Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Trends of the genetic connectedness measures among Nelore beef cattle herds

N.T. Pegolo[1], D. Laloë[2], H.N. de Oliveira[3], R.B. Lôbo[1] & M.-N. Fouilloux[4]

1 Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, USP, Ribeirão Preto, SP, Brazil
2 INRA - Génétique animale et biologie intégrative, Jouy-en-Josas, France
3 Departamento de Zootecnia, Faculdade de Ciências Agrárias e Veterinárias, UNESP, Jaboticabal, SP, Brazil
4 Institut de l'Élevage, INRA - Génétique animale et biologie intégrative, Jouy-en-Josas, France

Keywords
Beef cattle; coefficient of determination; connectedness; genetic evaluation.

Correspondence
N.T. Pegolo, Rua Goiás, 880, Centro, Avaré, CEP 18700-140, SP, Brazil. Tel: 55 14 37328340, 55 14 37320574, 55 14 97070686; Fax: 55 16 32092678; E-mail:newton.pegolo@gmail.com

Received: 11 July 2010;
accepted: 1 May 2011

Summary

Validity of comparisons between expected breeding values obtained from best linear unbiased prediction procedures in genetic evaluations is dependent on genetic connectedness among herds. Different cattle breeding programmes have their own particular features that distinguish their database structure and can affect connectedness. Thus, the evolution of these programmes can also alter the connectedness measures. This study analysed the evolution of the genetic connectedness measures among Brazilian Nelore cattle herds from 1999 to 2008, using the French Criterion of Admission to the group of Connected Herds (CACO) method, based on coefficients of determination (CD) of contrasts. Genetic connectedness levels were analysed by using simple and multiple regression analyses on herd descriptors to understand their relationship and their temporal trends from the 1999–2003 to the 2004–2008 period. The results showed a high level of genetic connectedness, with CACO estimates higher than 0.4 for the majority of them. Evaluation of the last 5-year period showed only a small increase in average CACO measures compared with the first 5 years, from 0.77 to 0.80. The percentage of herds with CACO estimates lower than 0.7 decreased from 27.5% in the first period to 16.2% in the last one. The connectedness measures were correlated with percentage of progeny from connecting sires, and the artificial insemination spread among Brazilian herds in recent years. But changes in connectedness levels were shown to be more complex, and their complete explanation cannot consider only herd descriptors. They involve more comprehensive changes in the relationship matrix, which can be only fully expressed by the CD of contrasts.

Introduction

Genetic connectedness is an important factor to consider in comparisons between expected breeding values obtained from BLUP (Best Linear Unbiased Prediction) procedures (Henderson 1973). The breeding values have valid comparisons across groups with different fixed-effect levels if there are genetic links between the groups. Beef cattle herding is growing in economic importance in Brazil. Herds are usually composed of Nelore cattle and are characterized by their large size (usually >100 animals). Some ranchers have an elite herd (sire and dam production) mixed with a commercial herd (meat

  doi:10.1111/j.1439-0388.2011.00934.x

production), where pedigrees are sometimes poorly registered. Many breeders participate in the Nelore cattle breeding programme managed by ANCP (Associação Nacional dos Criadores e Pesquisadores, or National Association of Breeders and Researchers) to improve growth and fertility traits. In collaboration with the University of São Paulo, ANCP estimates the growth genetic merit of Nelore cattle by BLUP of breeding values on weight traits. Associated with this, the use of artificial insemination (AI) has been increasing in recent years. Semen sales in Brazil rose from 5 568 194 doses in 1999 to 8 204 783 in 2008 (ASBIA 2009), or approximately 47%. The evolution in management and technology application, such as improvement of the quality of pedigrees and the use of AI, is known to affect the measurement of genetic connectedness among herds (Fouilloux *et al.* 2008).

Different methods have been proposed to evaluate the connectedness of data, based on the prediction error variance (PEV) (Foulley *et al.* 1992; Kennedy & Trus 1993) or functions of it, such as the coefficient of determination (CD) (Laloë 1993; Laloë *et al.* 1996). The latter method considers the whole data design, as well as the balance between the decrease of PEV and the loss of genetic variability because of genetic relationships among animals. The CD is also related to the potential biases in the comparison between animals of different management groups with different genetic means (Laloë & Phocas 2003). Kuehn *et al.* (2007) examined the importance of connectedness and showed a consistent relationship between CD and different connectedness scenarios. CDs of comparison can be calculated by inverting the coefficient matrix of the mixed model equations (Henderson 1973). This procedure has restrictions when performed within large and complete data designs because of the complex matrix computation required. To avoid this problem, Fouilloux *et al.* (2008) proposed estimating CDs of comparisons between pairs of herds using a sampling-based method. A second step was added to define clustering groups of connected herds (the CACO method). Since 2002, this method has been the benchmark in France for estimating connectedness among the herds involved in onsite genetic evaluation of beef cattle from 2002, and for genetic evaluation of goats from 2007 onwards. The CD criterion was also used by Nakaoka *et al.* (2009) to improve national evaluation in Japanese Black cattle.

The objective of this study was to evaluate the evolution of the connectedness among Brazilian Nelore beef cattle herds by measuring CD of contrasts using the CACO method. Genetic connectedness levels were analysed by using simple and multiple regression analyses on herd descriptors to understand their relationship and their temporal trends from the 1999–2003 to the 2004–2008 period.

