Bayesian Network Analysis for the Factors Affecting the 305-day Milk Productivity of Holstein Friesians

Volkan SEVİNÇ[a], Özge AKKUŞ[a], Çiğdem TAKMA[b], Öznur İŞÇİ GÜNERİ[a]

[a]*Muğla Sıtkı Koçman University, Faculty of Science, Department of Statistics, Kötekli, Muğla, TURKEY*

[b]*Ege University, Faculty of Agriculture, Department of Animal Science, Biometry and Genetics, Bornova, İzmir, TURKEY*

AUTHORS ORCID ID
(Volkan SEVİNÇ: 0000-0003-4643-443X), (Özge AKKUŞ: 0000-0002-3077-0896), (Çiğdem TAKMA: 0000-0001-8561-8333), (Öznur İŞÇİ GÜNERİ: 0000-0003-3677-7121)

ABSTRACT

The variables affecting the milk productivity have been discussed in various articles through different methods. A recent study using path analysis shows that three variables significantly affect the 305-day milk yield of Holstein Friesian cows. These variables are parity, first calving year and lactation length. Calving season is another variable which appears to be significant in a different study. The aim of this study is to provide a simultaneous multilateral analysis among the milk yield, these three variables and a new variable calving season. The analysis was realized through a Bayesian network built over the findings of the path analysis. 17,109 records of Holstein Friesian cows calved between 2001-2011 years were analyzed. The estimated Bayesian network showed that younger cows produced more milk. Lactation length and parity do not depend on each other. Cows reached their highest amount of milk yield on their 4th parities. Milk yield is mostly affected by lactation length. Finally, first calving year, parity, lactation length and calving season should be considered as criteria in a selection study to increase the milk yield.

Keywords: 305-day milk yield; Holstein Friesians; Bayesian networks; Structure learning

1. Introduction

Yield is an important economic trait for dairy cattle. In order to reach a good level of productivity, factors affecting milk yield and identifying and analyzing their effects are important issues. In the literature, various studies about milk yield exist. Perochon et al (1996) have modeled the lactation curves of dairy cows with an emphasis on individual variability. They have found that the quality of the individual prediction was better for French Friesian and Montbeliarde cows than for pure or crossbred Holstein cows. İnci et al (2007) have investigated the milk yield and reproductive traits of Brown Swiss cattle raised at Altınova State Farm and they found that the milk yield is highly affected by lactation length and period. Aktürk et al (2010) have studied on the factors affecting milk yield and milk production cost and they have found that silage maize (20%) and barley (14%) have the biggest effect on the milk yield. Tahtalı et al (2011) performed a path analysis to examine the factors affecting milk productivity of Brown Swiss cows and they calculated the percentages of direct effects of factors on actual milk yield as 17.3% for lactation length and 1% for 305-day milk yield. A study performed by İşçi et al (2015) showed that milk yield was significantly affected by three variables: parity, first calving year and lactation length. They applied a path analysis to indicate the significance of those three variables affecting the 305-day milk yield of Holstein

Friesian cows. They also provided information about some bilateral relations among these variables. Iqbal et al (2016) have proposed a manually constructed Bayesian network model which assumes that the body condition parameter, the number of days after calving, and parity number affect the estrous cycle of the cattle. Using this model, they analyzed the dairy cattle productivity. Mote et al (2016) examined the effect of different macroclimatic variables on lactation milk yield and lactation length of Holdeo (Holstein Friesian x Deoni) crossbred cattle. Their regression analysis showed that calving season significantly affects the lactation milk yield. Eetvelde et al (2017) studied the factors associated with first-lactation milk yield in Holstein Friesian heifers and they found that the season of birth, but not calving, had a significant influence on the milk yield. Verma et al (2017) reported that, cows reach their highest amount of milk on their 4th parities. Eastham et al (2018) examined the associations between age at first calving and subsequent lactation performance of UK Holstein Friesian cows. They reported that lower first calving age leads to more amount of milk yield. However, there is not a previous work in the literature concerning milk productivity through a Bayesian network model.

