agriculture

MDPI

Article

A Vector Representation of Lactation Curves for Dairy Cows

Seonghun Lee and Jaehwa Park *

Department of Software, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea; gnstjdok@cau.ac.kr
* Correspondence: jaehwa@cau.ac.kr

Abstract: Machine learning techniques provide efficient data analysis tools without mathematical derivations. Data-centric LC representations are highly demanded to use these tools for LC-related research. A novel data-oriented LC representation model using piecewise linear regression (PWLR) is presented. This representation is intended to be used directly as data for machine learning along with other associated data at an individual base. An LC is represented in vector form as a series of connected line segments and the location and number of segments are determined by the maximum residual. The critical points are determined at the rapid transit point in the LC. The Bayesian information criterion was used to choose the proper number of line segments to avoid the overfitting problem. To demonstrate the validity of the PWLR model as an LC descriptor, its approximation accuracy and representation generality were tested experimentally. The results revealed that the PWLR model is advantageous for representing the LCs of an individual or a large herd that are directly applicable to data-driven approaches.

Keywords: lactation curve; piecewise linear regression; vector representation

check for updates

Citation: Lee, S.; Park, J. A Vector Representation of Lactation Curves for Dairy Cows. *Agriculture* **2022**, *12*, 395. https://doi.org/10.3390/agriculture12030395

Academic Editor: Tomas Norton

Received: 21 January 2022
Accepted: 9 March 2022
Published: 11 March 2022

1. Introduction

The lactation curve (LC) is a periodic record of daily milk production in dairy cows for a given time period. The value of milk yield per day is usually collected during the birth cycle from delivery to dry off. It helps estimate the total milk yield of a farm and is used as primary information to monitor the health conditions of individuals.

Many LC modeling studies have been conducted. Cunha et al. [1] compared various empirical and mechanistic models to test the fitting performance of various LC models. Models described by Dijkstra et al. [2], Wood [3], and Wilmink [4] displayed good fitting performance for the high, medium, and low milk production groups. Hossein-Zadeh [5] showed the efficiency of the models of Wood, Dijkstra, and Rook [6] in modeling productivity with a large dataset.

The LC models in previous studies were derived based on the one model fits all LCs concept. Early models aimed to describe general lactation patterns based on mathematical functions with several control parameters. The model represents the LC characteristics for the entire group. The LCs are generated as the regression results of the parameter estimation using the least square error approach. A typical shape, called the standard pattern, is a convex curve that has a vertex with rapid transit from the delivery day and a relatively slow drop after the vertex until the day of dry off. Conventional LC models provide statistically optimal solutions. The models may be one of the basic references for controlling mass milk production for large groups, such as a nation.

It was found that, for individual LCs, not all cases follow conventional LC models. Typical patterns were detected in most of the cases [3]. Some atypical patterns, such as a plat without a significant vertex, a valley with quick drop and recovery, or a convex shape but with jagged lines, were also detected in the remaining cases [7]. Abnormal breeding circumstances, such as sickness, relocation, or changes in feedstuff, generate diverse patterns in LCs [8]. However, not all atypical LCs were included in these cases. Some of these may have originated from individual biological characteristics. Atypical cases are not the majority. However, they are not trivial minorities to be ignored as outliers in regression-based LC modeling.

To overcome the limitations of the early models, non-parametric approaches have been proposed [9,10]. More general mathematical functions not specially designed to fit the standard LC are used as the primary structure. Orthogonal-polynomials or regression splines are typical tools. In orthogonal-polynomial-based approaches, an LC is formulated in a linear combination structure with control coefficients. In regression spline models, an LC is represented in a set of segmented splines joined at points called knots. Linear, quadratic, and cubic polynomials were used to describe the segmented splines. The segmented spline is found using regressions in each intervals between the two knots. The increase in the complexity of the regression and higher-order splines with a large number of knots results in a high computational burden.

Recently, IoT technology has enabled the collection of large data corpora from the terminal data-acquisition points. Livestock is not an exceptional subject, and various smart farming applications have been developed. Large-sized data can be easily collected owing to the diffusion of automatic milking and breeding systems [11]. In addition, various types of data, which could be external or internal clinical factors, such as temperature, feeding amount, motion quantity, and rumination time can be collected with various types of sensory devices on an individual basis. Individual identification with RFID tags allows huge heterogeneous data collection, conserving the correlation among the data fields with synchronized time marks. An increase in data size and dimension with high association on the individual allows the research of modeling from average patterns to individual deviations [12,13].

Machine learning algorithms make it possible to obtain the desired information or classification system, unlike conventional modeling, which uses fixed mathematical functions. Artificial neural networks (ANNs), recently inherited advanced as convolutional neural networks (CNNs) or deep learning, have been proven to be very successful for LC-related research [14,15]. The learning phase of the algorithms is derived from regression analysis; however, the primitive processing architectures are iterative bottom-up structures from individual base datasets. To feed the ANNs for the learning phase, large preprocessed datasets that preserve the association among data fields for individual specific cows, usually called feature vector sets, are essential.

In order to use data-driven machine learning methods for LC-related research, a data-oriented LC representation is essential. The intrinsic LC pattern also needs to be well preserved in the representation to analyze correlations with other associated data. A novel LC model based on piecewise linear regression (PWLR) is presented. An LC is represented in feature vector form as a series of connected line segments in the PWLR model, and the location and number of segments are determined by the maximum residual. The critical points are determined at the rapid transit point in the LC. Linear regression is performed for each divided sub-period of the LC between two critical points. If the regression error for the sub-period is larger than a predetermined threshold, a new critical point is added, and the regression process is repeated until the termination condition is satisfied. This model expands the LC representation ability up to cases that do not follow conventional LC patterns and can provide feature vectors of LCs to employ machine learning techniques directly.