Material and methods

Data

This study was based on data from the Brazilian breeding programme – Nelore Brasil – conducted by ANCP. It involves evaluation of the genetic merit of weaning weight of Nelore cattle using a single-trait animal model for 210-day adjusted weight (W210). These data were analysed for connectedness in this study.

The original pedigree dataset is composed of animals born from 1974 to 2008, registered in 144 Brazilian herds. In this dataset, the average number of records per herd was 794, with a maximum of 15 584 records per herd. For the connectedness analyses, the relationship matrix was adapted to a sire model, and the dataset was used as shown in Table 1, according to the following method.

The CACO method

The Criterion of Admission to the group of COnnected Herds ('CACO') method was described by Fouilloux *et al.* (2008) and consists of estimating genetic connectedness among herds using a two-step procedure.

In the first step, the CDs of comparison between genetic levels of pair-wise herds are estimated by using a sampling method. In the second step, these

Table 1 Description of data used in FS08 and FS03 analyses. Heritabilities (h^2), number of performances (Nb perform), number of sires (Nb sires), number of contemporary groups (Nb CG) in the Brazilian breeding program, using the whole dataset for an animal model best linear unbiased prediction. Period, number of compared herds (Nb group) and number of performances (Nb perform) for the comparing group definition in the sire model of the connectedness evaluation

	FS08	FS03
Brazilian program (whole dataset)		
h^2	0.25	0.25
Nb perform	239 125	141 896
Nb sires	4954	3643
Nb CG	18 680	10 632
Connectedness analyses		
Period	2004–2008	1999–2003
Nb group	68	69
Nb perform	96 302	86 347

CDs are summarized with a clustering method so that each herd obtains a single CACO measure depending on the level of genetic connectedness. This CACO is used to determine the group of connected herds wherein EBVs can be compared.

Initially, the records were fitted in a mixed linear model, with one random factor and a residual effect:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \quad (1)$$

with the following variance structure:

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \mathbf{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{A}\sigma_{\mathbf{a}}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_{\mathbf{e}}^2 \end{pmatrix}\right] \quad (2)$$

where **y** is the performance vector, **b** the fixed-effect vector, **u** the random effect vector, **e** the residual vector, **X** and **Z** are the incidence matrices that associate elements of **b** and **u** with those of **y**; **A** is the numerator relationship matrix; the scalars $\sigma_a^2$and $\sigma_e^2$are the genetic and the residual variances, respectively. The Best Linear Unbiased Estimate (BLUE) of **b**, denoted $\mathbf{b^o}$, and BLUP of **u**, denoted $\hat{\mathbf{u}}$, are obtained by solving:

$$\begin{pmatrix} \mathbf{b}^\circ \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \lambda\mathbf{A}^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{pmatrix}$$

where

$$\lambda = \sigma_{\mathbf{e}}{}^2/\sigma_{\mathbf{a}}{}^2 \quad (3)$$

Accordingto Laloë (1993), the precision of a comparison between the genetic values of animals or groups of animals is assessed by the CD of the corresponding contrast. The contrast is written as a linear combination of the breeding values (**c'u**). So, for any linear contrast **c'u**, one can define the following CD:

$$\mathrm{CD}(\mathbf{c}'\hat{\mathbf{u}}) = \frac{\left(\mathrm{cov}\left(\mathbf{c}'\mathbf{u}, \mathbf{c}'\hat{\mathbf{u}}\right)\right)^2}{\mathrm{var}(\mathbf{c}'\mathbf{u})\mathrm{var}(\mathbf{c}'\hat{\mathbf{u}})} \quad (4)$$

with the vector **c** represented by a null vector except in the positions corresponding to the animals to be compared. We considered two herds connected if the CD of contrast between their genetic levels was greater than an 'a priori' threshold (χ).

The CD estimates were obtained using the method presented by Fouilloux & Laloë (2001) and Fouilloux *et al.* (2008), where variances and covariances of true and predicted linear combinations of breeding values were estimated from a simulated n-sample. The procedure was as follows:

(i) The animals involved in the simulation were sorted from the oldest to the youngest.

(ii) The direct genetic value u_i of animal *i* was calculated according to the status of its sire (*j*). If j was unknown, u_i was generated from N [0, $\sigma_a{}^2$]. If j was known, u_i was calculated by $u_i = 0.5\ u_j + \varphi_i$ where φ_i was drawn from *N* [0, $3\ \sigma_a{}^2/\ 4$].

(iii) Performance of each performance-tested animal (l) was simulated using the generated breeding value of its sire (j). Fixed effects were set to 0. Consequently, $y_l = s_i + \varepsilon_l$, with $s_j = 0.5 \times u_j$ and the residual ε_l, was drawn from *N* [0, σ_ε^2],where $\sigma_\varepsilon^2 = 3\sigma_\varepsilon^2/4 + \sigma_\varepsilon^2$.

(iv) The vector ŝ was obtained by solving the mixed model equations (MME) using **y**. This process repeated 1000 times led to vectors of genetic values $\{u^{(k)}\}_{k\ =\ 1,1000}$ and $\{\hat{u}^{(k)}\}_{k\ =\ 1,1000}$, where $\hat{u} = 2 \times \hat{s}$.

(v) The CDs of contrast of interest were estimated by computing their empirical variances and covariances and substituting them in the CD formula, according to Fouilloux *et al.* (2008).

