2024-09-17

Uncovering service gaps and patterns in smallholder dairy production systems: A data mining approach

Nyambo, Devotha

Science Direct

https://doi.org/10.1016/j.sciaf.2024.e02392

Scientific African 26 (2024) e02392

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Uncovering service gaps and patterns in smallholder dairy production systems: A data mining approach

Devotha G. Nyambo

Nelson Mandela African Institution of Science and Technology, P. O. Box 447, Tengeru, Arusha, Tanzania

ARTICLE INFO

Editor: DR B Gyampoh

Keywords:
Association rules
Frequent patterns
Smallholder farmers
Milk production
On-farm decisions

ABSTRACT

Traditional clustering algorithms have often been used to categorize farmers but tend to overlook the underlying reasons for these groupings. Typically, clusters are formed based on common metrics such as dispersal and centrality, which provide limited insights into the relationships among key attributes. This study introduces an innovative approach using pattern and association rules analysis to better understand the characteristics of dairy production clusters. Focusing on Tanzanian smallholder farmers, the research moves beyond identifying clusters to uncovering the hidden relationships within them. Through pattern analysis, the study logically examines the behavioral mechanisms that define these clusters, highlighting service gaps that, if addressed, could enhance smallholder dairy farmers' productivity. Frequent patterns with support ranging from 57 % to 93 % and confidence levels between 85 % and 100 % were identified, revealing critical challenges faced by these farmers. For instance, farmers using Artificial Insemination—typically younger or new entrants—face constraints related to farm size, land holdings, fodder production, lack of farmer groups, and insufficient formal training in dairy care. Meanwhile, seasoned farmers deal more with institutional barriers such as limited access to marketplaces, extension services, and distant water sources. The study highlights the diverse challenges faced by different farmer groups and provides strategic recommendations for improving dairy productivity. Enhancing access to formal training, improving fodder production, supporting the formation of farmer groups, and addressing institutional barriers are key actions that could help Tanzanian smallholder dairy farmers increase milk yield and overall productivity.

Introduction

More than 150 million farm households worldwide are supported by smallholder dairy farming, which is characterized by small herds of 1–2 milking cows [1]. Nearly half of all cattle produced in Africa's livestock farming industry is produced by smallholder farmers [2]. According to [1], the majority of farmers in developing nations operate small farms, where Tanzania is home to 24 million cattle, which accounts for 1.67 % of the world's total cow population. Livestock raising is the primary source of income for about 50 % of Tanzanians [3]. Despite having a significant impact on milk production and satisfying market demand, smallholder dairy farmers face challenges that lower productivity [4]. For the majority of farmers who are actively engaged in farming, smallholder dairy farming initiatives are a crucial source of daily sustenance [2].

The rising demand for milk and dairy products has prompted research projects on how to increase milk productivity for smallholder farmers [4]. To ensure that the necessary services that aid smallholder dairy farmers in maximizing their productivity are easily

E-mail address: devotha.nyambo@nm-aist.ac.tz.

https://doi.org/10.1016/j.sciaf.2024.e02392
Received 14 December 2022; Received in revised form 11 September 2024; Accepted 16 September 2024
Available online 17 September 2024

available, the proper framework and substructure have to be implemented [2]. Collaboration among smallholder dairy producers helps them develop new skills and cope with challenges more effectively [5]. Practically, heterogeneous groupings hamper service delivery, information sharing, and technology dissemination, especially for those wishing to maximize productivity and profitability, due to pertinent characteristics of farmers based on management practices [2]. It is necessary to identify homogenous groupings of dairy producers for more approachable intervention [6].

Farm typologies present a way of disaggregating individuals based on prominent attributes that define them. Characteristics derived from the farm typologies represent some of the challenges and supporting attributes attached to each production cluster. Through these characteristics, relevant interventions and policy support to the farmers can be defined. In Tanzania, the livestock industry is generally characterized by low production and unmet demands for meat and dairy products [7–9]. Existing data shows that the majority (86 %) of the smallholder dairy farmers in Tanzania live in rural with cattle herds dominated by traditional breeds [3].

Generally, the defining attributes of smallholder dairy farmers are a small number (3–5) of animals kept per farm, small pieces of land (2 ha or less), which are used for cash and food cropping [10,11] and low yields [12,8]. The reasons underlying low dairy production have been studied extensively and categorized as feeds [12], animal health, breed type, farm size, and distance to markets [9] among others. Knowledge of the nature of our dairy production systems has been identified through cluster-based characterization.

Characterization techniques through clustering, rely on common features using dispersal and centrality metrics but fail to capture the intricate relationships among the attributes under study. This paper introduces an innovative approach to unraveling the distinctive characteristics of dairy production clusters through pattern and association rules analysis. By delving into pattern analysis, we aim to provide a logical assessment of the behavioral mechanisms that underlie the commonalities observed in cluster analysis.

Traditional clustering algorithms have long been employed to understand and characterize various patterns in smallholder farming systems [13]. These algorithms typically group farms based on similarities in management practices and on-farm parameters, resulting in subgroups with high within-group similarity and low between-group similarity. However, despite their widespread use, these methods have significant limitations when it comes to providing comprehensive insights into the complex relationships and patterns within smallholder farming systems.

For instance, Herrero et al. [14] highlight that while clustering can identify groups with similar farm management practices, it falls short of uncovering the underlying features that influence group formation, farmer evolution patterns, and predictive modeling of future changes. This gap in knowledge hampers the ability to understand how various farm parameters and management practices are interlinked, which is crucial for devising effective intervention strategies and improving farm productivity [15].

Additionally, studies such as those by Sirsikar and Wankhede [16] have shown that traditional clustering methods often fail to account for the dynamic nature of farm management practices and their evolving impact on farm productivity. These methods typically provide a static snapshot of farm clusters without capturing the temporal changes and adaptations made by farmers in response to environmental and economic factors.

Moreover, work by Cai et al. [17] has demonstrated that clustering techniques are insufficient in revealing the socio-economic dimensions that influence decisions. These dimensions include factors like access to markets, credit availability, and extension services, which play a crucial role in shaping farm practices but are often overlooked in purely quantitative clustering approaches.

