

JOINT INSTITUTE FOR NUCLEAR RESEARCH

Laboratory of Information Technologies

**FINAL REPORT ON THE**

**START PROGRAMME**

**Remote Sensing and Heavy Metal Data Analysis:**

Approaches to soil elemental composition

**Supervisor:**

Dr. Alexander Ayriyan

**Co-supervisor:**

Dr. Alexander Uzhinskiy

**Student:**

Leen Abdeen, Egypt Menofia University

**Participation period:**

August 18 – October 12,

Summer Session 2024

Dubna, 2024

**Contents**

**1. Abstract ............................................................................................. 3**

**2. Mastering technical tools ................................................................. 4**

**3. Introduction ..................................................................................... 5**

1. **Data collection .................................................................................. 7**

**4.1** Map sampling .......................................................................... 7

**4.2** Satellite data ........................................................................… 9

**5. Data analysis ………………............................................................ 10**

**5.1** Convert satellite data to python platform …………………… 10

**5.2** Correlation function and filtering results …………………… 11

**6. Data visualization ……………………………………………….... 12**

**6.1** Scatter plot …………………………………………………... 12

**6.2** Linear plot …………………………………………………... 13

**7. Future plans and final results ......................................................... 15**

**8. Refrences............................................................................................ 16**

**Abstract**

This report presents a comprehensive analysis of the relationship between heavy metal (HM) concentrations in soil and satellite data using remote sensing techniques. The study focuses on specific sampling locations across a designated area, where soil samples were collected between August 2022 and September 2022 to measure the concentrations of heavy metals. Remote sensing data, obtained through **Google Earth Engine (GEE)**, was utilized to gather information on environmental conditions within the study area. The collected data was then processed and analyzed using **Python** to investigate potential correlations between satellite-derived features and heavy metal concentrations in the soil. The findings of this study aim to deepen our understanding of the relationship between soil composition and satellite data, demonstrating the effectiveness of integrating remote sensing and data analysis techniques for environmental research.

**Mastering technical tools**

Over the past two months, I have developed proficiency in utilizing the **Google Earth Engine (GEE)** platform for remote sensing and data collection. In addition to mastering GEE, I have gained expertise in several key Python libraries that have been instrumental in this project, including:

* **Earth Engine API (ee):** for accessing and processing satellite imagery.
* **Matplotlib:** for visualizing data through various graphical representations.
* **Pandas:** for efficient data manipulation and analysis.

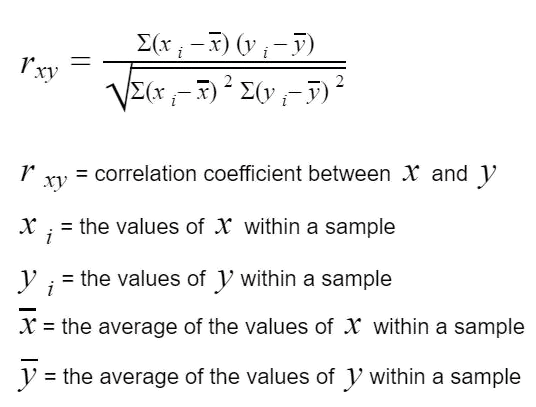
These tools have significantly enhanced my ability to analyze environmental data and derive meaningful insights from satellite imagery.

**Introduction**

The integration of remote sensing and soil analysis offers a powerful framework for understanding environmental conditions, particularly in assessing heavy metal concentrations in soil. As global pollution levels continue to rise, it is increasingly important to develop methodologies that can efficiently monitor and analyze the impact of contaminants on ecosystems and human health. This report seeks to explore the connection between satellite imagery and soil element analysis, with a specific focus on heavy metals that pose significant health risks.

To begin this study, data was collected using the **Google Earth Engine (GEE)** platform, a cloud-based geospatial analysis tool that provides access to an extensive repository of satellite imagery. Various image collections, such as **Landsat**, **MODIS**, and **Sentinel**, were utilized to ensure that the data corresponded accurately to the geographic areas of interest. This process involved displaying specific sampling points on a map, ensuring that the satellite data aligned with ground truth observations. This step was essential to guarantee that the collected satellite imagery reflected the real-world environmental conditions, thereby enhancing the reliability and accuracy of the analysis.

Once the data collection was completed, **Python** was employed to conduct a detailed analysis of the relationship between soil elements and satellite-derived features. The **correlation coefficient** was used to quantify the correlation between these variables, providing a statistical measure of the strength and direction of the linear relationship between two continuous variables. The Pearson correlation coefficient r is calculated using the following formula:



This calculation enables us to assess the strength of the relationship between soil elements and satellite data, offering valuable insights into potential environmental impacts.

To refine the analysis, the dataset was filtered to focus exclusively on soil elements that exhibited significant correlations with satellite-derived features. This selective approach ensured that the study concentrated on the most relevant variables, enhancing the overall accuracy and interpretability of the results.

To visualize the relationships between soil elements and satellite data, **scatter plots** and **linear regression plots** were employed. Scatter plots provided a graphical representation of the correlation between two variables, offering a clear and immediate visual understanding of their relationship. Linear regression plots further illustrated the trend lines, helping to identify the strength and direction of these correlations.

By integrating remote sensing data with soil analysis, this study aims to deepen our understanding of environmental dynamics and the factors contributing to soil contamination. The methodologies discussed in this report lay a solid foundation for future research to further investigate the implications of heavy metal pollution. These findings can inform the development of effective environmental management strategies to mitigate the risks posed by heavy metal contamination.

