Predicting Male Fertility From Semen Quality Traits: Performance Of Different Statistical Procedures For Model Selection

M. Piles[*], J.P. Sánchez[†], M., García-Tomás[*], O. Rafel[*], J. Ramon[*], and Ll. Tusell[*]

Introduction

Male fertility and prolificacy could be improved by indirect selection through its components, such as the parameters defining semen quality. The relationship between the characteristics of the ejaculate and the result of insemination is also very important in order to find immediate, simple and no expensive assays to evaluate the ejaculates and optimize the use of males in AI centers. However, this relationship is still not clearly established since most of the research works have shown that the proportion of the observed variance explained by different models was very low. This could be due to: i) the experimental design regarding AI conditions related to ejaculate selection and dose preparation, ii) the variables used as descriptors of semen quality, and the time -with respect to the AI time- and manner in which they were evaluated, and iii) the procedures used for variable selection and parameter estimation. Most of the published researches assumed a linear relationship between fertility and semen quality traits and give information concerning the goodness of fit of the model and the estimates of the effects of the different semen quality traits, but they don't show its predictability power (Brun et al., 2002, Gadea et al., 2004., García-Tomás et al., 2006) in an independent set of data.

The objective of this research was to apply different statistical methods to a set of fertility data and semen quality traits to determine the model and the procedure that provides the best predictive performance of fertility on new data. We compare the classical multiple regression models with two other algorithms: The modified ant colony algorithm for variable selection proposed by Shen et al. (2005) and the neural network algorithm (Bishop, 1995). The first one has the advantage with respect to sequential forward regression that avoids the problem of nesting since in each step of the optimization process the combination of variables selected not only depends on the selected variables in the previous step but also in the variables selected in the best model along the optimization process. The Neural Network model has the advantage in respect of all the other models that it is a nonlinear function (being the linear model a particular case) from a set of input variables to a set of output variables controlled by a vector of adjustable parameters, defining the weights for each input. This model can be considered more flexible than the others since it includes any kind of relationship between the independent and the dependent variables.

---

[*]Unitat de Cunicultura, Institut de Recerca i Tecnologia Agroalimentàries, Torre Marimon s/n, 08140, Caldes de Montbui, Barcelona, Spain
[†] Departamento de Producción Animal, Universidad de León, 24071, León, Spain

Material and methods

Bucks came from a sire line. Two ejaculates per male and per week were collected with artificial vagina. The following semen characteristics were evaluated: Concentration (**C**) and volume (**V**) of the ejaculate, individual motility (**IM**); percentages of: viable spermatozoa (**Vi**), spermatozoa with normal apical ridge (**NAR**), morphological abnormalities of head (**HAP**), neck-midpiece (**NAP**) and tail (**TAP**) and spermatozoa with presence of proximal (**PD**) and distal (**DD**) cytoplasmic droplet. Dose concentration (**DC**) and dilution factor of the AI dose (**DF**) were also recorded. Artificial insemination was performed under limited conditions (i.e. low sperm dosage and a small pre-selection of the ejaculates) in order to avoid the masking effect that a very high number of spermatozoa may have on compensating deficiencies in semen characteristic (Saacke et al., 2000). Ejaculates were rejected for AI only if individual motility was lower than 2 (in a scale from 0 to 5), or if they had urine or calcium carbonate deposits. Artificial inseminations were performed at two different dose concentration (10 and 40 millions of spermatozoa per mL) over crossbred females 24 h after collection. Diagnosis of pregnancy was made by palpation, 14 d after AI, and the result was confirmed at parity. Fertility (**F**) was defined as % of kindling rate by male. A total of 271 data was obtained from 135 males. Variables were standardized to be able to compare its effect on fertility. The correlation between all pairs of independent variables was lower than 0.65.

In order to achieve the objective of the work data were divided in 3 subsets of equal size. Then, the 3-fold cross-validation was used: 2 groups were used to train the different models (training set, **Tset**) that were then evaluated on the remaining group (validation set, **Vset**), repeating this process for the 3 possible choices for the Vset. The number of data for the three processes was: 189, 167 and 186 for the Tset and 82, 104 and 85 data for the Vset in process 1, 2 and 3, respectively. Four approaches were used for the statistical analysis: 1) Multiple regression including all the independent variables (**MR**); 2) Stepwise Forward Regression (**SFR**); 3) Neural Network algorithm (**NN**) where the model is a nonlinear function from a set of input variables to a set of output variables controlled by a vector w of adjustable parameters. The objective of training is to find the set of weights that minimize the differences between the output and the observed value given a Tset. This was achieved according to the criteria of minimizing the mean square error, by using the back propagation algorithm. After several trials to determine the best settings, the procedure was implemented by using 1 hidden layer, 4 neurons and a learning rate equal to 0.2. 4) The modified ant colony algorithm for variable selection proposed by Shen et al. (2005). In this work we applied the algorithm in two ways: 1) Using as a measure of the goodness of fit during the estimation process (training) the mean square error for the current model in the whole Tset (**AC**). 2) Using the same measure of the goodness of fit but this time calculated in an independent set of data, obtained including the 33% of the data from the Tset ($\mathbf{AC_{cross}}$). The last approach was applied expecting to obtain a greater predictive ability of the model selected in the Vset. After several trials, the Ant Colony algorithm was implemented with