2. Materials and Methods

2.1. Conventional Regression Model

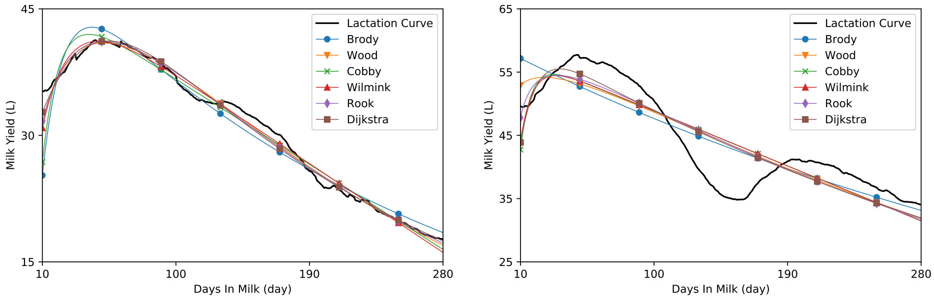
Table 1 lists well-known LC models [1,8]. The models of Brody [16], Wood [3], Cobby [17], Wilmink [4], and Rook [6] are derived mathematically from empirical observations. The Dijkstra [2] model is a mechanistic model that considers both biological and physiological characteristics of the mammary gland. These models were designed with the underlying assumption that the basic LC shapes are convex with a vertex [18]. These models exhibited good approximation performance for typical LC patterns, although there were minor differences among them.

Table 1. Conventional lactation curve models.

Model	Function for LC	
Brody (1924)	$y = a \cdot e^{-b \cdot t} - a \cdot e^{-c \cdot t}$	[16]
Wood (1967)	$y = a \cdot t^{b} \cdot e^{-c \cdot t}$	[3]
Cobby (1978)	$y = a - b \cdot t - a \cdot e^{-c \cdot t}$	[17]
Wilmink (1987)	$y = a + b \cdot e^{-k \cdot t} + c \cdot t$	[4]
Rook (1993)	$y = a \cdot \left[\frac{1}{1 + \frac{b}{c+t}} \right] \cdot e^{-d \cdot t}$	[6]
Dijkstra (1997)	$y = a \cdot e^{b \cdot \frac{1-e^{-c \cdot t}}{c} - d \cdot t}$	[2]

a, b, c, d, and k are model parameters and t is time variable.

Figure 1 shows an example of the actual LC data and the fitting results using the LC models of Table 1. As shown in Figure 1a, all the models showed good fitting performance for the typical LC pattern. However, the approximation performance is poor for the atypical case shown in Figure 1b. This undesirable phenomenon results in a novel representation that can handle such LCs.



(**a**) A typical convex shape

(**b**) An atypical shape

Figure 1. Approximated LCs of the models listed in Table 1.

2.2. Piecewise Linear Regression Representation

An actual LC is usually a data sequence of daily milk yield for a given lactation period. An LC can be represented as a vector, denoted as $\mathbf{Y}$, as:

$$\mathbf{Y} = [y_0, y_1, \ldots, y_{n-1}], \tag{1}$$

where n is lactation period in days.

Piecewise linear regression (PWLR) is a non-parametric approach compared to conventional regression models [19,20]. It depicts an LC as a vector of critical points. An LC is divided into a set of connected line segments as shown in Figure 2. The critical points of the line segments are represented as a vector and denoted as:

$$\mathbf{V} = [\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_{p-1}]^T, \tag{2}$$

where $\mathbf{P}_i$ is the i-th critical point, defined as:

$$\mathbf{P}_i = [d_i, y_i, r_i], \tag{3}$$

where d_i is the day, y_i is the milk yield, and r_i is the total regression residual of the line segment period from $\mathbf{P}_{i-1}$ to $\mathbf{P}_i$. For $\mathbf{P}_0$, r_o is the maximum residual over the entire period.

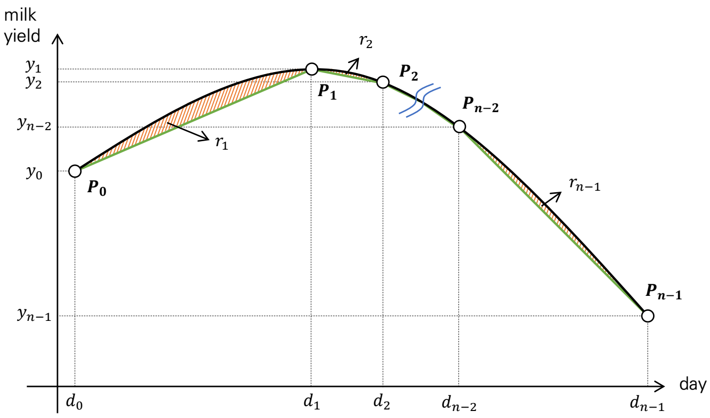


Figure 2. PWLR presentation of an LC.

The LC vector generation procedure for the PWLR model is illustrated in Figure 3. Initially, the starting, ending, and vertex days, which have the maximum value of milk yield for the lactation period, become the three anchors of critical points. The anchor points are shown in Figure 3a. Regression residuals are measured as the root-mean-square for the lactation period. If the size of the residuals is larger than the predetermined threshold, an additional critical point is inserted between two consecutive anchor points (Figure 3b). The new critical point is determined by the day with the maximum residual (or error). Linear regressions were then adopted for the two sub-lactation periods divided by the new critical points. This procedure was repeated until the termination condition was satisfied (Figure 3c). Figure 4 shows the approximation results with $p = 11$ for the same LCs in Figure 1.

