Computers and Electronics in Agriculture 211 (2023) 108002

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Multi-algorithm clustering analysis for characterizing cow productivity on automatic milking systems over lactation periods

Karina Brotto Rebuli [a], Laura Ozella [a,*], Leonardo Vanneschi [b], Mario Giacobini [a]

[a] Department of Veterinary Sciences, University of Torino, Grugliasco, TO, Italy
[b] NOVA Information Management School, Universidade Nova de Lisboa, Lisbon, Portugal

ARTICLE INFO

Keywords:
Multi-algorithm clustering
Cluster algorithms merging index
Automatic Milking System
Future lactation period
Milk production

ABSTRACT

The introduction of Automated Milking Systems (AMSs), or milking robots, represented a significant advancement in dairy farming techniques. AMSs enable real-time monitoring of udder health and milk quality during each milking episode, which provides a wealth of data that can be utilized to optimize herd management practices. ML algorithms are well-suited for handling large and multi-dimensional datasets, making them a valuable tool for analyzing the vast amount of data generated by AMSs. This study introduces a novel approach to characterize the milk productivity of Holstein Friesians cows milked by AMSs during individual lactation periods and evaluate their stability over time. Four unsupervised ML clustering algorithms were employed to cluster the cows within each lactation period, and a merging index was proposed to combine the clustering results. The dairy cows were grouped into clusters based on their productivity, and the stability of these Productivity Groups (PGs) over time was analyzed. The PGs were found to be weakly stable over time, indicating that selecting cows for insemination based solely on their present or past lactation productivity may not be the most effective strategy. In addition, the results revealed that the High Productivity Group exhibited lower levels of protein, fat, and lactose content in the milk. The proposed methodology was demonstrated using data from one farm with dairy cows that exclusively uses the AMS, however, it can be applied to any context and dataset in which a multi-algorithm clustering analysis is suitable, including data from conventional milking parlors. Understanding milk productivity and its factors in future lactation periods is essential for effective herd management. A comprehensive long-term analysis is of significant importance for the zootechnical sector as it could assists farmers in selecting cows for insemination and making decisions on which ones to retain for future lactation periods.

1. Introduction

The introduction of Automated Milking Systems (AMSs), or milking robots, in the early 1990s represented one of the major headways in dairy farming techniques. The main advantages of AMS rely on its potential to decrease labor, increase milk production and animal welfare, and improve herd management (Lyons et al. 2022). During each milking session, automatic sensors enable real-time monitoring of the udder's health and milk quality by providing detailed information about each cow. Such a level of information was not easily obtainable with previous conventional systems (Jacobs & Siegford, 2012). The extensive collection of information through AMSs has led to an exponentially growing amount of data. This can be used to optimize herd management, but, on the other hand, the sheer volume of data represents a challenge in terms of its processing and analysis. Machine Learning (ML) algorithms are well-suited for analyzing datasets of this nature, as they are designed to effectively handle multi-dimensional, heterogeneous, and large datasets (Dulhare, 2020). Indeed, the quality of the ML model's outcome tends to improve with an increasing amount of data, given that the data is of high quality and relevance to the specific problem being modeled (Brownlee, 2021). Hence, ML models have emerged as promising tools for effectively modeling AMS data, with the goal of improving herd management practices. There are two main approaches to ML techniques: supervised and unsupervised learning (Jo, 2021). In the supervised approach, the model is trained using annotated data, aiming to optimize the accuracy of label predictions. On the other hand, in the unsupervised approach the model is trained on unannotated data, with the objective of identifying patterns that capture the underlying structure of the data.

* Corresponding author.
E-mail address: laura.ozella@unito.it (L. Ozella).

https://doi.org/10.1016/j.compag.2023.108002
Received 27 April 2023; Received in revised form 12 June 2023; Accepted 16 June 2023
Available online 22 June 2023

Cluster Analysis (CA) is an unsupervised ML method that aims to identify instances of similar types and group them into respective categories or clusters (Frades & Matthiesen, 2010). In CA algorithms, the concept of similarity encompasses two key properties: homogeneity among objects within the same cluster and heterogeneity among objects from different clusters (Everitt et al., 2011). When the data structure is linear and well-defined, various clustering algorithms generally produce consistent results. However, as the complexity of the data increases, different clustering algorithms are more likely to emphasize different characteristics of the data, leading to divergent outcomes. In such cases, determining the optimal combination of hyperparameters and models for data clustering becomes challenging.

The primary objective of this study is to examine the long-term patterns of milk productivity in cows milked by robotic milking systems. While existing literature predominantly focuses on predicting daily milk yield for the current lactation period (Masía et al., 2020; Fuentes et al., 2020; Piwczynski et al., 2020; Klis et al., 2021; Ji et al., 2022) our research aims to explore patterns extending beyond the current period. Developing a deeper understanding of milk productivity and its associated factors in future lactation periods is crucial for effective herd management. Making informed decisions about which cows should be retained for production requires anticipating their performance in advance. This aspect is of significant importance for the zootechnical sector, as a comprehensive long-term analysis can assist farmers in selecting cows for insemination and deciding which ones to keep for future lactation periods. Traditionally, this answer is given by considering only the current lactation productivity. Considering this demand, the first research question addressed in this study was whether the clustering of the data within each lactation period could effectively capture differences in cow productivity. The second research question investigated in this study was if these clusters, and more specifically these Productivity Groups (PGs)PGs formed by the clusters, are stable over time. Specifically, we first investigated if clusters formed inside each lactation period can support the grouping of the cows in terms of levels of productivity, and, secondly, if so, if these PPGs remain consistent across multiple lactation periods. To find the PGs, a CA with four clustering algorithms was performed with the data of each lactation period. Different clustering algorithms have the potential of capturing different aspects of the data, and it is not always clear which is the univocal best solution. The design presented in this study utilized multiple clustering algorithms to provide a more robust methodology that can identify a representative structure of the data. In this scenario, the merging step is extremely important, as its outcome is more reliable and less dependent on each single decision in the modeling workflow. Thus, the four outcomes were combined into a unique result with a proposed merging index. The merged clusters were classified into Low and High PGs, according to their values of Milk Production by Day (MPD). Finally, the stability of the PGs over the lactation periods was analyzed. The results show that the proposed methodology is useful for characterizing the cows according to their levels of productivity over lactation periods, which is a valuable information for herd management. This can help in deciding to inseminate a cow, based on the decision if the cow should be kept in production in future lactation periods. Data from one farm with 240 Holstein cows equipped with AMS was used to demonstrate the methodology.

2. Materials and methods

2.1. Dataset

The data utilized in this study was collected from August 2018 to January 2022 on a farm located in the northern part of Italy by four milking robots (Astronaut robot, Lely®, The Netherlands). The farm's herd is composed of Holstein Friesian cows. Raw data was extracted from the milking robot's management software (T4C "Time-for-Cows" InHerd, Lely, Maassluis, The Netherlands) through reports (i) of aggregated data by lactation, (ii) aggregated data by day and (iii) of single milking events. The choice of using only data available through the milking robot's system, i.e. not integrating external data, was made to ensure that the proposed methodology can be easily applied by any farm utilizing AMS without requiring additional data sources. The definition of the data features used in the analysis is presented in Table 1 and their statistical description is presented in Table 2.

The productivity of the cows was determined by dividing the total milk production (kg) by the total number of days in the corresponding lactation period. To ensure that the dataset only comprised complete day records, we excluded the first and last days of each lactation period. Additionally, 7 instances with outlier values were removed (High CDT/100 milkings > 5, DIM < 249 and Separated milk by colostrum > 64). Four daily registers of Rumination time by day that were below 300 were replaced by the mean of the Rumination time by day of the corresponding lactation period. Most of the clustering methods are not robust to differences in the scale of the features, thus, each variable was scaled with a Z-score Normalization (Singh and Singh, 2020).

Even though there were some instances from the fifth to the eighth lactation periods, only data up to the fourth lactation were used because for the higher periods there were too few instances to be analyzed by the clustering algorithms. The number of instances in each lactation period is presented in Table 3.

2.2. Data analysis

Fig. 1 presents the workflow of the analysis. After extracting, preprocessing and splitting the data by lactation, the clustering analysis was made separately for each lactation period. The MPD of the clusters was used to define the groups of productivity. As the different clustering algorithms could outcome a different PG for the same cow, the four productivity results were merged by the merging index defined in Section 2.3.5. Finally, the continuity of the PGs was analyzed graphically. All analyses were done using the R Core Team (2021) with random seed 1111.

