Archiv Tierzucht 55 (2012) 4, 332-345, ISSN 0003-9438

Statistical modelling of somatic cell counts using the classification tree technique

Dariusz Piwczyński and Beata Sitkowska

Department of Genetics and General Animal Breeding, Faculty of Biology and Animal Breeding, University of Agriculture and Life Sciences in Bydgoszcz, Poland

Abstract

The research studied a sample of 455 Polish Holstein-Friesian Black and White cows. Its aim was to apply and compare two modern statistical methods, i.e. classification trees and a logistic regression in examination of the impact of selected lactation-related factors (successive lactation, herd size and production level, year of calving, calving season, test day season, lactation phases and the amount of milk obtained in a test milking) on the somatic cell counts. Two different division criteria were taken into account in the creation of classification trees, i.e. entropy reduction and Gini coefficient. The quality of classification trees and multiple regression models was compared taking into consideration the following criteria: an average squared error, cumulative lift, Kolmogorov-Smirnov statistics and the area under the ROC curve. Having conducted the research, it may be concluded that from among the statistical methods applied, the best modelling of the level of somatic cell counts was obtained using the classification tree technique when the division criterion was based on the entropy function. According to the results of the study, somatic cell counts were diversified by the following factors, in a decreasing order of importance: herd production level, year of calving, subsequent lactation, calving season, day of test milking, herd size and the month used to take milk samples. Using somatic cell count as an udder health benchmark, it may be concluded that cows requiring particular attention as a result of udder diagnosis are from those in herds with high milk production levels, with individual cows producing up to 15 kg of milk.

Keywords: somatic cell count, dairy cattle, data mining, logistic regression

Introduction

The reasons underlying variation in somatic cell counts (SCC) in cows' milk are widely discussed in the literature and a plenty of research have been conducted to indicate factors affecting SCC level (Baltay 2002, Koç & Kizilkaya 2009, Król *et al.* 2010). The most significant changes in SCC level are considered to originate from the presence or absence of bacterial infection inside the udder (Burston & Erskine 2003) and other factors that have a role to play are as follows: influence of a cowshed, cattle's age, stage and season of lactation, metabolic causes, physiological »stress« and daily variation (Olsen *et al.* 1999, Kuczaj 2001, Baltay 2002, Green *et al.* 2006, Sitkowska 2008, Hagnestam-Nielsen *et al.* 2009, Koç & Kizilkaya 2009, Król *et al.* 2010, Sitkowska & Piwczyński 2011).

In dairy cattle farming, a high level of somatic cells is associated with cows' susceptibility to mastitis (Bradley 2002, Burston & Erskine 2003, Forsbäck 2010). This inflicts losses resulting from an early cows' culling (Samoré *et al.* 2003), lowered profitability of milk production and deteriorated quality for processing (Halasa *et al.* 2007, Forsbäck 2010). Mastitis is one of the most frequent and costly diseases in the dairy industry (Halasa *et al.* 2007, Huijps *et al.* 2008). The susceptibility of cows to mastitis is one of the more important functional problems of dairy cattle farming and one that affects the profit to be made from milk production. Mastitis causes the greatest losses in dairy farming, arising from the increase in somatic cells in milk, which results in a lowered hygienic quality. At the same time, a decrease in milk production and an alteration in chemical components of the milk itself are observed. The frequency of mastitis increased for Holstein population due to genetic antagonism between milk production and clinical mastitis resistance (Gulyas & Ivancsics 2001). The presence of somatic cells in cows' milk can be indicative of udder health and milking hygiene, which translates into cows' health and the quality of milk (Baltay 2002, Bradley 2002, Seegers *et al.* 2003, Koç & Kizilkaya 2009, Skrzypek *et al.* 2003).

Clinical cases of mastitis occur due to a number of factors. It is possible to apply various graphical methods to model this problem (Bradley 2002, Gasqui & Barnouin 2003, Heringstad *et al.* 2006, Němcová *et al.* 2007). In a study performed by Miciński *et al.* (2009), the percentage of cows with mastitis increased in successive lactations and during each lactation. More than one third of cows suffered from mastitis. Despite decades of research in this field, mastitis continues to be a widespread problem in dairy farming (Seegers *et al.* 2003, Hagnestam *et al.* 2007, Forsbäck 2010).

In many countries, SCC is used as an indirect examination criterion for improving mastitis resistance (Fahr 2002, Kalm 2002, Kühn *et al.* 2008, Interbull 2010). In Poland and other European Union countries, a requirement laid down for raw milk for consumption establishes a 400 000 somatic cell count per cm^3 as the upper limit (European Commission 2004). Samples to be found in the second group may indicate a clinical condition of the udder. As a result of such a division, each sample may be giving information on udder health, i.e. »normal« or »abnormal«. From the statistical point of view, it is therefore a binomial variable.