Random numbers were generated by the NAG® (Numerical Algorithm Group, 1993) subroutines. BLUP was estimated using a successive overrelaxation iterative method, ceasing iteration when the convergence criterion (Fouilloux *et al.* 2008) was $<10^{-3}$. Breeding values are calculated by ANCP using BLUP procedures in a single-trait animal model. Genetic parameters were previously estimated: heritability ($h^2 = 0.25$), additive genetic variance ($\sigma_a^2 = 92.83$ kg^2),maternal additive genetic variance ($\sigma_m^2 = 38.55$ kg^2)and phenotypic variance ($\sigma_p^2 = 367.27$ kg^2). They generated the sire model parameters used in the connectedness evaluation: environmental variance ($\sigma_\varepsilon^2 = 344.06$ kg^2),additive genetic variance or sire variance ($\sigma_s^2 = 23.21$ kg^2)and lambda ($\lambda = \sigma_\varepsilon^2/\sigma_s^2 = 14.83$).

For the second step, sets of connected herds were built by using a clustering method, in such a way that any pair-wise CD of contrast between those herds was greater than χ. In this study, χ was defined as 0.4, following the French programme definition. The CACO method (Fouilloux *et al.* 2008) estimated connectedness across herds in the genetic evaluation using an alternative agglomerative clustering procedure, which was explicitly designed for building compact clusters for large datasets. According to this method, initially, each herd begins in a cluster by itself. Next, the two herds linked by the highest CD of contrast are clustered together, and they define the main cluster. A similarity index is calculated for each herd outside the main cluster. The similarity index of a given herd is equal to its lowest CD with the herd currently in the cluster.

The herd with the highest similarity index is added to the main cluster. The CACO of this new clustered herd is equal to its similarity index at this step.

The CACO method was performed using the in-house software developed by INRA and Institut de l'Elevage. Estimation of CDs of comparison was carried out by running BLUP analyses 1000 times in a sire model, simulating the whole dataset considered in the ANCP genetic evaluation on W210 except (i) animals without weight records and out of sires' pedigree and (ii) sires with no progeny after the previous exclusions. Performances of sires without pedigree information were removed. Sire model required that performance of animals with unknown sires should be removed. This removal should bias the estimation of the genetic level of herds, and hence of the measurement of genetic connectedness. To avoid these biases, Fouilloux *et al.* (2008) suggested creating 'fictitious sires' to replace the unknown sire information. Therefore, 'fictitious sires' were created as founder animals with genetic values randomly generated from N $[0, \sigma_a^2]$ for the sire model. Each herd with this situation was assigned one fictitious sire per year, so that there was no increment of connectedness between different herds.

Because the number of herds in this study was relatively small, the CACO method's results could be compared to those of the original complete linkage method (CLM), a hierarchical agglomerative clustering method that finds small and compact clusters that do not exceed a diameter threshold (Everitt 1974). In this method, the distance between two clusters is defined by:

$$D_{KL} = \max_{i \in C_k} \max_{i \in C_l} d(X_i, X_j) \quad (5)$$

The combinatorial formula is:

$$D_{JM} = \max(D_{JK}, D_{JL}) \quad (6)$$

where $d(x_i, x_j)$ is equal to 1 minus the CD of contrast between herds i and j, D_{KL} is the difference between clusters C_k and C_l, and D_{JM} is the difference between a new cluster C_m, originated from the next joining clusters C_k and C_l. In this method, the distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. Here, the results were summarized using dendrograms. CLM and dendrogram building were performed using the 'PROC CLUSTER' and 'PROC TREE' procedures in the SAS software (SAS Institute Inc., 2004), version 9.1.3, Cary, NC, USA. In Figure 2, the y axis is limited to the interval [0.1, 1.0].

Herd descriptor analyses

As the aim of this work was to evaluate the evolution of connectedness and its relation to herd descriptors along the time vector (years), and not the accumulative accuracy of the programme genetic evaluation, only animals born within 5 years in each herd composed the groups to be compared. This situation allows assuming that dams were responsible for perfect connectedness within the herds during this period. This can accentuate the differences between the sire model's results and those from a possible animal model approach (Kennedy & Trus 1993), where pedigrees of females are taken into account. The sire model assumed that most of connectedness among herds was created by the relationships among sires.

Two CACO analyses were performed. The first analysis (FS08) represented the current situation, and its groups were composed of animals born from 2004 to 2008. The second analysis (FS03) showed the past level of connection among herds involved in the ANCP programme until 2003, and the groups were composed of animals born from 1999 to 2003.

The data description for each analysis is shown in Table 1. Herd descriptors were analysed in an attempt to explain the CACO values and to deduce the importance of some herd features to improve the connectedness:

(i) Number of animals in the herd (NH).
(ii) Number of sires used in the herd (NS).
(iii) Percentage (in NS) of Connecting Sires (CS%). 'Connecting Sires' were defined as sires with calves in three or more herds in a single year, and with an average of more than two calves/herd/year. They were considered probable AI

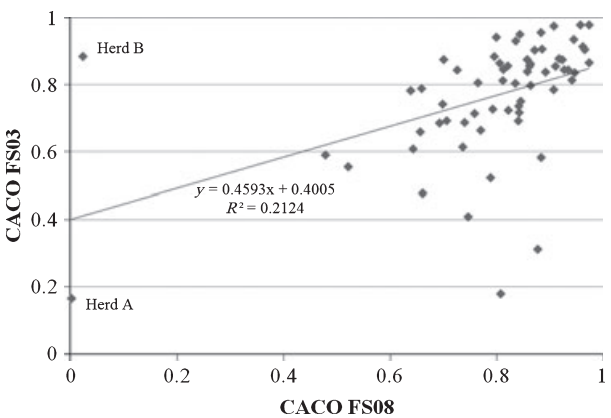

Figure 1 Regression analyses for CACO FS03 estimates on CACO FS08 estimates. Herd A and Herd B's corresponding points are assigned.

