Using Python to Evaluate Sedimentation Parameters in Puerto Rico

Kolin Tyler Bilbrew

8 December 2024

1 Abstract

This project involves a secondary verification analysis of InVEST outputs used in the "Going with the flow: the supply and demand of sediment retention ecosystem services for the reservoirs in Puerto Rico". De Jesus Crespo estimated sediment retention ecosystem services for the watersheds upstream to major drinking water and power generation reservoirs in the commonwealth of Puerto Rico (De Jesus Crespo et al., 2023). InVEST Sediment Delivery Ratio model produced (estimates of the ranges of sediment retention). The manuscript validated those estimates by comparing the deposition rates reported by the Puerto Rico Sewage and Water Authority (PRSA), which manages the dams. Subsequently, et al developed validation methods using water quality data from stream gauges monitoring stations reported in the USGS EPA Water Quality Portal. To validate and calibrate the InVEST SDR Models, the developers suggest comparing, "total amount of sediment exported to the stream per watershed [estimated by the model, with] …. any observed sediment loading at the outlet of the watershed" (De Jesus Crespo et al., 2023).Here we obtain those data from WQP observations and compare them to invest outputs and PRASA deposition data used in the dataset provided(De Jesus Crespo et al., 2023).

2 Introduction

Puerto Rico is an island located within the Atlantic Ocean, covered with forested land, mountains, and agricultural crops. However, hurricanes and increased precipitation has become a threat to the island due to flash flooding (Murphy & Stallard, n.d.). It has affected the land by increasing runoff and sediment transport that creates erosion and polluting the water quality of rivers, lakes, and streams, which are used for drinking sources across the island (Yuan et al., 2015).

The island provides many environmental services that create many opportunities to improve environmental benefits that can impact many residents (De Jesus Crespo et al., 2023). The importance of reservoirs, which are used to manage water resources used for human consumption and agricultural use, provides a solution to improve water quality and building aquatic habitats. However, sedimentation affects the function of reservoirs and can reduce the flow capacity due to accumulation of sediments and pollutants. While evaluating sediment retention services, we need to understand the flows of the watersheds and how they can provide services for conservation and sustainable land practices. This requires our understanding of how different geological processes such as erosion, deposition, and sedimentation are intensified by the effects of climate change and infrastructure.

The main component of understanding their importance is collecting data based on the main parameters that can be utilized and create environmental benefits based on the site and how sediment retention services can be useful for building vegetation and sustainable land for the island of Puerto Rico(De Jesus Crespo et al., 2023). ArcGIS Pro is a great tool to visualize the locations of the watersheds and utilizing various datasets provided by the Water Quality Portal (WQP) and the United States Geological Survey. Furthermore, statistical analysis will be conducted in Python, a scientific programming system that is used

to construct data manipulation, cleaning, merging, and visualization to enhance the understanding of the scientific research involving sedimentation.

The primary objective of the project is to thoroughly understand the complexities of the environmental issue by visualizing and analyzing the amount of total dissolved and suspended solids based on the extensive data provided. This process involves evaluating the correlation between these various factors by comparing multiple parameters within the watershed, including water quality, sediment composition, and environmental conditions. By undertaking this in-depth analysis, we aim to generate valuable insights that could inform policy decisions and provide strong evidence to support the creation of additional vegetated land in Puerto Rico, ultimately contributing to enhanced land management strategies. Furthermore, our findings could help guide the development of innovative techniques for utilizing sediment trapped in reservoirs, offering long-term solutions to both environmental and economic challenges in the region.

3 Methods

3.1 Data Cleaning and Organizing Datasets via ArcGIS Pro and Excel

In this experiment, we used the data collected based on various parameters that understand sedimentation within different various drainages across the island of Puerto Rico for the year 2000 by using different databases. I was provided 4 datasets by Dr. Rebeca De Jesus Crespo, Mariam Valladares-Castellanos and Dr. Thomas Douthat, which conduct research under the Department of Environmental Sciences at Louisiana State University in Baton Rouge, Louisiana. After evaluating the datasets and creating a flow chart of the steps I needed to accomplish before the coding process in Python, we presumably recognized that the dams and monitoring stations in the datasets do not have common codes. I began the process by matching them spatially using ArcGIS Pro and selected the parameters that we want to provide, evaluate, and contained a significant amount of data for this project. Therefore, we imported four files that provided the following parameters: USGS identification number (USGS-MS), drainage name, area in square kilometers, total dissolved solids, total suspended solids, and sediment exports provided from InVEST, and dam deposition.

