The Journal of Animal & Plant Sciences, 25(6): 2015, Page: 1520-1526
ISSN: 1018-7081

COMPARISON OF MULTIVARIATE LOGISTIC REGRESSION AND CLASSIFICATION TREE TO ASSESS FACTORS INFLUENCING PREVALENCE OF ABORTION IN NIGERIAN CATTLE BREEDS

A. Yakubu[1], A. D. Awuje and J.N. Omeje[2]

[1]Department of Animal Science, Faculty of Agriculture, Nasarawa State University, Keffi, Shabu-Lafia Campus, P.M.B. 135, Lafia, Nasarawa State, Nigeria.
[2]Department of Veterinary Medicine and Surgery, University of Abuja, Nigeria
Corresponding author's email: abdulmojyak@gmail.com

ABSTRACT

The aim of the study was to investigate the application of binary logistic regression and classification tree to assess the potential factors associated with the prevalence of abortion in two indigenous cattle breeds in Nigeria. Three thousand, six hundred and four calving records of cows from a total of 115 cattle herders within Nasarawa, Benue and Plateau, north central Nigeria from the year 2010-2013 were utilized in the study. The cattle which were of Bunaji and Sokoto Gudali genetic groups originated from different flocks and were reared under the traditional extensive system. The risk factors investigated were dam's breed-group, season and parity. The incidence of abortion was higher in Bunaji (16.9%) compared to Sokoto Gudali (13.8%) cattle. Abortion was also more in the rainy (19.6%) than the dry (13.0%) and harmattan (lower temperatures with winds)(13.5%) seasons respectively. Animals which have calved once and twice aborted (23.9% and 21.1%, respectively) more than their counterparts which have recorded up to three (9.8%) and four births (5.8%). However, the multivariate logistic model (χ^2 = 8.56, P= 0.197) and classification tree (risk value = 0.123±0.004) revealed that parity number and breed were the most important parameters affecting the prevalence of abortion in cattle. The present information may be exploited in management practices to attenuate the incidence of abortion, thereby increasing the productivity of the animals in Nigeria, sub-saharan Africa.

Key words: Abortion, cattle, logistic regression, classification tree, Nigeria.

INTRODUCTION

In Nigeria, cattle production plays an important role in the economic improvement of poor farmers and contributing to poverty alleviation (Yakubu *et al*., 2010). However, one of the major constraints to a successful development of cattle industry is the menace of abortion. Abortion implies expulsion of a fetus before full term and viability outside of the uterus. Antepartum death is characterized by variable degrees of autolysis, accumulations of blood-tinged fluids in body cavities, soft autolytic kidneys, and variable degrees of liquefaction of the brain (Holler, 2012). These early losses are associated with a wide range of physiologic, nutritional, environmental, and noninfectious causes that often go unrecognized. Abortion in livestock herds at a level that significantly affects productivity is a common clinical problem (Menzies, 2011). Despite ongoing research investigations to diagnose and evaluate the epidemiology of infectious factors that cause abortion, breeders are still being confronted with abortion continuously (Rafati *et al*., 2010). Reproductive failure due to abortion disease remains a significant revenue drain in many ruminant livestock production systems (Simsek *et al*., 2012).

