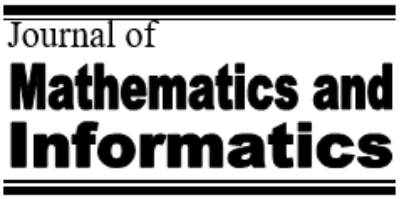
Journal of Mathematics and Informatics
Vol. 27, 2024, 9-26
ISSN: 2349-0632 (P), 2349-0640 (online)
Published 1 December 2024
www.researchmathsci.org
DOI:http://dx.doi.org/10.22457/jmi.v27a02247



Validating a Data-Driven Multi-Model Characterization Technique for Water Users: A Case of Pangani Basin in Tanzania

***Matimbila P Lyuba*[1*], *Devotha G Nyambo*[2], *Anael Sam*[2], *Seifu Tilahun*[4,5] and *Aneel Rahim*[6]**

[1,2,3] School of Computational and Communication Science and Engineering (CoCSE)
Nelson Mandela African Institution of Science and Technology (NM-AIST)
[4]International Water Management Institute, Accra, Ghana.
[5]Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology
Bahir Dar University, Ethiopia.
[6]School of Computing, Technological University of Dublin, Ireland.
[1]E-mail: lyubam@nm-aist.ac.tz
[2]E-mail: Devotha.nyambo@nm-aist.ac.tz
[3]E-mail: anael.sam@nm-aist.ac.tz; [4,5]E-mail: s.tilahun@cgiar.org
[6]E-mail: aneel.rahim@tudublin.ie
[*]Corresponding author

Received 20 September 2024; accepted 30 November 2024

Abstract. Through a literature review, it has been observed that water scarcity results from increased demand due to population growth, economic progress, and climate change, leading to disparities between required and available water resources. Addressing this challenge requires segmenting water users into homogeneous groups and thoroughly examining their characteristics regarding water utilization to develop efficient and effective water governance strategies. This study employed data-driven multi-model validation techniques to characterize water users in Pangani Basin in Tanzania. The K-means, Agglomerative Hierarchical, and Fuzzy C-means clustering algorithms were used to ascertain the efficacy of the characterization. Cluster validation showed that K-means outperformed Agglomerative hierarchy by owning a high Calinski–Harabasz Index and low Davies–Bouldin Index of 692.3 and 1.8, respectively, compared to Agglomerative hierarchy with values of 578.2 and 1.9, respectively. The clustered dataset was tested for prediction accuracy by fitting the logistic regression. K-means showed a prediction accuracy of 98.2% over 97.5% of the Agglomerative Hierarchical method. The four clusters identified were large-scale irrigation water users, moderate irrigation water users, community water supply entities, and domestic water users. We argue that understanding users' characteristics could efficiently and effectively add value to water governance along the basins.

Keywords: characterization; multi-model; water users; river; clustering algorithms; Pangani Basin

AMS Mathematics Subject Classification (2010): 00A71

1. Introduction

River basins serve as vital spatial boundaries to manage water resources for various socioeconomic activities [1]. Activities that depend on water availability in river basins include water for irrigation, livestock, domestic water supply, hydropower generation, industrial usage, and mining [2,3]. For instance, approximately 12 million tons of freshwater fish, accounting for approximately 25% of global production, comes from rivers annually [2]. In Tanzania, the Pangani Basin supports the livelihoods of at least 4.7 million people through various socioeconomic activities [4]. In the recent past, water scarcity has emerged as a significant global concern, exacerbated by population increase, climate change, and economic growth [5,6]. These trends have led to imbalances between the increasing demand for water to meet diverse user needs and the availability of water from natural resources [7]. Consequently, ensuring sustainable water management practices within river basins has become imperative to address the challenges posed by water scarcity and promote socioeconomic development in affected regions.

Despite the scarce water resources within the basin, governing bodies such as the Pangani Basin Water Board (PBWB) face challenges related to water governance due to a lack of understanding of water users' characteristics which influence their water utilization behaviors [8,9]. Reports show that such challenges include conflicts among water users, disagreements between formal and informal governmental institutions, unequal water resource sharing between the upstream and downstream water users, ineffective and poorly planned strategies for water allocation and rationing, and cooperative prioritization for profit over the community's welfare [8–12]. For the water governing bodies to create suitable interventions that enable water users to control their water abstraction, it requires characterizing the various groups of water users [13]. Characterizing the water consumption systems throughout the basin and identifying homogeneous units that, in terms of management, represent modern groupings enable us to comprehend the particular characteristics linked to drivers that are essential to water allocation and rationing. This is the key to revealing the components of sound water governance, which have been developed via careful planning, the adoption and application of suitable water abstraction, and crucial governmental support. This study aimed to test a robust mechanism that allows water users with similar water utilization characteristics or performance to be categorized into homogenous clusters that characterize their organizational structure and water abstraction needs.

Clustering algorithms offer valuable insights into understanding the characteristics of water resources [14,15]. In clustering, similar groups, items, or objects are aggregated [14]. Previous studies in water resource management have utilized clustering to categorize data into various groups, such as end-user water consumption [16,17], hydrological segmentation [18], groundwater pattern recognition [19], water quality [20], and rainfall and risk monitoring [21,22]. For example, Jenifer and Jha [23] utilized a rainfall time series in southern India to cluster with varying rainfall trends using the Agglomerative Hierarchical algorithm, and Shahfahad et al. [24] addressed urban flooding using the Fuzzy C-means and K-means algorithms. Regarding water use segmentation, Gao et al. [25] characterized water usage across various economic sectors in China using K-means clustering algorithms and classified 139 sectors into 5 clusters with distinct features. To formulate effective water conservation and management

strategies, a recent study by Rahim et al. [26] is an example in which digital water meter data from 306 homes in Melbourne, Australia, over ten months were used to study water user behaviors, habits, and preferences. In addition, mathematical approaches such as Fuzzy logic have been employed to assess water quality as well as analyze data that are vague in nature [27–29]. Furthermore, classification models such as Random Forest (RF), Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) have been used to analyze various parameters related to water resources [30–33]. Umair Ahmed [34] applied both the classification and regression algorithms on various water parameters such as temperature, pH, and water dissolvent ability to predict the water quality index. Cluster validation techniques such as the Davies–Bouldin Index (DBI), Silhouette Coefficient (S.S.), Calinski–Harabasz (C.H.), and Dunn Index (DI) have been applied to assess cluster separations and identify the level of cluster homogeneity [35,36]. The application of these algorithms highlights that machine learning could largely contribute to addressing challenges related to water resource management along the basins.