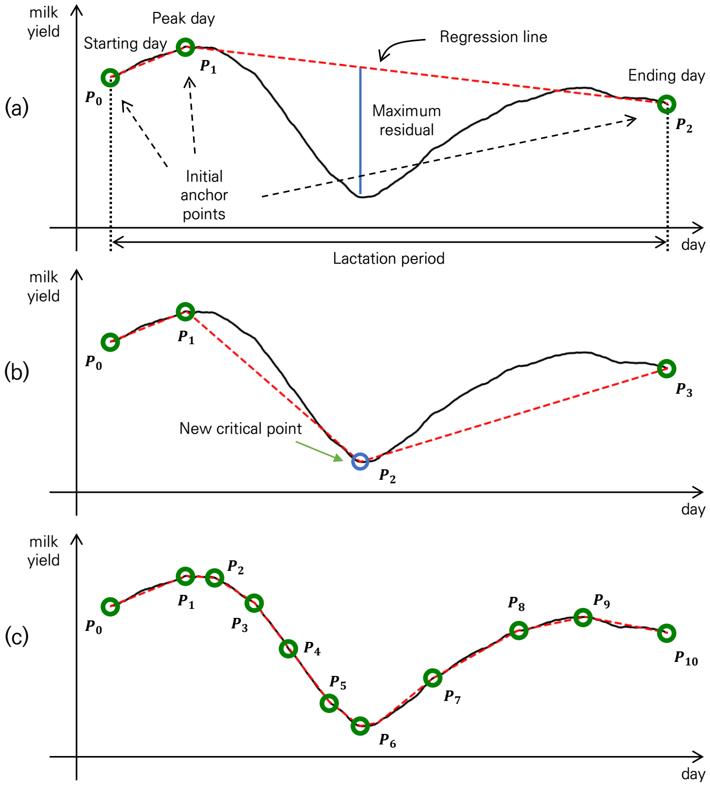
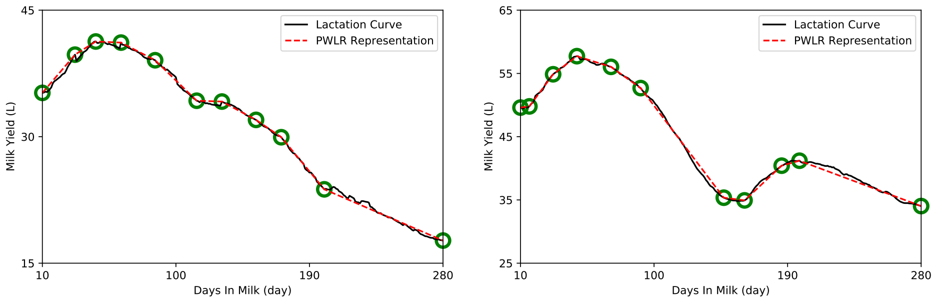


Figure 3. Fitting procedure of the PWLR model. (**a**) initial anchor points setting, (**b**) adding a new critical point, (**c**) final fitting result.



(**a**) a typical convex shape (**b**) an atypical shape

Figure 4. Approximation results using PWLR when $p = 11$ for the same LCs in Figure 1.

The fitting error can be reduced by increasing the number of critical points. However, this leads to a phenomenon known as overfitting, in which models lose generalization. The Akaike information criterion (AIC) [21] and Bayesian information criterion (BIC) [22] are metrics used to measure the appropriateness of fittings. They were devised in an attempt to assess a model in terms of complexity and simplicity [23,24]. A model with lower metric values is preferred to avoid the negative effect of overfitting. BIC was selected to determine the appropriate vector size for the PWLR because the sample size was not considered for AIC.

BIC is defined as:

$$BIC = p \cdot ln(n) - 2ln(\widehat{L}), \tag{4}$$

where p is the number of model parameters, $\widehat{L}$ is the likelihood of the model, and n is the sample size. The first term of Equation (4) indicates the model complexity: the number of parameters involved in the model. In the PWLR model, p is given by the number of critical points, since it is the most compatible counterpart for the number of parameters. If we assume that the residuals are independent and follow a normal distribution, the model likelihood can be simplified using the residual sum of squares (RSS) as $ln(\widehat{L}) = -nln(RSS/n)/2$ in the linear regression [23,25,26].

2.3. Data Resources

The same dataset of [7] is used in this study. Samples were collected from four commercial farms in Chungcheong Province, South Korea, from 2016 to 2018. Data were collected daily from an automatic milking system (Astronaut A4; Lely Industries NV, Maassluis, the Netherlands, three farms) and a conventional milking parlor system in one farm (DeLaval international AB, Tumba, Sweden). In total, 330 LCs were obtained from 286 cows. All the individuals were healthy during the data collection period (subclinical status were not considered). A total of 175 (66.4%) individuals were multiparous.

Data records for only 10–280 days after the delivery day were considered to unify the LC datasets in our experiments and to minimize the differences in the days-in-milk (DIM) caused by the breeding management of the four farms. LCs were z-normalized and clustered into six groups using the k-medoids algorithm. The number of clusters was determined using the elbow method, and each cluster was analyzed as in our previous study [7]. The approach has superior noise immunization compared to the k-means algorithm, although the computation burden is greater [27–29]. The initialization method of the k-means++ algorithm was adopted to achieve fast initial convergence [30]. The method chooses the initial points using a weighted probability distribution of inter-point distances.

A statistical summary of LC data is presented in Table 2. Group A was the largest group, comprising 36.1% of the total LCs, and group F was the smallest group (3.6%). Groups E and F had a relatively larger number of LCs obtained from primiparous individuals than the other groups. In particular, in groups E and F, the peak day was delayed and the milk yield amount on the peak day was also relatively small compared to the other groups.

Table 2. Statistics for the LC groups.

Group	**A**	**B**	**C**	**D**	**E**	**F**
Number of cows	119	64	50	47	38	12
- distribution	36.1%	19.4%	15.2%	14.2%	11.5%	3.6%
- primiparous ratio	14.3%	43.8%	36.0%	21.3%	73.7%	83.3%
Parity	2.61	2.16	2.30	2.57	1.63	1.25
	(0.12) †	(0.17)	(0.19)	(0.20)	(0.20)	(0.18)
Total Milk Yield (liter/cow)	10,713	9930	10,351	10,504	9509	9806
	(165)	(261)	(252)	(275)	(305)	(459)
Peak Milk Yield (liter/cow)	53.08	46.39	47.14	54.25	42.70	45.82
	(0.83)	(1.24)	(1.08)	(1.34)	(1.33)	(2.26)
Peak Day (day)	59.92	86.25	119.68	54.94	144.00	119.92
	(2.73)	(4.61)	(8.42)	(6.11)	(9.67)	(25.73)

† standard error.

