CLUSTER ANALYSIS OF THE ESTIMATES FROM GROWTH CURVES

Rita de Cassia Oliveira Barboza[1]
Fabiane de Lima Silva[1]
Kuang Hongyu[2]

ABSTRACT: (CLUSTER ANALYSIS OF THE ESTIMATES FROM GROWTH CURVES). The main of this study was to assessment the use of nonlinear mixed model to growth using a data of weight-age simulated of pigs by QTLMAS2009 workshop. In addition, we also aimed to identify the groups of animals with the growth patterns similar over time. In the first step, the mixed nonlinear model, Gompertz, was fitted with parameters as random or fixed effect. Comparisons between the fixed and mixed effects models are made. In two-step, an analysis of cluster was applied with the estimate the parameters of each animal in order to recognize similar groups of animals. It was shown that the Gompertz model with random effect was suitable to explain the growth than nonlinear model with fixed effect. The cluster analysis of the parameters of the Gompertz discriminated animals into three groups. The statistical analysis methodologies used contributed to the analysis of the data.

Keywords: Growth curves, longitudinal data, nonlinear models, swine,

[1]Centro de Ciências Agrárias, Ambientais e Biológicas, Universidade Federal do Recôncavo da Bahia – UFRB, Rua Rui Barbosa, Campus Universitário, CEP 44380-000, Cruz das Almas, BA, Brasil. E-mail: fabianesilva@ufrb.edu.br/ rita.cassia.93@hotmail.com

[2]Professor Doutor do Departamento de Estatística/Instituto de Ciências Exatas e da Terra. Universidade Federal de Mato Grosso, Av. Fernando Corrêa da Costa, nº 2367, Bairro Boa Esperança. CEP: 78060-900, Cuiabá, MT, Brasil. E-mail: prof.kuang@gmail.com

INTRODUCTION

The growth of domestic animals can be described over a sigmoid curve and can be fitted by nonlinear regression models. Different nonlinear models are used in animal production to describe the animal growth over time with different biological interpretation meaning related to initial conditions, growth rate, or adult body weight, all linked to economic aspects of production (Fernanda de Mello et al., 2015).

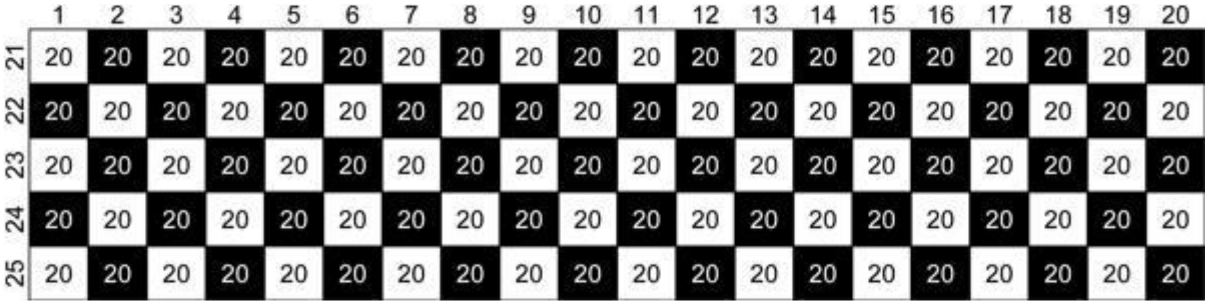
The study of the selection criteria is important for the breeding programs that aim to change the shape of the animal growth curves (Silva et al., 2011; Gonçalves et al., 2011). According to Freitas (2005), traditionally the models used in animal production are Brody, Richards, Von Bertlanffy and two alternatives are Gompertz and Logistic. These models are attractive because they allow the inclusion of both fixed and random effects, because they have a flexible covariance structure and allow adjustment in unbalanced data situations, which is limiting in the traditional regression approach (Lindstrom and Bates, 1990).

The multivariate clustering technique allows grouping individuals into homogeneous groups through which allows to gather the data in a number of groups with evaluation of characteristics of simultaneous interest, so that there is homogeneity within each group and heterogeneity between them (Silveira et al., 2012). It allows explanatory the existing variability among the animals, grouping them into homogeneous groups for a given set of data.