The first dataset contained the United States Geological Survey's watershed identification numbers and the names of the drainages, which provided 25 drainages. However, the locations within the survey consisted of duplicated drainages with different identification numbers, which required us to pivot into finding other alternatives to how we want to display and provide accurate data to figure out the sedimentation issue. We decided to add the sum and the average of the dissolved and suspended based on the provided data, which I will be displayed specifically in the Python Process. In the Water Quality Portal dataset, we converted the two parameters, total dissolved solids and suspended solids, from kilograms to tons per year based on the year 2000. The InVEST dataset, provided by Dr. De Jesus Crespo, were provided with the area of the drainage by square kilometers and the sediment export per tons. Lastly, we organized the dam deposition dataset by drainages and ensuring the data was converted to tons per year.

3.2 Working with Anaconda and Python

Once the datasets were clean, organized, and arranged, we began importing the data through Pandas to organize the data for the joining process. I was struggling to import the data because I didn't have the correct path for Python to find my datasets, so I created a folder to make it easily accessible. I installed the os package to make sure that the path and files needed for the project existed. I imported the Pandas package to import all of the datasets through the DataFrame feature to structure the data. From this process, I practice data manipulation by learning to join the datasets, which was a difficult step. The first dataset, which included the United States Geological Survey's watershed identification numbers and the names of the drainages, still included the duplicated drainages and showed that I had a total of 26 watersheds. I continued to import the second dataset containing the total dissolved and suspended solids and imported the third dataset including

After importing the data, I rearranged how I wanted to start the left joining process by placing an order of how I wanted to display and understand my data easier. First, I used the Pandas merge feature to combine the USGS and drainage names to the Water Quality Portal data to place the correct data with the location. However, there were still duplicates of certain locations, so I had to utilize an alternative to calculate the sum and mean of the total dissolved and suspended solids. I decided to group the different parameters to reduce the duplicates into one location and organize the table to make a graph. In this code, I organized the dataset to group the table by the following: drainage, total dissolved load sum, total dissolved load mean, total suspended load sum, and total suspended load mean. After successfully completing the code, the duplicates were removed and reduced the number of drainages went down to 12 drainages. Next, a graph was formed to understand the differences of total dissolved and suspended solids in various drainages. I also decided to create a graph to display the differences in the averages of the two parameters for the available parameters. Next, the merged file created was coded into the invest file to include the sediment exports in the year 2000. Lastly, the merged invest file was merged into the last dataset, which is the dam deposition dataset, which completes the process of importing files into Python to begin data visualization through Matplotlib and Seaborn.

3.3 Data Visualization through Matplotlib and Seaborn

After merging the 4 datasets into one, we created 3 different figures to understand the parameters of each watershed and their impact on Puerto Rico's sedimentation issue. After completing the merging processes, I decided to create bar graphs to understand the differences by categorizing them, which is a simple and easy way to compare the data and figure out the effects of the watershed in this particular year by the mean and sum of the total dissolved and suspended solids. Soon, I wanted to see the correlation and the accuracy of each parameter to recognize their relationships, so we created a matrix map to represent the positive and negative correlations of the parameters. In this graph, I created this graph through the Python package named Seaborn, which simplifies the complexity of statistical analysis and is easily compatible with Pandas DataFrames. The matrix is represented by the positive correlation in purple and the negative correlation in gold to emphasis the skills I learned through the class and Louisiana State University. Lastly, we created a regression analysis to analyze the predicted and actual values of each parameter. Creating data visualization of this project provides the results to better understand the benefits and consequences of various processes and how we can use the data to form sustainable practices.

