2024-12-07

Characterizing Water Users through Frequent Patterns and Association Rules by Using Apriori Algorithm: A Case of Pangani Basin Tanzania

Lyuba, Matimbila

IJST

https://doi.org/10.17485/IJST/v17i45.3526

INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Characterizing Water Users through Frequent Patterns and Association Rules by Using Apriori Algorithm: A Case of Pangani Basin Tanzania

Matimbila P Lyuba[1]*, Devotha G Nyambo[1], Anael Sam[1], Seifu Tilahun[2,3]

1 School of Computational and Communication Sciences and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania
2 International Water Management Institute, Accra, Ghana
3 Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Ethiopia

OPEN ACCESS

Received: 26-10-2024
Accepted: 11-11-2024
Published: 07-12-2024

Citation: Lyuba MP, Nyambo DG, Sam A, Tilahun S (2024) Characterizing Water Users through Frequent Patterns and Association Rules by Using Apriori Algorithm: A Case of Pangani Basin Tanzania. Indian Journal of Science and Technology 17(45): 4694-4703. https://doi.org/10.17485/IJST/v17i45.3526

* **Corresponding author**.

lyubam@nm-aist.ac.tz

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To identify the hidden patterns in the K-means clustered dataset for the Pangani Basin using the Apriori algorithm through frequent patterns and association rules to enrich cluster characteristics. **Methods:** Frequent patterns and association rule mining were used to discover the hidden attributes in the K-means clustered dataset. Measures of minimum support ranging from 0.5% to 5% and minimum confidence ranging from 50% to 100% were used to generate a manageable number of rules which were then filtered for redundancy. Lift value >1.0 was used to determine the rule's interestingness while Arules and ArulesViz in R were used to visualize generated rules. **Findings:** Clusters one to four generated 25, 31, 47, and 49 rules respectively at a minimum confidence of 50% and minimum support of 2% in the first two clusters and 1% in other clusters. Furthermore, water users in cluster one were observed to abstract more water than the three clusters, while their water use fee also reflected on the amount they abstracted. In clusters two and three, water users identified the same amount of water source capacity but differed in the amount requested and water use fee. Water users in cluster four were identified with less water source capacity and fewer amounts abstracted than other clusters. However, their water use fee identified was higher than those in cluster three, with high water source capacity and high amount requested. Such a difference is attributed to the type of water use for cluster three users being domestically supplied through community water supply entities to help villagers access water. In contrast, the water use for users in cluster four is domestic and commercial. **Novelty:** When aggregated with the clustering observations, the identified association rules mining results provide a broad understanding of water users' characteristics for better water allocation and rationing.

Keywords: Association rule; Frequent Patterns; Apriori; Characterization; Pangani Basin

1 Introduction

Pangani Basin is the main water source for the population around its boundaries [1]. Individuals and groups of water users abstract water from rivers, streams, springs, lakes, and wetlands within the Basin. Water abstracted supports the social-economic activities undertaken along the basins, such as agriculture through irrigation systems, livestock keeping, home use, and mining [2]. The Pangani Basin water resource data sheet informs that approximately 75% of users abstracting water suffer from water stress [3]. Despite the scarce water resources, the population along the Basin relies on the Basin's water to drive their social-economic activities [4,5]. Therefore, an in-depth understanding of water user characteristics related to water utilization based on their devise formations and how they are relied on with particular formations is paramount for strategized water allocation and rationing.

Water users' characteristics in Basins can be well understood when categorized into homogeneous groups or clusters [6]. Based on the defined homogeneous groups, various attributes can be identified and analyzed to determine their similarities and dissimilarities among the clusters. The K-means clustering algorithm categorizes the Pangani Basin water users into homogeneous groups and informs on the characteristics of each cluster identified based on three attributes: amount_abstracted, water_use and water_use_category. Despite informing the cluster formation and their characteristics for the Pangani Basin dataset, the clustering algorithm could not identify hidden attributes among many attributes in the dataset, which should describe the clusters' characteristics in detail.

Association Rule Mining (ARM) is a technique used to find intriguing connections between objects in a given dataset [7]. ARM makes it possible to determine the relationship between attributes based on the frequency of certain items. The data analyst interprets the patterns found through ARM as knowledge by seeing them as rules, trees, or clusters. Association rules that satisfy the measures of minimum support and minimum confidence are described as strong rules, while those that satisfy the minimum lift value are the rules of interestingness [7]. Normally, ARM in analyzing various dataset's insights aims to reveal the patterns of interest, which could inform how the dataset attributes relate. ARM approaches have proven effective in several industries, including natural language processing [8], market basket analysis [9] education [10], healthcare [7], and recommendation systems, among others.