In this context, the objective of this work was to use the mixed nonlinear model approach to analyze weight-age data in pigs, and to clustering the animals with similar growth by cluster analysis.

MATERIALS AND METHODS

The simulated data comes from QTLMAS2009 (Workshop of Quantitative Trait Loci Mapping and Marker Assisted Selection). The dataset contains 2025 two-generation individuals who have complete marker information. There are 453 markers SNPs that are randomly distributed over five chromosomes. The first 25 individuals are parents, 20 females and 5 males. The remaining 2000 individuals are descendants, 100 families of complete siblings, one of each combination between males and females. Each full sibling family has 20 descendants. Of the 100 families reported, 50 (training population containing 1000 individuals) have phenotypic production records, the other 50 (validation population containing 1000 individuals) do not have phenotypic information. The phenotypes were recorded at five different times (0, 132, 265, 397, 530 days). More details in Coster et al. (2010; Silva et al., 2013). The entire dataset used is available at the following by address: http://www.qtlmas2009.wur.nl/UK/Dataset/. Figure 1 is a graphical representation of the combination structure used to simulate generation 2.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
22	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
23	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
24	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
25	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20

Figure 1. Graphical representation of the simulated second generation. Each i, j represents a family of complete siblings simulated by the combination of a female i and a male j. Black cells represent families of complete siblings whose phenotypic data were simulated; white cells represent families of complete sibs whose phenotypic data were not simulated. Each complete sibling family consists of 20 descendants (Coster et al., 2009).

In the first step, a descriptive statistical analysis of the data was performed. Successively, the statistical analysis consisted of the adjust of the growth curve using the nonlinear model Gompertz $y_i = \beta_1 * e^{-\beta_2 * e^{-\beta_3 * x_i}} + \epsilon_i$ in which the parameters presented in this model have the following meanings y_i describes the weight of the animal at a given age or time; $\beta_1$asymptotic value of y_i when the time x_i tends to ∞; β_3 indicates the speed at which the growth of the animal tends to adult weight; $\beta_2$refers to the degree of maturity of the animal at birth, that is, in t=0. Four nonlinear mixed model were adjusted to describe the growth pattern of the animals (MSE, MF1, MF2 e MF3) (Table 1).

Table 1 – Gompertz models used

Models	Expression	Random effects
MSE*	$y_i = \beta_1 * e^{-\beta_2 * e^{-\beta_3 * x_i}} + \epsilon_i$	-
MEF1**	$y_i = (\beta_1 + b_1) * e^{-\beta_2 * e^{-\beta_3 * x_i}} + \epsilon_i$	β_1
MEF2**	$y_i = \beta_1 * e^{-\beta_2 * e^{-(\beta_3 + b_3) * x_i}} + \epsilon_i$	β_3
MEF3**	$y_i = (\beta_1 + b_1) * e^{-\beta_2 * e^{-(\beta_3 + b_3) * x_i}} + \epsilon_i$	β_1, β_3

MSE: fixed model; **MEF: random models effects.

The strategy for the building of the models was to start with a model in which all the parameters as fixed effect (MSE), and the other models (MF1, MF2 and MF3) were included just one or two parameters as of random effect. The MF1 model (where only one parameter was considered as a random effect) presented a better fit to the weight-age data. In order to select models, the following criteria were used: i) Akaike information criterion (AIC): AIC = -2ln (mv) + 2p, where, ln = neperian logarithm, mv = likelihood function and ep = number of parameters of model; ii) Bayesian information criterion (BIC): BIC = -2ln (mv) + p ln (n), where n is the number of observations used to adjust the curve.

Hierarchical clustering and partitioning

In two-step, once the most appropriate model was selected, we required to group the animals into groups with similar growth behavior within the population, we applied cluster analysis with the estimates the parameters of each animal.