Bayesian statistics assumes that parameters are not constants but variables. Moreover, in this approach, along with the sample information (likelihood), any existing information or experience about the parameter to be estimated are taken into account in a distribution form which is briefly called prior information. After combining the prior information with the likelihood function obtained from the observations, the posterior distribution is achieved. Therefore, using a Bayesian approach in a milk yield estimation study, one can add the results obtained in previous studies into the analysis to improve the quality of the estimations. Hence, the aim of this article is to provide a detailed multilateral analysis of the relationships among milk yield, these three variables affecting milk yield and a new variable calving season as well as to introduce the use of the Bayesian network model in milk productivity.

2. Material and Methods

The material of this study consisted of the 305-day milk record of the Holstein Friesian cows that calved between 2001-2011 years. 17,109 records, obtained from 1840 herds belonging to the members of Cattle Breeders Association in Isparta province, Turkey, were analyzed by building an appropriate Bayesian network.

Bayesian networks are a kind of probabilistic graphical models based on the Bayes' formula. Using a Bayesian network model, all conditional dependencies can simultaneously be calculated for a set of variables. Moreover, any prior knowledge about the relations among the variables can be added into the analysis. Bayesian networks correspond to the Directed Acyclic Graph (DAG) structure. Hence, a Bayesian network model has two main parts. The first part is the graphical part including the nodes (variables) with non-overlapping levels and directed arrows (edges) among the nodes. The second part consists of the conditional probabilities (parameters) usually displayed in a table called Conditional Probability Table (CPT). A CPT of a node gives the probability distribution of the variable represented by that node, conditional on its parent node. A directed arrow from one node to another indicates dependence among the variables represented by those two nodes. Any node from which an arrow comes out is called the parent node and the node that the arrow goes in is called the child node. All nodes following a path formed by the direction of certain arrows are called relative nodes. Among the relative nodes, the nodes on the path behind a certain node are called ancestors, and the ones afterwards are called descendant nodes. In a DAG, it is impossible to return to the same node following the path of an arrow directed from that node which means it is impossible for a node to be its own ancestor or descendant. Nodes are independent of their non-descendant nodes, given the state of their parents. An example of a Bayesian network model is presented in Figure 1. The Bayesian network shown in Figure 1 has five nodes named C1, C2, C3, C4 and C5. The conditional probabilistic relations among them are represented by the directed arrows. For example, C1 is a parent node of C2, while C2 is a child node of C4. It is also seen that, in accordance with the nature of the DAG structure, it is not possible to return to the same node following the path of an arrow directed from that node.

Let $X = (X_1, X_2, \ldots, X_n)$ be the set of variables of a Bayesian network. Then, a Bayesian network simply represents a joint probability function such as

$$P_B(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i \backslash \pi_i) \tag{1}$$

Where; X_i are the nodes representing variables and π_i is any conditioned state of a node or an evidence. Evidence can also be defined as an observed or occurred value of a variable. Inference in Bayesian networks is made by the propagation of the evidences through the network. For any evidence, π_i the joint posterior probabilities $P_B(X_1, X_2, ..., X_n)$ in Equation 1 can be obtained.
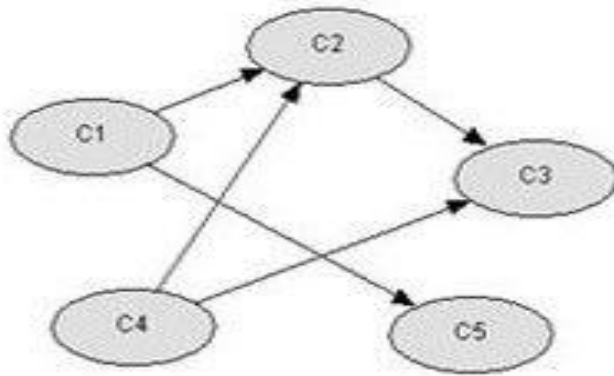