Various statistical methods are used for analysing traits expressed on a binomial scale, e.g. the χ^2 test, a logistic regression (Nash *et al.* 1996, Binns *et al.* 2002, Piwczyński 2009). According to recent studies (Piwczyński *et al.* 2011), a possible alternative method which can be used is the classification tree technique, which belongs to the analysis representation area known as data mining. An analysis of clusters and artificial neuron networks are distinguished in this area of statistics (SAS Institute Inc. 2009).

The classification tree technique has already been used in human medicine (Austin 2007). Its suitability in the field of sheep farming for modelling lamb mortality has been proved (Piwczyński *et al.* 2011). One very significant advantage of using the tree structure is an easy interpretation of data through the graphic representation of a model depicting very complex effects arising from various factors and the interplay within those factors, by which high-order interactions may take place, influencing visual traits measured on the nominal and ordinal scale of the tree presentation. Owing to the classification tree technique, it is viable to indicate clearly the best and worst solutions available, i.e. present a set of these levels of individual experimental factors that guarantee the lowest and highest level of a trait (Piwczyński *et al.* 2011).

It may be thus concluded that the proposed method can be utilised by those researchers who study an alteration in SCC level (but also other production traits measured on a bracket, ordinal or nominal scale). Furthermore, the method allows tracing the differences in SCC level depending on the factors that are recorded in each herd undergoing assessment.

The aim of the present paper was to apply and compare modern statistical methods: classification trees and a logistic regression in examination of the impact of selected lactation-related factors (successive lactation, herd size and production level, year of calving, calving season, test day season, lactation phases and the amount of milk obtained in a test milking) on the somatic cell counts.

Material and methods

The research studied a sample of 455 Polish Holstein-Friesian Black and White cows born in the years 1999-2000 and bred in 2003-2008. The cattle under the study were from highly efficient herds located in the Kujawy and Pomerania region. The cows in the study were hybrids of Polish Black and White lowland cattle and the Holstein-Friesian breed. The Holstein-Friesian proportion in the cows' genotype exceeded 94 %. 8 868 milk samples from daily flows test day (TD) were checked in terms of somatic cell quantity. A numerical data used in the present paper was taken from a cow milk herd management application »Obora« (Zeto, Olsztyn, Poland).

In order to determine somatic cell counts, the Fossomatic method was applied. In accordance with the order of the Minister for Agriculture and Rural Development (2004), milk samples were divided to reflect the number of somatic cells, into milk of less than 400 000 grade (from the »normal« medical condition of udder) and milk containing more than 400 000 somatic cells – resulting from a sub clinical (i.e. »below normal«) condition. Consequently, a variable expressed on a binomial scale was obtained.

The following factors which could cause higher content of somatic cells were taken into consideration: successive lactation (1, 2, 3, ≥4), herd size (≤30, 31-50, >50 cows), herd production level (≤6 000, >6 000 ≤8 000, >8 001 ≤9 000, >9 000 kg), year of calving (2003-2008), calving and TD season (spring (III, IV, V), summer (VI, VII, VIII), autumn (IX, X, XI), winter (XII, I, II), month of TD, lactation phases (DIM-day in milk) and the amount of milk (kg) obtained in a test milking (≤15, >15 ≤30, >30 ≤45, >45).

In the first stage of the statistical analysis, a percentage distribution of milk samples below and above 400 000. SCC was calculated depending on herd size and herd production level, year of calving, calving season, TD season, lactation phases, amount of milk. The statistical analysis of correlations was carried out using the χ^2 test (SAS Institute Inc. 2009). Then, descriptive statistics were calculated with reference to somatic cell count. A statistical analysis of the effect of the experimental factors on somatic cell counts was conducted utilising classification trees.

From among the 8 868 observations analysed, 60 % were used to create a »training set«, and the remaining 40 % constituted a »validation set«. The sets were generated by means of simple random sampling. The training data is used for preliminary model fitting, whereas the validation set is used to prevent a modelling node from over-fitting the training data and to compare prediction model.

It was assumed that the minimum size of the final node (branch of the tree) must not be less than 30 observations and the depth (number of branches) no higher than 6. The criteria for a leaf size and depth were set in such a way so as to avoid an inaccurate over fitting (manipulation) of the tree to the training data, which could reflect random relations within the validation set. Missing values were assigned to one of the branches.

Building graphical models of trees results in separable, T-shaped subsets being created through a recurrent split of the observation set. This is carried out to achieve subsets with maximum homogeneity with respect to the value of a dependant variable. Optionally, the same independent variables can be used at different stages of the multi-trait split of the data set. The variable selected is the one ensuring the best splitting of a node, i.e. the most homogeneous sets are generated. So as to generate a classification tree, the entire data set, known as the root node, is indispensable. The nodes that follow, resulting from splitting, are called child nodes. When a subset is not subject to any further division, it is called a leaf.