Table 1
Description of the dataset features used to demonstrate the proposed methodology. Each instance corresponds to aggregated data for an entire lactation period of a cow. The dataset was obtained from the T4C 'Time-for-Cows' InHerd system by Lely, Maassluis, The Netherlands.

Feature	Definition
Lactation	The lactation period
Milk production by day	Average milk production (kg) by day
Milking frequency	Average number of times that the cow went to be milked by day
Refusals by milking	Average number of refusals (the robot refuses to milk the cow because the cow has gone too frequently to the robot)
Milking Robot Rate	Average quantity of milk (kg) by minute of use of the milking robot.
Milking Rate	Average quantity of milk (kg) by total time (sum of all quarter's times) in minutes milking
Errors every 100 milkings	Number of unsuccessful milkings (for example, if the cow moves too much) at each 100 milkings
High CDT every 100 milkings	Number of events of high conductivity in the milk at each 100 milkings
Watery milk every 100 milkings	Number of events of watery milk at each 100 milkings
Separated milk by colostrum	Total (kg) milk separated due to colostrum
Days in lactation	Duration of the lactation period in days
Dry days	Duration of the dry period in days before the lactation cycle starts
Protein	Average percent of protein in milk
Fat	Average percent of fat in milk
Lactose	Average percent of lactose in milk
Rumination time by day	Average minutes of rumination by day

Table 2
Summary statistics of the features of the dataset used to demonstrate the proposed methodology.

Feature	Min	Median	Max	Mean	Std
Lactation	2	2	8		
Milk production by day	15.36	39.47	59.72	39.50	6.71
Milking frequency	2.01	2.94	5.02	2.96	0.55
Refusals by milking	0.00	0.19	1.52	0.25	0.24
Milking Robot Rate	0.95	2.80	5.53	2.83	0.85
Milking Rate	0.29	0.86	1.62	0.87	0.26
Errors every 100 milkings	0.00	0.81	36.38	1.79	3.52
High CDT every 100 milkings	**0.00**	**0.00**	17.07	0.32	1.68
Watery milk every 100 milkings	**0.00**	**0.00**	6.35	0.09	0.52
Separated milk by colostrum	**0.00**	**5.90**	82.50	11.73	15.32
Days in lactation	249	313	478	326.12	42.72
Dry days	22	66	112	65.29	9.48
Protein	2.69	3.35	3.68	3.33	0.15
Fat	2.36	3.68	5.18	3.72	0.55
Lactose	4.71	4.89	5.06	4.89	0.06
Rumination time by day	420.33	538.45	638.55	531.62	47.84

Table 3
Number of instances in each lactation period in the dataset used to demonstrate the proposed methodology.

	Lactation							
	1	2	3	4	5	6	7	8
Number of instances	147	89	43	25	12	5	2	1

2.3. Clustering analysis

2.3.1. Clustering criteria

The results of the clustering algorithms were assessed by two clustering internal criteria. This kind of criteria measures how much the clusters reflect the patterns of the data without using any external information. The two internal criteria used in the present work were the Silhouette and the Dunn indexes.

The Silhouette index (Eq. (1) (Rousseeuw, 1987) is measured for each individual observation, it ranges from −1 to 1, and it assesses the pairwise difference of within and intra-cluster distances. Its average maximization can be used to define the optimal number of clusters (Liu et al. 2010). It can be interpreted as a measure of cluster consistency: the closer to 1, the more similar the observation is to its own cluster and the more different it is from the other clusters.

$$\boldsymbol{S(x)} = \sum_{x \in C_I} \frac{\boldsymbol{b(x)} - \boldsymbol{a(x)}}{\boldsymbol{max}|\boldsymbol{b(x)}, \boldsymbol{a(x)}|} \tag{1}$$

where x is the observation, C_I is the cluster to which x belongs, and $a(x)$ and $b(x)$ are defined below:

$$a(x) = \frac{1}{N_I - 1} \sum_{i \in C_I} d(x, x_i) \tag{2}$$

$$b(x) = \min_{J \neq I} \left(\frac{1}{N_J} \sum_{j \in C_J} d(x, x_j) \right) \tag{3}$$

where N_I is the number of instances in the cluster C_I and $d(x, x_i)$ is the distance between x and x_i, C_J are the clusters to which x do not belong and $d(x, x_j)$ is the distance between x and x_j.

The Dunn index (Eq. (4) (Dunn, 1974) ranges from zero to infinity and it is calculated by the ratio between the smallest inter-cluster and the largest intra-cluster distances:

$$\boldsymbol{D} = \frac{\boldsymbol{min}_{1 \leq I < J \leq M} \boldsymbol{\delta(C_I, C_J)}}{\boldsymbol{max}_{m \in M} \boldsymbol{\Delta(C_m)}} \tag{4}$$

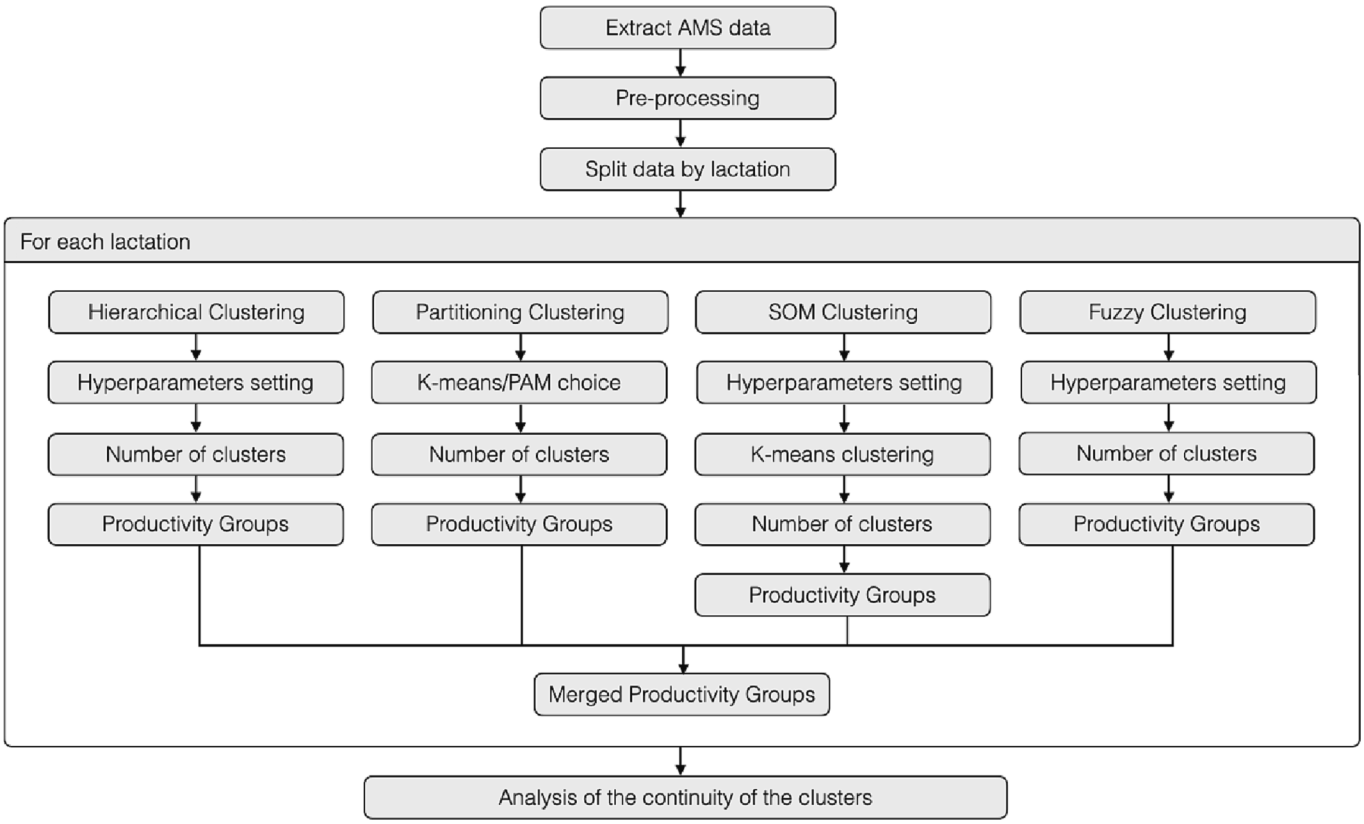

Fig. 1. Diagram of the steps of the analysis.

where M is the number of clusters, $\delta(C_I, C_J)$ is the distance between clusters C_I and C_J and $\Delta(C_m)$ is the diameter of the cluster C_m. More homogeneous, or compact, clusters and more well-separated clusters have a higher value of the Dunn index. However, it can be biased in a situation in which the clusters are not homogeneous, as it is assessed by the extreme values of the intra-cluster and inter-cluster distances.