Figure 1- Example of Bayesian network

Building a Bayesian network takes two main tasks: construction of the graphical structure of the network and calculation of the posterior probabilities. There are three main approaches to building the graphical structure of Bayesian networks. In the first case, dependencies among the variables, that is, directions of the arrows are determined by a specialist of the subject manually. In the second approach, however, the structure of the network is learned directly from the data set through some learning algorithms. This type of building a Bayesian network is called structure learning. The third approach called the hybrid approach is a combination of the first two ones. In this type of building a Bayesian network, a specialist constructs the graphical structure of the network manually then an algorithm is used to learn the parameters or the specialist provides background information to an algorithm. For more detailed information about the construction of Bayesian networks, see Jensen (2001).

In structure learning, there are various algorithms for building a Bayesian network. These algorithms can mainly be classified as constraint-based and score-based algorithms. Constraint-based algorithms are based on an algorithm called Inductive Causation algorithm by Verma & Pearl (1991). This type of the algorithms decides for the existence and direction of an edge through some conditional independence tests. More detailed information about constraint-based algorithms exists in Spirtes et al (1993), Cheng et al (2002), de Campos & Huete (2000). Score-based algorithms choose the graph structure having the highest score, after calculating a score value for all possible candidate Bayesian network structures. Detailed information about score-based algorithms can be found in Cooper & Herskovits (1991), Heckerman (1998), Chickering (2002).

In general, constraint-based algorithms are less time consuming and more powerful algorithms compared to the score-based algorithms (Natori et al 2015). There are various computer packages that can be used to build a Bayesian network such as GeNIe (2019), Hugin (2019), Netica (2019). To measure the estimation performance and uncertainty of Bayesian networks, there are various metrics. Marcot (2012) provides detailed information about those metrics and their applications. Among these metrics, the logscore value can be used as a performance evaluation measure for a Bayesian network and it is calculated as follows.

$$logscore(X_1, ..., X_n) = \frac{\sum_{i=1}^{n} log_2 P(X_i/model)}{nN} \tag{2}$$

Where; n is the number of cases in the test set and N is the total number of nodes in the model. The highest the

logscore is, the better is the model. A sensitivity analysis for a Bayesian network model can determine how much the nodes are affected by the changes in the other nodes on the network. Model sensitivity can be measured by entropy scores. Bayesian estimation approach takes into account the prior information unlike the classic approach.

In this research, the data consist of the values belonging to four variables: 1) first calving year, 2) lactation length, 3) parity and 4) calving season. The variable first calving year denotes the life year (first year, second year etc.) or simply the age of the animal in which the cow gave birth to its first calf. First calving year is an important variable showing the fertility rate of a cow. Because we need the non-overlapping nodes of these variables as the levels of the network, we classified each variable using appropriate interval in accordance with dairy science.

Furthermore, prior information from a former study about the relations among those variables is available such that, İşçi et al (2015) have shown that milk yield is significantly affected by three variables: first calving year, lactation length and parity. Similarly, Mote et al (2016) stated that calving season is also a significant variable affecting the milk yield. Therefore, in accordance with the nature of the Bayesian estimation, it is appropriate to adopt a hybrid approach to build the network. Hence, at the stage of deciding for an appropriate algorithm, with the help of GeNIe (2019) software, Bayesian Search algorithm provided by Cooper & Herskovits (1991), PC algorithm (named after its authors, Peter and Clark) given by Spirtes et al (1993) and Greedy Thick Thinning algorithm given by Cheng et al (2002) have been employed along with the prior information providing a constraint to be imposed on the initial directions of the edges among the nodes to be forming the network. Also, the entropy scores were calculated by Netica (2019) software. The main differences between these algorithms are: Bayesian Search is a score-based algorithm while PC algorithm is a constraint-based algorithm ordering the conditional independence tests from small to large which makes it an efficient algorithm. Greedy Thick Thinning algorithm is also a score-based algorithm based on the Bayesian Search algorithm. There are some differences among their mechanisms such as Bayesian Search algorithm produces a DAG that gives the maximum score which is proportional to the probability of the structure given by the data while PC algorithm uses the independences detected in data through some independence tests to estimate the structure of the network. Greedy Thick Thinning algorithm, however, starts with an empty graph and repeatedly adds the arcs until no arc addition causes a positive increase.

