A GP approach for precision farming

A GP Approach for Precision Farming

Francesca Abbona [1,4,*], Leonardo Vanneschi [2,3], Marco Bona [4] and Mario Giacobini [1]

[1] Department of Veterinary Sciences, University of Torino, Turin, Italy.

[2] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal.

[3] LASIGE, Departamento de Informàtica, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

[4] ANABORAPI, Associazione Nazionale Allevatori Bovini Razza Piemontese, Carrù, Italy

* Corresponding author.

E-mail addresses:

francesca.abbona@unito.it (F. Abbona), lvanneschi@novaims.unl.pt (L. Vanneschi), marco.bona@anaborapi.it (M. Bona), mario.giacobini@unito.it (M. Giacobini).

Abstract—Livestock is increasingly treated not just as food containers, but as animals that can be susceptible to stress and diseases, affecting, therefore, the production of offspring and the performance of the farm. The breeder needs a simple and useful tool to make the best decisions for his farm, as well as being able to objectively check whether the choices and investments made have improved or worsened its performance. The amount of data is huge but often dispersive: it is therefore essential to provide the farmer with a clear and comprehensible solution, that represents an additional investment. This research proposes a genetic programming approach to predict the yearly number of weaned calves per cow of a farm, namely the measure of its performance. To investigate the efficiency of genetic programming in such a problem, a dataset composed by observations on representative Piedmontese breedings was used. The results show that the algorithm is appropriate, and can perform an implicit feature selection, highlighting important variables and leading to simple and interpretable models.

Keywords—Genetic Programming, Precision Livestock Farming, Cattle Breeding, Piedmontese Bovines.

1. Introduction

In this article, the performance of the breeding farms of *Piemontese* bovines are investigated. The considered cattle farms are located in Piedmont, a region in Northwestern Italy. The Piedmontese cattle derives its name from this region, its cradle of origin, even if today it is spreading in several foreign countries. The bovines are usually bred in beef intensive farms, which are therefore provided with the installation of stables to control the animals, grazing for fattening purposes, the addition of different artificial fodder on feed and curative intents, and particular attention to the reproduction of the livestock. The main information that represents the yield of a Piedmontese cattle farm is given by the *count of calves per cow per year* [1, 2]. It is a quantity that is basically predicted considering the average calving interval *intp*, expressed in days, and the average perinatal mortality of the farm *m*, referred to the previous 12 months:

$$Y_a = \frac{365}{intp}\left(1 - \frac{m}{100}\right) \qquad (1)$$

However, this expression does not take into account the period following the birth of the calf. Calf mortality is an important cause of economic losses in Piedmontese cattle farms [2]. It represents for the farmer the loss of the economic value of the calf and the reduction of the herd's genetic potential. Furthermore, the high mortality rate reduces the number of young animals to be used to increase the size of the breeding. From the analysis of the recordings in 2017, described later in the manuscript, the difference between the number of dead calves at birth and those that did not survive through the weaning period is straightforward extremely significant. In Figure 1, the distributions of dead calves are represented for the selected group of breedings. Most of the farms report no death at birth (mean value: 0.13), whereas during weaning period records show up to 23 deaths per farm in the considered year (mean value 3.02), entailing that many of the newborns were not able to survive. This issue can lead to the need of buying animals and, therefore, to additional costs. Perinatal mortality is related to birth and the first few hours after it: it is mainly due to the delivery itself (difficulty of parturition of the cow and its health condition) and to the difficulty of birth and the weight of the newborn.

However, the complete development of the calf occurs in the 60 days following birth. During the weaning period, the physiological development process of the animal reaches completion and it is straightforward that the gestational phase alone is not exhaustive: it is therefore crucial to consider neonatal mortality, outlining the calf's ability to survive. Thus, in addition to the genetic factors previously mentioned, it is inevitable to consider congenital calf's defects, such as arthrogryposis or macroglossia. Together with environmental and food conditions, they affect the quality of life of the

newborn, denoting an important source of stress that can compromise the immune response, the growth rate, the disease resistance, and the well-being of the animals. It is therefore necessary to incorporate in expression (1) the factors that encapsulate the effect of the weaning period of the newborn. This issue leads to the reformulation of the problem into the following question:

"How many weaned calves per cow are produced per year?"

It is straightforward that a proper model should be formulated, with the encapsulation of other parameters, among those available in the dataset.

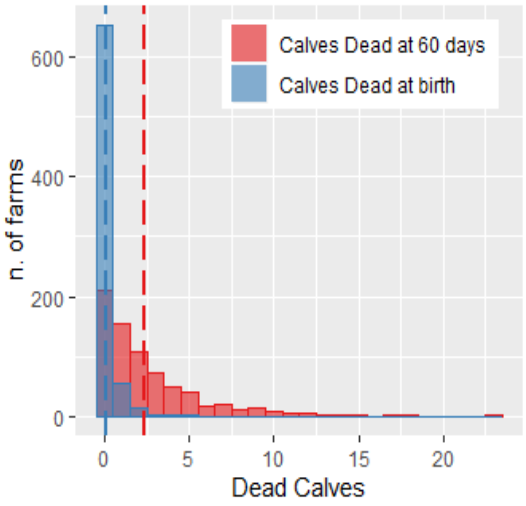

Fig.1 Distribution of the number of dead calves at birth and during the weaning period in 2017. Mean values are represented with the dashed line at the two different time reference. The data derive from the dataset described in Section III. All the breedings (725) show extremely different values between the dead calves at birth (in blue) and (in red) at 60 days after it (Kruskal-Wallis test: p-value << 0.001).