In recent studies, Zhong et al. [11] employed the K-means algorithm and ARM to analyze the causes and consequences of flash floods in the humid southern region of China. In their study, K-means clustered risk factors of flash floods into four homogeneous clusters, while ARM revealed rules indicating a weak correlation between the identified risk factors of 24-h rainfall and soil moisture as well as a strong correlation between soil type and soil moisture. They further observed that total rainfall, soil type, and soil moisture are the crucial risk factors for flash floods along the Upper Hanjiang River. Despite their dataset being limited to 31 events, it demonstrated ARM's ability to enhance the results of clustering techniques. Mirhashemi & Mirzaei [12] employed the K-means clustering technique and the Apriori algorithm to examine the impact of various climatic factors such as precipitation, temperature, humidity, evapotranspiration, water volume per crop, orchards posed on the amount of water that is delivered into the irrigation network. In the 6 clusters identified, 18 interesting rules were found describing varying influences of the climatic factors in the quantity of water supplied into the irrigation network. The clustering algorithm did not previously identify this knowledge; thus, it informs us of the benefits of using ARM to analyze the dataset deeply for knowledge mining. Another study by Liu et al. [13] used the K-means

algorithms and Apriori to identify 47 rules that describe patterns behind the reason for water supply fluctuations and the duration of the fluctuations in the Shenzhen sub-provincial of China for the three identified clusters. In their study, they introduced an objective function that optimizes the lift value of the valid rules to increase reliability. However, this approach has a performance impact as the dataset grows.

In another study, Apriori and FP-Growth were used to discover hidden patterns related to water quality parameters that frequently co-occur in a given raw dataset on water quality [14]. The results identified rules informing that sulphates positively correlate with ions that cause water hardness. Nyambo et al. [15] demonstrated the use of ARM to enrich the results of clustering techniques, thus providing a broad understanding of the subject of analysis and extraction of informed knowledge. Regarding the case of the Pangani Basin where users were clustered according to their water usage, this study employed the strength of ARM to address the gap left by the clustering algorithm. Through frequent patterns and ARM, this paper presents the Apriori algorithms to discover hidden attributes on the Pangani clustered dataset not identified by the clustering algorithm. The results improve the previous clustering results and provide a clear understanding of cluster characteristics. This knowledge is significant to the water governing organs and policymakers, particularly when strategizing water allocation and rationing along the Basin. It further contributes to the knowledge of the approach that could address clustering drawbacks. Despite the dataset obtained from the Pangani Basin, we urge that similar results for both clustering and association rule mining should prevail in other basins with a similar setup as the Pangani.

1.1 Background

Literature indicates that hidden patterns can be discovered using various ARM algorithms [16]. These algorithms differ in mining approaches depending on the knowledge to be discovered. There are algorithms that rely on frequent patterns such as Apriori, DIC, Apriori-Tid, FP-Growth, H-Mine, FP-Max, ECLAT, and Charm [17]. Other algorithms, such as SPADE and GSP, dealt with sequential pattern mining, while FSG and GSPAN mined structured patterns [17]. The choice of ARM depends on the data structure and knowledge to be discovered. The Apriori, FP-Growth, and ECLAT are the widely used algorithms in identifying hidden patterns. However, when candidate set generation is crucial in identifying frequent item sets, Apriori is preferred over the others [18]. FP-Growth is a tree-based algorithm that relies on a deep first search strategy, while the ECLAT works on a level-wise search in a vertically converted dataset [13]. To better visualize the generated rules, tools such as Arules, AruleViz, Orange, rapidMiner, Weka, and KNIME are applicable [19]. Despite numerous visualization tools, Arules and AruleViz provide flexibility in graph generation, parameter tuning, and additional extensions [15].

2 Methodology

2.1 Study Area and Data Preparation

The dataset was acquired from the Pangani Basin Water Board (PBWB) (see Figure 1) between January 2023 and June 2023. Based on preliminary analysis, a K-means clustered dataset with four (4) clusters of water users was used to study all frequent patterns to obtain interesting rules in each identified cluster. The clustered dataset comprises 3460 records.

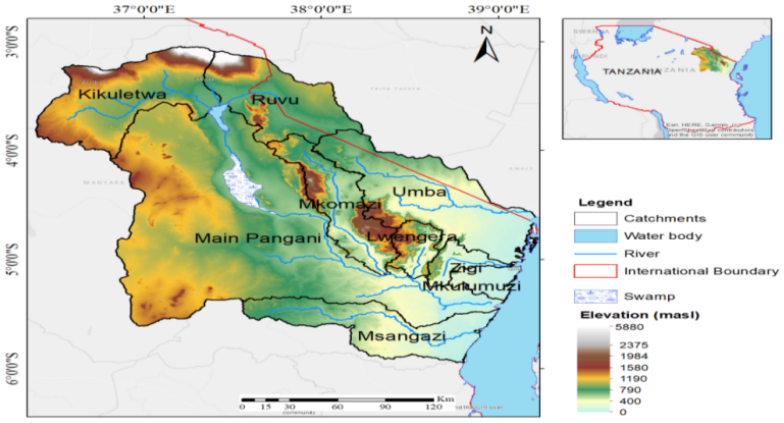

Fig 1. The Pangani Basin [4]

