Contents lists available at ScienceDirect

Livestock Science

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

Classification of goat genetic resources using morphological traits. Comparison of machine learning techniques with linear discriminant analysis

E. Rodero [a,*], A. González [a], M. Dorado-Moreno [b], M. Luque [c], C. Hervás [b]

[a] Animal Production Department, Campus de Rabanales, Universidad de Córdoba, Campus de Excelencia Internacional Agroalimentario ceiA3, 14071 Córdoba, Spain
[b] Computer Science and Numerical Analysis Department, Campus de Rabanales, Universidad de Córdoba, 14071 Córdoba, Spain
[c] Spanish Federation of Livestock Purebred Associations (FEAGAS), Castelló, 45, Madrid 28001, Spain

ABSTRACT

The aim of this study was to provide effective solutions for the nominal classification of twelve Spanish goat breeds using nine morphological traits and considering their aptitude (meat, dual purpose and milk). Different statistical and artificial intelligence algorithms were used to compare our hierarchical methodology with a representative of each Machine-Learning typology and several common statistical methods. The most appropriate tool to solve problems of classification, by considering the aptitude, would be the k-Nearest Neighbours used in a hierarchical model. For the first level of this hierarchy, the study was conducted using a 1-Nearest Neighbour classification of individuals by aptitude, and on the second level, the breeds were analysed again using three new 1-Nearest Neighbour classifiers, one for each aptitude. The results obtained improved the accuracy rates for assigning individuals to breed, compared with those usually employed using Linear Discriminant Analysis methodologies. Only 78% correct classification rate (Minimum Sensitivity=19%) was obtain with the Linear Discriminant Analysis, but this result increased to 89.18% with a 1-Nearest Neighbour+1-Nearest Neighbour (1NN+1NN). Hierarchical methodology, thus increasing the classification rate (Minimum Sensitivity=37.08%). Furthermore, the percentage of correct classification was 83.48% (Minimum Sensitivity=35.08%) with 1-Nearest Neighbour+Multi-Layer Perceptron, that justifies the use of hierarchical models. The new second level model (1NN+1NN) permit 100% of goats successful classified in *Pirenaica*, *Retinta* and *Malagueña* breeds, each of these belongs to a different aptitude. The improve of classification obtained of the majority of the breeds with the application of the hierarchical method, suggested that defining firstly the aptitude class, the unique and distinctive characteristics of the breed are identified more clearly.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Spain has 2,637,336 goats (MAGRAMA, 2013) that makes it the second-ranked EU country in the number of animals and in goat meat production (8830 t/year), but is the third-ranked (FAOSTAT, http://faostat.fao.org) in goat milk production (490 million tons/year). This difference is due to the variability in the farming systems and breeds that exist in Spain; most breeds are reared in the traditional extensive farming system.

The Breed is the unit or "primary focus" to classify the diversity in Domestic Animal Genetics Resources (AnGR), and the characterization and inventory of AnGR are fundamental for genetic improvement, or conservation breeds programmes and for the design of programmes to protect valuable resources at risk (FAO, 2007). Spanish goat breeds are, therefore, inventoried in the AnGR global and Spanish breed data banks (DAD-IS and ARCA, respectively). Spain has twenty-one autochthonous goat breeds, fifteen of which are classified as endangered and need to be preserved (Formal List of Spanish Breeds, RD 2129/2008). But inventory and characterization are two complementary processes (Tixier-Boichard et al., 2008). To measure domestic animal diversity, the FAO proposes the use of molecular and phenotypic procedures (FAO, 2011, 2012). Phenotypic characterization is understood as the process for identifying different breed populations and describing their external and production characteristics in a given environment and under given management, taking into account the social and economic factors that affect them. The breeds possess adaptive characteristics that allow them to produce a remarkable array of production systems, but many of them have become

endangered by competition with highly selected "global" breeds that are extremely productive but presented a lack of adaptational characteristics critical in extensive and subsistence production (Notter et al., 2013).

The list of research priorities for animal genetics resources proposed by Cardellino and Boyazoglu (2009) include the following thematic areas: Information Systems, Characterization, Genetic diversity, Functional index and Animal Breeding. These authors consider that the assessments of genetic diversity have to be done using molecular markers, but the characterization must be done by phenotypic methods. Accordingly Ollivier and Foulley (2005), Ruane and Sonnino (2006) and Tixier-Boichard et al. (2008) considered that when characterising a breed, genetic markers are the first option to analyse genetic diversity but should be integrated with more detailed phenotypic information. The best technique is to combine the results of both groups of variables (molecular and phenotypic) with appropriate statistical analysis (FAO, 2011). The distances between breeds, obtained using both types of variables can be different (Brito, 2002). The different environmental factors in the selection pressure can be responsible for these differences (Crepaldi et al., 2001).

