Livestock Science xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci

Classification of Spanish autochthonous bovine breeds. Morphometric study using classical and heuristic techniques

E. Rodero [a], A. González [a,*], M. Luque [b], M. Herrera [a], J.C. Gutiérrez-Estrada [c]

[a] *Departamento de Producción Animal, Campus de Rabanales, Universidad de Córdoba,Campus de Excelencia Internacional Agroalimentario CEIA3, 14071 Córdoba, Spain*
[b] *Departamento Técnico, Federación Española de Asociaciones de Ganado Selecto (FEAGAS), C/Castelló 45 2° Izda., 28001 Madrid, Spain*
[c] *Departamento de Ciencias Agroforestales, ETSI, Campus La Rábida, Universidad de Huelva, Campus de Excelencia Internacional Agroalimentario CEIA3, 21819 Palos de la Frontera (Huelva), Spain*

ARTICLE INFO

Article history:
Received 4 April 2011
Received in revised form 28 July 2011
Accepted 26 September 2011
Available online xxxx

Keywords:
Phenotypic characterisation
Multivariate analysis
Discriminant function analysis
Artificial neural networks
Support vector machines

ABSTRACT

Six morphometric traits (height at withers, height at rump, chest depth, width at hips, width at pins and rump length) were analysed to characterise from a breed point of view 518 females from four autochthonous Andalusian cattle breeds (Berrenda en Colorado, Berrenda en Negro, Cárdena Andaluza and Negra Andaluza). Four methods (one classical and three heuristic) were used to distinguish between the four breeds by morphometric traits: Discriminant Function Analysis (DFA), Multilayer Perceptrons (MLPs) (a type of neural network), Probabilistic Neural Networks (PNNs) and Support Vector Machines (SVMs). Results indicated not only that DFA was overall inferior to the other three methods, but also that it could not be used to distinguish one breed from another when they were genetically very close or related in terms of breeding. MLP and SVM had similar ability to discriminate, both being better than PNN. Sensitivity analysis carried out on the models found to have the best discrimination power indicated that the most important variables were: depth, height at rump and width at pins.

1. Introduction

Although it is difficult to agree on a single widely acceptable definition of breed, it seems clear that the various definitions of the concept have a feature in common, namely, populations that are considered to be of the same breed have similar characteristics (in terms of morphology and production, among others), defining and setting them apart from other groups within the same species. Thence, the need to establish a series of morphological characters that is related to and identifies each breed and may have the following functions: i) morphology as a criterion for describing a breed; ii) morphology as a criterion for differentiating between breeds; iii) morphology as a criterion for identifying breeds and individuals; and iv) morphology as the basis for distinguishing groups of animals and establishing breeds (Sierra, 2009).

Despite the fact that, in general, no breed classification should rely exclusively on biometric data, it is clear that these play a proxy or complementary role in the description of a breed (Parés i Casanova, 2009), in particular with regard to breeds of farm animals. This is due to the effect of environmental, geographical, physiological, nutritional and even pathological aspects of the morphological and phenotypic characters that identify a breed. In relation to this, analyses of morphometric variables that are easy to measure makes it possible to explore areas such as the structure of breeds, the degree of variability between various populations, the harmony of morphological models and the definition of morphological models for given breeds (Herrera, 2007). Hence it is important to accurately analyse the morphological variables that enable us to distinguish between breeds, as well as explore the use of various discrimination methods to assess the potential of each of the variables under study.

In the context of farming, the issue of discriminating between the various breeds using morphometric variables has

* Corresponding author.
E-mail address: v32gomaa@uco.es (A. González).

1871-1413/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.livsci.2011.09.022

been addressed by various authors in recent years. Herrera et al. (1996) and Luque et al. (2005) used classical discriminant function analysis (DFA) (Lachenbruch, 1975) to differentiate between various Andalusian goat breeds. In all these cases, the method was found to be relatively efficient and allowed differences between breeds and subpopulations to be detected, as well as the relative distance between them to be assessed. Further, the variables with greatest power to differentiate were identified enabling them to be weighted accordingly. However, despite the statistically significant results obtained in these studies, it is necessary to explore new classification methods that may be more accurate for distinguishing between breeds and allow a better understanding of the involvement of each variable in the process of classification.