Two different division criteria were taken into account when generating classification trees – entropy reduction (ENTROPY) (1) and Gini index (GINI) (2). The criteria above are the basic measurements of homogeneity of sets that are used in the data mining technique for generating classification trees. (SAS Institute Inc. 2009).
Entropy function:

$$H(p_1, p_2, ..., p_k) = -\sum_{j=1}^{k} p_j log_2(p_j) \quad (1)$$

Gini index (G(p)):

$$G(p) = 1-\sum_{j=1}^{k} p_j^2 \quad (2)$$

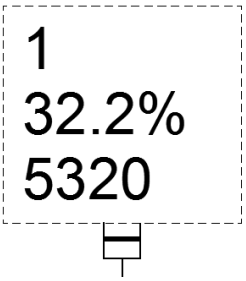
where p is the probability vector of object assignment to classes in the form of:

$$\boldsymbol{p} = (p_1, p_2, ..., p_k) = \left(\frac{1_1}{n}, \frac{1_2}{n}, ..., \frac{1_k}{n}\right) \quad (3)$$

k is the number of classes, l is the size of class and n is the size of analysed population (cows).

Both calculated parameters, i.e. entropy function, as well as the Gini index, share one property connected with the fact that these measures take on the value of zero when a trait distribution is focused on a single value. The higher the values they take, the more diverse a population is in terms of the examined trait.

The following information was added to each created node and a resulting leaf: node ID [1], percentage of samples indicating a clinical condition of udder (SCC level: ≤400 000, >400 000) [2], number of observations in a node or leaf [3] (Figure 1).



[1]

[2]

[3]

Figure 1
Description of a node 1 – root node

The ranking of variables in terms of their importance in creating data set splits was prepared based on the »importance« measure (SAS Institute Inc. 2010). This measure (I(v,T)) is calculated as a square root of the product of the Gini coefficient reduction (ΔGini), calculated based on the main and surrogate splits which use the *v* variable in all nodes (*T*) comprising the tree, and the »agreement« measure (*a(sv, t)*) for the rule using the v variable in the t node:

$$I(v,T) = \sqrt{\sum_{t=1}^{T} \Delta Gini(s(v,t))a(s_v,t)} \tag{4}$$

where *s(v, t)* is the best surrogate split in the *t* node using the *v* variable

$$a(s_v,t) = \begin{cases} 1 & \text{if } s_v \text{ is the main splitting rule} \\ agreement & \text{if } s_v \text{ is the surrogate splitting rule} \\ 0 & Otherwise \end{cases} \tag{5}$$

The »agreement« measure takes values from the (0,1) range and it designates the percentage of agreeing cases when the main and surrogate splits are compared. »Importance« measures calculated for particular variables were divided by the »importance« variable with the highest importance.

At the following stage of the statistical analysis, the effect of the aforementioned factors on the level of SCC was examined using a multiple logistic regression (SAS Institute Inc. 2009). Variables statistically related to SCC were selected by means of the forward method. The variables were included in the model at the significance level equal 0.05. Models generated this way, as a result of adding new variables, were compared using the Akaike Information Criteria (AIC). The significance of parameters in the created model was assessed with the Wald statistics (SAS Institute Inc. 2009).

The quality of information displayed on the tree models and multiple regressions was compared taking into consideration the following criteria: an average squared error, cumulative lift, Kolmogorov–Smirnov statistics and the area under the receiver operating characteristic (ROC) curve (SAS Institute Inc. 2009). The criteria listed above are the key ones used in assessing the quality of classification trees generated, but also for the purposes of a logistic regression. Decreasing values of the average squared error and increasing values of cumulative lift, Kolmogorov-Smirnov statistics, and the area under the ROC curve indicate a higher quality in any given model. A statistical analysis was conducted using the Enterprise Miner 6.2 software included in the SAS package (SAS Institute Inc. Cary, NC, USA).

Results and discussion

Somatic cell count

The conducted examination revealed that the mean number of somatic cells count (SCC) was 666010 cells/ml. In 50 % of the examined samples, the somatic cell count level did not exceed 200000 cells/ml. In the next 25 % milk samples, SCC was not higher than 573000 cells/ml (Table 1).

The experimental factors included in own research were of interest to other authors. It should be emphasised that the majority of them is included in the Polish model for estimating cattle's breeding value with reference to SCC (National Research Institute of Animal Production 2011).

Table 1
Statistic characteristic of the somatic cell count

No. of cows	455
No. of milk samples	8 868
Mean, cells/ml	666 010
Coefficient of variability, %	228.92
Lower quartile, cells/ml	82 500
Median, cells/ml	200 000
Upper quartile, cells/ml	573 000

From the analysis, it may be concluded that the highest level of somatic cell counts is to be expected in the samples from the fourth and subsequent lactations in herds of 30 to 50 animals, in herds with a production over 305 days of 8 000 to 9 000 kg of milk, from cows calved in 2008 and from cows calved during the winter months, but used for milk production in the summer (Table 2).