3. Results and Discussion

At the stage of building the Bayesian network, three different Bayesian networks based on the PC, Greedy Thick Thinning and Bayesian Search algorithms were developed. Afterwards, their estimation performances based on the logscore values given by the Equation 2 were calculated. Table 1 shows the model prediction performances of those algorithms in logscore values. The algorithm with a higher logscore builds a model that has a better estimation performance. In Table 1, it is seen that PC algorithm has the biggest logscore of -676.90. Hence, it can be concluded that the model estimated by the PC algorithm fits the data the best. The second best performing model was estimated by the Greedy Thick Thinning algorithm with a logscore of -678.04 and the third was the model estimated by the Bayesian Search algorithm with a logscore of -679.09. Therefore, the Bayesian network estimated based on the PC algorithm was adopted, visualized and given in Figure 2. The estimated Bayesian network displays the conditional relationships and the initial probabilities belonging to the states of the nodes first calving year, lactation length, parity, calving season and milk yield. Propagating any evidence into the network, the posterior probabilities of all the nodes can be calculated. Sensitivity analysis measures how much a node in a Bayesian network is affected by the changes in the levels of other nodes. That measurement can be made by the entropy reduction scores given by Pearl (1991).

Table 1- Candidate algorithms and their logscore values

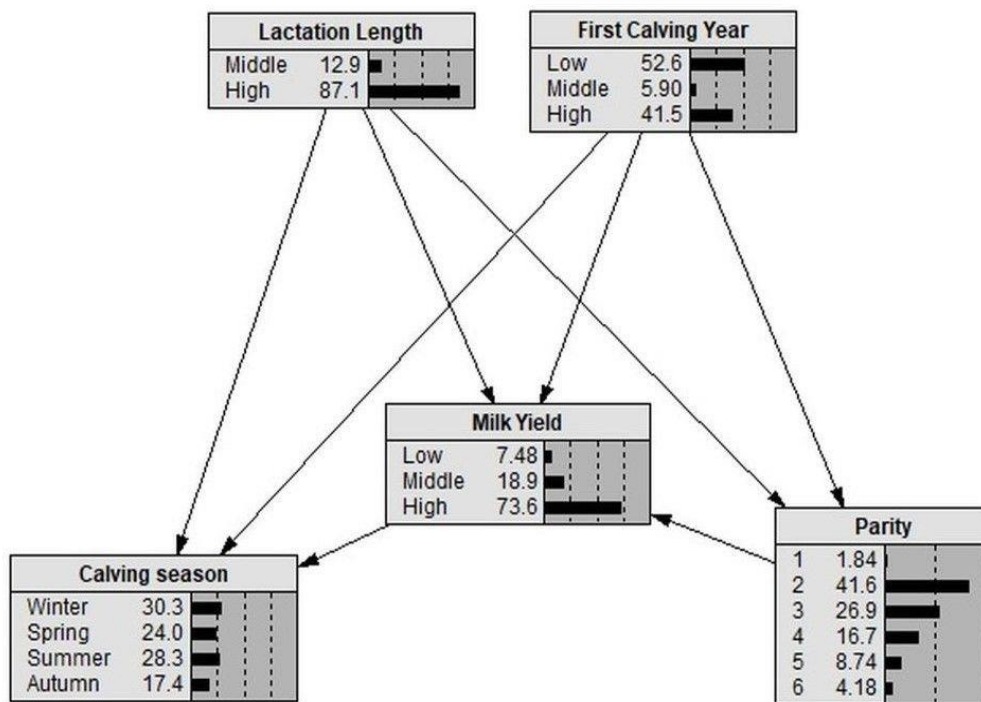| Algorithm	Logscore
PC	-676.90
Greedy thick thinning	-678.04
Bayesian Search	-679.09