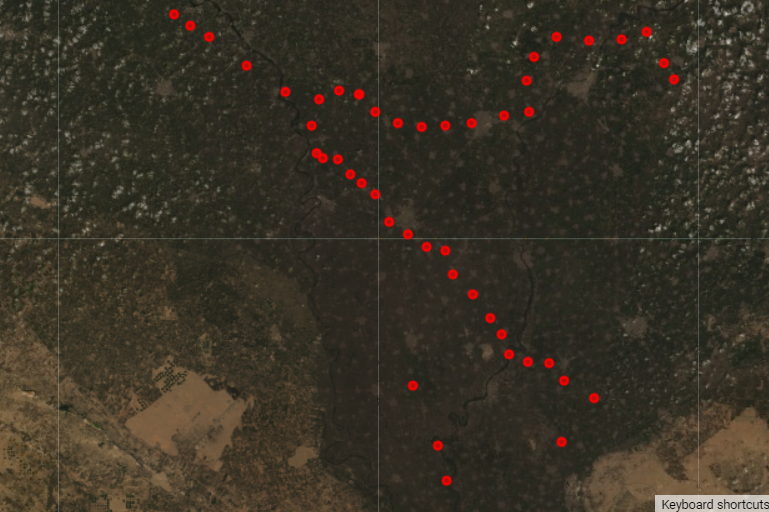
**Data collection**

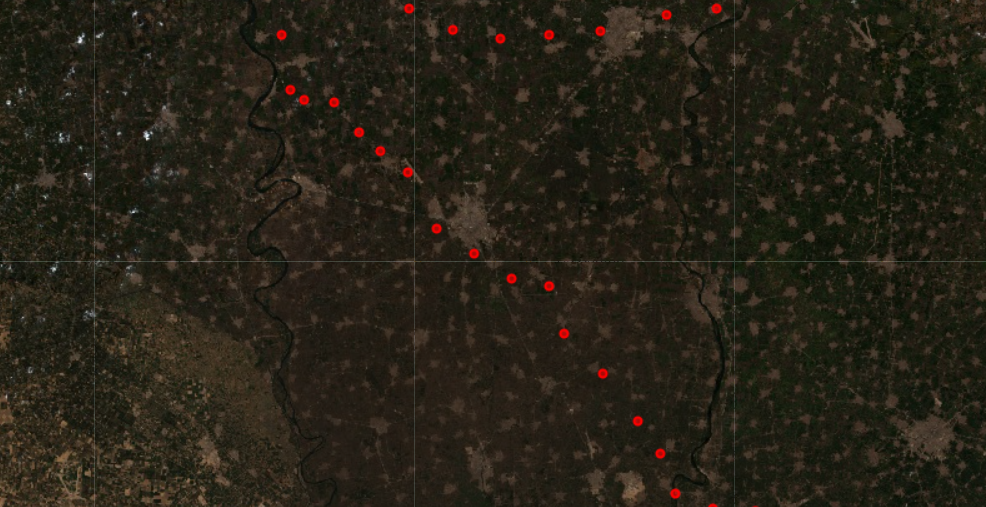
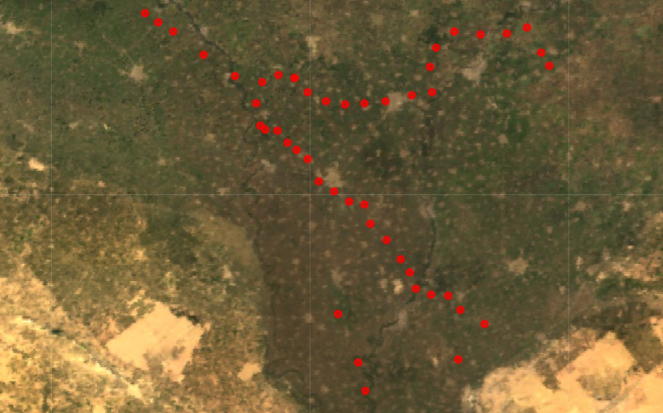
* 1. **Map sampling**

The data collection was executed utilizing Google Earth Engine (GEE), a robust platform for processing and analyzing geospatial data. The methodology encompassed several critical steps to ensure the accuracy and relevance of the collected information.

Initially, specific points on the map were sampled where soil samples were collected between August 30 and September 30, 2022. For these points, a feature collection was created, incorporating data from three distinct satellite sources: Landsat, MODIS, and Sentinel. Each satellite was selected based on its unique capabilities:

* **Landsat**: Provides high-resolution imagery (30 meters), ideal for detailed land cover analysis.
* **MODIS**: Offers daily global coverage with moderate resolution (250 to 1000 meters), making it suitable for monitoring larger-scale environmental changes.
* **Sentinel**: Particularly Sentinel-2, delivers high-resolution data (10 to 60 meters) with a comprehensive set of spectral bands, enhancing the analysis of vegetation and soil properties.