In recent years, comparisons of classical statistical and heuristic methods have generally shown better performance of the latter type across a range of different science and engineering applications. In the context of the differentiation of autochthonous breeds, the DFA technique has not been compared with heuristic methods like Artificial Neural Networks (ANNs), Fuzzy Logic Classification Functions (FLCFs) or Support Vector Machines (SVMs). In general, artificial neural networks, and particularly Multilayer Perceptrons (MLPs) and Probabilistic Neural Networks (PNNs), are well known techniques in some academic fields such as ecology (Goethals et al., 2007; Gutiérrez-Estrada and Bilton, 2010; Gutiérrez-Estrada et al., 2008; Lek and Guegan, 1999; Lek et al., 1996), fisheries sciences (Gutiérrez-Estrada et al., 2007, 2009; Haralabous and Georgakarakos, 1996; Robotham et al., 2010), agricultural sciences (Pulido-Calvo and Gutiérrez-Estrada, 2009; Pulido-Calvo et al., 2003, 2007) and hydrology (Adamowski and Karapataki, 2010; Fernando et al., 2009; Pulido-Calvo and Portela, 2007). On the other hand, the Support Vector Machine (SVM) method is a relatively new technique developed as a tool for recognising patterns or discriminating between groups (Haralabous and Georgakarakos, 1996; Vapnik, 1995). Therefore the main objective in this study was the classification of four Andalusian authochthonous bovine breed for what the performances of a classic DFA and three heuristic classification techniques (MLP, PNN and SVM) were evaluated. Also, as a second objective, we analysed the weight of each morphological variable in the discriminatory ability of the model. The discrimination study was carried out using morphological descriptors of four autochthonous cattle breeds: Berrenda en Colorado (BC), Berrenda en Negro (BN), Cárdena Andaluza (CA) and Negra Andaluza (NA.

2. Material and methods

2.1. Data collection and morphometric variables

The assessment of the morphological models was carried out in a population of 518 cows belonging to these four breeds autochthonous to Andalusia: Berrenda en Colorado ($n = 179$), Berrenda en Negro ($n = 214$), Cárdena Andaluza ($n = 48$) and Negra Andaluza ($n = 77$). For each individual the following morphometric variables, considered to be quantitative independent variables, were assessed: i) height at withers (HW, cm); ii) height at rump (tuber coxae) (HR, cm); iii) chest depth (ChD, cm); iv) width at hips (WH, cm); v) width at pins (WP, cm); and vi) rump length (RL, cm) (Alderson, 1999; Aparicio et al., 1986; Aparicio-Sánchez, 1960; Rodero et al., 1994). Further, the ratios between HW and HR, and between WH and WP were considered as independent variables.

2.2. Classical discriminant function analysis (DFA)

Discriminant function analysis is a statistical technique that allows new individuals to be assigned to previously established or defined groups. The analysis is based on a set of data from n individuals for which p quantitative variables have been measured (independent variables) as a profile for each of them. On the other hand, an additional qualitative variable (dependent variable), with two or more categories and defined by other means, groups each individual in a category. This produces an $n \times (p + 1)$ table in which each case has a profile and is assigned to one group. From this table, a discriminant model is obtained to compare to the profile of new individuals. A complete description of the method can be found in Dossa et al. (2007), Hair et al. (1999), Herrera et al. (1996), Lanari et al. (2003), Macciotta et al. (2002), Rodero et al. (2003), and Zaitoun et al. (2005) among others.

2.3. Artificial neural network models: MLP and PNN

Artificial neural networks (ANNs) are mathematical models inspired by the neural architecture of the human brain. The most widely studied and used type is the multilayer perceptron (MLP) (Rumelhart et al., 1986). These models 'learn' in an iterative way, with the data set being presented to the neural network the necessary number of times to reach a given level of error (each iteration in which all the data set is presented to the MLP is known as an epoch). These supervised ANNs allow complex data sets to be analysed and their non-linear separation into two or more groups. A detailed description of MLP performance can be found in Czerwinski et al. (2007), Gutiérrez-Estrada et al. (2000, 2007), Pulido-Calvo and Portela (2007), and Tsoukalas and Uhrig (1997).

A typical three or four layer MLP has one input layer, one or two hidden layers, and one output layer. The processing elements in each layer are called nodes or neurons. In our case, the input data to the MLP are morphological traits and the output corresponds to the classification results (breed). The neurons are connected through a set of connections referred to as weights, which are analogous to synapse strength in biological neural nets. There are many MLP calibration or learning methods. In this work the standard backpropagation algorithm was applied and solved using STATISTICA®, the Neural Network software package from StatSoft (2005).

Probabilistic neural networks (PNNs) are another type of ANN (Specht, 1990). The neural network architecture for PNNs contains a sequence of layers: input layer (features of breeds), pattern layer (cattle breeds for calibration samples), summation layer (density function value by breed) and output layer (results of classification). The PNN method provides a general solution for classifying cattle breeds based on Bayes classification technique. The idea is that, given a sample pattern, we can make a decision as to the most likely breed that sample is taken from. PNNs use a probability density function as the transfer function. The probability density function is

Please cite this article as: Rodero, E., et al., Classification of Spanish autochthonous bovine breeds. Morphometric study using classical and heuristic techniques, Livestock Science (2011), doi:10.1016/j.livsci.2011.09.022

estimated using multi-dimensional kernels in the pattern layer. In our work, as in Rutkowski (2004), a Gaussian kernel was used as the activation function; this being controlled for the standard deviation, the width of the activation function. Thus, a PNN essentially constructs an estimate of the probability density function of each breed (class) by adding together Gaussian curves located at each point in the calibration set. There is no training with PNN, in the sense that there is with MLP, since the set of weights are determined from the calibration data.

2.4. Support vector machines: SVM

The support vector machine method is a statistical classification technique proposed by Vapnik (1995); it belongs to the family of linear classifiers as it seeks to separate the space of input characteristics by hyperplanes. At an algorithmic level, the learning of SVM is modelled as a quadratic optimization problem with linear constraints, the size of which depends on the dimension of the characteristic space.

