Use of linear discriminant analysis to characterise three dairy cattle breeds on the basis of several milk characteristics

Roberto Leotta

Dipartimento di Produzioni Animali. Università di Pisa, Italy

Corresponding author: Prof. Roberto Leotta. Dipartimento di Produzioni Animali. Università di Pisa. Viale delle Piagge 2, 56124 Pisa, Italy - Tel. +39 349 1809326 - Fax: +39 050 3139401 - Email: rleot@vet.unipi.it

Paper received November 21, 2003; accepted July 14, 2004

ABSTRACT

To characterise individuals of differents breeds on the basis of milk composition and to identify the best set of variables a linear discriminant analysis (LDA), on 14 milk production traits, was performed on milk samples from 199 cows of different breeds (respectively, 127 subjects were Italian Friesians (IF), 62 were German Friesians (GF), and 10 were Jerseys (J) and all came from the same breeding farm in Tuscany. The variables were: test day milk yield (kg milk), % Fat, % Protein,% Lactose, % solid non fat (SNF), % total solid (TS), pH and titratable acidity (TA); five rheological variables: r, k_{20}, a_{30}, a_{45}, and somatic cell counts /ml (SCC); and one hygiene-related variable: total bacterial count (TBC). The analysis performed on the 14 variables, with regard to the three breeds, allowed us to identify 10 of these as variables useful for discrimination (leaving out kg milk, pH, a45, and TBC). The most important variables were the percentage of Fat and TS for the first canonical variate and SNF, Lactose and Protein for the second. Fat and TS play an important role since they present significant values (even if opposite sign) in the two variates. The resulting classification of subjects was satisfactory: 79% of the Italian Friesians, 73% of German Friesians and 100% of the Jersey cows were classified correctly.

Key Words: Cattle, Milk quality, Breeds, Discriminant linear analysis.

RIASSUNTO

USO DELL'ANALISI DISCRIMINANTE LINEARE PER LA CARATTERIZZAZIONE DI TRE DIVERSE RAZZE BOVINE SULLA BASE DELLA QUALITÀ DEL LATTE

L'analisi discriminante lineare fu condotta su 14 variabili quanti-qualitative di campioni di latte provenienti da 199 bovine di diversa razza (rispettivamente, 127 soggetti di razza Frisona italiana (IF), 62 di Frisona tedesca (GF), e 10 Jersey (J) di un allevamento della Toscana. Le variabili erano: produzione di latte/mungitura (kg), % Grasso, % Proteine, % Lattosio, % SNF, % TS, pH e acidità titolabile (TA); 5 variabili reologiche: r, k_{20}, a_{30}, a_{45} e n. cellule somatiche/ml (SSC); e una variabile igienica: carica microbica totale (TBC). L'analisi condotta rispetto alle tre razze, sulle 14 variabili considerate, ha consentito di identificare 10 di queste come variabili utili alla discriminazione, scartando kg milk, ph, a45 e carica. Le variabili che risultano più importanti sono le percentuali di Grasso e TS per la prima variata canonica e per la seconda SNF, Lattosio e Proteine. Grasso e TS giocano un ruolo importante poiché assumono valori significativi (anche se di segno opposto) nelle due variate. La classificazione dei soggetti che ne deriva è soddisfacente: risultano classificati correttamente il 79% delle Frisone Italiane, il 73% delle Frisone Tedesche ed il 100% delle bovine di razza Jersey.

Parole chiave: *Bovini, Qualità del latte, Razze, Analisi discriminante lineare.*

Introduction

The linear discriminant analysis has long been known (Fisher, 1936) and can be used not only to examine multivariate differences between groups, but also to determine:

- which variables are the most useful for discriminating between groups,
- whether one subclass of variables works as well as another,
- which groups are similar and which are different.

Recently discriminant analysis has been used to distinguish the milk and cheese of various species (Fresno *et al.*, 1995); (Herrero-Martinez *et al.*, 2000), (Martin-Hernandez *et al.*, 1992), (Rodriguez *et al.*, 1999). In cows it has been used to attempt to identify preventively those subjects which were about to give birth, according to milk composition (Harwood *et al.*, 1991); it has also been used to distinguish between two different diets (Favretto *et al.*, 1994), to distinguish the different physiological conditions of the animals and the different season based on the metabolic profile (Biagi *et al.*, 1990; Biagi *et al.*, 1991).

In previous studies concerning the same animals used in this study (Cecchi and Leotta, 2002, Cecchi *et al.*, 2002a, Cecchi *et al.*, 2002b), differences between breeds were brought to light, especially regarding the relationships between the chemical and technological parameters of cow's milk, while in other studies the sources of environmental and genetic variability were analyzed only in milk from Italian Friesians (Leotta *et al.*, 2003).

