USING CLASSIFICATION TREES IN STATISTICAL ANALYSIS OF DISCRETE SHEEP REPRODUCTION TRAITS

ZASTOSOWANIE DRZEW KLASYFIKACYJNYCH W STATYSTYCZNEJ ANALIZIE DYSKRETNYCH CECH ROZRODU OWIEC

Piwczyński DARIUSZ

University of Technology and Life Sciences, Faculty of Animal Breeding and Biology Department of Genetics and General Animal Breeding, Bydgoszcz 85-084, Mazowiecka 28, POLAND, +48 (052) 3749721, darekp@utp.edu.pl

ABSTRACT

The research material covered 2-8 year-old 6,586 Polish Merino ewes. Ewes were obtained from ten flocks located in the Pomorze and Kujawy Region (Poland). The reproductive performance index (the number of reared offspring from mated mother/year) was analyzed. The data collected were verified statistically using the classification tree technique, followed by the analysis of variance. The average number of the offspring reared by ewes was 1.208 lambs and was similar to the corresponding data presented in the applicable literature reports on Polish Merino sheep breed. To identify the factors which were responsible for the variation in the number of offspring reared by the mated ewe, the classification tree technique was applied. As a result of this statistical procedure, the ewe lambing information set was divided according to the factors which demonstrated the highest power of the 'importance' index: mother's age, flock, birth type and the body weight of ewes at the age of 12 months. The effect of the factors identified with the use of classification tree technique on the offspring number per mated ewe was significant, which was seen from the analysis of variance. The results suggest that the classification tree technique can be used to provide the statistical analysis of discrete reproduction traits.

Keywords: sheep/reproductive performance/classification trees

STRESZCZENIE

Materiał badawczy stanowiło 6,586 owiec matek rasy merynos polski w wieku 2-8 lat. Maciorki pochodziły z dziesięciu stadach zlokalizowanych w regionie Pomorza i Kujaw (Polska). Analizowano wskaźnik użytkowości rozpłodowej (liczba odchowanego potomstwa od pokrytej matki/rok). Zgromadzone dane zostały statycznie opracowane za pomocą techniki drzew klasyfikacyjnych a następnie analizy wariancji. Średnia liczba odchowywanego potomstwa przez maciorki wyniosła 1,208 sztuki i była zbliżona do odpowiednich danych prezentowanych w literaturze naukowej przedmiotu w odniesieniu do rasy merynos polski. W celu wyłonienia czynników odpowiedzialnych za zmienność liczby potomstwa odchowanego przez maciorkę pokrytą zastosowano technikę drzew klasyfikacyjnych. W efekcie tego postępowania statystycznego został dokonany podział zbioru informacji o wykotach maciorek ze względu na czynniki, które charakteryzowały się najwyższą siłą wskaźnika" importance": wiek matki, stado, typu urodzenia i masa ciała maciorek w wieku 12 miesięcy. Wpływ czynników wyłonionych z użyciem techniki drzew klasyfikacyjnych na liczbę uzyskanego potomstwa od pokrytej maciorki został statystycznie potwierdzony za pomocą analizy wariancji. Uzyskane wyniki sugerują, że technika drzew klasyfikacyjnych może być wykorzystana w celu statystycznej analizy cech reprodukcyjnych o charakterze dyskretnym.