Table 2
Udder health in respect of investigated factors

Factor	Level	n	≤400 000, %	*P*
Successive lactation	1	3 121	75.49	<0.0001
	2	2 508	66.51	
	3	1 765	65.84	
	≥4	1 474	56.04	
Herd size	≤30	609	66.83	0.196
	>30 ≤50	2 666	66.58	
	>50	5 593	68.48	
Herd production level	≤6 000	1 798	80.87	<0.0001
	>6 000 ≤8 000	895	62.12	
	>8 000 ≤9 000	2 088	61.64	
	>9 000	4 087	66.43	
Year of calving	2 003	3 170	75.36	<0.0001
	2 004	2 008	65.84	
	2 005	1 723	68.02	
	2 006	1 254	60.21	
	2 007	688	52.91	
	2 008	25	40.00	
Season of calving	Spring (SG)	2 150	67.95	<0.0001
	Summer (SR)	2 322	71.27	
	Autumn (AN)	1 800	68.17	
	Winter (WR)	2 596	64.29	
Season of TD	Spring (SG)	2 157	67.69	0.141
	Summer (SR)	2 208	65.99	
	Autumn (AN)	2 290	68.34	
	Winter (WR)	2 213	69.14	
Amount of milk obtained in a test milking, kg	≤15	1 505	71.03	<0.0001
	>15 ≤30	3 592	63.70	
	>30 ≤45	2 888	68.70	
	>45	883	75.99	

SG, SR, AN, WR: marking used on tree diagrams

Models comparison

Based on a measure of quality of the fitting of a statistical model to data, it may be concluded that the classification tree model built, based on the reduction of entropy and Gini coefficient, were characterised by favourable statistical values, assessing the quality of a model rather than any logistic regression model (Table 3). In the case of both classifications of tree models, quality values were similar, with the more favourable found in the tree built based on the entropy function.

The ROC index values (0.656-0.675) obtained mean that the forecasting capability of all models is moderate. A similar assessment of the value of this index was obtained in the area of sheep farming through earlier studies on lamb collapse (Piwczyński *et al.* 2011) with the use of classification trees (0.645-0.648) and a logistic regression (0.609-0.634). Austin (2007), however, cites higher values for the ROC index, when modelling mortality among patients, caused by acute myocardial infarction (0.779-0.849).

Table 3
Model comparisons

Statistics Label	Entropy	Gini	Logistic regression
Average Squared Error	0.1994	0.1996	0.2030
Cumulative Lift	1.2486	1.2445	1.1581
Kolmogorov-Smirnov Statistic	0.2630	0.2530	0.2470
Roc Index	0.6750	0.6720	0.6560

In Table 4, the values of the »importance measure« are presented. This measure carries information concerning the significance of particular variables to be expected in the course of creating classification tree models. The variable ranking obtained in relation to both tree models leads to the conclusion that the amount of milk from test milking, herd production level and the year of lactation are the most significant differentiating factors. In the case of the model built based on the entropy function, the subsequent factors, ranked according to their importance were the following: subsequent lactation, calving season, test milking day, herd size and the month of the test milking. In the other model of the tree we created, the ranking was slightly different; however, calving season, subsequent lactation, herd size, test milking day and the month of the test milking were significant in their importance.

When referring the results presented in Table 4 to those in Table 2, it has been generally demonstrated that the variables considered when generating a classification tree were also indicated as statistically related to SCC using the χ^2 test.

Table 4
Variables importance

Factor	Entropy	Gini
Amount of milk obtained in a test milking	1.000	1.000
Herd production level	0.931	0.811
Year of calving	0.883	0.879
Successive lactation	0.823	0.569
Season of calving	0.621	0.592
DIM	0.349	0.261
Herd size	0.348	0.329
Month of TD	0.210	0.176
TD season	0.000	0.000

Table 5
Summary of forward selection

Step	Variable	Walda χ^2	*P*	AIC
1	Year of calving	138.2953	<0.0001	6 554.1
2	Herd production level	155.5448	<0.0001	6 396.6
3	Amount of milk obtained in a test milking	154.4595	<0.0001	6 251.6
4	Season of calving	19.3646	0.0002	6 237.5

Using a logistic regression to analyse the level of SCC, the following variables in the order of their inclusion into the model were established: year of calving, herd production level, amount of milk obtained in the test milking and the calving season (Table 5).

Construction of classification tree

The graphical effect of statistical analysis using the classification tree technique is shown in Figures 2 to 5. The information provided therein concerns the set utilised to create the (training set) tree. The tree, as presented, was built with the use of the entropy function as the division criterion. The resulting tree consists of 27 leaves, and is 6 levels deep.