The aim of this study is to find the linear combination of characteristics of milk production that best differentiates between the three breeds examined. In fact, we know that strong correlations exist between the variables which are potential candidates to serve as predictors for estimating the linear discriminant function, and we are interested in learning which of these subsets would be the most useful.

Material and methods

Animals.

A trial was carried out on 199 cows of different breeds (127 Italian Friesian, 62 German Friesian and 10 Jersey); animals were farmed in a herd located in the province of Pisa, and they were all fed the same diet. Milk samples for quantitative/qualitative analysis were taken over a period of 1 year; only one sample from each animal was taken from the morning milking and yield production (*kg milk*) was recorded. Sampling was performed on data related to samples collected on animals of various conditions (parity 2.8 ± 0.17, parturition distance in months 5.2 ± 0.29 and age at parturition in months 45.5 ± 2.23) in order to evaluate the response of the LDA to raw data to allow generalization.

Chemical analysis

Milk samples were analyzed for Fat, Protein and Lactose content by infrared analysis (Milkoscan, Foss Electric, Italy), somatic cell count (SCC) (Fossomatic 250), total bacterial count (TBC), titratable acidity (TA) by Soxhlet-Henkel and pH. Rheological parameters, rennet clotting time (r), rate of firming (k_{20}) and curd firmness after 30 (a_{30}) and 45 minutes (a_{45}) were also measured (Formagraph apparatus, Foss Electric), (ASPA, 1995).

Statistical analysis

The data underwent screening, and to meet the assumptions of normal distribution of the classifying variables (Fisher, 1936) the following transformations were applied:

- r → inverse (1/r);
- TBC → logarithmic (Log10);
- SCC → logarithmic(Log10);
- k20 → inverse (1/k20).

Linear Discriminant Analysis (LDA) provides a linear function of the variables that 'best' separate cases (individuals) into two or more predefined groups. LDA require that one know the groups share a common covariance matrix whose values are used to calculate distances between cases we want to classify and the center of each group in a multidimensional space. The closer a

case is to the center of one group (relative to its distance to other groups), the more likely it is to be classified as belonging to that group.

The variables in the linear function can be selected in a forward or a backward stepwise manner. In the forward method, begins with no variables in the model. At each step, the variable with the F greater than the specified value (F-to-enter limit) is added to the model (if tolerance permits).

The process go on since the significance (on the basis of R^2-adjusted) of the model increases.

In the backward method, all the candidate variables are first forced in the model. At each step the variable with the F less than the specified value (F-to-remove limit) is removed from the model.

Here LDA was applied with the method of backward stepping automatic elimination of the variables, with the value of F-to-remove=3.9 and F-to-enter=4.0 and with tolerance limit value for the matrix inversion (T=0.0001). As a measure of distance between individuals and the centroids of single groups the statistic D2 of Mahalanobis (Systat® 9, 1999) was used, calculated on the variance-covariance matrix.

The tolerance index measures the correlation of a candidate variable with the variables included in the model, and its values range from 0 to 1. If a variable is highly correlated with one or more of the others, the value of tolerance is very small and the resulting estimates of the discriminant function coefficients may be unstables.

The Jackknifed Classification Matrix is an attempt to approximate (nonparametrically) cross-validation. Tukey (1958) proposed computing n subsets of (x_1…,x_n), each consisting of all the cases except the ith deleted case (for i = 1, …, n). He produced standard errors as a function of the n estimates from these subsets.

Results and discussion

Table 1 shows results for the 14 variables considered in the three breeds.

The differences revealed by the between group F-matrix on the full data set (14 variables), measuring distances between centroids relative to the three breeds calculated by the D^2 statistics of Mahalanobis for all 14 variables, were highly significant (P<0.00005). These indicate that the multivariate distance between centroids for IF and J is low (F=7.165) (more similar breeds), while the higher value (F=7.201) (less similar

Table 1. Statistics for the 14 variables in the three breeds (mean ± SE).