Słowa kluczowe: owce/użytkowość rozpłodowa/drzewa klasyfikacyjne

DETAILED ABSTRACT IN POLISH

Materiał zwierzęcy stanowiło 6586 owiec matek rasy merynos polski w wieku od 2 do 8 lat użytkowanych rozpłodowo w latach 1993-2000. W celu wyłonienia czynników determinujących zmienność liczby uzyskanego od maciorki potomstwa, wykorzystano drzewa klasyfikacyjne zbudowane w oparciu o wskaźnik zróżnicowania Gini (SAS) [2]. Wykonana analiza za pomocą drzew klasyfikacyjnych wykazała, jako istotnie różnicujący, podział na 3 grupy wiekowe matek o odmiennych proporcjach analizowanej cechy. Procentowy udział odchowanych bliźniąt w poszczególnych grupach wahał się od 20,08 % (przystępki) do 35,30 % (maciorki w wieku 4 i więcej lat). Kolejny podział, jaki dokonał się został przeprowadzony w oparciu o zmienną zależną stado. Spośród 9 powstałych węzłów potomków najbardziej korzystne wyniki stwierdzono w odniesieniu do podzbioru, maciorek najstarszych użytkowanych w stadach: B, D, E, F, J, K (Node 13) – udział wykotów mnogich wyniósł ponad 44 %. Dalsze podziały miały miejsce już tylko w odniesieniu do węzłów 12 i 13. W pierwszym z nich, w dalszym ciągu zmienną grupującą okazało się ponownie stado. Podział ten doprowadził do powstania 3 podzbiorów (Node 32, 33, 34). Spośród nich najlepsze wskaźniki rozrodu stwierdzono w stadzie A (Node 32). Jeśli chodzi o węzeł 13 to podział, jaki w nim się dokonał utworzył podzbiór obserwacji pochodzących od maciorek urodzonych jako jedynaczki i bliźniaczki. Maciorki z bliźniąt uzyskały bardziej korzystne wyniki rozrodu niż przystępki. W grupie maciorek najstarszych, użytkowanych w stadzie M (Node 34) kolejnym czynnikiem grupującym okazała się masa ciała maciorek w wieku 12 miesięcy – największy udział wykotów bliźniaczych stwierdzono w grupie maciorki o masie ciała do 55, 5 kg. W węźle 35 nastąpił ponadto podział na 2 grupy o różnej masie ciała, tj. poniżej 48,5 kg i o masie równej lub wyższej od 48,5 kg. Podział węzła 36 utworzył 3 grupy stad: K, D (Node 54), B, E, J (Node 55) i F (Node 56). Wykonana w oparciu o uzyskane reguły trójczynnikowa analiza wariancji potwierdziła istotność różnic między liczbą uzyskanego potomstwa w badanych grupach oraz wykazała wysoko istotne interakcje między stadem a wiekiem oraz między stadem a typem urodzenia maciorek (tab. 2).

INTRODUCTION

The basic factor determining the sheep production profitability in Poland is the number of full-value offspring obtained per mother in the flock. Considering a single breeding season, therefore, we deal with a discrete trait which usually assumes the following values: 0, 1, 2, 3. In earlier reports by Piwczyński [12], this type of traits was exposed to probit transformation, and then a multi-factor analysis of variance was made. And in the case of statistical analysis of binary reproduction traits, Piwczyński [13] utilized multivariate logistic regression analysis. Urioste and Danell [18] suggested that litter size in sheep may be distributed as a Poisson distribution. In such a case generalized linear models with log link function can be very useful [17]. Whereas Matos et al [7] modeled genetic variability of reproduction traits, not only by means of nonlinear Poisson model but also by means of threshold model and negative binomial distribution.

The objective of this paper was to establish factors responsible for the number of lambs reared from the fertilized mother with the use of two alternative statistical methods, i.e. modern techniques of classification trees and analysis of variance.

The classification trees are modern analytic techniques which are data-mining group tools [1, 2, 3, 6]. They allow for building graphic easily-comprehensible models used to describe and to predict the phenomenon expressed in both the nominal and the ordinal scale. The classification trees can also be used for the purpose of preliminary selection of the traits which have a statistical effect on the dependent variable. As part of the graphic tree model created, there is a recurrent division of the observation set into 'n' disjoint subsets. The aim of the divisions is obtaining such subsets which are maximum homogenous due to the values of the dependent variable. At every stage of this multi-trait division of the data set, there can be used different independent variables, and the selected variable is the one which ensures the best division of the node, that is the one which creates the most homogenous sets. The structure of the classification tree starts with the entire information set (root node) (Fig. 1). The subsets which emerge as a result of division are referred to as child nods. The final subsets which are not exposed to further divisions are called leaves. The number of leaves determines the tree size, while the number of edges between the tree top and the most distant leaves informs about the tree depth. The classification trees have already been applied in financial management [2], medicine [1] as well in animal farming [15]. As for animal farming, Sawa et al. [15] used the classification tree technique to determine the genetic-and-physiological-and-environmental parameters which ensure obtaining milk rich in protein and poor in somatic cells.