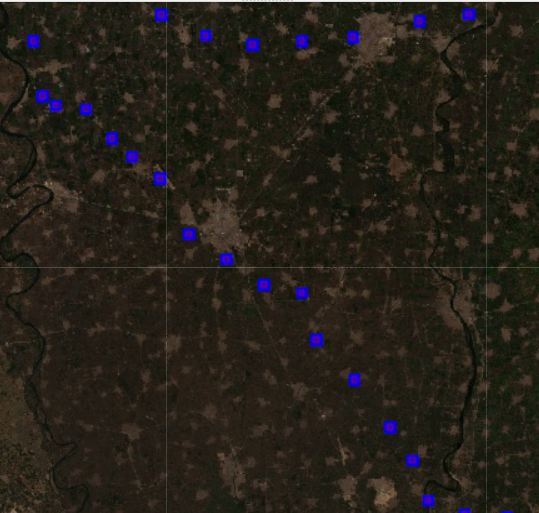
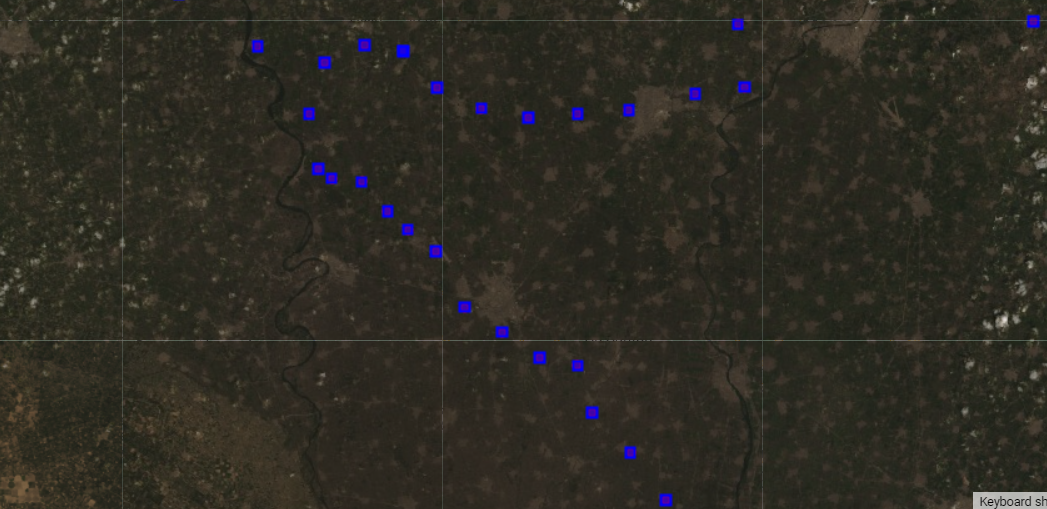
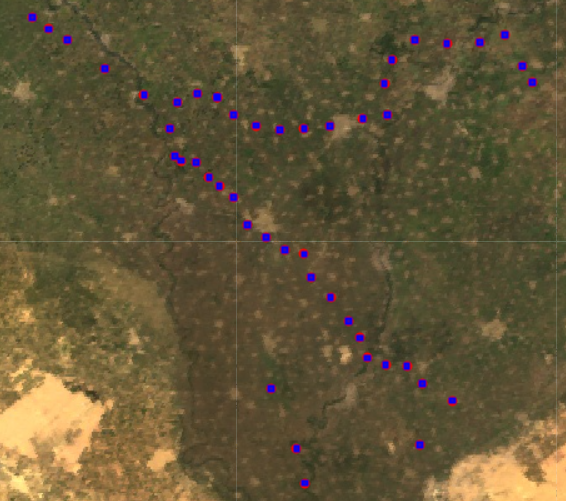


The combination of these satellites allows for a comprehensive assessment of environmental conditions, capturing both fine-scale details and broader patterns.

**Next**, I extracted data from each of these satellites within a defined area. To achieve this, I used rectangular bounding boxes to delineate a 10 km² area around each sampled point. This approach ensured that I captured relevant satellite imagery and environmental data, providing context for the soil samples collected. By leveraging the strengths of Landsat, MODIS, and Sentinel, I was able to gather a diverse set of data that would facilitate a more robust analysis of the relationship between soil elements and satellite observations.

This systematic approach to data collection not only enhanced the reliability of the analysis but also laid a solid foundation for subsequent data processing and correlation analysis.

* 1. **Satellite data**

Following the establishment of sampling points and the definition of rectangular areas, satellite data was collected to obtain relevant statistics for both the sampled points and the surrounding areas. This process involved calculating the median values of the spectral bands for each satellite within the defined 10 km² rectangles.

To achieve this, the capabilities of Google Earth Engine (GEE) were utilized to extract pixel values from the selected satellite images. GEE provides powerful tools for processing and analyzing geospatial data, enabling efficient computation of statistics across various image collections. For each rectangular area, the median values for the spectral bands of interest were calculated, including red, green, blue, near-infrared, and other relevant bands depending on the satellite source.

The median was chosen as the statistical measure due to its robustness against outliers and its ability to provide a better representation of central tendency, particularly in environmental data that may be skewed or affected by anomalous values. By focusing on the median, the results reflect the typical conditions of the area rather than being disproportionately influenced by extreme values.

Once the median values for each band were calculated, the results were organized into a structured table. This table included the median values for the various bands across Landsat, MODIS, and Sentinel, creating a comprehensive dataset that would facilitate subsequent correlation analysis between the satellite data and the heavy metal concentrations in the soil. Each entry in the table represented a unique sampling point, allowing for easy reference and comparison across different locations.