To construct an optimal hyperplane, SVM employees an iterative training algorithm, which is used to minimise an error function. Typically, when the objective is to classify categorical variables the form of the error function can be classified into two groups: Classification SVM Type 1 (also known as C-SVM classification) and Classification SVM Type 2 (also known as ρ-SVM classification). In this work, the classification SVM type used was C-SVM. For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2}\, w^T w + C \sum_{\mathrm{i}=1}^{N} \delta_\mathrm{i}$$

subject to the constraints:

$$y_i \left[w^T \varphi(x_i) + b \right] \geq 1 - \delta_i \quad \delta_i \geq 0,\ i = 1, \ldots, N$$

where C is the capacity constant, w is the vector of coefficients, b is a constant and δ_i are parameters for handling non-separable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ is the class labels and x_i is the independent variables. The kernel φ is used to transform data from the input (independent) to the feature space.

In order to find a classifier function using the support vector machine, we must first determine what type of kernel function is going to be used, as this should reflect a priori knowledge of the problem. If it seems that the data might not be linearly separable, for example, in our case we have two different breeds namely Berrenda en Colorado (BC) and Berrenda en Negro (BN) in which the height at withers of the individuals is very similar, kernel functions developed for non-vectorial structures should be used (e.g., polynomial, Gaussian, sigmoidal, or inverse multiquadratic kernels). On the other hand, for classification, Gaussian kernels are widely recommended in the literature (e.g., Scholkopf and Smola, 2002), since only the γ parameter of the kernel must be estimated and also, they tend to be more stable. Another parameter that needs to estimated, whichever the kernel, is C; this represents a trade-off between the size of the margin and the training error.

2.5. Data processing: calibration and testing samples

The general procedure employed for the calibration and testing of DFA, MLP, PNN and SVM is outlined in Fig. 1. Before the calibration of any model the data set was randomly divided in two subsets: the first one (calibration subset, CS) was composed of 75% of total patterns and the second one (test subset, TS) was composed of 25% remaining. In the case of MLP, 25% of calibration subset was selected as control or select subset (also randomly selected) to avoid overtraining phenomena.

For each model and configuration, the previous procedure was carried out 30 times (Iyer and Rhinehart, 1999). Therefore, we adopted a K-Fold cross validation method with randomly selected subset. Once each model and configuration were calibrated and tested a model ensemble process was carried out. Ensemble is the most important means of combating the over-calibration and improving the generalisation capacity of the models (Watts and Worner, 2008). The best approach was then selected as the ensemble with the smallest error in the validation phase.

Obviously, each technique has different features which depend on their mathematical natures. For example, for MLP is necessary know the number of hidden layers and the number of neurons by layer. Therefore different procedures were conducted to reach the optimum configuration of each model. This way, for MLP one and two hidden layers were tested and the number of neurons in each hidden layer oscillated between 5 and 20.

In the case of PNN, as noted above, there is no training phase (in the sense of that is required for the MLP method) since all the network parameters (units and weights) are determined directly from the calibration data. In our study, smoothing factors (variance) between 0.1 and 0.3 were used; the value of this variance controlling the width of the activation function (Gaussian curve).

For the SVM method, two parameters need to be estimated: C representing the trade-off between the size of the margin and the calibration error, and γ representing the parameter of the Gaussian kernel. There is no single accepted procedure for estimating these two parameters. In this study, we investigated all the combinations of parameters in the range [1, 500] with step sizes of 50 (in the case of C) and in the range [0.05, 10] with step sizes of 0.05 (for γ). Each trial consisted of 100 experiments with the same set of parameters (C,γ) and the average classification error for the testing data was calculated. In each of the iteration, the data set being used for calibration and testing changed, i.e., each time 75% of the data were chosen at random for calibration and the remaining 25% were used for testing and calculating the classification errors. Once the best values for the parameters (C,γ) were found, smaller ranges were explored around these values, with step sizes of 0.01 for γ and 1 for C. The same 30 calibration and testing samples used with the rest of techniques was then used with the best pair of values for C and γ.

2.6. Data processing: multi-class SVM

Support vector machines (SVM) were originally designed for binary classification. However, they can also be used for

Please cite this article as: Rodero, E., et al., Classification of Spanish autochthonous bovine breeds. Morphometric study using classical and heuristic techniques, Livestock Science (2011), doi:10.1016/j.livsci.2011.09.022

