Vegetation indices: Indices that quantify vegetation health, such as the Normalized Difference Vegetation Index (NDVI).

Landsat 9 Dataset

The Landsat 9 dataset is a collection of high-resolution, multi-spectral imagery collected by NASA's Landsat 9 satellite, which was launched in September 2020. Landsat 9 is the latest satellite in the Landsat series, which has been providing continuous coverage of the Earth's surface since the 1970s. The Landsat 9 instrument captures images of the Earth's surface at a spatial resolution of 30 meters, making it suitable for studying a wide range of environmental phenomena

GEE provides a range of Landsat 9 products, including:

1. Top-of-atmosphere reflectance: The reflected solar radiation from the Earth's surface, measured in 10 spectral bands.

2. Surface reflectance: The reflected solar radiation from the Earth's surface, corrected for atmospheric effects.

3. Land cover classification: Maps of land cover classification based on machine learning algorithms.

***

Initially, we were provided with a set of coordinates, which served as the foundation for spatially-located data collection. By utilizing the supplied coordinates, we identified specific points on a map where research was conducted, obtained visual data and corresponding coordinates, and subsequently performed data collection at each point. The resulting data included heavy metal concentrations, which were subsequently employed to develop a statistical model. Subsequently, we randomly selected 50 points within the designated region, collected data at each point, and applied the pre-existing statistical model to the obtained data, yielding a predictive outcome.
We employed a statistical modeling approach to investigate the relationship between environmental pollution and satellite imagery, as direct field research was not feasible in many locations. We leveraged existing datasets and extracted relevant spatial data from areas where the research was conducted, comprising two databases. Subsequently, we analyzed the correlation between the data, identifying regions with higher and lower levels of dependence, and developed a statistical model that detects patterns and associations between pollution and satellite image data.

Database from MODIS

This dataset contains satellite image data for environmental monitoring, specifically the reflected radiance values for seven bands (B01 to B07) of the surface. The data is collected from various locations and is used to study the relationships between environmental pollution and satellite imagery.

The columns:
* sur_refl_b01: Reflected radiance values for band B01 (visible light)
* sur_refl_b02: Reflected radiance values for band B02 (blue light)
* sur_refl_b03: Reflected radiance values for band B03 (green light)
* sur_refl_b04: Reflected radiance values for band B04 (red light)
* sur_refl_b05: Reflected radiance values for band B05 (near-infrared light)
* sur_refl_b06: Reflected radiance values for band B06 (mid-infrared light)
* sur_refl_b07: Reflected radiance values for band B07 (thermal infrared light)
* latitude: The geographic latitude of the observation point in decimal degrees.
* longitude: The geographic longitude of the observation point in decimal degrees.
* .geo: A GeoJSON object containing the coordinates of the observation point.

The rows in the dataset represent different observations, with each observation containing the reflected radiance values for the seven bands. The values are represented as floating-point numbers, with units of W/m^2/sr/μm (watts per square meter per steradian per micrometer).

This dataset is suitable for researchers and analysts interested in environmental monitoring, remote sensing, and climate change. It can be used to analyze the relationships between environmental pollution and satellite imagery, as well as to identify patterns and trends in environmental data.

Database from Landsat

This dataset contains a combination of environmental monitoring data and geospatial information, including satellite imagery, latitude, and longitude coordinates. The data is collected from various locations and is used to study the relationships between environmental factors and their spatial distribution.