**Data analysis**

**5.1** Convert satellite data to python platform

For the data analysis phase, the collected satellite data was transferred to the **Python** environment using Jupyter Notebook, a versatile platform for data manipulation and visualization. To facilitate this process, the Earth Engine API (ee) library was employed, enabling seamless interaction with Google Earth Engine’s extensive geospatial datasets.

After successfully authenticating the account, cloud-based data was accessed to extract the median values for the spectral bands from the three selected satellites: Landsat, MODIS, and Sentinel. This extraction was performed for all sampled points and their surrounding 10 km² rectangles, ensuring the capture of a comprehensive dataset representing the environmental conditions of each area.

Once the data was retrieved, it was saved in CSV format, which is widely compatible with various data analysis tools and software. This format allowed for straightforward manipulation and analysis in subsequent steps. The organized dataset included detailed statistics categorized by satellite source and spectral bands, providing a rich resource for further exploration

**5.2** Correlation function and filtering results

Next, the correlation analysis was conducted using the **Pandas** library to assess the relationship between the satellite data and the concentrations of heavy metal (HM) elements in the soil. Three types of correlation coefficients were calculated: Pearson, Spearman, and Kendall. Each method offers distinct insights into the data:

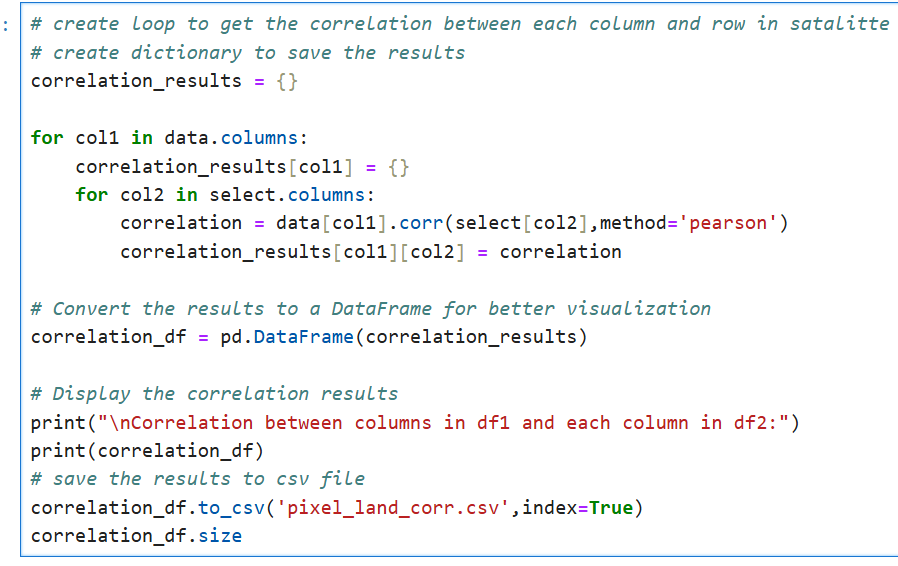
* **Pearson correlation** assesses linear relationships between variables.
* **Spearman correlation** evaluates monotonic relationships, making it useful for non-parametric data.
* **Kendall correlation** measures the ordinal association between variables, providing a robust alternative for smaller datasets.

By applying these correlation methods, significant connections between the satellite-derived features and the heavy metal concentrations were identified.

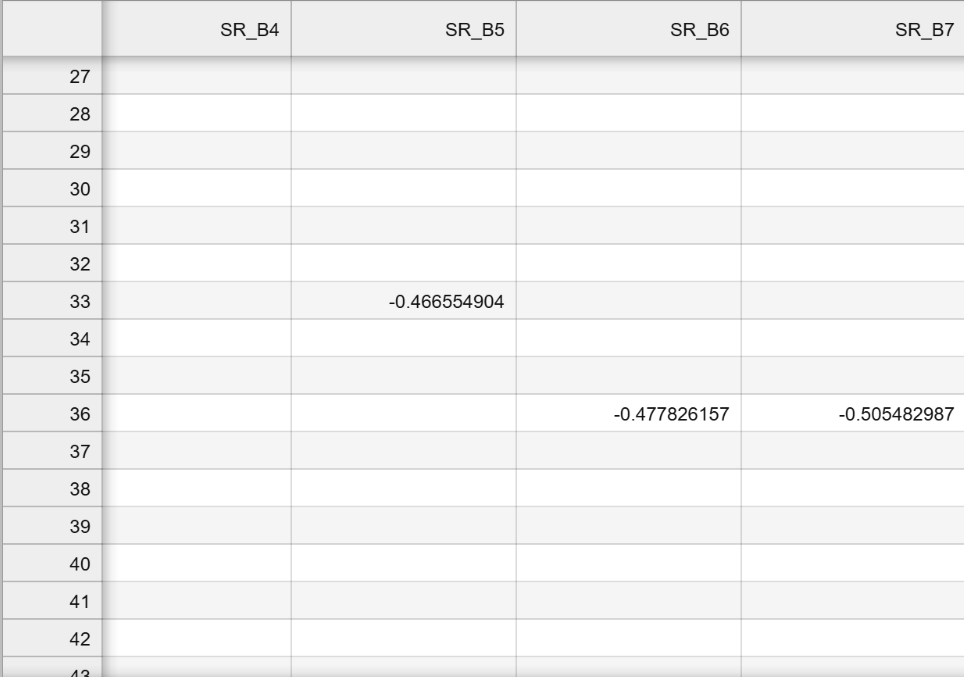
Following the correlation calculations, the results were filtered to focus solely on elements exhibiting strong connections, specifically those with correlation coefficients greater than 0.45 or less than -0.45. This threshold was chosen to highlight the most relevant relationships, ensuring that only the most significant correlations were considered for further analysis.

