Analyzing and Helping Understanding Reservoir Level Using K-Means Clustering

Jian Yoo

Received September 07, 2024
Accepted November 08, 2024
Electronic access November 30, 2024

This paper examines the application of K-means clustering to analyze reservoir storage data in California, utilizing data sourced from the United States Geological Survey (USGS). The study aims to categorize reservoir storage levels into distinct clusters, revealing patterns in water storage behaviors and providing insights into similarities between reservoirs that may not be obvious from individual data analysis. The methodology involved data collection, preprocessing to ensure accuracy, and implementing K-means clustering using a Java program to identify meaningful patterns and trends. The results revealed three clusters representing reservoirs with similar storage behaviors. The findings demonstrate the potential of K-means clustering to enhance the interpretation of complex environmental data, although further research incorporating additional variables and longer time frames is recommended. A limitation of relying on absolute water levels also existed. Further research is recommended with normalized data and longer timeframes to uncover additional trends. This study highlights the potential of clustering to enhance water management in the context of increasing climate variability.

Keywords: California reservoirs, K-means clustering, water resource management, data analysis

Introduction

Effective water resource management is a growing concern in regions facing significant climatic variability, particularly California. The state is known for its prolonged droughts and fluctuating water availability, making reservoirs crucial for water storage, flood control, hydropower generation, and agricultural irrigation. Managing reservoir storage levels effectively is vital for sustaining the state's economy and environment, especially in the face of climate change, which adds to the unpredictability of water resources. While optimizing water distribution, forecasting capabilities, and long-term conservation strategies rely heavily on understanding reservoir storage, the challenge lies in finding patterns that can enhance decision-making in these areas.

Existing research, such as studies [1–4], has primarily focused on predictive methods to forecast reservoir levels using machine learning models, statistical techniques, and hydrological simulations. These approaches aim to predict future water levels based on climatic, geographic, and hydrological factors. However, a notable gap exists in applying unsupervised learning techniques—specifically K-means clustering—to reservoir data. Unlike predictive models, clustering methods do not require predefined labels or outcomes and can reveal unexpected patterns in the data. This study seeks to explore what insights can be uncovered by applying K-means clustering to time-series data of reservoir water levels. The key question is not necessarily about confirming known patterns but about exploring what clustering might reveal.

Clustering techniques, particularly K-means clustering, automatically segment complex datasets into groups. Previous studies [5–7], have shown the efficacy of K-means clustering in environmental and water resource management contexts. The motivation behind using K-means clustering in this context is to answer: *What patterns emerge when we group reservoirs based on their water level trajectories over time*? Rather than approaching this analysis with predefined expectations, we apply K-means clustering to investigate whether reservoirs with seemingly different geographic or capacity characteristics might exhibit similar storage behavior over time. These similarities could reflect broader environmental factors or shared operational practices, insights that may not be easily visible through traditional analysis. Additionally, by analyzing clusters formed from sequential water level data, we aim to understand whether clustering reveals trends that provide useful information for water resource management. This study seeks to apply K-means clustering to reservoir storage data from California, sourced from the United States Geological Survey (USGS) [8].

To our knowledge, this is one of the first studies to apply K-means clustering to sequential data where water levels at different time steps are used as features. By experimenting with this clustering technique, we aim to uncover any underlying trends or behaviors that might inform water management strategies. While the results are exploratory, they provide an

© The National High School Journal of Science 2024

opportunity to see whether the method yields surprising or unexpected groupings of reservoirs based on their temporal water level patterns. Any insights gained could open new pathways for managing California's reservoir systems in the face of growing environmental challenges.

The purpose of this study is to investigate whether clustering can contribute to our understanding of reservoir behaviors. By applying K-means clustering to California's reservoir storage data, we hope to shed light on patterns that might inform future water management policies, including drought mitigation, flood prevention, and conservation planning. Ultimately, this exploratory analysis aims to contribute to the broader field of environmental data analysis by showing how clustering techniques can be applied in novel ways to time-series data.

Methods

Data Collection

The dataset for this study was sourced from the United States Geological Survey (USGS) website, specifically from the 'Water Quality' section under 'Historical Observations.' The dataset comprises reservoir storage levels, measured in acre-feet, across various locations in the United States. I specifically selected California as a target location for analysis. The data was downloaded in CSV format, encompassing specific date ranges and locations as selected from the USGS database. Each data entry includes six columns: agency code, site number (a unique identifier for each reservoir location), date and time, timezone code, storage level, and data value qualification code, which indicates the approval status of the data. An example of a data entry is 'USGS 09427500 2024-03-11 00:00 MST 560200 A,' where the elements represent the respective columns.

