Livestock Science 183 (2016) 28–32

Contents lists available at ScienceDirect

Livestock Science

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

Short communication

Cluster analyses to explore the genetic curve pattern for milk yield of Holstein

Rodrigo Pelicioni Savegnago [a], Guilherme Batista do Nascimento [a], Guilherme Jordão de Magalhães Rosa [b], Raul Lara Resende de Carneiro [c], Roberta Cristina Sesana [c], Lenira El Faro [d], Danísio Prado Munari [a,*]

[a] Departamento de Ciências Exatas, Faculdade de Ciências Agrárias e Veterinárias/Universidade Estadual Paulista FCAV/UNESP, Jaboticabal, São Paulo 14884-900, Brazil
[b] Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA
[c] CRV Lagoa, Sertãozinho, São Paulo 14174-000, Brazil
[d] Agência Paulista de Tecnologia dos Agronegócios (APTA) Centro Leste/Secretaria de Agricultura e Abastecimento (SAA), Ribeirão Preto, São Paulo 14001-970, Brazil

ARTICLE INFO

Article history:
Received 10 April 2014
Received in revised form
5 November 2015
Accepted 10 November 2015

Keywords:
Breeding value
Dairy cattle
Multivariate analysis
Persistency

ABSTRACT

Animal selection in dairy cattle can vary depending on the objectives of the breeding programs. The objective of this study was to explore the genetic curve pattern of EBVs for test day milk yields (TDMY) in Holstein cows using cluster analyses to identify the most suitable animals for selection based on their genetic curve for milk yield. A data set with 29,477 monthly TDMY records from 3543 first lactations of Brazilian Holstein cows were used to predict the breeding values for TDMY with random regression model. Hierarchical and non-hierarchical cluster analyses were performed based on the EBVs for 30, 60, 90, 120, 150, 180, 210, 240, 270, and 305 days in milk (DIM) to explore the genetic curve patterns of milk production of animals within the population. At first moment, the population was divided into three groups based on animals' genetic curve pattern for milk yield using hierarchical cluster analysis. According to non-hierarchical cluster analysis, one of those groups had EBVs along the lactation curve above the population average. Further cluster analysis done only with those animals with genetic curve pattern above the population mean showed specific subgroups of animals with different genetic curves for milk yield despite of all of those animals had EBVs above the population average, along the lactation curve. It indicated that specific subgroup of animals with a specific genetic curve pattern for milk yield can be chosen depending on the objectives of the breeding program. It was concluded that the cluster analyzes could be used to select animals based on the shapes of the genetic curve for milk production together with the EBV for milk yield at 305 days in milk. Thus, it can be possible to select at the same time more productive animals with genetic curves that met the goals of breeding programs that take into account the milk production in other parts along the milk production curve.

1. Introduction

Random regression models have been used for genetic evaluation of cows for test day milk yield (TDMY) (Schaeffer, 2004; Druet et al. 2005; Strabel and Jamrozik, 2006). This model is suitable for fitting the additive genetic, non-additive genetic and residuals covariance structures of quantitative traits measured along time, as milk production (Kettunen et al., 2000). The Holstein's milk production in Brazil is influenced by many environmental variations, once it is a tropical country, and for many kinds of production systems due to the particularities of local economy. Thus, the selection goals based on the genetic curve of milk production could vary depending on the region of the country in order to attend the local needs.

In general, it is desirable to select animals with higher predicted breeding values (EBVs) over the lactation curve to improve the milk yield. However, animals with different genetic curve pattern could have almost the same EBV for milk yield at 305 days in milk (EBV_{MY305}), e.g. a cow with high EBVs on the beginning and subsequent decrease of them until the end of the lactation curve

* Correspondence to: Via de Acesso Prof. Paulo Donato Castellane s/n., 14884-900 Jaboticabal, SP, Brazil.
E-mail addresses: rodrigopsa@yahoo.com.br (R.P. Savegnago), guilhermefcav@gmail.com (G.B. Nascimento), grosa@wisc.edu (G.J.M. Rosa), lenira@iz.sp.gov.br (L. El Faro), danisio@fcav.unesp.br (D.P. Munari).

http://dx.doi.org/10.1016/j.livsci.2015.11.010
1871-1413/

can have, in average, the same EBV_{MY305} as a cow with low EBVs on the beginning and high EBVs on the end of the lactation curve. A third cow could have almost the same EBV_{MY305} as the previous examples by having constant breeding values along the lactation curve in an intermediate genetic level between the EBVs of the beginning of the curve of the first cow and the EBVs of the end of the curve of the second cow.

