Revista Ciência Agronômica, v. 51, n. 3, e20196977, 2020
Centro de Ciências Agrárias - Universidade Federal do Ceará, Fortaleza, CE
www.ccarevista.ufc.br

Scientific Article
ISSN 1806-6690

Principal component and cluster analyses to evaluate production and milk quality traits[1]

Análise de agrupamento e componentes principais para avaliar características de produção e qualidade do leite

Bueno da Silva Abreu[2]*, Severino Benone Paes Barbosa[2], Elizabete Cristina da Silva[2], Kleber Régis Santoro[3], Ângela Maria Vieira Batista[2] and Rafael Leonardo Vargas Martinez[4]

ABSTRACT - Using multivariate analyses, this study attempted to identify important traits explaining the relationship between milk production and quality produced by Holstein cows. Monthly milk records from three commercial dairy farms located in the Agreste region of Pernambuco, Brazil, collected in the period from 2007 to 2017, were used. A total of 5,872 observations regarding milk production, milk components and somatic cell score (SCS) were analyzed using principal component analysis (PCA) and cluster analysis. According to the former analysis, the first three principal components explained 79.69% of the total variation. Total solids content contributed 29.66% of the variation in the first principal component, while lactose content contributed 49.43% of the variation in the second principal component. According to the latter analysis, three clusters differed for all characteristics ($p<0.001$) and cluster 2 concentrated 43.15% (2,534) of the information with lower SCS and higher lactose content and milk production. Total solids, lactose and fat were considered the most representative traits explaining the variability of the data set. The multivariate techniques used in this study proved useful in obtaining effective characteristics, with three factors considered important in explaining the relationship between Holstein cows' milk production and quality.

Key words: Multivariate analysis. Dairy cattle. Milk composition.

RESUMO - Objetivou-se com este estudo identificar, por meio da análise de componentes principais e análise de agrupamento, as variáveis capazes de explicar a variabilidade na qualidade e na produção de leite de vacas Holandesas. Foram utilizados dados mensais de controle leiteiro, de três fazendas comerciais localizadas na região Agreste do estado de Pernambuco, Brasil, obtidos no período de 2007 a 2017. Foram analisadas 5.872 informações de produção e componentes do leite, e de escore de células somáticas (SCS), quanto à possibilidade de formação de grupos que pudessem ser destacados pela similaridade e verificar a capacidade discriminante dessas características nos grupos. Os métodos K-means e Ward.D2, baseados na análise da distância euclidiana e dos componentes principais (PCA), foram utilizados para indicar as fontes de variação que diferenciaram os grupos. Foi observado que os primeiros três componentes principais explicaram 79,69% da variabilidade dos dados. A variável que mais contribuiu no primeiro componente foi o teor de sólidos totais com 29,66%. No segundo componente, a lactose, se destacou com uma contribuição de 49,43%. Na análise de agrupamento, três *clusters* diferiram em relação a todas as características ($p<0,001$), o *cluster* 2, por exemplo, concentrou 43,15% (2.534) das informações, agrupando animais com um menor SCS e maior lactose e produção de leite. As variáveis sólidos totais, lactose e gordura foram as que mais contribuíram dentro dos três componentes selecionados. A ACP e agrupamento podem ser ferramentas úteis na obtenção de características efetivas, sendo três fatores considerados importantes para explicar a relação entre produção e qualidade do leite.

Palavras-chave: Análise multivariada. Bovinocultura leiteira. Composição do leite.

DOI: 10.5935/1806-6690.20200060
*Author for correspondence
Received for publication 07/10/2019; approved on 04/02/2020
[1]Parte da Tese do primeiro autor, apresentado ao Curso de Pós-Graduação em Zootecnia, Universidade Federal Rural de Pernambuco/UFRPE
[2]Departamento de Zootecnia, Universidade Federal Rural de Pernambuco/UFRPE, Recife-PE, Brasil, abreu607@hotmail.com (ORCID ID 0000-0001-7549-5448), severino.pbarbosa@ufrpe.br (ORCID ID 0000-0003-4987-078X), bete_zootec@hotmail.com (ORCID ID 0000-0001-6698-6528), angelamvbatista@gmail.com (ORCID ID 0000-0001-6133-2795)
[3]Departamento de Zootecnia, Universidade Federal do Agreste de Pernambuco/UFAPE, Garanhuns-PE, Brasil, kleber.santoro@ufrpe.br (ORCID ID 0000-0002-7592-8423)
[4]Associação de Criadores de Pernambuco/ACP, Recife-PE, Brasil, snc@uol.com.br (ORCID ID 0000-0001-8020-9522)

INTRODUCTION

Dairy activity has stood out in national and global agribusiness for playing a particularly important role in providing food for human consumption, generating jobs and incomes, and increasing various countries' gross domestic product (GDP). Besides constituting a key component of the economic sector, milk is a food of high nutritional value for humans: indeed, it is rich in nutrients essential to growth (especially in childhood), helps in maintaining a healthy life (GÓMEZ-CORTÉS; JUÁREZ; DE LA FUENTE, 2018) and reduces bone problems in older people (THORNING *et al*., 2016).