Table 1. Variables used on frequent pattern and association rule mining

SSN	Feature Name	Encoding	Feature Type	Feature description and metrics
1	source_type	0-9	Discrete	Type of source where water is abstracted, i.e. river, spring, borehole, etc.
2	source_name	0-1169	Discrete	Name given to the specific water source
3	water_use_fee	100000-2918734	Continuous	Amount paid (TZS) by the water user as the fee for abstracting water
4	amount_requested	0.1-140500	Continuous	Amount of water (L/s) requested by the user once logged in to an application. Users can be granted less than or equal to the amount requested depending on the water assessment conducted at the source.
5	water_source_capacity	0.1-140500	Continuous	Amount of water (litres/s) available in the source
6	permit_status	0(invalid)-1(valid)	Boolean	Whether the permit is active or expired
7	catchment	0-9	Discrete	Catchment where a user is abstracting water
8	cluster_label	0-3	Discrete	The homogeneous group, a user, belongs among the four clusters

Table 1 lists the variables used to study the frequent pattern, thus discovering rules of interestingness. These variables were chosen for analysis since the K-means clustering did not reveal them even though they influence water allocation rationing manually conducted by the PBWB when attending to water users' applications.

2.2 Water Users Clusters

The K-means clustered dataset was studied and characterized by the amount_abstracted, water_use, and water_use_category. The missing values were handled using a mean imputation approach to retain data representation. Outliers were removed using the interquartile range (IQR) with a capping. The dataset was standardized to have a unit standard deviation (σ = 1) and a mean of roughly zero (μ = 0) after the nominal characteristics were transformed into discrete values. The Principal Component Analysis (PCA) was used to address feature correlation while Hopkins test >0.9 validated the dataset clusterability. The clustered dataset was evaluated using the Calinski–Harabasz Index (CHI) and the Davies–Bouldin Index (DBI) metrics among the candidate clustering algorithms which were the K-means and Agglomerative Hierarchical (AH). The K-means outperforms the AH by prevailing a CHI of 692.3 compared to 578.2 and a DBI value of 1.8 compared to 1.9 of the AH. The two algorithms were further validated for generalization to unseen data by fitting each clustered dataset to the logistic regression. The K-means clustered dataset outperforms AH with a prediction accuracy of 98.2% over 97.5% respectively and becomes the dataset for use. The K-means characteristics of each cluster are detailed in Table 2.

Table 2. Attributes characterizing each cluster type by K-means algorithm

Cluster #	Cluster Size	Characteristics based on dominant attributes	Cluster Name
1	22.4%	amount_abstracted:23.18l/s water_use_category: irrigation, water_use: large scale irrigation	Large-scale irrigation water users
2	29.3%	amount_abstracted: 3.98l/s, water_use_category: irrigation , water_use: small scale irrigation	Small-scale irrigation water users
3	25.4%	amount_abstrated: 2.87l/s, water_use_category: Community Water Supply, water_use: home(individual)	Community water supply entities
4	22.9%	amount_abstracted: 2.62l/s, water_use_category: domestic, and water_use: home (individual).	Domestic water users

2.3 Association Rule Mining (ARM)

The R programming tool was used for rule mining and data analysis. The Arules package was used to generate the association rules, and the ArulesViz package was used to visualize rules using group-matrix [15]. Different values for minimum support

(minSup) ranging from 0.5% to 5% and minimum confidence (minConf) ranging from 50% to 100% were used to generate rules that can be visible (~100) for evaluation of relevance and interestingness of the rules. Support, Confidence, and Lift measures were used to validate the generated rules [20].

Support: This metric counts the number of times an item has occurred within the dataset.

Count: This is the total number of observations in the dataset that satisfy a specific rule.

Confidence: Measures how true the rule is. i.e. for a rule $[A \cup B] \rightarrow C$, the confidence measure is calculated as the ratio of the support of $[[A \cup B] \cup C]$ to the support of $[A \cup B]$

$$Confidence\ [A \cup B] \rightarrow C = \frac{support\ ([A \cup B] \cup C]}{support\,[A \cup B]} \tag{1}$$

where of $[A \cup B]$ is the antecedent and C is the consequent.

Lift refers to the deviation of the support of a whole rule from the support expected under independence, given the support of the antecedents and that of the consequent. Lift ($[A \cup B] \rightarrow$ C) is computed as;

$$Lift = \frac{support\,[[A \cup B] \cup C]}{support[A].support[B].support[C]} \tag{2}$$

where $[A \cup B]$ is the antecedent and C is the consequent.

When the lift value of an association rule is high, it indicates that the frequent items have greater collective strength than when separated. Furthermore, a lift value greater than 1 implies that the two items have a positive correlation, a lift value less than one means that items have a negative correlation, and a lift value of 1 implies no correlation between items. Interesting rules exceed the values of minSup, minConf and possess a lift value greater than 1 [20].

2.4 Estimation of the minSup, minConf, Lift value, and number of rules

For this work's purpose, graphs with different values of minSup between 0.5% and 5% and the minConf between 50% and 100% were plotted to indicate the possible number of rules that can be generated in each cluster followed by the removal of redundant rules. We urge that rules with minConf greater or equal to 50% are strong rules, while those with less are weak rules. Rules with a lift value greater or equal to 1 are the interesting rules. Table 3 details the number of interesting rules obtained in each cluster with corresponding measures.