Data mining methods are special statistical instruments which are applied to detect relationships between attributes in datasets (Gerjets *et al*., 2011). However, extracting meaningful information from the avalanche of biological data is a formidable task. Data mining approaches based on decision tree methods have been successfully exploited in elucidating importance of features affecting important biological processes such as growth and reproduction. Classification and regression trees, which were first introduced by Breiman *et al.* (1984), are effective and simple means for sifting complex biological data for hidden explicit patterns and information. More and more biological studies are harnessing classification and regression tree methodologies owing to its simplicity and ability to handle missing values (Banerjee *et al*., 2008; Yakubu, 2012; Eyduran *et al*., 2013). The classification tree technique has been used in the analysis of discrete sheep reproduction traits (Dariusz, 2009). It has also been used to assess herd specific risk factors for coliform mastitis in sows (Gerjets *et al*., 2011). It is a tree-structured solution in which a constant or a relatively simple regression model is fitted to the data in each partition (Loh, 2008). The structure of the classification tree starts with the entire information set (root node). The subsets which emerge as a result of division are referred to as child nodes. The final subsets which are not exposed to further divisions are called leaves. CHAID (Chi-square Automatic Interaction Detector) is certainly nowadays the most popular among earlier statistical supervised tree growing techniques (Ritschard, 2010). The main characteristics of CHAID that contributed to its popularity are: At each node, CHAID determines for each potential predictor the optimal n-ary split it would produce, and selects the predictor on the basis of these optimal splits. CHAID uses p-values with a Bonferroni correction as splitting criterion.

Logistic regression allows the prediction of group membership from a set of categorical and/or continuous variables (x). Generally, the dependent variable is dichotomous and can take the value 1 (member of the group) with a probability of success y, or the value 0 (non-member) with probability of failure 1 – y. The relationship between the dependent and independent variables is not a linear function. Instead, the logistic regression function is used, which is the logit transformation of y (Dossa *et al*., 2008). It has been applied to the analysis of abortion and stillbirth in goats (Yakubu *et al*., 2014).

The present study therefore aimed at using multivariate logistic regression and classification tree to assess some factors [breed (Bunaji and Sokoto Gudali), season (rainy, dry and harmattan) and parity number (1, 2, 3 and 4)] that influence prevalence of abortion in Nigerian cattle breeds.

MATERIALS AND METHODS

Study area and animal sampling: Calving records (n=3, 604) of cows from a total of 115 cattle herders within Nasarawa, Benue and Plateau, north central Nigeria from the year 2010-2013 were utilized in the study. The cattle which were of Bunaji and Sokoto Gudali genetic groups originated from different flocks and were reared under the traditional extensive system. Sampling was restricted to only farmers that were able to give information on calves, bull and cow identification as well as occurrence of abortion, calving date or period and parity. Three seasons of abortion were generated according to the month of the year, Rainy season: May to October; Dry season: February to April; and Harmattan season: November to January. The rainy season is characterized by high temperatures, rain and abundant pasture. The dry season is characterized by high temperatures, lack of rain and scarce pasture. The harmattan season has lower temperatures with winds. No etiological diagnoses were made in aborted fetuses.

Statistical analysis: The logit of the probability of an abortion was modelled using logistic regression assuming an asymptotic binomial distribution. First, the univariate analysis for all hypothesized risk factors (dam breed group, season, parity and number of fetuses) and the occurrence of abortion in the present study was carried out using Pearson’s Chi-square (χ^2) test. Subsequently, a multivariate model was built by including every hypothesized risk factor which had p-value of $P<0.200$ from the univariate analysis, following the description of Santos *et al*. (2012) and Ryan *et al*. (2012). Variables were retained, if p-value from the logistic regression was $P<0.05$, otherwise they were removed from the final model. Backward stepwise elimination based on Wald method was applied (Noordhuizen *et al*., 2001). Model validity and reliability was assessed using the Hosmer-Lemeshow goodness of-fit test (Hosmer and Lemeshow, 2000). The multivariate model employed (Czopowicz *et al*., 2012) was:

$$P(Y=1) = \frac{1}{1 + \exp[-(B_0 + B_1 \times X_1 + ... + B_n \times X_n)]}$$

where,

P(Y=1) = probability of a final outcome (abortion)

B_0 = intercept

B_1, B_n = regression coefficients for individual risk factors

X_1, X_n = risk factors (dam breed group, season, parity and number of fetuses)

As the response variable in the present study (prevalence of abortion) is a categorical variable, chi-square splitting criterion was used in CHAID supervised tree growing technique. The criterion is in that case the p-value of the F statistic for the difference in mean values between the g nodes generated by the split (Ritschard, 2010):

$$F = \frac{BSS/(g-1)}{WSS/(n-g)} \sim F_{(g-1),(n-g)}$$

where,

F = F statistic

WSS = within sum of squares

BSS = between sum of squares = TSS (total sum of squares) minus WSS

n = optimal splits of each predictor at each node

g = final number of optimally merged categories.