Milk productivity and quality vary between production systems, as each property has specific characteristics for obtaining milk in terms of milking types, nutritional and sanitary management (CUNHA *et al*., 2008), beyond the effects of weather conditions (temperature, humidity and precipitation) (QI; BRAVO-URETA; CABRERA, 2015) and physiological factors (age at first calving, parity order and lactation period) (CHEGINI *et al*., 2017).

Therefore, studying the behavior of milk components is crucial to determining the various sensory and industrial properties of milk quality and production (RIBAS *et al.*, 2014), beyond the profitability of the farm (CINAR *et al*., 2015). For industry, a milk with high value in terms of its somatic cell count (SCC) is directly associated with reduced production of dairy products and their shelf life (CHEGINI *et al*., 2017). In some countries the parameters of milk protein and milk fat content are used to pay milk producers (e.g. Australia, New Zealand, Canada, the United States).

To evaluate the variables of milk production, composition and quality, univariate analyses can be considered limited, as they evaluate each of the variable individually, whereas multivariate analyses concomitantly evaluate a set of characteristics considering the correlations between variables, enabling better interpretations of the information extracted from a data set (SANTOS *et al*., 2010). Among the multivariate analysis techniques of great applicability in animal production, principal component analysis (PCA) and cluster analysis are both commonly used in several studies involving production animals (BODENMÜLLER FILHO *et al*., 2010; RIBEIRO *et al*., 2018; VENTURA *et al*., 2012).

As regards Brazilian dairy cattle, a vast literature addresses these statistical tools. Alessio *et al*. (2016), evaluated the factors that influence milk lactose variation in herds in Santa Catarina and verified the relationship between SCC and calving order. Santos *et al*. (2017), used multivariate analysis to characterize the production systems of the Brazilian Amazon. Haygert-Velho *et al*. (2018), analyzed monthly controls of milk production, composition and microbiological quality in Rio Grande do Sul and observed the formation and the separation of groups by season.

The objective of this study was to use PCA and cluster analysis to identify and analyze the variability in the milk production, components and somatic cell score (SCS) of Holstein cows reared in the Agreste region of Pernambuco, Brazil.

MATERIAL AND METHODS

All the information used in the present study was obtained from existing databases, so no approval was required from the Ethics Committee with the Use of Animals (CEUA).

To carry out this study, 5,872 milk production and composition observations of three commercial dairy herds of Holstein cows, located in the cities of Gravatá (8º12'04" S, 35º33'53" W) and São Bento do Una (08º31'22" S, 36º26'40" W) from the state of Pernambuco in Brazil were used. The climatic characteristics for the 11 years of data collection according to the National Institute of Meteorology (INMET) were: average annual temperature of 24.21 ºC (minimum of 20.43 ºC and maximum of 30.03 ºC); average annual precipitation of 551.7 mm (minimum of 376.9 mm and maximum of 859.4 mm); and average annual relative humidity of 75.92% (minimum of 73.30% and maximum of 78.14%).

The herds were characterized by milk production under a semi-arid climate and with similar food management (as concentrate, commercial feed based on corn and soybean; for forage, forage palm, corn silage and Tifton hay) over the years. Mineral salt and water were offered *ad libitum.* The farms were selected according to the number of animals controlled, the period in which the herds were in control and the frequency of monthly milk control. The facilities were similar, comprising a shed with feeders and drinking fountains as well as a free access area with a sand bed, except on farm two, which had use of a free stall, fans, sprinklers and rotating brushes.

Milk production, composition information and somatic cell count were obtained monthly from the dairy control reports of the selected herds. The data used were acquired from the official reports of the Northeast Dairy Herd Management Program (PROGENE) and the Pernambuco Breeders' Association (ACP), containing information according to the Brazilian Association of Holstein Cattle Breeders (ABCBRH).

The data were previously analyzed using the *outlierKD* function of R to exclude extreme values (outliers) that could interfere with the results. The variables of milk production (MY), protein (PROT), fat (FAT), lactose (LAC), total solids (ST), non-greasy solids (FNS) and SCS were analyzed using mean descriptive analyses, standard deviations and coefficients of variation. A Pearson's correlation analysis for complete observations was also used to study the relationship between the variables using the *rcorr* function in the *Hmisc* package together with the *p.adjust* function to obtain the p-values corrected by the Holm method (HOLM, 1979) between correlations.

PCA was used to indicate the sources of variation differentiating the groups, serving as a useful tool for identifying and understanding the patterns of association between the variables. PCA is a multivariate statistical technique that analyzes several variables to reduce a large dimension of data to a smaller number of dimensions and components (linear combination of variables), linearly independent of each other, representing a percentage of total covariance (SANTOS *et al*., 2019). As a result, it may reveal unperceived aspects in the univariate analysis of the original characteristics (RIBEIRO *et al*., 2018).

The number of main components to be retained (i.e. the number of principal components needed to explain the variability of milk components) was considered as eigenvalues (variances) greater than 1, following the Kaiser Rule (KAISER, 1960). Therefore, only components with own values greater than 1 were considered significant, while all components with own values below were considered insignificant and discarded. The relationship between the original variables and the main components was analyzed by Pearson's correlation. These analyses were carried out with the statistical packages *FactoMineR* (for analysis) and *factoextra* (for viewing the results).