This study aims hence at investigating the production performances of Piemontese calves and its optimization for fattening purposes but also for the calf's reproductive career. In particular, the intention is to extend the horizon by investigating which administrative and production variables available in the dataset may influence the production of calves. Studies conducted so far within the association are based on traditional statistical identification approaches. Actual modelling involves only two variables, without exploiting the huge number of parameters in the dataset [1, 2, 3]. Without making a priori bio-, epi-, or eco- logical assumptions about data or the relationship between the response and the independent variables, even if still uncommon for this class of problems, Machine Learning (ML) techniques may provide interesting feature selection characteristics, representing a flexible and robust alternative in predictors

identification. Specifically, the potential of Genetic Programming (GP) [4, 5] is investigated to create and to analyze predictive models for the number of weaned calves in Piedmontese cattle breedings, which could improve the analysis of Piedmontese breeding performance. Inside the ML arena, we chose to use GP, because this technique has a set of interesting characteristics, that distinguish it from many other methods. First of all, it assumes no hypothesis about the shape of the final model, which is very important for the problem under investigation, considering that no a priori knowledge is given. Secondly, using some precaution, GP can be able to generate readable and interpretable models, which is crucial for our application. Finally, GP is able to perform an automatic feature selection, thus relieving us from any pre-processing task. These models are compared with the predictive model that is currently adopted by the National Association of Piedmontese Cattle Breeders ANABORAPI to monitor the progress of each farm.

The paper is organized as follows: in Section II, the background is described. Then, the dataset is analyzed and some basic assumptions on the model are made in Section III. GP models and their performance are illustrated in Section IV, where hypothesis tests and results are examined. Finally, discussions are presented, and further developments are highlighted in Section V.

2. Background

The 'Piemontese' is an Italian bovine breed native of Piedmont and represents a characteristic element of the territory. It is the major bred breed among beef cattle in the region, showing both organoleptic and zootechnical remarkable qualities. If, on one side, it results in greater tenderness of the meat, on the other hand, it is a breed with exceptional character skills, such as meekness, maternal attitude, resistance to diseases, little stress, and great adaptation to pasture. It, therefore, allows easy management and, not less important, the use and development of the local area [1, 3]. The association ANABORAPI is responsible for promoting the breed through the study of the productive, reproductive and management processes of the Piemontese breeding [6]. The activity is carried out with the management of the Herd Book of the Race, a complex database that preserves the pedigrees of all the registered animals and a series of additional information, such as validation of breed characters, reproductive career, morphological studies, and genetic values. Nowadays, these activities must deal with new needs, increasingly connected to the sustainability of breeding and well-being of animals, in the perspective of monitoring every animal. The contribution of each individual is the concept behind Precision Livestock Farming (PLF). It is the solution to avoid imprecise or non-objective farmers evaluations and to facilitate management methods, to obtain hence the best profit both for the individual and the community. ANABORAPI offers to its members a wide section of statistics, which provide a detailed analysis of various

parameters of technical and economic efficiency of the farm and can contribute to identifying the breeding strengths and critical points for possible improvements or developments. In particular, the average situation of breeding due to the main fertility parameters is monitored, summarized by the average number of calves per cow produced in the last year, net of mortality and calving interval (the period between two deliveries of a cow). This is then translated into a brief economic summary, which compares the gross revenue with the mortality losses, providing the farmer with an economic indicator of breeding performance.

A large amount of data is now collected through the use of sensors, ear tags, collars, images and video recordings in many fields, and livestock sector is not different [7-9]. It is increasingly common in farms to monitor each animal: as already mentioned, the PLF approach aims for greater accuracy on the quantity and quality of information, to achieve the economic and environmental sustainability of farms. The breeder must generally deal with animals' problems like their health conditions and social behavior, that affect the quality of the product, the life of the animal, and the performance of the farm. Indeed, the PLF approach provides the offset of incurred costs, as these issues are identified in advance, allowing decisions to be made in time [10, 11]. The creation of prediction models on a specific result in the zootechnical field is increasingly addressed with the use of ML techniques [10-20]. These approaches are suitable for the management of large data sets and are used to predict livestock issues such as the time of disease events, risk factors for health conditions, and failure to complete a production cycle. Studies have been conducted, based on the application of ML techniques, to model the individual intake of cow feed [12], optimizing health and fertility, to predict the rumen fermentation pattern from milk fatty acids [13], which influence the quantity and composition of the milk produced but also the sensorial and technological characteristics of the meat. The use of ML techniques is also often exploited to identify potential disease predictors, e.g. Bovine Viral Diarrhoea Virus (BVDV), Infectious Bovines Rhinotracheitis (IBR), Bovine Tuberculosis (TB), lameness, and mastitis [14-16, 21], to classify grazing and social behavior [17-19], and to predict carcass conformation [20], an important component of price negotiations between beef producers and market operators. These works are mostly carried out on dairy cattle, which are more critical to manage from a health point of view. Dairy animals generally have a shorter average life compared to the lifespan of beef bovines and are usually affected by diseases and metabolic problems. In the beef cattle sector, and in particular in the Piedmontese cattle breed, animals are more resistant and exposed to fewer stress factors. This is an explanation why meat farms show moderate use of devices. However, individual information is already recorded and loaded by the technicians during the checks, and therefore the management of big data is necessary.