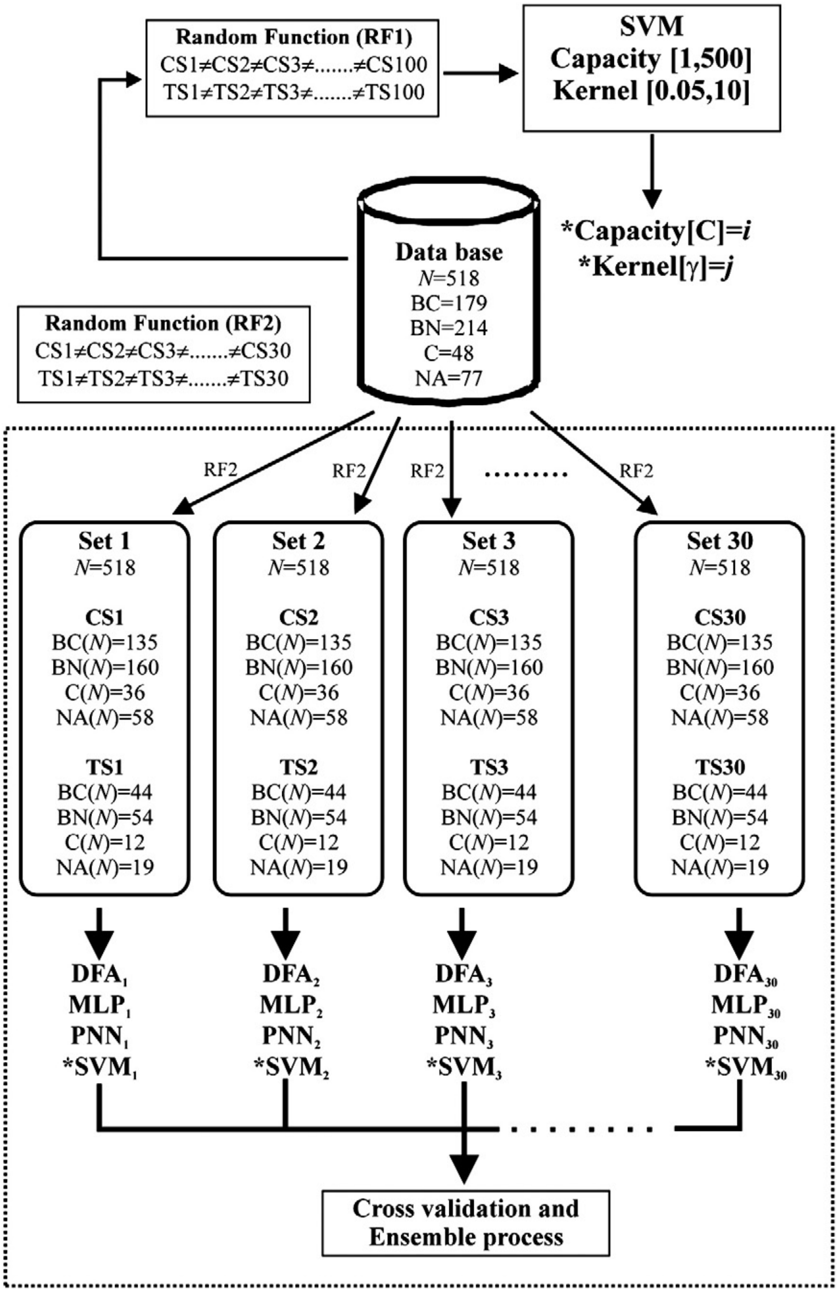

Fig. 1. Schematic representation of the general procedure followed for the calibration of models. CS = calibration set, TS = validation set. Berrenda en Colorado (BC), Berrenda en Negro (BN); Cárdena Andaluza (C) and Negra Andaluza (NA). *Indicate the previous selection of the parameters Capacity[C] and Kernel[γ] for SVM.

multi-class problems. In general, two strategies are used to approach multi-class SVM problems (Hsu and Lin, 2002). In the first, a series of binary classifications is solved in distinguishing between two approaches: these would be one breed compared to another (1-to-1) and one breed compared to the rest (1-to-R), given that we are analysing breeds. The second strategy is to consider all the data in a single optimization formulation, resulting in a problem that is far more difficult to solve numerically. In our study, we used the first approach to classify our multi-class SVM, applying both 1-to-1 and 1-to-R criteria. In the case of 1-to-1, $k(k-1)/2$ classifier functions have to be created between two breeds, where k represents the total number of breeds being classified, in our case $k=4$ and, accordingly, we designed four classifier functions. A voting strategy is then used to decide to which breed a given individual corresponds. Votes are obtained depending on the evaluation of each classifier function, in the vector of descriptors for each individual used for testing. With the 1-to-R binary classification, the classifiers are defined by labelling the breeds to be identified $+1$ and the other breeds -1. The classification corresponds to the breed for which the classifier functions, evaluated in the

Please cite this article as: Rodero, E., et al., Classification of Spanish autochthonous bovine breeds. Morphometric study using classical and heuristic techniques, Livestock Science (2011), doi:10.1016/j.livsci.2011.09.022

vector of descriptors for each individual used for testing, had the highest value.

2.7. Sensitivity analysis

The sensitivity analysis was carried out by replacing each variable (descriptor) by missing values and assessing the effect of this on the output error. Subsequently, the new error calculated was compared with the original error to obtain a ratio (error of the model with a variable with missing values/ error of the model with all variables complete). For a given variable, a ratio with close or equal to 1 indicates that this variable has a very low weight in the general structure of the model (Hunter et al., 2000).