The tree building process continued until it became impossible. The maximum tree value was obtained after the tree reached a maximum dimension.10-fold cross-validation was used as an error estimation method; this was to provide estimates of the future prediction error for each sub-tree (Camdeviren *et al*., 2005). The statistical package employed in the analysis was SPSS (2010).

RESULTS

The incidence of abortion was higher in Bunaji (16.9%) compared to Sokoto Gudali (13.8%). Abortion was also more in the rainy (19.6%) than the dry and harmatan season (Table 1). Animals that have calved once and twice aborted more than their counterparts which have recorded up to three and four births. However, breed, season and parity number were all found to be associated with abortion according to the univariate chi-square analysis. These three factors were thus fitted in the subsequent multivariate logistic regression analysis.

In the multivariate logistic regression analysis, parity number and breed were found to be the variables of utmost importance to predict the incidence of abortion (Table 2). Cows that have calved only once (P=0.000; Odds ratio (OR) = 1.681) appeared to be more susceptible to abortion compared to those that carved in later years (P=0.058-0.009; Odds ratio (OR) = 0.746-1.216). The logistic regression model fitted well as revealed by the Hosmer and Lemeshow test (χ^2 = 8.56, P= 0.197).

The classification tree also revealed the importance of parity number and breed in predicting the occurrence of abortion in cattle (Figure 1). Node 0, which is called the root node, contains descriptive statistics related to the prevalence of abortion. Animals in this node were on the basis of parity divided into three nodes: Nodes 1, 2 and 3. However, there was a subdivision of node 1 into Nodes 4 and 5, respectively. Node 3 was equally divided into two homogenous (terminal) subgroups: Nodes 6 and 7, respectively. Node 4 which depicts Bunaji cattle was a homogenous Node with higher abortion prediction rate (0.265) compared to Nodes 5 (heterogeneous, 0.184), 6 (0.037) and 7 (0.080), respectively. The risk value of the CHAID analysis was 0.123±0.004.

Table 1. The association between risk factors and the prevalence of abortion in Nigerian cattle

Parameters	**Number**	**No.of abortion (%)**	**Chi-square (P-value)***
Breeds of cattle			
Bunaji	1802	304 (16.9)	6.71 (0.01)
SokotoGudali	1802	248 (13.8)	
Season			
Rainy	1168	229 (19.6)	24.76 (0.00)
Dry	1183	154 (13.0)	
Harmattan	1252	169 (13.5)	
Parity number			
1	876	209 (23.9)	155.80 (0.00)
2	976	206 (21.1)	
3	876	86 (9.8)	
4	876	51 (5.8)	

Table 2. Logistic regression predicting the prevalence of abortion in Nigerian cattle.

Risk factor	B	S.E.	Wald's χ^2	P-value	Odds ratio	CI (95%)
Breed (1)	–0.212	0.074	8.254	0.008	0.809	0.700-0.935
Parity	-	-	6.539	0.000	-	-
Parity (1)	0.520	0.103	25.33	0.000	1.681	1.373-2.059
Parity (2)	0.195	0.103	3.591	0.058	1.216	0.993-1.488
Parity (3)	–0.294	0.112	6.854	0.009	0.746	0.599-0.929
Constant	–0.894	0.084	113.451	0.000	0.409	-

B= regression coefficient, S.E.= standard error of B, CI= confidence interval
Hosmer and Lemeshow test: χ^2 = 8.56, P= 0.197.