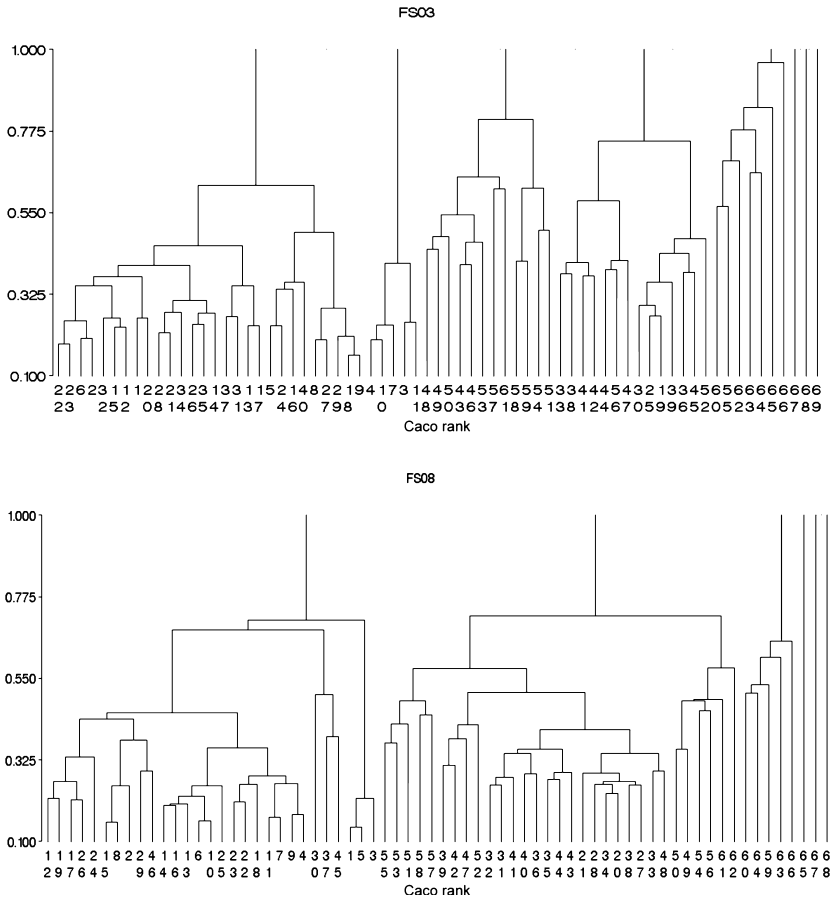

Figure 2 Tree graphs (dendrograms) resulting from the complete linkage method (CLM) elaborated with CD of contrasts estimates in FS03 and FS08. Distances between herds (heights) are defined by 1−CD, and the identification of herds corresponds to their position in the CACO ranking (Caco-Rank).

sires. There were 103 Connecting Sires found in the dataset.

(iv) Percentage of Progeny from Connecting Sires (PCS%).

(v) Percentage of Progeny from Connecting Paternal Grandsires (PCGS%).

(vi) Percentage of Calves with Unknown Sires (CUS%). In the database, 33.8% of the herds presented records from animals without sire information.

The parameters of herd descriptor distributions in FS08 and FS03 are presented in Table 2. The multivariate regression analyses had the following general model:

$$\mathrm{CACO} = b_0 + b_1\mathrm{NH} + b_2\mathrm{NS} + b_3\mathrm{CS}\% + b_4\mathrm{PCS}\% + b_5\mathrm{PCGS}\% + b_6\mathrm{CUS}\% + \varepsilon$$

where b_0, b_1, b_2, b_3, b_4, b_5 and b_6 were the linear coefficients for the regressions. They were performed in the FS08 and FS03 situations.

The evolution of the herd descriptors was analysed by their changes from 1999–2003 to 2004–2008 (ΔNH, ΔNS, ΔCS%, ΔPCS%, ΔPCGS% and ΔCUS%), and these differences were used to analyse the respective changes in CACO estimates (ΔCACO) over time.

The multivariate regression analyses had the following general model:

$$\Delta\mathrm{CACO} = b_0 + b_1\Delta\mathrm{NH} + b_2\Delta\mathrm{NS} + b_3\Delta\mathrm{CS}\% + b_4\Delta\mathrm{PCS}\% + b_5\Delta\mathrm{PCGS}\% + b_6\Delta\mathrm{CUS}\% + \varepsilon$$

where b_0, b_1, b_2, b_3, b_4, b_5 and b_6 were the linear coefficients for regressions.

In both analyses, their significances were determined by the *F*-test ($p < 0.05$). Multicollinearity was evaluated by variance inflation factors ($VIF = 1/(1-R^2)$) for each coefficient. VIF values above ten indicate significant multicollinearity (Belsley *et al.* 1980). In fact, the herd descriptors were expected to be redundant, so a matrix of correlations was defined considering each combination of variables in single regression analysis ($p < 0.05$).