Figure 2- Bayesian network model for the factors affecting the milk productivity of Holstein Friesian cows

$$I = H(Q) - E(Q\backslash F) = SUM \sim qSUM \sim fP(q,f)log(P(q,f)/[P(q)P(f)]) \qquad (3)$$

In Equation 3, Q is the query variable, F is the varying variable, q is a state of the query variable and f is a state of the varying variable. $H(Q)$ is the entropy of Q before any new findings, $E(Q\backslash F)$ means the expected real value of Q after new finding f for node F. $SUM \sim q$ means the sum over all states q of Q. Hence, to see how much milk yield is affected by the changes in the levels of other nodes i.e. lactation length, calving season, parity and first calving year, the calculated entropy scores for each node are given in Table 2. As seen from the entropy scores provided in Table 2, milk yield is mostly affected by the lactation length (0.12126). Afterwards, the second most effective variable is calving season (0.08136), the third is the parity (0.06567) and the last is the first calving year (0.00091). After estimating the Bayesian network for the factors affecting the milk productivity of Holstein Friesian cows, it is possible to perform a simultaneous analysis among those factors using this network. That is, it is possible to observe how the levels of the other variables will change simultaneously, when an evidence for any node is propagated through the network. Below are some remarkable results of the analysis.

Table 2- Entropy scores of the variables affecting the milk yield

Variable	Entropy score
Lactation length	0.12126
Calving season	0.08136
Parity	0.06567
First calving year	0.00091

- The highest amount of milk yield is obtained from the cows having low first calving years. The second highest amount of milk is produced by the cows having high first calving years. At middle first calving ages, cows are more likely to produce middle amount of milk. Hence, heifers should be bred at young ages. This result is in accordance with the results given by Kaya et al (1998) who found that breeding the heifers at young ages will increase the milk yield as well as lowering the cost of raising. Eastham et al (2018) also stated that lower first calving ages leads to more amount of milk yield. This result is also supported by Şekerden & Özkütük (2000)

pointing out that for the productivity of a cow per day to be maximum, its first breeding age must be low. First breeding can be done when the cows reach 67% of their adult age (Tümer 2001). Holstein Friesian heifers raised in Turkey can be bred between 14th and 15th months with high level care and feeding (Yüksel et al 2000).

- Cows reach their highest amount of milk yield on their 4th parities. This result is in accordance with the result given by Verma et al (2017). Cows on the first and the sixth parity tend to produce less amount of milk than other periods.

- Cows having middle lactation length tend to produce the middle amount of milk, while the ones having high lactation length produce a high amount of milk. Cows calving in spring and summer seasons tend to have the middle lactation lengths. However, the ones calving in winter and autumn seasons mostly have high lactation length. Therefore, this result complies with the other finding showing that the most amount of milk is obtained in winter and autumn seasons compared to the other seasons. These results are also supported by the findings given by Mote et al (2016) who observed that the average lactation milk yield and lactation length is more in winter season. However, Eetvelde et al (2017) stated that season of calving did not actually influence the milk yield and the seasonal differences were due to the psychological features caused by the seasonal metabolic adaptations of the animals.

- Cows having high lactation lengths calf in winter at the highest rate while the ones having middle lactation lengths mostly calf in spring. This result is also in accordance with the previous comment that points out the highest milk yield is obtained in winter with the highest rate.

- Cows calving for the first time in low ages seem to calf in winter with the highest rate. However, the ones calving in middle and high ages mostly seem to calf in summer.

- Cows having low first calving year and high first calving year tend to have more high level of milk yield than the ones having middle first calving year.