Variables		Breed					
		IF		GF		J	
Milk	kg	12.9	± 0.46	14.6	± 0.65	10.0	± 0.96
pH		6.69	± 0.012	6.67	± 0.017	6.68	± 0.035
TA	°SH	3.19	± 0.035	3.45	± 0.040	3.76	± 0.110
r	min	27.0	± 0.87	30.0	± 1.30	17.9	± 1.32
k20	"	37.9	± 3.60	34.1	± 4.85	12.9	± 9.68
a30	mm	12.2	± 1.14	10.1	± 1.64	26.8	± 3.05
a45	"	19.7	± 1.17	20.0	± 1.88	29.2	± 4.61
TBC	n./ml	16,900	± 2,360	19,400	± 2,630	13,900	± 3,220
SCC	"	1,050,000	± 161,000	890,000	± 164,000	350,000	± 162,000
TS	%	12.28	± 0.096	12.58	± 0.154	14.00	± 0.351
SNF	"	8.94	± 0.047	9.27	± 0.057	9.53	± 0.096
Lactose	"	4.79	± 0.023	4.91	± 0.023	4.88	± 0.045
Protein	"	3.32	± 0.036	3.45	± 0.048	3.84	± 0.109
Fat	"	3.37	± 0.068	3.34	± 0.120	4.53	± 0.268

IF: Italian Friesian; GF: German Friesian; J: Jersey.

breeds) was relative to the distance between IF and GF and that between GF and J (F= 7.91) was intermediate.

The differences between the three breeds, tested using the lambda statistics of Wilks, was highly significant (P<0.00005)

The variables with the lowest values of F-to-remove, and therefore less useful for the discrimination, were the pH, TBC, a45 and kg milk (respectively, 0.26; 0.48, 0.99, 1.87).

Greater initial differences between the breeds were found respectively for the following variables: Protein, Lactose, SNF and TA (F≥ 9.99), then Fat, k20, TS and SCC (values ranging from 7.34 ≤ F<5.80).

The very low tolerance values (T) indicate the possibility of redundance, high correlation, or the possibility of linear combination of other variables and in this study, were found respectively to be for SNF, TS, Protein, Fat, Lactose and a30 (T<0.07). This was not surprising because these variables (except a30) are, by definition, quasi-linear combinations.

The variables removed with the application of the procedure of discriminant analysis were in the following order: pH, TBC, a45, and kg milk.

Likewise, the exclusion of a45 has rendered a30 more useful as a discriminant variable, since the two are highly correlated (r=0.86) and the Tolerance value of a30 is rather low (T=0.076).

After discarding the less useful variables, the comparisons performed with the F- test on the D^2 values of Mahalanobis proved to be highly significant (P< 0.00005). Unlike the situation before the elimination of the variables, and as expected, the multivariate distance between centroids for IF and GF is lower (F= 9.535), (more similar breeds), while the greater value (less similar breeds) was

Table 2. Classification matrix (subjects in the rows categories, classified in columns).

		In genetic type			
		IF	GF	J	Total
From genetic type	IF	100	22	5	127
		79%	17%	4%	
	GF	15	45	2	62
		24%	73%	3%	
	J	0	0	10	10
		0%	0%	100%	
	Total	115	67	17	199
	Correct Total				78%

'Jackknifed' Classification Matrix

		In genetic type			
		IF	GF	J	Total
From genetic type	IF	94	27	6	127
		74%	21%	5%	
	GF	16	42	4	62
		26%	68%	6%	
	J	1	2	7	10
		10%	20%	70%	100%
	Total	111	71	17	199
	Correct Total				72%

relative to the distance between IF and J (F= 9.873); this speaks favourably for the usefulness of the analysis in interpreting the relationships and the relative importance of the variables.

The percentage of individuals classified correctly in the classification matrix (78%) shown in Table 2, is not better than that classified by the "jackknifed" classification matrix (72%). Since with the "jackknifed" matrix breed classification is performed by preventive elimination from the classification procedure, it can be confirmed not only that no redundance exists between the 10 variables identified in the model and therefore these all compete usefully for the classification, but that the identified discriminant function is stable for IF and GF, while a larger set of data would be necessary for J.

The first eigenvalue is not very far from the second (respectively, 0.528 and 0.510) and this indicated that the first canonical variate alone does not manage to capture the greater part of the differences between the groups. This comprises 50.9% of the total dispersion, while the second is about 49.1%. Both are necessary and account for nearly all the total variation (approximately 100%).

The first canonical variable is the linear combination of the variable tha best discriminate among the groups.

The canonical correlation between the first canonical variate and the two dummy variables (the number of the dummy variables is the number of groups minus 1) representing the groups is 0.5888.; a value that is not much different from that between the second variate and the same dummy variables (0.581), a further confirmation of the previous observation.

The multivariate tests for the equality of groups mean for the 10 variables in the discriminant functions were analysed with the lambda statistic of Wilks, the trace of Pillai, and the trace of Lawley-Hotelling, and all were very significant ($P<0.00005$).