Table 2 Distribution parameters (mean, standard deviation, minimum and maximum values) of herd descriptors (NH, NS, CS%, PCS%, PCGS% and CUS%)[a] from 68 herds in FS08 and FS03 analyses

	Mean		Standard deviation		Min		Max	
	FS08	FS03	FS08	FS03	FS08	FS03	FS08	FS03
NH	1416.2	1251.4	1403.9	1354.3	46	9	7797	7647
NS	59.6	60.5	35.5	47.4	12	5	200	210
CS%	31.2	37.5	14.4	15.6	5.1	0.0	66.7	73.1
PCS%	38.7	47.1	21.7	24.8	0.1	0.0	78.3	92.7
PCGS%	31.1	43.1	14.1	20.1	0.0	0.0	61.8	83.5
CUS%	3.0	4.1	13.5	13.6	0.0	0.0	86.4	83.8

[a]NH, Number of animals in the herd; NS, Number of sires used in the herd; CS%, Percentage (in NS) of connecting sires; PCS%, percentage of progeny from connecting sires; PCGS%, percentage of progeny from connecting paternal grandsires; CUS%, percentage of calves with unknown sires (CUS%).

Results and discussion

Connectedness evaluation

The CDs of contrast were estimated between the 68 herd genetic levels (2278 pair-wise comparisons two by two) in FS08 and 69 herd genetic levels (2346 comparisons) in FS03. The main CD of contrast statistics is in Table 3. The CD average increased slightly, from FS03 to FS08. The average of CACO estimates also increases slightly, from FS03 to FS08. With the threshold χ maintained at 0.4, as in French evaluation, the number of unconnected herds presented few changes. This shows that the Brazilian evaluation programme was and continues to be well connected. But the percentage of herds with CACO values below 0.7 decreased from 27.5% in the first period (FS03) to 16.2% in the last period (FS08), showing there was a positive evolution in connectedness among the herds in the Brazilian programme. Fouilloux *et al.* (2008) found a CACO average of 0.53 in the analysis of the Charolais breed and 0.297 in the Bazadais breed, showing lower levels of connectedness. Nakaoka *et al.* (2009) found contrast CD levels among three Japanese prefectures in Japanese Black cattle evaluation from 0.59 to 0.91, depending on the strategy of using a link provider in the analyses. Tarrés *et al.* (2010) found intermediate levels of average CD of contrast per herd (0.455) in the Bruna dels Pirineus breed, when heritability was 0.25 and where there was low level of AI practices.

The changes in CACO estimates from FS03 to FS08 for each herd were pronounced. A regression analysis between these variables showed $R^2 = 0.21$ and a significant linear coefficient of 0.4593, with a correlation coefficient equal to 0.46 (Figure 1). Two herds stood out in the graphs, exhibiting very odd behaviour, with extreme decreasing changes of CACO values from FS03 to FS08. They were specified as herd A and B. They are atypical herds, with very large size and a higher percentage of animals with unknown sires in FS08. They deserve a special attention because they can affect the subsequent regression analyses.

The complete linkage clustering method was also applied to the distances between herds (calculated by 1−CD). The resulting dendrograms for FS03 and FS08 are in Figure 2, with branch heights related to the distances between herds. The identification of the herds in the graphs was made by the ranking of CACO estimates from each analysis (CacoRank). Shorter branch heights connect higher CACO values identified by lower CACO ranks. The correlation coefficients between CLM heights and CACO values were −0.81 and −0.92 in FS08 and FS03, respectively. In fact, very low levels of connectedness were badly correlated, because the CLM method has no

Table 3 Description of distributions of contrast CD and CACO estimates for FS08 and FS03 analyses. The minimum and the maximum values are the same for both estimates

	CD of contrasts estimates			CACO Estimates				
Analyses	Total	Mean	SD	Total	Mean	SD	Min	Max
FS08	2278	0.81	0.19	68	0.80	0.17	0.00	0.97
FS03	2346	0.78	0.16	69	0.77	0.18	0.16	0.98

upper bound to the height values. If a limit of 1.0 was considered for CLM branch heights (Figure 2), only three herds would be totally unconnected, altering those correlation coefficients to −0.91 and −0.93, for FS08 and FS03, respectively. The high absolute correlations confirmed the validity of the CACO clustering method for identifying comparable herds. Two main connected groups were observed in FS08, while three herds were totally unconnected. Two of them, CacoRank 67 and 68, corresponded to Herd B and Herd A, respectively. The third one corresponded to the CacoRank 65, meaning that the CACO clustering method also assigned it among the lower connected herds. Looking to the past, the CLM distances in FS03 were notably larger, and there were five different connected groups found, reinforcing the positive evolution of herd connectedness.