2.4. Evaluation

In the evaluation, the distance D between two LCs is the root-mean-square of the element-wise difference of the daily milk yield amount of the two LCs as:

$$D(\mathbf{Y}, \hat{\mathbf{Y}}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}, \tag{5}$$

where $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are the LC vectors of the n-day lactation period defined in Equation (1).

If $\mathbf{Y}$ is a real LC and $\hat{\mathbf{Y}}$ is an approximated LC generated by an LC model, $D(\mathbf{Y}, \hat{\mathbf{Y}})$ is considered the *fitting error* of the approximated LC, denoted as $e_f = D(\mathbf{Y}, \hat{\mathbf{Y}})$. The performance of an LC model is evaluated based on the approximation accuracy of a real LC. The approximation accuracy can be measured using the fitting error.

If an LC model describes a real LC, it is best if the model has a minimum fitting error. If an LC model depicts multiple LCs for a group, it would be the same as in the single LC case. To fit multiple LCs together, the fitting error should be replaced by the mean of the individual fitting errors as follows:

$$e_g = \frac{1}{m}\sum_{i=0}^{m-1} D(\mathbf{Y}_i, \hat{\mathbf{Y}}), \tag{6}$$

where $\mathbf{Y}_i$ is the ith member LC in the group, which is defined as $\mathbf{Y}_i = [y_{i,0}, y_{i,1}, \ldots, y_{i,n-1}]$, and m is the number of LCs in the group, which is called group size. The mean, rather than the sum, is preferred to be independent of group size variations.

Evaluating the conventional LC models by e_g is straightforward because they use least-square-based regressions. If the PWLR model fits only an LC, the minimum fitting error with a limited number of critical points may be compatible. However, for a group, the PWLR cannot directly adopt the least-square method as the conventional method. This is because PWLR works only for one LC and not for multiple LCs simultaneously.

A prototype LC is necessary from the target LC group to fit multiple LCs together, or the PWLR model. The prototype LC should accurately represent the entire group of LCs. In our experiments, the mean LC, denoted by $\bar{\mathbf{Y}}$, was used. The mean LC is generated by the daily mean values of the yield amounts of LCs in a group as follows:

$$\bar{\mathbf{Y}} = [\bar{y}_0, \bar{y}_1, \ldots, \bar{y}_{n-1}], \tag{7}$$

where $\bar{y}_i$ is given as:

$$\bar{y}_j = \frac{1}{m}\sum_{i=0}^{m-1} y_{i,j}. \tag{8}$$

For performance evaluation, another metric for measuring the fitting error of a prototype LC was devised. In this evaluation, e_m is a fitting error to measure the approximation accuracy of an LC model for a prototype LC instead of a group LC, where e_m is directly given by the Equation (5):

$$e_m = D(\bar{\mathbf{Y}}, \hat{\mathbf{Y}}). \tag{9}$$

The performance evaluation of the PWLR model by e_m is intuitive. However, for conventional methods, it is an indirect metric, conversely to e_g. In the evaluation, e_m is the mean fitting error and e_g is the intra-fitting error. In our experiments, both e_g and e_m were measured for objective performance evaluation.

The PWLR model risks the overfitting problem when the number of critical points is excessively large. If an approximation of PWLR is overfitted to an LC, the intra-fitting error e_g would be much larger than the standard deviation of all member LCs. Representation generality of an LC model can be tested by comparing the intra-fitting errors.

The representation generality of an LC model can be clearly evaluated through k-fold cross-validation. In the k-fold cross-validation, all LCs in a group are equally divided into k subgroups. A subgroup becomes a test set, while the others are merged into a training set. All LCs in the training set are grouped through k-means clustering, the same as when the dataset was generated in our previous study [7]. Subsequently, all LCs in the test set were mapped to their closest group using z-normalized Euclidean distance [7]. Because the LC groups of each fold are all naturally different, they are also re-labeled as the closest original group, as shown in Table 2 for comparison under the same standards. The distance between the two groups was measured as the z-normalized Euclidean distance between their mean LCs.

A prototype LC, the mean LC in our experiments, is generated from each group of the training set, and a PWLR model is approximate to each prototype LC. The intra-fitting error of the test set was then calculated. It becomes, for example, $e_{g,i}$ of the ith trial. This process is repeated k times until every k subgroup becomes the test set. Then, the fitting errors for the cross-validation, e_c is given by the mean value of the errors, for example, $e_{g,*}$ obtained from k trials:

$$e_c = \frac{1}{k} \sum_{i=0}^{k-1} e_{g,i}. \tag{10}$$

The representation generality of an LC model can be evaluated using e_c, which is called the cross-fitting error.

3. Results

To demonstrate the validity of the PWLR model as an LC descriptor, approximation accuracy and representation generality were tested using the LC dataset presented in Table 2. First of all, the BIC test was performed to determine the appropriate number of critical points. For each group in Table 2, BIC was calculated according to the increment in the number of critical points, p, by one. The results are shown in Figure 5.

When p increased, BIC values dropped rapidly until $p = 7$ in all the groups. After $p = 11$, the BIC values were stable and remained near the minimum values in all groups. The first p for each group reached minimum BIC value of 11, 12, 13, 9, 10, and 12, respectively. Based on this observation, 11 critical points were chosen for all groups. The following experimental results were obtained with $p = 11$ at all times.

The vectors of the PWLR model applied to each group are listed in Table 3. Figure 6 shows the LCs of the six groups generated using the vectors in Table 3. The mean LC and LC generated by the Wilmink model are simultaneously shown for the approximation performance comparison in each group figure. To simplify the figures, the Wilmink model was chosen as a representative from other models in Table 1 because it has the minimum fitting errors. As shown in Figure 6d,f, the PWLR model maintains a the good description ability, even when the LCs do not follow the typical convex style compared to the Wilmink model.