Table 3. Estimation of interesting rules in each cluster type

Cluster #	minSup	minConf	Lift range	Number of rules
1	2%	50%	2.23 to 4.46	25
2	2%	50%	1.70 to 3.40	31
3	1%	50%	1.97 to 3.94	48
4	1%	50%	2.18 to 4.36	49

3 Results and Discussion

3.1 Results

3.1.1 Cluster 1: Large-scale irrigation water users

The K-means characterized this cluster by amount_abstracted (23.18L/s), water_use (large-scale irrigation), and water_use_category (irrigation). In this cluster, 25 nonredundant rules were extracted at 2% minSup and 50% minConf, describing a positive correlation between the antecedent and consequent. Among the 25 rules, 9 were observed with a minSup ranging between 0.02 to 0.13, minConf of 100%, and lift >4.46. Rules discovered characterize that this cluster by water_source_capacity (2636.30L/s to 6587L/s), water_use_fee (1761440TZS to 2918734TZS), amount_requested (15L/s to 26.975L/s), catchment (Pangani Mainstream, Kikuletwa, Ruvu), source_type (river, spring) and permit_status (valid, invalid). The orderly of the antecedents by lift value is depicted in Figure 2. We urge that we are at least 50% confident that attributes identified through the 25 discovered rules characterized water users belong to cluster one. Both sets of attributes describe the requirements of the majority members in this cluster type, hence significant in guiding water allocation along the basin.

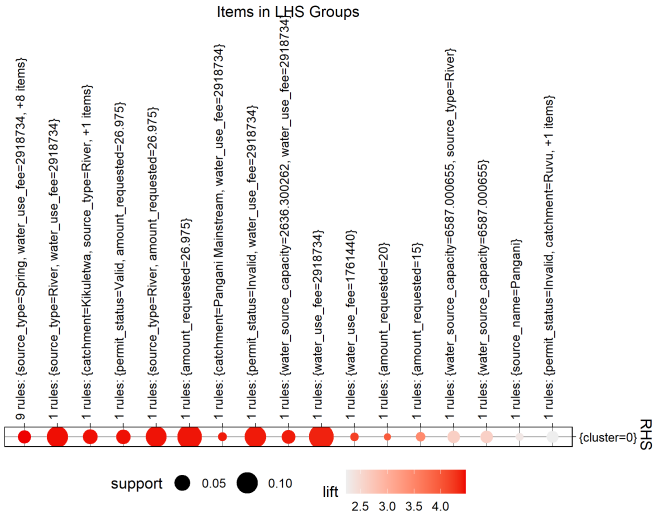

Fig 2. Association rules for the cluster 1

3.1.2 Cluster 2: Small-Scale irrigation water users

The K-means characterized this cluster by amount_abstracted (3.98L/s), water_use (small-scale irrigation), and water_use_category (irrigation). ARM extracted 31 rules at 2% minSup and 50% minConf, describing a positive correlation between the antecedent and consequent. Among the 31 rules the first 2 rules; R1 {source_type=Spring, water_use_fee=300000, permit_status=Invalid} => {cluster=1} minSup 0.03, minConf 100% and R2 {water_use_fee=300000, water_source_capacity=2636.300262, permit_status=Invalid} => {cluster=1} minSup0.04, minConf 100%, were observed with a lift value of 3.40 each. The identified rules characterize this cluster by water_source_capacity (2636.30L/s), water_use_fee (300000TZS to 1130720TZS), amount_requested (10L/s), catchment (Kikuletwa, Ruvu), source_type (river, spring) and permit_status (invalid). The orderly of the antecedents by lift value is depicted in Figure 3. We urge that we are at least 50% confident that attributes identified through the 31 discovered rules characterized water users belong to cluster two. The two sets of attributes are essential in directing water allocation since they define the requirements of the majority members in the formed cluster.

