Brazilian Journal of Biometrics, **43**, 1–16 (2025)
doi:10.28951/bjb.v43i4.869

ARTICLE

Clustering calf growth curves using quantile regression and unsupervised learning

Gabriela Maria Rodrigues,* Taciana Villela Savian, and Fábio Prataviera

Department of Exact Sciences, Luiz de Queiroz College of Agriculture, University of Sao Paulo, Piracicaba, SP, Brazil
*Corresponding author. Email: gabrielar@usp.br

(Received: March 31, 2025; Revised: August 15, 2025; Accepted: September 1,2025; Published: December 2, 2025)

Abstract

The study of growth characteristics can be crucial to the profitability of animal and plant production. An important aspect to be considered in this type of modeling is the potential presence of heterogeneous sample variances. The Quantile Regression (QR) methodology does not impose any distributional assumptions on the model error, such as normality or constant variances, making it an interesting alternative to conventional regression models. Additionally, it can provide more information about the relationship between the independent variable and the response by fitting different quantiles. This study analyzes data related to the weights in kilograms of 28 calves over a period of 26 weeks after birth. The objectives were to examine QR as an alternative to conventional methods for growth data, considering asymmetry and heterogeneity of residual variances, and to use it to classify animals into groups with different growth patterns. Furthermore, the clusters obtained by QR are compared with clusters obtained by unsupervised machine learning algorithms, a widely used statistical tool nowadays. QR proved to be a more robust alternative to conventional regression models and provided clustering that competes with unsupervised machine learning algorithms. Therefore, it can be recommended for inference purposes as well as for reference in clusters.

Keywords: Animal science; Quantile regression; K-means; Hierarchical algorithms; Residual analysis.

1. Introduction

The study of growth characteristics is of fundamental importance for both animal and plant production. Understanding and enhancing the growth of living organisms play a crucial role in ensuring food security, optimizing agricultural production, and managing natural resources sustainably. A detailed growth analysis provides valuable information about the health, quality, and productivity

of plants and animals, enabling the adoption of more effective agricultural and zootechnical practices. Furthermore, it contributes to science and research, fueling the development of innovative strategies to meet the growing global demands for food, fibers, and agricultural products. For example, in the field of cattle farming, the performance of growth characteristics can be decisive for the profitability of production (Laureano *et al.,* 2011; Barbosa *et al.,* 2017).

In this context, in order to enhance zootechnical indices, optimize the production system, and meet the economic needs of cattle sales, it becomes important to acquire information on the potential for weight gain and reproductive precocity (de Rezende *et al.,* 2014; Silveira *et al.,* 2014). Particularly, the study of the initial phase is of great significance in identifying growth efficiency.

An important aspect to be considered in modeling growth data is the potential presence of heterogeneous sample variances. In this type of data, heteroscedasticity among measurements is common, as growth is regulated by extrinsic or environmental factors and intrinsic or organic factors. Therefore, as the fruit, plant, or animal develops, the variation in its size and/or weight becomes greater (Fernandes *et al.,* 2014). For instance, in identifying animals with higher growth efficiency, it is necessary to fit regression models using a robust method to the variability of individual weight-age data, capable of producing estimates that accurately represent the precocity of the animals (da Silva *et al.,* 2006).

This paper presents an analysis of calf growth up to 6 months of age. According to the descriptive analysis of the response variable (y_i =Weights) versus the covariate (x_i =Time), presented in Figure 1, the following facts are observed: heterogeneous variances over time; different growth rates among animals, that is, at the end of the study period, animals reach distinct weights and most animals exhibit a quadratic trend in individual profiles, and all show an increase in weight over time.

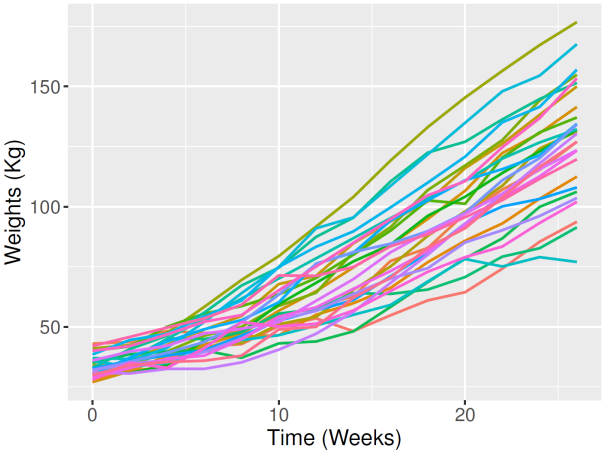

Figure 1. Individual growth profiles throughout the weeks.

In this context, the quantile regression (QR) proposed by Koenker & Bassett Jr (1978) represents an interesting alternative for modeling the type of problem presented above. Unlike typical mean regression models, which focus only on the conditional mean, QR allows the assessment of covariate effects across different points of the conditional distribution of the response variable. QR further enables the characterization of the entire conditional distribution without relying on distributional assumptions such as normality or constant variance, offering a more comprehensive understanding of variability in the response.

Based on quantile regression modeling, it is possible to explore different parts of the conditional distribution, such as the lower tail (e.g., the $\tau = 0.1$ quantile) or the upper tail (e.g., the $\tau = 0.9$

quantile). When researchers require information about specific subpopulations, QR can provide a more comprehensive view of the dataset (Hao & Naiman, 2007; Oliveira & Rios-Neto, 2006). Moreover, this methodology exhibits favorable properties that justify its application: it is robust to the presence of outliers; its models are invariant under monotonic transformations; it is flexible with incomplete data and efficient in parameter estimation, even in cases of asymmetry or heterogeneity of variances. According to Koenker & Bassett Jr (1978), QR is a more robust alternative to the least squares (LS) estimation method.