MATERIALS AND METHODS

The animal material was made up of 6,586 Polish Merino sheep mothers used for breeding over 1993-2000. The animals evaluated were 2- to 8-year-old and were obtained from ten flocks located in the Pomorze and Kujawy Region. The evaluation involved a total of 21,121 production seasons. The variety in the number of the offspring reared by the mated ewe was evaluated depending on: the flock (A-1378 samples, B-891, D-918, E-764, F-616, G-4154, J-1309, K-3071, L-3431, M-4589), ewe birth year (1985-1998), lambing year (1993-2000), ewe age (2-8 year-old), birth type of ewe and its parents (1, 2), ewe rearing type (1, 2), production sector (public, private), ewe body weight at the age of 70 days and 12 months.

To identify the factors which were responsible for the variation in the number of offspring reared up to the age of 100 days by the mated ewe, the classification tree technique was used. To do so, Enterprise Miner 4.3 (SAS) and Tree Desktop Application were used [16]. Of the 21,121 observations analyzed, 67 % was used to create the training set, while the other 33% of the data was a validation set. It was assumed that the minimum size of the final node cannot be lower than 100 observations and the tree depth cannot be bigger than 6. The missing values were allocated to one of the branches. In reference to each node formed, there was given a percentage of cases qualified to respective groups (0, 1, 2 lamb reared and the number of observations per node). In the algorithm of the tree division, as a division rule, Gini index (G(p)) was used [2]. The index is a measure of variability for categorical data, and a measure of node impurity at the same time.

$$G(p) = \sum_{j<k}^{r} p_j p_k \;,$$

where p_1, p_2, p_r – the relative frequencies of each target class in a node.

The ranking of variables, as for the weight they played in the created model of population division was made with the measure 'Importance' (SAS)[16](Fig. 2). At the further step with three-factor analysis of variance [17] there was verified the effect of factors identified earlier thanks to the classification tree mechanism on the number of the offspring obtained: mother's age, flock, ewe birth type, as well as accompanying variable – the body weight of ewe at the age of 12 months. Due to the discrete character of the number of reared lambs, this character was earlier exposed to probit transformation. The analysis of variance, considering the interaction of the first degree between factors was made using the data already transformed (GLM procedure - SAS) [17].

RESULTS

The average number of the offspring per ewe in the production cycle was 1.208, while the standard deviation for that trait – 0.601 (tab. 1). The graphic effect of the analysis in the tree form is given in Fig. 1. The classification tree has 17 leaves and is 4 layers deep, uses 4 independent variables (Fig 1). Figure 2 demonstrates the ranking of "Importance" of independent variables (0 to 1 scale). The table lists the input variable name (Variable), the number of nodes that use the input variable as the primary splitting rule (Nodes), the measure of importance computed from the training (Training) and validation data set (Validation). In the "Importance" column, the top horizontal bars represent the training estimates and the bottom bars represent the validation estimates of variable importance. The highest results of the ranking show the greatest effect of the independent variable on the dependent variable. According to the ranking, the age of mothers was the factor most differentiating the data set, and then the flock was qualified. The differentiating power of the birth type and the body weight of ewe was similar.

The first tree splits was made based on the age of mothers, which created three age groups: primparous (Node 2), three-year-old mothers (Node 3), four-year-old mothers and older (Node 4). The statistics which cover the oldest mothers demonstrate that in that group there was the biggest number of twin lambing (35.30%) and at the same time the lowest number of infertile seasons (8.96%). The

Table 1. Statistical characteristic of reproductive performance of ewes
Tabela 1. – Charakterystyka statystyczna użytkowości rozpłodowej maciorek

Variables Zmienne	n	Mean Średnia	Standard deviation Odchylenie standardowe
Reproductive performance (lambs/ewe mated) Użytkowość rozpłodowa (sztuk/matkę)	21,121	1.208	0.601