Table 3. The risk estimate associated with the prediction of the prevalence of abortion in Nigerian cattle using CHAID

Method	Estimate	Standard Error (SE)
Resubstitution	0.123	0.004
Cross-Validation	0.123	0.004

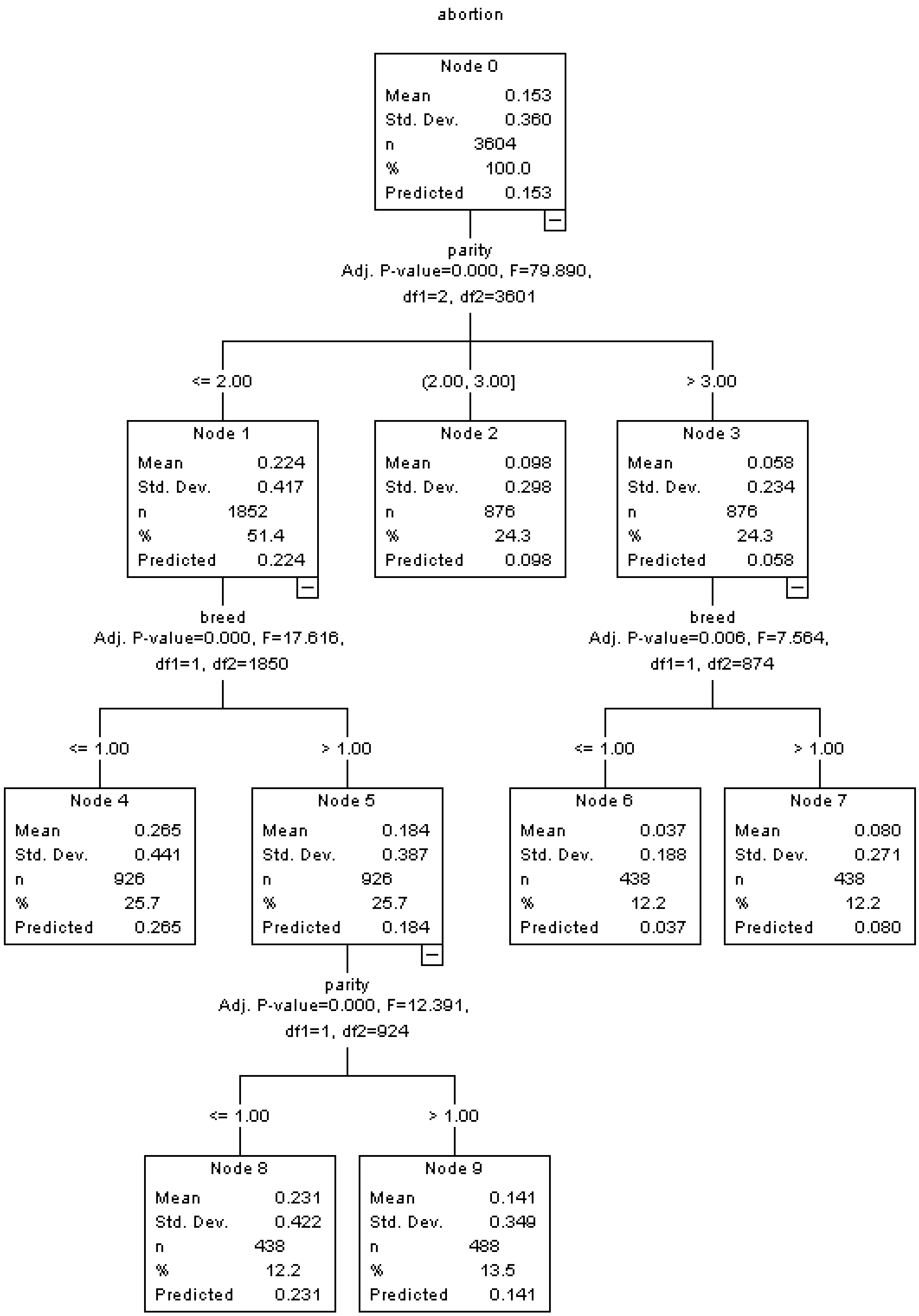

Figure 1. Classification tree showing the prediction of abortion in cattle.