Although research indicates the popular clustering algorithms used to address various problems in water resources are OPTICS, Fuzzy C-means, K-means, Self-Organizing Map (SOM), Agglomerative Hierarchical, DBSCAN, Neural Networks, and Genetic K-means [13,35], it is noted that Agglomerative Hierarchical, Fuzzy C-means, and K-means have been frequently used to understand characteristics related to various attributes in water resources and have demonstrated promising results compared to other algorithms [13,24,37]. Despite their widespread use, clustering algorithms encounter challenges related to consistency and predictability in such a way that identifying the most suitable algorithm for a given dataset and validating the resulting clusters becomes challenging [38,39]. Although other strategies such as applying several algorithms to cluster datasets and choosing the ones with the most homogeneous groupings have been tried to address this shortcoming, their number is still limited [17,35].

The present study used data from Tanzania's Pangani Basin to categorize and examine the characteristics of water users using three unsupervised machine learning clustering models. To the best of the authors' knowledge, no features associated with water consumption have been analyzed using basin water consumption data at this level of granularity. The objective of the study was to apply a data-driven approach to validate the most reliable model for correctly classifying different water users into homogeneous groups that correspond to variations in water use. The findings of this study contribute to an effective approach that validates the characteristics of waters users along the basin for proper water allocation and rationing.

2. Materials and methods

2.1. Study area

The case study area was northern Tanzania's Pangani Basin (see Fig. 1). The basin comprises about 58,400 km^2, of which 93% (54,600 km^2) is in Tanzania, and 7% (3800 km^2) is in Kenya, with Nyumba ya Mungu reservoir as its primary source. The two main tributaries of the Pangani Basin are the Ruvu, with an average flow of 5.339 m^3/s, and the Kikuletwa, with an average long-term flow of 11.502 m^3/s. The Kikuletwa, Ruvu, Pangani Main stem, Mkomazi, Luengera, Msangazi, Zigi, Mkulumuzi, and Umba

catchments are the nine catchments that make up the basin according to the water shade boundary [4]. The basin spans four regions: Arusha, Manyara, Kilimanjaro, and Tanga. It crosses 21 districts (Pangani, Tanga City, Mkinga, Muheza, Handeni, Handeni Town, Kilindi, Lushoto, Moshi Municipal, Moshi Rural, Rombo, Mwanga, Same, Korogwe, Korogwe Town, Silanjiro, Siha, Hai, Moshi City, Moshi Municipal, and Moshi Rural). It serves a population of 4.7 million people [4]. According to the 2022 census, the population within the basin has increased to more than 6 million people [40].

The amount of resources invested by blue water users, i.e., water drawn from aquifers, lakes, rivers, and artificial reservoirs, is more significant than that of green water users, i.e., water use that depends on rainfall in the basin [9,11]. The social–economical activities conducted in the basin include crop irrigation such as horticulture, rice, maize, hydroelectricity power production (~97 megawatts), mining such as Tanzanite, gold, ruby, Sunstones, moonstones, aventurine quartz, aquamarine, spessartite, tourmaline, green garnets, red garnets, anyolite, corundum, and rhodolite [4]. Moreover, there are significant and small-scale industries such as bonite bottlers and TPC sugar, and domestic users such as hotels, homes, etc. The basin also includes several controlled areas and national parks for both tourism and wildlife, including the Amani, Chome, Nilo, and Magamba Forest Nature Reserves and the Kilimanjaro, Arusha, Saadani, and Mkomazi National Parks [4]. The Pangani Basin was selected because it is the first basin to be introduced in the country, and therefore, water user data are available and records are well kept. It has diverted socioeconomic activities and created high competition for water use. As per the water resource fact sheet, approximately 75% of the population in the basin experiences water stress [41].

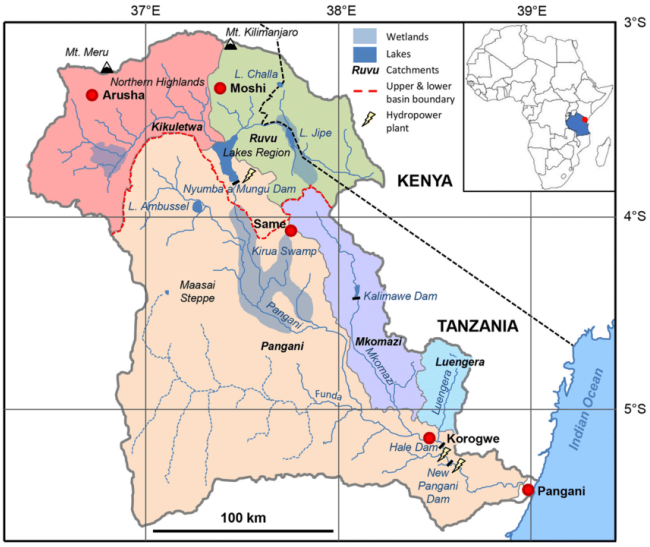


Figure 1. The Pangani Basin source: [4].

2.2. Data collection and pre-processing

The dataset was collected as a secondary source at the PBWB located in Moshi, Kilimanjaro. Data collection was carried out between January 2023 and June 2023. The dataset was organized into a Microsoft Excel sheet comprising rows and columns. The

columns include details such as I.D., file number, applicant name, applicant postal address, applicant email address, applicant village and ward, file number, region, district, type of water sources, water source name, amount of water abstracted, amount of water requested by the applicant, amount of water available in the source, water uses, water use category, water use fee paid by users for water abstraction, water use history, activity status, permit status, and catchment.