3. The Dataset

The content of the dataset elaborated by the ANABORAPI system covers a total of over 4000 active farms, keeping historical records for all of them. The elaboration processed by the ANABORAPI system to evaluate Y_a (see Equation (1)) goes back 365 days, starting from the last check, to process the average summaries. A first restriction is therefore the isolation of the data of a whole year (in our case 2017) and to consider the target we want to infer for the following year (2018). Since the performance of the farm mainly focuses on fertility, the data concerning multiparae cows were considered to elaborate the number of deliveries and the calving intervals. In the same way, data on bulls used for artificial insemination were maintained (i.e. selection indices, that represent namely estimations of the additive genetic effect of a subject). Information referred to inbreeding levels between animals were not incorporated into the study, since they required more investigations. However, they will be included in future developments, for a more accurate inspection on the consanguinity of unborn calves. Finally, restrictions on farms were imposed to obtain a solid representative subset: filters on breeding located in Piedmont with at least 30 cows and percentage of artificial insemination between 90% and 100% were applied. This last condition means that a part of the considered farms actually own bulls and carry out natural impregnations. Thereby, two main groups of farms result from the selection: a smaller one, containing 330 farms, and a larger one, consisting of 395 breedings, resulting in a total of 725 breedings. The difference between the two sets results in a major use of the breeding bull: this means that instead of recording the date on which the insemination took place, breedings belonging to the second group use to set a period of several days, followed by the diagnosis of the pregnancy. As both datasets are representative for the Piedmontese breeding reality, where the second dataset features a more diffused situation and the first one the most accurate one, we used both groups in the study, as propaedeutic to the objective. Since the aim is the building of predictive models via a ML technique, we therefore decided to designate the first set of farms (size 330) as a learning set, as the algorithm can learn on precise recordings, while the second set (size 395) was designated as a test set. Each record of the final datasets stands for a single farm and variables {1 – 19} refer to year 2017, whereas Y is the actual number for weaned calves recorded in 2018 (Table I). All variables can only assume positive values.

4. Application of GP

GP main traits.

The GP technique is a tree-based algorithm, in which the initial population evolves through mechanisms of selection, mutation, and recombination of individuals (i.e. mutation and crossover), as in a biological evolutionary process.

Subtrees at each generation are recombined and recursively evaluated. The best candidates are eligible for the new generation and they are on average fitter than previously generated individuals, i.e. show a smaller error. The error is measured with a fitness function, that is an objective function used to evaluate the distance from the experimental target. with a regression problem we have chosen as fitness function the Root Mean Square Error (RMSE) between the expected and predicted numbers as the measure for the fitness of the models: lower values represent better solutions (i.e. expressions fitting well correspond to low error levels).

Dataset partitioning.

For our experimental study, we used the GPLab Toolbox of MATLAB [4]. As mentioned in the previous section, the first group of farms (size 330) was used as a learning set, while the second one (size 395) as a test set. We considered the possibility of dividing the datasets through a *k*-fold cross validation approach. However, the reduced set of data does not allow us to find a suitable *k* value: for instance, if we chose a *k* smaller than 10, we would obtain a small number of subsets, leading to a small number of runs (i.e. less than 10). On the contrary, with a *k* greater than 10, we would have a restrained number of records within the test sets (i.e. less than 39 test farms for each run). We used the splitting of the learning set into 30 different subsets, with constant training-validation partitioning (75%-25%). Each division was carried out with a random choice of records at each run with uniform distribution and without repetition, keeping separate training and validation. In other words, among the total 330 learning records, 83 records were chosen to form the validation set, and the remaining 247 were labeled as training ones, reiterating the process with different sets for all the 30 runs. For each run, the individuals obtained on the training set were evaluated on the validation set, in order to select the best ones (i.e. models with the lowest error among the validation set). Finally, the generalization ability of the latters was checked, by analyzing the respective error achieved on the test set.

Terminals, nodes and operators

The GP individuals were generated using a tree-based representation, where the trees were built using a set of terminal symbols T and a set of primitive functional symbols F. The set T was composed by the previously described variables, plus a set of random constants between 0 and 1 generated during the initialization process. The set *F* was equal to *{plus; minus; times; mydivide},* where *plus*, *minus* and *times* indicate the usual operators of binary addition, subtraction and multiplication, respectively, while *mydivide* represents the protected division, that returns the numerator when the denominator is equal to zero. In order to limit overfitting and maintain the models as simple as possible, besides

crossover and mutation, operators such as *shrinkmutation* and *swapmutation* (predefined in GPLab) were used. These two operators respectively exchange a subtree with a terminal node and permutate binary non-commutative functions' elements. Table II reports the employed experimental setting.

Performance evaluation.

The performance of the simulations is reported in Figure 2, where the fitness among the 30 runs on the training, the validation and the test sets are presented. The Lilliefors test, performed with significance level $\alpha=0.05$, showed that a normal distribution can be assumed only on the training set. Hence, we applied a Kruskal-Wallis test ($\alpha=0.05$), under the alternative hypothesis that, at the end of the runs, the RMSEs do not have equal medians. Results entailed that there is no significant difference between the three distributions: given a p-value $p=0.17$, the null hypothesis was not rejected, that is the median values of the errors committed on the three sets are not different. The median value obtained on the test set allows us to affirm that the obtained models are able to generalize well, on unseen data.