Cluster analyses could be used to group animals based on their EBVs along the lactation to explore the genetic curve pattern for milk yield. This analysis groups similar individuals based on a set of traits, minimizing the heterogeneity of the animals within the groups and maximizing heterogeneity between groups (Hair et al., 2009). Thus, cluster analysis could help to access group of animals with similar genetic curve pattern for milk yield within the population that may be suitable for selection. The objective of this study was to explore the genetic curve pattern of EBVs for TDMY in Holstein cows using cluster analyses to identify the most suitable animals for selection based on their genetic curve for milk yield.

2. Materials and methods

2.1. Description of the data and random regression model

A data set with 29,477 monthly TDMY records from 3543 first lactations of Brazilian Holstein cows were used on the analyses. The TDMY were measured between 5 and 305 DIM divided into ten classes. The first class included the milk yield between the 5 and 30 DIM, the second included milk yield between 31 and 60 DIM, and so on until the last class, which were from 270 and 305 DIM. The pedigree contained 4288 animals, with 443 sires with 8 progenies on average.

Analyses were performed using a single-trait RR model. The model included the fixed effects of contemporary group (herd-month-year of TDMY), the covariate calving age (linear and quadratic effect), and the additive genetic and non-genetic animal random effects. A fourth-order regression on Legendre orthogonal polynomials of DIM was used to model the population-based mean curve. The fixed effects and the covariates were significant ($P<0.01$) on the monthly TDMY using the least-squares method by the GLM procedure of the SAS software (SAS 9.2, 2008). The additive genetic and non-genetic animal random effects were estimated using RR on Legendre polynomials of DIM.

The random regression model used for test-day milk yield was:

$$y_{ijk} = HMY_j + \sum_{p=1}^{2} b_p X_t + \sum_{m=1}^{4} \beta_m \varphi_m(w_t) + \sum_{m=1}^{3} \alpha_{im} \varphi_m(w_t) + \sum_{m=1}^{6} p_{im} \varphi_m(w_t) + (e_{ijk})_r$$

where y_{ijk} is the kth recorded on test day t of animal i in HYM j; HYM_j is the effect of the jth contemporary group (herd-month-year of TDMY); X is the calving age of animal i as linear and quadratic covariate ($p=1, 2$); β_m is the set of m fixed regression coefficients to model the average trajectory of the population; $\phi_m(w_t)$ is the mth Legendre polynomials of standardized day in milk t (w_t), which DIM at t were standardized in the range –1 to +1 representing 5 to 305 DIM; α_{im}, p_{im} are sets of m additive genetic and permanent environmental random regression coefficients for each animal i; $(\varepsilon_{ijk})_r$ is the residual random effect of the model associated with each test day record, where r is the number of residual classes ($r=4$; from 5 to 30 DIM; from 31 to 60 DIM; from 61 to 210 DIM; and 211 to 305 DIM).

The DIM were standardized between −1 and 1 using $w_t = -1+2[(t-t_{min})/(t_{max}-t_{min})]$ (Kirkpatrick et al., 1990, 1994), in which w_t is the standardized DIM at time t; $t=t_{min}, \ldots, t_{max}$. The residual variance was considered heterogeneous, and four classes were used: 5–30 DIM, 31–60 DIM, 61–210 DIM, and 211–305 DIM. Restricted maximum likelihood (REML) method was used to estimate the covariance components for the monthly milk production using the WOMBAT software (Meyer, 2007), by the AIREML algorithm, which enabled error calculations for the estimates of variance and heritability components (Fischer et al., 2004). Convergence was met when the change of value of the logarithm of the likelihood function in two consecutive iterations was lower than 10^{-6}.

Fourteen different RR models were compared to identify the best order for Legendre polynomials for additive genetic and non-genetic animal effects. The order of Legendre polynomials ranged from second to fourth-order for the additive genetic effect (k_a= 3–5 coefficients) and from second to sixth-order for the non-genetic animal effects (k_c=3–7 coefficients). Bayesian information criterion (BIC) (Schwarz, 1978) was used for model selection. The model with the lowest BIC value was $k_a=3$ (second-order Legendre polynomial), $k_c=6$ (fifth-order Legendre polynomial) and four classes of residual variance (from 5 to 30 DIM; from 31 to 60 DIM; from 61 to 210 DIM; and 211 to 305 DIM). So, this model was chosen to estimate the genetic parameters of TDMY. More details about the data structure, data edition, and genetic parameters estimates were described in Savegnago et al. (2013).