In contrast, association rules mining offers a promising complementary technique to traditional clustering. This method focuses on the relationships among farmers, common management practices, and on-farm parameters, revealing frequent patterns that can inform decision-making processes. Arumugam [18] demonstrates the feasibility of using association rules mining in precision agriculture, where it helps farmers predict the outcomes of their management decisions and enables policymakers to design farm-specific interventions.

A review by Oyewole & Thopil [19], underscored the role of clustering in various application domains with an assessment of how such techniques can be combined with other analytical techniques for robust outputs. This paper presents an innovative approach that employs pattern and association rules ;'/analysis to uncover the characteristics of dairy production clusters. By moving beyond the mere identification of clusters, this study aims to understand the hidden relationships and compositions within dairy production clusters, particularly in the context of Tanzanian smallholder farmers. Pattern analysis offers a logical assessment of the behavioral mechanisms underlying the commonalities identified by cluster analysis, thereby providing a deeper understanding of the service gaps that, if addressed, could significantly enhance the productivity of smallholder dairy farmers.

For the case of Tanzania dairy farmers, a recent cross-sectional survey indicates a mismatch and unreliability for some services such as the use of Artificial Insemination [20]. Through pattern and association analysis, this paper identifies service gaps for different categories of farmers to improve adoption of interventions by various stakeholders. By dis-aggregating the farming systems, our approach clearly indicates service gaps for six production clusters which is an effort to minimize the use of generalized solutions for challenges in our farming systems.

Material and methods

The presented study is quantitative with Machine Learning techniques used in data analysis and characterization of dairy farms. Unsupervised Learning techniques were used for both clustering and association rules mining objectives.

Study area and data preparation

The data used in this study was acquired from the PEARL[1] project, which conducted a baseline survey of smallholder dairy farmers in six regions of Tanzania in 2015–2016. The survey study sites were; Arusha, Kilimanjaro, Tanga, Iringa, Njombe, and Mbeya. A meta-analysis was done to identify commonly used attributes in clustering smallholder dairy farmers based on a review paper [21] the attributes could be grouped into five main categories; feeding systems, breeding, health, milk yield, and sales, and farmer groups. These attributes have long been used in studying dairy production systems [22].

Based on cluster analysis with the Self-Organizing Maps (SOM) algorithm (a.k.a. the Kohonen SOM) [23], a pre-clustered data set with six farm typologies/production clusters was used to analyze frequent patterns associated with the farmers within and across groups. Data was available from a total of 3317 smallholder dairy households. Before association rules mining, data were transformed as follows: four continuous variables (total land available for dairy activities, peak milk yield, amount of milk sales, and amount of reserved milk for home consumption) were scaled into categories by deviating individual observations from the mean with upper and lower bounds of the boxplot distribution used as the thresholds to obtain three categories, high, medium and low for each trait.

These attributes were also used to provide an overview of households' distribution in the dataset with a standard deviation (± SD) (Table 1). It is shown that the number of households with high values in all the four variables is smaller than number of households who are medium and low in the four variables. Table 1 details the distribution of households for the four variables.

Clusters were generated by using the Self-Organizing Maps (SOM) algorithm. In previous research [24] the Tanzania dairy production clusters were derived and validated by considering the algorithm's clustering consistency and total variance explained by the clusters on milk yield, sales, and choice of breeding method. One-way analysis of variance (ANOVA) was adopted as a means of comparison test across the clusters. To further understand the clustered dataset, Exploratory Factor Analysis (EFA) was adopted to show how the weight of all variables in the six clusters.

Production cluster one contained the highest number of farmers from three regions located in the southern highlands of Tanzania. Production cluster two consisted of farmers from Tanga and some from Iringa. The remaining clusters contained farmers from the north-east of Tanzania i.e., Arusha and Kilimanjaro (Table 2). The distribution of farmers in the production clusters across regions indicated the necessity of studying shared characteristics of farmers across different locations because farmers from the same location are not necessarily the same. Farm-level characteristics based on the attributes given in Tables 3 and 4 are further detailed in Appendix A.

Association rules mining

Mining for association rules proceeded with the Apriori algorithm based on the fact that: i) frequent items and patterns can be observed from the first parts of the data set and solutions are rare, so applying a depth-first search will require more time to generate all patterns, ii) The number of nodes in the data is finite even though the depth of the tree to be generated is infinite. Thus, depth-first search might fail to locate all children of nodes as it goes down, and iii) the primary interest was to find out frequent patterns and then find out how the patterns are related by visualizing them. Considering the existing literature [25,26] and the given assumptions, this research considered the use of the Apriori algorithm for association rules mining.

Data analysis was done in R software [27], Version R.4.1.0. Four measures of rules interestingness were: support, count, confidence, and lift. Support is a count of the number of times an item appears in the data set. The count is a number of observations in the data set supporting a particular association rule. Confidence is a measure of likelihood of occurrence of a rule; for example, considering a rule $[AUB] \rightarrow$ C, confidence measure indicates the likelihood of this association rule by taking the ratio of the support of $[[AUB]\ UC]$ to the support of $[AUB]$ as represented in Eq. (1). Lift is a measure of the deviation of the support of a whole rule from the support expected under features' independence, given the support of the antecedents and the consequent. Therefore, the higher the deviation the stronger the rule. From the example, $[AUB] \rightarrow C$, Lift is given by Eq. (2). Association rules were generated using the Arules package and visualized using the ArulesViz package (See figures in Appendix B), by using graph and grouped matrix visualization [28,29]. In the graph and grouped matrix visualization, strong rules were indicated by higher lift values (strong color intensity) and high support levels denoted by the size of bubbles.

Minimum support denotes the least number of times an item/pattern has appeared in the dataset, and the minimum confidence denotes the least likelihood of occurrence for the item on the Right-Hand Side (RHS) upon occurrence of the item in the Left-Hand Side (LHS). During analysis, the values were set to 0.1 and 0.5, respectively. These values were adjusted on each run to produce a manageable number of rules (maximum 60) for visualization visualization.

$$Confidence\ [AUB] \rightarrow C = \frac{Support\ [[AUB]UC]}{Support\ [AUB]} \tag{1}$$

Where; [AUB] is the antecedent and C is the consequent.

$$Lift = \frac{Support\ [[AUB]UC]}{Support[A].Support[B].\ Support\ [C]} \tag{2}$$

[1] Program for Emerging Agricultural Research Leaders

Table 1
Distribution of households into three categories used to transform continuous variables.