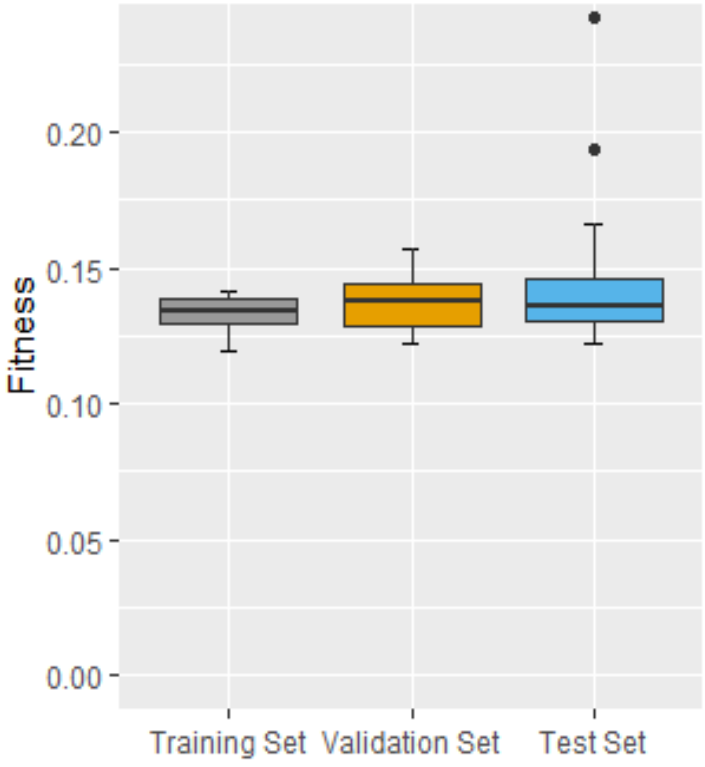

Fig.2 Performance of the best 30 selected models, respectively, on the training, validation and test sets. There is no significant difference between the results (Kruskal-Wallis test: $p = 0.17$, with $\alpha=0.05$), i.e. the median values of the errors committed on the three phases are not different.

Table I. Final set of variables used in the studied dataset. The last line (variable Y) represents the dependent variable, target of the predictive models generated by GP.

	Reference Year	***Variable***	***Description***
1	2017	*COWS*	Consistency for cow (n. of cows in the farm)
2	2017	*C_AGE*	Mean age of cows, expressed in days.
3	2017	*INT_P*	Mean value of calving interval, i.e. the average number of days that elapse between a parturition and the following one.
4	2017	*C_PAR*	Mean number of parturitions of cows.
5	2017	*N_PAR*	Number of occurred deliveries.
6	2017	*C_EASE*	Number of easy parturitions, that did not require human intervention and that did not cause stress to the cow nor the calf.
7	2017	*C_GRAV ID*	Number of pregnant cows.
8	2017	*C_INS*	Number of inseminated cows.
9	2017	*BIRTHW_M*	Mean birth weight of male calves.
10	2017	*BIRTHW_F*	Mean birth weight of female calves.
11	2017	*IND_PAR*	Mean Genetic selection index referred to facility of parturition of the cows.
12	2017	*TFA_BIRTH*	Mean Genetic selection index referred to facility of birth of the bulls, which semen has been used on artificial inseminations.
13	2017	*TFA_PAR*	Mean Genetic selection index referred to facility of parturition with which the bulls, which semen has been used on artificial inseminations, have been born.
14	2017	*N_ELIM*	Number of calves dead within 60 days after birth.
15	2017	*N_TOT*	Total number of newborns.
16	2017	*N_BALIV E*	Total number of calves born alive.
17	2017	*N_CORRECT*	Percentage of calves born without birth defects, such as Macroglossia or Arthrogryphosys.
18	2017	*ABORT*	Percentage of abortions.
19	2017	*MORT*	Mean neonatal mortality.
20	2018 range= [0.26;1.24]	*Y*	Number of calves per cow per year. It is obtained on data from 2018 with the following: $Y = \frac{N\_BALIVE - N\_ELIM}{COWS}$

Table II. Parameters used in our experimental study

Parameter	Description
Maximum number of generations	20
Population size	500
Selection Method	Lexicographic Parsimony Pressure
Elitism	Keepbest
Initialization Method	Ramped half and half
Tournament Size	2
Subtree Crossover Rate	0.8
Subtree Mutation Rate	0.1
Subtree Srinkmutation Rate	0.05
Subtree Swapmutation Rate	0.05

Key role variables result in non-null median frequencies among the best solutions on all the runs, whereas negligible ones correspond to null estimations: values greater than zero suggest that the corresponding variables were used in over 50% of the final solutions, namely the number of cows (*COWS)*, the number of occurred deliveries in the farm during the year (*N_PAR)*, and the number of calves that were born alive (*N_BALIVE).* The information was confirmed also by the equivalent percentage, reported in the second column of Table III.

Finally, we investigated the interpretability of the expressions, considering the number of variables involved in each one of the best final models and the corresponding fitness. In order to compare the performance of the GP models, we examined the number of parameters encapsulated in each one, paying attention to the corresponding fitness obtained on the test set (Table IV). Observing Table IV, we can identify a general trend: models that use less variables tend to have a worse fitness (i.e. a larger error) on the test set than those that use more variables. Among the 19 variables in the dataset, the obtained models encapsulate from a minimum of 3 to a maximum of 10 variables. An intermediate situation is represented by models involving 4 of these parameters, since in this case the error is small and, as shown later, the expression is interpretable. We selected two models in order to make comparisons, the one showing the best fitness among all the evolved expressions (*GP*3 in Figure 3) and the one with the best fitness among the models that use 4 variables (*GP*8 in Figure 3). The choice of the second model was entailed, as shown below, as a consequence of its interpretability. For both Models *GP*3 and *GP*8, the distance values between predictions based on 2017 and target values Y_i recorded in 2018 are represented through boxplots, that is:

$$\Delta_{model,i} = Y_{model,i} - Y_i$$

for each record *i* = 1…,395 in the test set. Predictions obtained with the two models *GP3* and *GP8* are not significantly different (Kruskal Wallis: p-value = 0.2372).