2.3.2. Clustering algorithms

In the present work, four clustering algorithms were used, namely Agglomerative Hierarchical Clustering (HC), Partitioning Clustering, Self-Organized Maps (SOM) Clustering and Fuzzy Clustering. Each of them uses a different strategy to form the clusters. In HC, the clusters are formed by grouping together the closest clusters, one by one. Initially, the clusters are each single data observation. The method that calculates the closeness between clusters is a hyperparameter of the model and in the present work the Ward with squared dissimilarities (Ward D2) (Murtagh and Legendre 2014) was used. In Partitioning Clustering, the clusters are formed by instances that are closer to the center of the clusters. In K-means (Lloyd, 1957; MacQueen1967), this center is the mean of all the instances in the cluster. In Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 1987), this center is the most centrally located observation of the cluster. In the present work, the Silhouette and Dunn indexes were used to choose between K-means and PAM in each lactation period. SOM (Kohonen, 1982) clustering is a two-steps clustering method. First, the SOM is generated. The SOM is an unsupervised neural network in which each node is iteratively approximated to the closest data observation and to its neighboring nodes. The number of iterations, the grid size of the network and neighborhood radius are hyperparameters of this method. Once the SOM is defined, its nodes are then clustered with a clustering algorithm. In the present work the K-means was used to cluster the nodes of the SOM. In Fuzzy clustering, the instances can belong to more than one cluster, with different membership degrees. For each observation, the membership degrees should sum up 1. The Fuzzy K-means algorithm used in the present work was the Kaufman and Rousseeuw (1990). It consists in minimizing the following sum:

$$\boldsymbol{\Sigma}_{v=1}^{k} = \frac{\boldsymbol{\Sigma}_{i,j=1}^{N} \boldsymbol{u}_{i,v}^{r} \boldsymbol{u}_{j,v}^{r} \boldsymbol{d}(\boldsymbol{i},\boldsymbol{j})}{2 \times \boldsymbol{\Sigma}_{j=1}^{N} \boldsymbol{u}_{j,v}^{r}} \tag{5}$$

where N is the number of instances, k is the number of clusters, $u_{i,v}$ is the membership of the object i in cluster v, r is the membership exponent and $d(i,j)$ is the dissimilarity between instances i and j. Thus, the membership exponent, r, is an extra hyperparameter of this algorithm that affects the convergence rate of the algorithm. Table 4 presents the values of the specific hyperparameters used for each clustering algorithm.

2.3.3. Choice of the number of clusters

The Silhouette and Dunn indexes were used to support the choice of the number of clusters of each clustering algorithm. First, both indexes were used to define the most promising solutions. As an example of this first step of the analysis, Fig. 2 displays the plots of these indexes for the data from the first lactation period using the HC algorithm. The left plot shows the value of the mean of the Silhouette index of all instances for different numbers of clusters (k), from 2 to 11. The right plot shows the Dunn index values for the same number of clusters. The vertical dotted lines indicate the solutions with higher values of the respective index.

After that, the individual Silhouette indexes of all data points of the best solutions found in step 1 were graphically analyzed. As an example of this second step of the analysis, Fig. 3 presents the plots of the individual Silhouette index values for the solutions with 2, 3, 7, 8, and 11 clusters, specifically for the data from the first lactation period using the HC algorithm. In these plots, the vertical bars represent the Silhouette index of each observation, and the colors represent the clusters. Notice that the y-axis is centered in zero, which makes it easy to have an idea of the scale of the positive and negative Silhouette indexes. A good solution will not have a cluster in which most of the solutions have negative Silhouette indexes, and it will also not have solutions with very negative values of this index. In this example, the solution with $k = 3$ was chosen. Besides these properties, solutions in which the size of the clusters was not exceedingly irregular were preferred. The plots of these two first steps of the analysis for the other algorithms and lactation periods are available in the Supplementary Material.

Table 4
Hyperparameters used in the clustering algorithms.

Clustering algorithm	R function	Hyperparameter	Lactation	Value
HC	stats::hclust()	–	–	–
Partitioning	ClusterR:: KMeans_rcpp() for K-means ClusterR:: Cluster_Medoids() for PAM	Algorithm	All	K-means
SOM	kohonen::supersom()	Iterations	All	1000
		Neighborhood radius	1	2.34
			2	2.02
			3	2.00
			4	1.99
		Map units	1 and 2	8×8
			3 and 4	4×4
Fuzzy	Cluster::fanny()	Iterations	All	1000
		Membership exponent	All	1.2

2.3.4. Productivity groups

The significance of the differences in MPD among the clusters was assessed with a Kruskall-Wallis test (Kruskal & Wallis, 1952) with a level of significance of 0.05. The cows belonging to the clusters with the smaller MPD were assigned to the Low Productivity Group and the cows belonging to the clusters with higher MPD were assigned to the High Productivity Group. Notice that the definition of the clusters refers to each lactation and each clustering method. Therefore, a cow *could* belong to the Low *and* to the High productivity groups in the same lactation period, if considering different clustering methods.

2.3.5. Merging index

In order to assign cows to a unique productivity group, a merging index (Eq. (6)) is proposed. To the best of our knowledge, although very useful, there is no such index for combining the results of different clustering algorithms in literature.

$$m_i^{MPD_g} = \frac{\sum ca\left[d_N^{ca} \times \frac{sil_i^{ca}+1}{2}\right]}{n_{CA}} \tag{6}$$

where i is the i^{th} cow, $MPD_g = \{low, medium, high\}$are the MPD groups, $ca = \{hierarchical, partitioning, SOM, fuzzy\}$are the clustering algorithms, $d_N^{ca}$is the normalized Dunn index of the cluster to which the cow I was assigned using the algorithm ca, $sil_i^{ca}$is the individual Silhouette index of the cow I using the algorithm ca and $n_{CA}$is the number of clustering algorithms that were used. This index is an average of the results of the different clustering algorithms, weighted by the quality of the solution, which is assessed by the Dunn index of the whole clustering solution and the Silhouette index of the individual instances. The index gives the degree of membership of each cow to each PG. The higher the index, the more the cow belongs to that group. The final PG of a cow was determined by the group with the highest merging index.

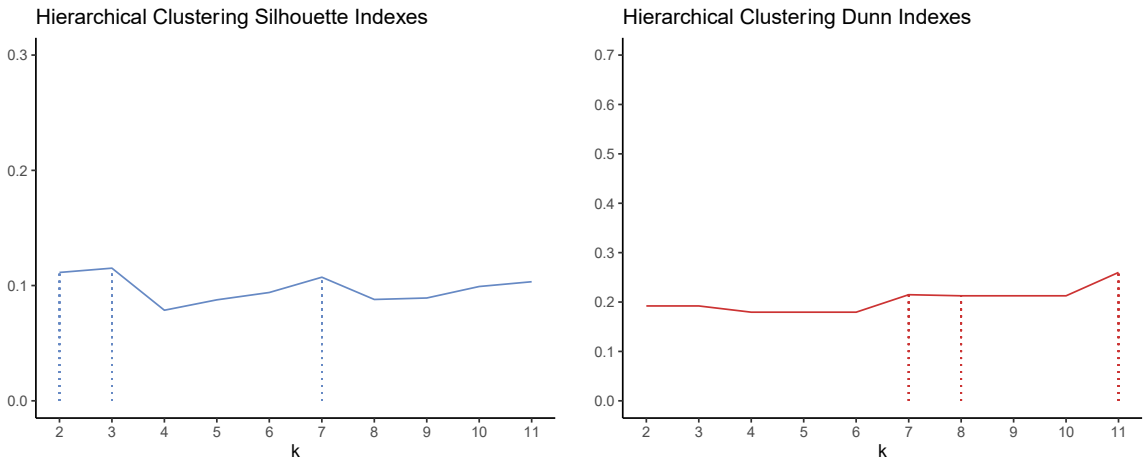

Fig. 2. Mean of Silhouette index values (left) and Dunn index values (right) for the Hierarchical Clustering solutions with the data from the first lactation period using from 2 to 11 clusters.