Hierarchical trees considered in this paper use the Ward's criterion. This criterion is based on the Huygens theorem which allows to decompose the total inertia (total variance) in betweenand within-group variance. The total inertia can be decomposed (HUSSON; JOSSE; PAGÈS, 2010):

$$\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\left(x_{iqk}-\bar{x}_k\right)^2=\sum_{k=1}^{K}\sum_{q=1}^{Q}\left(\bar{x}_{qk}-\bar{x}_k\right)^2+\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\left(x_{iqk}-\bar{x}_{qk}\right)^2,$$

Total inertia = Between inertia + Within inertia

with x_{iqk} the value of the variable k for the individual *i*of the cluster q, $\bar{x}_{qk}$ the mean of thevariable k for cluster q, $\bar{x}_k$ the overall mean of variable k and I_q the number of individuals incluster q.

The Ward's method consists in aggregating two clusters such that the growth of within-inertia is minimum (in other words minimising the reduction of the between-inertia) at each step ofthe algorithm. The within inertia characterises the homogeneous of a cluster (HUSSON; JOSSE; PAGÈS, 2010).

All analyzes of the data were performed with software free R, using the packages nlme (*linear and Nonlinear Mixed Efects Models*), and factorextra (*Extract and Visualize the Results of Multivariate Data Analyses*).

RESULTS AND DISCUSSION

The current study to examine one aspect of the heterogeneity of variance in the growth of animals using nonlinear mixed models. In Figure 1, the growth behavior of the animals increase over time, and the variability between individuals also increase, which suggests that the variances increase over time (Figure 1A e 1B). Experiment carry out with repeated measures over time, one must take into account that increasing of variance over time, there will be dependence between the errors, and correlation between measurements measured in time (Cestari et al., 2012).

In the correlation analysis between the weight and the days it is possible to verify positive relation between the variables, meaning that with the increase of the age will also increase of the weight; the correlation is higher than 0.9, indicating a strong association between the traits. The Durbin-Watson test revealed the presence of correlated residues (DW = 1.0218, p-value < 2.2e-16).

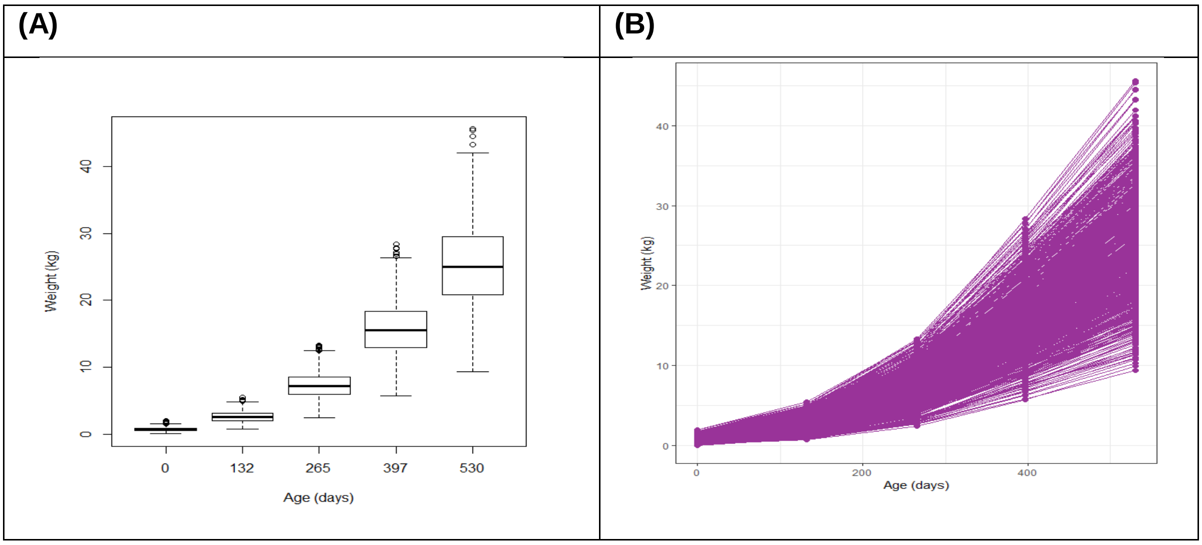


Figure 1 - Temporal analysis of weights of animals. (A) Box plot of weights over time; (B) individual animal profiles over time.