Variable	High	Medium	Low
Total land (acre) (3.9 ± 20.1)	204	987	2126
Peak milk (Liters) (12.7 ± 4.9)	556	2382	379
Milk sold (Liters) (10.5 ± 9.6)	308	1893	1116
Reserved milk (Liters) (1.6 ± 2.6)	298	1383	1624

Table 2
Production cluster densities and location of dairy households.

Production cluster	Density	Regions
1	1180	Iringa, Njombe, Mbeya
2	952	Tanga, Iringa
3	203	Arusha
4	295	Kilimanjaro
5	516	Kilimanjaro
6	171	Arusha

Table 3
Production cluster characteristics based on factor loadings.

Factor	Production cluster loadings					
	1	2	3	4	5	6
Vaccination frequency	2.038123	0.797619	1.258333	1.614583	1.738938	1.155039
Watering frequency	1.501466	1.869048	2.533333	1.75	2.146018	1.945736
Experience in dairy farming	1.283459	-0.40917	-1.27027	-0.36134	-0.6117	-0.89951
Distance to buyers	-1.04301	1.145007	0.23089	1.11271	-0.49803	1.428073
Total land size	-0.93538	0.62411	0.368615	0.60047	-0.56729	1.902952
Land for fodder	-0.91741	1.610701	-2.64E-13	1.56806	-4.58E-13	-3.51E-13
Distance to breeding service	0.371558	1.384379	-1.56509	1.397133	-0.42205	-1.27521
Grazing land	0.901289	-1.58885	7.36E-11	-1.54591	7.36E-11	7.36E-11
Years of schooling	0.872069	-1.43285	-1.13005	-1.38332	-0.05523	0.867026
Sale of bulk milk	-0.62271	-0.87143	2.347676	-0.90305	0.534935	0.353077
Visits by extension officers	0.210158	1.272907	-1.87464	1.2979	-0.54204	-0.37712
Number of milking cows	-0.57839	-0.23662	-1.4048	-0.23724	1.433476	1.133085
Herd size	-0.84404	0.584678	-0.81133	0.571074	-0.09112	2.086942
Liters of milk sold	-0.59993	-0.48079	0.790488	-0.50688	1.753072	-0.62056
Distance to market	0.824873	0.493123	0.147668	0.500794	-1.8976	-0.42586
Total crop sale	0.431172	1.252638	-1.26026	1.266918	-0.28039	-1.69281

Table 4
Means for all farm-level variables across farm typologies.

Factor	Production cluster means for the analysis variables					
	1	2	3	4	5	6
Vaccination frequency	1.56	1.37	2.1	1.99	2.15	2.08
Watering frequency	2.33	2.2	1.8	1.42	1.67	1.63
Experience in dairy farming	10.82	11.73	16.87	19.06	20.03	20.37
Distance to buyers	2.55	3.13	1.22	1.59	1.57	0.95
Total land size	4.26	5.18	1.86	2.42	2.98	2.23
Land for fodder	0.93	0.91	0.64	0.59	0.65	0.72
Distance to breeding service	1.65	1.57	0.89	1.23	2.46	3.32
Grazing land	0.02	0.11	0	0	0	0
Years of schooling	8.41	8.64	8.09	8.22	9.48	9.59
Sale of bulk milk	15.72	145.1	389.01	278.3	515.1	429.4
Visits by extension officers	7.57	7.08	4.86	5.32	9.85	9.89
Number of milking cows	2.3	2.27	2.24	2.13	2.17	2.23
Herd size	5.03	5.29	4.74	4.56	4.57	5.1
Liters of milk sold	12.57	10.38	8.2	6.00	8.13	12.49
Distance to market	2.16	2.68	3.01	2.99	2.86	3.28
Total crop sale	4.34	4.44	4.12	4.78	5.77	5.02

Where; [AUB] is the antecedent and C is the consequent. An association rule is stronger if its lift value is high, meaning that; the frequent items are much stronger together than when they are apart. Generally, good lift values must be greater or equal to 1.

Results

Association rules mining on production clusters

Factor loadings were used to study the characteristics of the farmers within production clusters by considering high and low loadings as positive and negative attributes, respectively. The naming of the production clusters was according to the attributes in both categories (Tables 3 and 4). The values in Table 3 are resulting from data scaling (mean of 0 and 1 standard deviation) as used during cluster analysis. For actual values, Table 4 shows the means for all cluster-level variables. One-way analysis of variance (ANOVA) was adopted as a means comparison test. For all farm-level variables, results indicated that there were significant differences across production clusters at $p < 0.0001$.

Production cluster one contained the majority of the households (36 %) consisting of "*non-commercial dairy production households*", termed *non-commercial* since milk or crop sale characteristics did not come up for this production cluster. Farmers in this cluster were characterized by high vaccination and watering frequency, many years of experience in dairy farming, small distances to milk buyers, small land sizes, and small lands under fodder production. Production cluster two which had 29 % of the sampled households consisted of "*Semi-intensive commercially oriented medium production households with limited access to improved breeding services*", and were characterized by high frequencies of cattle watering, large areas under fodder production, long distances to improved breeding service providers, low grazing lands, few years of formal education, and low sales of bulk milk.

Production cluster three contained 6 % of the households and consisted of the "*commercially oriented and self-reliant high production households*". Production cluster three was characterized by high frequencies of cattle watering per day and vaccination rounds per year, high sales of bulk milk, low frequency of visits by extension officers, small distances to breeding service providers, and a small number of milking cows. Production cluster four consisted of 8.9 % of the households categorized as "*Semi-intensive commercially oriented medium production households*". Production cluster four was characterized by high frequencies of watering and vaccination, large areas under fodder production, small grazing lands, a few years of schooling, and low sales of bulk milk. Clusters 2 and 4 were highly similar but their access to improved breeding services was found to be similar to production cluster 2 (long distances to service providers).