- According to the sensitivity analysis, milk yield is mostly affected by lactation length, secondly it is affected by calving season and thirdly by parity and finally by the first calving year. This result slightly differs from the results that are given by İşçi et al (2015) who have found that the effect of the lactation length on milk yield is of second importance. This difference can be explained by the difference in the herd management factors as location, climate and diets of the animals under consideration. Furthermore, İşçi et al (2015) remarks that in a selection study to increase the milk yield, first calving year and lactation length should be considered as selection criteria. However, we suggest that calving season and parity should also be included in those criteria.

4. Conclusions

In conclusion, using a Bayesian network has some advantages compared to the other analysis methods. Unlike the other methods, previous findings of the variables and their known relations can be added into the Bayesian model under the name "prior information". In this study, the estimated Bayesian network was built on two different prior information. The first one is that the milk yield was significantly affected by three variables: parity, first calving year and lactation length and secondly calving season is another factor affecting milk yield significantly. The prior information was inserted into the network by directing arrows manually from those four nodes into the milk yield node. Afterwards, when the real structure of the Bayesian network was obtained based on the current data set, the prior relations among the variables were updated. This feature of the Bayesian networks allows current data to update the findings of the old information. After this updating process, new findings can also be used as prior information for another study. Hence, Bayesian networks have the property of recycling the information. Therefore, instead of ignoring or excluding the previous information, Bayesian networks include them in the analysis. Another advantage of the Bayesian networks are they do not need to assume any probabilistic distribution. They can be applied to continuous, discrete or categorical data. Moreover, as well as observing the effects of the variables on the target variable (the milk yield in this study), one can also observe how variables affect each other as well. In this research, the data consist of the structure of the Bayesian network observing the relations among the variables is easy and can be made simultaneously. When the corresponding levels of the desired nodes are selected on the network (propagation process), the response in the other nodes conditional to the selected values

can immediately be observed. Basically, this can be though as a very fast conditional probability calculation based on complex conditional probabilistic relations represented by the Bayesian network estimated.

The theoretical and practical advantages of Bayesian networks given above can be successfully adapted in milk productivity or any other farming area. In fact, Bayesian networks are quite in accordance with the natural mechanism of milk production. There are not just direct relations between the milk yield and the factors affecting it. Instead, there are also complex relations among the variables affecting the milk yield. Hence, in farm management, to increase the milk productivity, while changing a factor that is known to be effective on milk yield, one must also consider its indirect effects due to the changes in other factors caused by the factor under consideration. For example, cows calving in spring and summer seasons mostly seem to have middle lactation length, and as the Bayesian network indicates middle lactation length is a factor causing middle amount of milk. Hence, it is obviously seen that calving season has both direct and indirect effects on the milk yield.

Examining the factors affecting milk productivity and the relations among them will increase the milk productivity as well as the success of future improvement studies. The study showed that in a selection study choosing the cows having high lactation lengths and parity for breeding or adjusting the calving season and first calving year time will be useful. These improvements will increase the genetic quality of the cows and the milk productivity.

References

Aktürk D, Bayramoğlu Z, Savran F & Tatlıdil F F (2010). The factors affecting milk production and milk production cost: Çanakkale case-Biga. *Kafkas Üniversitesi Veterinerlik Fakültesi Dergisi* 16(2): 329-335

Cheng J, Greiner R, Kelly J, Bell D A & Liu W (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* 137: 43-90

Chickering D M (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* 2: 445-498

Cooper G & Herskovits E (1991). A Bayesian method for constructing Bayesian belief networks from databases. In: *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, 13-15 July, Los Angeles, California, USA, pp. 86-94

de Campos L M & Huete J F (2000). A new approach for learning belief networks using independence criteria. *International Journal of Approximate Reasoning* 24: 11-37

Eastham N T, Coates A, Cripps P, Richardson H, Smith R & Oikonomou G (2018). Associations between age at first calving and subsequent lactation performance in UK Holstein and Holstein-Friesian dairy cows. *PLoS One* 13(6): 1-13

Eetvelde M V, Kamal M M, Vandaele L & Opsomer G (2017). Season of birth is associated with first-lactation milk yield in Holstein Friesian cattle. *Animal* 11(12): 2252-2259