The herd production level (Figure 2) proved to be the most significant factor in the diversification of SCC level. Based on its value, milk samples were divided into those from herds with up to 6 000 kg (node 2), and over 6 000 kg (node 3) of milk output, over a 305-day lactation. In the former group, the percentage of samples indicating an upper level of SCC was 18.4 %, and in the second it was nearly twice as high – 35.6 %. According to Hagnestam *et al.* (2007), high milk yield results in a predisposition to clinical mastitis. In samples taken from cows calving since 2005 (node 7), the frequency of increased number of somatic cells was higher by approximately 20 percentage points as compared to those cows calving before and in 2005 (node 6).

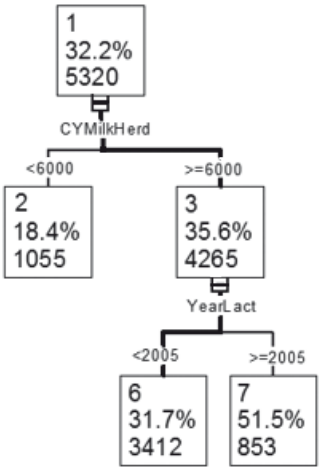


Figure 2
Splits of node 1 – first split

Marking used on tree diagrams: sMilk: amount of milk obtained in a test milking, CYMilkHerd: herd production level, YearLact: year of calving, Lact: successive lactation, SeasCalv: season of calving, DIM: day in milk, SizeHerd: herd size, MthMilk: month of TD, SeasMilk: TD season

In Figure 3, further divisions of node 2 can be seen. These refer to milk samples from herds in which production over a 305-day lactation was no higher than 6 000 kg of milk. Using the

division algorithm of the classification tree, a set consisting of samples from cows calved in autumn (node 4) and other seasons (node 5) was created. The percentage of milk samples with increased somatic cell content in the first collection was twice as low as in the second one (node 4 vs. 5). Analysing subsequent node divisions, it was found that the highest percentage of upper level of SCC was present in cows in their third lactation and subsequent lactations, from herds of less than 100 animals – 37.1 % (node 20). At the same time, it was observed that the lower level of SCC among all subsets of the entire classification tree was present in samples from the most numerous herds, in which cows calved for the third and subsequent times – 1.5 % (node 17). According to Król *et al.* (2010), the increase in the somatic cell count in cows' milk caused a progressive decrease in daily milk yield. Oleggini *et al.* (2001) observed that the larger herds had lower somatic cell counts compared with smaller herds.

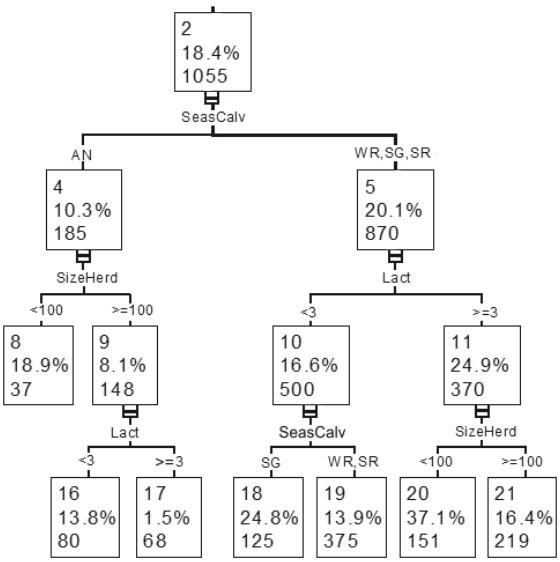


Figure 3
Splits of node 2

The number of somatic cells in milk samples obtained from cows used in herds with production of over 6 000 kg and calving before the year 2005 (node 6) were most strongly diversified by the amount of milk obtained in the sample milking (Figure 4). In the samples of obtained milk, with the maximum amount of 30 kg (node 12), the percentage of lower level of SCC by 13.4 percentage points more than in the case of amounts higher than 30 kg (node 13). A negative impact of subsequent lactation on the level of somatic cells was observed. In samples obtained from multiparas, the number of conditions of increased number of somatic cells approached 49 % (node 23). This relationship was also observed by Miciński *et al.* (2009) – a high level of somatic cells that may be indicative of mastitis was diagnosed on day 81 after calving in the third lactation, while in primiparous cows, the symptoms of this disease were observed two months later. Clinical mastitis was also more frequent in cows which had calved more than once rather than in cows which had only calved once in a study of Hagnestam *et al.* (2007) and Hagnestam-Nielsen *et al.* (2009). Sawa *et al.* (2007) reported that the lowest (13 %) proportion of cows with recurrent mastitis was found in herds of less than 10 cows, whereas the proportion of cows with recurrent mastitis in larger herds (100-200 cows) increased to 30 %.

Further division of node 23 led to the formation of a sub consideration of milk samples, in which the frequency of higher level of SCC was approximately 84 % (node 73). These were samples with the maximum quantity of milk at 15 kg, obtained between June and December. The increasing percentage of cows with recurring mastitis in subsequent lactations may result from the increase of somatic cells in milk as cows progress in age, which was also reported in a study by Sawa *et al.* (2007). Cows that have calved more than once, especially in late lactation, displayed a larger number of somatic cells in the research of Olechnowicz & Jaśkowski (2010).