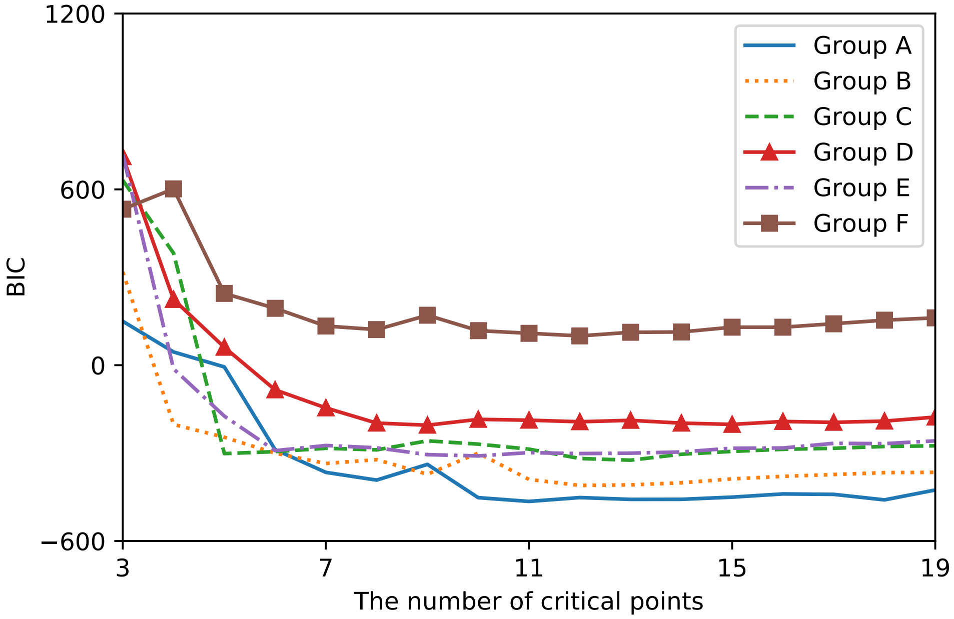
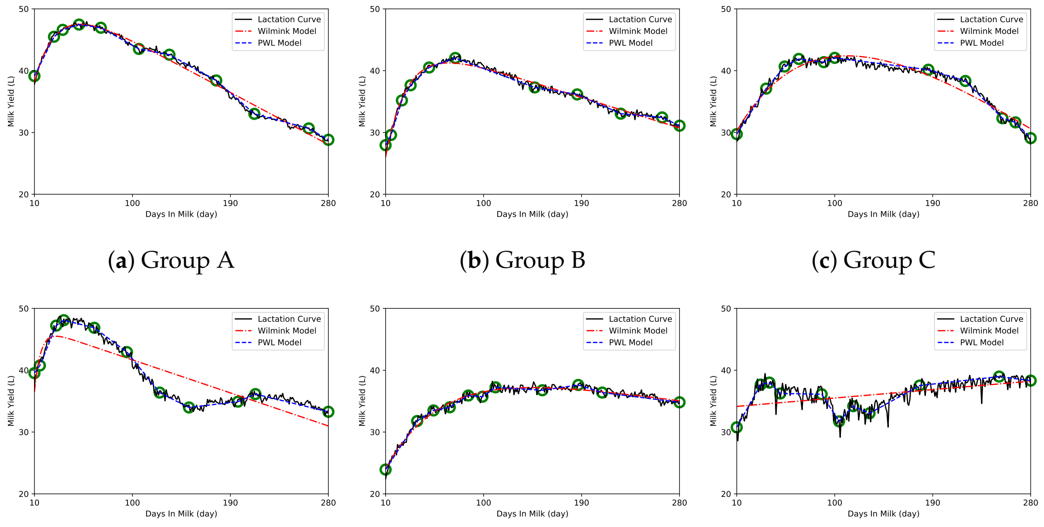


Figure 5. BIC test results of the PWLR model for the LC groups in Table 2.

Table 3. LC vectors for the LC groups in Table 2.

V	**Group A**			**Group B**			**Group C**		
	d	*y*	*r*	*d*	*y*	*r*	*d*	*y*	*r*
$\mathbf{P}_0$	10	39.13	12.77	10	27.95	27.63	10	29.74	34.71
$\mathbf{P}_1$	28	45.49	5.94	15	29.59	2.62	37	37.10	10.19
$\mathbf{P}_2$	36	46.60	1.62	25	35.18	3.84	54	40.67	5.92
$\mathbf{P}_3$	51	47.46	3.83	33	37.64	1.91	67	41.90	4.58
$\mathbf{P}_4$	71	46.94	3.42	50	40.51	3.53	90	41.38	7.59
$\mathbf{P}_5$	106	43.52	8.32	74	42.05	6.03	100	42.07	2.90
$\mathbf{P}_6$	134	42.58	7.05	147	37.28	27.63	186	40.15	34.71
$\mathbf{P}_7$	177	38.41	9.05	186	36.16	9.63	220	38.33	10.18
$\mathbf{P}_8$	212	33.01	11.31	226	33.04	10.15	254	32.30	12.07
$\mathbf{P}_9$	262	30.66	12.77	264	32.44	9.72	266	31.63	4.86
$\mathbf{P}_{10}$	280	28.81	1.99	280	31.10	1.81	280	29.09	1.60
	Group D			**Group E**			**Group F**		
	d	*y*	*r*	*d*	*y*	*r*	*d*	*y*	*r*
$\mathbf{P}_0$	10	39.53	22.03	10	23.93	22.40	10	30.77	47.19
$\mathbf{P}_1$	15	40.77	3.54	39	31.78	10.24	32	37.67	13.38
$\mathbf{P}_2$	30	47.21	8.63	54	33.50	4.61	40	38.00	7.74
$\mathbf{P}_3$	37	48.10	2.91	69	34.01	5.05	50	36.24	7.64
$\mathbf{P}_4$	65	46.87	14.92	86	35.91	6.02	88	36.12	32.69
$\mathbf{P}_5$	95	42.94	16.07	99	35.72	5.53	104	31.73	12.87
$\mathbf{P}_6$	125	36.38	15.30	111	37.24	4.81	117	34.19	11.03
$\mathbf{P}_7$	152	33.97	9.34	154	36.76	16.66	132	33.03	13.41
$\mathbf{P}_8$	197	34.92	18.58	187	37.61	10.16	178	37.52	33.21
$\mathbf{P}_9$	213	36.14	6.41	209	36.38	9.08	251	38.97	47.19
$\mathbf{P}_{10}$	280	33.27	22.03	280	34.81	22.40	280	38.32	10.07

d, *y* and *r* are each represented for days in milk, milk yield and residual.