The initial application of Quantile Regression occurred in econometrics (Buchinsky, 1998; Zietz *et al.,* 2008; Dufrenot *et al.,* 2010; Fitzenberger *et al.,* 2013). Since then, it has emerged as an inferential approach for statistically analyzing linear to nonlinear models across various knowledge domains. Several authors have employed quantile regression in several studies. For instance, in plant growth studies, Muggeo *et al.* (2013) estimated growth curves for the marine alga Posidonia oceanica, Pollice *et al.* (2014) examined the growth of sorghum roots, and Sorrell *et al.* (2012) analyzed the growth of three wetland plant species in response to water depth.

In the context of animal growth, Nascimento *et al.* (2019) identified pigs with different growth rates and proposed a classification of these animals based on their overall Euclidean distance between each observed weight and the weight estimated from quantile regression growth curves.

Some studies have also proposed quantile regression to predict the yield of eucalyptus stands (Carvalho, 2013; Farias *et al.,* 2018). Recently, Puiatti *et al.* (2018) investigated the accumulation of dry matter in garlic plants over time, categorizing different accessions based on their growth rate and asymptotic weight. Additionally, Santos *et al.* (2018) advocates for the use of quantile regression in estimating genomic breeding values for pigs, which have phenotypes with asymmetric distributions.

According to the referenced works, this study employs quantile regression to analyze different calf growth patterns and provide clustering based on these characteristics. The clustering problem involves clustering elements from a dataset based on their characteristics. The clusters are determined to achieve internal homogeneity and external heterogeneity. This means that the most similar elements should be in the same cluster, and the least similar ones in different clusters. There are various clustering methods, many of which fall under unsupervised learning, a subclass of machine learning (ML) that is currently a highly popular topic. In this type of learning, variables are not distinguished between predictors and responses, and the objective is to discover associations and patterns among them. In this context, we are concerned with comparing the clusters generated based on quantile regression with two classes of ML algorithms widely used in the literature: hierarchical and partitioning algorithms.

The present study aimed to: (i) analyze quantile regression models as an alternative to conventional methods for growth data in the presence of residual asymmetry and heterogeneity of variances; (ii) estimate calf growth curves and classify them into groups through quantile regression; and (iii) compare the clusters obtained by quantile regression with those obtained by unsupervised machine learning algorithms.

2. Materials and Methods

2.1 Description of the data

The study of animals, exemplified here by monitoring the weight of 28 calves (in kg) over a 26-week period following birth, plays a fundamental role in understanding and enhancing agricultural production. This research, conducted by the renowned Brazilian Agricultural Research Corporation (EMBRAPA), provides valuable insights for the livestock sector by examining the evolution of calf weights during their initial six months of life.

In this regard, the collected data assumes a central role in guiding cattle management and nutrition practices. Despite the presence of missing values in the biweekly weighings, the frequency of measurements allows a detailed understanding of weight fluctuations in the animals throughout the analyzed period. For this study, the data used were extracted directly from the article by Singer *et al.* (2017). As is well known, animal growth curves are commonly analyzed by nonlinear regression models. Here, as well as in Singer *et al.* (2017), due to the presence of observations only from the initial months of life, the data are analyzed using linear models, as described below.

2.2 Statistical methods

2.2.1 Quantile regression

Quantile regression is an extension of ordinary least squares (OLS) linear regression, but it remains a semiparametric method as it does not necessarily assume a parametric distribution for the error process. In contrast, conventional linear regression focuses on the mean relationship between a set of explanatory variables denoted as x and the dependent variable Y, based on the conditional mean function, $E(Y|x)$. This implies that conditional mean regression analysis is confined to only one part of the conditional distribution, providing limited information about the entire distribution.

This information gap is overcome by the quantile approach, which employs a continuous spectrum of quantile functions to describe the relationship at different points of the conditional distribution of Y. Furthermore, it is essential to recognize that causality in the tails of the distribution may differ from average causality, necessitating the exploration of all quantiles. The quantile framework provides us insights into how extreme values can affect the dependent variable (Hao & Naiman, 2007; Troster *et al.,* 2018).

Let $\boldsymbol{y} = (y_1, ..., y_n)^T$ be the n-dimensional vector of the response variable observation (continuous) and $\boldsymbol{x}_i = (\mathbb{1}, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ $(p+1)$-dimensional vectors of covariates for $i = 1, \ldots, n$. Thus, the function that relates the τ-th quantile of the response variable and the covariates is called quantile regression function, and under the assumption of a linear relationship it can be expressed by

$$Q_\tau(y|\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}(\tau), \tag{1}$$

where $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \ldots, \beta_p(\tau))^T$ is the $(p+1)$-dimensional vector of unknown regression coefficients for τ-th fixed quantile.

One way to represent the quantile regression model Eq. (1) is given by:

$$y_i = \beta_0(\tau) + \sum_{j=1}^{p} x_{ij}\beta_j(\tau) + \varepsilon_i \tag{2}$$

where ε_i is the random error.

The $\boldsymbol{\beta}(\tau)'s$ are estimated by solving the minimization problem

$$\min_{\boldsymbol{\beta}(\tau)\in\mathbb{R}^\rho} \sum_{i=1}^{n} \rho_\tau\left(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau)\right), \tag{3}$$

where $\rho_\tau(w) = \tau\max(w, 0) + (1-\tau)\max(-w, 0)$. The $\rho_\tau(w)$ is referred to as the check loss function and provides the weighted least absolute error (LAE) (Koenker & d'Orey, 1987). For every quantile level represented by τ, solving the minimization problem results in a unique collection of regression coefficients. It's important to observe that when $\tau = 0.5$, this corresponds to median regression, and the absolute value function is given by $2\rho_{0.5}(w)$.

