Slovak J. Anim. Sci., 41, 2008 (3): 126 - 132

USEFULNESS OF REGRESSION MODELS FOR 100-DAY MILK YIELD ESTIMATION IN DAIRY COWS

W. P. PERZ[1,*], Z. SOBEK[1], P. FĽAK[2]

[1]Department of Genetics and Animal Breeding, University of Life Sciences, Poznań, Poland;
[2]Slovak Agricultural Research Centre, Nitra, Slovakia

ABSTRACT

Twenty three models of approximation functions have been analyzed for accuracy of milk yield estimation in dairy cows over a period of 100 days from the date of parturition. The assessment is based on the method of counts and a comparison of estimation error variance. It was found that an increased number of test day yields markedly improves the accuracy of a 100-day yield estimation. Similarity, between the course of the function and the actual lactation curve no significant affect was found for the estimation of a 100-day milk yield based on test yields. The best results of estimation were obtained using the rescaling function. This method, however, requires data from other cows in order to construct a mean lactation curve. Among functions that only use data corresponding to the cow under study, the best results were obtained using linear regression and the difference of two exponential functions. The models were assessed for quality using data on daily milk yields of pure Holstein-Friesian breed of cows and crossbred cows with a Holstein-Friesian admixture.

Key words: milk yield estimation, lactation curves, regression models

INTRODUCTION

Keeping milk yield records is important for the genetic improvement and management of a dairy herd. Due to the necessity of cutting the costs, in many countries methods based on test day yields have been developed. One of the most frequently used methods is based on the monthly test-day record (Koonawootrittriron et al., 2001)

In the most widespread model of cow and bull assessment, a 305-day yield is used, estimated from 10 test yields. Errors in yield estimation are due to environmental conditions and the overall physical fitness of the herd. The methods of yield estimation based on 10 test yields tend to employ the results of automatic measurements from the milking pool (Bellamy, 1999).

Models describing lactation can be divided into two basic groups: linear and non-linear (Masselin et al., 1989). In recent years, non-linear models have become more popular as they enable the description of a relatively wider gamut of lactation curves. Iterative procedures of fitting non-linear regressions implemented in statistical packages seem to solve the problem of model fitting (Vargas et al., 2000).

Currently, one of the major problems involved in dairy cattle breeding concerns early prediction of the lactation yield of cows, as the milk yield level largely determines the profitability of production. Faster assessment means a smaller generation gap, and thus also better breeding progress on an annual scale. Moreover, assessment at an earlier stage also decreases the costs. An early assessment of a cow's yield would enable a faster decision concerning its breeding. At the same time, a more accurate assessment of the yield of particular individuals due to the choice of better models for the estimation of the lactation curve would make it

Correspondence: *E-mail: wpperz@jay.au.poznan.pl

ISSN 1335-3686

possible to take into account a smaller number of daughters for the evaluation of their sire. All the above factors were considered in setting the goal of this study, which attempts to develop a faster and more accurate mode of assessment of the total milk yield in cows. The study is restricted to the assessment of the milk yield, since in the course of lactation, changes in milk composition are not as great as changes in its quantity.

MATERIAL AND METHODS

The material used for the verification of results obtained from approximating models included empirical data from the daily milking of cows of pure Holstein-Friesian breed and a cross-breed with an H-F (Holstein-Friesian) admixture. The cows were selected at random from two herds of approximately 1000 animals each, reared at the Experimental Station of the National Research Institute of Animal Production in Pawłowice, Poland.

The cows subjected to the study were divided into three groups: the first group, numbering 29 animals, consisted of cows whose lactation started in winter, including 11 cows after first parturition, 9 cows after second parturition, and 9 cows after third parturition. For cows in the first and second lactation, parturition times ranged within two months, whereas in the third lactation the range of parturition period was three months. The second group comprised 31 cows which began lactation in summer; 19 cows in the first lactation, 5 in the second, 5 in the third, and one in the fourth and fifth lactation each. The third group consisted of 41 cows with parturition in the winter season (between October and December) selected at random from the second herd: 9 cows in the first lactation, 18 in the second, 7 in the third, 5 in the fourth, and one after the fifth and sixth parturition each. The cows under the study were milked twice daily. Altogether 20200 data have been collected.