Fifteen percent (15 %) of the households formed production cluster five which was categorized as "*Commercially oriented high production entrant households*". Key characteristics of the farmers in production cluster five were: high frequencies of watering, vaccination, and milk sales, small distances to market, low experience in dairy farming, and small land sizes. Production cluster six comprised 5.1 % of the households that were categorized as "*Commercially oriented mixed crop-dairy low production entrant households*". Production cluster six was characterized by: long distances to water sources, large herd sizes, low crop sales, short distances to breeding service providers, and low experience in dairy farming. Characteristics of the production clusters are further described in sub-sections below.

Production cluster one: Non-commercial dairy production households

Based on the clustering results, this cluster consisted of farmers who have been practicing dairy farming for a long time (loading high on experience in dairy farming) with no trace of their commercialization focus (Table 3). Investigating the production cluster by using association rules revealed that, farmers in this group practice traditional dairy farming based on stall feeding, few numbers of milking animals (1–3), prefer the bull method for breeding and they do not have land for fodder production. These experienced dairy farmers have not attended any formal training and received very few visits for extension support (1 – 9 visits per year). Strong rules concentrated on the stall-feeding system appearing on both antecedent and consequent positions.

Characteristics of the smallholder dairy farmers in the production cluster were generated with good quality measures in terms of support, confidence, and lift. High lift values (>1) were observed in rules involving several milking cows 1 - 3, lack of training, stall feeding, preference and use of bull breeding, and lack of areas for fodder production. Clustering results (Table 3) indicated low loading on the distance to milk buyers. With these characteristics, the commercial orientation in this cluster was assumed to be low. These results indicate with high confidence (> 85 %) that more than 63 % of the farmers practice subsistence dairy farming. Table 5 summarizes quality measures for the generated rules.

Table 5
Summary of quality measures for rules in production cluster one.

	Support	Confidence	Lift	Count
Minimum	0.6364	0.8506	0.9798	751
1st Quartile	0.6415	0.9126	1.0047	757
Median	0.6508	0.9405	1.0142	768
Mean	0.6538	0.9377	1.0816	771.5
3rd Quartile	0.6619	0.9636	1.0813	781
Maximum	0.6729	0.9987	1.2957	794

Production cluster two: semi-intensive commercially oriented medium production households with limited access to breeding services

Clustering results characterized production cluster two as semi-intensive commercial farming with medium production. Farmers in this group plant fodder but also utilize small grazing lands. Association rules indicated their preference for bull breeding, which is associated with long-distance to improved breeding service providers. High lift values (>1.2) were observed in rules involving: stall feeding system, preference and use of bull breeding, average milk production, lack of farm laborers, no use of purchased fodder, short distance to water source (< 1 km), and 1–3 milking cows. Clustering results indicated low amounts for bulk milk sales, the presence of areas for fodder production, and small grazing lands (Table 3).

A graph of rules further indicated that lack of farm laborers and absence of purchased fodder are associated with stall feeding and preference for bull breeding. Given that these farmers have average milk production and low sales of bulk milk with a mixed feeding system (dominated by stall feeding), their production system is semi-intensive and commercially oriented. Therefore, it can be demonstrated at > 88 % confidence level that, 57 % of farmers in production cluster two have a medium commercial orientation. Further details on the quality measures are presented in Table 6.

Production cluster three: commercially oriented high dairy production households

From the clustering results, this production cluster had the high dairy production households with high loadings on the sale of bulk milk, vaccination frequencies per year, and low loadings on the frequency of extension visits and distance to improved breeding services. High lift values (>1) were observed for association rules covering: stall feeding, preference and use of bull breeding, two vaccination rounds per year, lack of farm labor, and no membership in farm groups. Farmers in production cluster three are geographically located in the same region as the Tanzania National Artificial Insemination Center (NAIC).

A grouped matrix for the rules in production cluster three shows that only the feeding system, breeding method, and the number of milking cows have appeared as antecedents and consequents of rules, indicating that they influence and are influenced by other attributes. Attributes such as lack of employees, lack of farmer groups, vaccination frequency, and shorter distance to water sources (<1 km) were concentrated on the left-hand side as antecedents of the rules. Given that these farmers have high amounts of milk sold in bulk, vaccinate their cattle at least twice a year, receive few or no visits from extension officers, and do not belong to farmer groups their production system is considered highly commercial and self-reliant. As such, it can be stated at a 97 % confidence level that, 87 % of farmers in production cluster three practice commercial dairy farming. A summary of the quality measures is shown in Table 7.

Production cluster four: semi-intensive low commercial oriented average production households

The analysis characterized production cluster four with large areas under fodder production and some small areas available for grazing. This production cluster was identified with low commercial orientation based on a low loading on the sale of bulk milk. Association rules revealed high lift values (>1) in rules covering: stall feeding, preference and use of bull breeding, lack of farm labor, and ownership of 1–3 milking cows. Milk production and sales coefficients did not appear as frequent items for this group. However, 64 % of the farmers had average milk yield, 68 % had below-average milk sales and 59 % had below-average milk reserved for home consumption.

From the unique attributes given by the clustering results, a lowly educated farmer practicing a semi-intensive feeding system with average yield and low sales of bulk milk could be identified as an average producer with a low commercial orientation. Lack of training is assumed to be a complementing factor to the low commercial orientation for production cluster four. Preference and use of the bull breeding method appeared dominant for production cluster four which had farmers from Kilimanjaro region.

Quality measures for production cluster four rules indicate that minimum confidence and support were 0.90 and 0.84, respectively (Table 8).

Production cluster five: commercially oriented high production entrant households

Clustering results revealed this production cluster with high production, low experience, and with limited land sizes. High lift values (>1) were revealed in rules covering: stall feeding system, lack of farmer groups, distance to market 1–5 km, lack of formal training on dairy care, and reason for choice of stall feeding being insufficient land. Frequent pattern analysis revealed that the intensification system appeared to be a result of insufficient land available to the farmers. Although patterns show that, the number of milking animals in this production cluster is one to three, the commercial orientation of entrant dairy farmers is demonstrated by their high loadings on liters of milk sold and the short distance to markets (1 – 5 km) as given by the cluster analysis. Being close to formal markets, the commercial orientation for production cluster five appears to be on formal traders rather than neighbors who buy retail.