In this work, based on Eq. (2), we consider the following quantile regression structure given by

$$\text{Structure} = \begin{cases} \mathcal{M}_1 : y_i = \beta_0(\tau) + \beta_1(\tau)x_i + \varepsilon_i(\tau), \\ \mathcal{M}_2 : y_i = \beta_0(\tau) + \beta_1(\tau)x_i + \beta_2(\tau)x_i^2 + \varepsilon_i(\tau), \end{cases}$$

where y_i are the weights of the animals, x_i are the weeks of observation, $\beta_0(\tau)$, $\beta_1(\tau)$ and $\beta_2(\tau)$ are the parameters to be estimated at the τ-th quantile, $\mathcal{M}_1$ is the linear quantile regression model and $\mathcal{M}_2$ is the quadratic quantile regression model, respectively.

The asymptotic distribution of $(\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau))$ is given by $N_{(p+1)}(0, I(\boldsymbol{\beta})^{-1})$, where $I(\boldsymbol{\beta})$ is a matrix derived from the asymptotic theory of quantile regression, playing a role analogous to the Fisher information matrix in the classical maximum likelihood framework. This formulation allows us to asymptotically estimate the covariance matrix and construct approximate confidence intervals (C.I.) for the individual parameters (Kocherginsky *et al.,* 2005; Koenker & Bassett Jr, 1978).

The estimators for $\boldsymbol{\beta}(\tau)'s$ are obtained by using the `quantreg` package in `R` software (R Development Core Team, 2020). In addition to the model selection of the appropriate model, it is used the Akaike information criterion (AIC) and the coefficient of determination R^1 (Koenker & Machado, 1999). For R^1, the larger values indicate better fits, and lower AIC values indicate better fits. We also employ the Anowar test for the models proposed by Chen *et al.* (2008) to compare the fit of a restricted model (under the null hypothesis) with the fit of the full model, assessing whether the quantile parameters tested actually contribute to explaining the data.

2.2.2 Model checking

In this paper, we employ a simulated envelope of the randomized quantile residuals, as introduced by Dunn & Smyth (1996), to assess the goodness-of-fit of the proposed model to the dataset. Thus, for the calculation of quantile residuals, we assume the asymmetric Laplace distribution (ALD) as the parametric distribution for the errors ε_i Eq. (2). Consequently, for the ALD distribution, the maximum likelihood estimators (MLE) coincide with the estimators of weighted least absolute error. Therefore, the quantile residual is defined as

$$r_{q,i} = \Phi^{-1}\left(F(y_i|\hat{\mu}_i, \hat{\sigma}, \tau)\right) \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\hat{\mu} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}(\tau)$ and $\hat{\sigma}$ are the MLE's, $\tau \in (0, 1)$ is fixed, and $F(\cdot)$ is the cumulative distribution of ALD is given by

$$F(y|\mu, \sigma, \tau) = \begin{cases} \tau \exp\left(\frac{1-\tau}{\sigma}(y-\tau)\right), & \text{if } y \leq \mu \\ 1 - (1-\tau)\exp\left(-\frac{\tau}{\sigma}(y-\mu)\right), & \text{if } y > \mu \end{cases} \tag{5}$$

Therefore, this distribution came to be employed as the error distribution in the quantile regression model, enabling a more comprehensive inferential approach and the incorporation of more sophisticated model structures (Yu & Moyeed, 2001; Yu & Zhang, 2005; Yuan & Yin, 2010; Hinostroza, 2017; Santos & Bolfarine, 2018; Li *et al.,* 2010; Reich *et al.,* 2010; Geraci & Bottai, 2007; Geraci & Bottai, 2014; Morales, 2015).

2.2.3 Clustering by quantile regression curves

After fitting the model, animals were categorized into groups based on their growth dynamics. To capture not only the central tendency but also the variability and extreme patterns in weight change over time, we employed quantile regression. This approach allows modeling different quantiles of the response distribution, providing a more complete characterization of the growth trajectories compared to mean regression alone. Specifically, we considered quantiles representing the

central (e.g., median) and extreme positions (e.g., lower and upper quartiles) of the data distribution. This enables the detection of both typical growth patterns and atypical behaviors, such as calves that gained weight more slowly or more rapidly than the majority, or those that ended the study period with particularly low or high final weights.

Thus, we consider as a dissimilarity measure for classification, the Euclidean distance between the average rate of change ($\widehat{ARC}_{i(\tau)}$) of each curve obtained through quantile regression for the τ-th quantile, and the observed average rate of change (ARC_i) in the weights of each calf, defined as

$$ARC_i = \frac{\text{Final weight}_i - \text{Initial weight}_i}{\text{time period}_i}. \tag{6}$$

The dissimilarity measure is then given by:

$$d_{i(\tau)} = \sqrt{(\widehat{ARC}_{i(\tau)} - ARC_i)^2} = |\widehat{ARC}_{i(\tau)} - ARC_i|. \tag{7}$$

In this way, the animal is classified as belonging to the group corresponding to the smallest obtained distance.

2.3 Clustering by algorithms

We compared the groups obtained from quantile regression curves using two unsupervised machine learning algorithms: K-means and Hierarchical Clustering. Next, we provide a brief description of the K-means and Hierarchical Clustering methods, respectively.

- **K-means**: The *k*-means algorithm belongs to the class of partitioning algorithms and requires the space to be Euclidean. The most widely used version of this algorithm was proposed by Hartigan & Wong (1979). It provides *k* clusters for the dataset, such that the sum of the squares of the errors' distances to the cluster centers is minimized.
- **Hierarchical clustering (agglomerative)**: Consider *n* elements and construct a matrix with the distances between all pairs (distance matrix). Typically, this matrix is symmetric with zeros on the diagonal. Usually, the number of groups is defined based on the partitions constructed by the algorithms. Then, the types of clusters can be the following (Johnson, Wichern, *et al.,* 2002):
  - **Single linkage clustering**: In this method, the distance between two groups is defined as the shortest distance between any pair of points belonging to different groups. In other words, it measures the proximity between groups based on the two closest points among them. This method tends to create elongated clusters and can be sensitive to outliers, as a single distant point can significantly impact the cluster structure.
  - **Complete linkage clustering**: This method assesses the dissimilarity between groups by considering the maximum distance between any two data points in different groups. In other words, it measures the similarity between groups based on the most distant data points. Complete-linkage clustering tends to create compact and spherical groups, which can be advantageous for certain types of data. However, it is sensitive to outliers and may lead to the "chaining effect," where some data points are grouped together simply because they are relatively close to each other.
  - **Average linkage clustering**: This method calculates the average pairwise distances between all data points in two clusters when merging them. It is an agglomerative approach, starting with each data point as a separate cluster and iteratively merging clusters with the

smallest average distance until forming a single cluster containing all data points. Average Linkage Clustering tends to create clusters with relatively uniform sizes and is sensitive to outliers.