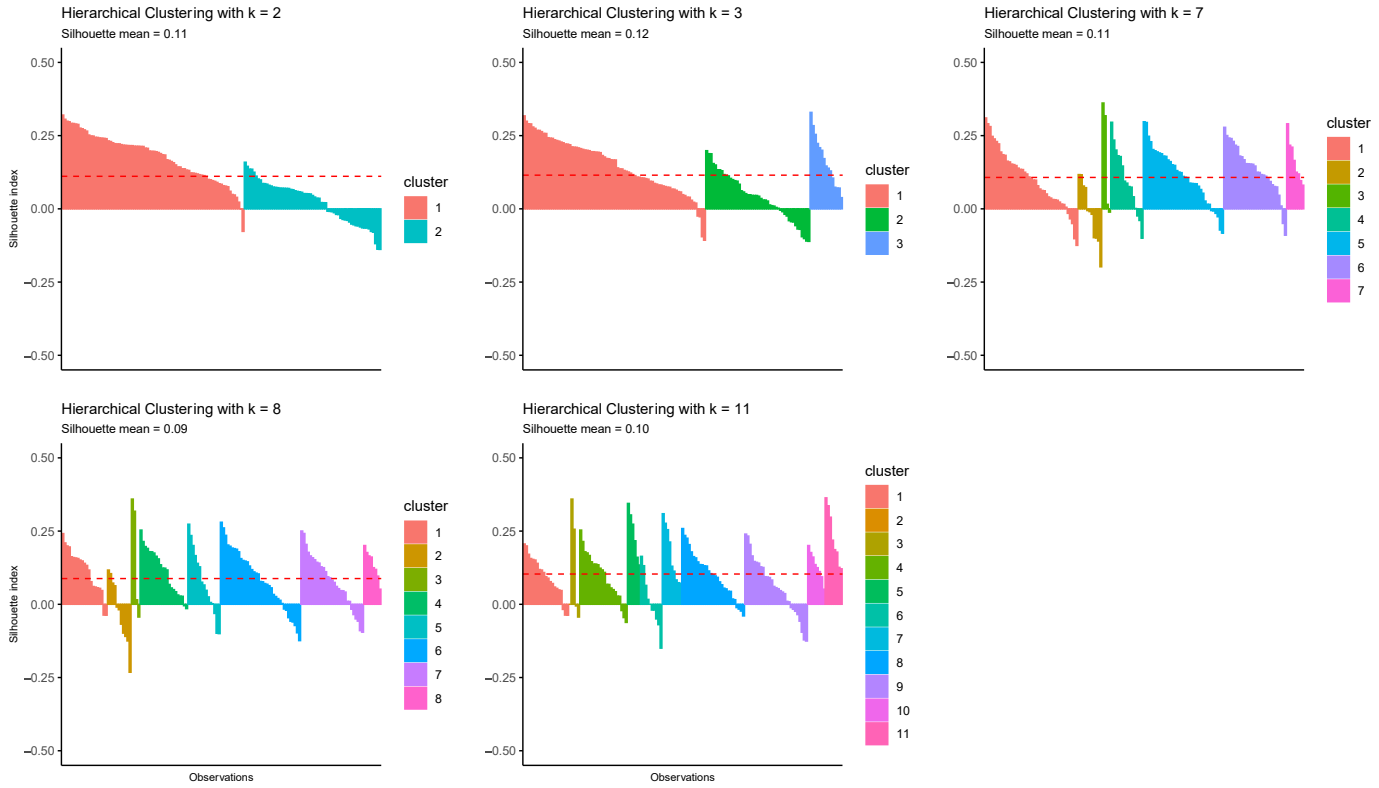

Fig. 3. Individual Silhouette index values for the best solutions found in the previous step of the analysis for the Hierarchical Clustering with data from the first lactation period.

3. Results and discussion

3.1. Number of clusters

Table 5 shows the number of clusters used in each lactation period for all clustering algorithms. Fuzzy Clustering was the most different among the four algorithms compared. The HC, Partitioning and SOM Clustering used the same number of clusters for the second lactation period, indicating that the pattern that they found in the data of this lactation period was very strong. Interestingly, the SOM algorithm uses the Partitioning K-means clustering at its second step, but it did not contribute to a higher similarity between the solutions of these two algorithms.

Table 5
Number of clusters for each lactation period with the four clustering algorithms used in the analysis.

	Clustering Algorithms			
Lactation	Hierarchical	Partitioning	SOM	Fuzzy
First	3	2	3	2
Second	4	4	4	2
Third	3	5	7	4
Fourth	4	4	5	3

For the first lactation, the best solutions of the HC were found with $k = 2$ and $k = 3$. The k = 3 was chosen based on two criteria. First, the third cluster had only instances with positive values of the individual Silhouette index. Second, the absolute value of the negative values of the Silhouette index decreased when using three clusters in comparison to the solution using two clusters. The best solution of the Partition Clustering algorithms was the one using $k = 2$ and the K-means algorithm. In this solution, almost all instances had positive values of the Silhouette index and the negative values were all very small. For the SOM Clustering algorithm, the best solution was found with $k = 3$. This solution had a good Silhouette mean and none of the instances presented a very negative Silhouette index. For the Fuzzy Clustering, solutions using $k = 2$ and $k = 3$ were good in terms of the individual Silhouette indexes. The solution with $k = 2$ was chosen because its Silhouette mean and Dunn indexes were higher.

For the second lactation, the optimal solution for the HC, Partitioning and SOM Clustering algorithms was achieved with $k = 4$. This choice was driven by the low values of the negative Silhouette index observed for individual instances. Furthermore, the distribution of individual Silhouette index values exhibited remarkable similarity among these solutions. In fact, there was a significant 67% correspondence observed among the three clustering approaches. and it was chosen because of the small values of the negative Silhouette index values of individual instances. The distribution of the individual Silhouette index values was also very similar among these solutions, that had 67% of correspondence. However, the mean Silhouette index for the HC was smaller, 0.10 vs. 0.12 for Partitioning and SOM Clustering algorithms. The K-means was used in the Partitioning clustering. For the Fuzzy Clustering algorithm, the best solution was also chosen because of the small values of the negative Silhouette index of individual instances, but it used $k = 2$.

For the third lactation, the best solutions of the HC were found with $k = 2$ and $k = 3$. The $k = 3$ was chosen because most of the instances of the third cluster had positive values for the Silhouette index. For the Partitioning Clustering the solution with $k = 5$ was chosen because its mean Silhouette did not decrease in comparison to the solutions with $k = 6$ and $k = 7$ and it splits the data better than the solutions with $k = 2$. The best solution of the SOM Clustering was found using $k = 7$. It had just one observation with a negative Silhouette index and a higher mean Silhouette index than the solutions with $k = 8$ and $k = 9$. For the Fuzzy Clustering algorithm, the solution with $k = 4$ was chosen because it had higher mean Silhouette and Dunn indexes. There were other solutions with more clusters and fewer instances with negative individual Silhouette index values, however, they had instances forming a single cluster, which is not useful for the analysis.

For the fourth lactation, the solution with $k = 4$ was chosen for the HC and Partitioning Clustering algorithms, both because the negative values of the individual Silhouette index were smaller using this number of clusters. Both solutions had two clusters with one single observation, with cows ID 505 and 581, indicating that the solutions were very similar. For the SOM Clustering, the best solution used $k = 5$. It has one observation with a very negative value of the individual Silhouette index. However, this caused a decrease of just 0.01 in the mean of the Silhouette index, in comparison to the solution with $k = 4$. This indicates that, even though the solution was worse for this specific observation, it was better overall. The solution of the Fuzzy Clustering using $k = 3$ was chosen because it had fewer negative individual Silhouette index values.

3.2. Productivity groups within lactation periods

3.2.1. Clustering algorithm solutions

The first research question addressed by this study was whether the clustering of the data within each lactation period could effectively capture differences in cow productivity. A positive answer to this question helps in characterizing the cows according to their productivity levels. The PGs of each algorithm for the first, second and third lactation periods are presented next. Two PGs, Low and High PG, were defined in each lactation, separately for each clustering algorithm, and, lastly, for the Merged solution. The PGs of the fourth lactation period could not be defined because the number of instances was not enough for finding any difference statistically significant in the MPD among the clusters of this lactation period.

3.2.1.1. First lactation. Fig. 4 shows the distribution of the MPD of each cluster, the size of the clusters and the PG of the clusters in the first lactation. Cluster 1 of the HC, Partitioning and Fuzzy algorithms formed the Low PG. The other clusters formed the High PG. For the HC solution, clusters 2 and 3 were combined into the High PG because they were not statistically different, indicating that other features than the MPD were important for the definition of these clusters. In other words, the split of the cluster 2 of the $k = 2$ solution into clusters 2 and 3 of the $k = 3$ solution did not improve the characterization of the clusters in terms of milk productivity. Clusters of the SOM algorithm could not be classified in terms of their productivity because the MPD of its clusters was not statistically different.