Herd descriptor correlations

The results of multiple regression analyses considering the CACO estimates and herd descriptors are shown in Table 4. The coefficients of determination (R^2) were relatively high in FS08 and FS03 (0.82 and 0.71, respectively), but lower for ΔCACO (0.25). We tested the hypothesis that herds A and B were outliers and should be considered apart in ΔCACO. They were excluded from one regression analysis to verify how much impact they could account for. In this case, R^2 was higher, at 0.38, showing that those herds' descriptors should be analysed in more detail. The most important herd descriptor was PCS%, which had significant effects in CACO FS08 and CACO FS03. This result agrees with other studies that have found the use of connecting sires as a main factor to determinate connectedness levels (Laloë *et al.* 1996; Fouilloux *et al.* 2008; Nakaoka *et al.* 2009; Tarrés *et al.* 2010). Also, the change (ΔPCS%) was important to explain the ΔCACO without herds A and B, but not for the complete ΔCACO. The NH was also significant in CACO FS08 and CACO FS03, but its change was not significant in both ΔCACO analyses. The CUS% was important in CACO FS08, and NS was important in the complete ΔCACO. The low significance of other herd descriptors indicated they were not important to explain connectedness if they were not highly correlated (multicollinearity aspects).

The correlation coefficients (r) between CACO estimates in FS08 and FS03 and all the herd descriptors were also calculated in simple regression analyses. The results are shown in Table 5. The highest correlations were related to PCS%, CS% and PCGS% in all analyses, but they were strongly mutually correlated. The correlations between CACO estimates and NH were low in FS08. Fouilloux *et al.* (2008) had comparable results, where the CACO of a herd depended only slightly on the herd size ($r = 0.16$), while it increased much more with the number of sires used ($r = 0.57$). The percentage of unknown sires in a herd in that study also tended to decrease the CACO value ($r = -0.27$), while it increased with AI link sires used across herds ($r = 0.76$) at a very similar level.

The ΔCACO estimates from FS03 to FS08 were also analysed and compared to the herd descriptor changes. The highest correlation coefficient (r) was between ΔCACO and ΔPCS%, while lower values were found for ΔNH, ΔNS, ΔCS% and ΔPCGS%. As

Table 4 Linear coefficients (b_0, b_1, b_2, b_3, b_4, b_5 and b_6) to the corresponding herd descriptor (NH, NS, CS%, PCS%, PCGS% and CUS%), with the significance probability ($p < 0.05$, indicated by *) and the variance inflation factor (VIF). Coefficients of determination (R^2) for each multiple regression analysis (CACO FS08, CACO FS03, ΔCACO and ΔCACO without herds A and B) are shown in the last row

	CACO FS08		CACO FS03		ΔCACO		ΔCACO without A and B	
Coefficient	Value	VIF	Value	VIF	Value	VIF	Value	VIF
b_0	0.7052*		0.4516*		0.0608*		0.0938*	
$b_{1\ (NH)}$	3.2E−05*	3.35	4.3E−05*	2.43	−1.4E−05	1.88	2.4E−05	2.23
$b_{2(NS)}$	−0.0005	2.78	0.0002	2.89	0.0022*	2.61	0.0011	2.88
$b_{3(CS\%)}$	−0.0004	3.32	−0.0002	3.58	0.0026	2.91	0.0010	3.03
$b_{4(PCS\%)}$	0.0033*	3.12	0.0054*	2.71	0.0030	1.99	0.0035*	2.02
$b_{5(PCGS\%)}$	−9.8E−05	1.41	0.0003	1.82	−0.0013	1.32	−0.0001	1.37
$b_{6(CUS\%)}$	−0.0108*	2.13	−0.0010	1.42	−0.0127	1.17	0.0060	1.18
R^2	0.8172		0.7142		0.2477		0.3831	

CUS, calves with unknown sires; PCS, progeny from connecting sires; PCGS, progeny from connecting paternal grandsires.

Table 5 Correlation coefficients (*r*) between CACO estimates (in FS08 and FS03) and herd descriptors (NH, NS, CS%, PCS% and PCGS%)[a] and between its change (ΔCACO) and herd descriptor changes (ΔNH, ΔNS, ΔCS%, ΔPCS and ΔPCGS%) in independent analyses

	NH	NS	CS%	PCS%	PCGS%	CUS%
CACO FS08	−0.32	0.04	0.42	0.58	0.35	−0.81
NH	1.00	0.62	−0.43	−0.30	−0.07	0.49
NS		1.00	−0.34	−0.23	0.28	−0.08
CS%			1.00	0.80	0.27	−0.23
PCS%				1.00	0.30	−0.30
PCGS%					1.00	−0.35
CUS%						1.00
CACO FS03	0.29	0.26	0.52	0.76	0.43	−0.32
NH	1.00	0.73	−0.25	−0.10	−0.09	0.11
NS		1.00	−0.31	−0.07	0.04	−0.11
CS%			1.00	0.77	0.58	−0.34
PCS%				1.00	0.53	−0.34
PCGS%					1.00	−0.46
CUS%						1.00
	ΔNH	ΔNS	ΔCS%	ΔPCS%	ΔPCGS%	ΔCUS%
ΔCACO	0.07	0.19	0.17	0.36	0.09	−0.20
ΔNH	1.00	0.65	−0.24	0.01	0.06	0.24
ΔNS		1.00	−0.54	−0.11	−0.03	0.12
ΔCS%			1.00	0.64	0.36	−0.09
ΔPCS%				1.00	0.37	−0.09
ΔPCGS%					1.00	−0.26
ΔCUS%						1.00

[a]NH, Number of animals in the herd; NS, Number of sires used in the herd; CS%, percentage (in NS) of connecting sires; PCS%, percentage of progeny from connecting sires; PCGS%, percentage of progeny from connecting paternal grandsires; CUS%, percentage of calves with unknown sires (CUS%).

shown in Table 4, there were VIF values smaller than ten for all coefficients in all analyses. They showed that besides the high correlation between some descriptors, multicollinearity was not a problem in the multiple regression analyses.