In R software, the algorithms can be implemented using the following functions from the base package: to obtain the distances `hclust`; for k-means `kmeans`; for single linkage clustering`"single"`; complete linkage clustering `"complete"`; for average linkage clustering `"average"`. For all algorithms, the metric used was the Euclidean distance.

2.4 Silhouette Score

To assess the results obtained from the clustering, we employed the Silhouette coefficient, which is an evaluation metric for clustering that quantifies the quality of separation between clusters in a clustering analysis *Silhouette coefficient*. It ranges from -1 to 1 and is calculated for each data point i as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{8}$$

where a_i is the average distance between observation i and the other observations within the same group, and b_i is the average distance between observation i and the other observations within the nearest group.

Therefore, we interpret the potential values of s_i as follows: a value close to 1 indicates that the point is well separated from other clusters and, therefore, belongs to the appropriate cluster; a value close to zero suggests that the point is near the boundary between two clusters; and a value close to -1 indicates that the point may have been assigned to the wrong cluster.

Thus, the Silhouette Score is calculated by taking the average of the coefficients for each observation across the entire dataset, and it serves as an indicator of the quality of the generated clusters. The closer to 1, the higher the accuracy in forming the groups.

3. Results and Discussion

The Table 1 presents a descriptive analysis of calf weights over weeks. It can be observed that the variation among individuals increases over time and is substantially lower at the beginning of the study, allowing us to conclude that there is heterogeneity of variances. The coefficients of variation also indicate lower variability in the initial phase. Regarding skewness coefficients, it is noted that some weeks exhibit a slight positive skewness, meaning there is a greater concentration of observations on the left side of the distribution. These analyses support the illustration in Figure 1 presented in the introduction of this study.

Table 1. Means, standard deviations (SD), asymmetry coefficients and coefficients of variation (CV) of calf weights (in kilograms) for each week of observation

Weeks	Mean	SD	Skewness	CV%
0	33,74	4,57	0,56	13,54
2	36,73	4,41	0,53	12,01
4	40,38	5,34	0,20	13,22
6	45,45	6,67	0,23	14,68
8	50,47	8,83	0,42	17,49
10	58,11	10,50	0,43	18,08
12	64,48	13,80	0,56	21,40
14	71,45	14,86	0,40	20,79
16	79,92	16,39	0,57	20,51
18	90,05	17,48	0,48	19,41
20	98,87	18,27	0,37	18,48
22	109,87	20,89	0,20	19,02
24	119,16	21,68	0,06	18,19
26	128,55	23,73	-0,10	18,46

In Table 2, the following measures are presented: the p-values of the anowar test, the coefficient of determination, and the Akaike Information Criterion (AIC) for quantile regression models $\mathcal{M}1$ and $\mathcal{M}2$ at various quantiles. It is observed that the quadratic effect (model $\mathcal{M}_2$) was significant at all quantiles and also exhibited the lowest AIC values, indicating that this is the most suitable model.

Table 2. P-value of the anowar test, coefficient of determination R^1 and Akaike information criterion (AIC) for the quantile regression models at different quantiles

	P-value	R^1		AIC	
		$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_2$
$\tau = 0,10$	0,007	0,47	0,49	3122,75	3086,58
$\tau = 0,20$	<0,001	0,51	0,54	3089,68	3040,60
$\tau = 0,30$	<0,001	0,54	0,58	3084,87	3019,10
$\tau = 0,40$	<0,001	0,56	0,61	3084,66	3006,11
$\tau = 0,50$	<0,001	0,59	0,62	3081,26	3016,80
$\tau = 0,60$	<0,001	0,60	0,63	3093,99	3041,89
$\tau = 0,70$	<0,001	0,61	0,63	3117,41	3078,02
$\tau = 0,80$	<0,001	0,61	0,63	3165,79	3117,13
$\tau = 0,90$	<0,001	0,60	0,62	3257,03	3212,73

This way, we proceed with the analyses considering the model $\mathcal{M}2$. Thus, the effects of the variable x_i are examined across different points of the y_i variable distribution in Figure 2. It is observed that the parameter estimates β_0 and β_1 increase across quantiles, indicating that the effects of the "week" variable vary at different quantiles. For the parameter β_2, an approximately constant behavior is observed. Confidence intervals (CIs) have slightly smaller amplitudes at quantiles below $\tau = 0.50$, suggesting that the conditional distribution of y_i as a function of x_i exhibits a slight right skewness. It can be concluded that the x_i and x_i^2 are statistically significant for all quantiles.

The quantile residuals for some quantiles are analyzed in Figure 6 in Appendix 4. From the 0.50 quantile onwards, many points are observed outside the confidence bands, indicating that these

models do not have an adequate fit. This fact shows that the conditional distribution of weight as a function of weeks presents a small positive asymmetry, corroborating the results above.