Each variable used in this study was analyzed to check the possibility of forming groups based on similar characteristics of the milk analyzed as well as to verify the discriminant capacity of these characteristics in the formation of groups manifesting homogeneity within and heterogeneity between them (VENTURA *et al*., 2012). After the application of PCA, cluster analysis was used to improve the characterization and interpretation of the groups formed, as this method facilitates comparison of individuals within and between groups (TREMBLAY *et al*., 2016).

Thus, for the cluster analysis, the hierarchical method Ward.D2 was used, which aims to minimize the total variation within groups (MURTAGH; LEGENDRE, 2014) based on Euclidean distance using the statistical packages *cluster* and *factoextra*. The number of actual groups was determined by the K-means grouping method, which partitions *n* observations into *k* groups, with each observation designated in the group closest to the mean (MACQUEEN, 1967). After determining the actual number of groups, the *fviz_cluster* function was used to visualize the results in a scatterplot). PCA and cluster analysis were performed with standardized data to obtain a mean of 0 and a standard deviation of 1.

From the results of the cluster analysis it was possible to designate each observation to a specific group composed of similar samples and differing from the observations assigned to the other groups. This stage was also important to observe whether there was a grouping of observations according to the cows' specific physiological stage, according to their order of parity (1, 2, 3 e $\geq$4) and their month of lactation (1, 2, ..., 10).

To identify the variables that contributed to the differentiation of the groups formed in the cluster analysis, the assumptions of normal distribution of residues and homogeneity of variances were tested using the Kolmogorov-Smirnov and Levene tests, respectively. The Kruskal-Wallis nonparametric test was used to confirm the differences between the groups obtained with the cluster analysis. Moreover, Dunn's (1964) *post-hoc* test was used to identify the variables that helped differentiate the groups ($p<0.05$), with the p-values adjusted by the Bonferroni correction. In addition, the hierarchical clustering on principal components (HCPC) method was used to confirm the most important variables within each group using the package *FactoMineR* (HUSSON; JOSSE; PAGÉS, 2010).

All the statistical analyses mentioned above were performed using the R version programme 1.2.1335.

RESULT AND DISCUSSION

The descriptive statistics for daily milk production, milk components and SCS are presented in Table 1. Holstein cows reared in the semi-arid region of Northeast Brazil were found to have an average milk production of 32.50 kg/day. Such high production is mainly a reflection of the genetic quality of the animals, in addition to the good adaptive characteristics of the herd and the management conditions used in the properties studied in the Agreste region.

Regarding milk components, the mean values of fat, protein, lactose and total solids were 3.33%, 3.21%, 4.58% and 12.12%, respectively. The mean SCS was 3.89, corresponding to approximately 185,000 cells/ml (Table 1). These results are like those found by Ludovico *et al*. (2015), who obtained an average daily milk production of 31.78 kg and an average SCS of 3.46.

Table 1 - Average, minimum, maximum, standard error values (EP) and coefficient of variation (CV) of the variables evaluated in Holstein cows reared in the Agreste region of Pernambuco

Variable	Average	Minimum	Maximum	EP	CV
MY (kg)	32.50	10.30	61.00	0.107	25.12
FAT (%)	3.33	1.50	6.05	0.010	22.57
PROT (%)	3.21	2.00	4.99	0.005	12.09
LAC (%)	4.58	3.31	5.20	0.003	4.98
TS (%)	12.12	9.52	15.19	0.012	7.75
FNS (%)	8.79	6.46	10.13	0.006	4.98
SCS	3.89	0.00	9.64	0.031	6172

*MY: Milk yield; PROT: Protein; FAT: Fat; LAC: Lactose; TS: Total solids; FNS: Non-fat solids; SCS: Somatic cell score

The correlation coefficients between milk production and composition and SCS ranged from -0.43 to 0.89 (Table 2). Holm's method revealed that these coefficients were significant ($p<0.05$), except for the variables FNS and SCS ($p>0.05$), which did not present significant correlations. Milk production showed negative correlations of weak to moderate magnitude with fat, protein, total solids, non-greasy solids and SCS and a weak positive correlation with lactose content ($r = 0.17$). Positive correlations of strong magnitude were observed between fat and total solids ($r = 0.89$) and protein content and non-greasy solids ($r = 0.79$). High correlations between total solids and their components have been observed in several studies (BONDAN *et al*., 2018; LUDOVICO *et al*., 2015; SILVA *et al*., 2018), that these parameters are important in the breeding programmes of dairy cattle because they promote milk quality, since the payment of milk by dairy components aims to improve the quality of the raw material by increasing the industrial yield for the manufacture of various dairy products (CHEGINI *et al*., 2017).

SCS showed a negative correlation with milk and lactose production but a positive correlation with the variables such as fat, protein and total solids. In general, high SCS values is indicative of breast inflammatory processes, which can reduce milk production and increase solids concentration, resulting in the two characteristics becoming positively correlated (CINAR *et al*., 2015).