Table III. Median frequencies and percentage of use of each variable among the best 30 individuals found by GP.

Variable	*Median*	*% of use on 30 runs*
X_1 - *COWS*	1	73%
X_2 – *C_AGE*	0	27%
X_3 - *INTP*	0	43%
X_4 – *C_PAR*	0	27%
X_5 – *N_PAR*	1	53%
X_6 – *C_EASE*	0	40%
X_7 – *C_GRAVID*	0	23%
X_8 – *C_INS*	0	17%
X_9 – *BIRTHW_M*	0	13%
X_{10} – *BIRTHW_F*	0	10%
X_{11} – *IND_PAR*	0	37%
X_{12} – *TFA_BIRTH*	0	13%
X_{13} - *TFA_PAR*	0	23%
X_{14} – *N_ELIM*	0	37%
X_{15} – *N_TOT*	0	43%
X_{16} – *N_BALIVE*	0.5	50%
X_{17} – *N_CORRECT*	0	37%
X_{18} - *ABORT*	0	23%
X_{19} - *MORT*	0	13%

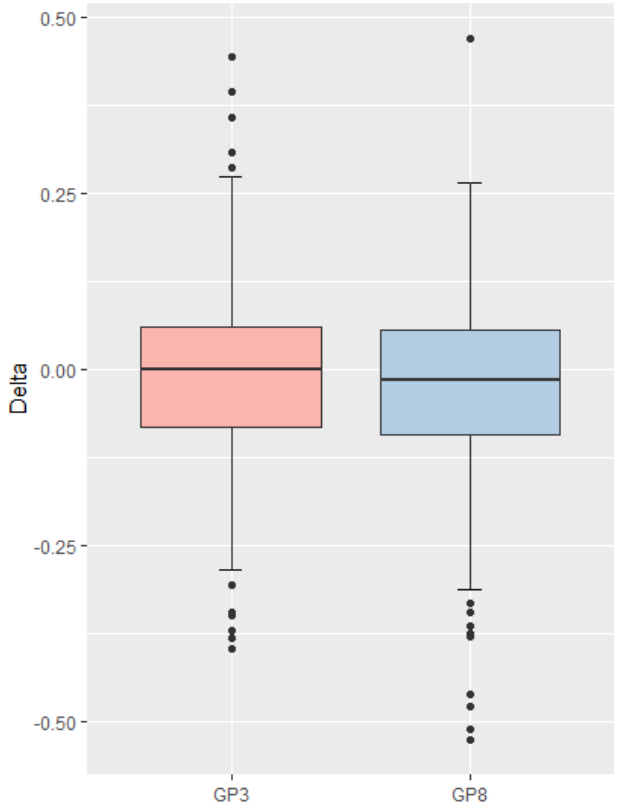

Fig.3 Comparisons between GP models on the test set. Distributions of the differences between predicted and real values are plotted. Both GP predicted values are not significantly different (Kruskal-Wallis: *p - value* = 0.2372). *GP*3 shows a median value equal to -0.0005928782, smaller than the median value obtained with GP8 (-0.0146762341).

Table IV.Fitness on the test set, number of involved variables and corresponding percentage are reported for each model evolved by GP in each one of the 30 performed runs.

Prediction model	***Fitness on Test***	***N. of variables***	***% of variables***
model 1	0.1379	5	26%
model 2	0.1418	3	16%
model 3	0.1218	9	47%
model 4	0.1354	8	42%
model 5	0.1660	3	16%
model 6	0.1290	8	42%
model 7	0.1370	4	21%
model 8	0.1321	4	21%
model 9	0.1258	8	42%
model 10	0.1357	3	16%
model 11	0.2422	9	47%
model 12	0.1461	3	16%
model 13	0.1286	7	37%
model 14	0.1548	4	21%
model 15	0.1320	9	47%
model 16	0.1261	7	37%
model 17	0.1285	8	42%
model 18	0.1371	9	47%
model 19	0.1610	3	16%
model 20	0.1571	4	21%
model 21	0.1355	9	47%
model 22	0.1450	3	16%
model 23	0.1291	7	37%
model 24	0.1426	4	21%
model 25	0.1935	5	26%
model 26	0.1330	10	53%
model 27	0.1305	6	32%
model 28	0.1543	3	16%
model 29	0.1308	7	37%
model 30	0.1361	9	47%

We therefore concluded that the two models, whose expression is provided in Equations (3) and (5), perform likewise, incorporating different variables with respect to Y_a (see Equation (1)). Parameters such as *MORT* and *N_ELIM* used in Equation 1 were encapsulated also in GP expressions, i.e. mortality at 60 days (*GP*8) and number of calves born alive (*GP*3 and *GP*8). Regarding *GP*3, the expression in infix notation to obtain the predictions is:

$$Y_{GP3} = \frac{X_{11}}{X_{17}+\frac{X_3}{X_{16}}+\frac{X_3}{X_6 \cdot \frac{2 \cdot X_{18}+X_{16}}{\frac{X_9}{X_{19}}+X_1}}}, \qquad (3)$$

where

X_1-*COWS*,
X_3-*INTP*,
X_6-*C_EASE*,
X_9-*BIRTHW_M*,
X_{11}-*IND_PAR*,
X_{16}-*N_BALIVE*,
X_{17}-*N_CORRECT*,
X_{18}-*ABORT*,
X_{19} – *MORT*.