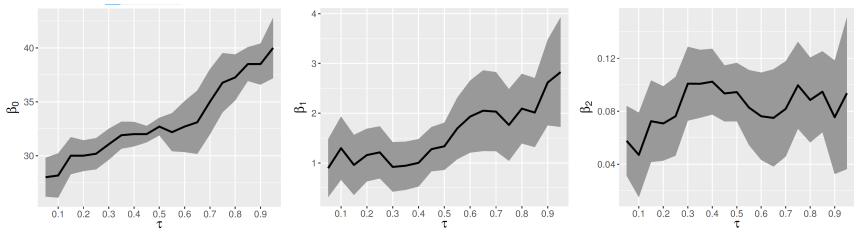

Figure 2. Parameter estimates and respective confidence intervals, considering a significance level of 5%, along quantiles equally spaced between [0.05, 0.95].

After selecting the model (model $\mathcal{M}_2$), the next step involves conducting a clustering of the animals based on quantile regression curves. It is important to emphasize that the number of groups, as well as the quantiles to be used as references, can be arbitrarily chosen by the researcher according to the research interests. In this study, we employed the Silhouette Score validation measure (Section 2.4) to choose three potential quantiles. Additionally, we examined these values for groups based on the K-means and Hierarchical clustering algorithms.

In Table 3, the total average Silhouette Score of clustering ($\bar{s}_i\cdot$), its corresponding total quantity of negative s_i's obtained, and the average value for each group ($\bar{s}_i$) are presented. It can be observed that the clusters based on the quantiles $\tau = 0.10, 0.45, 0.90$ and the K-means algorithm were the only ones that did not obtain negative s_i's. This result indicates that these clusters correctly allocated the animals. Note that some clusters obtained lower s_i's, indicating dissimilarities among individuals within the group, as characterized by negative s_i's.

Table 3. Average silhouette coefficient (s_i) for each of the clusters and total number of negative s_i's for the clustering based on quantile regression curves and k-means and hierarchical algorithms

		Amount of	Cluster 1		Cluster 2		Cluster 3	
			Size	$\bar{s}_i$	Size	$\bar{s}_i$	Size	$\bar{s}_i$
$\tau = 0,25; 0,50; 0,75$	0,56	2	11	0,37	9	0,82	8	0,55
$\tau = 0,10; 0,50; 0,90$	0,58	1	9	0,45	11	0,67	8	0,59
$\tau = 0,25; 0,45; , 0,75$	0,56	2	11	0,37	9	0,82	8	0,55
$\tau = 0,10; 0,45; 0,90$	0,57	0	8	0,49	12	0,60	8	0,61
$\tau = 0,25; 0,55; 0,75$	0,56	2	1	0,37	9	0,82	8	0,55
$\tau = 0,10; 0,55; 0,90$	0,58	1	9	0,45	11	0,67	8	0,59
K-means	0,57	0	8	0,49	12	0,60	8	0,61
Complete linkage	0,53	1	18	0,42	2	0,78	8	0,70
Average linkage	0,35	5	24	0,26	2	0,82	2	0,96
Single linkage	0,35	5	24	0,26	2	0,82	2	0,96

In this way, we proceed with the analysis considering the quantiles $\tau = 0.10, 0.45$, and 0.90 for the model $\mathcal{M}_2$. The groups obtained through the quantile regression curves were identical to those obtained by the K-means algorithms. The Single Linkage and Average Linkage algorithms also produced identical clusters. In this context, we present in Figure 3 the groups obtained through the quantile regression curves and by the complete linkage and single linkage algorithms.

We observed in 3 the significant portion of coefficients for each observation above the mean coefficient. On the other hand, in 3b-c, we can verify the animals that were not allocated to the correct groups. The dendrograms resulting from the hierarchical groupings can be visualized in Figure 4.

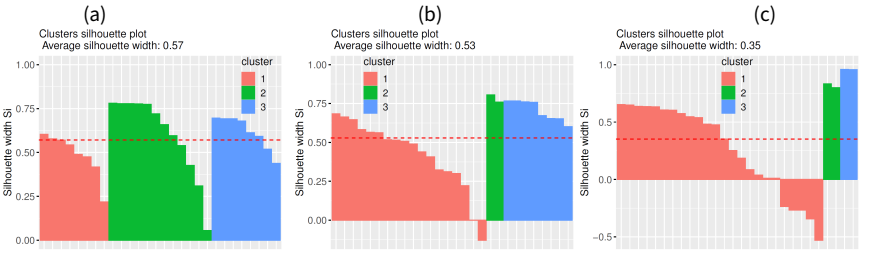

Figure 3. Silhouette analysis for generated clusters: (a) quantile regression curves and k – *means* algorithm; (b) Complete linkage algorithm and (c) Single linkage algorithm.

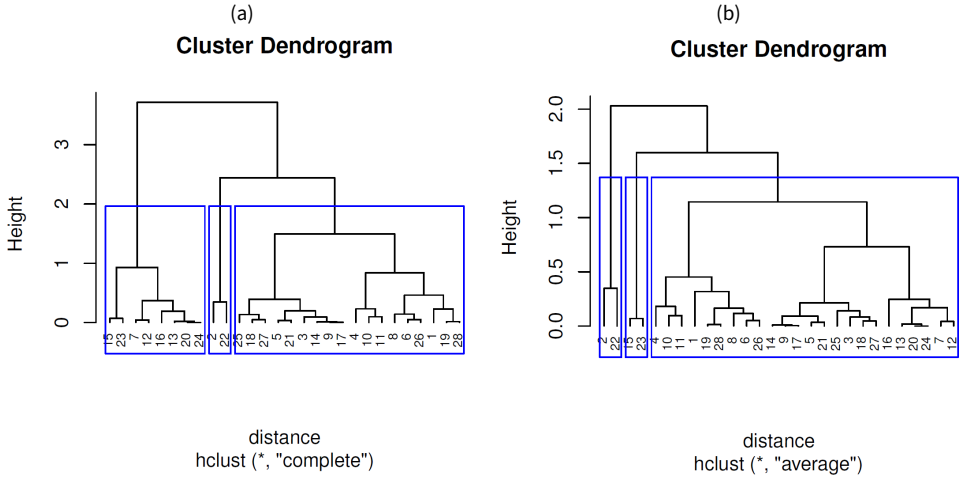

Figure 4. Dendrogram of hierarchical algorithms: (a) Complete linkage algorithm and (b) Single linkage algorithm.