The filtered results were then organized into a comprehensive file, which included the correlation coefficients for all three types of correlation. This file also specified the corresponding spectral bands and the satellite sources from which the data was derived. By documenting these connections, a valuable resource was created for subsequent analysis.

Understanding the relationships between soil elements and satellite observations, paving the way for more detailed analysis and interpretation of the findings.



**“Correlation function”**



**“Filtered data”**

**Data visualization**

After completing the data collection and preprocessing phases, the next crucial step was to visualize the data through a series of graphical representations using **Matplotlib** in Python. Visualization plays a pivotal role in data analysis, as it transforms raw numerical data into intuitive and interpretable visual forms. In this project, both scatter plots and linear plots were utilized to explore and reveal the relationships present within the data

**6.1** Scatter plot

Scatter plots were employed to visualize the relationship between heavy metal concentrations and satellite-derived spectral indices. Each data point represented a unique sample location, illustrating the variation in elemental concentrations relative to the corresponding satellite measurements. To aid in the analysis, lines of best fit were incorporated into each scatter plot. These included both a linear regression line and a polynomial curve (specify the degree of the polynomial, e.g., a second-order polynomial).

The linear regression line was calculated using the method of least squares, aiming to minimize the sum of squared differences between the observed and predicted values. The equation for the linear regression line is:

**y = mx + c**

Where:

* y represents the predicted heavy metal concentration.
* x represents the satellite-derived spectral index value.
* m is the slope of the line, representing the change in heavy metal concentration per unit change in the spectral index.
* c is the y-intercept, representing the predicted heavy metal concentration when the spectral index is zero.

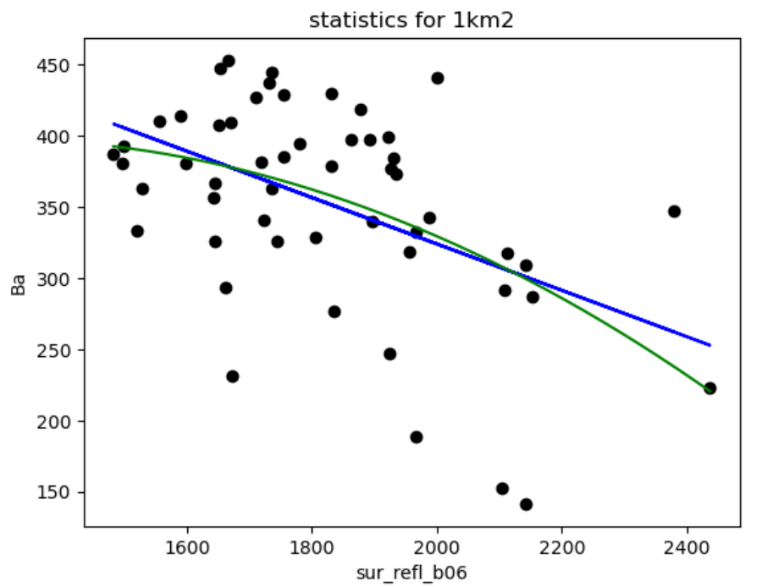
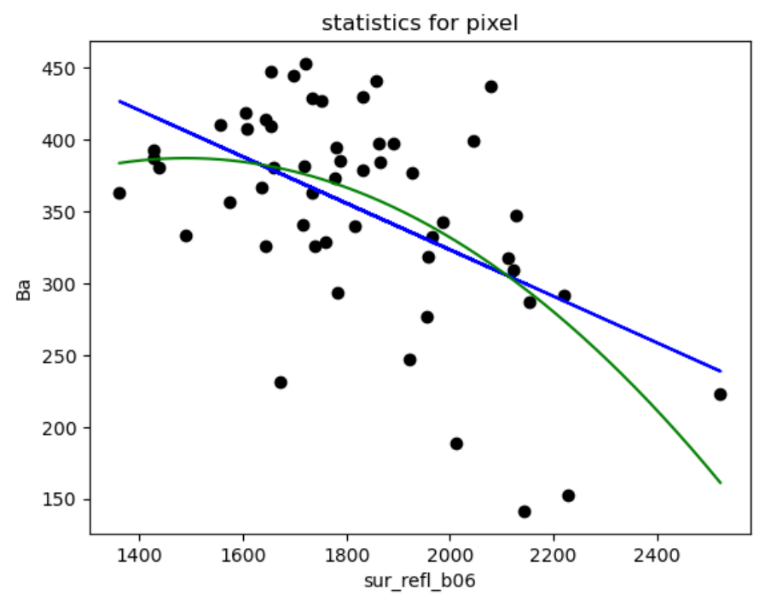
The polynomial curve (e.g., a second-order polynomial) was fitted to capture non-linear relationships. The general equation for a second-order polynomial is:

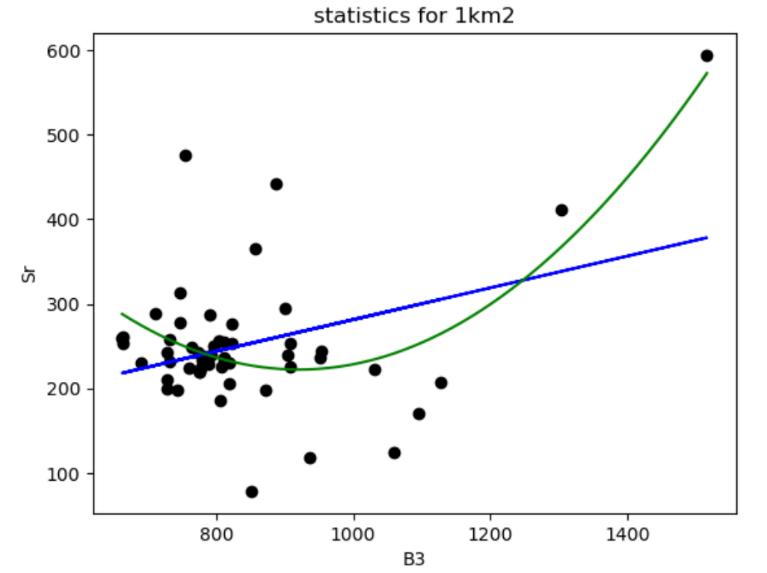
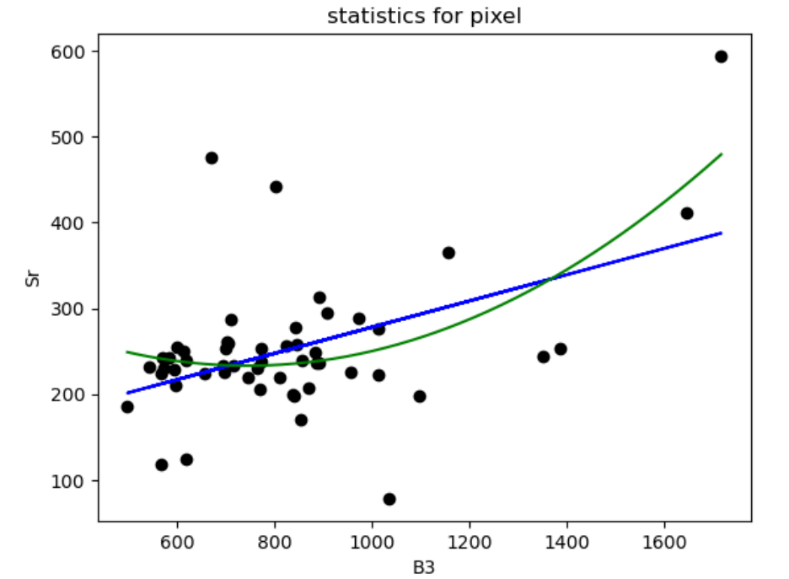
**y = ax² + bx + c**

Where:

* a, b, and c are coefficients determined through regression analysis.

Visual inspection of these plots facilitated the identification of correlations, trends, and outliers, providing a foundation for further statistical analysis. The inclusion of both linear and polynomial regression lines allowed for the assessment of both linear and non-linear relationships between the heavy metal concentrations and satellite-derived data.

****

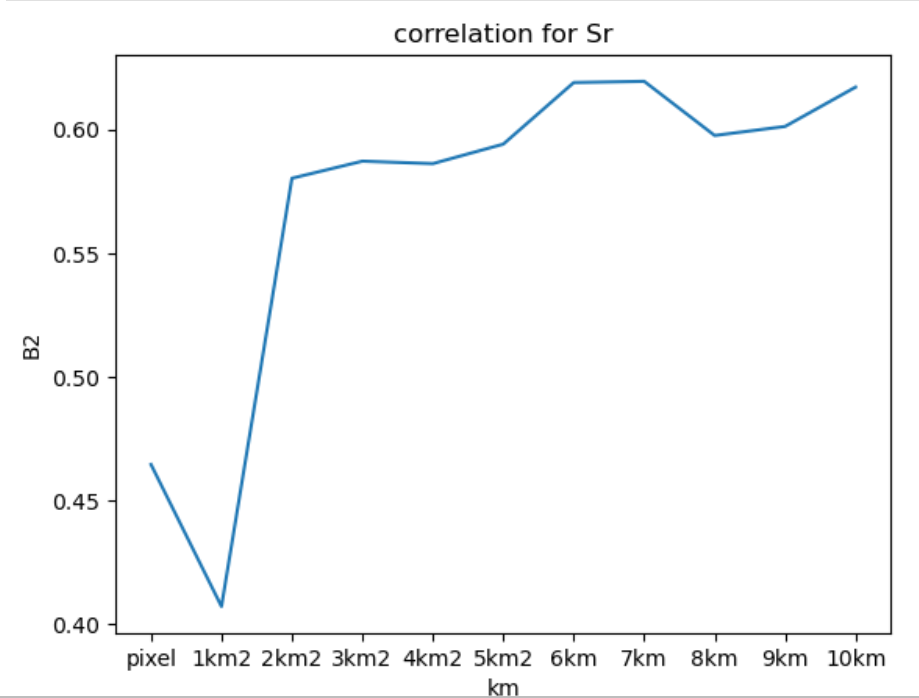
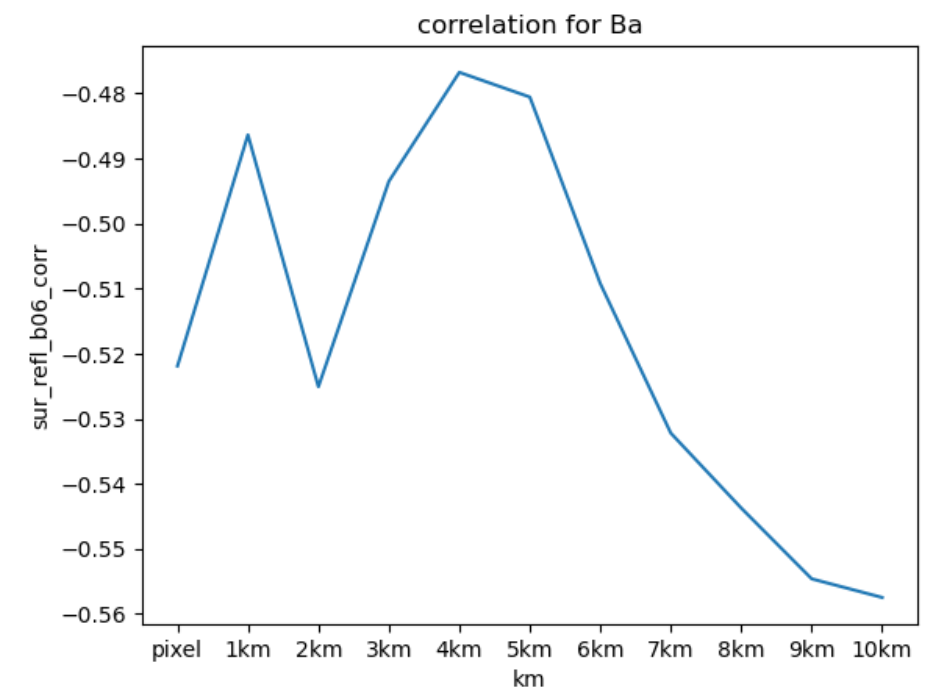
****