Data Processing

To ensure the dataset's integrity and usability for K-means clustering, several preprocessing steps were undertaken:

1. **Counting Total Entries and Verifying Data Structure:**
   Initially, the dataset was processed to count the total number of entries and verify that each followed the expected six-column structure. This was achieved using a Java program that read the CSV file line-by-line, incrementing a 'totalEntries' counter. The data was split using the 'split("")' method, and the 'parseLine' method verified that each entry contained exactly six fields ('if (parts.length != 6) return null;'). The 'siteNumber' and 'dateTime' fields were validated by parsing them as an integer and a 'Date' object, respectively. Entries failing these checks were excluded from further analysis.

2. **Handling Missing Values:**
   Entries with missing or incomplete fields were identified and excluded to maintain data integrity. The 'nullValueCount' variable tracked these entries. The 'parseLine' method checked for empty strings in critical fields like 'source,' 'timeZone,' and 'flag' ('if (source.isEmpty() —— timeZone.isEmpty() —— flag.isEmpty())'), and entries with missing data were marked as null.

3. **Identifying and Removing Duplicate Entries:**
   Duplicate entries were identified and removed to prevent skewing the analysis. A 'HashSet¡String¿' called 'uniqueEntries' was used to track unique data entries based on a combination of 'siteNumber' and 'dateTime'. For each valid entry, a unique key was generated ('String uniqueKey = entry.getSiteNumber() + "_" + entry.getDateTime().toString();'). If this key was already in the 'uniqueEntries' set, the entry was deemed a duplicate and added to the 'duplicateEntries' list. Only unique entries were retained in the 'dataMap.'

4. **Data Cleaning and Interval Selection:**
   A further data cleaning step involved standardizing the date intervals and ensuring consistency across locations. The 'DataCleaner' class was utilized to identify common dates across all locations. The 'cleanData' method filtered the dataset to retain only those entries on these common dates. The method used a 'HashSet¡Date¿' to store and intersect dates from all data entries, ensuring only common dates remained. Additionally, the dataset was reduced by setting a specific time interval (e.g., 14 days) between data points for each location. This was done by iterating through the data and selecting entries that matched the set interval, removing intermediate entries to standardize the data points across all locations. This step was crucial to ensure comparability across different locations and periods. The 'totalDeletedEntries' variable tracked the number of entries removed during this cleaning process, providing an overview of data reduction.

K-Means Clustering Implementation

The K-means clustering algorithm was employed to categorize the reservoir storage data into distinct groups, providing a simplified representation of storage levels across various California locations. Unlike traditional clustering with just a few features, we used reservoir levels measured at multiple time points as the features for clustering. Each feature corresponds to a measurement of the reservoir level at a specific time, making the number of features equal to the number of dates the water level was measured. For this analysis, the reservoir levels were measured at 15 different times, making it effectively clustering in 15-dimensional space. This approach is novel in that it applies

© The National High School Journal of Science 2024

K-means clustering to sequential data, capturing patterns in the variations of reservoir levels rather than clustering based on static attributes.

1. **Feature Selection:**
   The feature used for clustering was the storage level measured over time. Each reservoir was treated as an entity with 15 features, one for each water level measurement across different time intervals. This multi-dimensional approach allows us to cluster reservoirs based on their water level trajectories and variations over time, rather than just static attributes like total storage capacity or geographic location. This allows for identifying reservoirs with similar behavior in terms of how their water levels change over time. We hypothesized that clustering this way would create clusters that capture high-level patterns in reservoir water level trends that we might not have detected otherwise. The clustering here is performed on high-dimensional data (15 dimensions in this case), where each dimension represents a measurement of the water level at a specific point in time. While clustering based on just two features might enable a simple 2D plot, such an approach would miss the complex relationships captured over time. K-means clustering is well-suited to this high-dimensional space because it groups reservoirs that have similar water level trajectories across multiple time points, something that cannot be effectively visualized in two dimensions. This allows for a more robust analysis of water level patterns.

2. **Initialization of Clusters:**
   The number of clusters, k, was determined using the Elbow method, which involved plotting the within-cluster sum of squares (WCSS) against different values of k and selecting the point where the rate of decrease in WCSS diminishes. For this study, k was set to 3, indicating three distinct clusters. We implemented a KMeans algorithm in Java to implement the algorithm, initializing clusters with random centroids selected from the dataset. This random initialization helps start the clustering process by providing initial reference points for each cluster.