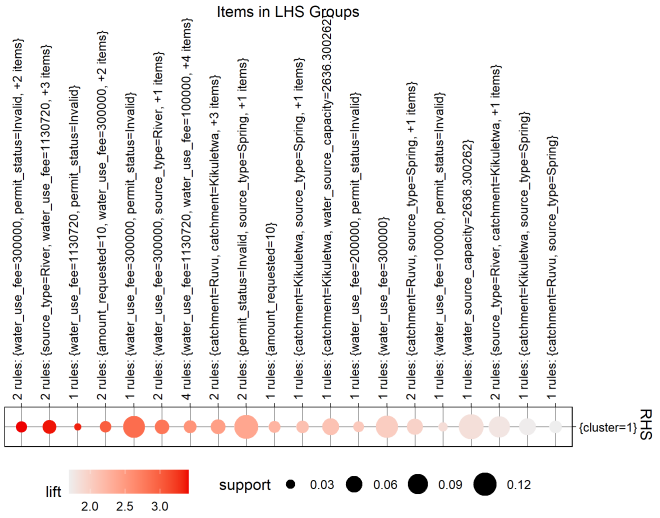

Fig 3. Association rules for the cluster 2

3.1.3 Cluster 3: Community water supply entities

The K-means characterized this cluster by amount_abstracted (2.87L/s), water_use (individual/home), and water_use_category (Community water supply entities). Community water supply entities abstract water and distribute to villagers for domestic use. They abstract from multiple water sources. ARM extracted 47 rules at 1% minSup and 50% minConf, describing a positive correlation between the antecedent and consequent. Among the 47 rules the first 5 rules; R1 {source_name=Borehole, permit_status=Valid, catchment=Zigi} => {cluster=2} minSup 0.01, minConf 100%, R2 {source_type=Borehole, permit_status=Valid,catchment=Zigi} => {cluster=2} minSup 0.01, minConf 100%, R3 {water_use_fee=200000, permit_status=Valid, catchment=Ruvu} => {cluster=2} minSup 0.01, minConf 100%, R4 {source_name=Borehole, permit_status=Valid, catchment=Pangani Mainstream} => {cluster=2} minSup 0.02, minConf 100%, R5 {source_type=Borehole, permit_status=Valid, catchment=Pangani Mainstream} => {cluster=2} minSup 0.02, minConf 100%, were observed with a lift value of 3.94 each. The 47 discovered rules characterize this cluster type by water_source_capacity (0.5L/s to 2636.30L/s), water_use_fee (100000TZS to 300000TZS), amount_requested (0.5L/s to 5L/s), catchment (Pangani Mainstream, Ruvu, Zigi), source_type (Borehole, river, spring), source_name (borehole) and permit_status (valid, invalid). The orderly of the antecedents by lift value is depicted in Figure 4. We urge that we are at least 50% confident that attributes identified through the discovered rules characterized water users belong to this cluster type. Since the two sets of attributes define the needs of most of this cluster type, they are essential for directing water allocation along the Basin.

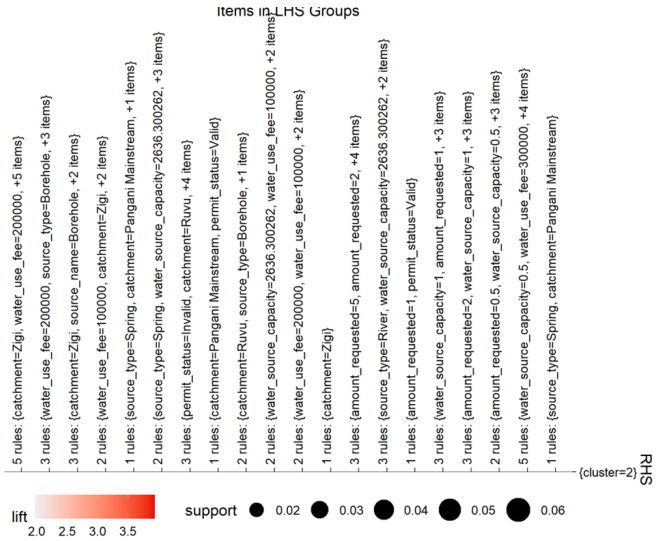

Fig 4. Association rules for the cluster 3

3.1.4 Cluster 4: Domestic water users

The K-means characterized this cluster by amount_abstracted (2.62L/s), water_use (individual), and water_use_category (domestic). In this cluster, 49 rules were extracted at 0.01 minSup and 50% minConf, describing a positive correlation between the antecedent and consequent. Among the 49 rules the first 2 rules; R1 {source_name=Borehole, water_use_fee=200000, permit_status=Valid,catchment=Kikuletwa}=>{cluster=3} minSup 0.03, minConf 100%, R2{source_type=Borehole,water_use_fee=200000,permit_status=Valid,catchment=Kikuletwa}=>{cluster=3} minSup 0.03, minConf 100% ,were observed with a lift value of 4.36 each. ARM discovered that cluster 4 is characterized by water_source_capacity (0.5L/s to 2L/s), water_use_fee (200000TZS to 300000TZS), amount_requested (0.5L/s to 2L/s), catchment (Kikuletwa), source_type (borehole, spring), source_name(borehole) and permit_status (Valid). The orderly of the antecedents by lift value is depicted in Figure 5. We urge that we are at least 50% confident that attributes identified through the discovered rules characterized water users belong to this cluster type. Water allocation is greatly aided by both sets of criteria, which define the needs of the majority members in this cluster type.