The EBV of the ith animal on the tth DIM were calculated as:

$$EBV_{it} = \varphi_m(w_t)'\alpha_i$$

In which $\phi_m(w_t)'$ is the transposed vector of the standardized DIM (w_t) with mth Legendre polynomials order at the tth DIM, and $\boldsymbol{\alpha_i}$ is the column vector of the additive effects of coefficients from the random regression model of the ith animal.

2.2. Cluster analyses

The cluster analyses were used to group individuals based on the EBVs for the milk yield along the lactation to explore the genetic curve pattern of this trait in each cluster. Two cluster analyses were performed to describe the additive genetic pattern of the population: hierarchical and non-hierarchical cluster analyses.

Hierarchical cluster analysis was used to choose the number of clusters into which the population could be separated. The Euclidian distance was used as distance measurement between the cows, and the Ward (1963) cluster algorithm was used to form the groups. After defining the number of groups using the previous analysis, non-hierarchical analysis using the k-means method was conducted to explore the genetic curve patterns based on the EBVs. In both analyses, the EBVs of 30, 60, 90, 120, 150, 180, 210, 240, 270, and 305 DIM were used to cluster the animals. The STATISTICA 8.0 software (Statsoft Inc., 2008) was used to carry out the cluster analyses.

3. Results and discussion

Three groups were found in the population using the hierarchical cluster analysis based on the EBVs (Fig. 1). There is no rule to choose the ideal number of clusters in this analysis. However, choosing few groups could lead no conclusion about them. Otherwise, choosing too many groups could be extremely confusing to interpret the results. Each of those groups presented different genetic curve patterns, obtained by the non-hierarchical cluster analysis (Fig. 2). There were 2084 cows in group one, 1113 into group two, and 1091 into group three, totaling 4288 cows. It was shown in Fig. 2 that the animals presented different genetic levels of milk yield within the population. The three curve patterns were almost constant along time, i.e. persistency of the EBVs, due to the selection process that the animals have been suffer along

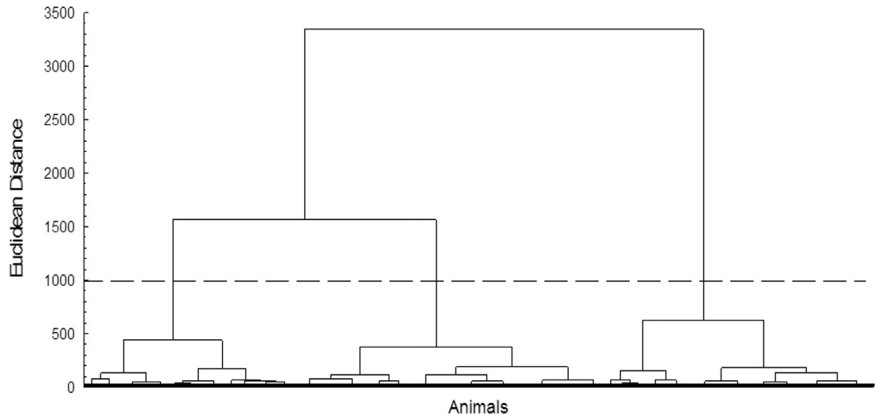

Fig. 1. Dendrogram based on the predicted breeding values (EBVs) of 30, 60, 90, 120, 150, 180, 210, 240, 270, and 305 days in milk using hierarchical cluster analysis with Ward's method.

years.

The non-hierarchical analysis could be used as a pre-screening to evaluate only the individuals desirable for selection based on their genetic curve pattern. In this case, the candidates to selection would be those classified on group 3 (Fig. 2). Although animals within groups are homogeneous due to the properties of the cluster algorithms, they could still have genetic variability. So, if desired, group 3 can be explored in more detail, looking for specific subgroups of animals, with specific genetic curve patterns for milk yield, within this group. In this second evaluation, cluster analysis was applied only with the animals classified in group 3, using the same traits in previous analyses to group the animals. The dendrogram with animals of group 3 showed that this group could be subdivided into two, four or five subgroups depend on how deep is desirable to explore the genetic curve patterns inside the group 3 (Fig. 3).