The columns:

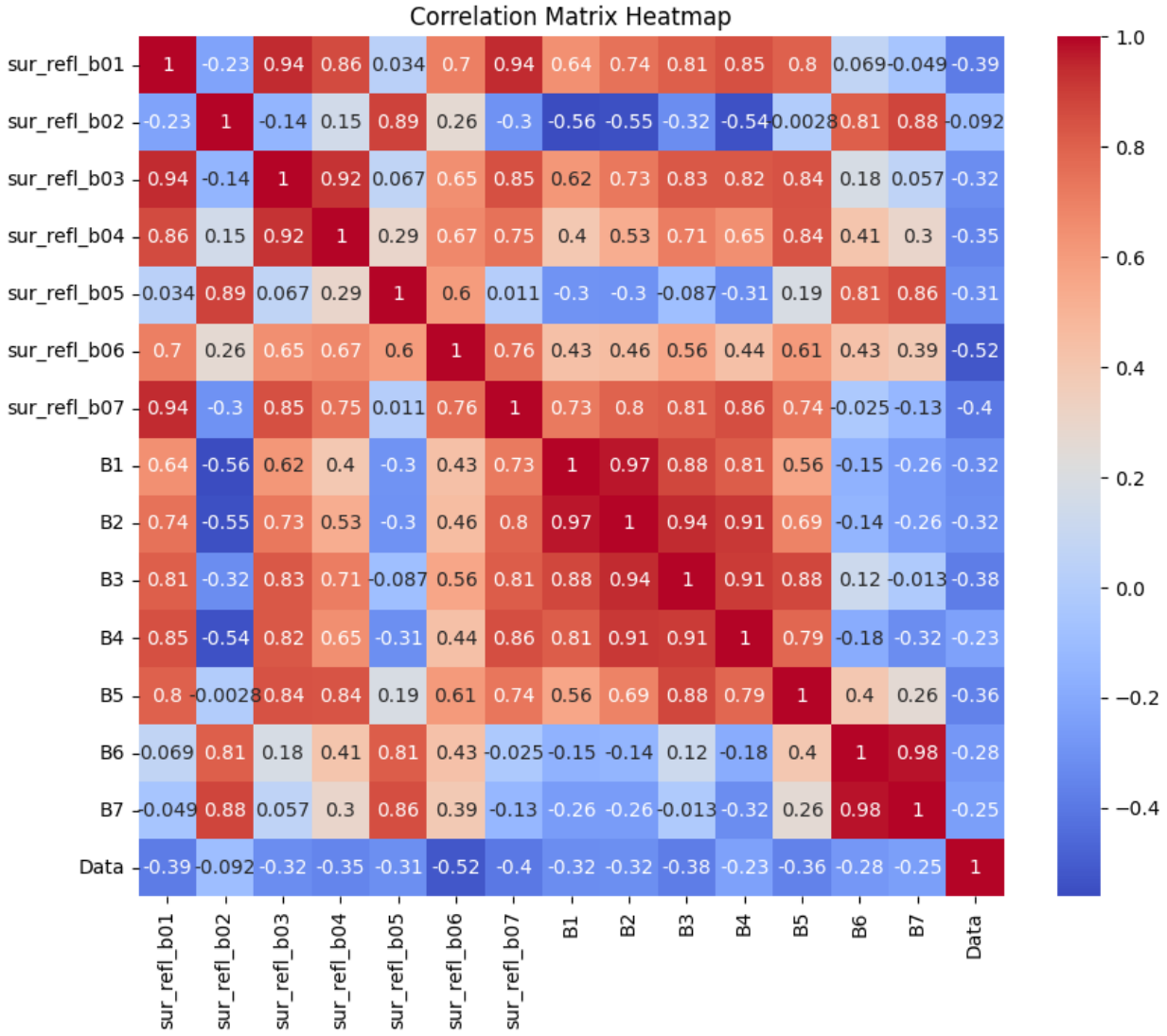
* B1 to B7: Reflected radiance values for seven bands (B1 to B7) of the satellite imagery, representing different spectral ranges.
* latitude: The geographic latitude of the observation point in decimal degrees.
* longitude: The geographic longitude of the observation point in decimal degrees.
* .geo: A GeoJSON object containing the coordinates of the observation point.

The dataset is structured as a table, with each row representing a single observation. The data is suitable for researchers and analysts interested in environmental monitoring, remote sensing, and spatial analysis. It can be used to analyze the relationships between environmental factors and their spatial distribution, as well as to identify patterns and trends in environmental data.