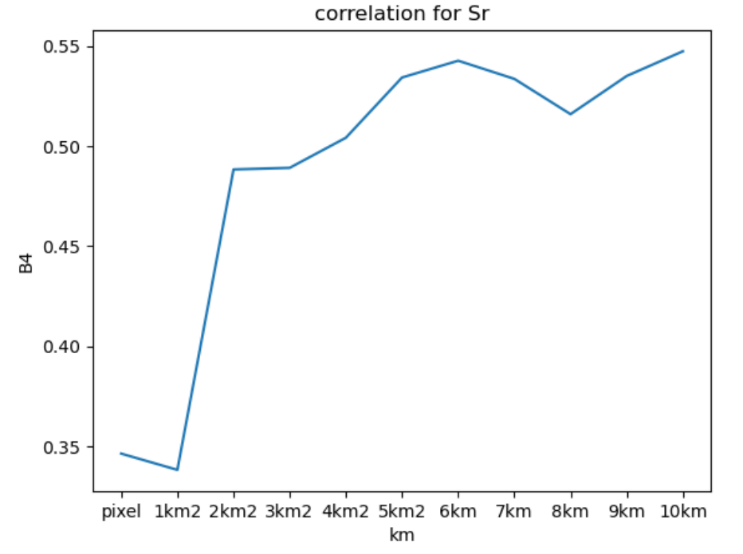
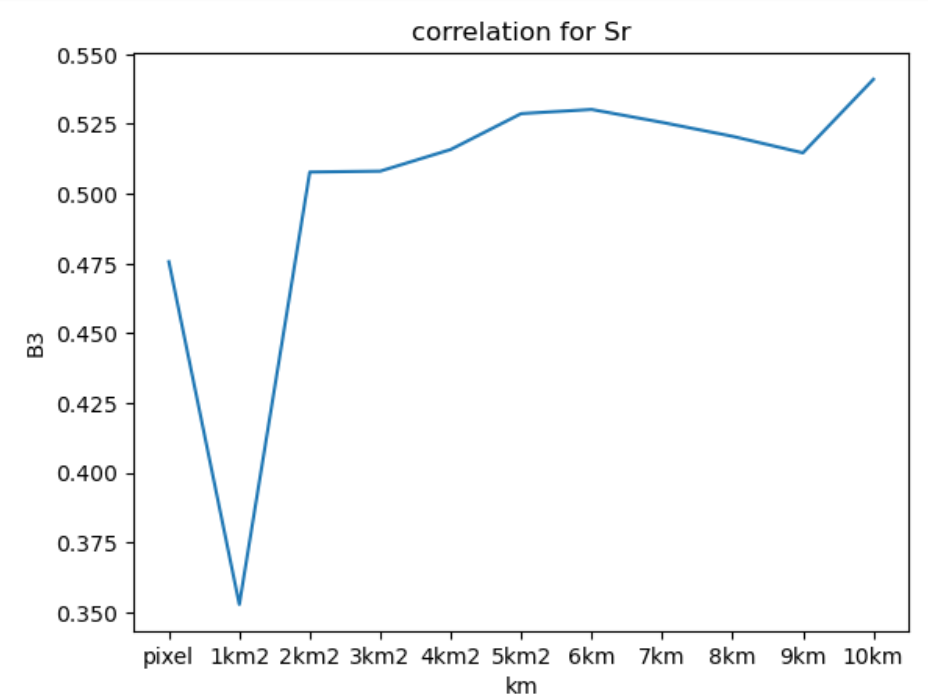
**“Scatter plots”**

**6.2** Linear plot

In addition to scatter plots, linear plots were employed to assess correlations across different spatial scales, ranging from the smallest scale of 1 pixel (1px) up to an area of 10 km². These plots were used to understand how correlations between the satellite data and element concentrations behaved across varying spatial extents. By examining these relationships over a range of spatial scales, it became possible to capture not just localized correlations but also broader patterns across larger regions.