In model *GP3*, the denominators of *mydivide* operator do not meet existence conditions, that is they can assume null values (e.g. perinatal mortality X_{19} is null for some records). It is not possible to assert that the *mydivide* operator is actually a division and the previous expression (3) cannot be further simplified. Contrarily to *GP*3, the model for *GP*8 is comprehensible:

$$Y_{GP8} = \frac{X_5}{\frac{(X_5 \cdot X_{14} + X_{16})}{X_1} + X_1}. \quad (4)$$

Since we previously set the constraint in the dataset on farms with more than 30 cows, and the other variables can even assume only positive values, the denominators of *mydivide* that appear in the latter model are also positive (in Model 4, the mentioned values cannot reach null levels, since the number of cows is added to a quantity, greater than zero). Existence conditions are in this case always verified and therefore the function *mydivide* is a division, leading to a simplified version:

$$Y_{GP8} = \frac{X_1 \cdot X_5}{X_1^2 + X_5 \cdot X_{14} + X_{16}}, \quad (5)$$

where

X_1 - *COWS*
X_5 – *N_PAR*
X_{14} – *N_ELIM*
X_{16} – *N_BALIVE*.

Model (5) can further be rewritten as

$$Y_{GP8} = \left(\left(\frac{N\_PAR}{COWS} \right)^{-1} + \frac{N\_ELIM}{COWS} + \left(\frac{N\_BALIVE}{COWS} \cdot \frac{1}{N\_PAR} \right) \right)^{-1}. \quad (6)$$

The first term can be expressed as the invers of the number of mean value of the yearly deliveries occurred in the farm, since the number of all parturitions is divided by the total number of cows ($\overline{N\_PAR}$). Likewise, the second and third terms contain, respectively, the yearly number of calves per cow that did not survive during the weaning period ($\overline{N\_ELIM}$) and the yearly number per cow of calves born alive ($\overline{N\_BALIVE}$), that is:

$$Y_{GP8} = \left(\frac{1}{\overline{N\_PAR}} + \overline{N\_ELIM} + \frac{\overline{N\_BALIVE}}{N\_PAR}\right)^{-1}. \quad (7)$$

Stated otherwise, by renaming the terms and performing basic operations, we obtained the following:

$$1 = n_j v_{1,j} + n_j v_{2,j} + n_j v_{3,j}, \quad (8)$$

for j=1,…,725, since we considered the complete dataset with all the selected farms (see Section III), and where:

$$\boxed{\begin{aligned} n_j &= Y_{GP8,j}\,, \\ v_{1,j} &= \left(\overline{N\_PAR_j}\right)^{-1}, \\ v_{2,j} &= \overline{N\_ELIM_j}\,, \\ v_{3,j} &= \frac{\overline{N\_BALIVE_j}}{N\_PAR_j}. \end{aligned}}$$

It is straightforward that Equation (8) can be formulated as the sum of rescaled variables

$$1 = \tilde{v}_{1,j} + \tilde{v}_{2,j} + \tilde{v}_{3,j}, \quad (9)$$

where $\tilde{v}_{i,j} = n_j v_{i,j}$ for *i={1,2,3}*. Thereby, it was possible to measure the contribution of each term in the sum expressed in Equation (9). The distributions of each $\tilde{v}_{i,j}$ was statistically analyzed and the three boxplots were displayed (Figure 4). Extremely significant difference is verified between all variables (Wilcoxon test with Bonferroni correction: *α=0.017*, *p<<0.001*). Moreover, we inspected how far the mean value of each variable is from the unit. We compared, one by one, the three distributions via a single sample Wilcoxon test. We set alternative hypothesis that the distribution shows a mean value *μ≠1*, with *α=0.05*. Once again, we found an extremely statistical difference between the mean value of $\tilde{v}_{i,j}$ from the value 1. Similarly, we compared the distributions with respect to 0. The results of the test were analogous to the previous ones: with extremely significant p-values (p<<0.001), we could deduce that $\tilde{v}_{2,j}$ and $\tilde{v}_{3,j}$ remain relevant parameters, even assuming values close to zero, providing hence a minimal contribute in Equation (8). In other words, we could assert that all the variables in Equation (9) are influent: in particular, $\tilde{v}_{1,j}$ is the most important one, since its mean value was μ_1*=0.951*, whereas $\tilde{v}_{2,j}$ and $\tilde{v}_{3,j}$ respectively showed μ_2*=0.032* and μ_3*=0.021*. Model (5) can be simplified, to the point of being expressed as the sum of three parameters. We verified that these three parameters are the average number of parts occurred during the year in the herd ($(\overline{N\_PAR})^{-1}$), the number of calves per cow that have not passed the weaning phase ($\overline{N\_ELIM}$) and finally the number of calves per cow live births compared to the total number of parts of the herd during the year ($\overline{N_{BALIVE}}/N\_PAR$). From a zoological point of view, these are actually the main parameters that intuitively can give an idea of the economic performance of the farm. All of them play a significant role with respect to the response variable: more importance is given to the parameter $(\overline{N\_PAR})^{-1}$, associable to the

inverse of the mean calving interval (days between two deliveries) of the farm, whereas $\overline{N\_ELIM}$ and $\overline{N_{BALIVE}}/N\_PAR$ give a smaller contribute.

Summing up, the most frequent variables in the models, obtained with GP, are the number of cows in the farm (*COWS)*, the number of deliveries occurred in the breeding (*N_PAR)* and the number of calves born alive (*N_BALIVE)*. The calving interval (*INTP)* and the number of dead calves at 60 days (*N_ELIM)* are slightly less frequent. Perinatal mortality is not so recurring, meaning that it could play a minor role in the prediction. The most frequent variables encapsulated in the expression (3) are *COWS* and *N_BALIVE*, followed by *INTP* and *MORT*. Then there are 5 less frequent additional parameters that could, therefore, be relevant in the refinement of the prediction. The median error of predictions obtained with model *GP*3 is slightly smaller than the one obtained with model *GP8*. The latter however processes less variables, exploiting exactly the three most frequent ones, listed in Table III.