3. **Assignment of Points to Clusters:**
   Each data point was assigned to the nearest cluster based on the Euclidean distance between the point and the cluster centroids. The 'assignPointsToClusters' method iterated through all points, calculating the distance to each centroid using the 'calculateDistance' method. Points were then assigned to the cluster with the closest centroid. This step was crucial for grouping data points based on similarity, as measured by their distance from the centroids.

4. **Recalculation of Centroids:**
   After assigning all points to clusters, new centroids were calculated for each cluster by averaging the coordinates of all points within the cluster. The 'calculateCentroid' method was used to compute these new centroids, which became the reference points for the next iteration of point assignment. The centroids were updated iteratively, with the process repeating until the centroids stabilized (i.e., the positions of the centroids did not change significantly between iterations). This iterative process is a standard feature of the K-means algorithm, aimed at refining the clusters to represent the data better.

5. **Convergence and Final Clusters:**
   The algorithm continued iterating, reassigning points, and recalculating centroids until the centroids ceased to change significantly, indicating convergence. The final set of clusters was then obtained, each containing a group of points representing similar reservoir storage levels. The 'getClusters' method provided access to the final clusters, each represented by a centroid and a list of data points.

Determining the Optimal Number of Clusters (k)

To determine the most appropriate value for k in the K-means clustering algorithm, I employed the elbow method, a commonly used technique for identifying the optimal number of clusters. The elbow method involves plotting the within-cluster sum of squares (WCSS), which measures the total variance within each cluster against different values of k. The goal is to identify a point called the "elbow," where increasing k provides diminishing returns in reducing WCSS, indicating that additional clusters do not significantly improve the model.

The following process was implemented in the main method of the ReservoirLevels class:

1. **Calculation of Average Distance:** For each value of k, I calculated the average distance between data points and their respective cluster centroids. This was done by assigning points to the nearest centroid, recalculating the centroids, and repeating the process until convergence. This process was repeated for k values ranging from 1 to 10, averaging over 10 random initializations per k value to ensure consistent results. I recorded the average distance for each k value and plotted the results.

2. **Elbow Method Analysis:** The graph generated from this analysis, as shown in Figure 1 (the elbow plot), revealed that the WCSS decreased sharply as k increased from 1 to 3. However, after k=4, the decrease in WCSS became less pronounced, forming an "elbow" in the graph. This indicated that while increasing k beyond 3 or 4 clusters would further reduce the WCSS, the benefit of doing so diminished rapidly. Furthermore, considering the relatively small dataset of 21 reservoir data points, having more than four

clusters could result in overfitting, where clusters become too specific and lose meaningful interpretation.

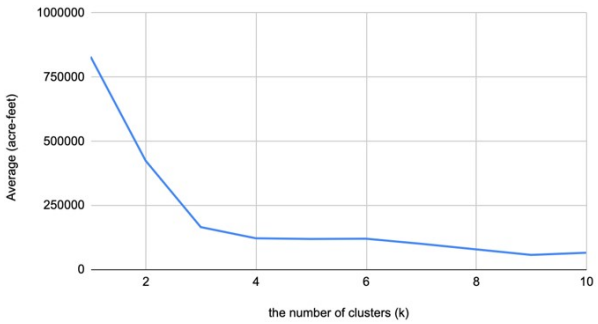

Fig. 1 Average reservoir level for each number of cluster. A sharp decline between k=1 and k=3 is evident, followed by a more gradual reduction after k=4

3. **Final Choice of k:** Based on the elbow method and the limited dataset size, I determined that k=3 or k=4 would be optimal for this analysis. Both values provided a balance between reducing WCSS and maintaining a reasonable number of clusters that would capture the underlying patterns in the data without overcomplicating the model. I selected k=3 for the final clustering analysis for this paper. Since we have relatively few data points (21), using fewer clusters is simpler and sufficient for our analysis.

Mapping and Graphing

To visualize the geographic distribution of the reservoirs across the different clusters, I utilized Google My Maps. The reservoirs were plotted as points on the map, with each point corresponding to a specific reservoir location. The reservoirs were then grouped into layers according to their assigned clusters from the K-means analysis. Each cluster was assigned a unique color to differentiate the groups visually, allowing for a clear comparison of the spatial patterns across California. This approach helped to identify geographic trends and potential regional factors influencing reservoir behavior.

I used Google Sheets to graph the data for the temporal analysis of reservoir levels. The process began with exporting the reservoir-level data from the Java program, which had already been processed and clustered. This data was then organized into a table within Google Sheets, with each reservoir's data placed in rows corresponding to the time series from January 1 to July 15, 2024. I then plotted line graphs for each cluster, with the y-axis showing the reservoir level (acre-feet) and the x-axis representing the date.