The dataset was cleaned by removing erroneous data, formatting text, and the water user's identity, such as water user names, phone numbers, email addresses, postal addresses, and office file numbers, as per research ethical requirements. The remaining dataset comprised 3460 records with 15 features, as depicted in Table 1. The mean imputation technique handled missing values and avoided minimizing data representation. No duplicate value was observed in the dataset.

The outliers were removed using the interquartile range (IQR) with a capping approach. The nominal features were converted into discrete values, and the dataset was standardized to have a mean of approximately zero ($\mu = 0$), and a unit standard deviation ($\sigma = 1$). The heatmap found a correlation among the features within the dataset. The 15 correlated features were reduced to 13 uncorrelated features using Principal Component Analysis (PCA). We applied the Hopkins (h) statistics test to verify the dataset cluster-ability tendency. A Hopkins score of $h > 0.90$ was observed and justified the dataset's clustering tendency.

Table 1. Description of the dataset features.

S.No	Feature Name	Encoding	Feature Type	Feature Description
1	Region	0–4	Discrete	Water user region of residence
2	District	0–18	Discrete	Water user district of residence
3	source_type	0–9	Discrete	Type of source where water is abstracted, i.e., river, spring, borehole, etc.
4	source_name	0–1169	Discrete	Name given to the specific water source
5	water_use	0–18	Discrete	Specific water uses to the end user
6	water_use_category	0–5	Discrete	Categories of water users set by the PBWB.
7	water_use_fee	100,000–1,059,478,880	Continuous	Amount paid (TZS) by the water user as a fee for abstracting water from the basin
8	amount_abstracted	0.1–140,500	Continuous	The amount of water approved by the PBWB abstracted by the user from the source measured in L/s

9	amount_requested	0.1–140,500	Continuous	Amount of water (L/s) requested by the user once the application is logged. Users can be granted less than or equal to the amount requested depending on the water assessment conducted at the source.
10	water_source_capacity	0.1–140,500	Continuous	Amount of water (L/s) available in the source
11	permit_history	0 (new)–1 (water right)	Boolean	Previous history of the permit owned by the water user. Permits change time after a time.
12	permit_type	0–2	Discrete	The current type of permit owned by the water user
13	activity_status	0 (Active)–1 (Inactive)	Boolean	The activity of the water user, i.e., either active: currently abstracting water, inactive: does not presently abstract water
14	permit_status	0 (invalid)–1 (valid)	Boolean	Permit status, i.e., valid or expired
15	catchment	0–9	Discrete	Catchment where a user is abstracting water

2.3. Clustering

We employed the elbow method to determine the ideal number of clusters for the dataset with $2 \leq K \leq 10$ where K is the number of clusters. The K-means, Agglomerative Hierarchical, and Fuzzy C-means clustering algorithms were utilized. Clustering the dataset started with the K-means algorithm. This algorithm uses the Euclidian distance metric to create K clusters from M points in N dimensions (M × N matrix) to minimize the sum of squares within each cluster [42]. K-means is a complex clustering algorithm since it assigns a data point to one cluster only. The second algorithm was Agglomerative Hierarchical. Agglomerative Hierarchical with ward criterion is the most used variant of hierarchical clustering. It considers the minimization of the distance between data points [43]. Agglomerative Hierarchical starts with n number of clusters and combines similar ones until they become one. The third algorithm was Fuzzy C-means. Fuzzy C-means is a soft clustering algorithm that allocates a data point into more than one cluster with varying degrees. It evaluates the solution's inter-cluster cohesiveness and fuzziness using a Silhouette separation coefficient and a Dunn coefficient [24].

2.4. Cluster evaluation and validation

2.4.1. Cluster evaluation

The multi-models were evaluated using the DBI, CHI, S.S., and DI for their robustness to the dataset. We focused on identifying the clustering algorithm that generates clusters that best fit the dataset by producing highly homogeneous clusters. DBI is a statistical measure that evaluates how far apart and compact a cluster is. It is predicated on the notion that high between-cluster separation and low within-cluster variance characterize

better clusters than clusters with low between-cluster separation and high within-cluster variance. The low value is preferred. The CHI considers K and N, the relative magnitudes between sum square (BSS) and within sum square (WSS). For a given value of K, a significant value of this index indicates a clustering solution with considerable variability between clusters and low variability inside the cluster. A high value is preferred. The S.S. measures a proportion of an object's cohesion (its similarity to its cluster) to its separation from other clusters. The silhouette's value is a number between 1 and −1, where a high number means the object matches its cluster well and neighboring clusters poorly. The DI computes and compares the ratios between the lowest distances between clusters to the most significant distance between clusters. A high DI value is preferred.

2.4.2. Clustering validation

We urged the clustering algorithm with high prediction accuracy tested in training and testing datasets generalizes the unseen dataset well. Since clustering is an unsupervised learning approach, we fitted the clustered dataset into the logistic regression model and determined the model prediction accuracy for the test dataset portion. The clustered dataset, which gives high prediction accuracy, signifies the suitability of its corresponding clustering model for the unseen dataset upon the model's deployment.

2.5. Cluster characterization

2.5.1. Inter-cluster characterization

The heatmap plot grouped the clustered dataset by water use category and aggregation with the cluster size, and the water abstraction mean rate was used to identify the between clusters behaviors. These characteristics essentially distinguish one cluster from the other. We spotted parameters with high domination of the cluster, differentiating it from others.