The parameter estimates and measures of goodness of fit for all growth models are presented in Table 2. Overall, the three models with random effects provided highly coefficient of determination (R^2: 0.97 – 0.98) compared to the fixed-effect model (R^2: 0.88). Considering that, the data set of the present study comes from a simulation study composed of animals of different genetic compositions and therefore shown great variability, the R^2 value for the four models with values above 0.88 shows that all the four models were well adjusted.

The use of criterion R^2 to evaluate the quality of fit of the models is important, however, other more accurate criteria have been used, for example Akaike information (AIC) and Bayesian criterion (BIC). The models with random effect in the parameters presented lower values for AIC and BIC. The MEF1 model was more suitable for adjusting weight-age data and was indicated to represent the growth curve of the animals. These findings agreed with those Craig and Schinckel (2001) that observed improvement in precision when estimating the growth rate by model with random effect. Das et al. (2017) used the models logistic, gompertz and von bertalanffy fixed and mixed models with data from pig growth and was observed that the logistic model fitted well for the data. The authors concluded that nonlinear mixed effects models provide more accurate and precise estimation of growth functions than the traditional fixed effects models.

Table 2 - Estimates of the parameters of the models

Model	Parameters					
	β_1	B_2	B_3	AIC	BIC	R^2
MSE*	67.67	4.86	0.0030	26068.56	26094.55	0.88
MEF1**	63.89	4.89	0.0031	11703.33	11735.81	0.98
MEF2**	84.23	4.90	0.0026	12484.48	12516.97	0.98
MEF3**	63.89	4.89	0.0031	11707.33	11752.81	0.97

MSE: fixed model; **MEF: random models effects.

In the Figures 1A, 1B and 1C we have the graphical representation of the values adjusted according to the predicted values of the models. Note that the results are quite similar for the models with the inclusion of the random effect in the parameters; the MSE model follows a trend different from the other models, and it is observed that the residues present an irregular distribution over time, being indicative of heterogeneity of variance. The lack of fit observed in the MSE model may be attributed to heterogeneity of variance not independent of residues that do not fit the features of longitudinal experiments.

Thus, a new approach of mixed nonlinear models is necessary to flexibilize these assumptions and allows modeling the variance between and within individuals. It is possible to observe that the inclusion of random effects to the parameters in the model has considerably improved the fit of the model.

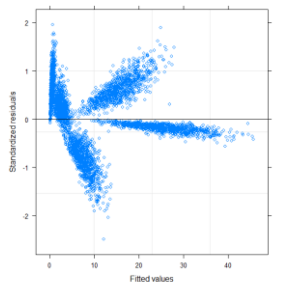
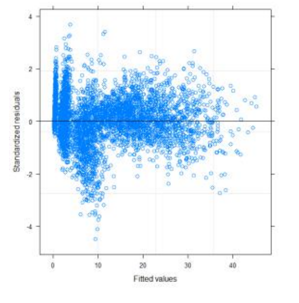
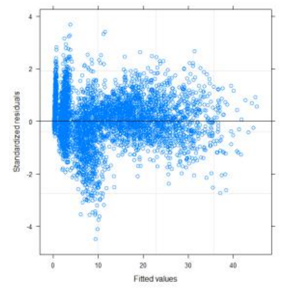
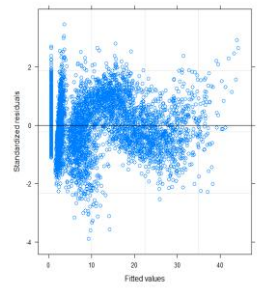
MSE	MEF1	MEF2	MEF3
			

Figure 2 - Graphic of the standardized residuals according to the predicted values of the models. MSE: fixed model; **MEF: random models effects.