Results

The K-means clustering analysis of the California reservoir storage data yielded three distinct clusters, as represented in the figures provided. The clustering process aimed to categorize reservoirs based on their storage levels from January 1 to July 15, 2024, to uncover patterns in water storage behaviors across different regions. The results are summarized below, with each cluster's characteristics described based on the analysis of their reservoir levels and geographic distribution.

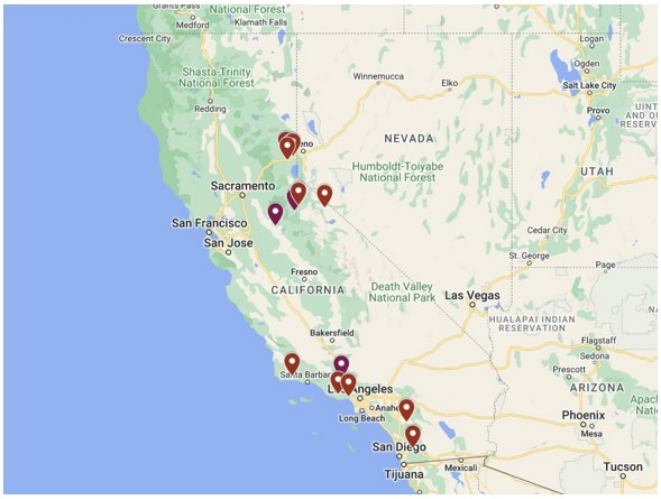

Fig. 2 Geographic distribution for cluster 1 (k=3). It shows the zoomed-out version of the California map.

Reservoirs in Cluster 1: This cluster includes reservoirs such as Alisal Reservoir, Vail Lake, Independence Lake, Bridgeport Reservoir, Beardsley Lake, Lake Piru, Lake Eleanor, El Capitan Reservoir, Prosser Creek Reservoir, Donnell Lake, Tulloch Reservoir, Boca Reservoir, Santa Ynez Reservoir, and Donner Lake. The storage levels generally range from 4,000 to 80,000 acre-feet. As shown in the maps, these reservoirs are widely distributed across California, with a noticeable concentration in the northern and southern regions, including areas around the Sierra Nevada and coastal regions near Los Angeles and Santa Barbara.

Reservoirs in Cluster 2: Don Pedro Reservoir is the sole member of this cluster. It exhibits high stability with minor fluctuations in its water levels, maintaining storage between 1.5 and 2 million acre-feet. Geographically, Don Pedro is located in central California, serving as a significant water storage facility for the region, heavily reliant on agricultural production.

Reservoirs in Cluster 3: This cluster includes Hetch Hetchy Reservoir, Lake Havasu, Stampede Reservoir, Indian Valley Reservoir, Cherry Lake, and San Vicente Reservoir. The reservoirs in this cluster generally have water levels ranging from 200,000 to 500,000 acre-feet. They are located across California, with Lake Havasu in the southeast near the Arizona border,