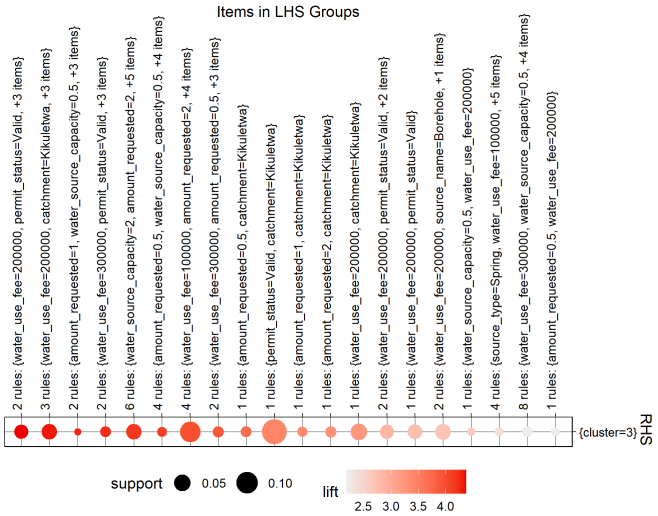

Fig 5. Association rules for the cluster 4

3.2 Discussion

The K-means algorithm has been used to describe water users with the primary objective of comprehending the characteristics of homogeneous groups within the Pangani Basin for effective water allocation and rationing. However, the clustering algorithm is limited in explaining the contribution of the least observed attributes in characterizing the homogeneous clusters along the Pangani Basin. This study used ARM to highlight recurring themes and elements in a clustered dataset of the Pangani Basin water users to discover hidden attributes. The results from this study are key inputs in understanding various attributes that characterize water users and chances to predict their water allocation and rationing based on the application of interesting rules.

The method adopted in this study to reveal hidden attributes of the clustered dataset is similar to that used by Zhong et al. [11] to analyze the causes and consequences of the flash flood happened in China. ARM indicated additional attributes apart from those depicted in Table 2, and for the large-scale irrigation water users cluster 1, K-means characterized it with amount_abstracted (23.18L/s), water_use_category (irrigation) and water_use (large-scale irrigation). ARM revealed that this cluster type can additionally be characterized by water_source_capacity (2636.30L/s to 6587L/s), water_use_fee (1761440TZS to 2918734TZS), amount_requested (15L/s to 26.975L/s), catchment (Pangani Mainstream, Kikuletwa, Ruvu), source_type (river, spring) and permit_status (valid, invalid) and for the small-scale irrigation water users cluster two, K-means characterized it with amount_abstracted (3.98L/s), water_use_category (irrigation) and water_use (small scale irrigation). ARM discovered that additional attributes were water_source_capacity (2636.30L/s), water_use_fee (300000TZS to 1130720TZS), amount_requested (10L/s), catchment (Kikuletwa, Ruvu), source_type (river, spring) and permit_status (invalid), for the Community water supply entities cluster three, K-means characterized it by amount_abstracted (2.87L/s), water_use_category (Community Water Supply) and water_use (individual/home). ARM informs additional attributes were water_source_capacity (0.5L/s to 2636.30L/s), water_use_fee (100000TZS to 300000TZS), amount_requested (0.5L/s to 5L/s), catchment (Pangani Mainstream, Ruvu, Zigi), source_type (borehole, river, spring), source_name(borehole) and permit_status (valid, invalid). Lastly, for the domestic water users cluster 4, K-means characterized it by the amount_abstracted (2.62L/s), water_use_category (domestic), and water_use (individual). ARM discovered additional attributes were water_source_capacity (0.5L/s to 2L/s), water_use_fee (200000TZS to 300000TZS), amount_requested (0.5L/s to 2L/s), catchment (Kikuletwa), source_type (borehole,spring), source_name(borehole) and permit_status (Valid).

In the present study, ARM complemented the characteristics identified by the K-means algorithm by revealing hidden patterns in the Pangani water user's dataset. The frequent items discover a wide set of attributes that help to understand the highlighted categories of users abstracting water from the Basin for various socioeconomic activities. The clustering algorithm characterizes the clusters based on three attributes, largely informing about the mean amount_abstracted and usage of water abstracted. ARM informed us about the types and water levels available in water sources where the user's abstract water, the location where the sources can be found, the amount requested, the amount paid as the water abstraction fee, and the validity of a user's permit. For instance, users belonging to cluster one are granted to abstract a high amount of water 23.18L/s which is within the range of their requested amount between 15L/s to 26.975L/s, they abstract water from sources with high water levels

between 2636.30L/s to 6587L/s compared to other clusters. Users in cluster two, despite abstracting a low amount of 3.98L/s, they abstract water in the source with a capacity of at least 2636.30L/s. Despite requesting 10L/s, they are permitted to abstract 3.98L/s compared to users in cluster one who are granted an abstraction amount close to their requests.

Water users in cluster three requested an amount ranging from 0.5L/s to 5L/s but granted to abstract 2.87L/s. They abstract from sources with a capacity of less than 2636.30L/s. Users in cluster four requested amounts ranging from 0.5L/s to 2L/s but are granted to abstract 2.62L/s. This is contrary with the first three clusters which were granted less than the requested amount. This is due to the fact that users in this cluster have high priority according to the Water Use Act 2009. We observed that water sources with a capacity of 2636.30L/s are shared with users in the first three cluster types, whereas users in cluster type one abstract more than other clusters. This prompt the need for water allocation and rationing to meet the demand of many applications submitted.