We can conclude that quantile regression and the k-means algorithm achieved appropriate classifications for this dataset; thus, some results from this clustering are provided in Table 4. The initial mean weights were 33.39, 33.82, and 34.17 kg for quantiles τ = 0.10, 0.45, and 0.90, respectively. It is noteworthy that these values are close, which is justified by the substantially lower variance in the initial weeks of observation. The final mean weights for each group, corresponding to six months of age, are as follows: 100.2, 128.75, and 156.58 kg for quantiles τ = 0.10, 0.45, and 0.90, respectively. In this case, differences in weights are observed, due to the greater variation at the end of the observed time.

The Figure 5 displays the fitted QR curves for the three quantiles and individual growth curves according to the groups. It is observed that animals transition between quantiles until approximately the 20th week. The average rates of change for the fitted curves are 2.52, 3.71, and 4.58 for the quantiles τ = 0.10, 0.45, and 0.90, respectively. These results indicate that animals exhibit different and increasing growth variations across quantiles over the studied time interval (0 to 26 weeks), in

accordance with the quantile estimates plotted in Figure 2.

Finally, it is concluded that the quantiles were satisfactorily selected to describe this dataset. In this sense, quantile regression curves can be appropriate references for identifying animals with different growth potentials. For instance, at the lower quantile ($\tau = 0.10$), animals presented lower body weights over time, while at the higher quantile ($\tau = 0.90$), animals presented higher weights. Measures related to inference from quantile regression models are provided in Table 5.

Table 4. Number of individuals classified in each subgroup based on quantile τ = (0.10,0.45 and 0.90) and their descriptive summary of weights in the initial and final weeks

Individuals	Initial week					
	Min.	1º Qu.	Median	Mean	3º Qu.	Max.
8 ($\tau = 0,10$)	28,00	30,00	32,70	33,39	35,50	42,00
12 ($\tau = 0,45$)	29,00	30,50	31,00	33,82	36,50	43,00
8 ($\tau = 0,90$)	27,00	31,42	32,50	34,17	38,88	41,00
Individuals	Final week					
	Min.	1º Qu.	Median	Mean	3º Qu.	Max.
8 ($\tau = 0,10$)	77,00	93,12	102,80	100,20	106,65	119,70
12 ($\tau = 0,45$)	112,50	126,12	130,85	128,75	132,70	137,10
8 ($\tau = 0,90$)	141,50	151,125	154,15	156,58	159,65	176,80

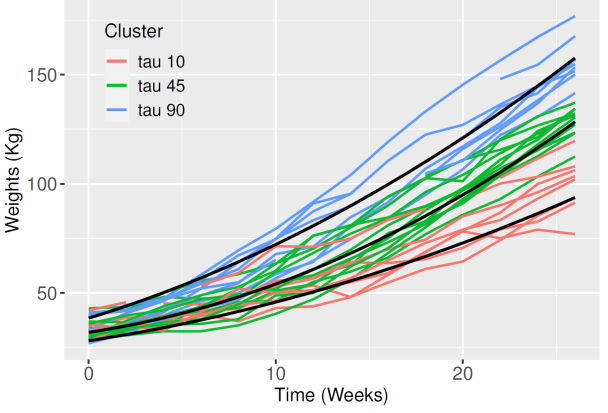

Figure 5. Individual growth curves over the weeks according to classification.

Table 5. Parameter estimates for quantile regression models at $\tau = 0.10, 0.45$ and 0.90, and confidence intervals considering a 95% level of significance level

		Estimates	Standard error	Pr(> |t|)	95% C.I.
$\tau = 0,10$	β_0	28,16	1,05	< 0,01	[26, 09; 30, 22]
	β_1	1,30	0,33	< 0,01	[0, 65; 1, 93]
	β_2	0,05	0,02	< 0,01	[0, 01; 0, 07]
$\tau = 0,45$	β_0	32,00	0,39	< 0,01	[31, 23; 32, 766]
	β_1	1,28	0,23	< 0,01	[0, 83; 1, 72]
	β_2	0,09	0,01	< 0,01	[0, 07; 0, 11]
$\tau = 0,90$	β_0	38,50	0,98	< 0,01	[36, 58; 40, 41]
	β_1	2,62	0,44	< 0,01	[1, 75; 3, 47]
	β_2	0,08	0,02	< 0,01	[0, 03; 0, 11]

4. Conclusions

This study employs quantile regression to analyze various patterns of calf growth and to provide clustering based on these characteristics. Additionally, we perform clustering using the K-means and hierarchical clustering algorithms for comparison with the proposed methodology.

The weight of the animals exhibits heterogeneous variances across weeks and slight right asymmetry. The studied variables show different relationships across quantiles, meaning that animals display different growth patterns over time. This fact justifies the use of quantile regression to analyze such relationships. Analyzing only the mean, as in usual regressions, can ignore important considerations regarding the dataset.

Quantile regression curves and the K-means algorithm produced the best and identical clustering. Hierarchical algorithms did not produce satisfactory results. We observed different growth rates for the three proposed groups. Homogeneity of variances would likely result in approximate rates of variations.

The method to be used must consider the research objective. The k-means algorithm provided the same clustering as the QR in a straightforward manner. However, to describe weight as a function of time and to understand how time affects the variation and asymmetry of calf weights, the use of regression becomes necessary. Therefore, for inferential purposes, the use of quantile regression is recommended. For instance, QR curves can serve as a reference to identify animals with distinct growth patterns, such as lower and higher body weights over time. Additionally, the silhouette coefficient proved to be suitable for selecting the quantiles to be fitted, and it can be employed in future research.