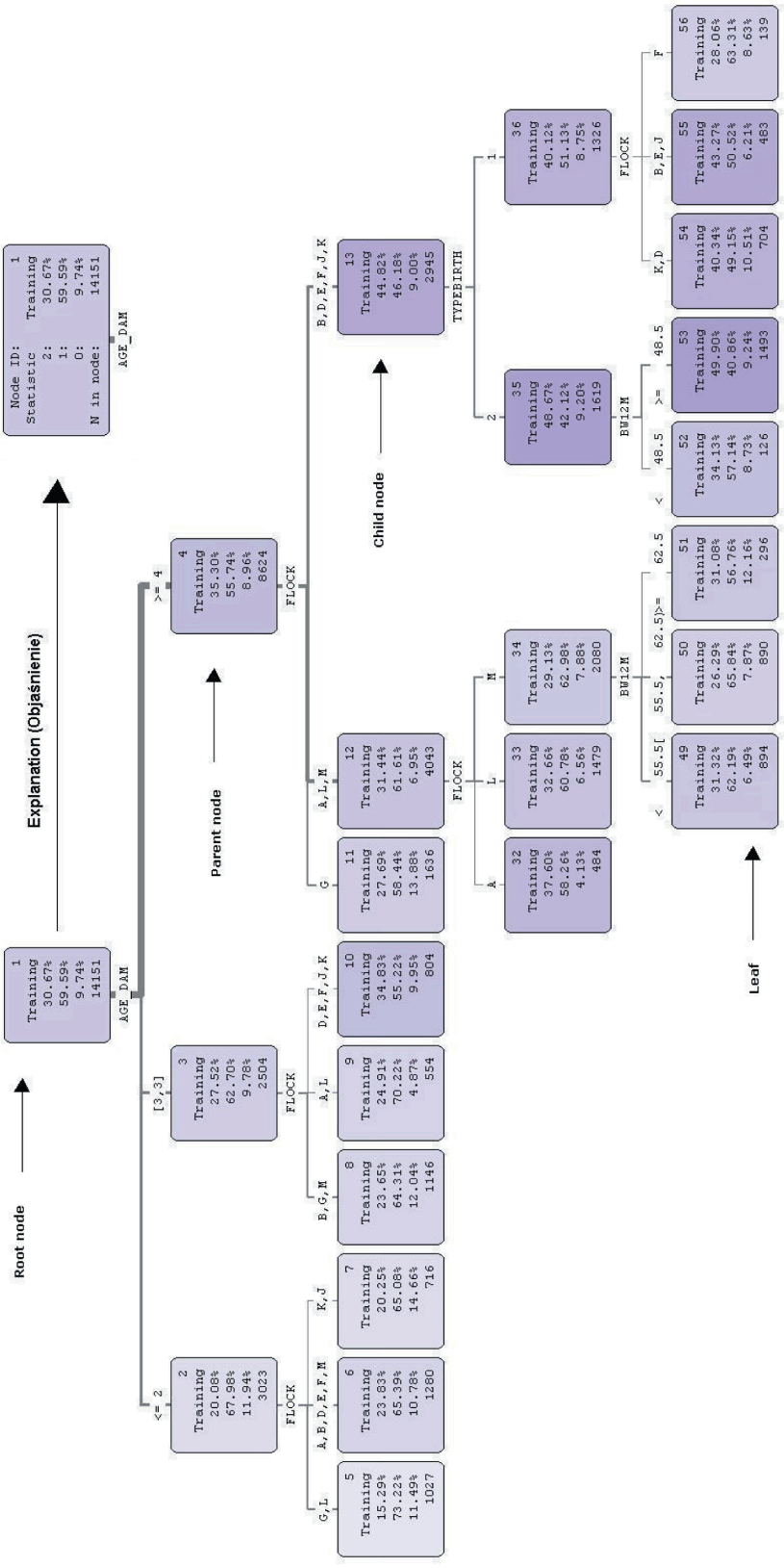

Figure 1. Reproductive performance – structure of population
Rycina 1. Użytkowość rozpłodowa – struktura populacji

worst results were recorded for primparous, respectively, 20.08% and 11.94%. Another division was made based on the dependent variable: the flock. Of 9 offspring nodes formed the most favorable results were reported for the subset, the oldest ewes used in flocks: B, D, E, F, J, K (Node 13) – the share of multiple lambing over 44 %. Further divisions were only reported for nodes 12 and 13. In the first of them, still the grouping variable was again the flock. This division led to the creation of 3 subsets (Node 32, 33, 34). Of them, the best reproduction indices were found for flock A (Node 32). As for node 13, the division created the subset of observations from ewes born as single litter and twins. Ewes from twins obtained more favorable reproduction results than from singles. In the group of the oldest ewes, used in flock M (Node 34), the next grouping factor was the body weight of ewes at the age of 12 months – the greatest share of twin lambing was identified in the group of ewes of the body weight up to 55.5 kg (Node 35). In node 35 there was also recorded the division into 2 groups of a different body weight, that is below 48.5, and the weight equal or higher than 48.5 kg. The division of node 36 created 3 groups of flocks: K, D (Node 54), B, E, J) (Node 55) and F (Node 56).