4 Results

4.1 Analyzing the Total Dissolved and Suspended Solids data

After the dataset, I evaluated the correlation matrix heatmap to compare the different parameters in the final merged code. The metric is calculated based on the relationships and patterns between the variables. The first observation from the correlation matrix is that the area by square kilometer has a strong relationship with all the parameters measured throughout the research (Figure 1). The area is important because it is the capacity in which sediment can transport based on the size of the watershed. The highest total dissolved loads is La Plata, which transports 106,548.76 tons per year. However, comparing the area in square kilometers, the watershed meausres at 467.45 square kilometers, which is ranked 2nd compared to Loiza that has a area of 537.91 square kilometers. It's clear that the area plays a huge factor in sedimentation and leads to other questions about topography and velocity flow. The second observation was that muliple parameters had negative correlations with the average and total suspended solids. The InVEST sediment exports per year and dam deposition displayed the lowest numbers with the average and total suspended solids due to the reserviors trapping the sediment and disrupting the flow into drainage areas. The reason for higher total suspended solids could be caused by steeper slopes, increased precipatation, or clay soils with low erodibility (Murphy & Stallard, n.d.). Places with reduced vegetative cover has a higher percentage of soil loss. In the "Sum of Total Dissolved and Suspended Loads by Location" bar graphs, we evaluated that La Plata and Dos Bocas had the highest total dissolved solids out of the other watersheds with 106,548.76 and 89,018.09 tons in the year 2000 (Figure 2). Canovanas and RioBlanco had the lowest total dissolved solids with 3235.13 and 577.38 tons. The highest average in total dissolved solids was LaPlata and the highest in suspended solids were Loiza (Figure 3). This allows us to look into the watersheds with the most deposition with 554,400 tons is Coamo, which could be used for dredging, which is ane effective way to create wetlands and forested areas (De Jesus Crespo et al., 2023). Three other watersheds were over 100,000 tons, which were Guajataca, LaPlata, and Loiza.

5 Discussion

5.1 Limitations of Other Factors Within Datasets

In this project, we were limited to a small amount of parameters within the datasets due to the amount of time I had to complete the project. I learned about the importance of time manage,ent and creating a plan that allows me to focus on different parts of research. Technological issues and learning new coding techniques allowed me to display the parameters that were given. There were other parameters, such as turbidity, that I could have included, but most of the watersheds lacked the data to be used in this project. Furthermore, the mapping process in ArcGIS Pro was excluded due to the batch delineation and caibration process taking longer due to the technological issues experienced over the last few weeks.

5.2 Challenges of Pandas and Merging Datasets

The coding process also had some challenges because this is my first project importing data and utilizing skills such as data manipulation and visualization into a topic related to environmental science and policy. The first challenge I encountered was importing the data into Python. In the beginning, I tried to import the data as it was given without thinking about organizing the datasets to make sure it contained similar parameters, converting the data into the same units, and finding solutions to reduce the number of drainages due to various limitations. After organizing the datasets, with the help of Mariam Valladares-Castellanos and Dr. Thomas Douthat, I began the process of coding in Python. The next challenge was realizing the process of the path to find the datasets, which caused me to download the os package to ensure that figure out the location and if the files existed. I solved the issue and began to import data based on a flowchart to start the joining process of the project.

6 Conclusion

Utilizing ArcGIS Pro and Python can provide different tools to evaluate and analyze environmental and hydrological parameters that impact the water quality and land loss in areas across different regions in Puerto Rico. In this project, we use the various skills that we've learned throughout the semester, including data collection, spatial analysis, and coding techniques, to properly display how different watersheds can provide sediment retention services by protecting water quality, restore vegetation, and mitigate flood risk to protect the residents. This emphasizes the importance of protecting the well-being and historical significance of Puerto Rico by using scientific programming to foster sustainable, long-term practices that provide long-term benefits over time.

This analysis evaluates the relationships between various watershed parameters and sediment transport, providing valuable insights into how factors like area, topography, and vegetation impact sedimentation patterns. The correlation matrix revealed that larger watershed areas strongly influence sediment transport capacity, as observed in watersheds like La Plata and Loiza. While La Plata had the highest total dissolved loads, Loiza's larger area emphasized the role of watershed size in sediment dynamics. Negative correlations between dam deposition and suspended solids underscore the impact of reservoirs in trapping sediment and altering natural flows, with steep slopes, precipitation, and soil types influencing suspended solid levels. Watersheds with reduced vegetative cover experienced greater soil loss, highlighting the importance of vegetation in erosion control. Bar graphs comparing total dissolved and suspended loads across watersheds revealed that La Plata and Dos Bocas led in dissolved loads, while Loiza dominated in suspended solids. Coamo, with the highest deposition at 554,400 tons, offers potential for sustainable practices like dredging to create wetlands or forested areas. These findings emphasize the interconnectedness of watershed characteristics and sediment transport, informing effective management and conservation strategies.