Lack of membership in farmer groups, lack of formal training on dairy care, and intensive feeding appeared as strong patterns in

Table 6
Summary of quality measures for rules in production cluster two.

	Support	Confidence	Lift	Count
Minimum	0.5735	0.8782	0.9935	546
1st Quartile	0.5851	0.9283	1.0332	557
Median	0.6024	0.9728	1.0691	573
Mean	0.6101	0.9609	1.1257	580.8
3rd Quartile	0.6229	0.993	1.2665	593
Maximum	0.6828	1	1.2733	650

Table 7
Summary of quality measures for rules in production cluster three.

	Support	Confidence	Lift	Count
Minimum	0.8719	0.9727	0.998	177
1st Quartile	0.8818	0.984	1.009	179
Median	0.8867	0.9944	1.02	180
Mean	0.8918	0.9923	1.033	181
3rd Quartile	0.9015	1	1.085	183
Maximum	0.931	1	1.086	189

Table 8
Summary of quality measures for rules in production cluster four.

	Support	Confidence	Lift	Count
Minimum	0.8441	0.9088	0.9932	249
1st Quartile	0.8441	0.934	1.0043	249
Median	0.8508	0.9862	1.0098	251
Mean	0.8544	0.965	1.0296	252
3rd Quartile	0.8576	0.9961	1.0889	253
Maximum	0.9051	1	1.0967	267

both RHS and LHS. It can be demonstrated at 88 % confidence that at least 69 % of farmers who have high amounts of milk sales in formal markets, practice intensive feeding in small land sizes, have no formal training in dairy care, and do not belong to farmer groups, are commercial oriented high production entrant farmers. Table 9 details further quality assessment of the rules.

Production cluster six: commercially oriented mixed crop-dairy low production entrant households

Production cluster six consisted of commercial entrants who loaded high on several cattle owned, walked long distances to water sources, were located near service providers for improved breeding (Artificial Insemination), and had few years of experience in dairy farming. Lift values from association rules were high (>1) in rules covering: land below average, number of milking cows being 1–3, preference to Artificial Insemination (AI), stall feeding system, lack of formal training in dairy care, and lack of farmer groups. Loading low on the distance to breeding service providers, association rules indicated that, this group of farmers prefer and use Artificial Insemination for breeding. Stall feeding is the default system used during rainy and dry seasons, associated with small pieces of land which are assumed to be used for crop farming.

Although cluster loadings indicated long distances to water sources, association rules revealed that, the majority of farmers in cluster six walk less than a kilometer to water sources.

Thirty rules from production cluster six were generated at a 100 % confidence level. The remaining rules (22) ranged between 95 % - 98 % confidence level for at least 83 % of the farmers. These quality measures imply that the rules can be used to generalize the characteristics of production cluster six. It could be demonstrated that at least 83 % of farmers in cluster six are commercial entrants in the mixed crop-dairy farming system. Table 10 details more on the quality measures for production cluster six rules.

Summary of findings

Farm characteristics

Dairy production cluster one was characterized by two times vaccination rounds per year, two times cattle watering per day, large land sizes, available areas for fodder production, stall feeding system, 1–3 milking cows, and bull breeding. Production cluster two was characterized by high frequencies of cattle watering per day, available areas for fodder production, small grazing land, use of bull breeding, no employed laborers, and no use of purchased fodder. Production cluster three was characterized by high frequencies of watering per day and cattle vaccination per year, 1–3 milking cows, bull breeding, no employed laborers, and two times cattle vaccination per year.

Production cluster four was characterized by high frequencies of watering per day, two vaccination rounds per year, available areas

Table 9
Summary of quality measures for rules in production cluster five.

	Support	Confidence	Lift	Count
Minimum	0.6919	0.883	0.9898	357
1st Quartile	0.7422	0.9496	1.0042	383
Median	0.7539	0.9719	1.0078	389
Mean	0.7591	0.9645	1.011	391.7
3rd Quartile	0.7868	0.9858	1.0128	406
Maximum	0.8256	1	1.0479	426

Table 10
Summary of quality measures for rules in production cluster six.

	Support	Confidence	Lift	Count
Minimum	0.8304	0.9586	0.9978	142
1st Quartile	0.8538	0.9655	1	146
Median	0.9123	1	1	156
Mean	0.8972	0.9879	1.0008	153.4
3rd Quartile	0.924	1	1	158
Maximum	0.9474	1	1.0075	162

for fodder production, small grazing lands, use of bull breeding, no employed labors, and 1–3 milking cows and stall-feeding system. Production cluster five was characterized by high frequencies of watering per day and two vaccination rounds per year, small land sizes, 1–3 milking cows, and a stall-feeding system due to insufficient land sizes for dairy production. Cluster six was characterized by large herd sizes, high frequencies of watering per day, small land sizes, milking cows 1–3, use of artificial insemination for breeding, and stall-feeding system.

Farmer characteristics

Dairy production cluster one was characterized by many years of experience in dairy farming and no training in dairy care. Production cluster two was characterized by a few years of formal schooling. Production cluster three was characterized by non-membership in farmer groups. Production cluster four was characterized by a few years of formal schooling and a lack of formal training in dairy care. Production cluster five was characterized by a few years of experience in dairy farming, no formal training in dairy care, and non-membership in farmer groups. Production cluster six was characterized by a few years of experience in dairy farming, no formal training, and non-membership in farmer groups.

Farm income

The results did not indicate an association between income parameters and production cluster one. Production cluster two was characterized by low sales of bulk milk, average milk yield, and regular sales. Production cluster three was characterized by high amount of milk sold in bulk implying their high commercial orientation. Production cluster four was characterized by low sales of bulk milk implying regular sales. Production cluster five was characterized by high amounts of regular milk sales. Production cluster six was characterized by low crop sales implying their crop-dairy low production.