3. Results and discussion

The results obtained clarify certain questions arising with respect to differentiating and characterising breeds, and the possibility of distinguishing between close breeds, as well as identifying the appropriate morphometric characters to be taken into account for this characterisation. Specifically, Table 1 shows the rates of correct classification for each of the models used in two different configurations: i) one breed compared directly with another (1-to-1), and each breed compared to the rest of the breeds (1-to-R). Overall, it can be observed that the classical statistical method (DFA) is less able to correctly classify the animals than any of the heuristic methods. The weakness of DFA for classifying individuals is clear in the validation phase with respect to discriminating between Cárdena Andaluza and Negra Andaluza and between these two breeds and the rest, and also between Berrenda en Colorado and Berrenda en Negro. With regard to the first two breeds, the results may be conditioned by the influence of Negra Andaluza on Cárdena Andaluza, some farms being in the same geographical area and the breeds having common ancestors (Pastor-Fernández, 2010).

From the methodological point of view, the results observed may be the consequence of attempting to carry out the classification using DFA on an unbalanced data set. On the other hand, our findings might also be taken to indicate that for these breeds, a discriminant function able to distinguish between them needs to have non-linear characteristics, which would be incompatible with DFA. This is, in part, confirmed by the results obtained using the three heuristic models, which are able to model highly nonlinear relationships between entry (independent) and the outcome (dependent) variables. Overall, the DFA method seems to be inadequate for statistically accurate classification of populations belonging to these four breeds.

With all of the heuristic methods, the rates of correct classification in both the calibration and validation phases are statistically acceptable for both configurations (1-to-1 and 1-to-R), poorer results being found for differentiating Berrenda en Colorado from Berrenda en Negro and also the former from the rest of the breeds. In any case, the rates of correct classification are always above 61% in the validation phase using PNN. This difficulty of differentiating between Berrenda en Colorado and Berrenda en Negro is not surprising since the ancestry of the two breeds is uncertain, several studies indicating that they diverged relatively recently (13.22 generations, that is 165 years ago) (Azor et al., 2006).

The degree of effectiveness of each method in distinguishing between the four breeds is clear when the models are calibrated and validated with a multi-class approach, that is, when the model is used to attempt to classify an individual into one of four classes. Table 2 shows the confusion matrices for each of the four types of models. The lowest average in the main diagonal in the validation phase (49.0%) was obtained with DFA. In this case, the greatest level of error was again found in the classification of Berrenda en Colorado and Cárdena Andaluza.

However, the correct classification rate was significantly improved when heuristic approaches were used. Among the three heuristic models, the best absolute average rate for correctly classifying individuals was achieved with MLP (62.1%), although this value was only 1.5 points better than obtained with SVM. The PNNs, although they performed better than DFA, only achieve an average of just over 53% of accuracy in the validation phase. Overall, these results once again demonstrate the difficulty of classifying the four breeds with DFA, at least with the entry variables used. Therefore, a multi-class approach seems to indicate to us that the most nearest breeds and consequently most difficult to classify

Table 1
Correct classification rates (%) in validation phase using discriminant function analysis (DFA), multilayer perceptrons (MLP), probabilistic neural networks (PNNs) and support vector machines (SVMs) according to the type of binary partition one-to-one (1-to-1) and one to the rest (1-to-R). Berrenda en Colorado (BC), Berrenda en Negro (BN), Cárdena Andaluza (C) and Negra Andaluza (NA).

Type of partition	DFA	MLP	PNN	SVM
1-vs-1				
BC-vs-BN	47.7-vs-72.2	72.7-vs-63.0	70.2-vs-61.0	72.2-vs-63.0
BC-vs-C	86.4-vs-66.7	95.5-vs-83.3	95.1-vs-82.0	95.5-vs-83.3
BC-vs-NA	88.6-vs-84.2	93.2-vs-89.5	91.2-vs-83.5	93.2-vs-89.5
BN-vs-C	98.2-vs-75.0	98.1-vs-100.0	98.1-vs-98.0	98.1-vs-100.0
BN-vs-NA	96.3-vs-94.7	94.4-vs-94.7	94.1-vs-93.7	94.4-vs-94.7
C-vs-NA	33.3-vs-94.7	75.0-vs-94.7	72.0-vs-94.0	74.9-vs-94.1
1-vs-R				
BC-v-R	15.9-vs-91.8	65.9-vs-79.6	61.6-vs-69.4	65.5-vs-79.4
BN-v-R	61.1-vs-69.3	75.9-vs-72.0	72.6-vs-69.0	75.1-vs-71.0
C-v-R	16.7-vs-95.7	66.7-vs-97.4	66.5-vs-97.4	66.7-vs-97.4
NA-v-R	47.4-vs-93.6	89.5-vs-93.6	89.3-vs-93.1	89.3-vs-93.1

Table 2
Confusion matrices for multi-class classifications. Percentages of successful recognitions by breed in the validation phase. Berrenda en Colorado (BC), Berrenda en Negro (BN), Cárdena Andaluza (C) and Negra Andaluza (NA).