The analysis of variance made with the probit transformation of data demonstrated a significant effect of independent variables identified with the classification tree method on the number of reared offspring up to the age of 100 days. The analysis of variance also showed highly significant interactions between the age of the mother and the flock and the birth type (Table 2).

DISCUSSION

In Poland the reproduction performance index is usually calculated as the quotient of the number of reared offspring and the number of all the mothers per flock [4]. In the present research, however, the analysis involved the number of reared offspring up to the age of 100 days, but by each mated mother. The average number of reared offspring was 1.208 (tab. 1). It is an index similar to the upper border of the reproduction performance value reported in literature for Polish Merino sheep breed in the Kujawy and Pomorze Region (1.19 – 1.20 lambs/mother) [4, 5, 11, 12, 14].

The results of studies reported by many authors show that the poorest reproduction indices should be expected among primparous, but with the age these abilities improve [8, 9, 12]. Patkowska-Sokoła and Barczyńska [9] found that the most favorable lambing for Merino sheep mothers, as for the fertility and productivity, are 4 and 5. Piwczyński [12] demonstrated that 4 and 5 lambing is most optimal, also as for reproduction performance. The results reported by the above authors [8, 9, 12] correspond to the present results generated with classification trees

Table 2. Analysis of variance results
Tabela 2. Wyniki analizy wariancji

Source of variation - Źródło zmienności	F value - Wartość F	p
Flock – Stado (F)	21.79	<.0001
Age dam - Wiek matek (A)	37.71	<.0001
F*A	1.93	<.0001
Type birth – Typ urodzenia (T)	34.36	<.0001
F*T	5.06	<.0001
A*T	1.01	0.4141
Body weight at 12 mths Masa ciała w wieku 12 miesięcy	11.06	0.0009

Variable	Nodes	Training	Validation	Importance
AGE_DAM	3	1.000	0.960	
FLOCK	3	0.943	1.000	
BW12M	2	0.391	0.302	
TYPEBIRTH	1	0.296	0.408	

Figure 2. Variables importance
Rycina 2. Ważność zmiennych

(Fig. 1).

The reproductive performance traits of sheep are low-heritability [14], and for that reason their variation is strongly affected by the environment conditions. A typical factor identified with the effects of the environment is the flock. In the present research the flock appeared to be a very important factor affecting the number of the offspring reared. A significant effect of the flock on reproductive performance traits was noted in e.g. the reports by Kowaliszyn and Mroczkowski [5] as well as in earlier reports of the present author [11, 12].

The effect of the ewe birth type on the number of offspring is not quite cleat-cut. Kowaliszyn and Mroczkowski [5] and Niedziółka [8] showed a significant effect of the birth type on productivity and reproduction performance of ewes. On the other hand, in the earlier reports by the author [12] there were found no significant differences between single litter and twins in reproduction performance, however, similarly as in the reports by other authors [5, 8] more favorable results were reported by twins. The graphic tree model (Fig. 1) facilitates understanding what could be the reason for varied results reported by different authors. It seems that the effect of the birth type of ewe on its reproduction indices is additionally conditioned by other variables, including: the ewe age upon lambing, flock or the body weight.

In the present research the factor creating data subsets was the ewe body weight. The present results suggest that good reproduction indices can be recorded for four-year or older ewes which at the age of 12 months showed a low body weight (Node 49). The reports by Pięta and Patkowski [10] show that the reproduction indices, including fertility, productivity, lamb rearing, can be an effect of the body weight and the condition of the ewe, but concerning the period right before the tupping. A good condition expressed as a higher body weight results in a more favorable reproduction index. In the present research it was observed that the effect of the body weight at the age of 12 months can be important also in further breeding seasons.

The decision tree technique identified independent variables which can affect the number of reared offspring, that is the ewe age, flock, birth type and the ewe weight at the age of 12 months. The effect of these variables was identified to be significant also with the analysis of variance, which can demonstrate that this method can be used interchangeably with traditional methods, e.g. the analysis of variance.

Linear models describing the variability of a dependant trait usually take into account the first degree interactions between factors. Whereas the actual cooperation between factors might be more complex. An excellent tool to understand this seems to be the very classification tree technique. Classification trees can present in an easy-to-understand graphical form even the most elaborate divisions or interactions, for instance, in the said research the influence of the type of birth on the number of progeny was revealed solely in the oldest mothers for the flock groups B, D, E, F, J, K. The advantage of data mining over the traditional methods, such as variance analysis, is also due to the fact that they also make possible to examine variability of discrete reproduction depending on the traits expressed in the continuous, ordinal or nominal scale. Classification tree techniques can be used for transformed data as well.