According to the behavior of the correlations' estimates, it is difficult to define strategies for the selection and the management of animals with the aim of improving the quality of the milk produced. Characteristics that present negative correlations indicate that the impact of the selection performed may act differently, while one characteristic increases the other may decrease. According to Knob *et al*. (2018) the stronger selection for increase milk production of Holstein cows in recent years, achieving high genetic gain for this trait, has impaired certain characteristics, such as the concentration of solids in milk.

Correlation analysis is an important source for understanding the degree of association between two or more variables, besides being an essential premise for use in multivariate analyses, considering that in order to use

Table 2 - Pearson's correlation matrix between milk production and composition characteristics in Holstein cows reared in the Agreste region of Pernambuco

	MY	FAT	PROT	LAC	TS	FNS	SCS
MY	1	-0.1674	-0.3260	0.1702	-0.2278	-0.2022	-0.0860
FAT	<.0001	1	0.2522	-0.0644	0.8948	0.2014	0.0402
PROT	<.0001	<.0001	1	-0.1870	0.5685	0.7929	0.2822
LAC	<.0001	<.0001	<.0001	1	0.1209	0.3714	-0.4346
TS	<.0001	<.0001	<.0001	<.0001	1	0.6130	0.0426
FNS	<.0001	<.0001	<.0001	<.0001	<.0001	1	0.0225
SCS	<.0001	0.0107	<.0001	<.0001	0.0065	0.2534	1

*MY: Milk yield; FAT: Fat; PROT: Protein; LAC: Lactose; TS: Total solids; FNS: Non-fat solids; SCS: Somatic cell score. Significant according to Holm's method ($p<0.05$)

PCA (for example), the variables present some degree of correlation.

In the PCA analysis, seven main components were obtained, of which the first three were selected according to the Kaiser Rule, because they presented self-values greater than 1. These three components were able to explain 79.69% of the total variation of the data, representing approximately 20% of the loss of explanation of the total variation. This finding demonstrates that the main component technique was effective in summarizing the variables responsible for the variability in the productive characterization of the milk of Holstein cows reared in the harsh region of Pernambuco (Table 3).

The results observed in this study corroborate those of other researchers who have also used PCA. For example, Bodenmüller Filho *et al*. (2010), evaluated the differences between production systems in the Northern region of Paraná using seven characteristics of milk production and quality and verified that three main components were sufficient to explain 70.52% of the total variance of the characteristics. Moreover, Fraga *et al.* (2016), assessed the relationship between the productive characteristics and the genotypic proportions of Holstein and zebu crossbred dairy cattle, observing that the first two principal components explained 89.4% of the total variance that represent the production and genotypic components.

In the present study, the first main component (CP1) represented 40.44% of the total variance and included the variables with the highest weighting coefficients (autovectors) and contributions (Figure 1A), represented by: total solids (0.916; 29.66%), followed by protein (0.823; 23.91%) and non-greasy solids (0.793; 22.23%). These variables were closely related to each other, characterizing CP1 as an index for determining solid milk content (Table 4).

The solid part of milk is an important indicator of the nutritional quality offered to the cows, besides indicating the quality of milk for the industrial process, as the higher the solids content, the higher the yield of the derivatives. This will consequently enable greater remuneration for the agents involved in the production chain (BELLI *et al*., 2017; CINAR *et al*., 2015).

In the second component (CP2), the variables of major autovectors and contributions were lactose (0.893; 49.43%) and SCS (-0.748; 34.67%), which explained 23.05% of the total variance (Figure 1B). This component

Table 3 - Eigenvalues, individual and cumulative proportion of the variation in milk production and composition characteristics explained by each main component evaluated in Holstein cows reared in the Agreste region of Pernambuco

Main Component	Eigenvalues	Total Variance (%)	Accumulated Variance (%)
CP1	2.831035411	40.443363	40.44336
CP2	1.613952848	23.0564693	63.49983
CP3	1.133622075	16.1946011	79.69443
CP4	0.870276506	12.4325215	92.12695
CP5	0.506521498	7.2360214	99.36298
CP6	0.042972737	0.6138962	99.97687
CP7	0.001618925	0.0231275	100

Table 4 - Autovectors for the seven descriptive variables according to the three main components retained in the cluster analysis

Variable	CP1	CP2	CP3
MY	-0.417	0.297	-0.128
FAT	0.690	0.026	-0.721
PROT	0.823	-0.214	0.420
LAC	0.030	0.893	0.178
TS	0.916	0.162	-0.355
FNS	0.793	0.309	0.493
SCS	0.185	-0.748	0.143

*MY: Milk production; FAT: Fat; PROT: Protein; LAC: Lactose; TS: Total solids; FNS: Non-fat solids; SCS: Somatic cell score

can be defined as the index determining the hygienic-sanitary quality of the herd, the contrast between lactose and SCS being evident here. One of the main osmotic regulators of the mammary gland is lactose: with greater SCS there is a change in osmotic pressure, increasing the permeability of the alveolus separating the milk from the blood, and the results is a loss of lactose to the bloodstream (ALESSIO *et al*., 2016).