Non-hierarchical cluster analysis only with the animals of group 3 revealed subgroups of animals with different genetic curve patterns for milk yield (Fig. 4). Thus, the subgroup that would be more desirable to be considered for selection will depend on the objectives of the breeding program. Animals from subgroup 1 of Fig. 4a, subgroup 4 of Fig. 4b and subgroup 3 of Fig. 4c would probably be most suitable for selection depending on how many divisions were made in the group 3. So, the animals in those groups could be ranked by their EBVs, calculated from random regression models, to be selected after the pre-screening of the previous cluster analyses.

Some breeding programs have cows with very high peak of lactation and posterior decrease in milk production, i.e., cows with low persistency. Higher peak yields at the beginning of the lactation causes negative energy balance leading the cows to mobilize more body fat reserves to increase the demand of nutrients to produce milk (Tamminga, 2000). Thus, many problems could occur at the beginning of the lactation due to the metabolic stress. Cows with higher persistency need less feed, the health and reproduction costs are lower, and they are more profitable (Dekkers et al., 1998). Cows with flatter curve of lactation are more persistent than cows with the same total milk yield but with a curve that decreases rapidly after the peak yield (Grossman et al., 1999; Harder et al., 2006). In this case, the animals that are most indicated for selection would be the ones with genetic curve pattern similar to subgroup 2 of Fig. 4a, subgroup 1 of Fig. 4b, or subgroup 2 of Fig. 4c. After that, animals of those subgroups must be ranked by their EBVs and selected. Those genetic curve patterns could be more desirable for breeding programs that suffer with reproductive problems due to high milk peak yield.

In short, cluster analysis could be help to previous explore the EBVs of traits within the population and take into account only the animals with desirable genetic patterns that meet the objectives of the breeding program. After that, the animals would be selected based on their EBVs considering only those within the desirable cluster.

4. Conclusions

It was concluded that the cluster analyzes can be used to choose animals candidates for selection based on the shapes of the

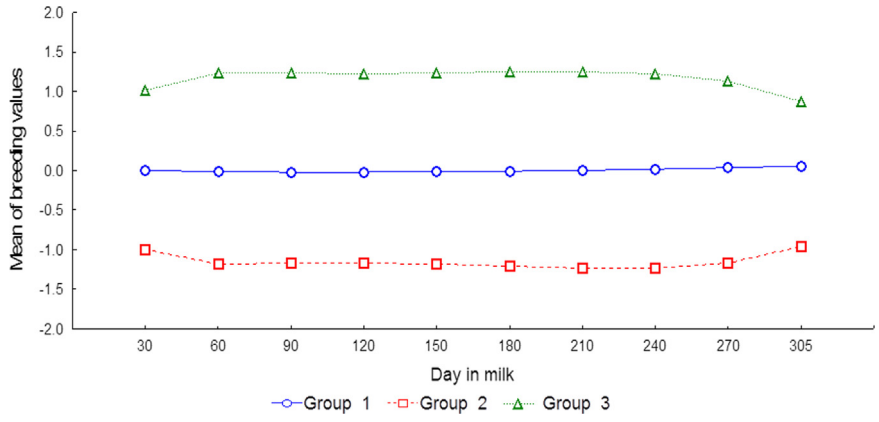

Fig. 2. Genetic curve patterns obtained by non-hierarchical cluster analysis with the K-means method using the predicted breeding values (EBVs) of 30, 60, 90, 120, 150, 180, 210, 240, 270, and 305 days in milk.