2.5.2. Intra-cluster characterization

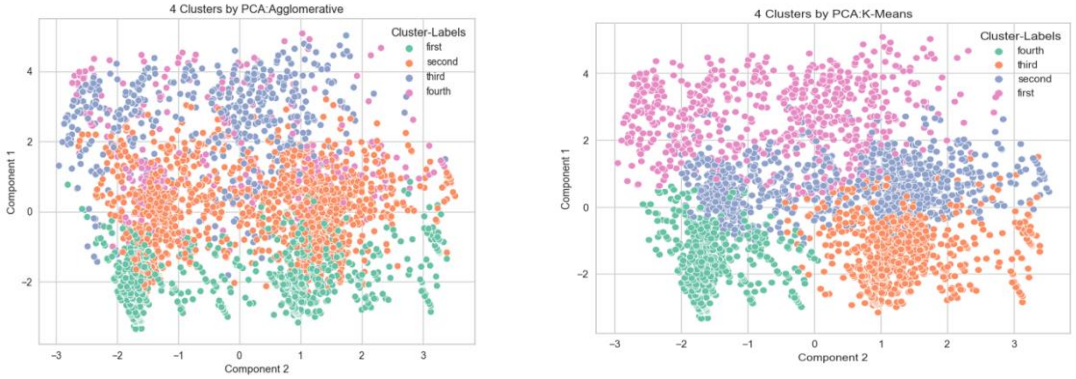
To identify differences between members within the cluster, we plotted heatmap visualization. We grouped the clustered dataset based on the specific water use and aggregated them by the cluster size and water abstraction mean rate. We spotted parameters with high domination within each cluster.

3. Results

3.1. Clustering

The elbow method showed that four (4) clusters are the optimal number of clusters with the lowest within-cluster sum square errors (WSSE) of 32160 compared to other clusters options. When tested using the Gap Statistics test and assessing the Silhouette Score, two (2) clusters with high WSSE above 35000 were observed. K-means and Agglomerative Hierarchical algorithms clustered the dataset into four clusters. Fuzzy C-means clustered the dataset into three clusters with no members in the fourth cluster. Table 2 depicts the cluster members obtained after clustering. K-means reveals that the first cluster comprised users with a high water abstraction mean rate of 23.18 L/s. A relatively low difference of 0.25 L/s between the third and fourth clusters was observed. However, the second cluster had many users compared to other clusters.

For the Agglomerative Hierarchical algorithm, the third cluster comprised users having a high water abstraction mean rate of 23.18 L/s. The clusters in Agglomerative Hierarchical clusters showed a substantial difference in the abstraction mean rates. The first two clusters had many members relative to the third and fourth clusters. The Fuzzy C-means algorithm showed a high water abstraction mean rate of 17.89 L/s in the second cluster, while its first and third clusters had a relatively low abstraction mean difference of 0.34 L/s. Fig. 2 depicts cluster visualization using the first two Principal Components for Agglomerative Hierarchical and K-means. Clusters generated by K-means are seen to be better compact separated and have a significant intra-cluster adhesion compared to the clusters generated by the Agglomerative Hierarchical algorithm.



(**a**) (**b**)

Figure 2. Clusters' visualization using PCA (**a**) Agglomerative Hierarchical (**b**) K-means.

Table 2. Cluster segmentation for the three algorithms.

Cluster #	**K-Means**		**Agglomerative Hierarchical**		**Fuzzy C-Means**	
	User	**Mean Rate (L/s)**	**User**	**Mean Rate (L/s)**	**User**	**Mean Rate (L/s)**
1	775	23.18	1115	1.86	911	2.73
2	1015	3.98	1352	5.58	1163	17.89
3	878	2.87	589	23.18	1386	2.39
4	792	2.62	404	8.22		

3.2. Cluster evaluation and validation

3.2.1. Cluster evaluation

We computed the DBI, S.S., CHI, and DI for the K-means and the Agglomerative algorithms to evaluate the cluster algorithms' fitness to the dataset. We left Fuzzy C-means since it fails to provide members in the fourth cluster. Table 3 summarizes the metrics indexes.

Table 3. Clusters evaluation scores

Metric	K-Means	Agglomerative
DBI	1.8	1.9
S.S.	0.2	0.2
CHI	692.3	578.2
DI	0.1	0.1

Both algorithms scored equally in the S.S. and DI with 0.2 and 0.1, respectively. This informs us that the two algorithms produce clusters owing members with high feature similarity. The K-means score was better for the DBI and CHI, with values of 1.8 and 692.3 respectively, thus outperformed the Agglomerative which scores a DBI of 1.9 and a CHI of 578.2 respectively. These results showed that the clusters formed by K-means had high variability between clusters and low variability within clusters; thus, they were more homogenous. Therefore, we choose the K-means for the water user's characterization.

3.2.2. Cluster validation

The two clustered datasets for K-means and Agglomerative Hierarchical were fitted to the logistic regression model for prediction accuracy analysis. Both clustered datasets were divided into 70% and 30% segments for the training and testing segments. The K-means scored a prediction accuracy of 98.2%, while Agglomerative Hierarchical scored a prediction accuracy of 97.5%. Furthermore, the classification reports show that the third cluster scored a precision, recall, and F1-score of 0.99, 0.97, and 0.98, respectively, for the K-means compared to scores of 0.93, 0.94, and 0.94 for Agglomerative Hierarchical. The K-means' results validate it over Agglomerative Hierarchical for the clustered dataset.

3.3. Cluster characterization

3.3.1. Inter-cluster characterization

Results show that the first cluster was dominated by irrigation users who abstract water at a mean rate of 23.11 L/s. Similarly, the second cluster was dominated by irrigation users who abstract water at a mean rate of 4.6 L/s. The third cluster was dominated by community water supply entities who abstract water at a mean rate of 3.54 L/s, and the fourth cluster was dominated by domestic users who abstract water at a mean rate of 1.03 L/s. We observed that the cluster sizes and water abstraction mean rate are the two parameters that broadly distinguish clusters. Fig. 9 depicts the four clusters' overall water abstraction mean rates as 23.18 L/s, 3.98 L/s, 2.87 L/s, and 2.62 L/s, respectively. We deducted cluster types based on the inter-cluster characterization analysis of cluster sizes and the water mean abstraction rate. Table 4 summarizes the inter-cluster characteristics.

Table 4. Inter-cluster characteristics for each cluster.