The subsequent lactation also diversified the SCC level in the case of node 24, namely milk samples of over 30 kg from cows used in herds with lactation yield of less than 9 000 kg. Milk from cows which have calved more than once (node 47) was characterised by a frequency of increased SCC condition twice as high as in heifer cows (node 46). With node 6 divisions, the DIM variable, i.e. the day of test milking, was used in the tree division algorithm. It was observed that milk samples obtained at an early stage of lactation (nodes 74 and 76) were characterised with lower SCC as compared with those at a more advanced stage (nodes 75 and 77). In the research reported by Sitkowska (2008), it was observed that the lowest content of somatic cells was found in the test yields at the beginning of lactation and as lactation progressed, the content of somatic cells in milk increased.

The research showed a similar effect in a herd of heifer cows production level (nodes 68-71). A higher level of SCC was observed in samples from herds producing over 9 000 kg of milk (nodes 69, 71) as compared with a lower output (68, 70). It is worth noting that when milk samples of over 30 kg (node 13) were divided, the trend was reversed (nodes 24, 25).

Figure 5 shows the division of samples which most frequently confirm a higher level of SCC, i.e. those obtained later than in 2004 from herds with an average production of over 6 000 kg of milk (node 7). The most important factor differentiating this set of samples was the quantity of milk obtained in one milking. It was observed that a high frequency of samples with an increased number of somatic cells was accompanied by the output of under 45 kg – approximately 57 % (node 14). The corresponding frequency in samples with the quantity of milk of less than 45 kg was lower by approximately 24 percentage points (node 15). Another factor differentiating the two subsets (nodes 14 and 15) was the calving season. In both cases, the subsets identical in terms of their structure were created, i.e. samples from cows calving in the summer season (nodes 26 and 29) as well as in the other seasons – nodes 27 and 28. It was shown that calving in summer was clearly associated with lower frequency of samples indicative of higher level of SCC, as compared with the other seasons. In the research by Kuczaj (2001), the lowest mean content of somatic cells was noted in cows calving in autumn and the highest in those calving in summer similar as in the research of Green *et al.* (2006). In the studies by Neja & Sawa (2006), the highest proportion of samples with the SCC indicating clinical or subclinical mastitis was found from September to November. Olsen *et al.* (1999), examining data from among the Norwegian cattle population, established that the lowest content of somatic cells found in milk was from cows calving in the summer and early autumn.

In further subdivisions of the above four subsets, the following factors were taken into consideration: the quantity of obtained milk (twice), the test milking day and again the calving season. Among the created subsets, a subset was formed containing samples with