Points depending on given coordinates where heavy metals are

After conducting data preprocessing and manipulation tasks, the dataset was optimized for subsequent analysis and modeling applications. Here is a rewritten version of the line in a scientific tone:
The concentrations of heavy metals were incorporated as a target variable (“Data”) in the combined dataset, thereby enabling the identification of this

parameter using the testing dataset. This allows for the evaluation of the model's ability to accurately forecast the concentrations of heavy metals based on the input features, providing valuable insights into the relationship between these variables.
The work then performs a correlation analysis using the seaborn library to identify relationships between the remaining columns. The correlation matrix is visualized using a heatmap.



As we were dealing with a regression problem we used Random Forest Regressor and Gradient Boosting Regressor algorithms on our training data. The models are evaluated using mean squared error (MSE) and R-squared score on the testing data.

- Mean Squared Error (MSE): MSE measures the average of the squares of the errors between predicted and actual values.
- R-Squared (R^2): R² represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

The results of the model evaluation are printed, including the MSE and R-squared score for both models. The Gradient Boosting Regressor model is found to perform better than the Random Forest Regressor model in terms of MSE and R-squared score.

```
MSE is 2128.8472093124988
R2 is -0.015382443099297927

Gradient Boosting - Mean Squared Error: 2276.7177737424754
Gradient Boosting - R-squared Score: 0.48203232417034425
```