The discriminant equation (calculated on standardized values and adjusted to the general mean equating to zero and with intra-group variances equal to 1), for the first canonical variate is:

0.1436*TA + 0.756*r + 0.980*k20 – 1.455*a30 – 0.465*_TBC – 4.222*TS + 1.475*SNF – 0.108*Lactose + 0.708*Protein + 3.695*Fat

For the second canonical variate we have:

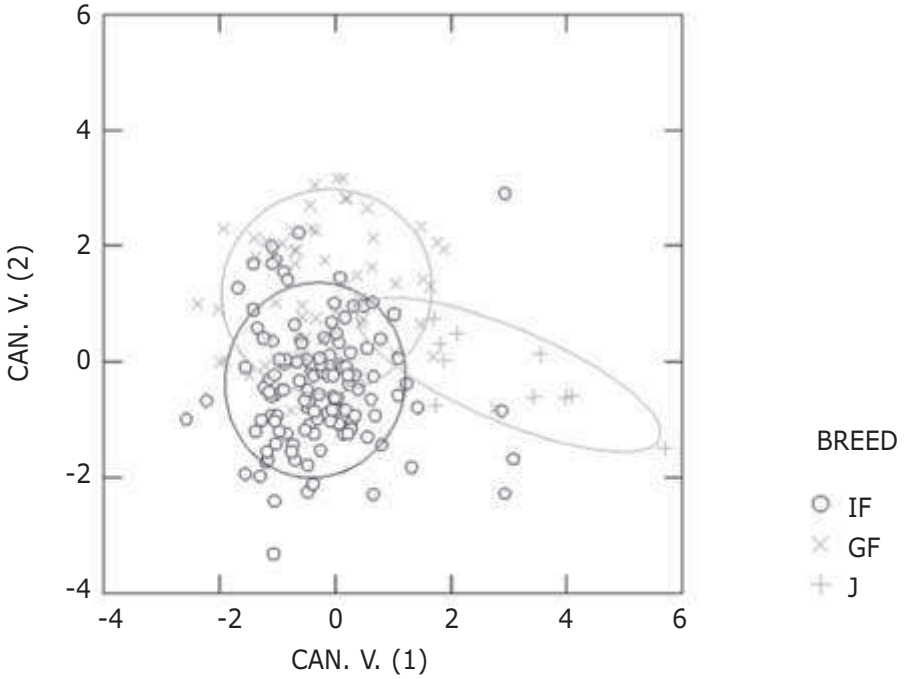
0.484*TA - 0.412*r – 0.210*k20 – 0.141*a30 – 0.272*TBC + 1.600*TS + 5.783*SNF – 3.206*Lactose – 4.932*Protein – 1.040*Fat

The observation of their values indicates that the variables that have highest relative weights on the first canonical variate are Fat (positive values) and TS (negative values), which also present very low tolerance values (respectively, 0.0014 and 0.0082), and SNF and a30 (also these with low tolerance values), respectively 0.009 and 0.102.

For the second variate, SNF (with positive values) is found as a 'guide', as well as Protein and Lactose (with negative values). With a certain importance, although with values carrying the opposite signs to those of the first variate, are TS and Fat. It can be noted that some variables contribute preponderantly to the differentiation both in the first and second canonical variates, which indicates that the set of variables is not optimal, nor is the fact that several among them have strong correlations.

The first canonical variate (Fig. 1), sets the two Friesians (IF and GF) against the Jersey (J), while the second sets the German Friesians(GF) against the Italian (IF). This graph permits us to perceive more quickly the differences between the three breeds and shows how much closer the two breeds of Friesians are (and therefore more similar) compared to the Jerseys. At any rate it is obvious that the similarities (and dissimilarities) between the three groups are less evident than might be expected (with the Friesian breeds relatively more separate from the Jersey), and that can also be attributed to various factors, some more obvious such as the different origins of the two strains of Friesian and others, more complex, related to choices made by breeders of the original strains (the definition of breeding objective and selection criteria are the first and more important steps to be taken in genetic improvement and they can vary too much), and finally, but not of minor importance, to the different size of the samples.

Figure 1. Ellipsoids of discrimination between groups.



Conclusions

The discriminant analysis carried out with respect to the three breeds on the 14 variables considered allowed us to identify 10 of these as useful discriminant variables, discarding milk yield (kg milk), pH, total bacterial count (TBC), and a45. The classification of the subjects derived in this way was satisfactory: 79% of the Italian Friesians, 73% of the German Friesians and 100% of the Jersey cows were classified correctly. As expected, the classification of the two strains of Friesians was less accurate, due to their greater genetic similarity. The most important variables for the two canonical variates were, respectively, the percentage of Fat and TS for the first canonical variate and SNF, Lactose and Protein for the second. Fat and TS play an important role since they assume values of an opposite sign in the two variates. The genetic strains that were more markedly different are the Italian Friesian and the Jersey. The results form an interesting pattern of the relationships between several of the variables considered that claim further investigation.