The correspondence between the PGs of the Partitioning and Fuzzy Clustering algorithms was 95.92%, with only 6 of 147 instances clustered in a different group. The correspondence of the PGs of these two clustering algorithms with the PGs of the HC was 82.31%. The correspondence between the PGs of the HC with the Partitioning and Fuzzy PGs was the same because of the 6 instances that were not corresponding in these two groups, 3 were classified by the HC solution in the same PG as the Partitioning solution, and the other 3, in the same PG as the Fuzzy solution. It is worth mentioning that even though the HC split the instances into three clusters, the PGs were highly correspondent to the other algorithms. These results show that the MPD was among the features that mostly contributed to the definition of the clusters, except for the SOM algorithm.

Table 6 shows the mean and standard deviation of MPD of the PGs of the first lactation period. The differences between the means and the standard deviation of the groups were very similar, indicating that the separation of the PGs among the clustering algorithms was equivalent.

3.2.1.2. Second lactation. Fig. 5 shows the distribution of the MPD of each clusters, the size of the clusters and the PG of the clusters in the second lactation. Cluster 3 of the HC, cluster 2 of the Partitioning and SOM, and cluster 1 of the Fuzzy algorithm formed the High PG. The other clusters formed the Low PG. For the HC solution, clusters 1, 2 and 4 were combined into the High PG, and for the Partitioning and SOM solution, clusters 1, 3 and 4 were combined into the High PG because they were not statistically different. Thus, in this lactation period, the Fuzzy clustering was more efficient, as its optimal split of the data into clusters already reflected the PGs.

The highest correspondence between PGs was seen among the groups of the Partitioning and SOM Clustering algorithms, 93.26%. The lowest correspondence was seen among the groups of HC and Fuzzy Clustering, 84.27%. All other correspondences were 86.51%. Overall, the groups had high correspondence. This can be seen in Fig. 5 by the number of instances in the High and Low PGs.

In addition, in these plots it is possible to notice that the clusters of the Partitioning, SOM and Fuzzy solutions were better separated in terms of the MPD than the clusters of the HC solution. Table 7 shows the mean and standard deviation of MPD of the PGs, which confirm this observation. This is also indicated by the minimum MPD of the High PG, 38.80 for the HC, 40.98 for Partitioning and SOM, and 39.13 for the Fuzzy solution. Notice that the minimum MPD of the Fuzzy High PG is greater, even though this PG of the Fuzzy solution has more instances than the High PG of the HC solution. In other words, it was observed that although the High PG of the Fuzzy clustering solution had a larger number of instances, it exhibited higher cohesiveness compared to the High PG of the HC solution.

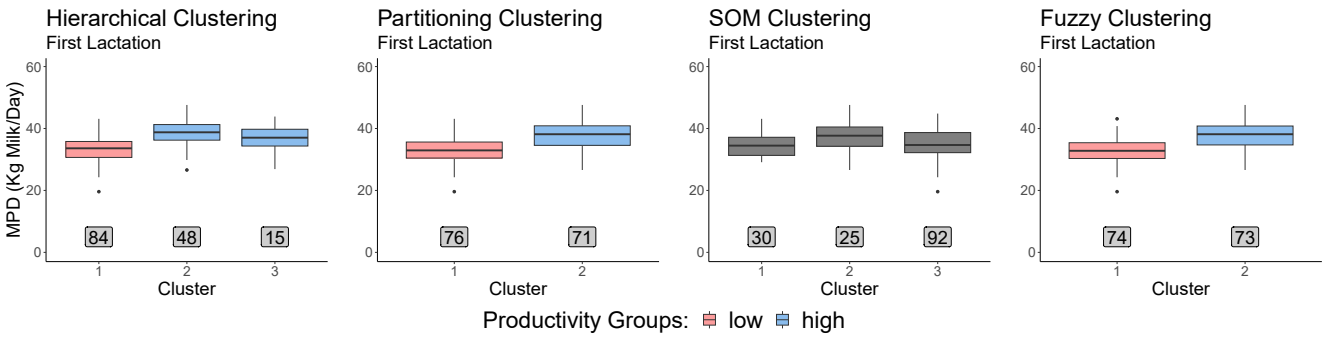

Fig. 4. Distribution of Milk Production by Day (MPD) of the clusters of Hierarchical, Partitioning, SOM and Fuzzy Clustering algorithms for the first lactation period. Clusters of the SOM Clustering solution were not statistically significant in terms of MPD. At the bottom of the plots, there is the number of instances in each cluster.

Table 6
Comparison of Productivity (Milk Production by Day) between High and Low Productivity Groups in the first lactation period. The values represent the means, and the values in parentheses indicate the respective standard deviations. The "Mean difference" row displays the numerical difference between the mean values of the High and Low Productivity Groups.

Productivity Group	HC	Partitioning	SOM	Fuzzy
High	38.08 (4.31)	37.84 (4.20)	–	37.90 (4.02)
Low	33.31 (4.14)	33.03 (4.18)	–	32.85 (4.21)
Mean difference	4.77	4.81	–	5.05

3.2.1.3. Third lactation. Fig. 6 shows the distribution of the MPD of each cluster, the size of the clusters and the PG of the clusters in the third lactation. In this lactation period, the instances were split into more clusters, except for the HC solution. Consequently, most of the clusters were not significantly different in terms of MPD.

For the HC solution, only clusters 2 and 3 had statistically different MPD. Cluster 1 was not statistically different from none of the other clusters because it had just 5 instances. It was joined to cluster 2 to form the Low PG. For the Partitioning solution, only clusters 2 and 5 were statistically different between them. Clusters 3 and 4 were joined to cluster 2 to form the Low PG, and cluster 1 was joined to cluster 5 to form the High PG. For the SOM solution, only clusters 3 and 4 were statistically different between them. Clusters 1, 6, and 7 were joined to cluster 3 to form the Low PG, and clusters 2 and 5 were joined to cluster 4 to form the High PG. For the Fuzzy solution, only clusters 2 and 4 were statistically different between them. Clusters 1 and 3 were joined to cluster 2 to form the Low PG. This recombination of the clusters produced PGs statistically different for all clustering algorithms. The *p*-values of the Kruskall-Wallis tests between the Low and High PGs of the third lactation are presented in Table 8. The need for this recombination indicates that the number of instances in this lactation period was close to the minimum limit for the proposed methodology.

As for the second lactation, the higher correspondence between the PGs was seen among the groups of the Partitioning and SOM Clustering algorithms, 93.02%. The correspondence between the groups of HC and Fuzzy was 88.37%. The correspondence between the groups of Partitioning and HC and Fuzzy was 69.77%. Finally, the correspondence between the groups of SOM and HC and Fuzzy was 67.44%. Overall, the correspondence among the groups was smaller for the third lactation period than the correspondence seen in the first two lactation periods.

In addition, the differences among the MPD of the PGs were smaller than those observed in the second lactation period, mostly due to a higher mean of the Low PGs, as can be seen in Table 9.

In general, the clusters formed in the first and second lactation periods have shown to be effective for creating the groups of productivity of cows, except for the SOM solution in the first lactation period. The PGs of the different clustering algorithms in these two first lactations were also compatible, indicating that the individual solutions were consistent enough. The only solution that was clearly better than the others was the Fuzzy solution of the second lactation period. For the third lactation, none of the algorithms generated a solution that could be directly used to form the PGs. Additionally, the correspondence among the solutions obtained from different clustering algorithms was lower compared to the other lactation periods. This was probably due to the

Table 7
Comparison of Productivity (Milk Production by Day) between High and Low Productivity Groups in the second lactation period. The values represent the means, and the values in parentheses indicate the respective standard deviations. The "Mean difference" row displays the numerical difference between the mean values of the High and Low Productivity Groups.