(**a**) Group A (**b**) Group B (**c**) Group C

(**d**) Group D (**e**) Group E (**f**) Group F

Figure 6. Fitting results of the PWLR model for the groups.

Table 4 shows experimental results for evaluating the approximation accuracy. The column group mean is for the simple average of group fitting errors without group size weighting. The column whole set is for the fitting errors to the entire dataset. The LC description ability of PWLR improved in all groups. The approximation accuracy of PWLR dramatically improved for the groups with atypical LC shapes. For e_m, improvements of 76% and 48% were achieved compared to the Dijkstra model on group D and Wilmink model on group F, respectively. However, the performance of PWLR did not improve for fitting to the mean LC of the whole set. The accuracy inferiority was somewhat marginal compared to performance improvements obtained in the group experiments. This is because the conventional methods work with the least square to minimize the whole errors where the PWLR model fits to the mean LC.

The advantages of the PWLR model are clearly shown in the experimental results. The representation generality can be evaluated via e_g, as well as the approximation accuracy, as e_g is calculated from the distance measurement between a model LC and real LCs individually. The lower the value of e_g, the better is the representation generality that is achieved, and the better is the approximation accuracy. Although the e_g reduction is not high as e_m, the description performance of the PWLR model is superior to that of other conventional methods in all cases.

The largest e_m of group F is associated with the following two observations: (i) the BIC values are relatively large compared to other groups, and (ii) the proper number of critical points is 12. These findings show that BIC works as a metric to determine the appropriate number of critical points for PWLR. The larger e_g of groups D and F in Table 4 is also associated with the BIC test results.

The experimental results of the k-fold cross-validation are listed in Table 5. The PWLR model has the lowest e_c in groups A, B, D, and F, and remains at a lower value than the other model in the other cases. For groups D and F, atypical LC groups showed 8% e_c reduction. Based on this observation, the PWLR model has a superior, or at least compatible, LC description ability for both approximation accuracy and representation generality, while the polymorphic paradigm is maintained.

Table 4. Mean and intra-fitting errors for the dataset.

	LC	Group							Whole
	Model	**A**	**B**	**C**	**D**	**E**	**F**	**Mean**	**Set**
e_m	Brody	2.553	0.788	2.073	2.682	1.261	2.019	1.896	0.742
	Wood	0.668	0.953	1.147	2.675	0.507	1.978	1.321	0.609
	Cobby	1.009	0.767	1.999	2.547	1.261	2.019	1.600	0.663
	Wilmink	0.668	0.524	1.006	2.546	0.551	1.881	1.196	0.244 *
	Rook	0.635	0.600	1.053	2.563	0.524	1.881	1.209	0.313
	Dijkstra	0.693	0.464	1.121	2.368	0.578	2.018	1.207	0.248
	PWLR	0.338 *	0.388 *	0.470 *	0.563 *	0.459 *	0.973 *	0.532 *	0.310
e_g	Brody	4.400	3.559	4.214	5.010	3.395	4.302	4.147	4.176
		(0.166) †	(0.184)	(0.185)	(0.244)	(0.213)	(0.527)	(0.253)	(0.093)
	Wood	3.887	3.431	3.896	5.040	3.166	4.245	3.944	3.894
		(0.145)	(0.16)	(0.177)	(0.241)	(0.191)	(0.495)	(0.235)	(0.085)
	Cobby	4.033	3.390	4.178	5.052	3.351	4.303	4.051	4.007
		(0.153)	(0.163)	(0.185)	(0.257)	(0.193)	(0.527)	(0.246)	(0.089)
	Wilmink	3.898	3.296	3.686	5.017	3.063	4.167	3.854	3.822
		(0.154)	(0.163)	(0.168)	(0.255)	(0.188)	(0.499)	(0.238)	(0.089)
	Rook	3.812	3.328	3.809	4.958	3.095	4.178	3.863	3.812
		(0.145)	(0.159)	(0.176)	(0.239)	(0.19)	(0.492)	(0.234)	(0.085)
	Dijkstra	3.760	3.253	3.717	4.846	3.045	4.148	3.795	3.741
		(0.146)	(0.159)	(0.167)	(0.238)	(0.187)	(0.504)	(0.234)	(0.085)
	PWLR	2.721 *	2.524 *	2.898 *	3.049 *	2.479 *	2.984 *	2.776 *	2.738 *
		(0.097)	(0.115)	(0.135)	(0.169)	(0.171)	(0.379)	(0.178)	(0.058)

† standard deviation; * min in each column; e_m and e_g represent mean fitting error and intra-fitting error each; *Mean* indicates the error average of group A–F, whereas *Whole set* means the error average of entire LCs.

Table 5. Cross-fitting errors of *k*-fold cross-validation.

	LC	Group							Whole
	Model	**A**	**B**	**C**	**D**	**E**	**F**	**Mean**	**Set**
e_c	Brody	4.436	3.956	4.068	4.605	3.709	7.460	4.289	4.524
	Wood	4.036	3.821	3.763 *	4.682	3.486 *	7.442	4.095	4.445
	Cobby	4.115	3.931	4.030	4.598	3.709	7.460	4.211	4.508
	Wilmink	4.035	3.787	3.780	4.601	3.513	7.210	4.075	4.429
	Rook	4.021	3.796	3.773	4.618	3.505	7.205	4.073	4.431
	Dijkstra	4.024	3.782	3.824	4.540	3.519	7.434	4.078	4.424 *
	PWLR	3.987 *	3.776 *	3.765	4.141 *	3.542	6.619 *	3.952 *	4.450

* min in each column; e_c represents cross-fitting error; *Mean* indicates the error average of group A–F, whereas *Whole set* means the error average of entire LCs.

4. Discussion

The shape of the LC is a phenotypic expression of the biological processes in cows. The basic traits of LC are prototyped by the two rates of increase and decrease in milk production dichotomized by the time at peak occurrence. Most previous studies on LC modeling have focused on describing this phenomenon. LCs were modeled to determine specific parameters, and the productivity factors for a herd were generated from the parameters to estimate their relationship with other clinical factors. These models are still widely used for average curves, which follow the standard pattern obtained from a large herd [9].