Fat presented a contribution of 45.86% and an autovector of -0.721 in the third component (CP3), explaining 16.19% of the total variation of the data. Fat content proved to be the component with the highest variability of the analyzed milk, being influenced by nutritional and environmental factors. Therefore, it is essential to supply good quality forage in the feed of lactating cows as a way of improving the composition, thereby producing milk with a higher fat content, one of the productive characteristics most valued by dairy industry to obtain higher prices for better quality, being a determinant of the product's yield (BELLI *et al*., 2017).

PCA can help producers to interpret the relationships between variables and consequently support their decisions regarding the selection of animals for a given characteristic, being able to define and identify the main variables and associations between them (PAIVA *et al*., 2010).

Through a projection graph (Figure 1C), the PCA allowed the behaviour of each variable to be evaluated according to the correlations inherent to the distribution of the components and as a function of the angle formed between the vectors. In such a projection graph, if the angle between the variables (vectors) is close to zero, the correlation is very high and positive and will be especially close; if the correlation is close to 180º, the correlation is also high, but negative and will be more distant; if the angle formed is about 90º, the variables are less correlated (BODENMULLER FILHO *et al*., 2010). A strong correlation could be observed between the contents of total solids, protein, non-fat solids and fat, characterizing the solid composition of milk and these variables were close to axis 1 and in the same quadrant (1), except for protein, which appeared in the quadrant 4. Therefore, we can verify that the variables with the greatest length vectors were the most important. In addition, the angle formed between the variables was less than 45º, indicating a strong relationship between those characteristics. The inverse situation was observed for milk yield, which formed an angle close to 180º and appeared in opposite quadrants, presenting a strong negative correlation. In the graph, different colors (C) are used to represent the correlations between the analyzed characteristics (*Cont.Var*) within a component, the colors black and blue represent positive and negative correlations respectively.

Similarly to CP1, in the CP2 the lactose ratio and milk yield were close while somatic cell score was related to opposite quadrants, indicating the negative association between these characteristics and represent a decrease in lactose and milk yield when somatic cell score increases. Lactose content is influenced by several sources of variation, mainly number of parities, lactation stage, health status of the udder and the individual animal in question (ALESSIO *et al*., 2016). The angle formed between lactose and SCS indicated a negative correlation (-0.43) between the variables, indicating that the lower the load of somatic cells in milk, the healthier is the cows, resulting in greater milk yield (SILVA *et al*., 2018).

In the cluster analysis, the definition of the number of groups was performed by the K-means method, in which the 5,872 observations resulted in the formation of three different groups composed of 1,357 (C1), 2,534 (2) and 1,981 (C3) observations (Table 5). A significant effect ($p<0.01$) between the formed group by test of Kruskal-Wallis was observed and the groups formed (clusters) represent a milk yield group and the composition of milk and SCS group (Table 5).

In Figure 2, the formation of three clusters by K-means (gap statistic*)* was observed. This test was performed through the partitioning method, defined so that objects within the same cluster would be as similar as possible, specifying each centroid in the group and assigned according to the average information. The cluster 1 comprises the variables related to the solid composition of milk, maintaining the same behavior as demonstrated by the first component (CP1) obtained by principal component analysis.

In the studied population, milk samples from cows of different calving orders were used. Cluster analysis rendered it possible to identify different groups according to the physiological age of the cows and the month of lactation (Table 6). In C1 grouped multiparous cows above the third order of calving. Samples with lower solids concentration and higher number of somatic cells in milk were commonly found in this study (Table 5).

Milk component contents varied with the number of lactations and the increase in SCS. Bondan *et al*. (2018), and Ludovico *et al*. (2015), observed declines in milk solids content with increased lactations and SCS, respectively. With the increase in the order of calving, the animals become more susceptible to infection, so the increase in SCS in milk may be partially justified by the increase in epithelial cells of mammary gland flaking present in milk in multiparous cows (CUNHA *et al*., 2008; GALVÃO JÚNIOR *et al*., 2010).

Primiparous cows and animals in the first months of physiological lactation tend to produce better milk

quality (i.e. with a lower number of somatic cells), mainly because they are not exposed to the factors responsible for infection of mammary gland. In cluster 2 (C2), the samples were lower SCS and higher lactose content and milk yield, and concentrated cows of first and second orders of calving in the five months of lactation (Table 6).

Figure 1 - Contribution (A and B) and projection of variables (C) in the main components (CP1 and CP2): milk yield (MY), lactose (LAC), Non-fat solids (FNS), total solids (TS), fat (FAT), protein (PROT) and somatic cell score (SCS)