100 ants and a dissipation rate of 0.7. The prediction ability of the models was measured as the correlation between the observed and the predicted value and the mean squared error in the Vset.

Results and discussion

In general, the predictive performance of all the models was similar and the results were consistent in the three data groups. The correlation between the observed and the predicted value in an independent set of data is not high, probably because of the percentage of variation in fertility explained by this group of semen characteristics is very low (Brun., 2002; Gadea et al., 2004; García-Tomás et al., 2006) and it is necessary to find other descriptors of semen quality, or it may be needed to evaluate some of them in a different way or time, i.e. closer to the AI time.

The best performance was obtained with MR, SFR and NN (Table 1), being NN the one with the smaller uniformity of results in the different groups of data maybe because it could need a greater amount of data. The relationship between the evaluated semen characteristics and fertility seems not to be markly nonlinear since the use of nonlinear models (as in NN) did not suppose an advantage over the linear models. Multiple regression including all the semen quality traits was the procedure with the best predictive ability in an independent set of data since it had the greatest correlation between the observed and the predicted values and the lowest mean square error in that data set. Considering the fact that the maximum correlation between explanatory variables is 0.65 that means that almost no penalty would be introduced in the mean squared error as a consequence of collineality, and this method would make a full usage of the predictive ability of all the explanatory variables. The ant colony algorithm was the procedure with the worst predictive performance, especially AC_{cross} which could require a greater amount of data for training given that it uses a part of the Tset to obtain the measure of goodness of fitting during the training procedure. However the results were consistent for the three data grouping.

All the procedures were consistent with the estimates of the effect of semen characteristics on fertility, being those with the greater effect the ones selected by the procedures making a selection of variables (AC and SFR). They were PD and DD, which had a negative significant effect in all the methods, NAR with a significant positive effect in most of the models and TAP with a negative effect.

Table 1. Predictive ability of the different methods. Correlation between the observed and the predicted data (ρ) and mean square error (MSE) in the validation set.

Method	Selected Variables	ρ[1]		MSE[2]	
		mean[1]	cv[1]	mean[1]	cv[1]
MR	All	0.45	0.04	0.72	0.09
SFR	NAR, TAP, PD, DD, V	0.43	0.12	0.73	0.10
NN	All	0.43	0.17	0.75	0.15
AC	All	0.37	0.11	0.80	0.08
[2]**AC_{cross}**	PD(3/3) DD(3/3) C(2/3) DC(2/3) NAR(2/3) DF(2/3) Vi(1/3) HAP(1/3) NAP(1/3) TAP(1/3) V(1/3)	0.35	0.06	0.82	0.08

[1] Mean and coefficient of variation of the values in the three groups of data

[2] in brakets is the number of times that the variable was selected

Conclusion

The predictability of male fertility from the semen characteristics usually included in classical spermiogram is not high enough and therefore it would be necessary to include new parameters or modify the manner and time in which they are evaluated. All the tested procedures for model selection and estimation perform similarly but multiple regression, including all the semen quality traits, seems to have the greater ability for predicting new data and show the most uniform results.

References

Bishop, C.M. 1995. Neural networks for pattern recognition, Oxford University Press

Brun, J.M., Theau-Clement, M., and Bolet, G. 2002. Anim. Reprod. Sci., 70:139-149.

Gadea, J., Sellés, E., and M.A. Marco. 2004. Reprod. Dom. Anim., 39:303-308.

García-Tomás, M., Sánchez, J., Rafel, O., Ramon, J., and Piles, M. 2006. Livest. Sci., 104:233-243.

Saacke, R. G., J. C. Dalton, S. Nadir, R. L. Nebel, and J. H. Bame. 2000. Anim. Reprod. Sci., 60: 663-677.

Shen, Q., Jiang, J.H., Tao, J.C., Shen, G.L., Yu, R.Q. 2005. J. Chem. Inf. Model, 45:1024-1029.