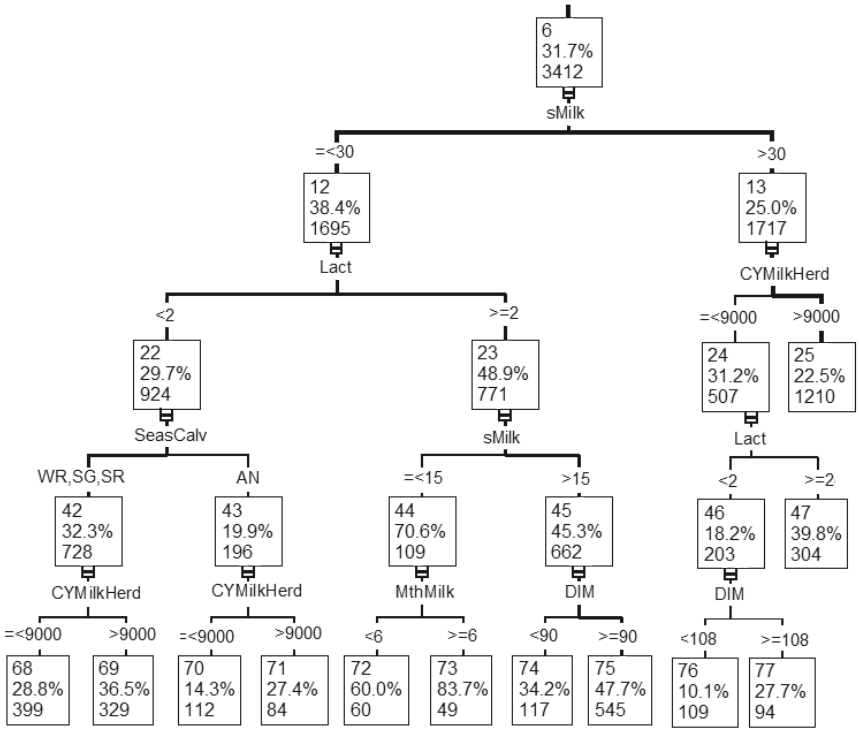


Figure 4
Splits of node 6

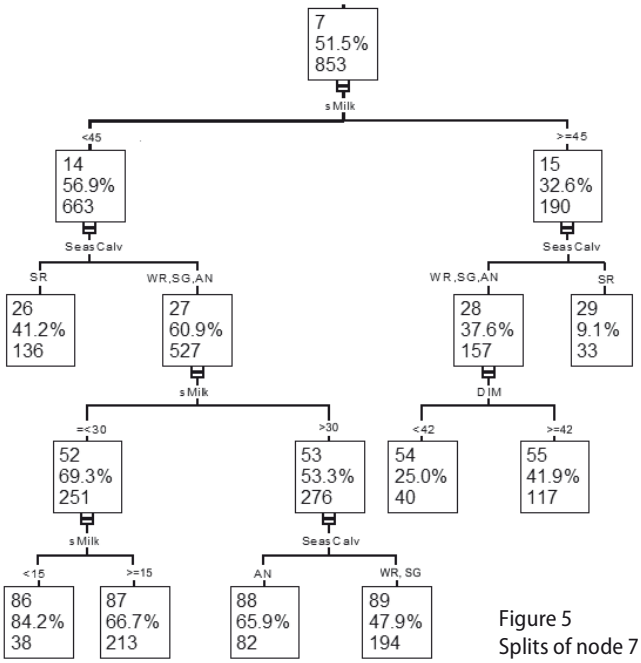


Figure 5
Splits of node 7

the highest frequency (84.2 %) of increased SCC – node 86. It comprised the samples which, in terms of milk output, were no higher than 15 kg, and which came from cows calving in the spring, autumn, or summer and in herds with an average milk output of over 6 000 kg.

Having conducted the research it may be concluded that from among the statistical methods applied, the best modelling was obtained using the classification tree technique when the division criterion was based on the entropy function.

The research showed that somatic cell counts were diversified by the following factors, in an decreasing order of importance: herd production level, year of calving, subsequent lactation, calving season, day of test milking, herd size and the month used to take milk samples.

Using somatic cell count as an udder health criterion, it may be concluded that cows requiring particular attention as a result of udder diagnosis are from those in herds with high milk production levels, with individual cows producing up to 15 kg of milk.

The graphical classification tree model we created indicates that there exists, among dairy herds, a very complex condition for the understanding of the level of somatic cell counts, proving therefore that the statistical techniques applied could have a practical application in this field. The statistical model put forward in the present research can be applied in cattle farming with an aim of quick and effective tracking of an increased SCC in milk.

References

Austin PC (2007) A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med 26, 2937-2957

Baltay Z (2002) Influence of time of day the milk and season on the somatic cell count under Hungarian conditions. Arch Tierz 45, 349-357

Binns SH, Cox IJ, Rizvi S, Green LE (2002) Risk factors for lamb mortality on UK sheep farms. Prev Vet Med 52, 287-303

Bradley A (2002) Bovine mastitis: an evolving disease. Vet J 164, 116-128

Burston JL, Erskine RJ (2003) Immunity and mastitis. Some new ideas for an old disease. Vet Clin North Am Food Anim Pract 19, 1-45

European Commission (2004) Regulation (EC) No 853/2004 of the European parliament and of the council of 29 April 2004 laying down specific hygiene rules for food of animal origin. OJ EU L 47, 206

Fahr RD (2002) [Influencing milk quality and composition: options and limits]. Arch Tierz 45 SI, 51-59 [in German]

Forsbäck L (2010) Bovine udder quarter milk in relation to somatic cell count. Diss. Uppsala. Sveriges lantbruksuniv., Acta Universitatis agriculturae Sueciae, 1652-6880

Gasqui P, Barnouin J (2003) Statistical modelling for clinical mastitis in the dairy cow: problems and solutions. Vet Res 34, 493-505

Green MJ, Bradley AJ, Newton H, Browne WJ (2006) Seasonal variation of bulk milk somatic cell counts in UK dairy herds: investigations of the summer rise Prev Vet Med 74, 293-308

Gulyas L, Ivancsics J (2001) [Relationships between the somatic cell count and certain uddermorphologic traits]. Arch Tierz 44, 15-22 [in German]

Hagnestam C, Emanuelson U, Berglund B (2007) Yield losses associated with clinical mastitis occurring in different weeks of lactation. J Dairy Sci 90, 2260-2270

Hagnestam-Nielsen C, Emanuelson U, Berglund B, Strandberg E (2009) Relationship between somatic cell count and milk yield in different stages of lactation. J Dairy Sci 92, 3124-3133

Halasa T, Huijps K, Østerås O, Hogeveen H (2007) Economic effects of bovine mastitis and mastitis management: a review. Vet Q 29, 18-31