The estimates of the parameters of the mixed nonlinear model selected in the first stage of the study (MEF1), the cluster analysis was performed by the Ward method. The clustering analysis makes possible the formation of homogeneous groups with the simultaneous evaluation of the characteristics of interest, this technique can be used to reduce the size of a dataset, reducing a wide range of objects to information of the center of its set. Since clustering is an unsupervised learning technique, can be useful to extract hidden characteristics of the data and develop the hypotheses about its nature (Lindeon, 2009).

The clustering analysis aimed at the classification of the animals in different groups, among which it has similarity according to the estimates of the parameters. The clustering analysis made it possible to group the 1000 animals into three distinct groups (Figure 3): i) Cluster1: 425 animals with high asymptotic weight, high growth rate; ii) Cluster2: 246 animals with low asymptotic weight and growth rates; iii) Cluster3: 309 animals with moderate asymptotic weight and growth rate.

The cophenetic correlation was 0.67, indicating how well the dendrogram reflects the grouping of the animals into three distinct groups. Additionally, the average silhouette width of cluster is 0.53, so all of the groups are quite well defined.

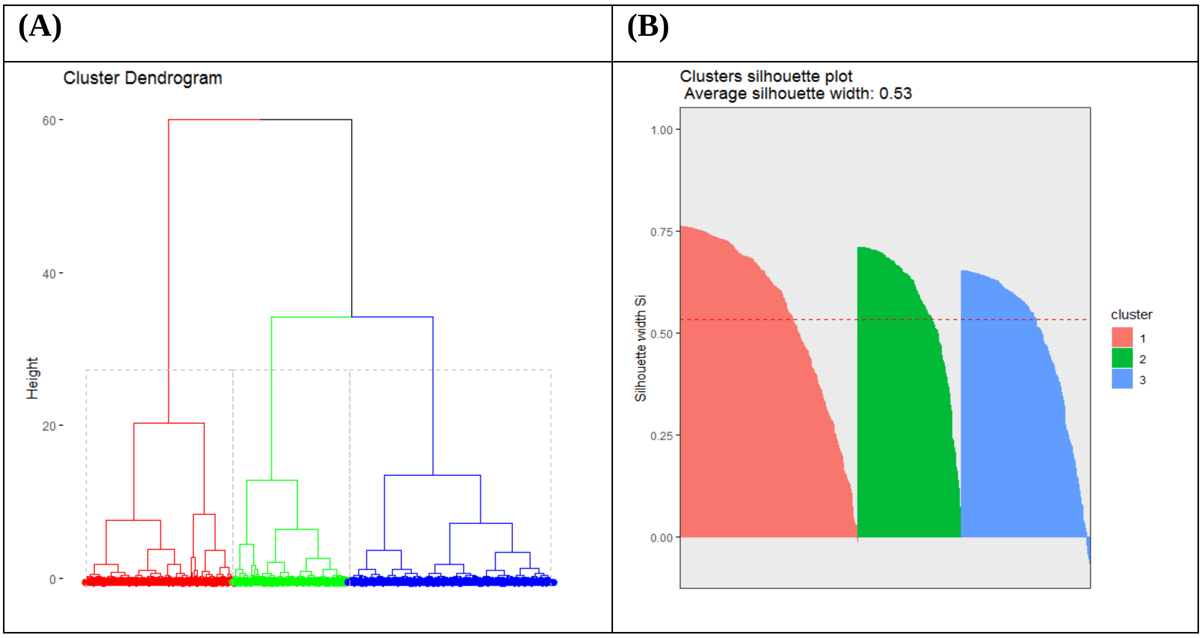


Figure 3 – Cluster analysis obtained with Ward's hierarchical, based on the Euclidean distance from parameters estimates animals of MF1 model.

Figure 4 shows the growth curve for each group of animals formed by cluster analysis. It is observed that Cluster1 and Cluster3, respectively, showed higher values for asymptotic weight. Cluster2, however, had a distinct and inferior behavior to the other groups. The inclusion of random effects in the model was important in describing the differences for the animal groups. In the early stage of pig growth, the weight variation was small, and with the change of age, this difference was greater among the animals formed by the clusters.