Institutional settings

Dairy production cluster one was characterized by small distances to milk buyers and water sources and received 1–9 visits from extension officers per year. Production cluster two was characterized by long distances to improved breeding service providers and received 1–9 visits from extension officers per year. Production cluster three was characterized by low frequencies of visits from extension officers per year and short distances to improved breeding service providers. Production cluster four was characterized by low distances to water sources. Production cluster five was characterized by a short distance to formal markets (1–5 km). Production cluster six was characterized by long distances to water sources and short distances to improved breeding service providers.

Discussion

Pattern analysis and association rules mining on clustered datasets

This paper presents the use of pattern and association rules analysis as an approach to unveil the characteristics of dairy farm typologies as a complement to cluster analysis. Cluster analysis groups farmers in big blocks based on common features using dispersal and centrality metrics. However, Pattern analysis provides a logical assessment of the behavioral mechanisms underlying the commonality exhibited by the cluster analysis. Since evolvement and technological sophistication are a factor of behavior, the extension from cluster analysis methods are to find logical patterns that drive the farm typologies. Additionally, usability of the clustering results on different case studies can be grounded on the confidence of assigning individuals to the farm typologies without re-running a production cluster analysis. That confidence is given by the association rules mining through determining the individual's belongness into a production cluster when several pre-identified attributes are seen together.

By applying the proposed techniques, this paper presents features of six dairy farming systems in Tanzania which could not be identified through cluster analysis alone.

Features of Tanzania's dairy farming

The cluster solution indicated that farmers had similar features across different production systems for attributes with high loadings, while differences among the production systems were observed in attributes with low loadings. Similar features for Tanzania dairy farmers were adherence to best health management practices and cattle watering. Summary statistics revealed that at least 87 % of 3317 households indicated that they deworm and vaccinate their cattle. From that population, at least 37.42 % deworm their cattle

thrice per year ($p < 0.0001$) and at least 82.13 % vaccinate their animals once per year ($p < 0.0001$). Low milk production and sells could be marked as a mutual challenge for Tanzania farmers except for production cluster three. Association rules indicate stall feeding is dominant (at 90 % confidence) and is used even in the absence of localized fodder production. These finding concedes with other research on the dominant feeding system in Tanzania [30]. Identification of similarities on feeding systems could not be revealed through cluster analysis, underscoring the need to combine techniques for efficient characterization [19]. By complementing clustering results with patterns and association analysis, this paper discusses prominent service gaps that should be addressed for farmer in the six analyzed typologies.

Dairy farms typologies in Tanzania and service gaps

The structure of the dairy production clusters considered the farm typologies attributes and the associations among the production cluster variables. Some characteristics support the development of the farm enterprises and others limit improvements in the same. A recent study by Kashoma & Ngou [20] done in Tanzania elaborated on gaps in Artificial Insemination services which are hindered by severe irregularities of services, unreliability of liquid nitrogen supplies, and unreliable transport. In agreement with our study, these examples of service gaps demonstrate areas of improvement for dairy production. Altogether, such areas of improvement or service gaps are termed as the farms' evolvement determinants. The determinants have been grouped into four categories following a decision framework for smallholder dairy farmers [31]. The four categories are labeled as farm characteristics, farmer characteristics, farm income, and institutional settings.

For production cluster one, the 1–9 extension support visit was associated with the preference of using bull breeding at 94 % confidence. The use of bull breeding was associated with a lack of formal training in dairy care at 89 % confidence. Among the service gaps facing these farmers from Iringa, Njombe, and Mbeya are training in dairy care and handling, adequate extension service, and feed production. Formal training will help the farmers to adopt a commercial orientation in dairy farming. Such training can also foster changes in their choices of breeding methods. This can be achieved by providing adequate extension services to educate the farmers on best practices. Due to the shortage of extension officers, electronic platforms can be utilized to deliver knowledge as proposed in poultry farming [32,33]. Best practices can feature in the importance of fodder production which is a gap in this typology. By focusing on the identified gaps, three pillars of the decision framework (farm, farmer characteristics, and institutional settings) that limit dairy production in production cluster one will be solved.

Farmers from Tanga and Iringa forming production cluster two had a commercial orientation which could be observed through the feeding system. Adequate extension service is a limiting factor for growth in the cluster. High loadings in distance to improved breeding services indicated another gap on access to breeding service providers although farmers prefer and use Artificial Insemination (AI). Being confident in their feeding system, production cluster two highly depends on institutional factors (availability and reliability of AI breeding services and extension) for its development. The result agrees with [31] where AI usage was observed in farmers residing in the northern part of Tanzania.

Farmers in production cluster three limit their herd sizes to lower labor costs (number of milking animals associated with lack of labor at 99 % confidence). There appear to be common practices in dairy for these farmers as the use of bull breeding could be associated with two times vaccination at 99 % confidence. An institutional service gap is observed in breeding services where the production cluster loaded low in distance to improved breeding service and yet they use bull breeding. The National Artificial Insemination Center (NAIC) is located in Arusha, implying that production cluster three could be the frontiers in transforming breeding practices. Although education level, experience, or training could not be associated with the high yield for farmers within the cluster, farmer-based initiatives for high milk yield and public institutional failure were identified. To improve the experience of farmers in this production cluster, capacity building could be tailored in forming farmer groups where they can advocate for issues related to institutional setup.

Production cluster four was mostly similar to cluster two in terms of production cluster formation. The feeding system appears to be sufficient as in cluster two. Factors accounting for low sales of milk could be banked on lack of improved breeding and extension services. Kilimanjaro region is located at least 60 Km from the NAIC, therefore, the lack of proper breeding services highlights a service gap. Training in dairy care and handling was also underlined as key areas for improvement. Production cluster five from Kilimanjaro highlights the differences between experienced subsistence farmers and entrants with a commercial orientation. Both categories being untrained and lacking farmer groups, the production difference made by the commercial entrants reveals an un-winded potential of dairy farming. Although the use of AI could not be observed in frequent item sets, further investigation of the individual records revealed that, 72 % of the 516 members use AI for breeding. Capacity building for these farmers who are presumably youth/young adults should be channeled through farmer groups and training to improve their productivity.