© The National High School Journal of Science 2024

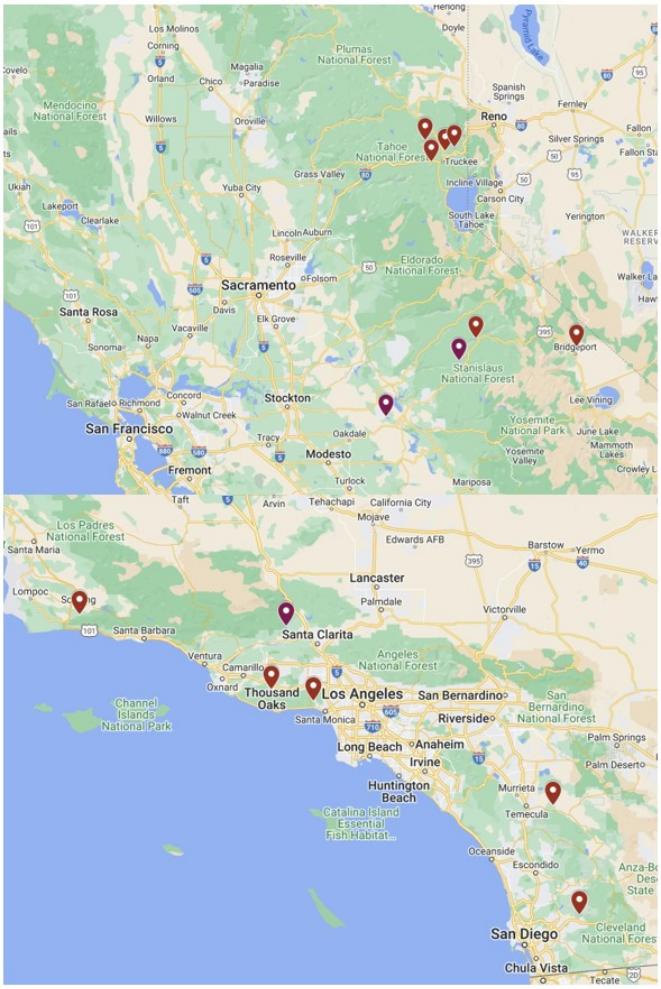

Fig. 3 - 4: Geographic distribution for cluster 1 (k=3), showing the zoomed-in version for detail.

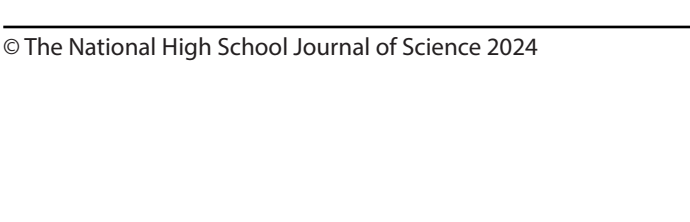

Fig. 5 Reservoir levels for cluster 1 (k=3). Unit indicated in acre-feet.

Hetch Hetchy in the Sierra Nevada, and others scattered across both northern and southern regions.

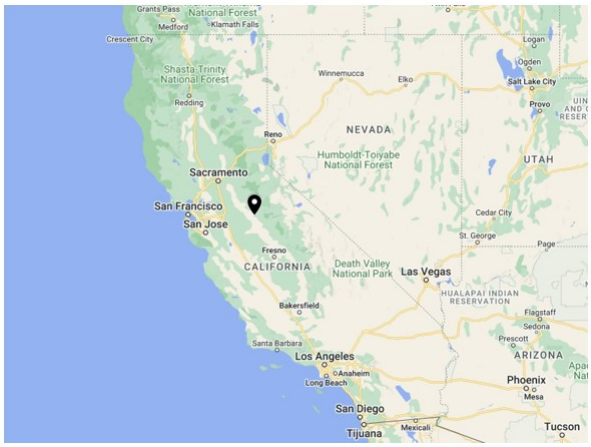

Fig. 6 Geographic distribution for cluster 2 (k=3), showing the zoomed-out version of the California map.

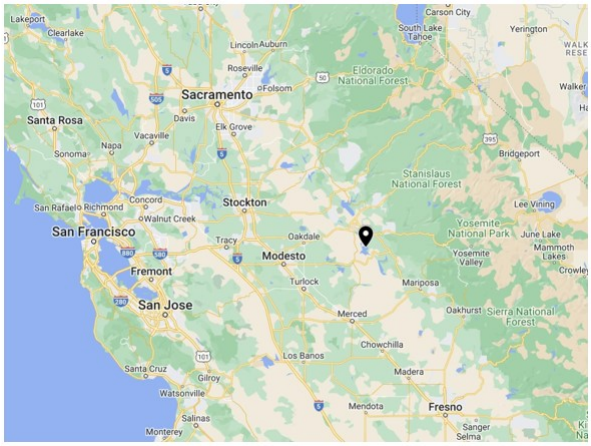

Fig. 7 Geographic distribution for cluster 2 (k=3), showing the zoomed-in version for detail.

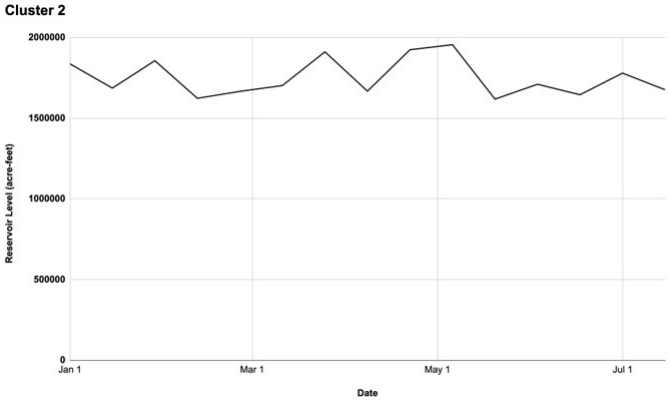

Fig. 8 Reservoir levels for cluster 2 (k=3). Unit indicated in acre-feet.