It is concluded that the QR proved to be a more robust alternative to conventional regression models and provided competitive clustering compared to unsupervised machine learning algorithms. Finally, this methodology can be applied to other growth studies in general, involving other animals or plants, as well as to describe data with high variability or asymmetry.

Acknowledgments

The financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES) is gratefully acknowledged. We also thank reviewers and editors for their comments.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Data curation**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Formal analysis**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Funding acquisition**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Investigation**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Methodology**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Project administration**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Software**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Resources**:RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Supervision**:RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Validation**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.G; **Visualization**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Writing - original draft**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F.; **Writing - review and editing**: RODRIGUES, G. M., SAVIAN, T. V., PRATAVIERA, F..

References

1. Barbosa, A., Carneiro, P., Rezende, M., Ramos, I., Martins Filho, R & Malhado, C. M. Parâmetros genéticos para características de crescimento e reprodutivas em bovinos Nelore no Brasil. *Archivos de zootecnia* **66,** 449–452. doi:https://doi.org/10.21071/az.v66i255.2523 (2017).
2. Buchinsky, M. The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of applied econometrics* **13,** 1–30. doi:https://doi.org/10.1002/(SICI)1099-1255(199801/02)13:1<1::AID-JAE474>3.0.CO;2-A (1998).
3. Carvalho, S. d. P. C. *Estimativa volumétrica por modelo misto e tecnologia laser aerotransportado em plantios clonais de Eucalyptus sp* PhD thesis (Universidade de São Paulo, 2013).
4. Chen, K., Ying, Z., Zhang, H. & Zhao, L. Analysis of least absolute deviation. *Biometrika* **95,** 107–122 (2008).
5. Da Silva, N. A. M., de Aquino, L. H., Fonseca, F., Muniz, J. A., *et al.* Estudo de parâmetros de crescimento de bezerros Nelore por meio de um modelo de regressão linear: uma abordagem Bayesiana. *Ciência Animal Brasileira* **7,** 57–65 (2006).
6. De Rezende, M., da Silveira, M., da Silva, R., da Silva, L., Gondo, A, Ramires, G., de Souza, J., *et al.* Pre and post weaning weight gain in Nellore cattle raised in the Pantanal, Mato Grosso do Sul, Brazil. *Ciência Animal* **24,** 20–27 (2014).
7. Dufrenot, G., Mignon, V. & Tsangarides, C. The trade-growth nexus in the developing countries: A quantile regression approach. *Review of World Economics* **146,** 731–761. doi:https://doi.org/10.1007/s10290-010-0067-5 (2010).
8. Dunn, P. K. & Smyth, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5,** 236–244. doi:10.1080/10618600.1996.10474708 (1996).
9. Farias, A. A. *et al.* Uso de regressão quantílica na predição da produção de povoamentos de eucalipto (2018).
10. Fernandes, T. J., Pereira, A. A., Muniz, J. A. & Savian, T. V. Seleção de modelos não lineares para a descrição das curvas de crescimento do fruto do cafeeiro (2014).
11. Fitzenberger, B., Koenker, R. & Machado, J. A. *Economic applications of quantile regression* doi:https://doi.org/10.1007/s00181-021-02186-1 (Springer Science & Business Media, 2013).
12. Geraci, M. & Bottai, M. Linear quantile mixed models. *Statistics and computing* **24,** 461–479. doi:https://doi.org/10.1007/s11222-013-9381-9 (2014).

13. Geraci, M. & Bottai, M. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8,** 140–154. doi:https://doi.org/10.1093/biostatistics/kxj039 (2007).

14. Hao, L & Naiman, D. *Quantile Regression, Sage Publication* 2007.

15. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* **28,** 100–108. doi:https://doi.org/10.2307/2346830 (1979).

16. Hinostroza, A. A. A. *Regressão quantílica bayesiana em modelos de fronteira de produção estocástica* PhD thesis (Universidade Federal de Rio de Janeiro, 2017).

17. Johnson, R. A., Wichern, D. W., *et al.* Applied multivariate statistical analysis (2002).

18. Kocherginsky, M., He, X. & Mu, Y. Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics* **14,** 41–55. doi:https://doi.org/10.1198/106186005X27563 (2005).

19. Koenker, R. & Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society,* 33–50. doi:https://doi.org/10.2307/1913643 (1978).

20. Koenker, R. & Machado, J. A. Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association* **94,** 1296–1310. doi:https://doi.org/10.1080/01621459.1999.10473882 (1999).

21. Koenker, R. W. & d'Orey, V. Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **36,** 383–393. doi:https://doi.org/10.2307/2347802 (1987).

22. Laureano, M., Boligon, A., Costa, R., Forni, S, Severo, J. & Albuquerque, L. G. d. Estimativas de herdabilidade e tendências genéticas para características de crescimento e reprodutivas em bovinos da raça Nelore: Estimates of heritability and genetic trends for growth and reproduction traits in Nelore cattle. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia* **63,** 143–152. doi:10.1590/S0102-09352011000100022 (2011).

23. Li, Q., Xi, R., Lin, N., *et al.* Bayesian regularized quantile regression. *Bayesian Analysis* **5,** 533–556. doi:10.1214/10-BA521 (2010).

24. Morales, C. E. G. Quantile Regression for Mixed-Effects Models (2015).

25. Muggeo, V. M., Sciandra, M., Tomasello, A. & Calvo, S. Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and ecological statistics* **20,** 519–531. doi:https://doi.org/10.1007/s10651-012-0232-1 (2013).

26. Nascimento, M, Nascimento, A., Dekkers, J. & Serão, N. Using quantile regression methodology to evaluate changes in the shape of growth curves in pigs selected for increased feed efficiency based on residual feed intake. *Animal* **13,** 1009–1019. doi:https://doi.org/10.1017/S1751731118002616 (2019).

27. Oliveira, A. M. H. C. d. & Rios-Neto, E. L. G. Tendências da desigualdade salarial para coortes de mulheres brancas e negras no Brasil. *Estudos Econômicos (São Paulo)* **36,** 205–236. doi:https://doi.org/10.1590/S0101-41612006000200001 (2006).