Heringstad B, Gianola D, Chang YM, Ødegård J, Klemetsdal G (2006) Genetic associations between clinical mastitis and somatic cell score in early first-lactation cows. J Dairy Sci 89, 2236-2244

Huijps K, Lam TJ, Hogeveen H (2008) Costs of mastitis: facts and perception. J Dairy Res 75, 113-120

Interbull (2010) Description of National Genetic Evaluation Systems for dairy cattle traits as applied in different Interbull member countries. http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm [last accessed 23.07.2012]

Kalm E (2002) Development of cattle breeding strategies in Europe. Arch Tierz 45, 5-12

Koç A, Kizilkaya K (2009) Some factors influencing milk somatic cell count of Holstein Friesian and Brown Swiss cows under the Mediterranean climatic conditions. Arch Tierz 52, 124-133

Król J, Brodziak A, Florek M, Litwińczuk Z (2010) Effect of somatic cell counts in milk on its quality depending on cow breed and season. Annales UMCS 28, 9-17

Kuczaj M (2001) Interrelations between year season and raw milk hygienic quality indices. Electronic Journal of Polish Agricultural Universities, 4 Available from www.ejpaumedia.pl (accessed Jan 24 2011)

Kühn C, Reinhardt F, Schwerin M (2008) Marker assisted selection of heifers improved milk somatic cell count compared to selection on conventional pedigree breeding values. Arch Tierz 51, 23-32

Miciński J, Pogorzelska J, Barański W, Kalicka B (2009) Effect of disease incidence on the milk performance of high-yielding cows in successive lactations. Pol J Nat Sci 24, 102-112

Minister for Agriculture and Rural Development (2004) [Regulation of the Ministry of Agriculture and Rural Development of 18 August 2004 on veterinary requirements for milk and dairy products]. J Laws No. 188, item 1946 [in Polish]

Nash ML, Hungerford LL, Nash TG, Zinn GM (1996) Risk factors for perinatal and postnatal mortality in lambs. Vet Rec 139, 64-67

National Research Institute of Animal Production (2011) [Estimation of breeding bulls PHF variety]. http://www.wycena.izoo.krakow.pl/doc/metody_oceny_2011_1.pdf [last accessed 23.07.0212] [in Polish]

Neja W, Sawa A (2006) Cytological quality of milk from cows kept in different types of pens, according to the season of the year. Arch Tierz 49 Special Issue, 238-243

Němcová E, Štípková M, Zavadilová L, Bouška J, Vacek M (2007) The relationship between somatic cell count, milk production and six linearly scored type traits in Holstein cows. Czech J Anim Sci 52, 437-446

Oleggini GH, Ely LO, Smith JW (2001) Effect of region and herd size on dairy herd performance parameters. J Dairy Sci 84, 1044-1050

Olsen I, Lindhardt E, Ebbesvik M (1999) Effects of calving season and sire's breeding value in a dairy herd during conversion to ecological milk production. Livest Prod Sci 61, 201-211

Piwczyński D (2009) Using classification trees in statistical analysis of discrete sheep reproduction traits. J Cent Eur Agric 10, 303-309

Piwczyński D, Sitkowska B, Wiśniewska E (2011) Application of classification trees and logistic regression to determine factors responsible for lamb mortality. Small Rum Res 103, 225-231

Samoré AB, del Schneider P, Canavesi F, Bagnato A, Groen AF (2003) Relationship between somatic cell count and functional longevity assessed using survival analysis in Italian Holstein-Friesian cows. Livest Prod Sci 80, 211-220

SAS Institute Inc. (2009) Getting Started with SAS Enterprise Miner 6.1. SAS Institute Inc., Cary, NC, USA

Sawa A, Neja W, Bogucki M (2007) Relationships between cytological quality and composition of milk and the effect of some environmental factors on the frequency of recurrent mastitis in cows. J Cent Eur Agri 8, 295-299

Seegers H, Fourichon C, Beaudeau F (2003) Production effects related to mastitis and mastitis economics in dairy cattle herds. Vet Res 34, 475-491

Sitkowska B (2008) Effect of the cow age group and lactation stage on the count of somatic cells in cow milk. J Cent Eur Agric 9, 57-62

Sitkowska B, Piwczyński D (2011) Impact of successive lactation, year, season of calving and test milking on cows' milk performance of the Polish Holstein-Friesian Black-and-White breed. J Cent Eur Agric 12, 283-293

Skrzypek R, Wójtowski J, Fahr RD (2003) Hygiene quality of cow bulk tank milk depending on the method of udder preparation for milking. Arch Tierz 46, 405-411

Received 2 August 2011, accepted 16 March 2012.

Corresponding author:

Dariusz Piwczyński
email: darekp@utp.edu.pl

Department of Genetics and General Animal Breeding, Faculty of Biology and Animal Breeding, University of Agriculture and Life Sciences in Bydgoszcz, Mazowiecka Street 28, 85-084 Bydgoszcz, Poland