Cluster #	Cluster Size (in %)	Prevailing Inter-Cluster Characteristics	Cluster Type
1	22.4	Cluster with water abstraction mean rate of 23.18 L/s A large number of irrigation users with a water abstraction mean rate of 23.11 L/s	Large-scale irrigation water users

		A low number of construction users with a water abstraction mean rate of 26.30 L/s	
2	29.3	Cluster with water abstraction mean rate of 3.98 L/s	Moderate irrigation water users
		A large number of irrigation users with a water abstraction mean rate of 4.60 L/s	
		A low number of hydropower generation users with a water abstraction mean rate of 1.21 L/s	
3	25.4	Cluster with water abstraction mean rate of 2.87 L/s	Community water supply entities
		A large number of community water supply entities with water abstraction mean rate of 3.54 L/s	
		A low number of construction users with a water abstraction mean rate of 2.40 L/s	
4	22.9	Cluster with water abstraction mean rate of 2.62 L/s	Domestic water users
		A large number of domestic users with a water abstract mean rate of 1.03 L/s	
		A low number of construction users with a water abstraction mean rate of 6.97 L/s	

3.3.2. Intra-cluster characterization

We observed that large-scale irrigation water use characterized the first cluster, and the second cluster was characterized by small-scale irrigation water use with a significant number of individual (home) water users. The third cluster comprised individual water use, with many institutions and community-based water (CBWO) associations. The fourth cluster was characterized by individual water use, with many institutions using water. Table 5 summarizes the intra-cluster characteristics.

Table 5. Intra-cluster characteristics for each cluster.

Cluster Type	Prevailed Intra-Cluster Characteristics	Prevalence (%)
Large-scale irrigation water users	Many users perform large-scale irrigation activities with a water abstraction mean rate of 23.24 L/s	76.9
Moderate irrigation water users	Many users perform small-scale irrigation activities with a water abstraction mean rate of 4.29 L/s	44.3
	A significant number of users abstract water for individual uses (home) at a mean rate of 1.61 L/s	27.7
	A significant number of Community-Based Water Supply Organizations (CBWSOs) abstract water at a mean rate of 6.76 L/s	10.2
Community water supply entities	Many users abstract water for individual uses (home) with a mean rate of 1.17 L/s	24.6
	A significant number of institutions abstract water at a mean rate of 2.28 L/s	20.1
	A significant number of Community-Based Water Supply Organizations (CBWSOs) abstract water at a mean rate of 4.42 L/s	14.8
	A significant number of small-scale irrigators abstract water at a mean rate of 3.99 L/s	12.1

Domestic water users	Many users abstract water for individual uses (home) at a mean rate of 0.91 L/s	37.2
	A significant number of institutions abstract water at a mean rate of 2.80 L/s	22.8
	A significant number of small-scale irrigators abstract water at a mean rate of 3.76 L/s	13.8

.

3.3.3. Clusters' characterization as a result of the characteristics of the inter and intra-clusters

To determine the clusters' characteristics, we considered both the inter-cluster characteristics that differentiate a cluster from other clusters and the intra-cluster characteristics that provide a detailed description of members within the cluster. The characteristics of the four homogeneous clusters for the Pangani Basin water users were as follows:

Cluster Type One: *A total of 22.4% of water users conducting their social–economical activities along the basin are large-scale irrigation water users. Users in this cluster were characterized by high amount_abstracted, irrigation water_use_category, and large-scale irrigation water_use.*

Cluster Type Two*: Contained the majority of water users, 29.3%, consisting of moderate irrigation water users, characterized by moderate amount_abstracted, irrigation water_use_category and small-scale irrigation water_use.*

Cluster Type Three: *A total of 25.4% of users consisted of community water supply entities, characterized by low amount_abstracted, community water supply water_use_category, and home water_use.*

Cluster Type Four: *A total of 22.9% of users consisted of domestic water users characterized by the low amount_abstracted, domestic water_use_category, and home water_use.*

4. Discussion

4.1. Clustering

To characterize the water consumption dataset for the Pangani Basin, we used three clustering algorithms: Agglomerative Hierarchical, Fuzzy C-means, and K-means. Among the three algorithms, K-means and Agglomerative could both allocate users into four clusters, while Fuzzy C-means fails to allocate users in the fourth cluster. As suggested by Nyambo et al. [39] and Yuan and Yang [44], we used the elbow method to ascertain the ideal number of clusters with the lowest WSSE. Usually, varying values of K in a given set and observing the elbow bending shape are used to determine the ideal number of clusters. The elbow shape signifies that there is no a substantial difference in continued increases in K, which might give the best WSSE. In the present study, the elbow results showed four high homogeneous clusters with the WSSE ~32160. We further employed the Gap Statistics and Silhouette diagrams to visualize the clusters and

confirm their formation. Results from the Gap Statistics and Silhouette diagram indicated the presence of two widely separated clusters with a Silhouette Score of 0.21. However, the elbow method showed that the two clusters had a WSSE > 42000, indicating they were less homogeneous than the four clusters. Fuzzy C-means is a soft clustering algorithm, whereas K-means and Agglomerative are classified as complex clustering algorithms. A data point is only assigned to one cluster by a hard clustering algorithm, whereas a soft algorithm assigns a data point to several clusters. We argue that Fuzzy C-means works better in a dataset with high heterogeneity. This might be a reason for its failure to locate users in the fourth cluster since the four clusters determined by the elbow method are highly homogeneous.

We observed differences in the clusters produced by the three algorithms. The K-means produced three clusters with a comparable distribution of members (cluster1: 775, cluster3: 878, and cluster 4: 792) and one cluster with a high density of members (cluster2: 1015). The Agglomerative Hierarchical produces two pairs of clusters with a comparable distribution of members (cluster1: 1115, cluster2: 1352) and (cluster3: 589, cluster4: 404). In both algorithms, the three clusters possess a comparable water abstraction mean rate, and one cluster with a high water abstraction mean rate. Despite its failure to locate members in the fourth cluster, Fuzzy C-means produces two clusters with a comparable distribution of members (cluster2: 1163, cluster3: 1384) and one with low density (cluster1: 911)