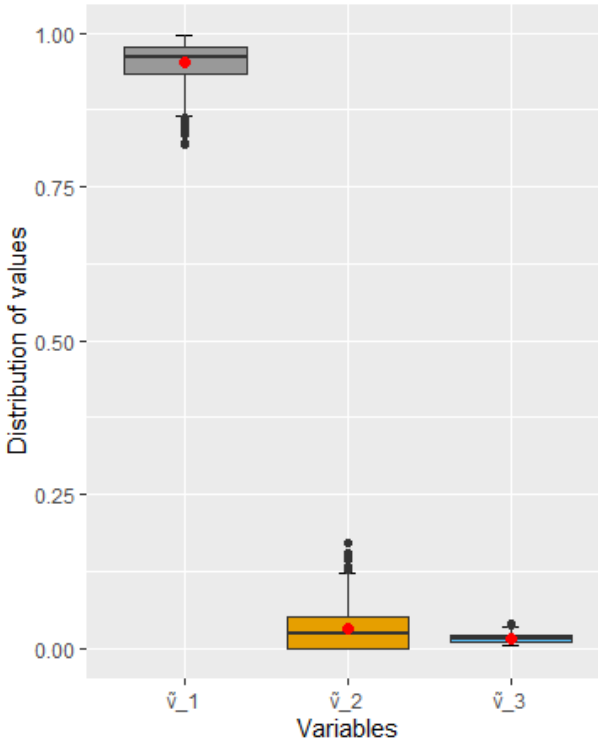

Fig.4 Boxplots of the distributions of the variables in Equation (9). Wilcoxon test with Bonferroni correction at *α=0.017* reported *p<<0.001.* Hence, the variables are significantly different. The *s*ingle sample Wilcoxon test, with *α=0.05,* showed for each distribution a mean value $\mu \neq 1 (p<<0.001)$. Therefore, all the variables are extremely significant in Equation (9). Mean values are respectively $\mu_1=0.951$, $\mu_2=0.032$ and $\mu_3=0.021$*(red dots).*

5. Conclusions and Future Works

In this study, we investigated the performance of medium to large farms located in Piedmont of Piedmontese cattle, starting from the model implemented in the systems of the National Association of Piedmontese Bovines (ANABORAPI) [1-3]. The currently used model (reported in Equation (1)) predicts the number of calves per cow per year. However, it is not completely suitable to represent the performance of the farms. In fact, during the weaning period, many calves do not survive, entailing great losses to the economic revenues of the breedings. The reasons for those deaths are various and difficult to identify objectively. It is hence necessary to take into account crucial parameters, that encompass the calf's weaning in the output, as, for example, the number of calves born alive and those dead after the weaning period, within 365 days. Although biologically acceptable, these hypotheses could not be sufficiently informative or they could be informative enough, but not exhaustively combined in the formulation of a model. It is therefore difficult to build a model with only zootechnical speculations and an automatic learning method has been applied, which can meet the requirements. In addition, it is necessary to research and propose a simple model, which can be easily interpreted by the breeder. The expression to target should be a simplification and an added value to the management of the farm. The breeder should be able to easily read the information, in order to identify the critical points and strengths in production.

Given its ability to perform an automatic feature selection, a Genetic Programming approach (GP) was used applied [4, 5] to build predictive models, trained, validated and tested on data recorded in 2017 and 2018. Accurate models were achieved, and this means that GP can learn from a smaller dataset composed by representative farms and predict good results on the selected test set. Moreover, the algorithm was able to select and process important variables, without previous assumptions on the zoological aspect. The variety of expressions obtained by GP is composed of well-performing models that involve more parameters, resulting in a more complex expression, hardly reducible to a simpler one. However, other predictive models were also achieved that encapsulate fewer variables. Although these expressions have a slightly larger error, their formula can be extremely simple and possibly easier to interpret from the zoological point of view.

It is therefore worth investigating further the application of GP to a larger dataset. In this first study, we focused on data directly referred to parturitions and artificial insemination, in order to process sound and solid data. The dataset was filtered and resized, and 19 variables were kept among 210: many were duplicate fields, aggregates of several variables,

and even incomplete ones, because introduced lately in the database of ANABORAPI. Parameters such as those on heifers, i.e. bovines that did not give birth yet, were not considered, since we focused on data directly referred to cows, i.e. bovines that gave birth at least once. In breeding farms, heifers are mostly intended to the production of calves and are going to contribute to the restock of the herd. The behavior of GP and its features selection ability among these variables will be investigated, as well as among parameters on the bulls used for natural insemination. To this purpose, their genetic indexes will be added to the analysis, as well as the levels of consanguinity of calves that will be born from ongoing pregnancies. Comparisons with other machine learning methods will be performed, to inspect better the potential of GP in the zootechnical field, and to explore possibly better models.

In future developments, data regarding environmental conditions inside the farm will also be taken into account, such as the size of the boxes and the surface available to the animals, air and water quality and the composition of the food ration. These factors are usually considered as marginal. It is common to think that cow-calf problems are almost exclusively induced by genetic and pathological factors associated to pregnancy and childbirth. Indeed, not enough importance is given to the period after the birth, in which the cow and the calf need feeding and environmental conditions, suitable for the respective postpartum and weaning phases. In this context, once again, the ability of GP to automatically select features will be very important to understand if and which of these variables are influential.