The linear plots were particularly useful for examining satellite bands of interest—those that exhibited significant correlations with element concentrations across these different scales. By visualizing data from 1px to 10 km², the analysis could determine whether trends observed at the pixel level were consistent across larger areas, or if the correlations changed as the spatial extent increased. This scale-based analysis offered a comprehensive view, revealing hidden patterns and helping to draw meaningful insights about how spatial resolution impacts the relationship between satellite data and element concentrations.

****

****

**“Linear plots”**

**This visualization step** was crucial for confirming hypotheses, refining the focus on significant satellite bands, and ensuring that the analysis covered a broad range of spatial scales, all of which were necessary for gaining a comprehensive understanding of the data.

In this project, visualization served as the final and most important step of the analysis, as it provided a clear and insightful understanding of how the concentration of elements interacted with the satellite data. By transforming raw numerical data into intuitive visual forms, the visualization process enabled the identification of significant patterns and relationships, thereby enhancing the overall interpretability and impact of the analysis.

**Future plans and final results**

The correlation analysis yielded significant insights into the relationship between heavy metal (HM) elements in the soil and satellite-derived features. Notably, six to seven elements, such as Barium (Ba) and Strontium (Sr), exhibited connections, indicating the potential for using satellite data to predict the location of heavy metal contamination. This finding opens up exciting possibilities for future research and applications in environmental monitoring and management.

By identifying these key elements and their associated satellite bands, predictive models can be developed using machine learning techniques. These models can leverage readily available satellite imagery to estimate heavy metal concentrations in soil, even in areas where ground-based sampling is limited or expensive. This approach offers several advantages:

* **Cost-effective monitoring**: Satellite data provides a cost-effective and efficient way to monitor large areas for potential heavy metal contamination. Traditional methods, such as soil sampling, can be time-consuming, labor-intensive, and expensive, especially for vast areas. Using satellite imagery can significantly reduce these costs and allow for more frequent monitoring.
* **Early detection**: Predictive models can help identify areas at risk of contamination before it becomes a serious environmental issue. This early detection capability is crucial for preventing widespread contamination and mitigating its potential impacts on human health and ecosystems.
* **Targeted remediation**: By pinpointing areas with high contamination levels, remediation efforts can be focused on the most critical locations, optimizing resource allocation. This targeted approach ensures that remediation efforts are directed towards the areas where they are most needed, maximizing their effectiveness and minimizing unnecessary expenditures.

The results of this study provide a strong foundation for developing robust and reliable predictive models that can inform environmental management strategies and contribute to a more sustainable future. By integrating satellite data into our understanding of heavy metal contamination, we can enhance our ability to monitor, predict, and mitigate environmental risks, ultimately protecting human health and the environment.

Further research can explore the development of these predictive models, focusing on:

* **Model validation**: Evaluating the accuracy and precision of the models using independent datasets and field validation.
* **Model optimization**: Fine-tuning the models to improve their performance and predictive accuracy.
* **Integration with other data sources:** Combining satellite data with other environmental data, such as soil properties and land use information, to enhance model accuracy and provide a more comprehensive understanding of heavy metal contamination.

**By pursuing these research avenues, we can harness the power of satellite data to create a more sustainable and environmentally responsible future.**

**References**

* Radočaj, D.; Obhođaš, J.; Jurišić, M.; Gašparović, M. Global Open Data Remote Sensing Satellite Missions for Land Monitoring and Conservation: A Review. *Land* **2020**, *9*, 402[. https://doi.org/10.3390/land9110402](.%20https:/doi.org/10.3390/land9110402)
* Uzhinskiy A.; Vergel K.; Central Russia heavy metal contamination model based on satellite imagery and machine learning, Computer Optics 2023; 47(1): 137-151. DOI: 10.18287/2412-6179-CO-1149. [https://www.computeroptics.ru/KO/PDF/KO47-1/470116.pdf](%20https:/www.computeroptics.ru/KO/PDF/KO47-1/470116.pdf)
* Qiao, Z.; Sun, S.; Jiang, Q.; Xiao, L.; Wang, Y.; Yan, H. Retrieval of Total Phosphorus Concentration in the Surface Water of Miyun Reservoir Based on Remote Sensing Data and Machine Learning Algorithms. *Remote Sens.* **2021**, *13*, 4662. <https://doi.org/10.3390/rs13224662>
* Lovynska, V.; Bayat, B.; Bol, R.; Moradi, S.; Rahmati, M.; Raj, R.; Sytnyk, S.; Wiche, O.; Wu, B.; Montzka, C. Monitoring Heavy Metals and Metalloids in Soils and Vegetation by Remote Sensing: A Review. *Remote Sens.* **2024**, *16*, 3221. <https://doi.org/10.3390/rs16173221>
* Abdulraheem, M.I.; Zhang, W.; Li, S.; Moshayedi, A.J.; Farooque, A.A.; Hu, J. Advancement of Remote Sensing for Soil Measurements and Applications: A Comprehensive Review. *Sustainability* **2023**, *15*, 15444. <https://doi.org/10.3390/su152115444>