GeNIe (2019). BayesFusion, LLC. Retrieved in January, 15, 2019 from http://www.bayesfusion.com/

Heckerman D (1998). A tutorial on learning with Bayesian networks. In: M I Jordan (Eds), *Learning in Graphical Models*, NATO ASI Series (Series D: Behavioral and Social Sciences), Springer, Dordrecht, pp. 301-354

Hugin (2019). Hugin Expert. Retrieved in January, 15, 2019 from https://www.hugin.com/

Iqbal A, Fukuda O, Hiroshi O, Endo K, Arai K & Yamashita K (2016). Japanese dairy cattle productivity analysis using Bayesian network model (BNM). *International Journal of Advanced Computer Science and Applications* 7(11): 31-37

İnci S, Kaygısız A, Efe E & Baş S (2007). Milk yield and reproductive traits in Brown Swiss cattle raised at Altınova State Farm. *Tarım Bilimleri Dergisi* 13(3): 203-212

İşçi Ö, Takma Ç & Akbaş Y (2015). Siyah alaca sığırlarda 305 günlük süt verimini etkileyen faktörlerin path (iz) analizi ile belirlenmesi. *Kafkas Üniversitesi Veterinerlik Fakültesi Dergisi* 21(2): 219-224

Jensen F V (2001). Bayesian Networks and Decision Graphs**.** Springer-Verlag New York Inc., Secaucus, New Jersey

Kaya A, Yaylak E & Önenç A (1998). Süt sığırcılığında düzenli üreme ve önemi. *Hayvansal Üretim Dergisi* 38: 8-17

Marcot B G (2012). Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling* 230: 50-62

Mote S S, Chauhan D S & Ghost N (2016). Effect of environment factors on milk production and lactation length under different seasons in crossbred cattle. *Indian Journal of Animal Research* 50(2): 175-180

Natori K, Uto M, Nishiyama Y, Kawano S & Ueno M (2015). Constraint-based learning Bayesian networks using Bayes factor. In: J Suzuki & M Ueno (Eds), *Advanced Methodologies for Bayesian Networks*, Springer, Switzerland, pp. 15-31

Netica (2019). Norsys Software Corp. Retrieved in January, 15, 2019 from https://www.norsys.com/netica.html

Pearl J (1991). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, California, pp. 321

Perochon L, Coulon J B & Lescourret F (1996). Modelling lactation curves of dairy cows with emphasis on individual variability. *Journal of Animal Science* 63: 189-200

Spirtes P, Glymour C & Scheines R (1993). Causation, Prediction and Search. Lecture Notes in Statistics 81, Springer-Verlag, New York

Şekerden Ö & Özkütük K (2000). Büyükbaş Hayvan Yetiştirme. Ç. Ü. Ziraat Fakültesi Ders Kitabı, No: C-122 s, Adana

Tahtalı Y, Şahin A, Ulutaş Z, Şirin E & Abacı S H (2011). Esmer ırkı sığırlarda süt verimi üzerine etkili faktörlerin path analizi ile belirlenmesi. *Kafkas Üniversitesi Veterinerlik Fakültesi Dergisi* 17(5): 859-864

Tümer S (2001). Süt Sığırı Yetiştirmede Döl Verimi ve Sürü Yönetimi. Kurs Notları, İzmir

Verma T S & Pearl J (1991). Equivalence and synthesis of causal models. *Uncertain Artificial Intelligence* 6: 255-268

Verma M K, Sachdeva G K, Yadav A K, Gautam S, Ali M M & Kumar S (2017). Effect of genetic and non-genetic factors on milk yield and milk constituents in Murrah buffalo. *Indian Journal Animal Research* 51(2): 387-390

Yüksel A N, Soysal İ, Kocaman İ & Soysal S İ (2000). Süt Sığırcılığı Temel Kitabı. Hasad Yayıncılık, İstanbul