ACKNOWLEDGMENT

This work was partially supported by FCT, Portugal, through funding of LASIGE Research Unit (UID/CEC/00408/2019) and projects BINDER (PTDC/CCI-INF/29168/2017), GADgET (DSAIPA/DS/0022/2018), AICE (DSAIPA/DS/0113/2019) and PREDICT (PTDC/CCI-CIF/29877/2017), and by the Slovenian Research Agency (research core funding No. P5-0410).

REFERENCES

[1] Bona, M., Albera, A., Bittante, G., Moretta, A., Franco, G.: L'allevamento della manza e della vacca piemontese, Supplemento al n. 44 dei Quaderni della Regione Piemonte-Agricoltura, pp. 65-129. (2005).

[2] Lo svezzamento del vitello Piemontese,

pp. 3-5, http://www.anaborapi.it/images/media/pdf/rivista/2012/2012-05.pdf

pp. 9 -11, http://www.anaborapi.it/images/media/pdf/rivista/2012/2012-06.pdf

[3] Associazione Nazionale Allevatori Bovini Razza Piemontese, http://www.anaborapi.it

[4] Silva, S.: GPLAB a genetic programming toolbox for Matlab, (2007). http://gplab.sourceforge.net/index.html

[5] Poli, R., Langdon, W., McPhee, N.: A Field Guide to Genetic Programming. Lulu Enterprises, UK Ltd. (2008). https://doi.org/10.1007/s10710-008-9073-y

[6] Relazione Tecnica e Statistiche al 31.12.2018, http://www.anaborapi.it/images/media/pdf/stat/relazionetecnica2018.pdf

[7] Berckmans, D., Guarino, M., From the Editors: Precision livestock farming for the global livestock sector, Animal Frontiers, Volume 7, Issue 1, January 2017, Pages 45. https://doi.org/10.2527/af.2017.0101

[8] J. B. Cole, S. Newman, F. Foertter, I. Aguilar, M. Coffey,: BREEDING AND GENETICS SYMPOSIUM: Really big data: Processing and analysis of very large data sets, Journal of Animal Science, Volume 90, Issue 3, March 2012, Pages 723-733. https://doi.org/10.2527/jas.2011-4584

[9] Lokhorst, C., de Mol, R.M., Kamphuis, C.: Invited review: Big Data in precision dairy farming. Animal. 13(7):15191528. (2019). https://doi.org/10.1017/S1751731118003439

[10] Morota, G., Ventura, R. V., Silva, F. F., Koyama, M., Fernando, S. C.: BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. Journal of animal science, 96(4), 15401550. (2018). https://doi.org/10.1093/jas/sky014

[11] Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine Learning in Agriculture: A Review. Sensors (Basel). 2018;18(8):2674. Published 2018 Aug 14. https://doi.org/10.3390/s18082674

442 [12] Yao, C., Zhu, X., & Weigel, K. A.: Semi-supervised learning for genomic prediction of novel traits with small
443 reference populations: an application to residual feed intake in dairy cattle. Genetics, selection, evolution: GSE, 48(1),
444 84. (2016). https://doi.org/10.1186/s12711-016-0262-5
445
446 [13] M. Craninx, V. Fievez, B. Vlaeminck, B. De Baets Artificial neural network models of the rumen fermentation
447 pattern in dairy cattle. Comput. Electron. Agric., 60 (2008), pp. 226-238. https://doi.org/10.1016/j.compag.2007.08.005
448
449 [14] Williams, M.L., Parthalin, N.M., Brewer, P., James, W.P.J., Rose, M.T.: A novel behavioral model of the pasture-
450 based dairy cow from GPS data using data mining and machine learning techniques. J Dairy Sci., 99(3):20632075. (2016).
451 https://doi.org/10.3168/jds.2015-10254
452
453 [15] R. Dutta, D. Smith, R. Rawnsley, G. Bishop-Hurley, J. Hills, G. Timms, D. Henry, Dynamic cattle behavioural
454 classification using supervised ensemble classifiers. Comput. Electron. Agric., 18–28 (2015),
455 https://doi.org/10.1016/j.compag.2014.12.002
456
457 [16] O. Guzhva, H. Ard, A. Herlin, M. Nilsson, K. strm, C. Bergsten,: Feasibility study for the implementation of an
458 automatic system for the detection of social interactions in the waiting area of automatic milking stations by using a
459 video surveillance system. Computers and Electronics in Agriculture, Volume 127, Pages 506-509, ISSN 0168-1699.
460 (2016). https://doi.org/10.1016/j.compag.2016.07.010.
461
462 [17] Ortiz-Pelaez, A., Pfeiffer, D.U.: Use of data mining techniques to investigate disease risk classification as a proxy
463 for compromised biosecurity of cattle herds in Wales. BMC Vet Res.;4:24. (2008). https://doi.org/10.1186/1746-6148-
464 4-24
465
466 [18] Machado, G., Mendoza, M. R. & Corbellini, L. G.: What variables are important in predicting bovine viral
467 diarrhea virus? A random forest approach. Vet. Res. 46 (2015). https://doi.org/10.1186/s13567-015-0219-7
468

[19] Alonso, J., Castañón, Á.R., Bahamonde, A., Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter. (2013) Computers and Electronics in Agriculture, 91, pp. 116-120. https://doi.org/10.1016/j.compag.2012.08.009

[20] Amrine, D. E., White, B. J., & Larson, R. L.: Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. Computers and Electronics in Agriculture, 105, 9-19. (2014). https://doi.org/10.1016/j.compag.2014.04.009

[21] Bovine Diseases and Resources, http://www.cfsph.iastate.edu/Species/bovine.php