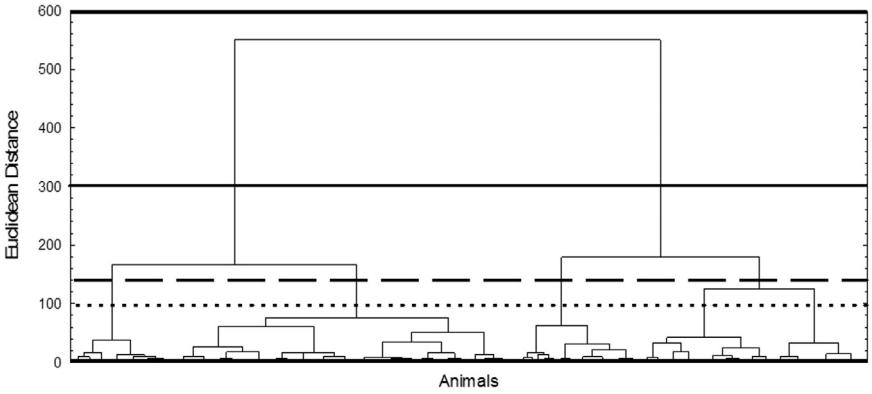

Fig. 3. Dendrogram using the Ward's method, with only the animals classified in group 3 by discriminant analysis, based on the predicted breeding values (EBVs) of 30, 60, 90, 120, 150, 180, 210, 240, 270, and 305 days in milk. The horizontal lines indicate the division into two (full line), four (dashed line) and five (dotted line) subgroups.

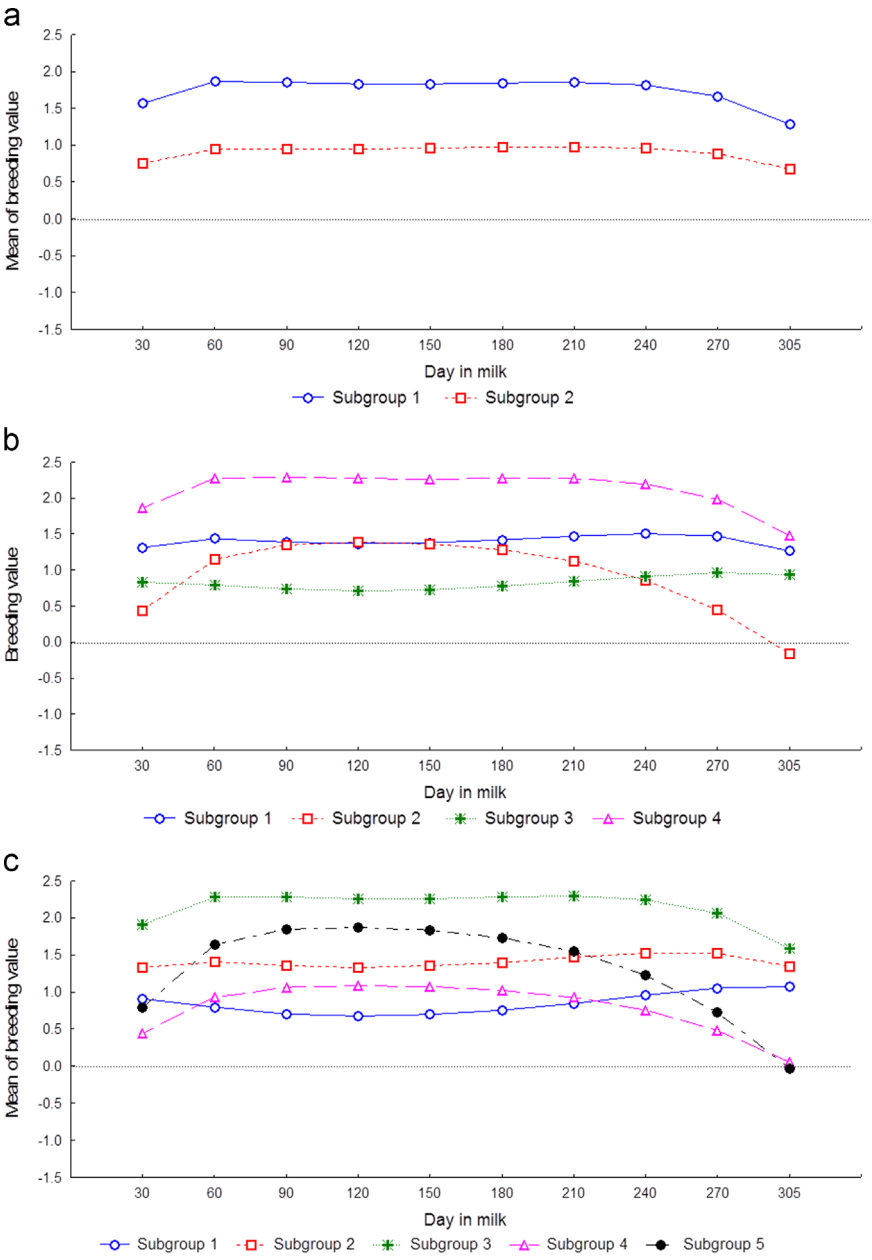

Fig. 4. Non-hierarchical cluster analyses only with the animals classified in group 3 by the previous non-hierarchical cluster analysis. Division in two (a), four (b) and five (c) subgroups.

genetic curve for milk production together with the EBV for milk yield at 305 days in milk. Thus, it can be possible to select at the same time more productive animals with genetic curves that met the goals of breeding programs that take into account the milk production in other parts along the milk production curve. The cluster analyzes should be used as an exploratory analysis to choose the best group that attend the selection goals of the breeding program and, after that, choose the best animals into the group of interest, based on their breeding values for milk yield. So, the cluster analysis should be used as complementary analysis together with the random regression.