DISCUSSION

Brucellosis is one of the world's major zoonotic diseases associated with reproductive disorders and a potential infection of human. Brucellosis leads to serious economic losses due to late-term abortion, stillbirth, weak calves, and sterility (Segura-Correa and Segura-Correa, 2009; Sylla *et al*., 2014). Abortions can occur as outbreaks, but more often, they are sporadic. In a similar study, Sylla *et al*. (2014) reported that the prevalence of brucellosis in Macenta and Yomou provinces in Guinea was 12 and 5.33 %, respectively. In both provinces, the prevalence mean was 8.67 %. In Zambian cattle however, Muma *et al*. (2009) reported abortion prevalence rate of 16.22%. An abortion rate between 2% and 5% suggests that endemic disease may be present *(*Menzies, 2011). Thus, high abortion rate in ruminants may also be a signal of the introduction of other zoonotic viral diseases that are currently spreading in numerous African countries (Bronner *et al*., 2013). The present study reveals that the incidence of abortion is an important reproductive problem in cattle breeding in Nigeria. Therefore, the evaluation of contributory factors is justified.

Parity number and breed were significant factors in the logistic regression model affecting the incidence of abortion. In a related study in dairy cattle, some risk factors of prenatal mortality identified were calving number, age at first calving, gestation length, sire of calf and inbreeding (Benjaminsson, 2007). However, Haileselassie *et al*. (2011) reported that parity number had no significant effect on the incidence of abortion.The classification tree equally revealed that parity number and breed were the most important factors that affect the incidence of abortion. This may be due to similarity in the differential power of each model. The implication of the information revealed in Nodes 1 and 4 is that Bunaji cows that have calved only once and twice are more likely to suffer from abortion. The risk value associated with the prediction of abortion in the present study appeared low which is an indication of better accuracy and reliability of prediction. According to Gerjets *et al*. (2011), using more observations for model building improves the evaluation accuracy; while lower prevalence and therewith more skewed data, made it easier to reach higher accuracies. For practical use, graphical trees should be smaller with clearly arranged decision steps to simplify interpretations for farmers and consultants (Yakubu, 2012). CHAID is a non-parametric method; therefore, no assumptions such as normality, constant variance, linearity, non-multicollinearity are required about the underlying distribution of independent variables (Stark and Pfeiffer, 1999). The tree diagram shows the probability of the occurrence of the events and vividly illustrates the structure of the risk factors and their complex interactions thereby making the findings easier to interpret, even to those with less statistical background (Harper, 2005). Regression tree methodology is increasingly being used in biological studies such as diagnosis decision processes and epidemiology (Lemon *et al*., 2003).

The use of new methodologies in veterinary clinical epidemiology is of prime importance as indicated by Saegerman *et al*. (2013) who used multivariate logistic regression analysis and classification and regression tree analysis to differentiate herds at risk of Q fever exposure (in ruminants, well-known manifestations of Q fever are abortion, stillbirth, delivery of weak offspring and premature delivery). The decision tree method differs from the logistic regression in that it examines the essential effect of a variable on the explanatory variable in correlation with another variable. Although this can be done by logistic regression as well, significant effects have to be revealed and included in the model; while for decision tree, the method extrapolates correlations (Kovacs, 2009).