Breed	BC	BN	C	NA
DFA classification (average rate in the validation set = 49.0%)				
BC	34.1	47.7	6.8	11.4
BN	29.6	68.5	0.0	1.9
C	41.7	0.0	25.0	33.3
NA	10.5	5.3	15.8	68.4
Multilayer perceptron (MLP; 8-20-1) classification (average rate in the validation set = 62.1%)				
BC	51.3	38.5	7.7	2.6
BN	31.0	65.5	0.0	3.5
C	14.3	0.0	71.4	14.3
NA	20.0	4.0	16.0	60.0
Probabilistic neural network (PNN; λ = 0.3) classification (average rate in the validation set = 53.4%)				
BC	43.6	46.2	7.7	2.6
BN	34.5	58.6	1.7	5.2
C	14.3	0.0	71.4	14.3
NA	16.0	1.0	40.0	40.0
Support vector machine (SVM; C = 100; γ = 0.1) classification (average rate in the validation set = 60.6%)				
BC	48.7	41.0	7.7	2.6
BN	34.5	62.0	0.0	3.5
C	14.3	0.0	71.4	14.3
NA	20.0	4.0	16.0	60.0

are Berrenda en Colorado and Cárdena Andaluza. This effect also can be observed between Cárdena Andaluza and Negra Andaluza although in a less pronounced form. This seems to corroborate the preliminary and not yet published results obtained with analysis of DNA molecular markers for the same breeds which show a very close genetic proximity between Berrenda en Colorado y Berrenda en Negro and a very marked genetic distance between Negra Andaluza and the rest of breeds.

Table 3 shows the result of the sensitivity analysis for each of the heuristic methods. As can be observed the two methods with the best average rate of classification (MLP and SVM), weight most variables in similar way. In particular, for both these models, the variable given the greatest weight is ChD, followed by WP and HR with the MLP, and HR and WP with the SVM method. Moreover, the same trend was observed with the PNN method. The effects of the HR–WP and HR–HW ratios are clearly non-significant, probably due to the fact that the underlying information is already contained in others of the variables used. These variables seems to be associated with greater variability, at least with respect to the two most heavily weighted variables, as can be seen in Table 4 (Rodero et al., 2008). On the other hand, HR (ranked third in MLP; and second in SVM) shows lower variability, which may indicate the need for further analysis of the relationship between co-variables, as well as of the input variables used in the models.

Table 3
Values of (1-Ratio) expressed in percentage. Each value indicates the increasing of error when the variable is removed. Ranking in brackets.

Descriptor	MLP	PNN	SVM
Height at withers (HW)	6.00 (4)	6.51 (4)	5.98 (4)
Height at rump (HR)	7.57 (3)	8.57 (2)	7.60 (2)
Chest depth (ChD)	15.58 (1)	11.58 (1)	17.2 (1)
Width at hips (WH)	3.31 (6)	3.65 (6)	2.99 (7)
Width at pins (WP)	7.70 (2)	6.23 (5)	7.51 (3)
Rump length (LG)	3.24 (7)	3.32 (7)	3.01 (6)
HR-HW	2.38 (8)	2.00(8)	2.58 (8)
WH-WP	5.60 (5)	6.70 (3)	5.70 (5)

Table 4
Variation coefficients of each morphological trait analysed. Berrenda en Colorado (BC), Berrenda en Negro (BN), Cárdena Andaluza (C) and Negra Andaluza (NA).

Descriptor	BC	BN	C	NA
Height at withers (HW)	3.78	4.92	2.88	5.57
Height at rump (HR)	4.14	4.90	2.64	5.25
Chest depth (ChD)	15.31	13.82	7.30	20.57
Width at hips (WH)	8.88	11.02	7.37	6.38
Width at pins (WP)	24.55	26.89	26.07	14.58
Rump length (RL)	8.94	7.89	7.13	6.00

4. Conclusion

There is no universal method for characterising and classifying breeds using morphological characteristics. In this paper, we present the results of a study attempting to classify four cattle breeds autochthonous to Andalusia using a classical statistical method and three heuristic methods. Our results demonstrate that when analysing data from autochthonous breeds, not subject to generations of selective pressure imposed by humans aiming to improve production, classification using linear methods is not possible or at least not efficient, probably due to non-linear relationships between the independent and dependent variables. Therefore, in these cases, we have to employ more powerful classification procedures (such as heuristic models) that allow for non-linear relationships in the modelling process. Among the heuristic models used, we particularly recommend the MLP and SVM methods since they have shown to be effective during the generalisation of the model.

On the other hand, although sensitivity analysis identified a similar pattern in the models with greater discrimination ability, further research is needed to deepen our understanding of the relationships between the co-variables as well as the causal relationships.

Finally, the results shown in this study indicate that the heuristic models, particularly multilayer perceptron neural networks and support vector machines, could be a promising tool for phenotypic characterisation, mainly for those countries needing to make comparisons between different breeds and who do not have financing channels and technologies such as molecular markers (FAO, 1992, 2010a, 2010b).