The research carried out allows us to conclude that the classification tree technique can be used interchangeably with traditional methods, e.g. the analysis of variance. At the same time, the present research demonstrated the best reproduction indices can be expected in the group of ewes from twin litter whose body weight at the age of 12 months was of at least 48.5 kg, of at least 4 year old, and used in flocks B, D, E, F, J, K, while the poorest results were represented in primparous in flocks G, L.

REFERENCES

1. Abu-Hanna A., de Keizer N., Integrating classification trees with local logistic regression in Intensive Care prognosis. Artificial Intelligence in Medicine. (2003) 29, 5–23.

2. Feldman D., Gross S., Mortgage default: classification trees analysis. The Pinhass Sapir Center for Development Tel-Aviv University. Discussion Paper. (2003) 3, 1-46.

3. Gatnar E., Nieparametryczna metoda dyskryminacji i regresji. PWN, 2001.

4. Hodowla Owiec i Kóz w Polsce, ROCZNIKI 1995-2003 (1996-2004) PZO, Warszawa.

5. Kowaliszyn B., Mroczkowski S., Influence of ewe's type birth and type of her parents birth on chosen production traits of sheep. Roczniki Naukowe Polskiego Towarzystwa Zootechnicznego. (2005) 1, supl. 2, 83-90.

6. Łapczyński M., Drzewa klasyfikacyjne w badaniach satysfakcji i lojalności klientów. StatSoft Polska. (2003) 93-102.

7. Matos C.A.P., Thomas D.L., Gianola D., Tempelman R. J., Perez-Enciso M., Young L.D., Genetic analysis of discrete reproductive traits in sheep using linear and nonlinear models: I. Estimation of genetic parameters. Journal of Animal Science. (1997) 75, 76-87.

8. Niedziółka R., Pieniak-Lendzion K., Influence

of age and birth type on reproductive performance of Berrichonne du Cher ewes. Roczniki Naukowe Zootechniki. (2005) 21, 45-49.

9. Patkowska-Sokoła B., Barczyńska E., Influence of age of Polish Merino ewes on their reproductive rate. Prace i Materiały Zootechniczne. (1985) 36, 45-51.

10. Pięta M., Patkowski K., Reproductive performance of the Uhrusk ewes, depending on the body weight and body condition before mating season. Zeszyty Naukowe Przeglądu Hodowlanego. (2002) 63, 19-25.

11. Piwczyński D., Selected traits of reproductive performance of Polish Merino sheep. Zeszyty Naukowe Przeglądu Hodowlanego. (2003) 70, 59-63.

12. Piwczyński D., Reproduction performance of merino sheep depending on the number of lambing and type of birth. Prace Komisji Nauk Rolniczych i Biologicznych BTN. (2004) 53, B, 167-172.

13. Piwczyński D., Mroczkowski S., Application of logistic regression for analysis of some reproduction traits on sheep. Roczniki Naukowe Polskiego Towarzystwa Zootechnicznego, (2005), 1, supl. 2, 49-58.

14. Piwczyński D., Mroczkowski S., Kowaliszyn M., Włodarczak M., Genetic parameters of reproduction traits of Polish Merino sheep estimated using different linear models. Zeszyty Naukowe Przeglądu Hodowlanego. (2004) 72 (3), 15-21.

15. Sawa A., Kowaliszyn B., Piwczyński D., Application of ID3 algorithm in choice of genetic-physiological-environmental parameters conditioning obtaining milk with low somatic cell count and high protein content. Zeszyty Naukowe Przeglądu Hodowlanego. (2004) 72, 1, 41-49.

16. SAS Institute Inc., Decision Tree Modeling. Course Notes. Cary, NC: SAS Institute Inc., 2001.

17. SAS Institute Inc. SAS/STAT(r) 9.1 User's Guide. Cary, NC: SAS Institute Inc., 2004.

18. Urioste J., Danell Ö., Variation in sheep litter size interpreted with the treshold model. In: I. Urioste Licentiate thesis. Report 74. Dept of Animal Breeding & Genetics. Uppsala, Sweden. (1987).