The significance of ΔPCS% to explain ΔCACO in the regression analysis confirmed that the use of connecting sires, linked to AI use, is an important factor in measuring changes of connectedness (Laloë *et al.* 1996). But the need to exclude herds A and B showed that herd descriptors are not sufficient to have a good evaluation of changes in connectedness levels among herds. In addition, our results present something of a paradox, because the increase of CACO averages was accompanied by a decrease in PCS% averages from FS03 to FS08 (Table 2). ''Although, the increasing use...in PCS% averages''. On the other hand, the increasing use of connecting sires (probably artificial insemination sires) in Brazilian herds from FS03 to FS08 might be an important factor to explain the increase in connectedness over time and should imply an increase in PCS% averages. Three factors can explain this situation: (i) PCS% is a relative index that considers the total number of animals in the herd (NH). The NH averages increased from FS03 to FS08. So, just maintenance of AI use numbers would cause a decrease in PCS%; (ii) a second factor is the smaller standard deviation in FS08, which appears to reduce the number of herds with very low connectedness levels. So, in spite of the reduction or the maintenance of connecting sires and AI use levels within the same herds, other herds began to adopt the use of connecting sires or AI techniques. More widespread use of AI instead of larger Brazilian herds where AI was already used can be linked to the improvement of lower connectedness herds. This appears to be a logical explanation for the smaller number of main branches in the CLM dendrograms, from 5 in 2003 to 3 in 2008; (iii) finally, it is important to consider the increase of endogamy and reduction of the effective number of sires in the programme. Faria *et al.* (2002) showed that there was a rate of inbreeding per generation of 0.73 from 1994 to 1998 in Brazil's Nelore cattle population. They also found an increase of 122% in the inbreeding coefficient (*F*) and a decrease from 866 to 68 in the effective population number (N_e) from 1979 to 2001. In our study, CS% and NS were not good explanatory variables, probably because higher endogamy between sires could not be explained without a relationship matrix.

Atypical herds

Finally, focusing on the atypical herds A and B, they had the lowest CACO estimates in FS08 and the most important decreases in CACO estimates from FS03 to FS08 (Table 6). They presented a very large number of animals (6165 records on herd A and 4135 on herd B), with a high percentage of them with unknown sires (86.4% in herd A and 71.7% in herd B) in FS08. The evolution of their connectedness measures and herd descriptors appears to show important discrepancies compared to the general results.

Herd A had decreasing CACO estimates and lightly increasing PCS%. In the original data, the number of sires changed from nine to 35 and the number of connecting sires increased from zero to two. This situation seems to be counterintuitive. But NH was strongly increased, and CUS% was lightly increased from an already high level. The number of progenies from connecting sires only increased from zero to eight. In FS03, there was no one progeny from a

Table 6 Herd A and Herd B descriptors (NH, NS, CS%, PCS%, PCGS%, CUS%)[a] and respective CACO estimates, in FS08 and FS03 analyses

	Herd A		Herd B	
	FS08	FS03	FS08	FS03
NH	6165	235	4135	3183
NS	35	9	43	37
CS%	5.7	0.0	23.3	40.5
PCS%	0.13	0.0	11.1	21.4
PCGS%	0.02	1.28	7.5	17.3
CUS%	86.4	83.83	71.7	61.6
CACO	0.00	0.16	0.02	0.88

[a]NH, number of animals in the herd; NS, number of sires used in the herd; CS%, percentage (in NS) of connecting sires; PCS%, percentage of progeny from connecting sires; PCGS%, percentage of progeny from connecting paternal grandsires; CUS%, percentage of calves with unknown sires (CUS%).

connecting sire and three progenies from connecting paternal grandsire for 235 animals in the total herd. In FS08, there were eight progenies from connecting sires and one progeny from connecting paternal grandsire for 6165 animals in the total herd, or in a better format to the comparison, approximately one progeny from connecting sire for each 769 animals in the herd.

Herd B showed strongly decreasing CACO estimates and the same occurred to the PCS% which dropped down to almost a half, showing an expected positive correlation. But the CACO value in FS03 was larger (0.88) than the average CACO value (0.77) although PCS% was lower (21.4%) than the average PCS% (47.1%). It is important to notice that herd B size increased from 2003 to 2008 mainly because of the increase of animals with unknown parents (CUS% increased from 61.6 to 71.7, whereas NH increased from 3183 to 4135).

In a more general way, those features suggest two explanations for the discrepancies in those herds. The first one is related to the 'quality' of the connecting sire. This 'quality' reflects not only the sire pedigree but also its progeny distribution across herds. For example, a sire with 99% of its progeny in just one herd has a less important effect on connectedness than a sire with a balanced distribution of its progeny among all herds. The CACO method takes this 'quality' into account, which is difficult to evaluate by using only the herd descriptors. A second explanation is related to the high percentage of calves with unknown parents and the use of 'fictitious sires'. In fact, higher CUS% would expect to generate lower connectedness, because it implies in proportional less pedigree links. So, the use of 'fictitious sires' is a coherent solution. But in herds A and B, the high levels of CUS% appeared to supplant other herd descriptors' effect on connectedness measures. This proportionality can be assessed, but a prior situation must be considered: the importance of mixing data of the commercial and elite herds. This is the explanation to the high level of unknown parents in those two herds. The inclusion of these data only can increase the precision of fixed-effect estimates, provided this information is broadly distributed among management groups. Otherwise, more biases can be generated. Hence, maintaining the complete dataset covering groups of animals with unknown parents is questionable, mainly if they will not be selected by the breeding programme. The separation between commercial herd and elite herd and the exclusion of the first one appear to be the most suitable solution in this case. In general, high CUS% large herds will be less connected, and their evaluations will require more attention.