Conclusion: In the present study, the incidence of abortion was higher in Bunaji than the Sokoto Gudali cattle. Abortion was also more in the rainy than the dry and harmatan season. Incidence of abortion was higher in animals that have calved once and twice than those have recorded up to three and four births. However, the multivariate logistic models and classification tree showed that parity and breed were the most important parameters affecting the prevalence of abortion in the two Nigerian cattle breeds. Prospective field studies of abortion are very expensive and not routinely applicable in the field. Therefore, the present study on Nigerian cattle has significant implications for farmers and veterinary practitioners/herd health consultants as informed risk analysis is the key to successful decision making in relation to reproductive problem control on farms. The present information may be exploited in management practices to reduce abortion rate in cattle in Nigeria thereby ensuring increased production level.

REFERENCES

Banerjee, A. K., N. Arora and U.S.N. Murty (2008). Classification and regression tree (CART) analysis for deriving variable importance of parameters influencing average flexibility of CaMK Kinase Family. Elect. J. Biol. 4: 27-33.

Benjamínsson, B.H. (2007). Prenatal death in Icelandic cattle. Acta Vet. Scand. 49 (Suppl 1): S16 doi:10.1186/1751-0147-49-S1-S16.

Bronner, A., V. Hénaux, T. Vergne, J.L. Vinard, E. Morignat, P. Hendrikx, D. Calavas and D. Gay (2013). Assessing the mandatory bovine abortion notification system in France using Unilist Capture-Recapture Approach. PLoS ONE 8: e63246. doi: 10.1371/ journal. pone. 0063246.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984). Classification and regression trees for bone marrow immunophenotyping. Cytometry 20: 210-217.

Camdeviren, H., M. Mehmet, M.M. Ozkan, F.T. Toros, T. Sasmaz T. and S. Oner (2005). Determination of depression risk factors in children and adolescents by regression tree methodology. Acta Med. Okayama 59: 19-26.

Czopowicz, M., J. Kaba, O. Szaluś-Jordanow, M. Nowicki, L. Witkowski and T. Frymus (2012). Multivariate model for the assessment of risk of fetal loss in goat herds. Pol. J. Vet. Sci. 15: 67-75.

Dariusz, P. (2009). Using classification trees in statistical analysis of discrete sheep reproduction traits. J. Cent. Eur. Agric. 10: 303-310.

Dossa, L. H., B. Rischkowsky, R. Birner and C. Wollny (2008). Socio-economic determinants of keeping goats and sheep by rural people in southern Benin. Agric. Hum. Val. 25: 581–592.

Eyduran, E., I. Yilmaz, M. M. Tariq and A. Kaygisiz (2013). Estimation of 305-d milk yield using regression tree method in Brown Swiss cattle. J. Anim. Plant Sci. 23: 731-735.

Gerjets, I., I. Traulsen, K. Reiners and N. Kemper (2011). Application of decision-tree technique to assess herd specific risk factors for coliform mastitis in sows. Vet. Devel.1: 27-31.

Haileselassie, M., Kalayou, S., Kyule, M., Asfaha and K. Belihu (2011). Effect of Brucella infection on reproduction conditions of female breeding cattle and its public health significance in Western Tigray, Northern Ethiopia. Vet. Med. Intern. 2011: doi:10.4061/2011/354943.

Harper, P.R. (2005).A review and comparison of classification algorithms for medical decision making. Health Pol.71: 315–331.

Holler, L.D. (2012). Ruminant abortion diagnostics. Vet. Clin. Food Anim. 28: 407–418.

Hosmer, D. W. and S. Lemeshow (2000). Applied Logistic Regression.2nd ed. John Wiley and Sons, New York, USA.p 375.

Kovacs, S. (2009). Methods for analysing technological risks in animal breeding.University Doctoral Dissertation Abstract.University of Debrecen, KarolyIhrig Doctoral School of Management and Business Administration. Pp. 27.

Lemon, S. C., J. Roy, M.A. Clark, P. D. Friedmann and W. Rakowski (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. Ann. Behav. Med. 26: 172-181.

Loh, W.Y. (2008). Classification and regression tree methods (In Encyclopedia of Statistics in Quality and Reliability, Ruggeri, Kenett and Faltin (eds.), 315–323, Wiley, 2008).