Selection of cows for the analysis from two herds living in similar environmental conditions and with similar dates of parturition aimed at unification of observation conditions. The division of cows into groups enabled a comparison of the results, and a test application of the average lactation curve based on the day yields of cows from one group, to the prediction of yield for cows from the other groups, using the scaling method (Perz, 1998).

The lactation curves for the first 100-day period following parturition were estimated by: regression line, two intersecting regression lines, polynomial regression (of second to sixth degree), power equation and exponential equation (equations 1 – 8):

$$y = \alpha + \beta x, \tag{1}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{2}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, \tag{3}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4, \tag{4}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5, \tag{5}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6, \tag{6}$$

$$y = \beta_0 x^{\beta_1}, \tag{7}$$

$$y = \beta_0 e^{\beta_1 x}. \tag{8}$$

The usefulness of Newton's interpolating polynomial has been considered (indicated as number 9 in tables). Moreover, the 100-day milk yields for particular cows were determined by the scaling method (Perz, 1998), shown against number 10 in tables.

The results of the above estimations were then compared with those obtained by means of other functions used for modeling the lactation curve (equations 11 – 22; Sherchand et al., 1995).

$$y = a \exp(bx) - a \exp(cx), \tag{11}$$

$$y = a \exp(bx - cx^2), \tag{12}$$

$$y = x / (a + bx + cx^2), \tag{13}$$

$$y = ax^b \exp(cx), \tag{14}$$

$$y = a - bx - c \ln(x), \tag{15}$$

$$y = a + bx + cx^2 + d \ln(x), \tag{16}$$

$$y = ax^{mc} \exp(-cx), \tag{17}$$

$$y = ax^b / \cosh(cx), \tag{18}$$

$$y = a[1 - \exp(bx)] / \cosh(cx), \tag{19}$$

$$y = a \arctan(bx) / \cosh(cx), \tag{20}$$

$$y = a \arctan(bx)] \exp(-cx), \tag{21}$$

$$y = a_1 \{1 - \tanh^2 [b_1 (x^k - c_1)]\} + a_2 \{1 - \tanh^2 [b_2 (x - c_2)]\}, \tag{22}$$

Usefulness of the Michaelis-Menten function – Rose and Bullock (1993) for the estimation of a 100-day yield has also been considered. It has been decided to include this function due to its simple analytical form (see 23), and its shape seemed appropriate to the first period of lactation.

$$y = \frac{\theta_0 x}{\theta_1 + x} \tag{23}$$

where θ_0 - horizontal asymptote, θ_1 - value x, for which the function assumes the value 0.5 θ_0.

In all the equations x denotes a subsequent day of lactation, whereas y stands for the milk yield. The other letter symbols stand for coefficients.

The scaling method essentially consists of the following steps:

1) Determination of the generalized lactation curve (the curve resulting from the mean daily yields of a selected group of cows within a specific period of time – in the case of our study it was the first 100 days of lactation).
2) Selecting the test days (numbers and dates).
3) Determination of the w_t – coefficients – quotients of the actual milk yield of the cow studied on a given day (a_t) and the milk yield from the generalized lactation curve (b_t) for the same day (the number of coefficients correspond to the number of sample milkings).
4) Calculation of the rescaling coefficient w – the arithmetic mean of w_t coefficients.
5) Finding out the estimated value of the cow's 100-day milk yield – product of the scaling coefficient w and the 100-day yield derived from the generalized curve (the sum of 100 mean daily yields for cows for which the generalized lactation curve has been plotted).

The terms used in the description of the above algorithms are presented in Fig. 1.

