Extended Abstract: Anomaly Detection of Records in a Reservoir Sedimentation Dataset

Alejandra Botero-Acosta, Research Scientist, WATER Institute, Saint Louis University, St. Louis, MO, alejandra.boteroacosta@slu.edu
Amanda Cox, Associate Professor, WATER Institute, Saint Louis University, St. Louis, MO, amanda.cox@slu.edu
Vasit Sagan, Associate Professor, Taylor Geospatial Institute, Saint Louis University, St. Louis, MO, vasit.sagan@slu.edu
Ibrahim Demir, Associate Professor, University of Iowa, Iowa City, IA, ibrahim-demir@uiowa.edu
Marian Muste, Research Engineer, University of Iowa, Iowa City, IA, marian-muste@uiowa.edu
Paul Boyd, Hydraulic Engineer**,** US Army Corps of Engineers, Omaha, NE, paul.m.boyd@usace.army.mil
Chandra Pathak, Senior Engineer/Policy Advisor, US Army Corps of Engineers, Washington, D.C. chandra.s.pathak@usace.army.mil

Introduction

Although our nation's reservoirs provide critical benefits to our communities, the sedimentation processes that continuously occur in reservoirs can jeopardize their water-related functions and compromise dam safety. The U.S. Army Corps of Engineers (USACE) operates several U.S. dams, with some approaching 100 years of operation (Pinson et al., 2016). Several reservoirs in the US are approaching the end of their economic design life, and climate change and associated hydrologic uncertainty are introducing additional stressors to these vital systems. The large life span of these structures and their hydraulic characteristics make them susceptible to significant sedimentation processes. The consequences of sedimentation on reservoir functionality include capacity loss, water abstraction prevention due to buried intakes, navigability reduction, and damage to recreational areas. Increased sedimentation rates translate into increased maintenance costs to recover reservoir functionality (Sholtes et al., 2018). In addition to the fact that most of these structures are approaching the end of their designed operating periods, uncertainties about the operations of U.S. reservoirs are continuously rising as many of them are experiencing an increasing frequency of extreme hydrologic events.

Watershed and water-resources managers are working on developing sustainable sediment management plans to ensure the continuation of reservoir functions and reduce maintenance costs under current and future climate conditions. Frequent reservoir capacity surveys are fundamental to assessing lost volume storage as well as sedimentation rates resulting from upstream erosion processes. The analysis of historical surveys enables the identification of past and current trends in reservoir sedimentation rates, as well as the assessment of aggradation trends, life expectancy, and reservoir vulnerabilities to climate change. This information is essential for the development of effective management plans for reservoir operation, maintenance, and upstream erosion control that include climate preparedness and resilience aspects. Considering the relevance of historical reservoir survey data for the nation's water resources, the USACE implemented the Enhancing Reservoir Sedimentation Information for Climate Preparedness and Resilience (RSI) system to compile and assess data for over 700 dams. This RSI system is primarily composed of elevation-capacity and elevation-surface area data derived from surveys.

The objective of this study was to identify anomalous and potentially erroneous data within the RSI dataset by applying machine learning techniques. Detecting anomalous records improves the quality of the RSI dataset and any research projects that will use this information. Furthermore, the extracted information can be utilized to better understand the mechanisms of sedimentation and capacity loss in U.S. reservoirs.

Methodology

Data from 184 reservoirs in the RSI dataset were combined with data related to watershed processes that affect erosion and sedimentation, such as basin topographic features, upstream reservoir properties, land use/landcover features, and precipitation descriptors. Multivariate relationships within the dataset were analyzed and interpreted to identify and investigate the presence of anomalous data. Machine learning techniques are particularly useful in this dataset given the numerous parameters involved in erosion and sedimentation processes, and the large number of reservoirs and records. Initially, preliminary filtering of the dataset was conducted to remove records with negative sedimentation rates and/or duplicate records. Subsequently, two unsupervised machine learning methods, the Autonomous Anomaly Detection (AAD) and the Kolmogorov-Smirnov and Efron (KSE) anomaly detection methods, were used to evaluate the remaining records to further identify possible erroneous data based on the multidimensional space and their relative location within the data cloud. The AAD, first proposed by Angelov (2014), is based on the Empirical Data Analytics (EDA) approach, while the KSE (Kim, 2013) employs multiple implementations of Kolmogorov-Smirnov tests with resampling. Preliminary results demonstrated that variable scale discrepancies and zero values impacted the performance of the automated anomaly detection. Therefore, data transformation and normalization techniques were applied to the composite dataset to reduce the bias from records having relatively large or zero values. Flagged records and variable relationships were analyzed through the Principal Component Analysis (PCA) and the K-means clustering method. A Principal Component Analysis (PCA) was initially conducted to explore and visualize the variability of the dataset and analyze relationships existing between variables. PCA, a multivariate and statistical method frequently applied to interpret the variability of large environmental datasets, was selected for this analysis because it does not require the data to follow any distribution.