Menzies, P.I. (2011). Control of important causes of infectious abortion in sheep and goats. Vet. Clin. Food Anim. 27: 81–93.

Muma, J.B., J. Godfroid, K.L. Samui and E. Skjerve (2009). The role of Brucella infection in abortions among traditional cattle reared in proximity to wildlife on the Kafue flats of Zambia. Rev. Sci. Tech. Off. int. Epiz. 26: 721-730.

Noordhuizen, J.P., M.V. Thrusfield, K. Frankena and E.A.M. Graat (2001). Application of quantitative methods in veterinary epidemiology. 2nd ed., Wageningen Pers, Wageningen, Holland.

Rafati, N., H. Mehrabani-Yeganehaand T.E. Hanson (2010). Risk factors for abortion in dairy cows from commercial Holstein dairy herds in the Tehran region. Prev. Vet. Med. 96: 170–178.

Ritschard, G. (2010). CHAID and earlier supervised tree methods. Département d’économétrie Université de Genève, 40 Boulevard du Pont d’Arve, CH - 1211 Genève 4http://www.unige.ch/ses/metri/.

Ryan, E.G., N. Leonard, L. O’grady, M.L. Doherty and S. J.More (2012). Herd-level risk factors associated with Leptospira Hardjosero prevalence in Beef/Suckler herds in the Republic of Ireland. Vet. J. 65: (http://www.irishvetjournal.org/content/65/1/6).

Santos, C.S.A.B., R.M. Piatti, S.S. Azevedo, C.J. Alves, S.S.S. Higino, M.L.C.R. Silva, A.W. Brasil and S.M. Gennari (2012). Seroprevalence and risk factors associated with Chlamydophilaabortus infection in dairy goats in the Northeast of Brazil. Pesq. Vet. Bras. 32: 1082-1086.

Saegerman, C., N. Speybroeck, F. Dal Pozzoand G. Czaplicki (2013). Clinical Indicators of Exposure to Coxiellaburnetii in Dairy Herds. Transbound. Emerg. Dis. doi: 10.1111/tbed.12070.

Segura-Correa J.C. and V.M. Segura-Correa (2009). Prevalence of abortion and stillbirth in a beef cattle system in Southeastern Mexico. Trop. Anim. Health Prod. 41: 1773–1778.

Simsek, S., A. Risvanli, A.G. Zonturlu, O. Demiral and N. Saat (2012). Absence of link between abortion and seropositivity of cystic hydatid disease in ewes and female goats in Turkey. Vet. Ital. 48: 323-327.

SPSS (2010). Statistical Package for Social Sciences.SPSS Inc., 444 Michigan Avenue, Chicago, IL60611, 2010.

Stark, K.D.C. and D.U. Pfeiffer (1999). The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology – An example. Intel. Data Anal. 3: 23-35.

Sylla, S., Y. Sidimé, Y. Sun, S. Doumbouya and Y. Cong (2014). Seroprevalence investigation of bovine brucellosis in Macenta and Yomou, Guinea.

Trop. Anim. Health Prod. (in press). DOI:10.1007/s11250-014-0625-2.

Yakubu, A., M. M. Muhammed and I.S. Musa-Azara (2014). Application of multivariate logistic regression models to assess parameters of importance influencing prevalence of abortion and stillbirth in Nigerian goat breeds. Biotech. Anim. Husb. 30: 79-88.

Yakubu, A. (2012). Application of regression tree methodology in predicting the body weight of Uda sheep. Anim. Sci. and Biotech. 45: 484-490.

Yakubu, A., K.O. Idahor, H.S. Haruna, M. Wheto and S. Amusan (2010). Multivariate analysis of phenotypic differentiation in Bunaji and Sokoto Gudali cattle. Acta Agric. Slov. 96: 75-80.