All the models considered in the present study were assessed on the basis of the error variance for estimation of 100-day milk yield. Models of simple regressions using higher degree polynomials including the scaling method were additionally assessed based on the number of yield counts within the assigned accuracy range (at least 96% in this study). For the counting method the value of the coefficient of a broadening of the count peaks was also taken into account. This coefficient was defined as the quotient of the mean of maximum counts (following subsequent maximum counts) by the total number of counts. The broader the maximum, the lower is the model's sensitivity to the choice of test day.

A 100-day milk yield for each cow within the assumed accuracy of estimation was calculated, for all the possible configurations of test days, for 3 to 7 test yields for all the above mentioned models. Polynomials of 3[rd] to 6[th] degree were the only exceptions as in their case the numbers of test days must be greater by at least one than the equation degree. The actual milk yield during the study period (100 days) was calculated by adding up the day yields for subsequent days. In turn the estimated 100-day milk yield was obtained either by adding up the respective 100 values of the function approximating the lactation curve or by calculating the definite integral if

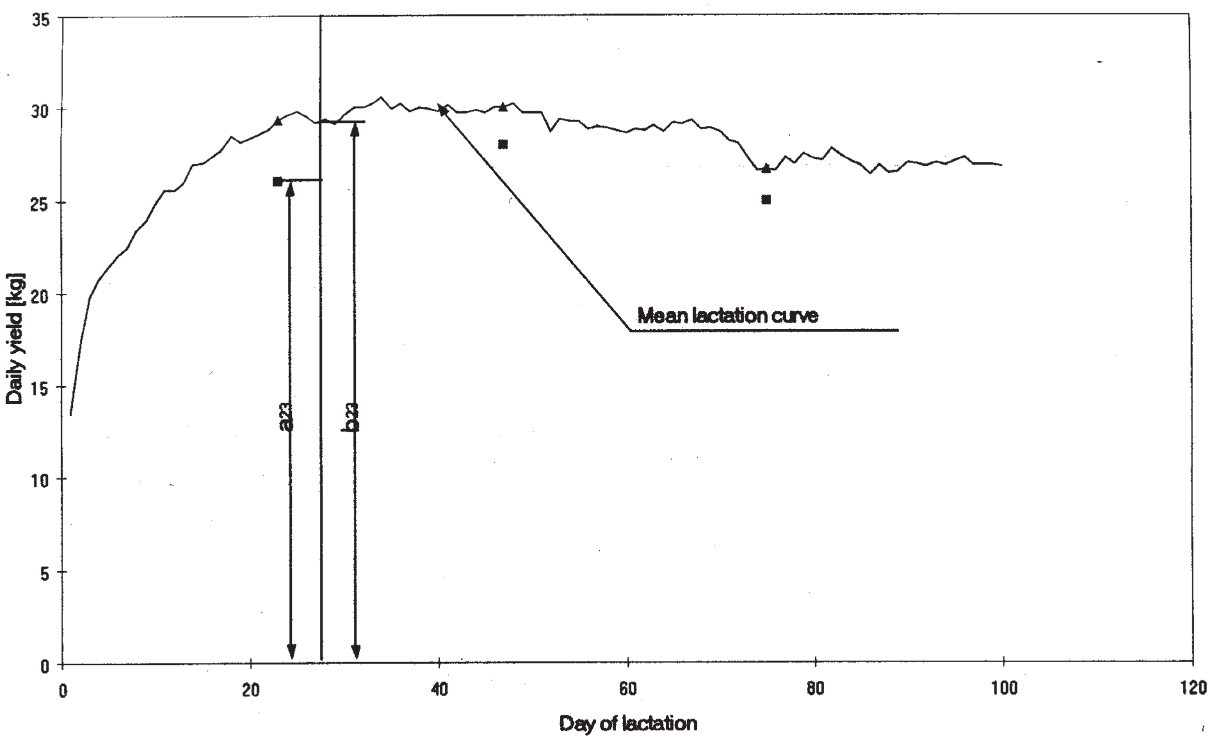

Figure 1: **Illustration of concepts used in the description of the algorithm**

this was the simpler way, as for example, for polynomial functions the integral is also a polynomial.