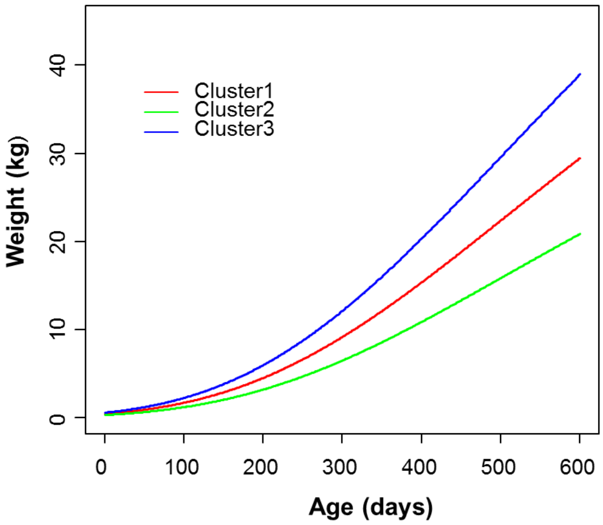


Figure 4 - Average curves of the three cluster of animals by the hierarchical grouping of Ward.

CONCLUSION

The statistical analysis methodologies used contributed to the analysis of the data. In addition, it was verified that the Gompertz model with random effect was adequate to explain the growth of animals. The animals were grouped into three groups based on the Gompertz model estimates.

REFERENCES

CESTARI A., SILVANO C. E MINHO, A. Análise de dados longitudinais em experimentação animal. **Semina: Ciências Agrárias**. 33: 1565-1580,2012.

COSTER A., BASTIAANSEN J.W. M.; CALUS M.P.L., MALIEPAARD C., BINK M.C.A. QTLMAS 2009: simulated dataset. BMC Proceedings, 4(Suppl 1): S3,2010.

CRAIG, B.A.; SCHINCKEL,P.A. Nonlinear Mixed Effects Model for Swine Growth. **The Professional Animal Scientist,** v17. 256-260, 2001.

FREITAS A.R. Curvas de crescimento na produção animal. **Revista Brasileira de Zootecnia**. 34: 786-795, 2005.

GONÇALVES T. M. DIAS M. A. D. AZEVEDO J.J. RODRIGUEZ M. A. P. TIMPANI V.D. OLIVEIRA A. I. G. Curva de crescimento de fêmeas da raça Nelore e seus cruzamentos. **Ciência e Agrotecnologia**. 35: 582-590, 2011.

HUSSON, F.; JOSSE, J.; PAGÈS, J. Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? **Technical Report – Agrocampus Applied Mathematics Department**, September 2010.

LINDSTROM, M.L. e BATES, D.M. Nonlinear mixed effects models for repeated measure data. **Biometrics.** v.46, p. 673 – 687, 1990.

SILVA, F. L.; ALENCAR, M. M.; FREITAS, A. R.; PACKER, I. U.; MOURÃO, G.B. Curvas de crescimento em vacas de corte de diferentes tipos biológicos. **Pesquisa Agropecuária Brasileira**, v.46, n.3, p.262-271, 2011.

SILVA F.F. ROCHA G. S. RESENDE M.D.V. GUIMARÃES S.E.F. PERTERNELLI L. A. DUARTE D.A.S. AZEVEDO C. Seleção genômica ampla para curvas de crescimento. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**. 64: 1519-1526, 2013.

SILVEIRA F.G. SILVA F.F. CARNEIRO P.L.S. MALHADO C.H.M. Classificação multivariada de modelos de crescimento para grupos genéticos de ovinos de corte. Revista **Brasileira Saúde Produção Animal**.13:62-73, 2012.

TEIXEIRA C.M. VILLARROEL A.B. PEREIRA E.S. OLIVEIRA S.M.P. ALBUQUERQUE I.A. MIZUBUTI I.Y. Curvas de crescimento de cordeiros oriundos de três sistemas de produção na região Nordeste do Brasil. 33: 2011-2018, 2012.