Recent advances in computing power, data quality, and industry requirements of precision farming have shifted the research focus from average patterns to individual deviations. The processing of rich associated data with high dimensions and precision provides a large amount of information available for management precision farming applications. Monitoring variation within individuals and controlling individual deviations from expected patterns may be the main concerns of new research issues.

Recently developed machine learning techniques have provided good tools for new research topics. Most machine learning algorithms extract the desired information directly from data without adopting specific mathematical functions that are essential in the previous method. The direct use of data draws out the causal relationships between the various types of information contained in the data. To use machine learning, the information contained in the data must be well-preserved. The larger the data size is, the more advantageous it is.

Enhancing the LC representation ability up to atypical cases and preserving the LC patterns is one of the major aims of this study. Eventually, the vectorized LC representation is intended to be used directly as data for machine learning, along with other clinical data at an individual base. The PWLR model is advantageous for representing the LCs of individuals or a small group of clusters. For general cases where an individual follows typical LC patterns, the PWLR model may have marginal advantages, but it has greater advantages for atypical cases of non-convex-shaped LCs.

The key idea is to represent LCs in the vector format of an array of critical points that approximates the actual LC dataset. As mentioned in the previous section, the larger the dimension of the vector, the better is the approximation accuracy. However, a higher dimension does not always guarantee a better representation power. The overfitting problem, referred to as the curse of dimensionality, is a well-known problem. Finding a proper dimension of the vector to represent essential information is the key to success in such data-driven approaches.

Two strategies can be used to complete the LC vector generation: keep recursion until residuals are satisfied below the predetermined threshold or reach the limit for the maximum allowable number of critical points. The first strategy regulates the regression errors but cannot control the dimensions of the LC vectors. The second strategy unifies the dimensions but cannot regulate the regression errors.

In this study, the dimension of the LC vectors (i.e., number of critical points in the PWLR model) was unified as a fixed number. The proper dimension of the LC vector was determined using BIC. The experimental results show a better representation ability compared to previous LC models in terms of regression performance. However, this does not mean that the PWLR model is better than the others in any application where the LC model is necessary. Statistical-based models, most of the previous ones, may be better for the prediction of milk yield amount in a large group, such as a local cooperative union of a country. Regression errors for individuals are usually compensated for by others in a large group.

This study only used a data-driven approach to represent LCs without feature selection methods. The minimum number of critical points that satisfies the maximum allowable regression residuals may be a good feature for LC classification. The distribution of the critical points divided by the peak yield points could be a useful feature for phenotyping of individuals. The fine feature selection procedure may be directly applicable to LC vectors, which may lead to further understanding of the individual milking characteristics.

Persistency of milk production, a measure of the ratio change in milk production within a period, is a valuable trait that is of direct economic interest because of its relationship with reproduction, health, and feed costs [31–33]. Persistency can be calculated based on the interval of the test day (generally one month) or using the parameters derived from the model [34]. Considering that the PWLR model approximates LCs with connected line segments, the traditional persistency measure can be applied and utilized in lactational and biological analyses [33,35,36].

5. Conclusions

Data collection in dairy science has become easier and richer than in the past, and various data-driven machine learning methods have been developed. In order to use these various tools for LC-related research, data-centric LC representation is highly demanded.

A vector representation of LC is presented without the use of a mathematical framework. It regards an LC as a continuous line segment, and the location and number of segments are determined by the maximum residual and the BIC. The descriptive ability of PWLR model was evaluated in terms of approximation accuracy and representation generality. The experimental results show that the presented method is superior to other conventional models in describing typical and atypically shaped LCs as well.

This method requires additional adjustment of the number of critical points to maintain an appropriate level of precision, and in some cases may lead to overfitting problems of the LC. However, since this method is a data-centric, various other features can be directly extracted without considering the mathematical derivations. Not only can this method be directly applicable to machine learning techniques, but it can also enhance the LC descriptive ability up to atypical cases while preserving the intrinsic LC patterns. The use of this method for machine learning with associated LC-related datasets is beyond the scope of this paper and remains as a future study.