The number of the possible configurations of k test yield days over an n-day period is the number of k-element combinations from an n-element set without repetition (k = 3-7, n = 100). In the case of three test yields it amounts to 161700, for four test yields it is 3921225, for five 75287520, for six 1192052400, and for seven 1.6×10^{10}. For three, four, and five test yields the approximation functions were determined for all the possible combinations of test days over a 100-day period starting with the day of parturition. Based on these analyses it was found that optimal results are obtained when test yields are chosen at regular time intervals. Therefore, the number of sets checked for six and seven test yields could be limited accordingly. The 100-day period was divided into six intervals in the case of six test yields, and into seven intervals for seven test yields. The intervals for six test yields were: (1;17), (18;34), (35;51), (52;68), (69;84), (85;100), whereas the intervals for seven test yields were: (1;16), (17;30), (31;44), (45;58), (59;72), (73;86), (87;100), respectively. The particular yields were selected from subsequent intervals (e.g. the first yield was taken from intervals 1-17, and not from intervals 1-90+ as previously). In this way the number of combinations checked was reduced to 21381376 for six yields, and to 120472576 for seven yields.

The equations used for calculating linear regression coefficients and higher degree equations were obtained by the least squares method (Martin, 1976). The remaining models were fitted to the data iteratively. For this purpose the Nreg-model module of the Lisp-Stat system was used (Tierney, 1990). Following the determination of model parameters, they were taken as a basis for estimating the 100-day value.

RESULTS

Table 1 includes the mean number of counts for the three groups of cows studied, for functions 1 - 10. From an analysis of the results given in this table it can be stated that the rescaling method is optimal for each of the test yields. However, its construction requires a historical database (Perz, 1998). Among the functions that use test yields only, good results were obtained for linear regression – 2nd in a ranking of 4 test yields, 3rd for 3 and 5, and 4th for 6 and 7 test yields. Quadratic function comes 2nd for 5 and 6 test yields, 3rd for 7 test yields, and 4th for 3 test yields. Third degree polynomial is 2nd for 7 test yields and 3rd for 6 test yields. Satisfying results have also been obtained using power and exponential functions for 3 and 4 test yields. Power function is 2nd for 3, and 3rd for 4 test yields, whereas exponential function is fourth for 4 and 5 test yields. The function consisting of two intersecting straight lines is the last out of an 11-position ranking for 3 to 6 test yields, and 9th for 7 test yields.

Table 1: Number of counts for analysed estimation models

No.	Function	Mean no. of counts (cases when the predicted yield fell within the assumed accuracy bounds) with ranking in bracket				
		Number of test yields				
		3 test yields	4 test yields	5 test yields	6** test yields	7** test yields
1.	Linear	13 (3)	4282 (2)	390231 (3)	1493595 (4)	24922050 (4)
1.a	Combination of two straight lines	-	331 (8)	56593 (9)	283379 (10)	4752606 (9)
2.	Quadratic	12 (4)	3638 (5)	421575 (2)	2208177 (2)	27307804 (3)
3.	3rd degree polynomial	-	1490 (6)	269182 (6)	2172432 (3)	27482578 (2)
4.	4th degree polynomial	-	-	68539 (7)	1138977 (6)	17271287 (6)
5.	5th degree polynomial	-	-	-	362308 (8)	8847409 (8)
6.	6th degree polynomial	-	-	-	-	2088226 (10)
7.	Power	22 (2)	4134 (3)	283526 (5)	1014655 (7)	14537821 (7)
8.	Exponential	11 (5)	3786 (4)	333170 (4)	1389218 (5)	20135554 (5)
9.	Newton's interpolating polynomial	5 (6)	785 (7)	63371 (8)	346231 (9)	1717377 (11)
10.	Rescaling*	67 (1)	8643 (1)	633094 (1)	3024163 (1)	36931745 (1)