As in production cluster five, farmers in cluster six are presumably youths or young adults who engage in dairy with a commercial orientation. Although cluster six loaded high on herd size, frequent patterns revealed ownership of 1–3 milking cows just as in other production clusters. The location of this production cluster (Arusha) highlights awareness of improved breeding services and the use of AI as the preferred breeding method. Limited access to cattle water could be observed from the high loading distance to water sources and low milk yields of these farmers who have adopted a stall-feeding system. These entrants are also limited in land for crop and fodder production. The low crop sales suggest a feed shortage for the animals. The strength of production cluster six is in the use of improved breeding services but highly constrained on farm and farmer characteristics.

For Kilimanjaro and Arusha regions where two production clusters were found for each, further studies need to be done to uncover the hidden trends i.e., the use of bull breeding by experienced farmers and the use of Artificial Insemination (AI) by entrants with a commercial focus. Although direct responses to these facts are beyond the scope of this paper results do agree with the risk assessment

done by Twine [34], where youth in the formal dairy value chain were observed to have the greatest risks. Here the risks could be under institutional settings (AI service provision, extension support) and absence/non-membership in farmer groups.

Conclusion

This paper investigated the characteristics of smallholder dairy farms by using patterns analysis and association rules mining approach. A clustered dataset of smallholder dairy producers was used. The clusters were referred to as dairy production clusters or dairy farm typologies in Tanzania's smallholder dairy system. By taking advantage of the clustered dataset, the paper demonstrates how the differences among the clusters can be identified through the use of pattern analysis and association rules mining, revealing features that could be identified through a cluster solution alone. With the presented approach, unique features of six production clusters in Tanzania are described and discussed. Observed differences among the production clusters can assist stakeholders in the identification of service gaps and the design of appropriate intervention strategies.

The analysis has shown that even though some dairy farmers can be in the same geographical location, they experience different challenges in dairy farming which calls for the design of specialized interventions to improve on-farm productivity. Service gaps have been identified for each farm typology/cluster; for example, farmers who prefer to use Artificial Insemination (new in dairy farming and presumed to be youth) are highly constrained by farm and farmer characteristics i.e., herd size, land holding, fodder production, lack of farmer groups and lack of formal training in dairy care. These highlight key areas that can be improved to empower youth in dairy farming. On the other hand, in production clusters where the experience on dairy farming was high and the traditional dairy keeping was for household subsistence, constraints were observed in institutional settings.

Future work is proposed on transforming the identified patterns and associations into a recommendation model that can be deployed to assist farmers in decision-making. Such a model will be implemented as conditional logic evaluating the inputs of the user. Developed association rules will be coded at the back end with incremental updates based on the accumulation of new inputs from users. This can be realized as a mobile-based solution.

Credit Author Statement

Dr. Devotha Godfrey Nyambo developed the concept from her PhD work, improved the analysis and the manuscript writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The following agencies are acknowledged: (i) Scholarship funders, International Development Research Centre (IDRC) and Swedish International Development Cooperation Agency (SIDA); (ii) Scholarship program, Artificial Intelligence for Development (AI4D) Africa; (iii) Scholarship fund manager, Africa Center for Technology Studies (ACTS).

Appendix A. Clusters, proportion in the pre-clustered dataset (n=3317), attributes and name of production clusters

Cluster	Proportion	Attributes with high loadings	Attributes with low loadings	Name of production cluster
1	36 %	Vaccination frequency, frequency of watering, experience in dairy farming	Distance to buyers, total land size, area under fodder production	Non-commercial dairy production households
2	29 %	Frequency of watering, area under fodder production, distance to breeding service providers	Area under grazing, years of schooling, sell of bulk milk	Small stock semi-intensive commercial dairy households with limited access to breeding services
3	6 %	Frequency of watering, sell of bulk milk, vaccination frequency	Frequency of visits by extension officers, distance to breeding service providers, number of milking cows	commercially oriented and self-reliant high dairy production households
4	8.9 %	Frequency of watering, vaccination frequency, area under fodder production	Area under grazing, years of schooling, sell of bulk milk	Semi-intensive commercially oriented medium production households
5	15 %	Frequency of watering, litres of milk sold, vaccination frequency	Distance to markets, experience in dairy farming, total land size	Commercially oriented high production entrant households
6	5.1 %	Distance to water sources, total cattle owned, frequency of watering	Total crop sales, distance to breeding service providers, experience in dairy farming	Commercially oriented mixed crop-dairy low production entrant households