Results and Conclusions

This study performed a multivariate analysis, diagnosis, and interpretation of the RSI composite dataset housing information from U.S. reservoirs managed by the USACE. The dataset is composed of 30 numerical variables that include reservoir capacity, reservoir sedimentation, and watershed variables related to reservoir sedimentation (e.g., hydrologic, geologic, and topographic parameters). Prior-knowledge filtering, two machine learning techniques, AAD and KSE, and a multivariate analysis, PCA, were used to identify likely erroneous data, as well as to investigate relevant information and relationships within this unique dataset. The initial filtering was performed based on the physical meaning of the variables, while the unsupervised machine learning methods evaluated the remaining records based on the multidimensional space and their relative location within the data cloud.

The variables holding most of the data cloud variance were grouped by the PCA as follows: 1) basin topographic features; 2) dam properties and basin extent; 3) forested area and average precipitation; and 4) geo-location descriptors and maximum precipitation. The PCA analysis indicated that sedimentation rates and capacity losses were primarily related to drainage basin size and potential runoff processes, while being independent of elevation related properties. Given the multidimensionality of the dataset, the PCA was a powerful analysis tool to visualize the location of categorical variables, K-means clusters, and records flagged as anomalous within the data cloud. EPA ecoregions with larger reservoir capacity losses either belonged to the Great Plains or the Eastern Temperate Forests, as opposed to Mediterranean California and the Northwestern Forested Mountains, which had smaller capacity losses.

The anomaly detection methods used in this study were designed to detect local and global anomalies identified from data clusters and data ensembles, respectively. The AAD and KSE methods flagged 20 reservoirs for having anomalous records, of which 5 were identified by both methods, and 6 had more than one record detected. The single-variable outlier analysis for anomalous records allowed the identification of variables potentially causing these records to be flagged. Certain variables related to elevation characteristics (channel slope, average drainage basin slope, and minimum drainage basin elevation), precipitation trends (median monthly precipitation between surveys and cumulative precipitation between surveys), dam properties (years since dam completion and initial reservoir trap efficiency), and watershed properties (curve number) were found to have values more than four times the standard deviation apart from the mean.

Further development of the RSI composite dataset could consider the addition of watershed variables that can potentially influence sedimentation and erosion processes. For example, the inclusion of hydrological variables such as mean and maximum streamflow, and percentage of agricultural land could provide new information associated with soil particle detachment and transport processes. In addition, the temporal variation of the reservoir variables, such as capacity, trap efficiency, and upstream reservoir storage, could be incorporated to reflect the updated state of the reservoir and drainage basin. Finally, the normalization of capacity loss and sedimentation rates by the basin's area could enable the identification of further relationships within the dataset.

The RSI dataset is a potential major data source for large scale studies related to sedimentation rates and suspended sediment loads in freshwater systems due to the spatial and temporal scale of its records. Apparent erroneous data, related to duplicate records or increases in reservoir capacities, can be manually flagged through visual inspection. However, the detection of anomalies in an automated and fully data-driven way represents a powerful tool for the maintenance and monitoring of this large and heterogeneous dataset. The flagged records should be analyzed and verified by managers and operations staff and handled with caution by RSI dataset users. This research highlights the challenges related to data analysis and anomaly detection of large datasets containing variables of a heterogeneous nature, such as the composite Reservoir Sedimentation Information System. Raw values and units facilitated the initial knowledge-based filtering, given the physical meaning of variables and the experience of the user developing the task. However, this heterogeneity prohibited the automated detection of anomalous records from raw data and required the implementation of transformation techniques that reduced the bias introduced by scale differences and null values, preserving the composition of the data cloud and the relative location of records within it.

Acknowledgments

This project was funded by U.S. Army Corps of Engineers. Federal Award Identification Number (FAIN): W912HZ-20-2-0059.

References

Angelov, P. (2014) Outside the box : an alternative data analytics framework. *Journal of Automation Mobile Robotics and Intelligent Systems,* Vol. 8, No. 2, 29-35. DOI 10.14313/JAMRIS_2-2014/16.

Kim, M. S. (2013) Robust, scalable anomaly detection for large collections of images. Institute of Electrical and Electronics Engineers (IEEE), 2013 International Conference on Social Computing. 8-14 September 2013. DOI: 10.1109/SocialCom.2013.170.

Pinson, A., Baker, B., Boyd, P., Grandpre, R., White, K. D. & Jonas, M. (2016) U.S. Army Corps of Engineers Reservoir Sedimentation in the Context of Climate Change. *Civil Works Technical Report, CWTS 2016-05*. Washington DC.: U.S. Army Corps of Engineers.

Sholtes, J. S., Ubing, C., Randle, T. J., Fripp, J., Cenderelli, D. & Baird, D. C. (2018) Managing Infrastructure in the Stream Environment. *Journal of the American Water Resources Association,* 54(6), 1172-1184. 10.1111/1752-1688.12692.