In the studies by Gao et al. [25], Nyambo et al. [39], and Sinaga et al. [42], the elbow method and Silhouette Score were combined to determine the number of clusters and cluster qualities. We argue that in obtaining highly homogeneous clusters, the WSSE should be minimized so that an increase in K does not provide a significant difference in the WSSE. In the present study, we chose the minimum WSSE to obtain the highly homogeneous clusters that better distinguish water users in the basin. The same elbow method was adopted by Rahim et al. [13], who identified five homogeneous clusters. During their analysis of the dataset linked to water quality for drinking consumption, Eskandari et al. [45] compared K-means and Fuzzy C-means and found five and six homogeneous clusters, respectively. In their investigation, the Fuzzy C-means algorithm outperformed the K-means algorithm, most likely due to the dataset's use of uncertainties in class boundary definition. Another study that employed similar techniques is that of Rahim et al. [13], where the Agglomerative Hierarchical and K-means algorithms were used to cluster residential water users in various households. K-means outperformed the Agglomerative Hierarchical algorithm. Water regulating bodies such as the PBWB will be able to understand their customers and create water rationing awareness initiatives if they can determine the ideal number of clusters and the study characteristics for each cluster. We argue that the four homogenous clusters identified in this study are significant to the water governance along the basin with a similar setup as the Pangani Basin.

4.2. Cluster validation

Cluster validations show that K-means clusters are robust and fit better to the dataset than the clusters produced by Agglomerative Hierarchical. The clusters are highly distinguished by the number of members and their water abstraction mean rates.

The DBI, S.S., CHI, and DI were used to validate cluster robustness. K-means outperformed Agglomerative with a DBI of 1.8 compared to 1.9 and a CHI of 692.3

compared to 578.2, respectively. Both algorithms scored similar values for the S.S. and DI metrics. Based on these metrics, we argue that clusters produced by K-means were more compact and better suited to the dataset than the Agglomerative ones. Furthermore, we tested the performance of both algorithms for their prediction accuracy in both the training and testing environment by fitting their clustered dataset into the logistic regression model to confirm the generalization of the clustering algorithm to the unseen dataset. We urge that the clustered dataset, which indicates high prediction accuracy once fitted to the logistic regression, be validated and the algorithm that best generalizes the unseen dataset in the production environment be generalized. The K-means clustered dataset scored a prediction accuracy of 98.2%, while the Agglomerative clustered dataset scored a prediction accuracy of 97.5% in the logit testing data segment.

The S.S. of 0.299 and CHI of 156.6 for K-means, and S.S. of 0.25 and CHI of 129.9 for Agglomerative, respectively, were obtained in the study by Rahim et al. [13] while measuring the clusters' qualities, and they concluded that the clusters formed by K-means are robust and better fit the dataset. These results are similar to those of the present study, where the K-means values were an S.S. of 0.2 and CHI of 692.3 compared to Agglomerative with an S.S. of 0.2 and CHI of 578.2. K-means gave high homogeneous clusters for our case due to the high CHI value. A CHI value measures a degree of variation between clusters and is a ratio between inter-cluster and intra-cluster convergence [46]. The higher the numerator, the higher the degree of dispersion between clusters and the lower the denominator, the closer the data points are within the clusters [13]. The S.S of 0.2 informs us that both algorithms gave clusters with similar purity levels regardless of their clusters' dispersion. The other related study of the segmentation of a flood-affected area in Jakarta, Indonesia, used the DBI, CHI, and S.S. to validate the K-means clusters and found three homogenous clusters [36]. A similar approach was employed to validate clusters in the present study. The logit prediction accuracy of 98.2% of K-means over 97.5% of Agglomerative cemented the reproducibility of our experiment to other basins with similar sets as the Pangani Basin. We suggest that our finding describes a data-driven approach in which clusters can be validated and tested for robustness.

4.3. Cluster characterization

We characterized the four clusters produced by K-means as follows: large-scale irrigation water users, moderate irrigation water users, community water supply entities, and domestic water users.

The large-scale irrigation water users' cluster is characterized by large-scale irrigation users conducting agriculture activities within the basin. These users' abstraction of water was at a mean rate of 23.24 L/s. The large-scale irrigators comprise 22.4% of all users in the basin. Most use tap water from the rivers using dedicated pipes or furrows. The moderate irrigation water user's cluster is characterized by small-scale irrigation users comprising 29.3% of basin water users. These were individuals who depended on agriculture as their income generation activities. Their water abstraction mean rate was 3.98 L/s. Most users in this cluster tapped water from rivers and springs through traditional furrows. The community water supply entities cluster consisted of 25.4% of users in the basin, cross-cutting a wide range of categories. In this cluster, we found the

CBWSO, private and public institutions, and water utility companies. Local communities form the CBWSO to abstract water and distribute it to community members (villagers) for various household purposes such as domestic use, house-scale farms, irrigation, and livestock keeping. The community member pays a monthly fee for administration purposes to the CBWSO's leadership. CBWSO users abstract water at a mean rate of 1.17 L/s. Institutions abstract water for their office use at a mean rate of 2.28 L/s.

Water utilities were categorized into three classes (A, B, and C) per their permit. They abstracted and sold water to users through distribution channels. Their water abstraction mean rates were 5.75 L/s, 3.48 L/s, and 3.13 L/s for the classes A, B, and C, respectively. The fourth cluster, named the domestic water user, consisted of 22.9% of all users in the basin who abstracted water for their home usage. We observed that domestic users abstracted water at a mean rate of 2.62 L/s. The last three (2–4) clusters had low water abstraction mean rates. We argued that these clusters were the diffusion of one among the two less homogeneous clusters depicted by the Silhouette diagrams and Gap Statistics. The four clusters are highly homogenous, thus best describing the basin's socioeconomic activities.

The study by Rahim et al. [13] employed clustering algorithms to segment datasets. They used two datasets: an engineered features dataset clustered using K-means and time of use and a weighted probability of use dataset clustered using the Agglomerative Hierarchical technique. Finally, they described the characteristics of each cluster identified. Unlike the present study, a single dataset (comprised of 3460 records) of water river basin users was used. One of the three clustering algorithms was identified as an appropriate technique to validate the clusters for robustness and generalization to an unseen dataset, thus describing each cluster's characteristics. Jenifer and Jha [23] used the Agglomerative Hierarchical clustering technique to segment the rain gauge station dataset into six clusters and highlight the clusters' characteristics. We argue that understanding water users' characteristics is paramount for water governing bodies and policymakers when establishing strategies that should guide efficient and effective ways towards water governance and rationing along the river basins.