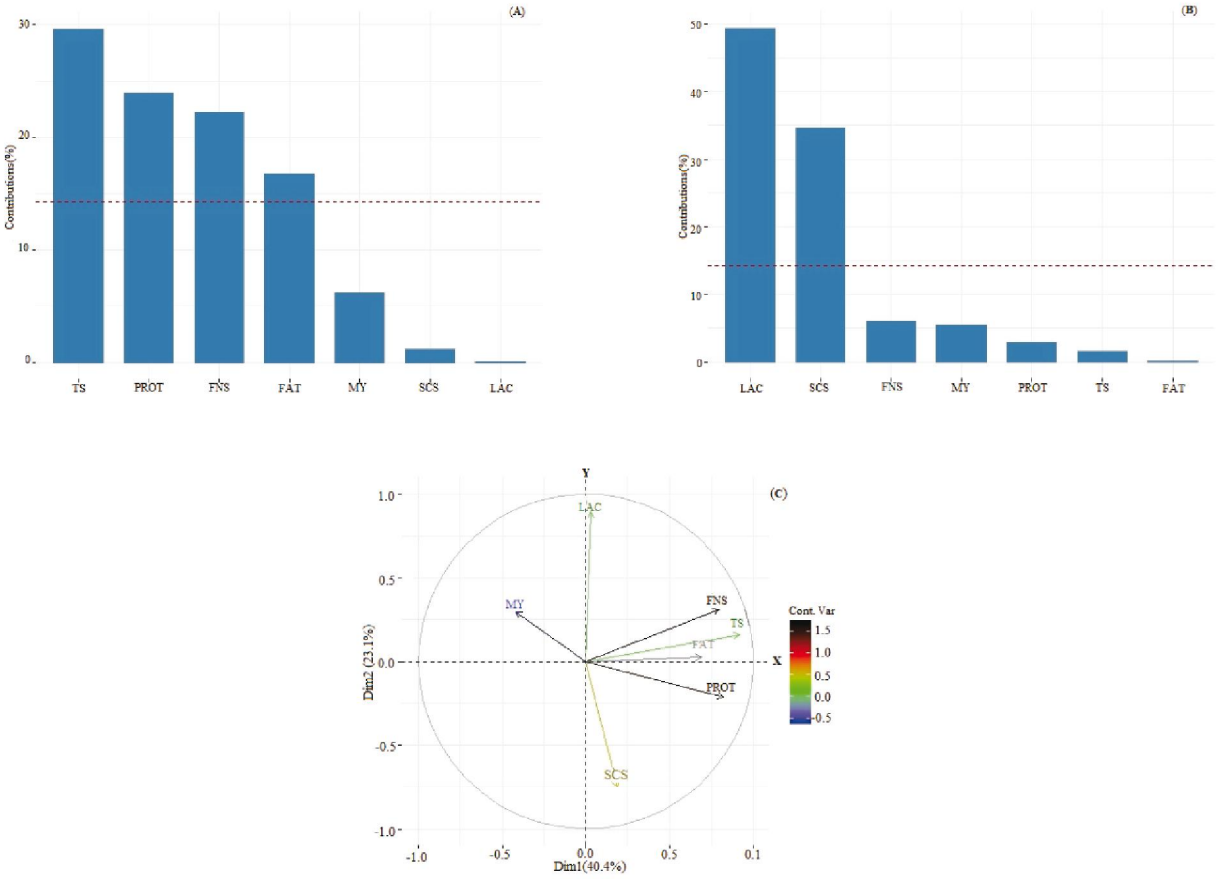

Table 5 - Comparison of the differences between the clusters formed by the variables evaluated in Holstein cows in the Agreste region of Pernambuco

	Cluster1 (n=1,357)	Cluster2 (n=2,534)	Cluster3 (n=1,981)	P-value
FAT	2.96 ± 0.017 c	3.08 ± 0.011 b	3.89 ± 0.016 a	<0.001
FNS	8.37 ± 0.010 c	8.72 ± 0.006 b	9.15 ± 0.007 a	<0.001
LAC	4.36 ± 0.005 c	4.70 ± 0.002 a	4.57 ± 0.005 b	<0.001
MY	33.33 ± 0.226 b	35.11 ± 0.156 a	28.58 ± 0.155 c	<0.001
PROT	3.05 ± 0.009 b	3.02 ± 0.005 c	3.56 ± 0.007 a	<0.001
SCS	5.69 ± 0.058 a	2.45 ± 0.032 c	4.49 ± 0.051 b	<0.001
TS	11.34 ± 0.018 c	11.80 ± 0.012 b	13.05 ± 0.015 a	<0.001

*MY: Milk yield; PROT: Protein; FAT: Fat; LAC: Lactose; TS: Total solids; FNS: Non-fat solids; SCS: Somatic cell score. Means followed by the same letter in the line do not differ from each other by Dunn's test (P<0.05)

Figure 2 - Definition of the number of clusters by K-means (gap statistic) (A) and the representation of the clusters formed (B) from the analysis of main components for the characterization of milk production and quality of Holstein cows reared in the Agreste region of Pernambuco