* Results given for the rescaling function were obtained by the scaling method
** Limited numbers of sets were examined (see text)

Table 2: Value of coefficient *p*

No.	Function	Value of coefficient *p* with ranking in bracket			
		4 test yields	5 test yields	6 test yields	7* test yields
1.	Linear	0.099 (2)	0.068 (2)	0.118 (7)	0.098 (4)
1.a	Combination of two straight lines	0.172 (8)	0.082 (5)	0.141 (10)	0.129 (9)
2.	Quadratic	0.115 (5)	0.073 (3)	0.095 (3)	0.095 (2)
3.	3rd degree polynomial	0.152 (6)	0.089 (7)	0.094 (2)	0.095 (3)
4.	4th degree polynomial	-	0.113 (9)	0.106 (5)	0.104 (6)
5.	5th degree polynomial	-	-	0.130 (9)	0.116 (8)
6.	6th degree polynomial	-	-	-	0.143 (10)
7.	Power	0.099 (3)	0.084 (6)	0.113 (6)	0.110 (7)
8.	Exponential	0.100 (4)	0.078 (4)	0.105 (4)	0.103 (5)
9.	Newton's interpolating polynomial	0.153 (7)	0.107 (8)	0.122 (8)	0.143 (11)
10.	Rescaling *	0.095 (1)	0.063 (1)	0.085 (1)	0.089 (1)

* Results given for the rescaling function were obtained by the scaling method.

Table 3: Ranking of models in the order of growing variance of estimation error

Ranking position	Variance estimation error for 4 to 7 test yields							
	4 test yields	Number of function	5 test yields	Number of function	6 test yields	Number of function	7 test yields	Number of function
1	6,84	10	4,29	10	4,77	1	2,50	1
2	7,04	1	4,30	1	4,99	11	2,52	8
3	7,14	11	4,39	2	5,07	10	2,55	2
4	7,16	21	4,42	8	5,07	2	2,59	10
5	7,38	8	4,50	12	5,14	3	2,68	3
6	7,69	2	4,58	21	5,33	12	2,71	11
7	8,21	12	4,89	11	5,43	8	2,73	12
8	8,96	7	5,05	7	5,89	14	3,13	21
9	9,02	3	5,18	14	6,91	4	3,14	4
10	17,32	14	5,25	3	7,35	15	3,19	14
11	17,51	15	6,27	15	7,47	7	3,68	15
12	23,59	20	9,38	4	7,62	13	4,38	7
13	26,33	17	16,69	16	7,98	21	5,02	5
14	27,26	13	18,21	20	11,22	5	6,06	16
15	30,82	19	22,84	13	11,79	16	14,40	20
16	71,60	23	23,13	19	19,02	20	20,96	17
17	87,30	18	23,18	17	24,83	19	26,90	6
18	138,33	22	48,99	18	28,74	23	57,12	18
19	-	-	107,93	23	30,27	17	85,19	23
20	-	-	126,43	22	118,98	18	95,99	13
21	-	-	-		123,22	22	127,38	22

Table 2 shows the values of the peak broadening coefficient (p) for the studied functions, for 4 to 7 test yields. Coefficient p was defined as a quotient (ratio) of the arithmetic mean of maximum values of counts (that is, average magnitude of peaks) to the total number of counts. A lower value of the p coefficient reflects a flattened distribution of counts and indicates that the model is less sensitive to the choice of test milking days. The value of the p coefficient for all the analyzed functions depends on the number of test yields. For the scaling method the coefficient assumes the lowest value, so test day is a function with the least influence on the accuracy of milk yield estimation. For 4 and 6 test yields the function made up of two intersecting straight lines is last in the ranking for 7 test yields, and 9th for 11 test yields. For this reason it has been omitted in further considerations.

Table 3 includes the ranking of models compared on the basis of the mean error variation for milk yield estimation in cows. The place of particular functions in the ranking changes relative to the growing number of test yields. Still, most of the functions in the top places remain stable, while minor changes are largely random (e.g. due to the choice of a certain set of test days).