Author Contributions: Formal analysis and writing—original draft, S.L.; Supervision and writing—review and editing, J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Advanced Production Technology Development Program (Project No. 318005-04-2-HD030), Ministry of Agriculture, Food and Rural Affairs, Korea; this research was supported by the Chung-Ang University Research Scholarship Grants in 2018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to extend their special thanks to Mingyung Lee, Seongwon Seo at Chungnam National University, Younghwa Ham at AgriRoboTech Co., Ltd. and Jae June Cho at Korea Dairy Committee for providing the data, comments, and help for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cunha, D.d.N.F.V.d.; Pereira, J.A.C.; Silva, F.F.e.; Campos, O.F.d.; Braga, J.A.L.; Martuscello, J.A. Selection of models of lactation curves to use in milk production simulation systems. *Rev. Bras. Zootec.* **2010**, *39*, 891–902. [CrossRef]
2. Dijkstra, J.; France, J.; Dhanoa, M.; Maas, J.; Hanigan, M.; Rook, A.; Beever, D. A Model to Describe Growth Patterns of the Mammary Gland During Pregnancy and Lactation. *J. Dairy Sci.* **1997**, *80*, 2340–2354. [CrossRef]
3. Wood, P.D.P. Algebraic Model of the Lactation Curve in Cattle. *Nature* **1967**, *216*, 164–165. [CrossRef]
4. Wilmink, J. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. *Livest. Prod. Sci.* **1987**, *16*, 335–348. [CrossRef]
5. Hossein-Zadeh, N.G. Application of nonlinear mathematical models to describe effect of twinning on the lactation curve features in Holstein cows. *Res. Vet. Sci.* **2019**, *122*, 111–117. [CrossRef]
6. Rook, A.J.; France, J.; Dhanoa, M.S. On the mathematical description of lactation curves. *J. Agric. Sci.* **1993**, *121*, 97–102. [CrossRef]
7. Lee, M.; Lee, S.; Park, J.; Seo, S. Clustering and Characterization of the Lactation Curves of Dairy Cows Using K-Medoids Clustering Algorithm. *Animals* **2020**, *10*, 1348. [CrossRef]
8. Murphy, M.; Zhang, F.; Upton, J.; Shine, P.; Shalloo, L. A review of milk production forecasting models. In *Dairy Farming: Operations Management, Animal Welfare and Milk Production*; Global Agriculture, Nova Science Publishers: New York, NY, USA, 2018; pp. 14–61.
9. Macciotta, N.P.; Dimauro, C.; Rassu, S.P.; Steri, R.; Pulina, G. The mathematical description of lactation curves in dairy cattle. *Ital. J. Anim. Sci.* **2011**, *10*, e51. [CrossRef]
10. Iewdiukow, M.; Lema, O.; Velazco, J.; Quintans, G. Is it possible to accurately estimate lactation curve parameters in extensive beef production systems? *Appl. Anim. Sci.* **2020**, *36*, 509–514. [CrossRef]
11. Nixon, M.; Bohmanova, J.; Jamrozik, J.; Schaeffer, L.; Hand, K.; Miglior, F. Genetic parameters of milking frequency and milk production traits in Canadian Holsteins milked by an automated milking system. *J. Dairy Sci.* **2009**, *92*, 3422–3430. [CrossRef]
12. Grandl, F.; Furger, M.; Kreuzer, M.; Zehetmeier, M. Impact of longevity on greenhouse gas emissions and profitability of individual dairy cows analysed with different system boundaries. *Animal* **2019**, *13*, 198–208. [CrossRef] [PubMed]
13. Masía, F.; Lyons, N.; Piccardi, M.; Balzarini, M.; Hovey, R.; Garcia, S. Modeling variability of the lactation curves of cows in automated milking systems. *J. Dairy Sci.* **2020**, *103*, 8189–8196. [CrossRef] [PubMed]

14. Murphy, M.; O'Mahony, M.; Shalloo, L.; French, P.; Upton, J. Comparison of modelling techniques for milk-production forecasting. *J. Dairy Sci.* **2014**, *97*, 3352–3363. [CrossRef] [PubMed]
15. Liseune, A.; Salamone, M.; Van den Poel, D.; van Ranst, B.; Hostens, M. Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning. *Comput. Electron. Agric.* **2021**, *180*, 105904. [CrossRef]
16. Brody, S.; Turner, C.W.; Ragsdale, A.C. The Relation between the initial rise and the subsquent decline of milk secretion following parturition. *J. Gen. Physiol.* **1924**, *6*, 541–545. [CrossRef]
17. Cobby, J.M.; Le Du, Y.L.P. On fitting curves to lactation data. *Anim. Sci.* **1978**, *26*, 127–133. [CrossRef]
18. Bouallegue, M.; M'Hamdi, N. Mathematical Modeling of Lactation Curves: A Review of Parametric Models. In *Lactation in Farm Animals*; M'Hamdi, N., Ed.; IntechOpen: Rijeka, Croatia, 2020; Chapter 6. [CrossRef]
19. Vieth, E. Fitting piecewise linear regression functions to biological responses. *J. Appl. Physiol.* **1989**, *67*, 390–396. [CrossRef]
20. Hamann, B.; Chen, J.L. Data point selection for piecewise linear curve approximation. *Comput. Aided Geom. Des.* **1994**, *11*, 289. [CrossRef]
21. Hirotogu, A. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*; Parzen, E., Tanabe, K., Kitagawa, G., Eds.; Springer: New York, NY, USA, 1998; pp. 199–213. [CrossRef]
22. Schwarz, G. Estimating the Dimension of a Model. *Ann. Statist.* **1978**, *6*, 461–464. [CrossRef]
23. Burnham, K.P.; Anderson, D.R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]
24. Wit, E.; Heuvel, E.v.d.; Romeijn, J.W. 'All models are wrong...': An introduction to model uncertainty. *Stat. Neerl.* **2012**, *66*, 217–236. [CrossRef]
25. Myung, I.J. The Importance of Complexity in Model Selection. *J. Math. Psychol.* **2000**, *44*, 190–204. [CrossRef]
26. Browne, M.W. Cross-Validation Methods. *J. Math. Psychol.* **2000**, *44*, 108–132. [CrossRef] [PubMed]
27. Liao, T.W. Clustering of time series data—A survey. *Pattern Recognit.* **2005**, *38*, 1857–1874. [CrossRef]
28. Xu, D.; Tian, Y. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
29. Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering—A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [CrossRef]
30. Arthur, D.; Vassilvitskii, S. K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, New Orleans, LA, USA, 7–9 January 2007; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2007; pp. 1027–1035.
31. Sölkner, J.; Fuchs, W. A comparison of different measures of persistency with special respect to variation of test-day milk yields. *Livest. Prod. Sci.* **1987**, *16*, 305–319. [CrossRef]
32. Dekkers, J.; Ten Hag, J.; Weersink, A. Economic aspects of persistency of lactation in dairy cattle. *Livest. Prod. Sci.* **1998**, *53*, 237–252. [CrossRef]
33. Western Canadian DHI Services. Lactation Curves. Available online: http://www.agromedia.ca/ADM_Articles/content/DHI_lactcrv.pdf (accessed on 20 January 2022).
34. Thornley, J.H.; France, J. *Mathematical Models in Agriculture: Quantitative Methods for the Plant, Animal and Ecological Sciences*; CABI: Wallingford, UK, 2007. [CrossRef]
35. Hickson, R.; Lopez-Villalobos, N.; Dalley, D.; Clark, D.; Holmes, C. Yields and Persistency of Lactation in Friesian and Jersey Cows Milked Once Daily. *J. Dairy Sci.* **2006**, *89*, 2017–2024. [CrossRef]
36. Cole, J.; Null, D. Genetic evaluation of lactation persistency for five breeds of dairy cattle. *J. Dairy Sci.* **2009**, *92*, 2248–2258. [CrossRef]