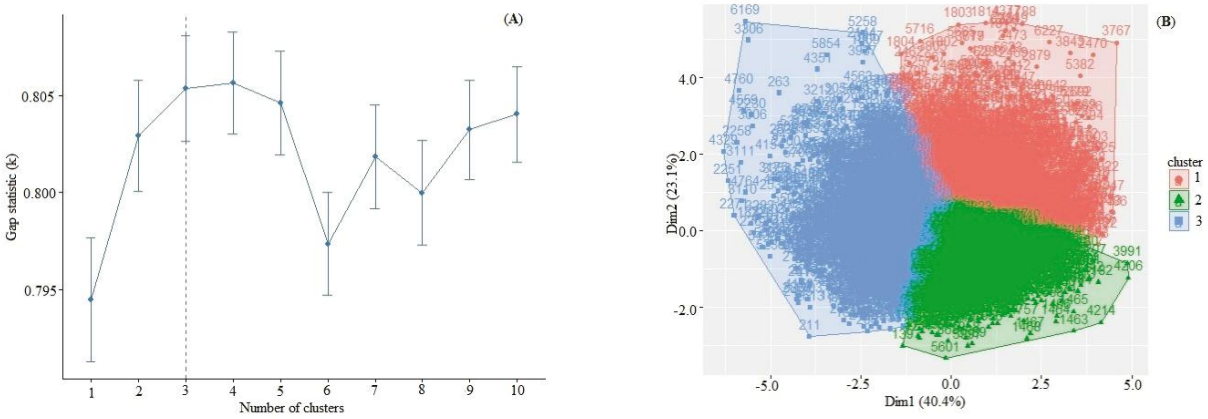

Table 6 - Percentages of distributions in clusters according to orders of deliveries (ORD1; ORD2...) and lactation month (1; 2; 3...) of Holstein cows reared in the Agreste region of Pernambuco

Characteristics	Cluster 1 (n=1,357)	Cluster 2 (n=2,534)	Cluster 3 (n=1,981)
ORD1	12.93	49.01	38.05
ORD2	22.09	42.62	35.29
ORD3	42.24	34.90	22.86
ORD≥4	43.71	29.30	26.99
1	23.46	58.77	17.76
2	27.29	64.08	8.64
3	26.40	63.99	9.62
4	25.95	58.92	15.14
5	25.14	53.39	21.47
6	24.82	32.52	42.66
7	22.65	36.08	41.27
8	23.23	30.75	46.02
9	21.94	27.25	50.81
10	18.01	21.76	60.23

According to Alessio *et al*. (2016), and Bondan *et al*. (2018), the number of lactations and the lactation phase are the most important predictors of lactose concentration in milk. Evaluating the milk characteristics of Holstein cows, Bondan *et al.* (2018) observed that lactose concentrations decreased with increases in SCS and the number of lactations and were affected linearly and negatively by the lactation phase. Indeed, the concentrations were higher in cows up to four months into the lactation cycle, before gradually decreasing until the end of lactation. The high milk production observed in C2 may have also been due to the lactation phase of the animals being the peak period, characterized by higher milk production, which usually occurs after around 60 days of lactation (CHEGINI *et al*., 2017).

The lactation phase is extremely important to producers to define the appropriate management of animals, mainly in terms of nutrient demand. Milk constituents typically tend to increase with lactation

time. Indeed, only lactose is reduced, a fact that coincides with lower milk yield. The third cluster (C3) contained the highest fat, protein and total solids content and the lowest milk production, grouping animals from the sixth month of lactation. Animals in the final lactation phase have generally been found to produce lower volumes of milk with higher numbers of total solids, fat and protein (BONDAN *et al*., 2018). Studying the relationship between milk chemical composition and lactation stage, Cabral *et al*. (2016), identified the lactation stage's significant influence on total dry extract, protein and fat, observing an increase in these parameters until the end of lactation. For example, the fat concentration was 3.91% in animals at 305 days of lactation, and cows in the initial phase of lactation presented an average of 3.24%.

CONCLUSION

1. Principal component and cluster analysis enabled a reduction in the number of characteristics evaluated in milk production, composition and quality data and made an important contribution to explaining the characteristics of Holstein cows in the Agreste region of Pernambuco. Indeed, it was possible to distinguish groups according to their physiological classes and month of lactation;
2. The variables total solids, lactose and fat were the most important among the three selected components, defining as important characteristics in the selection of animals to improve the milk quality of the herds.

REFERENCES

ALESSIO, D. R. M. *et al.* Multivariate analysis of lactose content in milk of Holstein and Jersey cows. **Semina: Ciências Agrárias**, v. 37, n. 4, p. 2641-2652, 2016.

BELLI, C. Z. P. *et al.* Qualidade do leite cru refrigerado obtido em unidades produtivas no Sudoeste do Paraná. **Revista de Ciências Agroveterinárias**, v. 16, n. 2, p. 109-120, 2017.

BODENMÜLLER FILHO, A. *et al.* Tipologia de sistemas de produção baseadas nas características de leite. **Revista Brasileira de Zootecnia**, v. 39, n. 8, p. 1832-1839, 2010.

BONDAN, C. *et al*. Milk composition of Holstein cows: a retrospective study. **Ciência Rural**, v. 48 n. 12, e20180123, 2018.

CABRAL, J. F. *et al*. Relação da composição química do leite com o nível de produção, estádio de lactação e ordem de parição de vacas mestiças. **Revista do Instituto de Laticínios Cândido Tostes**, v. 71, n. 4, p. 244-255, 2016.

CHEGINI, A. *et al*. Effect of somatic cell count on milk fat and protein in different parities and stages of lactation in Holstein cows. **Acta Agriculturae Slovenica**, v. 110, n. 1, p. 37-45, 2017.