5. Conclusions and future work

The optimal goal of this study was to obtain validated clusters that best-characterized water user along the Pangani Basin. The study employed a data-driven multi-model characterization approach. The approach applies to areas where data are available, and characteristics of the validated cluster types are described as large-scale irrigation water users, moderate irrigation water users, community water supply entities, and domestic water users. The study's findings would benefit water governing bodies and policymakers when strategizing water governance and rationing. Furthermore, researchers could use these findings as a stepping point for further studies. Since the dataset was acquired from a second dataset and the study was designed solely on the PBWB dataset, we propose future work to validate the clusters with the dataset, which includes a wide range of parameters such as the specific type of crop planted by the water user, particular number of livestock owned by the water user, number of tenants for domestic users as well as automated devices to measure water abstraction and sources' capacities in real-time bases. We also propose analyzing the clusters for hidden patterns not revealed by the clustering techniques using frequent pattern and association rule mining techniques.

Including a wide range of parameters and association rule mining will offer more relevant and valuable insights.

Acknowledgements. The authors extend gratitude to the referee for their valuable suggestions and comments, which have significantly enhanced the quality of the original manuscript.

Conflict of interest. The authors declare that they have no conflict of interest.

Author's Contributions: All authors have equal contribution.

REFERENCES

1. G.Grill, B.Lehner, M.Thieme, B.Geenen, D.Tickner, F.Antonelli, S.Babu, P. Borrelli, L.Cheng and H.Crochetiere, Mapping the world's free-flowing rivers, *Nature*, 569 (2019) 215–221.
2. S.M.Basak, M.S.Hossain, J.Tusznio and M.Grodzińska-Jurczak, Social benefits of river restoration from ecosystem services perspective: A systematic review, *Environ. Sci. Policy*, 124 (2021), 90–100.
3. A.M.Rodríguez-Pérez, C.A.Rodríguez-Gonzalez, R.López, J.A.Hernández-Torres and J.J.Caparrós-Mancera, Water Microturbines for Sustainable Applications: Optimization Analysis and Experimental Validation, *Water Resour. Manag*, 38 (2024) 1011–1025.
4. B.Delineation, Pangani Basin Water Board, 2023, 2009.
5. H.A.Ougougdal, M.Y.Khebiza, M.Messouli and A.Lachir, Assessment of futurewater demand and supply under IPCC climate change and socio-economic scenarios using a combination of models in Ourika watershed High Atlas Morocco, *Water*, 12 (2020) 1751.
6. M.J.Colloff, T.M.Doody, I.C.Overton, J.Dalton and R.Welling, Re-framing the decision context over trade-offs among ecosystem services and wellbeing in a major river basin where water resources are highly contested, *Sustain. Sci*, 14 (2019) 713–731.
7. A.Omer, N.A.Elagib, M.Zhuguo, F.Saleem and A.Mohammed, Water scarcity in the Yellow River Basin under future climate change and human activities, *Sci. Total Environ*, 749 (2020) 141446.
8. D.Zhang, M.S.Sial, N.Ahmad, J.A.Filipe, P.A.Thu, M.Zia-Ud-din and A.B.Caleiro, Water scarcity and sustainability in an emerging economy: A management perspective for future. *Sustainability*,13 (2021) 144.
9. G.R.Kattel, State of future water regimes in the world's river basins: Balancing the water between society and nature, *Crit. Rev. Environ. Sci. Technol*, 49 (2019) 1107–1133.
10. D.Garrick, L.De Stefano, W.Yu, I.Jorgensen, E.O'Donnell, L.Turley, I.Aguilar-Barajas, X.Dai, R.de Souza Leão and R.Punjabi, Rural water for thirsty cities: A systematic review of water reallocation from rural to urban regions, *Environ. Res. Lett*, 14 (2019) 043003.
11. S.S.Atef, F.Sadeqinazhad, F.Farjaad and D.M.Amatya, Water conflict management and cooperation between Afghanistan and Pakistan, *J. Hydrol*, 570 (2019) 875–892.