Discussion

This study aimed to explore the application of K-means clustering on reservoir storage data in California, using water level

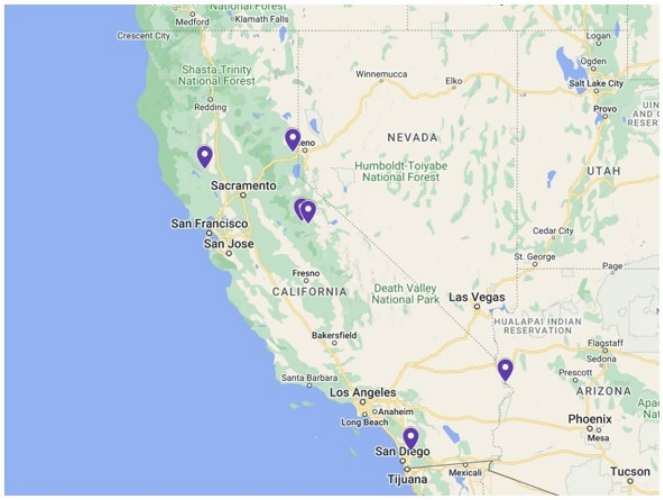

Fig. 9 Geographic distribution for cluster 3 (k=3), showing the zoomed-out version of the California map.

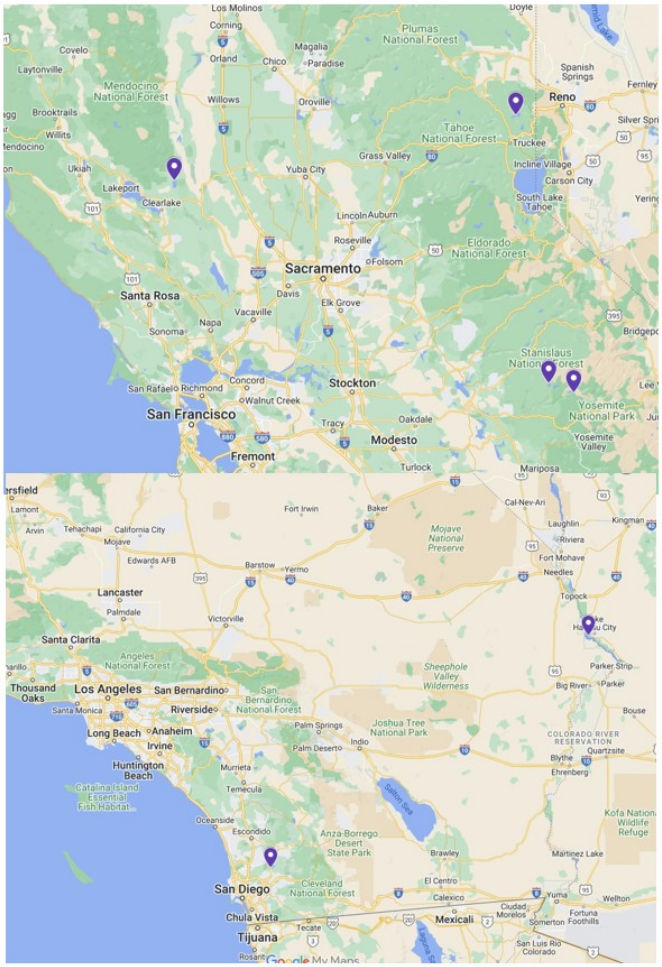

Fig. 10 - 11 Geographic distribution for cluster 3 (k=3), showing the zoomed-in version for detail.

trajectories over time as features. By doing so, we sought to investigate what patterns would emerge from the data without preconceived notions of what those patterns might reveal. As illustrated by the data analysis and geographical mapping, using k=3 for clustering, the clustering process produced distinct groupings that offer insights into reservoir behaviors across the state, but the primary goal was to see if clustering could simplify this complex dataset and provide a basis for future research into water resource management.

Spatial Distribution and Reservoir Clustering

The K-means clustering applied to reservoir storage data in California successfully grouped reservoirs based on their water levels from January 1 to July 15, 2024. Each cluster reflects distinct patterns in storage behavior, providing insights into the variability and trends across different reservoirs. However, this clustering analysis serves as an initial step, and several underlying factors, such as environmental conditions, were not incorporated, limiting the interpretive depth of the clusters. Also, it's essential to recognize that the clustering was primarily driven by the absolute water levels of the reservoirs, which influenced the grouping significantly.