Conclusion

The CACO estimates based on the CD of contrasts method were relevant as connectedness measures among Brazilian Nelore cattle herds. On average, these herds showed a high level of CACO estimates in both periods analysed, with the majority of them over the considered threshold of 0.4.

There was an increase in connectedness in the Brazilian programme from the 1999–2003 to the 2004–2008 period. Although the average connectedness measures of CACO estimates slightly increased, fewer herds were below the defined threshold of unconnected herds in the latter period.

Multivariate regressions showed that the CACO values were largely explained by the set of herd descriptors, with coefficients of determination above 0.7. But the evolution of the CACO values was poorly explained, with coefficient of determination below 0.3. Exclusion of commercial herds with an excessive number of animals with poor pedigree information was able to increase the coefficient of determination in the regression. Then, the connectedness increase could be explained by the variation of the percentage of progeny of connecting sires (ΔPCS%), indicating that the importance of AI sires was spread out to the herds at sufficient levels to raise the CDs of contrasts among less-connected herds, even with a decrease of average levels of PCS%. So, the connectedness was more related to the start or increasing use of AI than the increase of AI by farms already using it. NH values were

significantly associated to CACO values, but NH changes were not significantly related to the evolution of the connectedness. Herds with high CUS% showed discrepant results, and exclusion of commercial herd data seems to be a logical option to the breeding programme.

Changes in connectedness levels were shown to be a more complex situation, and their complete explanation cannot consider only herd descriptors. This implies more complete changes in the relationship matrix, which can only be fully expressed by the CD of contrast.

Acknowledgements

We thank Institut National de la Recherche Agronomique and Institut de l'Elevage, and all people at the Station de Génétique animale et biologie intégrative in Jouy-en-Josas who contributed to this study. We also acknowledge the assistance of Associação Nacional de Criadores e Pesquisadores (ANCP) for providing access to the database and to Faculdade de Medicina de Ribeirão Preto and CAPES for financial support.

References

ASBIA, (2009) Relatório de vendas da Associação Brasileira de Inseminação Artificial. http://www.asbia.org.br/novo/upload/mercado/relatorio2009.pdf, Accessed: 30/08/2010.

Belsley D.A., Kuh E., Welsch R.E. (1980) Regression Diagnostics : Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York.

Everitt B.S. (1974) Cluster Analysis. Heinemann, London.

Faria F.J.C., Vercesi Filho A.E., Madalena F.E., Josahkian L.A.. Pedigree analysis in the Brazilian zebu breeds. In: World Congress on Genetics Applied to Livestock Production, 7, 2002, Montpellier. *Proceedings*... Montpellier, 2002.

Fouilloux M.N., Laloë D. (2001) A sampling method for estimating the accuracy of predicted breeding values in genetic evaluation. *Genet. Sel. Evol.*, **33**, 473–486.

Fouilloux M.N., Clément V., Laloë D. (2008) Measuring connectedness among herds in mixed linear models: from theory to practice in large-sized genetic evaluations. *Genet. Sel. Evol.*, **40**, 145–159.

Foulley J.L., Hanocq E., Boichard D. (1992) A criterion for measuring the degree of connectedness in linear models of genetic evaluation. *Genet. Sel. Evol.*, **24**, 315–330.

Henderson C.R. (1973) Sire evaluation and genetic trend. In: Proc. Anim. Breed. Genet. Symp. In Honor of Dr. Jay L. Lush. Am. Soc. Anim. Sci. And Am. Dairy Sci. Assoc., Champaign, IL, USA.

Kennedy B.W., Trus D. (1993) Considerations on genetic connectedness between management units under an animal-model. *J. Anim. Sci.*, **71**, 2341–2352.

Kuehn L.A., Lewis R.M., Notter D. (2007) Managing the risk of comparing estimated breeding values across flocks or herds through connectedness: a review and application. *Genet. Sel. Evol.*, **39**, 225–247.

Laloë D. (1993) Precision and information in linear models of genetic evaluation. *Genet. Sel. Evol.*, **25**, 557–576.

Laloë D., Phocas F. (2003) A proposal of criteria of robustness analysis in genetic evaluation. *Livest. Prod. Sci.*, **80**, 241–256.

Laloë D., Phocas F., Ménissier F. (1996) Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genet. Sel. Evol.*, **28**, 359–378.

Nakaoka H., Gaillard C., Fujinaka K., Watanabe N., Ito M., Kawada K., Ibi T., Sasae Y., Sasaki Y. (2009) The use of link provider data to improve national genetic evaluation across weakly connected subpopulations. *J. Anim. Sci.*, **87**, 62–71.

Numerical Algorithm Group (1993) The NAG Fortran Library Manual, mark 16. Oxford: The Numerical Algorithm Group Limited.

SAS Institute Inc. (2004) SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc.

Tarrés J., Fina M., Piedrafita J. (2010) Connectedness among herds of beef cattle bred under natural service. *Genet. Sel. Evol.*, **46**, 6.