Productivity Group	HC	Partitioning	SOM	Fuzzy
High	46.93 (4.96)	48.46 (4.26)	48.00 (4.25)	47.31 (4.66)
Low	40.32 (5.10)	40.49 (4.87)	40.16 (4.86)	39.45 (4.42)
Mean difference	6.61	7.97	7.84	7.86

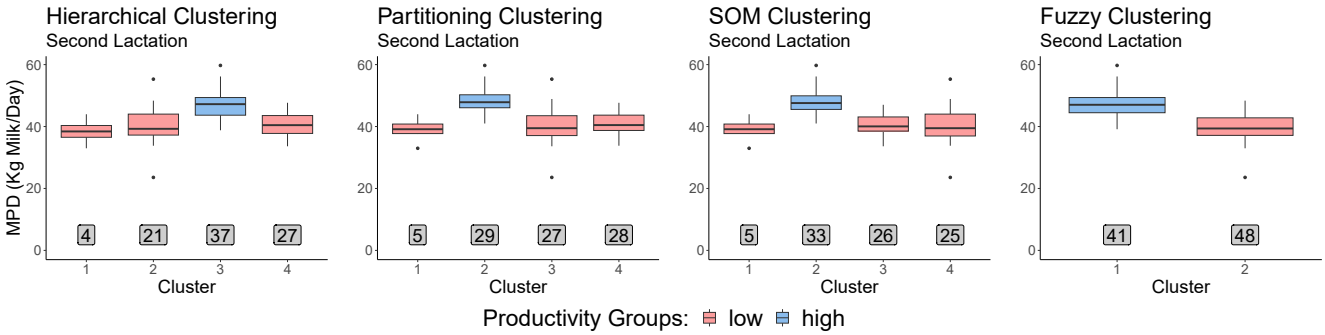

Fig. 5. Distribution of Milk Production by Day (MPD) of the clusters of Hierarchical, Partitioning, SOM and Fuzzy Clustering algorithms for the second lactation period. At the bottom of the plots, there is the number of instances in each cluster.

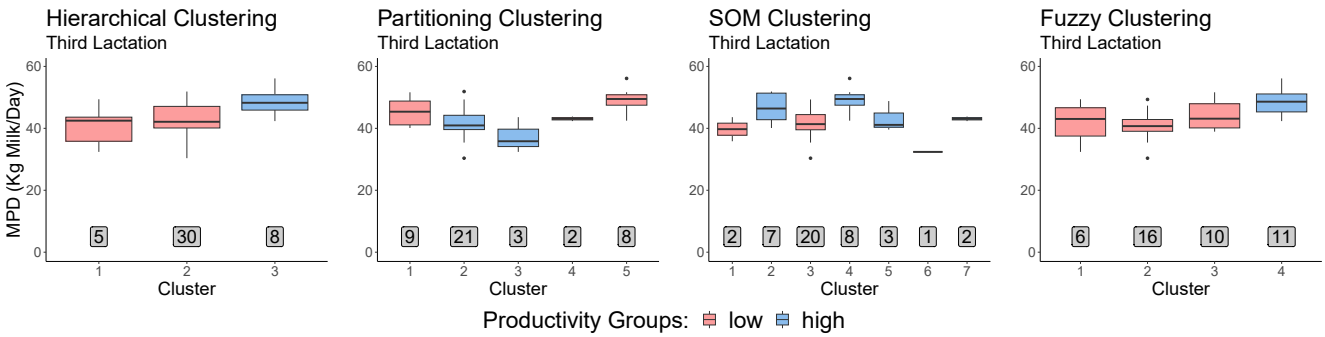

Fig. 6. Distribution of the Milk Production by Day (MPD) of the clusters of Hierarchical, Partitioning, SOM, and Fuzzy Clustering algorithms for the third lactation period. At the bottom of the plots, there is the number of instances in each cluster.

Table 8
Significance (p-value) of the Kruskal-Wallis test comparing the Milk Production by Day between Low and High Productivity Groups of the third lactation period.

	HC	Partitioning	SOM	Fuzzy
P-value	0.00730	0.00075	0.00065	0.00068

Table 9
Comparison of Productivity (Milk Production by Day) between High and Low Productivity Groups in the third lactation period. The values represent the means, and the values in parentheses indicate the respective standard deviations. The "Mean difference" row displays the numerical difference between the mean values of the High and Low Productivity Groups.

Productivity Group	HC	Partitioning	SOM	Fuzzy
High	48.45 (4.67)	46.86 (4.47)	46.83 (4.78)	48.37 (4.12)
Low	42.67 (5.31)	41.28 (5.22)	41.06 (4.89)	42.16 (5.18)
Mean difference	5.78	5.58	5.77	6.21

smaller number of instances for this lactation, and this reinforces the usefulness of a merged solution, which will give a more robust final solution even if there are fewer data or if the structure of the data is more complex. The subsequent section presents the results derived from the merged solution.

3.2.2. *Merged solution*

Fig. 7 shows the distribution of the MPD of each PG and the size of the groups of the Merged solution for the first, second and third lactation periods. As observed in the results of the clustering algorithms, the bigger difference between the MPD of the Low and High PG of the Merged solution was found in the second lactation, then in the third, and lastly in the first lactation period. In the Merged solution, for the first lactation, the number of instances is balanced between the Low and High PG. However, for the second and third lactation periods, there are more instances in the Low PG. It is important to keep in mind that this does not mean that the productivity in the second and third lactation was smaller. In fact, it can also be seen in these plots that the MPD of the second and third lactation was higher. The bigger Low PG means, though, that a smaller number of cows achieved a higher productivity in comparison to all selected cows in the respective lactation period.

Table 10 shows the p-values of the statistical tests for the differences between the Low and High PGs of the Merged solution on the values of all features of the dataset. As indicated in Section 2.3.4, the significance level used in the statistical tests was $\alpha = 0.05$, which with the Bonferroni correction for multiple comparisons became $\alpha = 0.0013$, in this case. In

Table 10
Significance (*p*-value) of the differences of the feature values in the Low and High Productivity Groups of the Merged solution. In bold, are the p-values that are below the critical value ($\alpha = 0.05$) after the Bonferroni correction.

	Lactation		
	First	Second	Third
Delivery age	0.0974	0.2680	0.3240
Days in milking	0.1210	0.4370	0.8740
Dry days	–	0.0327	0.0763
Milking frequency	**≪0.001**	**≪0.001**	0.1660
Milking robot rate	**≪0.001**	0.3290	0.3860
Milking rate	**≪0.001**	0.4810	0.8100
Refusals by milking	**≪0.001**	0.0163	0.419
Errors every 100 milkings	0.0349	0.6760	0.4460
Separated colostrum	0.2620	0.9790	0.4610
Rumination time by day	0.1450	0.0101	0.0513
Protein percent	**≪0.001**	**≪0.001**	0.0016
Fat percent	0.0015	**≪0.001**	**0.0011**
Lactose percent	**≪0.001**	**≪0.001**	0.0022

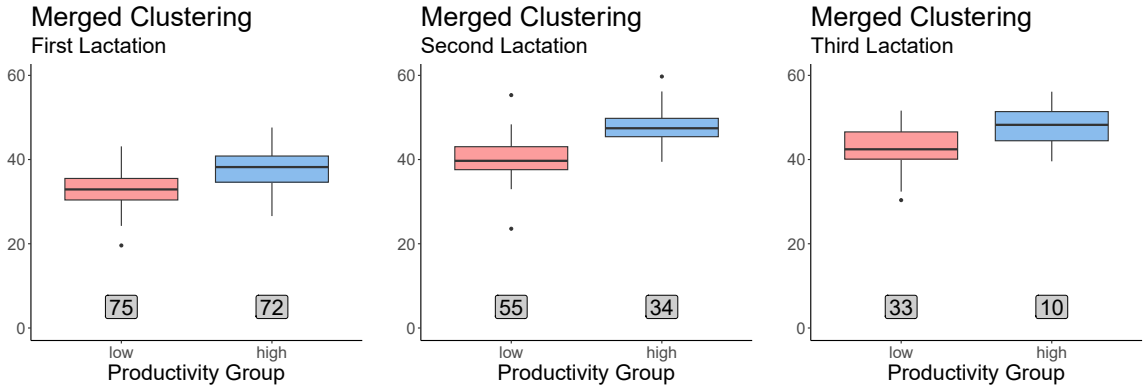

Fig. 7. Distribution of Milk Production by Day (MPD) of the Productivity Groups of the solution that merges the solutions of the Hierarchical, Partitioning, SOM and Fuzzy solutions in each lactation period.

the first lactation, Milking frequency, Milking robot rate, Milking rate, Refusals by milking, percentage of protein and percentage of lactose were significantly different. In the second lactation, Milking frequency, percentage of protein, percentage of fat and percentage of Lactose were significantly different. In the third lactation, only the percentage of fat was significantly different. Overall, the features with the smallest *p*-value were the percentage of protein, fat, and lactose. Even though the *p*-value of these features was not significant in all cases, it was always close to its critical value.