Furthermore, ARM revealed that the kikuletwa catchment has concentrated on users belonging to cluster four, where most hold valid water use permits. This differs from the first three clusters in which many users hold invalid permits. River and spring are observed as the frequent water source types where abstraction is conducted among the four clusters. Another unusual pattern observed is that users in cluster four are charged highly between 200000TZS to 300000TZS while abstracting 2.62L/s compared to users in cluster three who are charged between 100000TZS to 200000TZS with an abstraction rate of 2.87L/s. This informs that individuals who installed pipes or furrows to their homes for either domestic or commercial usage are charged highly compared to urban water supply entities that abstract water and distribute it to villagers who are considered as the least in terms of financial sustainability. This understanding solidifies our understanding of the characteristics of water users along the Pangani Basin. The results agree with Liu et al. [13] that ARM can be used to improve understanding described by the clustering algorithms. The use of ARM in identifying hidden patterns in the dataset using metrics such as minSup and minConf agrees with the majority conformant in the relationship rules that the idea of majority decisions should be preferred when making decisions [21].

The findings highlighted in this study point to significant characteristics that users within the examined cluster types possess that are not discernible from the clustering findings alone. According to Nyambo et al. [15] and Kusak et al. [19], the usage of association rules has given a thorough explanation for how the various clustering characteristics can be enriched, thus providing a broad picture of specific cluster characteristics. Thus, the features of the water user clusters are known as requirements for planning the allocation and rationing of water throughout the Pangani Basin. The identified requirements will be studied further in our future work of modelling water allocation and rationing to determine the appropriate abstraction rates that best utilize the scarce water resource and address challenges highlighted by Atef et al. [22] and Zang et al. [23].

ARM validation was limited on the metric of minimum support, minimum confidence, and lift values suggested by [20]. Other metrics such as leverage and conviction were not adopted since they are least used and accounting that the dataset was not huge. The clustered dataset used in this study had a maximum record count of 3460; rules filtering were not as challenging as it would be in huge databases, hence we did not encounter challenges with Apriori computational demands. For the huge dataset, other metrics, such as those suggested by Antonello et al. [24,25] can be used to mine hidden attributes effectively and address computation complexity. Despite that the dataset was collected from the Pangani Basin, both the clustering and association rule mining justify that the results obtained generalized the unseen dataset and were applicable to other basins with a similar setup as Pangani.

4 Conclusion

This study employed Apriori algorithm to analyze the frequent pattern and association rule hence identifying hidden attributes in the K-means clustered dataset for the Pangani basin. The discovered attributes enrich the present attributes identified by the K-means algorithm hence providing a clear understanding of water users characteristics along the Pangani basin and other basins with similar setups. Through applying ARM on the K-means clustered datasets, 25 rules and 31 rules were identified in cluster 1 and 2 at a minimum support of 2% and minimum support of 50% respectively. Likewise, 47 rules and 49 rules were discovered in cluster 3 and 4 at minimum support of 1% and minimum confidence of 50% respectively. We urge that ARM addresses the limitation of the K-means algorithm while characterizing water users along the basin, thus advising against relying on the clustering technique to prevent missing out on crucial information.

Despite its strength in identifying hidden attributes, the Apriori algorithm has drawbacks that should be improved in the future. The major limitation is its computational complexity and time consumption when handling large amounts of data. In such a situation, conviction and leverage metrics of measurement for rules interestingness are recommended.