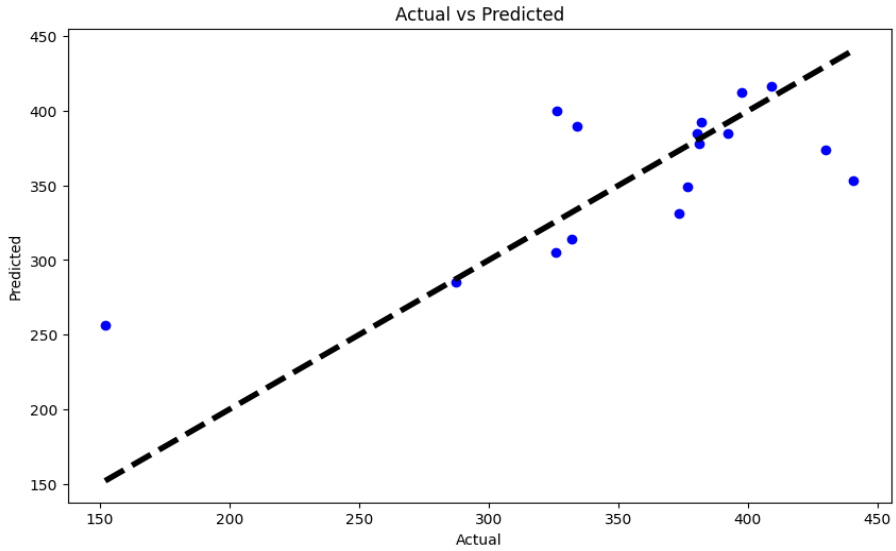
***

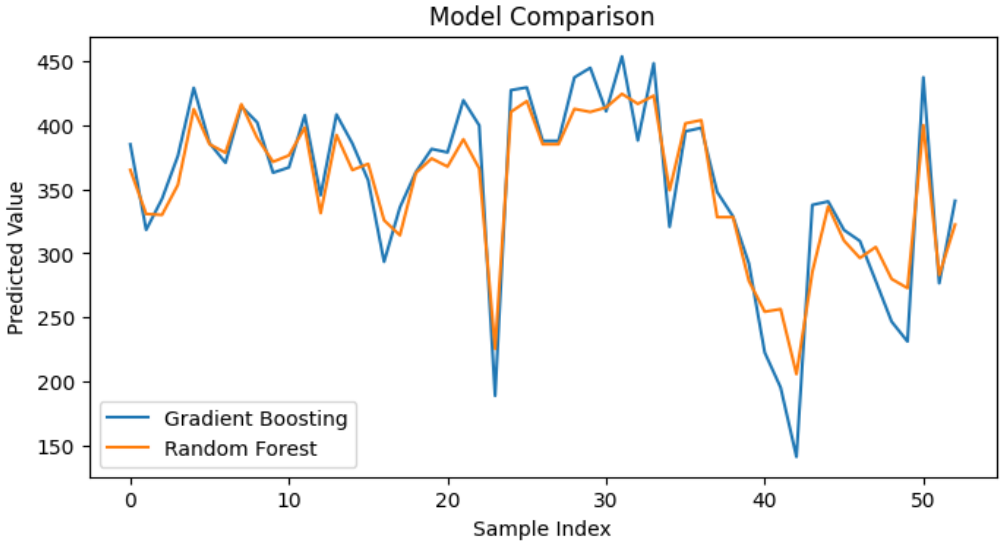
Following the development of the predictive model, a subsequent data collection effort was undertaken using the Google Earth Engine (GEE) platform, wherein 50 random points were selected to obtain additional observational data and make predictions. This sampling strategy enabled the collection of new information that can be used to validate and refine the model's performance.

Randomly chosen points

Following the random point selection, data collection was conducted using the GEE platform, yielding two databases. Subsequent data manipulation and processing procedures were employed to transform the collected data into a suitable testing dataset, ready for evaluation and validation of the predictive model's performance.

Actual target variable values	Predictions with Random Forest	Predictions with Gradient Boosting
384.9	364.99	384.90000268
318.3	330.634	318.30000927





Conclusion

In conclusion, this study has successfully demonstrated the potential of using satellite imagery and machine learning techniques for determination of heavy metal contamination in soil. The developed model, utilizing Google Earth Engine and machine learning algorithms, has shown promising results in identification of heavy metal concentrations at unmeasured locations, providing accurate estimates of contamination levels. The model's advantages, including cost-effectiveness, efficiency, timeliness, and comprehensiveness, make it a valuable tool for environmental monitoring and management strategies.

The application of Random Forest Regressor and Gradient Boosting Regressor algorithms demonstrated the potential of these techniques in accurately estimate heavy metal concentrations. The Gradient Boosting Regressor model, in particular, showed superior performance in terms of Mean Squared Error (MSE) and R-squared score, indicating its robustness and reliability for this specific application.

The model was further validated by predicting heavy metal concentrations at fifty randomly selected points within the study region in Egypt. The results underscore the model's ability to provide accurate estimates of contamination levels, highlighting its practical utility in real-world scenarios.

This study has several significant implications:

Firstly, it offers a cost-effective and efficient alternative to traditional physical sampling methods, which can be both expensive and time-consuming. By reducing the need for extensive field sampling and analysis, the model enables more comprehensive and timely monitoring and management strategies for heavy metal contamination.

Secondly, the use of satellite imagery and machine learning techniques allows for rapid results, facilitating timely decision-making and response to heavy metal contamination issues. This is particularly crucial in regions where immediate action is required to mitigate the adverse effects of heavy metal pollution on human health and the environment.

Thirdly, the model's ability to provide information on heavy metal concentrations over a large area enables more comprehensive monitoring and management strategies. This is essential for developing effective policies and interventions aimed at reducing heavy metal contamination and protecting public health.

However, it is important to acknowledge the limitations of this study. The accuracy of the model is contingent on the quality of the satellite imagery data and the laboratory analysis of the heavy metal concentrations. Additionally, the model's generalizability may be limited to regions with similar environmental conditions, and its complexity may require expertise in remote sensing and data analysis.

Despite these limitations, this study demonstrates the potential of using satellite imagery and machine learning techniques for predicting heavy metal contamination in soil. The results contribute to more comprehensive and timely monitoring and management strategies for heavy metal contamination in Egypt and potentially other regions with similar environmental challenges.