Acknowledgements

This project has been possible thanks to the funding for INIA projects from the Spanish Ministry of Agriculture (RZ00-017 and RZ2004-00013-00-00). We also want to thank the Federation of Associations of 'Raza Berrenda en Negro' and 'Raza Berrenda en Colorado' Cattle Breeders (ANABE) and also the Association for 'Raza Cárdena Andaluza' Cattle Breeders for their support.

References

Adamowski, J., Karapataki, C., 2010. Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ANN learning algorithms. J. Hydrol. Eng. 15 (10), 729–744.

Alderson, G.L.H., 1999. The development of a system of linear measurement to provide an assessment of type and function of beef cattle. Anim. Res. Inf. 25, 45–56.

Aparicio, M.J.B., Del Castillo, J., Herrera, M., 1986. Características estructurales del caballo español tipo Andaluz. Consejo Superior de Investigaciones Científicas, Madrid, Spain. (in Spanish).

Aparicio-Sánchez, G., 1960. Zootecnia especial. Etnología Comprendida, 4ª. Ed. Imprenta moderna, Córdoba, Spain. (in Spanish).

Azor, P.J., Rodero, E., Luque, M., González, A., Valera, M., Rodero, A., Molina, A., 2006. Algunas consideraciones sobre el proceso de formación de las razas bovinas andaluzas a partir de datos moleculares. V Encuentro de Científicos y Docentes Zooetnólogos Españoles, Sociedad Española de Zooetnología, Córdoba, Spain. (in Spanish).

Czerwinski, I.A., Gutiérrez-Estrada, J.C., Hernando-Casal, J.A., 2007. Short-term forecasting of halibut CPUE: linear and non-linear univariate approaches. Fish. Res. 86, 120–128.

Dossa, L.H., Wollny, C., Gauly, M., 2007. Spatial variation in goat populations from Benin as revealed by multivariate analysis of morphological traits. Small Rum. Res. 73 (1–3), 150–159.

FAO, 1992. The Management of Global Animal Genetic Resources. FAO Animal Production and Health Paper No 104 FAO, Rome.

FAO, 2010a. Draft Guidelines for Molecular Characterization of Animal Genetic Resources for Food Agriculture. FAO, Rome.

FAO, 2010b. Draft Guidelines of Phenotypic Characterization. FAO, Rome.

Fernando, T.M.K.G., Maier, H.R., Dandy, G.C., 2009. Selection of input variables for data driven models: an average shifted histogram partial mutual information estimator approach. J. Hydrol. 367, 165–176.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquat. Ecol. 41, 491–508.

Gutiérrez-Estrada, J.C., Bilton, D.T., 2010. A heuristic approach to predicting water beetle diversity in temporary and fluctuating waters. Ecol. Model. 221, 1451–1462.

Gutiérrez-Estrada, J.C., Pulido-Calvo, I., Prenda, J., 2000. Gonadosomatic index estimates of an introduced pumpkinseed (*Lepomis gibbosus*) population in a Mediterranean stream, using computational neural networks. Aquat. Sci. 62, 350–363.

Gutiérrez-Estrada, J.C., Silva, C., Yáñez, E., Rodríguez, N., Pulido-Calvo, I., 2007. Monthly catch forecasting of anchovy *Engraulis ringens* in the north area of Chile: non-linear univariate approach. Fish. Res. 86, 188–200.

Gutiérrez-Estrada, J.C., Vasconcelos, R., Costa, M.J., 2008. Estimating fish community diversity from environmental features in the Tagus estuary (Portugal): multiple linear regression and artificial neural network approaches. J. Appl. Ichthyol. 24, 150–162.

Gutiérrez-Estrada, J.C., Yáñez, E., Pulido-Calvo, I., Silva, C., Plaza, F., Bórquez, B., 2009. Pacific sardine (*Sardinops sagax*, Jenyns 1842) landings prediction. A neural network ecosystemic approach. Fish. Res. 100, 116–125.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1999. Multivariate Data Analysis, 5ª Ed. Prentice Hall International, Inc., Upper Saddle Rives, New Jersey, USA.

Haralabous, J., Georgakarakos, S., 1996. Artificial neural networks as a tool for species identification of fish schools. ICES J. Mar. Sci. 53, 173–180.

Herrera, M., 2007. Metodología de caracterización zooetnológica. La ganadería andaluza en el siglo XXI, patrimonio ganadero andaluz I, pp. 435–448 (in Spanish).

Herrera, M., Rodero, E., Gutiérrez, M.J., Peña, F., Rodero, J.M., 1996. Application of multifactorial discriminant analysis in the morphostructural differentiation of Andalusian caprine breeds. Small Rum. Res. 22, 39–47.

Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Netw. 13, 415–425.

Hunter, A., Kennedy, L., Henry, J., Ferguson, I., 2000. Application of neural networks and sensibility analysis to improved prediction of trauma survival. Comput. Meth. Programs Biomed. 62, 11–19.

Iyer, M., Rhinehart, R., 1999. A method to determine the required number of neural network training repetitions. IEEE Trans. Neural Netw. 10 (2), 427–432.

Lachenbruch, P.A., 1975. Discriminant Analysis. Hafner Press, New York, USA.