Fig. 8 shows the distribution of the features in which the differences between the Low and the High PGs were statistically significant in each lactation period. The percentage of protein, fat, and lactose are shown for all three lactation periods because even if their *p*-values were not below the significance level, they were all close to the critical value, as previously observed. The Milking frequency is higher in the High PG both in the first and second lactation periods. This result is in accordance with the study conducted by Lyons et al. 2013, in which they found that higher Milking frequency was associated with higher MPD for cows in all stages of lactation in a pasture-based AMS. In the present study, the non-significative difference in the third lactation can be due to the small number of instances in this lactation period. The Refusals by milking was also higher in the High PG in the first lactation, which can be explained by the greater Milking frequency during this period. The Milking robot rate and Milking rate were lower in the High PG of the first lactation. Both rates are associated, and this result indicates that it can be difficult to find cows that optimize the use of the milking robots in terms of Milk Harvest per Robot (MHR). In literature, this key performance indicator is evaluated by combining the Milking interval (the inverse of Milking frequency) and the number of cows per robot in the barn (Molfino, 2018). If cows belonging to High PG have a low Milking rate and vice-versa, it implies that high-productivity cows occupy the milking robots for a relatively longer duration compared to low-productivity cows. Consequently, the Milking rate should be considered as an additional factor when analyzing the efficiency of milking robot utilization. The High PG had lower values for all three features related to milk composition, suggesting that cows in this group may have a higher water content in their milk. It is well known that cows kept indoors tend to drink more water in response to high temperatures during summer months, resulting in higher milk content of water. Thus, this could have been a confounding factor in this result. It was verified using the differences in the month of the year in which the lactation cycle started, as the environmental data was not available. According to Masía et al. (2020), milk production is the highest in the first third of the lactation period. Thus, the weather in this initial phase can have a higher impact in the total production of the lactation cycle. The *p*-values of the statistical test in the difference on the initial month of the lactation period for the first, second and third lactation periods were 0.2137, 0.2957 and 0.0033. Thus, period this could be the case only in the third lactation period. Therefore, this still does not explain the observed result in the first and second lactation periods in terms of the differences between the High and Low PGs on the features related to milk contents, and further investigation is needed.

3.3. Continuity of the productivity groups

Once the clusters within each lactation period were defined and characterized by the PGs, the second research question investigated in this study was if these clusters, and more specifically these PGs, are stable over time. Fig. 9 shows the Sankey network of the flow of the cows belonging to the PGs in consecutive lactation periods. It is worth mentioning that only cows that reached the next lactation period are represented in the corresponding previous period. The blue and red vertical bars represent the High and Low PGs, respectively. The empty space to the right side of the vertical bars of the second lactation represent the number of cows that did not continue in production from the second to the third lactation. The width of the grey bars is proportional to the flow of the cows between the PGs, and the numbers into circles also show this information. Importantly, in this plot, the less grey bars split and cross, the more stable the PGs are.

In general, the PGs were unstable in both transitions analyzed, from the first to the second and from the second to the third lactation periods. The Low PG was less unstable than the High PG. For the Low PG, 93.10% of the cows remained in this group in the transition between the first and second lactation periods, and 79.17% remained in the transition between the second and third lactation periods. For the High PG, 73.53% of the cows remained in this group in the transition between the first and second lactation periods, and 62.50% remained in the transition between the second and third lactation periods. Therefore, proportionally, more cows moved from the High to the Low PG in both transition periods, from the first to the second and from the second to the third

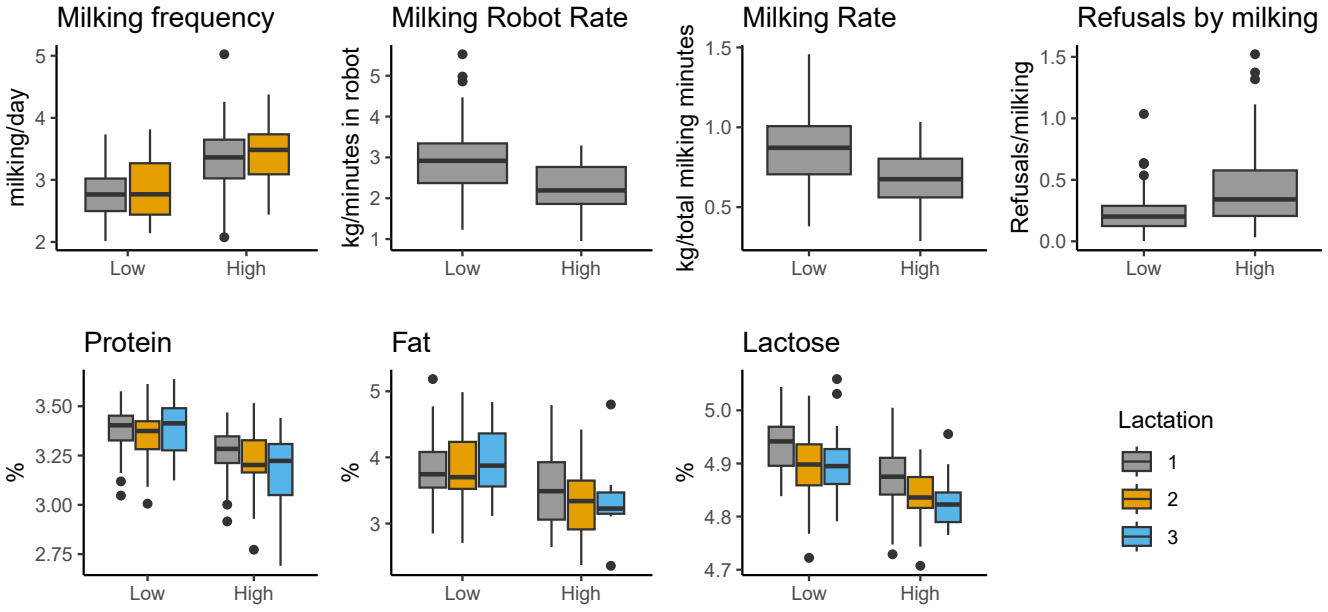

Fig. 8. Distribution of the features that were significantly different between the Low and High Productivity Groups. Only the lactation cycles in which the difference was significant are shown, except for the features related to milk content, which are shown in all three lactation periods.

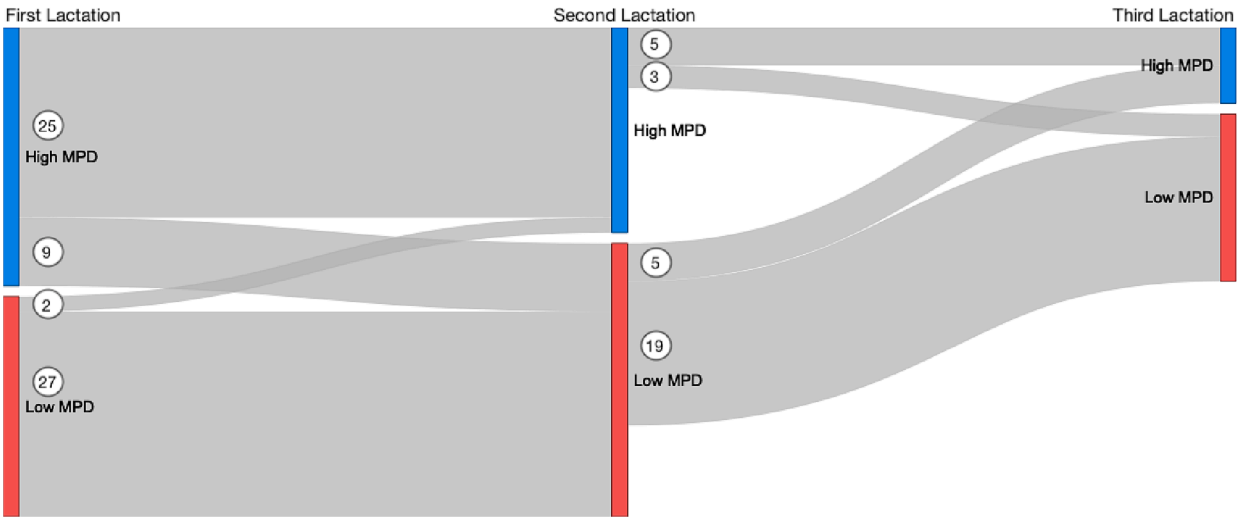

Fig. 9. Flow of the cows belonging to the Low and High Productivity Groups in each lactation period. Numbers into circles show the number of cows in the respective connection between the Productivity Groups of two consecutive lactation periods.

lactation. However, this does not necessarily imply a decrease in their productivity. Since the PGs were defined within each lactation period, they reflect the productivity of a cow in relation to the productivity of all other cows in the same lactation period. Three situations can make a cow switch from the High to the Low PG: (i) if the productivity of that cow was kept stable, the productivity of the other cows increased; (ii) if the productivity of that cow decreased, and either the productivity of the other cows increased, were kept stable or decreased at a lower rate; (iii) if the productivity of that cow increased, and the productivity of the other cows increased at a higher rate. To better understand which is the case in the analyzed data, the overall productivity in terms of MPD in each lactation is presented in Table 11.