7 Figures

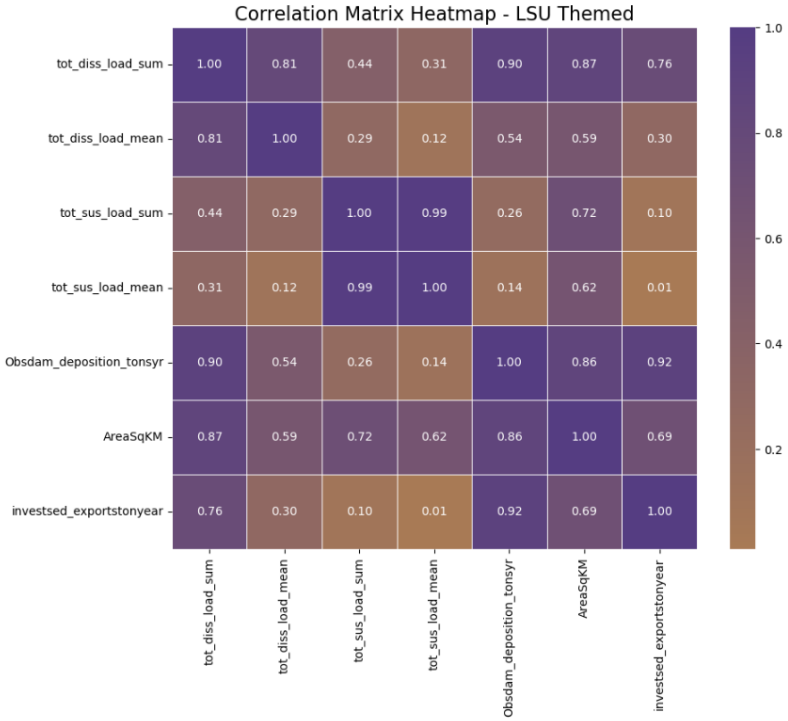

Figure 1: Python displaying the total dissolved and suspended loads of each watershed location in the year 2000. These parameters are measured by tons and represent the sediment transport over the course of the year. In the matrix, the positive correlations are represented in purple and the negative correlation in gold.

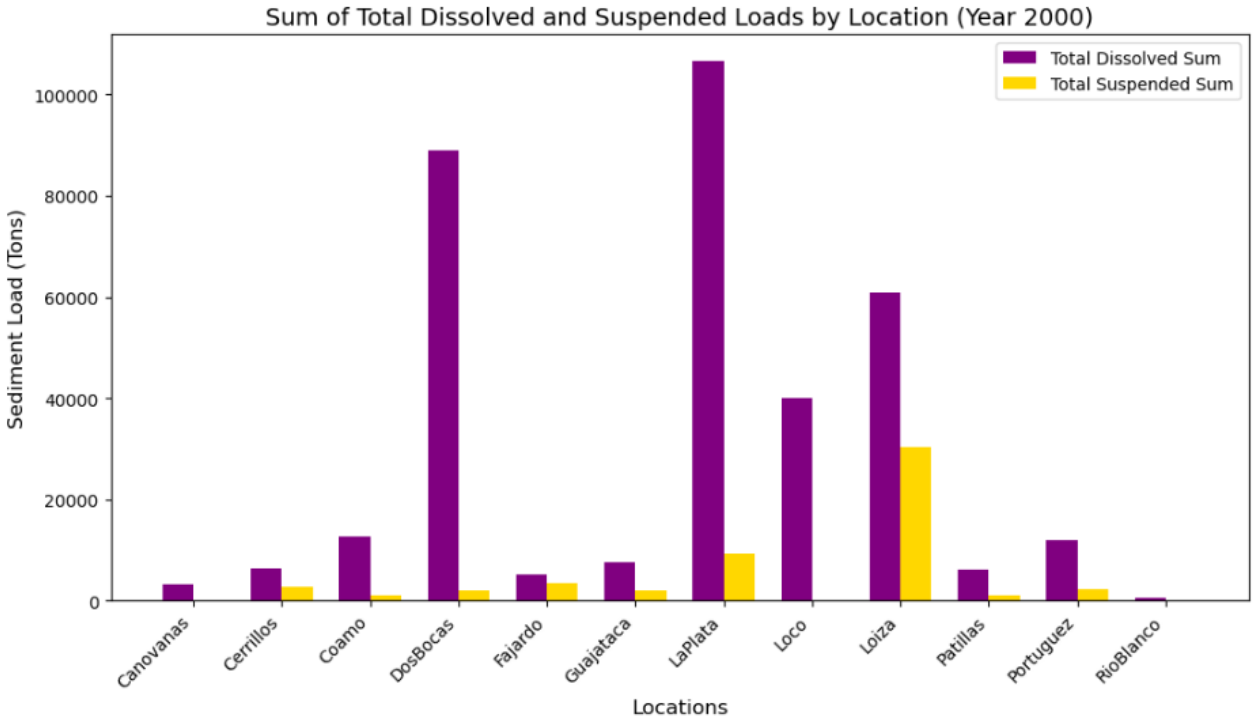

Figure 2: Python representing the average sediment transport of total dissolved and suspended solids in the year 2000. Three watersheds were not displayed because they didn't have data consisting of these parameters during the year (Canovanas, Loco, and RioBlanco).