Financial support by: Ministry of the University and the Scientific and Technologic Research (Italy), MURST 40%, 2000. Project "Analysis of the genetic variability of some dairy milk quality characteristics".

REFERENCES

A.S.P.A., 1995. Commissione metodologie di valutazione della produzione quanti-qualitativa del latte. Metodi di analisi del latte delle principali specie di interesse zootecnico. Università degli Studi di Perugia ed., Perugia, Italy.

BIAGI, G., VALENTINI, A., BAGLIACCA, M., CORAZZA, M., DEMI, S., SIGNORINI, G. C., GREPPI, G. F., ROMAGNOLI, A., 1990. Influenza del momento produttivo, dell'età e della stagione sul quadro lipidico nella capra Saanen. Ann. Fac. Med. Vet. Pisa, Italy, 43: 57-67.

BIAGI, G., VALENTINI, A., BAGLIACCA, M., GREPPI, G. F., SIGNORINI, G. C., NANNIPIERI, S., ROMAGNOLI, A., 1991. Il quadro proteico nella capra Saanen: influenza dello stato fisiologico, dell'età e della stagione. Proc. 1st Congr. FeMeSPRum, Alghero, Italy, 1: 331-335.

CECCHI, F., LEOTTA, R., 2002. Relazioni tra composizione chimica e parametri lattodinamografici

nel latte bovino di diversi tipi genetici. Ann. Fac. Med. Vet. Pisa, Italy, 55: 223-231.

CECCHI, F., LEOTTA, R., CIANCI, D., 2002a. Le fonti di variabilità della qualità chimica e tecnologica del latte bovino di diversi tipi genetici. Ann. Fac. Med. Vet. Pisa, Italy, 55: 233-254.

CECCHI, F., LEOTTA, R., SUMMER, A., 2002b. Effetti del tipo genetico sulle principali caratteristiche chimico-fisiche del latte e correlazioni con i parametri di coagulazione presamica. Sci. Tecn. Latt. Cas. 53: 427-437.

FAVRETTO, L., VOJNOVIC, D., CAMPISI, B., 1994. Chemometric studies on minor and trace elements in cow's milk. Anal. Chim. Act. 293: 295-300.

FISHER, R. A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics. 7: 179-188.

FRESNO, J. M., PRIETO B., URDIALES, R., SARMIENTO, R. M., CARBALLO, J., 1995. Mineral content of some Spanish cheese varieties. Differentiation by source of milk and by variety from their content of main and trace elements. J. Sci. Food. Agr. 69: 339-345.

HARWOOD, E. D., JENSEN, E. L., WIECKERT, D. A., CLAYTON, M., 1991. Milk yield variation concurrent with conception. J. Dairy Sci. 74: 2172-2179.

HERRERO-MARTINEZ, J. M., SIMO-ALFONSO, E. F., RAMIS-RAMOS, G., GELFI, C., RIGHETTI, P., MARTINEZ, J.E.A., RAMOS, G., 2000. Determination of cow's milk and ripening time in nonbovine cheese by capillary electrophoresis of the ethanol-water protein fraction. Electrophoresis. 21: 633-640.

LEOTTA, R., CECCHI, F., SUMMER, A., 2003. Heritability of milk coagulation parameters in Italian Friesian dairy cows. Page 85 (abstr. n. 392) in Proc. 54th Meet. EEAP, Roma, Italy.

MARTIN-HERNANDEZ, C., AMIGO, L., MARTIN-ALVAREZ, P., JUAREZ, M., 1992. Differentiation of milks and cheeses according to species based on the mineral content. Z. Lebensm. Unters. Forsch. 194: 541-544.

RODRIGUEZ, E. M. R., ALAEJOS, M. S., RODRIGUEZ, E. M. R. R., ALAEJOS, M. S., ROMERO, C. D., 1999. Chemometric studies of several minerals in milks. J. Agricult. Food. Chem. 47: 1520-1524.

TUKEY, J.W., 1958. Bias and confidence in not quite large samples. Ann. Math. Statistics. 29: 614- 619.

SYSTAT®, 1999. Statistic I, Version 9.01. SPSS Inc., 233 South Wacker Drive, 11th Chicago, IL, USA.