Lanari, M.R., Taddeo, H., Domingo, E., Pérez-Centeno, M., Gallo, L., 2003. Phenotypic differentiation of exterior traits in local Criollo goat population in Patagonia (Argentina). Arch. Tierzucht 46 (4), 347–356.

Lek, S., Guegan, J.F., 1999. Artificial neural networks as a tool in ecological modeling. An introduction. Ecol. Model. 120, 65–73.

Lek, S., Delacoste, M., Baran, P., Lauga, J., Aulagnier, S., 1996. Application of neural network for modeling in ecology. Ecol. Model. 90, 39–52.

Luque, M., Rodero, E., Peña, F., García, A., Sierra, I., Herrera, M., 2005. Application of discriminant analysis to the morphostructural differentiation of 7 extensive goat breeds. Annual Meeting of European Federation of Animal Science, Uppsala, Sweden.

Macciotta, N.P.P., Cappio-Borlino, A., Steri, R., Pulina, G., Brandano, P., 2002. Somatic variability of Sarda goat breed analysed by multivariate methods. Livest. Prod. Sci. 75 (1), 51–58.

Parés i Casanova, P.M., 2009. Zoometría. Valoración morfológica de los animales domésticos. Minist. Medio Ambiente Medio Rural Mar. 6, 171–196 (in Spanish).

Pastor-Fernández, J.M., 2010. Plan de conservación de las razas autóctonas andaluzas en peligro de extinción: análisis poblacional, caracterización demográfica y de sus condicionantes sanitarios. Tesis Doctoral, Universidad de Córdoba, Spain (in Spanish).

Pulido-Calvo, I., Gutiérrez-Estrada, J.C., 2009. Improved irrigation water demand forecasting using a soft-computing hybrid model. Biosyst. Eng. 102, 202–218.

Pulido-Calvo, I., Portela, M.M., 2007. Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds. J. Hydrol. 332 (1–2), 1–15.

Pulido-Calvo, I., Roldán, J., López-Luque, R., Gutiérrez-Estrada, J.C., 2003. Demand forecasting for irrigation water distribution systems. J. Irrig. Drain. Eng. 129 (6), 422–431.

Pulido-Calvo, I., Montesinos, P., Roldán, J., Ruiz-Navarro, F., 2007. Linear regressions and neural approaches to water demand forecasting in irrigation districts with telemetry systems. Biosyst. Eng. 97, 283–293.

Robotham, H., Bosch, P., Gutiérrez-Estrada, J.C., Castillo, J., Pulido-Calvo, I., 2010. Acoustic identification of small pelagic fish species in Chile using support vector machines and neural networks. Fish. Res. 102, 115–122.

Rodero, E., Delgado-Bermejo, J.V., Rodero, A., Camacho-Vallejo, M.E., 1994. Conservación de razas autóctonas andaluzas en peligro de extinción. Consejería de Agricultura y Pesca, Junta de Andalucía, Sevilla, Spain. (in Spanish).

Rodero, E., Herrera, M., Peña, F., Molina, A., Valera, M., Sepúlveda, N., 2003. Modelo morfoestructural de los caprinos lecheros Españoles Florida y Payoya en sistemas extensivos. Rev. Cient. FCV-LUZ 13 (5), 403–412 (in Spanish).

Rodero, E., González, A., Luque, A., 2008. Las razas bovinas andaluzas de protección especial: Berrenda en Negro, Berrenda en Colorado, Cárdena Andaluza, Negra Andaluza de las Campiñas, Pajuna y Marismeña. La ganadería andaluza en el siglo XXI, patrimonio ganadero andaluz II, pp. 53–120 (in Spanish).

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. "Learning" representations by backpropagation errors. Nature 323 (9), 533–536.

Rutkowski, L., 2004. Adaptive probabilistic neural networks for pattern classification in time-varying environment. IEEE Trans. Neural Netw. 15, 811–827.

Scholkopf, B., Smola, A., 2002. Learning with Kernels. The MIT Press, Cambridge, MA/London, UK.

Sierra, I., 2009. Importancia de la morfología y su valoración en los animales domésticos. Valoración morfológica de los animales domésticos. Minist. Medio Ambiente Medio Rural Mar. 1, 23–46 (in Spanish).

Specht, D.F., 1990. Probabilistic neural networks. Neural Netw. 3, 109–118.

Tsoukalas, L.H., Uhrig, R.E., 1997. Fuzzy and Neural Approaches in Engineering. Wiley Interscience, New York, USA.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York, USA.

Watts, M.J., Worner, S.P., 2008. Comparing ensemble and cascaded neural networks that combine biotic and abiotic variables to predict insect species distribution. Ecol. Model. 3, 354–366.

Zaitoun, I.S., Tabbaa, M.J., Bdour, S., 2005. Differentiation of native goat breeds of Jordan on the basis of morphostructural characteristics. Small Ruminant Res. 56 (1–3), 173–182.

Please cite this article as: Rodero, E., et al., Classification of Spanish autochthonous bovine breeds. Morphometric study using classical and heuristic techniques, Livestock Science (2011), doi:10.1016/j.livsci.2011.09.022