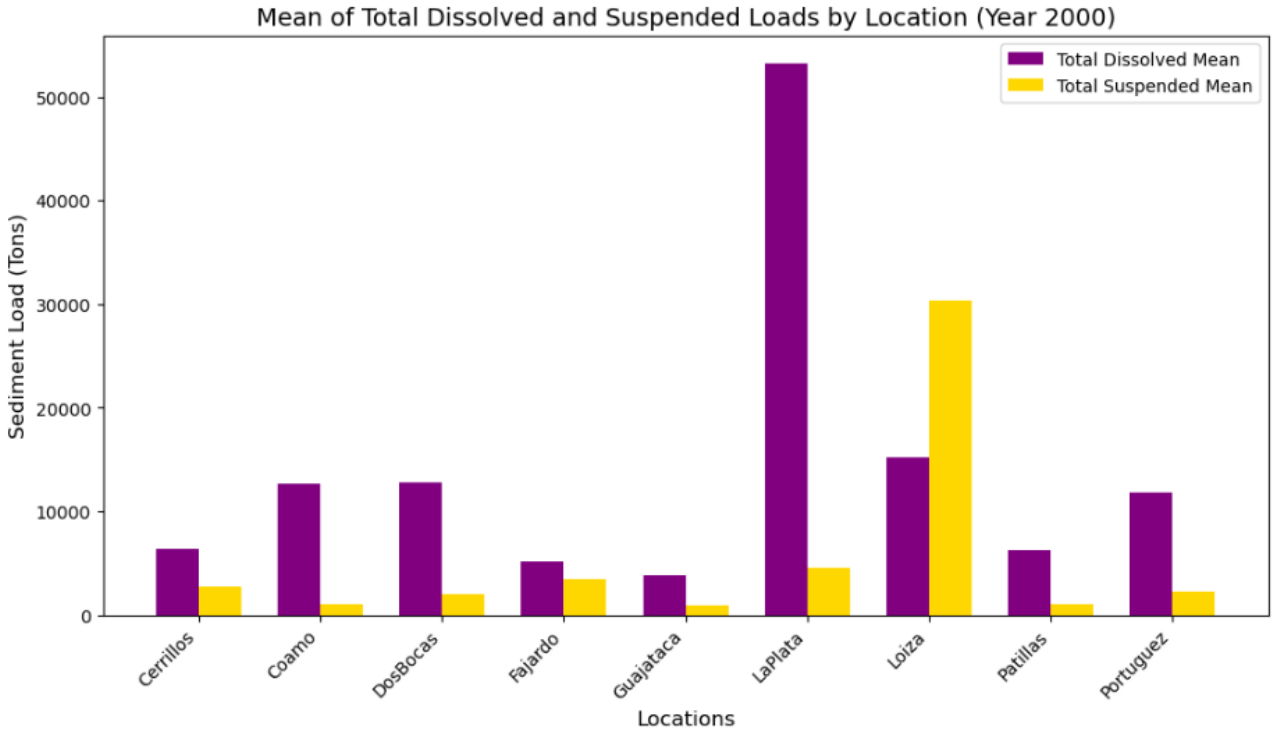

Figure 3: Bar graph represents the average of the total dissolved and suspended loads by location for the year 2000. Three watersheds are not shown due to lack of data (Canovanas, Loco, and RioBlanco),

8 References

De Jesus Crespo, R., Valladares-Castellanos, M., Mihunov, V. V., & Douthat, T. H. (2023). Going with the flow: The supply and demand of sediment retention ecosystem services for the reservoirs in Puerto Rico. *Frontiers in Environmental Science*, *11*. https://doi.org/10.3389/fenvs.2023.1214037

Murphy, S. F., & Stallard, R. F. (n.d.). *Hydrology and Climate of Four Watersheds in Eastern Puerto Rico*.

Yuan, Y., Jiang, Y., Taguas, E. V., Mbonimpa, E. G., & Hu, W. (2015). Sediment loss and its cause in Puerto Rico watersheds. *SOIL*, *1*(2), 595–602. https://doi.org/10.5194/soil-1-595-2015