References

[1] A. Chawala, G. Banos, A. Peters, M. Chagunda, Farmer-preferred traits in smallholder dairy farming systems in Tanzania, Trop. Anim. Health Prod. 51 (6) (2019) 1337–1344.
[2] F. Mavura, S.M. Pandhare, E. Mkoba, D.G. Nyambo, Rule-based engine for automatic allocation of smallholder dairy producers in preidentified production farm typologies, Scientific World J. (2022) 2022.
[3] Tanzania. (2016). *Tanzania livestock modernization initiative*.
[4] D. Nyambo, Doctoral Dissertation, NM-AIST, 2020, https://doi.org/10.58694/20.500.12479/895.
[5] Mudiwa, B. (2017). The influence of groups and group leadership on smallholder beef enterprise performance: a case of chipinge district in Zimbabwe.
[6] D.G. Nyambo, E.T. Luhanga, Z.Q. Yonah, A review of characterization approaches for smallholder farmers: towards predictive farm typologies, Sci. World J. 2019 (2019).
[7] F.D.N. Mujibi, J. Rao, M. Agaba, D. Nyambo, Performance evaluation of highly admixed tanzanian smallholder dairy cattle using SNP derived kinship matrix, Front. Genet. 10 (April) (2019) 375, https://doi.org/10.3389/fgene.2019.00375.
[8] Nzogela, B., Mwendia, S.W., Mwilawa, A.J., Kizima, J., & Bwire, J. (2022). *Farmers' perceptions on different forage types in southern highlands of Tanzania.* https://uknowledge.uky.edu/igc.
[9] E.S. Swai, E.D. Karimuribo, Smallholder dairy farming in Tanzania Current profiles and prospects for development, Outlook Agricult. 40 (1) (2011) 21–27, https://doi.org/10.5367/oa.2011.0034.
[10] E.S. Swai, P. Mollel, A. Malima, Some factors associated with poor reproductive performance in smallholder dairy cows: the case of Hai and Meru districts, northern Tanzania, Livestock Res. Rural Develop. 6 (26) (2014).
[11] S.K. Lowder, J. Skoet, T. Raney, The number, size, and distribution of farms, smallholder farms, and family farms worldwide, World Dev. 87 (2016) 16–29, https://doi.org/10.1016/j.worlddev.2015.10.041.
[12] D. Maleko, G. Msalya, A. Mwilawa, L. Pasape, Smallholder dairy cattle feeding technologies and practices in Tanzania: failures, successes, challenges and prospects for sustainability Smallholder dairy cattle feeding technologies and practices in Tanzania, Int. J. Agric. Sustain. 0 (0) (2018) 1–13, https://doi.org/10.1080/14735903.2018.1440474.
[13] P.K. Tapsoba, A.K. Aoudji, M. Kestemont, M.K. Konkobo, E.G. Achigan-Dako, Clustering smallholders' farmers to highlight and address their agroecological transition potential in Benin and Burkina Faso, Curr. Res. Environ. Sustain. 5 (2022) 100220, https://doi.org/10.1016/j.crsust.2023.100220.
[14] M. Herrero, P.K. Thornton, A. Bernués, I. Baltenweck, J. Vervoort, J. van de Steeg, S. Makokha, M.T. van Wijk, S. Karanja, M.C. Rufino, S.J. Staal, Exploring future changes in smallholder farming systems by linking socio-economic scenarios with regional and household models, Global Environ. Change 24 (1) (2014) 165–182, https://doi.org/10.1016/j.gloenvcha.2013.12.008.
[15] J. Mössinger, C. Troost, T. Berger, Bridging the gap between models and users: A lightweight mobile interface for optimized farming decisions in interactive modeling sessions, Agricult. Syst. 195 (2021) 103315, https://doi.org/10.1016/j.agsy.2021.103315.
[16] S. Sirsikar, K. Wankhede, Comparison of clustering algorithms to design new clustering approach, Proc. Comp. Sci. 49 (2015) 147–154.
[17] Q. Cai, M. Gong, L. Ma, S. Ruan, F. Yuan, L. Jiao, Greedy discrete particle swarm optimization for large-scale social network clustering, Inform. Sci. 316 (2015) 503–516.
[18] A. Arumugam, A predictive modeling approach for improving paddy crop productivity using, Turkish J. Electric. Eng. Comput. Sci. 25 (2017) 4777–4787, https://doi.org/10.3906/elk-1612-361.
[19] G.J. Oyewole, G.A. Thopil, Data clustering: application and trends, Artif. Intell. Rev. 56 (2023) 6439–6475, https://doi.org/10.1007/s10462-022-10325-y.
[20] I. Kashoma, A. Ngou, Insight into the adoption and success of artificial insemination services in smallholder dairy farming systems: a cross-sectional study, Tanzania Veterinary J. 38 (1) (2023).
[21] D.G. Nyambo, E.T. Luhanga, Z.O. Yonah, A review of characterization approaches for smallholder farmers : towards predictive farm typologies, Sci. World J. 2019 (2019) 9.
[22] M. Ngigi, The case of smallholder dairying in Eastern Africa, Environ. Prod. Technol. Division, EPT Discussion Paper 131 (February) (2005).
[23] T. Kohonen, The self-organizing map, Neurocomputing. 21 (1) (1988) 1–6.
[24] D. Nyambo, E. Luhanga, Z. Yonah, F. Mujibi, Application of multiple unsupervised models to validate farm typologies robustness in characterizing smallholder dairy farmers, Sci. World J. 12 (2019), https://doi.org/10.1155/2019/1020521. *2019*.
[25] D. Hunyadi, Performance comparison of Apriori and FP-Growth algorithms in generating association rules, in: Proceedings of the European Computing Conference, 2011, pp. 376–381.
[26] J. Heaton, Comparing dataset characteristics that favor the Apriori, Eclat, or FP-Growth frequent itemset mining algorithms, in: *Conference Proceedings - IEEE SOUTHEASTCON, 2016-July*, 2016, https://doi.org/10.1109/SECON.2016.7506659.
[27] Kabacoff, R.I. (2011). *R IN ACTION: Data analysis and graphics with R*.
[28] Hahsler, M., & Chelluboina, S. (2011). *Visualizing association rules: introduction to the R-extension package arulesViz*. R Project Module. http://www.comp.nus.edu.sg/~zhanghao/project/visualization/[2010]arulesViz.pdf.
[29] M. Hahsler, R. Karpienko, Visualizing association rules in hierarchical groups, J. Bus. Econ. 87 (3) (2017) 317–335, https://doi.org/10.1007/s11573-016-0822-8.
[30] E.D. Karimuribo, P.L. Gallet, N.H. Ng'umbi, M.K. Matiko, L.B. Massawe, D.G. Mpanduji, E.K. Batamuzi, Status and factors affecting milk quality along the milk value chain: a case of Kilosa district, Tanzania, Livestock Res. Rural Dev. 27 (3) (2015).
[31] G. Mwanga, F.D.N. Mujibi, Z.O. Yonah, M.G.G. Chagunda, Multi-country investigation of factors influencing breeding decisions by smallholder dairy farmers in sub-Saharan Africa, Trop. Anim. Health Prod. 2019 (51) (2019) 395–409.
[32] G. Msoffe, A. Chengula, M. Kipanyula, M.R.S. Mlozi, A.C. Sanga, Poultry farmers' information needs and extension advices in Kilosa, Tanzania: evidence from mobile-based extension, advisory and learning system (MEALS), Library Philosophy Practice (e-Journal) (2018).
[33] C. Sanga, e-Agriculture Promising Practice UshuariKilimo Information System Web and Mobile Phones for Extension Services in Tanzania, Sokoine University of Agriculture, 2018.
[34] E.E. Twine, A. Omore, J. Githinji, Uncertainty in milk production by smallholders in Tanzania and its implications for investment, Int. Food Agribusiness Manag. Rev. Vol. 21 (1) (2018) 53–72, https://doi.org/10.22434/IFAMR2017.0028.