The overall mean of the MPD increased from the first to the second lactation period. Hence, the greater number of cows switching from the High to the Low PG indicates that 26.47% of the cows did not keep up with the overall increase in productivity from the first to the second lactation period. On the other hand, the two cows that switched from the Low to the High PG had their productivity increased at a greater rate. In the transition between the second and third lactation period, the overall mean did not change. Thus, the three cows that switched from the High to the Low PG had their productivity decreased, and the five cows that switched from the Low to the High PG had their productivity increased.

Both PGs, Low and High, were more stable in the transition between the first and second lactation than in the transition between the second and third lactation periods. It is known from literature that the first lactation is different from the subsequent ones for several reasons. Firstly, it is the period in which the cow undergoes the greatest physiological changes, such as the transition from non-lactating to lactating state. Moreover, the mammary gland of the cow is still developing during the first lactation, which may affect the milk yield and composition (Gorewit, 1988). Thus, the result of the present study showed that the productivity pattern of the cows tended to change evenly from the first to the second lactation periods, with few exceptions. In contrast, no significant overall trend was observed in cow productivity from the second to the third lactation period. This can be explained by the differences in MPD between the PGs, presented in Section 3.2. The larger differences in terms of MPD were seen between the PGs of the second lactation, for all clustering algorithms, while in the third lactation, this difference decreased. With a smaller difference between the High and Low PG, the probability of a cow switching from one group to another is higher. Interestingly, from the 10 cows in the High PG of the third lactation period, 5 of them were in the Low PG and 5 were in the High PG in the second lactation period. Six of them were not in the data in the first lactation, and from the 4 remaining, 2 were in the Low PG and 2 were in the High PG in the first lactation period. This reinforces the conclusion that the PGs are not stable over lactation periods. Therefore, our results showed a weak relationship between the cow's productivity in past lactation periods and its potential productivity in future periods. Other factors, such as the cow's health, genetics, and environmental conditions, could significantly influence its future productivity.

Table 11
Mean and standard deviation of the Milk Production by Day (MPD) in each of the first three lactation periods.

	First	Second	Third
Mean	35.35	43.07	43.75
Standard deviation	4.82	5.99	5.60

4. Conclusions

This study presents a novel approach for characterizing Productivity Groups (PGs) of milking cows within each lactation period, and for assessing their stability over time through the use of clustering algorithms. Four algorithms were used to define the clusters in each lactation period, namely Agglomerative Hierarchical Clustering (HC), Partitioning Clustering, Self-Organised Maps (SOM) Clustering and Fuzzy Clustering. To combine the four outcomes into a univocal result, a merging index was proposed. This merging index is an average of the results of different clustering algorithms, weighted by the quality of the solution, which was assessed by two internal clustering criteria (i.e., Dunn and Silhouette indices). The proposed merging index provides a measure of the extent to which each cow belongs to each productivity group, with higher values indicating a stronger affiliation with that group. The clusters were categorized into High and Low PGs based on the Milk Production by Day feature values.

The final PG of a cow was determined by the group with the highest merging index. To demonstrate the methodology, data from first lactation periods from one farm with Holstein Friesians cows that exclusively uses the Automatic Milking System (AMS) was used. To ensure the applicability of the proposed methodology to other farms that use similar milking robots, only data available through the AMSs was utilized. However, it is important to note that the proposed methodology can be applied to other datasets as well, including those obtained from traditional milking systems or for addressing other scientific inquiries.

To address the scientific question regarding the stability of PGs over

time, data from the first, second, and third lactation periods were considered. The findings indicate that the PGs exhibited a relatively weak level of stability across lactation periods. Specifically, the Low PG displayed less variability during the transitions from the first to the second lactation period and from the second to the third lactation period. Furthermore, it was observed that the Low PG exhibited a larger size in both the second and third lactation periods. These findings indicate that the future productivity of a cow in subsequent lactation periods is not strongly correlated with its past productivity. Therefore, relying solely on the present or past lactation productivity of cows for the purpose of selecting them for insemination may not be the most effective strategy. The results found in this analysis are farm-dependent, and they may be subject to a selection bias, as not all cows were kept from the beginning to the end of the data collection. Instead, only those selected by the herd management were kept in production. A more informative study would involve examining the same cows across all lactation periods, without selectively choosing which cows to keep in production. This approach would provide a better understanding of the defining factors and dynamics of the PGs.

CRediT authorship contribution statement

Karina Brotto Rebuli: Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Visualization. **Laura Ozella:** Conceptualization, Methodology, Data curation, Writing – original draft. **Leonardo Vanneschi:** Conceptualization, Methodology, Writing – review & editing. **Mario Giacobini:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

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.

Data availability

The data that has been used is confidential.

Acknowledgements

This study is supported by Compagnia di San Paolo (ROL 63369 SIME 2020.1713) and by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2023.108002.

References

Brownlee, J., 2021. Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End. 176 pp.
Dulhare, U.D., 2020. Machine Learning and Big Data: Concepts, Algorithms. Tools and Applications, Wiley-Scrivener, p. 538.
Dunn, J.C., 1974. Well separated clusters and fuzzy partitions. J. Cybernetics 4, 95–104.
Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Cluster Analysis. Willey. 352 pp.
Frades, I., Matthiesen, R., 2010. Overview on techniques in cluster analysis. Bioinform. Methods Clin. Res. 81–107.
Fuentes, S., Viejo, C.G., Cullen, B., Tongson, E., Chauhan, S., S., Dunshea, F. R. 2020. Artificial Intelligence Applied to a Robotic Dairy Farm to Model Milk Productivity and Quality based on Cow Data and Daily Environmental Parameters. Sensors 2020, 20(10), 2975.
Gorewit, R., 1988. National Research Council (US) Committee on Technological Options to Improve the Nutritional Attributes of Animal Products. Designing Foods: Animal Product Options in the Marketplace. Washington (DC): National Academies Press (US). Lactation Biology and Methods of Increasing Efficiency.
Jacobs, J.A., Siegford, J.M., 2012. Invited review: The impact of automatic milking systems on dairy cow management, behavior, health, and welfare. J. Dairy Sci. 95 (5), 2227–2247.
Ji, B., Banhazi, T., Phillips, C.J.C., Wang, C., Li, B., 2022. A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. Biosyst. Eng. 216, 186–197.
Jo, T., 2021. Machine Learning Foundations. Springer Cham. 391 pp.
Kaufman, L., Rousseeuw, P.J., 1987. Clustering by means of medoids. Statistical Data Analysis Based on the L1-Norm and Related Methods, pp. 405-416.
Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.
Klis, P., Piwczynski, D., Sawa, A., Sitkowska, B., 2021. Prediction of Lactational Milk Yield of Cows Based on Data Recorded by AMS during the Periparturient Period. Animals 2021 (11), 383.
Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43 (1), 59–69.
Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583–621.
Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., 2010. Understanding of Internal Clustering Validation Measures. In: 2010 IEEE International Conference on Data Mining, Sydney, NSW, Australia, 2010, pp. 911-916.
Lloyd, S.P., 1957. Least squares quantization in PCM. Bell Telephone Laboratories Paper.
Lyons, N.A., Kerrisk, K.L., Dhand, N.K., Garcia, S.C., 2013. Factors associated with extended milking intervals in a pasture- based automatic milking system. Livest. Sci. 158 (1-3), 179–188.
MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate instances. In: Le Cam, L.M., Neyman, J. (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, University of California Press, California, pp. 281–297.
Masía, F.M., Lyons, N.A., Piccardi, M., Balzarini, M., Hovey, R.C., Garcia, S.C., 2020. Modeling variability of the lactation curves of cows in automated milking systems. J. Dairy Sci. 103, 8189–8196.
Molfino, J., 2018. Investigation into system and cow performance ef- ficiency in pasture-based automatic milking systems. The University of Sydney, Camden, Australia.
Murtagh, F., Legendre, P., 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J. Classif. 31 (3), 274–295.
Piwczynski, D., Sitkowska, B., Aerts, J., Schork, P.M., 2020. Forecasting the milk yield of cows on farms equipped with automatic milking system with the use of decision trees. Anim. Sci. J. 2020, 91.
R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Singh, D., Singh, B., 2020. Investigating the impact of data normalization on classification performance. Appl. Soft Comput. 97, 105524.