28. Pollice, A., Muggeo, V. M., Torretta, F., Bochicchio, R. & Amato, M. *Growth curves of sorghum roots via quantile regression with P-splines* in *47th Scientific Meeting of the Italian Statistical Society* (2014).

29. Puiatti, G. A., Cecon, P. R., Nascimento, M., Nascimento, A. C. C., Carneiro, A. P. S., Puiatti, M., Oliveira, A. C. R. d., *et al.* Quantile regression of nonlinear models to describe different levels of dry matter accumulation in garlic plants. *Ciência Rural* **48.** doi:https://doi.org/10.1590/0103-8478cr20170322 (2018).

30. R Development Core Team. *R: A Language and Environment for Statistical Computing* ISBN 3-900051-07-0. R Foundation for Statistical Computing (Vienna, Austria, 2020). http://www.R-project.org.

31. Reich, B. J., Bondell, H. D. & Wang, H. J. Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics* **11,** 337–352. doi:https://doi.org/10.1093/biostatistics/kxp049 (2010).

32. Santos, B. & Bolfarine, H. Bayesian quantile regression analysis for continuous data with a discrete component at zero. *Statistical Modelling* **18,** 73–93. doi:https://doi.org/10.1177/1471082X17719633 (2018).

33. Santos, P. M. d., Nascimento, A. C. C., Nascimento, M., Azevedo, C. F., Mota, R. R., Guimarães, S. E. F., Lopes, P. S., *et al.* Use of regularized quantile regression to predict the genetic merit of pigs for asymmetric carcass traits. *Pesquisa Agropecuária Brasileira* **53,** 1011–1017. doi:https://doi.org/10.1590/S0100-204X2018000900004 (2018).

34. Silveira, M., Souza, J. d., Silva, L., Freitas, J., Gondo, A & Ferraz Filho, P. Interação genótipo x ambiente sobre características produtivas e reprodutivas de fêmeas Nelore. *Archivos de zootecnia* **63,** 223–226. doi:https://dx.doi.org/10.4321/S0004-05922014000100026 (2014).

35. Singer, J. M., Rocha, F. M. & Nobre, J. S. Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review* **85,** 290–324. doi:https://doi.org/10.1111/insr.12178 (2017).

36. Sorrell, B. K., Tanner, C. C. & Brix, H. Regression analysis of growth responses to water depth in three wetland plant species. *AoB Plants* **2012.** doi:https://doi.org/10.1093/aobpla/pls043 (2012).

37. Troster, V., Shahbaz, M. & Uddin, G. S. Renewable energy, oil prices, and economic activity: A Granger-causality in quantiles analysis. *Energy Economics* **70,** 440–452. doi:https://doi.org/10.1016/j.eneco.2018.01.029 (2018).

38. Yu, K. & Moyeed, R. A. Bayesian quantile regression. *Statistics & Probability Letters* **54,** 437–447. doi:https://doi.org/10.1016/S0167-7152(01)00124-9 (2001).

39. Yu, K. & Zhang, J. A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics—Theory and Methods* **34,** 1867–1879. doi:https://doi.org/10.1080/03610920500199018 (2005).

40. Yuan, Y. & Yin, G. Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics* **66,** 105–114. doi:https://doi.org/10.1111/j.1541-0420.2009.01269.x (2010).

41. Zietz, J., Zietz, E. N. & Sirmans, G. S. Determinants of house prices: a quantile regression approach. *The Journal of Real Estate Finance and Economics* **37,** 317–333. doi:https://doi.org/10.1007/s11146-007-9053-7 (2008).

Appendix

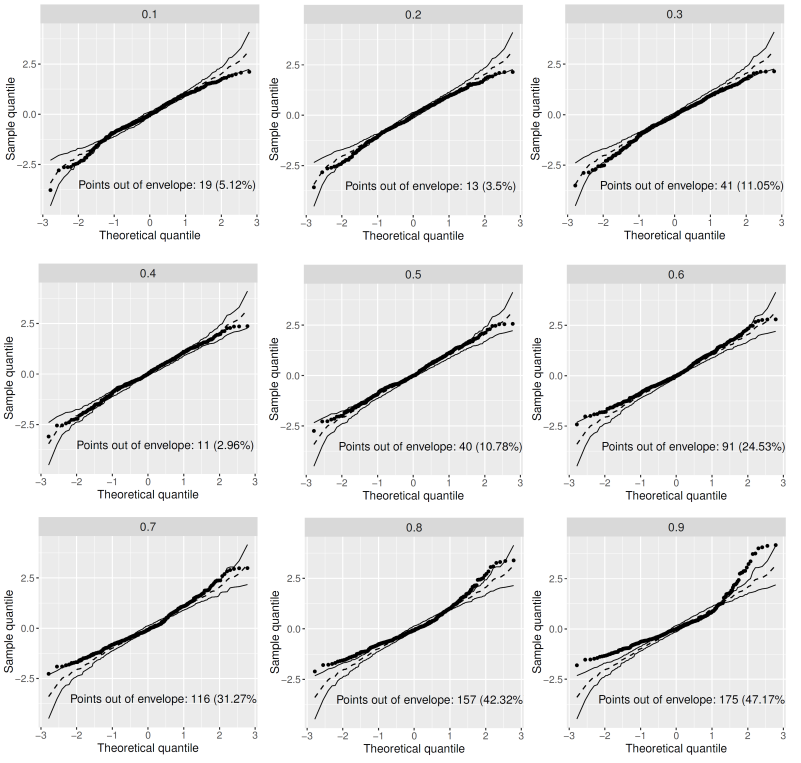

Figure 6. Normal probability plot of quantile residuals with simulated envelope of quantile regression models for equally spaced quantiles in the interval [0.10; 0.90].