Results improve markedly when the number of yields is increased to seven, but no radical differences in variance are observed between estimation from five and six yields. Interestingly, the results for six yields are even slightly worse than for five test yields.

In the ranking, original Wood's curve (see 14) is placed 10th for four and seven yields, 9th for five yields, and 8th for six yields. The most sophisticated models give relatively poorest estimation of results, which can probably be accounted for by a small number of records.

Not all of the testing functions are listed in each table. Using the Lisp-Stat package we were not able to find coefficients for equation 16 for four yields, and for equation 19 for 7 yields. Thus, we could not select the appropriate initial values to obtain a convergence of the iterative process.

DISCUSSION

Nonlinear functions usually require a denser division of the period under study. Functions that are more flexible in adjusting to data are more sensitive to fluctuations that occur in them. That is why the results obtained using the regression line are more accurate for a small number of test yields, compared to other models considered in this study (with the exception of the rescaling method).

In the above light it is worth noting that assessment of the fitness of the model based on the determination coefficient R^2 (Sherchand et al., 1995) need not lead to univocal conclusions as to its usefulness in estimating the yield in a given period. The determination coefficient denotes only the accuracy with which a model plots the measurement points. The shape of the function is significant only in those cases when we want to trace a unique and untypical course of the lactation curve.

Irrespective of the regression model chosen the number of test yields and intervals between them markedly affect the accuracy of 100-day estimation. Curvilinear equations to describe lactation curves have been used by Scott et al. (1996). They assessed five different models for lactation curves. In that experiment they studied the performance of cows from five herds of Holstein breed. The time intervals between subsequent test yields were from 12 to 31 days, and the average span between intervals was 15.6 days. The number of test yields in their studies was greater than six, but the authors did not justify the exact number. From the results reported in our study, it can be concluded that an increase of test yields from four to six does not lead to a marked decrease of the yield estimation error, whereas it is significantly reduced in the case of seven test yields. This can be accounted for by the fact that a growing number of test yields is accompanied by a flattening of record peaks distribution. This in turn is evidence that day of test yield has a minor influence on estimation error. An interesting method of milk yield prediction for incomplete lactations was presented by Jones (1997), who used prediction of future production by empirical Bayes fitting of Wood's curve. It consists in assembling a detailed database on the history of completed lactations in the herd. The database includes calculated Wood's curve coefficients assessed for each individual animal and all its lactations using the regression model. Yield assessment of a new cow is done by comparing its milk yield on a given day (test yield) with the yields determined for that day on Wood's curve from parameters recorded in the lactation history database. Importance is attached to those curves that best fit the yield of the cow studied. Parameters of Wood's curve for this individual are the weighted means of parameters recorded in the database. Yield assessment by means of the rescaling method (Perz, 1998) and in the method of prediction presented by Jones (1997) is based on a similar idea of assembling data on all the individuals in the herd, and adjusting them to the individual assessed. The advantage of both methods is the possibility of a dynamic adjustment of the function to changes in the shape of the lactation curve.

Jones (1997) also drew attention to the necessity of proper selection of the initial point of adjustment, as it should enable the choice of the appropriate curve from among all the curves available for the herd under study. Perz (1998) stated that the first test yield should be sampled following the 12th day from the date of parturition.