Cluster 1 (Figure 5) includes reservoirs such as Alisal Res, Vail Lake, and Bridgeport Reservoir, mainly small- to mid-sized reservoirs scattered across northern and southern California. These reservoirs have absolute water levels generally ranging from 4,000 to 80,000 acre-feet. The high variability in water levels over the study period, with significant fluctuations observed across all reservoirs, is a defining characteristic of this cluster. Many of these oscillations follow a "W" shape in the graph, indicating synchronized increases and decreases in water levels during similar periods. This synchronization likely reflects

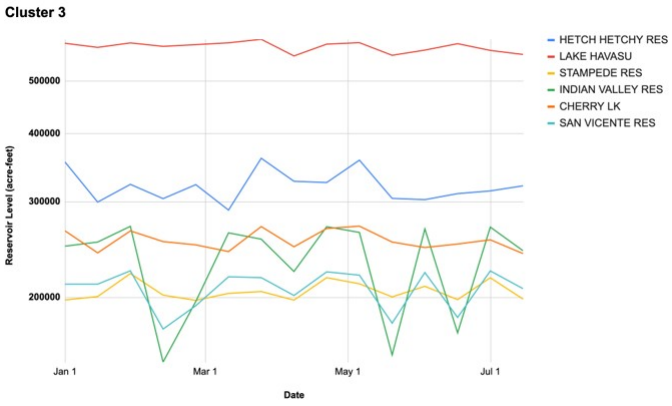

Fig. 12 Reservoir levels for cluster 3 (k=3). Unit indicated in acre-feet.

broader environmental conditions affecting multiple reservoirs simultaneously, but as this study is exploratory, we refrain from

© The National High School Journal of Science 2024

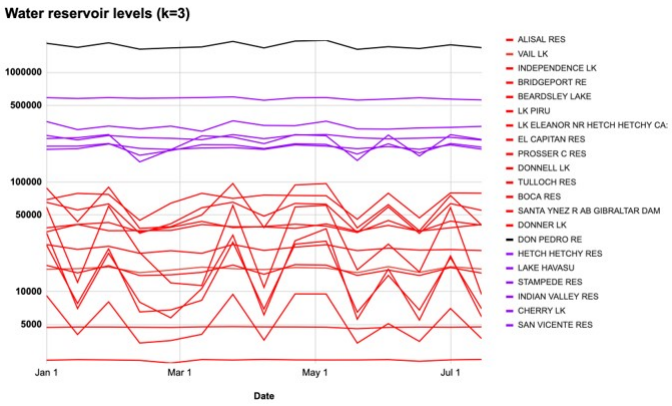

Fig. 13 Reservoir levels for all clusters. Cluster 1 is red, Cluster 2 is black, and Cluster 3 is purple. Unit indicated in acre-feet.

attributing these fluctuations to specific environmental conditions. The reservoirs in this cluster are distributed across both coastal and inland regions, with some located near significant urban centers like Los Angeles (e.g., Lake Piru, El Capitan Reservoir), where water demand may fluctuate more frequently.

Cluster 2 (Figure 8) is unique because it contains only Don Pedro Reservoir, which has significantly higher absolute water levels than the other reservoirs. Don Pedro maintains storage between 1.5 and 2 million acre-feet, exhibiting much more stable behavior than the reservoirs in Cluster 1. The high capacity and relative stability of Don Pedro indicate its critical role in state-level water management, likely servicing significant agricultural demand and municipal supply areas. The reservoir's location in the central valley further supports its necessary function, as this region experiences heavy agricultural water usage. Its isolation in this cluster underscores how the absolute size of the reservoir influenced the clustering.

Cluster 3 (Figure 12) encompasses a set of larger reservoirs, such as Lake Havasu, Hetch Hetchy Reservoir, and Indian Valley Reservoir, with water levels typically between 200,000 and 500,000 acre-feet. These reservoirs show moderate fluctuations in water levels, highlighting their roles as more stable, primary storage systems that provide a consistent water supply throughout the year. The behavior of these reservoirs is somewhere between the highly variable reservoirs in Cluster 1 and the stable Don Pedro in Cluster 2. The reservoirs in this cluster are located inland, closer to the Sierra Nevada and Colorado River regions, serving as critical water sources for regional and state-level supply chains. Like Cluster 1 and Cluster 2, this cluster's formation is primarily based on the absolute water levels of the reservoirs, which fall within a specific range.

Broader Implications and Limitations

The clustering analysis provides a valuable overview of how different reservoirs in California behave during the first half of the year. However, the clusters are mainly based on absolute water levels, which determine the grouping of reservoirs. This explains why Don Pedro Reservoir was isolated in its cluster due to its significantly higher capacity. Clusters 1 and 3 consist of reservoirs with absolute water levels below and above approximately 100,000 acre-feet. This reliance on absolute water levels limits the analysis to capacity-based trends rather than capturing dynamic behaviors.