Despite the fact that, in general, no breed classification should rely exclusively on biometric data, it is clear that these play a complementary role on the description of a breed. Hence, it is important to analyse accurately the morphological variables that enable us to distinguish breeds as well as explore the use of various discrimination methods to achieve this. When the aim is to identify and clarify poorly known populations, the most common methods are sequential, agglomerative, hierarchical and non-overlapping (SAHN). However, to compare and differentiate between known breeds, the discriminate analysis is used to validate a previous classification (FAO, 2012). Jordana et al. (1993), Herrera et al. (1996), Capote et al. (1998) and Rodero et al. (2003, 2012) have reported a comparison between autochthonous Spanish goat breeds according to morphological characteristics. Differences in morphological traits are still widely used to classify AnGR, especially in countries that have not yet clearly identified the biodiversity of local livestock populations (Crepaldi et al., 2001; Macciotta et al., 2002; Lanari et al., 2003; Nsoso et al., 2004; Zaitoun et al., 2005; Traoré et al., 2006; Dossa et al., 2007; Vargas et al., 2007; Traoré et al., 2008). In all cases, the data were analysed using Linear Discriminant Analysis (LDA). The method was found to be relatively efficient and could establish differences between all breeds and subpopulations Additionally, Zaitoun et al. (2005) and González et al. (2014), in previous studies using morphostructural variables, found that native goat breeds tended to cluster in function of purpose and farming system. Therefore, it is necessary to explore new classification methods that may be more accurate for distinguishing between breeds by taking into account the effects of aptitude in the process of classification. In a way that can assign individuals to their corresponding breed.

Recently, numerous algorithms for hierarchical classification have been developed (Dekel et al., 2004; Astikainen et al., 2008; Seeger, 2008). Machine-Learning techniques are used (Bishop, 2006) because the results of nonlinear classifiers k-Nearest Neighbour (k-NN) and Artificial Neural Network (ANNs) are more accurate than those of standard linear classifiers (LDA) and do not need the condition of linearity of variables. Machine-learning or heuristics methods such as Artificial Neural Networks (ANNs) or Support Vector Machine (SVM) are useful for taxonomists as a tool to recognize patterns or discriminate between groups of species (Haralabous and Georgakarakos, 1996; Gutiérrez-Estrada et al., 2007). But their application for classification of domestic AnGR biodiversity has been scarcely investigated. Rodero et al. (2012) found in Andalusian cattle breeds that the results obtained by heuristic classification techniques improved from those obtained using LDA. The same circumstance was observed by Bhatia et al. (2010) using computer learning techniques to study the possibility to identify AnGR. These authors applied scoring function and the PNC2 algorithm (positive and negative exemplar-based clustering) to assign animals to breeds based of phenotypic qualitative and quantitative traits.

Therefore, the main objective of this study was to test different classification techniques to arrive at the right techniques to provide the most effective solutions for the nominal classification of Spanish goat considering the influence of aptitude (meat, milk and dual purpose) on the body structure of the animals. The discrimination study was carried out using morphological descriptor of twelve autochthonous goat breeds.

2. Materials and methods

2.1. Data collection and morphometric variables

The assessment of the morphological models was carried out in a population of 2406 females from 392 flock of twelve breeds of Spanish goats: *Malagueña* ($n=142$), *Murciano-Granadina* ($n=102$), *Florida* ($n=63$), *Payoya* ($n=578$), *Pirenaica* ($n=152$), *Azpi Gorri* ($n=84$), *Moncaina* ($n=111$), *Blanca Andaluza* ($n=274$), *Blanca Celtiberica* ($n=398$), *Negra Serrana* ($n=266$), *Retinta* ($n=79$) and *Agrupacion de las Mesetas* ($n=157$). The goats were between 2 and 5 years old (according to teeth). The choice of flock for inclusion in the study was based on breed purity, localization (wide range of

Table 1
Characteristics of farming system and aptitude to group twelve autochthonous Spanish goat breeds.

Breeding systems[a]	Stocking (number of animal/ha)	Supplied concentrate (kg goat^{-1} yr^{-1})	Principal suckling system of kids	Breeds[b]	Aptitude[c]	Milk production levels (liter/lactation)
Intensive (feedlot)	< 1.5	> 150	Artificial	Ma MG F	Milk	350–600
Mixed	1.5–6	75–150	Natural	**P Pi AG Mo**	Hybrid or dual purpose	100–450
Extensive	> 6	< 75	Natural	**BA BC NS R AM**	Meat	> 100

Malagueña (Ma), *Murciano-Granadina* (MG), *Florida* (F), *Payoya* (P), *Pirenaica* (Pi), *Azpi Gorri* (AG), *Moncaína* (Mo), *Blanca Andaluza* (BA), *Blanca Celtiberica* (BC), *Negra Serrana* (NS), *Retinta* (R) and *Agrupación de las Mesetas* (AM).

[a] Breeding systems are defined by the space available for grazing, the amount of concentrate that will be supplemented and the suckling system of kids. Source: Vert and García Trujillo (2006).

[b] Breeds in danger of extinction have been highlighted. Source: RD 2129/2008.

[c] Aptitude breed classifications are in function of milk production levels. Source: Sañudo (2011).