5. Conflict of interest statement

The authors of the present study declare that they do not have any potential conflict of interest including any financial, personal or other relationships with other people or organizations within three years of beginning the work submitted that could inappropriately influence the present work.

Acknowledgments

We thank the CRV Lagoa for providing the data used in this study. R.P. Savegnago and G.B. Nascimento were Granted scholarships by the São Paulo Research Foundation (Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP Scholarships number 2010/05148-8, 2012/16087-5, 2012/23384-6, and 2013/20091-0). L. El Faro and D.P. Munari held productivity research fellowships from the National Council for Scientific and Technological Development (grant number 306888/2014-9) (CNPq; Conselho Nacional de Desenvolvimento Científico e Tecnológico).

References

Dekkers, J.C.M., Jamrozik, J., Ten Hag, J.H., Weersink, A., 1998. Economic aspects of persistency of lactation in dairy cattle. Livest. Prod. Sci. 53, 237–252.
Druet, T., Jaffrézic, F., Ducrocq, V., 2005. Estimation of genetic parameters for test day records of dairy traits in the first three lactations. Genet. Sel. Evol. 37, 257–271.
Fischer, T.M., Gilmour, A.R., Van der Werf, J.H.J., 2004. Computing approximate standard errors for genetic parameters derived from random regression models fitted by average information REML. Genet. Sel. Evol. 36, 363–369.
Grossman, M., Hartz, S.M., Koops, W.J., 1999. Persistency of lactation yield: a novel approach. J. Dairy Sci. 82, 2192–2197.
Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2009. Multivariate Data Analysis, 7th edition. Prentice Hall, Upper Sadde River, NJ, USA.
Harder, B., Bennewitzm, J., Hinrichs, D., Kalm, E., 2006. Genetic parameters for health traits and their relationship to different persistency traits in German Holstein dairy cattle. J. Dairy Sci. 89, 3202–3212.
Kettunen, A., Mäntysaari, E.A., Pösö, J., 2000. Estimation of genetic parameters for daily milk yield of primiparous Ayrshire cows by random regression test-day models. Livest. Prod. Sci. 66, 251–261.
Kirkpatrick, M., Lofsvold, D., Bulmer, M., 1990. Analysis of the inheritance, selection and evolution of growth trajectories. Genetics 124, 979–993.
Kirkpatrick, M., Hill, W.G., Thompson, R., 1994. Estimating the covariance structure of traits during growth and aging, illustrated with lactations in dairy cattle. Genet. Res. 64, 57–69.
Meyer, K., 2007. WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). J. Zhejiang Univ. Sci. B, 8815–8821.
Savegnago, R.P., Rosa, G.J.M., Valente, B.D., Herrera, L.G.G., Carneiro, R.L.R., Sesana, R. C., El Faro, L., Munari, D.P., 2013. Estimates of genetic parameters and eigenvector indexes for milk production of Holstein cows. J. Dairy Sci. 96, 7284–7293.
Schaeffer, L.R., 2004. Application of random regression models in animal breeding. Livest. Prod. Sci. 86, 35–45.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Statsoft Inc, 2008. Statistica version 8.0. StatSoft Inc, Tulsa, OK, USA.
Strabel, T., Jamrozik, J., 2006. Genetic analysis of milk production traits of Polish Black and White cattle using large-scale random regression test-day models. J. Dairy Sci. 89, 3152–3163.
Tamminga, S., 2000. Issues arising from genetic change: ruminants. In: Hill, W.G., Bishop, S.C., McGuirk, B., McKay, J.C., Simm, G., Webb, A.J. (Eds.), The Challenge of Genetic Change in Animal Production 27. British Society of Animal Science, Occasional Publication no., Edinburgh, Scotland, pp. 55–62.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.