12. A.Mianabadi, K.Davary, H.Mianabadi and P.Karimi, International Environmental Conflict Management in Transboundary River Basins, *Water Resour. Manag*, 34 (2020) 3445–3464.
13. M.S.Rahim, K.A.Nguyen, R.A.Stewart, T.Ahmed, D.Giurco and M.Blumenstein, A clustering solution for analyzing residential water consumption patterns, *Knowl.-Based Syst*, 233 (2021) 107522.
14. P.Ardarsa and O.Surinta, Water Quality Assessment in the Lam Pa Thao Dam, Chaiyaphum, Thailand with K-Means Clustering Algorithm, In Proceedings of the-2021 Research Invention Innovation Congress Innovation Electricals and Electronics RI2C 2021 Bangkok Thailand, 1–3 September 2021 pp. 35–39.
15. Z.K.Feng, W.J.Niu, R.Zhang, S.Wang and C.T.Cheng, Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization, *J. Hydrol*, 576 (2019) 229–238.
16. X.K.Bui, M.S.Marlim and D.Kang, Water network partitioning into district metered areas: A state-of-the-art review, *Water,* 12 (2020) 1002.
17. A.E.Ioannou, E.F.Creaco and C.S.Laspidou, Exploring the effectiveness of clustering algorithms for capturing water consumption behavior at household level, *Sustainability,* 13 (2021) 2603.
18. Y.Qin and Y.Lou, Hydrological time series anomaly pattern detection based on isolation forest, In Proceedings of the 2019 IEEE 3rd Information Technology Networking Electronic and Automation Control Conference (ITNEC) Chengdu China, 15–17 March 2019 pp. 1706–1710.
19. M.M.Jafari, H.Ojaghlou, M.Zare and G.J.P.Schumann, Application of a novel hybrid wavelet-anfis/fuzzy c-means clustering model to predict groundwater fluctuations, *Atmosphere*, 12 (2021) 9.
20. Tiyasha, T.M.Tung and Z.M.Yaseen, A survey on river water quality modelling using artificial intelligence models: 2000–2020, *J. Hydrol*, 585 (2020) 124670.
21. J.Li, D.Hassan, S.Brewer and R.Sitzenfrei, Is clustering time-series water depth useful? An exploratory study for flooding detection in urban drainage systems, *Water,* 12 (2020) 2433.
22. A.Mosavi, M.Golshan, B.Choubin, A.D.Ziegler, S.K.Sigaroodi, F.Zhang and A.A.Dineva, Fuzzy clustering and distributed model for streamflow estimation in ungauged watersheds, *Sci. Rep*, 11 (2021) 8243.
23. M.A.Jenifer and M.K.Jha, Assessment of precipitation trends and its implications in the semi-arid region of Southern India, *Environ. Chall*, 5 (2021) 100269.
24. Shahfahad, S.Talukdar, A.R.M.T.Islam, T.Das, M.W.Naikoo, J.Mallick and A.Rahman, Application of advanced trend analysis techniques with clustering approach for analysing rainfall trend and identification of homogenous rainfall regions in Delhi metropolitan city, *Environ. Sci. Pollut. Res*, 30 (2022) 106898-106916.
25. Z.Gao, Y.Li, S.Qu and M.Xu, Supply chain-wide sectoral water use characteristics based on multi-perspective measurements, *J. Clean. Prod*, 268 (2020) 122345.
26. M.S.Rahim, K.A.Nguyen, R.A.Stewart, D.Giurco and M.Blumenstein, Machine learning and data analytic techniques in digitalwater metering: A review, *Water,* 12 (2020) 2M.S.94.

27. R.Trach, Y.Trach, A.Kiersnowska, A.Markiewicz, M.Lendo-Siwicka and K.A.Rusakov, Study of Assessment and Prediction of Water Quality Index Using Fuzzy Logic and ANN Models, *Sustainability,* 14 (2022) 5656.
28. A.M.Rodríguez-Pérez, C.A.Rodríguez, A.Márquez-Rodríguez and J.J.C.Mancera, Viability analysis of tidal turbine installation using fuzzy logic: case study and design considerations, *Axioms*, 12 (2023) 778.
29. A.M.Rodríguez Pérez, C.A.Rodríguez, L.Olmo Rodríguez and J.J.Caparros Mancera, Revitalizing the Canal de Castilla: a community approach to sustainable hydropower assessed through fuzzy logic, *Appl. Sci.* 14 (2024) 1828.
30. M.H.Al-Adhaileh and F.W.Alsaade, Modelling and prediction of water quality by using artificial intelligence, *Sustainability,* 13 (2021) 4259.
31. Y.Chen, L.Song, Y.Liu, L.Yang and D.Li, A review of the artificial neural network models for water quality prediction, *Appl. Sci*, 10 (2020) 5776.
32. S.Kouadri, A.Elbeltagi, A.R.M.T.Islam and S.Kateb, Performance of machine learning methods in predicting water quality index based on irregular data set: Application on Illizi region (Algerian southeast), *Appl. Water Sci,* 11 (2021) 190.
33. H.Tyralis, G.Papacharalampous and A.A.Langousis, A brief review of random forests for water scientists and practitioners and their recent history in water resources, *Water,* 11 (2019) 910.
34. U.Ahmed, R.Mumtaz, H.Anwar, A.A.Shah, R.Irfan and J.García-Nieto, Efficient water quality prediction using supervised machine learning, *Water,* 11 (2019) 2210.
35. F.Ghobadi and D.Kang, Application of machine learning in water resources management: a systematic literature review, *Water* 15 (2023) 620.
36. I.F.Ashari, E.Dwi Nugroho, R.Baraku, I.Novri Yanda and R.Liwardana, Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta, *J. Appl. Inform. Comput,* 7 (2023) 89–97.
37. G.Shenbagalakshmi, A.Shenbagarajan, S.Thavasi, M.Gomathy Nayagam and R.Venkatesh, Determination of water quality indicator using deep hierarchical cluster analysis, *Urban Clim,* 49 (2023) 101468.
38. M.Z.Rodriguez, C.H.Comin, D.Casanova, O.M.Bruno, D.R.Amancio, L.D.F.Costa and F.A.Rodrigues, Clustering algorithms: a comparative approach, *PloS ONE,* 14 (2019) e0210236.
39. D.G.Nyambo, E.T.Luhanga, Z.O.Yonah and F.D.N.Mujibi, Application of multiple unsupervised models to validate clusters robustness in characterizing smallholder dairy farmers, *Sci. World J,* 2019 (2019) 1020521.
40. NBS Census Tanzania National Bureau of Statistics, 2022 (2022)
41. P.Basin and O.Basins, Pangani basin total water resources share of pangani basin in tanzania's renewable water resources basin area average surface water runoff in catchments pangani basin water infrastructure key figures water points by sources of water borehole shallow well. Volume, 870 (2015) p. 91.
42. K.P.Sinaga and M.S.Yang, Unsupervised K-means clustering algorithm, *IEEE Access*, 8 (2020) 80716–80727.
43. E.K.Tokuda, C.H.Comin and L.D.F.Costa, Revisiting agglomerative clustering, *Phys. A Stat. Mech. Appl,* 585 (2022) 126433.

44. C.Yuan, H.Yang, Research on K-Value Selection Method of K-Means Clustering Algorithm, *J*, 2 (2019) 226–235.
45. E.Eskandari, H. Mohammadzadeh, H.Nassery, M.Vadiati, A.M.Zadeh and O.Kisi, Delineation of isotopic and hydrochemical evolution of karstic aquifers with different cluster-based (HCA, KM, FCM and GKM) methods, *J. Hydrol*, 609 (2022) 127706.
46. X.Wang and Y.Xu, An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index, *IOP Conf. Ser. Mater. Sci. Eng*, 569 (2019) 052024.