CINAR, M. *et al.* Effect of somatic cell count on milk yield and composition of first and second lactation dairy cows. **Italian Journal of Animal Science**, v. 14, n. 1, p. 105-108, 2015.

CUNHA, R. P. L. *et al*. Mastite subclínica e relação da contagem de células somáticas com número de lactações, produção e composição química do leite em vacas da raça Holandesa. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 60, n. 1, p. 19-24, 2008.

DUNN, O. J. Multiple comparisons using rank sums. **Technometrics**, v. 6, n. 3, p. 241-252, 1964.

FRAGA, A. B. *et al.* Multivariate analysis to evaluate genetic groups and production traits of crossbred Holstein × Zebu cows. **Tropical Animal Health and Production**, v. 48, n. 3, p. 533-538, 2016.

GALVÃO JÚNIOR, J. G. B. *et al*. Efeito da produção diária e da ordem de parto na composição físico-química do leite de vacas de raças Zebuínas. **Acta Veterinaria Brasilica**, v. 4, n. 1, p. 25-30, 2010.

GÓMEZ-CORTÉS, P.; JUÁREZ, M.; DE LA FUENTE, M. A. Milk fatty acids and potential health benefits: an updated vision. **Trends in Food Science & Technology**, v. 81, p. 1-9, 2018.

HAYGERT-VELHO, I. M. P. *et al*. Multivariate analysis relating milk production, milk composition, and seasons of the year. **Anais da Academia Brasileira de Ciências**, v. 90, n. 4, p. 3839-3852, 2018.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian Journal of Statistics**, v. 6, n. 2, p. 65-70, 1979.

HUSSON, F.; JOSSE, J.; PAGÉS, J. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? **Technical Report - Agrocampus**, Applied Mathematics Department, 2010.

KAISER, H. F. The application of electronic computers to factor analysis. **Educational and Psychological Measurement**, v. 20, p. 141-151, 1960.

KNOB, D. A. *et al*. Growth, productive performance, and udder health of crossbred Holstein x Simmental cows and purebred Holstein cows. **Semina: Ciências Agrárias,** v. 36, n. 6, p. 2597-2606, 2018.

LUDOVICO, A. *et al.* Losses in milk production and quality due to milk somatic cell count and heat stress of Holsteins cows in temperate climate. **Semina: Ciências Agrárias**, v. 36, n. 5, p. 3455-3470, 2015.

MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations. **Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability**, v. 1, p. 281-297, 1967.

MURTAGH, F.; LEGENDRE, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? **Journal of Classification**, v. **31**, n. 3, p. 274-295, 2014.

PAIVA, A. L. C. *et al*. Principal component analysis in laying hen production traits. **Revista Brasileira de Zootecnia**, v. 39, p. 285-288, 2010.

QI, L.; BRAVO-URETA, B. E.; CABRERA, V. E. From cold to hot: climatic effects and productivity in Wisconsin dairy farms. **Journal Dairy Science,** v. 98, n. 12, p. 8664-8677, 2015.

RIBAS, N. P. *et al*. Escore de células somáticas e sua relação com os componentes do leite em amostras de tanque no estado do Paraná. **Archives of Veterinary Science**, v. 19, n. 3, p. 14-23, 2014.

RIBEIRO, M. J. B. *et al.* Principal components for the in vivo and carcass conformations of Anglo-Nubian crossbred goats. **Ciência Rural**, v. 48, n. 6, e20170771, 2018.

SANTOS, E. F. N. *et al*. Formação de grupos produtivos em vacas leiteiras por meio de componentes principais. **Revista Brasileira Biométrica**, v. 28, n. 3, p. 15-22, 2010.

SANTOS, M. A. S. *et al*. Caracterização do nível tecnológico da pecuária bovina na Amazônia Brasileira. **Revista de Ciências Agrárias: Amazonian Journal of Agricultutal and Environmental Sciences**, v. 60, n. 1, p. 103-111, 2017.

SANTOS, R. O. *et al.* Principal Component Analysis and Factor Analysis: differences and similarities in Nutritional Epidemiology application. **Revista Brasileira de Epidemiologia**, v. 22, e190041, 2019.

SILVA, J. E. *et al*. Effect of somatic cell count on milk yield and milk components in Holstein cows in a semi-arid climate in Brazil. **Revista Brasileira de Saúde e Produção Animal**, v. 19, n. 4, p. 391-402, 2018.

THORNING, T. K. *et al*. Milk and dairy products: good or bad for human health? An assessment of the totality of scientific evidence. **Food & Nutrition Research**, v. 60, 2016.

TREMBLAY, M, *et al*. Customized recommendations for production management clusters of North American automatic milking systems. **Journal of Dairy Science**, v. 99, p. 5671-5680, 2016.

VENTURA, H. T. *et al.* Use of multivariate analysis to evaluate genetic groups of pigs for dry-cured ham production. **Livestock Science**, v. 148, n. 3, p. 214-220, 2012.

This is an open-access article distributed under the terms of the Creative Commons Attribution License