Future work could involve scaling the reservoir levels so that all data points have a mean of 0 and a standard deviation of 1 to gain more actionable insights. This normalization process would shift the focus of clustering from absolute water levels to the patterns of variation in storage levels over time. By normalizing the data, the clustering would reveal reservoirs that exhibit similar trends and behaviors, regardless of their size or capacity, providing a deeper understanding of the factors driving reservoir behavior.

Additionally, clustering alone cannot explain why certain reservoirs fluctuate more than others. The synchronized "W" shape observed across clusters suggests common environmental factors, such as regional precipitation or snowmelt, that affect multiple reservoirs simultaneously. However, without further analysis, it is challenging to pinpoint the exact causes of these patterns. Incorporating meteorological data, such as rainfall and temperature, into future analyses would be critical for understanding the external drivers of reservoir behavior.

Another limitation of the current approach is that it treats each reservoir as an isolated entity even though many are part of interconnected systems. Exploring how reservoirs within and across clusters interact—through shared water transfers or coordinated management strategies—could offer deeper insights into California's broader water management framework.

Finally, the dataset spans only six and a half months (January to mid-July 2024), which restricts the ability to assess the full impact of seasonal cycles on reservoir levels, particularly during late summer and fall when many reservoirs experience the most strain. With only 21 reservoirs analyzed, the scope of this study is narrow. Expanding the dataset to include more reservoirs across California and extending the time series to capture a full year of reservoir behavior would provide a more comprehensive understanding of statewide water management dynamics. Future studies should consider adding variables such as rainfall, temperature, and water demand, which may offer a more nuanced understanding of the factors driving reservoir storage patterns.

Conclusion

This study successfully applied K-means clustering to categorize California's reservoir storage data based on water level trajectories. Using k=3, the clustering process produced meaningful distinctions between reservoirs, particularly in capacity and variability. The findings support the hypothesis that K-means clustering can be an effective tool for simplifying and analyzing complex datasets like reservoir water levels, allowing for identifying patterns that are not immediately visible through traditional data analysis methods. By grouping reservoirs based on their water level fluctuations, the study provides a foundation for future research into how these patterns might inform water resource management decisions.

While this research focused solely on water levels, it highlights the potential for K-means clustering to provide insights into large environmental datasets. However, it also reveals limitations—such as the reliance on absolute water levels and the narrow timeframe—that should be addressed in future work. Expanding the dataset to include more reservoirs over a longer period, incorporating additional variables, and normalizing the data would all contribute to a more robust understanding of reservoir behavior.

Future studies should explore these possibilities, building upon the results of this clustering analysis to develop more comprehensive water management strategies. Although this study is exploratory, it demonstrates that K-means clustering can be a valuable tool for identifying trends in reservoir storage data and opens the door to further research in this area, especially as California continues to face water management challenges due to climate variability.

Acknowledgments

Thank you for the guidance of Simon Alford, my mentor from Cornell University, in the development of this research paper.

References

1 W. Jiang, B. Pokharel, L. Lin and H. Cao, *Analysis and prediction of produced water quantity and quality in the Permian Basin using machine learning techniques, Science of The Total Environment*.

2 S. Yi, G. Kondolf and S. Sandoval-Solis, *Application of Machine Learning-based Energy Use Forecasting for Inter-basin Water Transfer Project*, https://doi.org/10.1007/s11269-022-03326-7.

3 A. J. Draper, A. Munévar and S. K. Arora, *CalSim: Generalized Model for Reservoir System Analysis*, https://doi.org/10.1061/(ASCE)0733-9496(2004)130:6(480).

4 Q. Wang and S. Wang, *Machine Learning-Based Water Level Prediction in Lake Erie*, https://doi.org/10.3390/w12102654.

5 C. Li, L. Sun and J. Jia, *Risk assessment of water pollution sources based on an integrated k-means clustering and set pair analysis method in the region of Shiyan, China, Science of The Total Environment*, https://doi.org/10.1016/j.scitotenv.2016.03.069., ISSN 0048-9697,.

6 L. Wu, Y. Peng and J. Fan, *A novel kernel extreme learning machine model coupled with K-means clustering and firefly algorithm for estimating monthly reference evapotranspiration in parallel computation, Agricultural Water Management*.

7 E. Aytaç, *Unsupervised learning approach in defining the similarity of catchments: Hydrological response unit based k-means clustering, a demonstration on Western Black Sea Region of Turkey, International Soil and Water Conservation Research*.

8 U. Survey, *National Water Information System: Web Interface*, https://waterdata.usgs.gov/nwis.

© The National High School Journal of Science 2024