References

1) Hao N, Sun P, He W, Yang L, Qiu Y, Chen Y, et al. Water Resources Allocation in the Tingjiang River Basin: Construction of an Interval-Fuzzy Two-Stage Chance-Constraints Model and Its Assessment through Pearson Correlation. *Water*. 2022;14(18). Available from: https://www.mdpi.com/2073-4441/14/18/2928.
2) A AUN, P BS. River Ecosystem Service in Settlement Development and History of Coastal Bangladesh: A Case Study on Kachua Upazilla. *J Res Archit Plan*. 2020;28(1):1–7. Available from: https://www.researchgate.net/publication/343240315_River_Ecosystem_Service_in_Settlement_Development_and_History_of_Coastal_Bangladesh_A_Case_Study_on_Kachua_Upazilla.
3) Basin P, Basins O, Pangani. 2015. Available from: http://www.cru.uea.ac.uk/data.
4) Delineation B. 2009. Available from: https://www.panganibasin.go.tz/.
5) Richards N. Water users associations in Tanzania: Local governance for whom? *Water (Switzerland)*. 2019;11(10). Available from: https://www.mdpi.com/2073-4441/11/10/2178#:~:text=WUAs%20are%20designed%20as%20the,irrigators%20through%20a%20permitting%20system.
6) Omer A, Elagib NA, Zhuguo M, Saleem F, Mohammed A. Water scarcity in the Yellow River Basin under future climate change and human activities. *Sci Total Environ*. 2020;749. Available from: https://doi.org/10.1016/j.scitotenv.2020.141446.
7) Telikani A, Gandomi AH, Shahbahrami A. A survey of evolutionary computation for association rule mining. *Inf Sci (Ny)*. 2020;524:318–352. Available from: https://doi.org/10.1016/j.ins.2020.02.073.
8) Alqahtani A, Alhakami H, Alsubait T, Baz A. A Survey of Text Matching Techniques. *Eng Technol Appl Sci Res*. 2021;11(1):6656–6661. Available from: https://etasr.com/index.php/ETASR/article/view/3968/2428.
9) Kurnia Y, Isharianto Y, Giap YC, Hermawan A, Riki. Study of application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm. *J Phys Conf Ser*. 2019;1175(1):1–7. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1175/1/012047/pdf.
10) Hartama D, Windarto P, Wanto A. The Application of Data Mining in Determining Patterns of Interest of High School Graduates. *J Phys Conf Ser*. 2019;1339(1):1–6. Available from: https://www.researchgate.net/publication/356427850_The_Application_of_Data_Mining_in_Determining_Patterns_of_Interest_of_High_School_Graduates.
11) Zhong M, Jiang T, Hong Y, Yang X. Performance of multi-level association rule mining for the relationship between causal factor patterns and flash flood magnitudes in a humid area. *Geomatics, Nat Hazards Risk*. 2019;10(1):1967–1987. Available from: https://www.tandfonline.com/doi/full/10.1080/19475705.2019.1655102.
12) Mirhashemi SH, Mirzaei F. Using combined clustering algorithms and association rules for better management of the amount of water delivered to the irrigation network of Abyek Plain, Iran. *Neural Computing and Applications* . 2022;34(5):3875–3883. Available from: https://link.springer.com/article/10.1007/s00521-021-06648-6.
13) Liu X, Sang X, Chang J, Zheng Y, Han Y. The water supply association analysis method in Shenzhen based on kmeans clustering discretization and apriori algorithm. *PLoS One*. 2021;16:1–21. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8341608/pdf/pone.0255684.pdf.
14) Jansi J, Jegathambal P, Arumainayagam SD. Identifying the underlying relationship between water quality parameters of the groundwater samples using association and clustering algorithms in Coimbatore district. *International Journal of Recent Technology and Engineering (IJRTE)*. 2019;8(2):177–185. Available from: https://www.researchgate.net/publication/364120624_Identifying_the_Underlying_Relationship_Between_Water_Quality_Parameters_of_the_Groundwater_Samples_using_Association_and_Clustering_Algorithms_in_Coimbatore_District.
15) Nyambo DG, Luhanga ET, Yonah ZO. Characteristics of smallholder dairy farms by association rules mining based on apriori algorithm. *International Journal of Society Systems Science*. 2019;11(2):99–118. Available from: https://www.researchgate.net/publication/368093995_Characteristics_of_smallholder_dairy_farms_by_association_rules_mining_based_on_apriori_algorithm.
16) Thurachon W, Kreesuradej W. Incremental Association Rule Mining with a Fast Incremental Updating Frequent Pattern Growth Algorithm. *IEEE Access*. 2021;9:55726–55741. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9399130.
17) Sharma A, Ganpati A. Association Rule Mining Algorithms: A Comparative Review. *Int Res J Eng Technol*. 2021;8(11):848–853. Available from: https://www.irjet.net/archives/V8/i11/IRJET-V8I11140.pdf.
18) Dhinakaran D, Prathap PMJ. Protection of data privacy from vulnerability using two-fish technique with Apriori algorithm in data mining. *The Journal of Supercomputing*. 2022;78:17559–17593. Available from: https://link.springer.com/article/10.1007/s11227-022-04517-0.
19) Kusak L, Unel FB, Alptekin A, Celik MO, Yakar M. Apriori association rule and K-means clustering algorithms for interpretation of pre-event landslide areas and landslide inventory mapping. *Open Geosciences*. 2021;13(1):1226–1244. Available from: https://www.researchgate.net/publication/355204627_Apriori_association_rule_and_K-means_clustering_algorithms_for_interpretation_of_pre-event_landslide_areas_and_landslide_inventory_mapping.
20) Hikmawati E, Maulidevi NU, Surendro K. Minimum threshold determination method based on dataset characteristics in association rule mining. *Journal of Big Data*. 2021;8. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00538-3.
21) Mercier H, Morin O. Majority rules: how good are we at aggregating convergent opinions? *Evolutionary Human Sciences*. 2019;1:1–17. Available from: https://dx.doi.org/10.1017/ehs.2019.6.
22) Atef SS, Sadeqinazhad F, Farjaad F, Amatya DM. Water conflict management and cooperation between Afghanistan and Pakistan. *Journal of Hydrology*. 2019;570:875–892. Available from: https://dx.doi.org/10.1016/j.jhydrol.2018.12.075.
23) Zhang D, Sial MS, Ahmad N, Filipe AJ, Thu PA, Zia-Ud-Din M, et al. Water Scarcity and Sustainability in an Emerging Economy: A Management Perspective for Future. *Sustainability*. 2020;13(1):1–10. Available from: https://dx.doi.org/10.3390/su13010144.
24) Antonello F, Baraldi P, Zio E, Serio L. A Novel Metric to Evaluate the Association Rules for Identification of Functional Dependencies in Complex Technical Infrastructures. *Environment Systems and Decisions*. 2022;42(3):436–449. Available from: https://dx.doi.org/10.1007/s10669-022-09857-z.
25) Salehpour HB, Javadi HHS, Asghari P, Abadi MESA. Improvement of Apriori Algorithm Using Parallelization Technique on Multi-CPU and GPU Topology. *Wireless Communications and Mobile Computing*. 2024;2024:1–14. Available from: https://dx.doi.org/10.1155/2024/7716976.