CONCLUSIONS

1. Test day yields should be sampled at regular time intervals.
2. The best results of the estimation of milk yield in cows for a 100-day period are obtained using the rescaling method. However, it requires data not just on the individual tested, but also on other cows in the herd (a generalized lactation curve).
3. Among the functions that use data from the test day yields of the tested cow only, the best results of 100-day yield estimation were obtained using linear regression and the difference of two exponential functions (equation 9).
4. More complex functions are unsuccessful in yield estimation over a shortened period and a small number of test yields (4-7). The model's considerable flexibility can even be a hindrance in accurate yield assessment based on few data.
5. The function for which the greatest number of counts is obtained (enabling a precise assessment of the yield for the biggest number of cases) does not necessarily guarantee the greatest accuracy for each data set.
6. An increase in the number of yields from three to seven leads to significantly higher accuracy of estimations, although a change from five to six yields gives ambiguous results (the change of error variance does not overlap with the reported increase of counts).
7. Due to its specific character, the Michaelis-Menten function is particularly sensitive to the shape of the lactation curve; the assessment of 100-day milk yield can only be fairly accurate for cows with a typical course of the curve.

REFERENCES

BELLAMY, K. 1999. Daily Milk Yields for Genetic Evaluation, http://www.mdc.org.uk/pdfs/fact017.pdf, The Milk Development Council UK, factsheet 17.

JONES T. 1997. Empirical Bayes prediction of 305-day milk production. In: *J. Dairy Sci.,* vol. 80, 1997, no. 6, p. 1060-1075.

KOONAWOOTRITTRIRON, S. – ELZO, M. A. – TUMWASORN, S.- SINTALA, W. 2001. Prediction of 100-d and 305-d Milk Yields in a Multibreed Dairy Herd in Thailand Using Monthly Test-Day Records, In: *Thai J. Agric. Sci.*, vol. 34, 2001, p. 163-174 http://www.Animal.ufl.edu/elzo/Publications/Refereed/TJAS_Paper_2-a.htm.

MARTIN, F. F. 1976. Wstęp do modelowania cyfrowego, Państwowe Wydawnictwo Naukowe Warszawa, 134-142 (in Polish); Computer Modeling and Simulation, NewYork : John Wiley & Sons, Inc., 1968.

MASSELIN, S. – SAUVANT, P. – CHAPOUTOT, D. – MILA, D. 1989. Les modeles d'ajustment des courbes des lactation. In: *Ann. Zootech.,* vol. 36, no. 2, p. 171-206.

PERZ, W. P. 1998. Evaluation of different statistical methods for assessment of 100-day milk yield of cows on the basis of test milking. PhD thesis, Faculty of Animal Husbandry, Agricultural University of Poznan, p. 29-30.

ROSE, S. – BULLOCK, S. 1993. Chemia życia, Wydawnictwo Naukowo Techniczne Warszawa, 100-102 (in Polish); The Chemistry of life, London : Penguin Books, 1991. ISBN 83-204-1636-1.

SCOTT, T. A. – YANDELL, B. – ZEPEDA, L. - SHAVER R. D. - SMITH T. R. 1996. Use of lactation curves for analysis of milk production data. In: *J. Dairy Sci.,* vol. 79, no. 10, p. 1885-1894.

SHERCHAND, L. – MCNEW, R. W. - KELLOGG, D. W. – JOHNSON, Z. B. 1995. Selection of a mathematical model to generate lactation curves using daily milk yields of Holstein cows. In: *J. Dairy Sci.,* vol. 78, no. 11, p. 2507-2513.

THIERNEY, L. 1990. LISP-STAT, An Object-Oriented Environment for Statistical Computing and Dynamic Graphics, New York : John Wiley & Sons, Inc., p. 119-159, ISBN 978-0-471-50916-5

VARGAS, B. – KOOPS, W. J. – HERRERO, M. - VAN ARENDONK J. A. M. 2000. Modeling Extended Lactation of Dairy Cows, In: *J. Dairy Sci*., vol. 83, no. 6, p. 1371-1380.

Authors' address: Wojciech Piotr Perz, Zbigniew Sobek, Department of Genetics and Animal Breeding, August Cieszkowski Agricultural University, Wołyńska 33, 60-637 Poznań, Poland, wpperz@jay.au.poznan.pl, zbigniew@jay.au.poznan.pl; Pavel Fľak, Slovak Agricultural Research Centre, Hlohovská 2, 949 92 Nitra, Slovak Republic, flak@scpv.sk