Comparação bayesiana de modelos de previsão de diferenças esperadas nas progênies no melhoramento genético de gado Nelore

Fabyano Fonseca e Silva[1], Thelma Sáfadi[2], Joel Augusto Muniz[2], Luiz Henrique de Aquino[2] e Gerson Barreto Mourão[3]

[1]Universidade Federal de Viçosa, Av. P.H. Rolfs, s/nº, CEP 36570-000 Viçosa, MG. E-mail: fabyano@dpi.ufv.br [2]Universidade Federal de Lavras, Caixa Postal 3037, CEP 37200-000 Lavras, MG. E-mail: safadi@ufla.br, joamuniz@ufla.br, lhaquino@ufla.br [3]Escola Superior de Agricultura Luiz de Queiroz, Av. Pádua Dias, nº11, Caixa Postal 9, CEP 13414-018 Piracicaba, SP. E-mail: gbmourao@carpa.ciagri.usp.br

Resumo – O objetivo deste trabalho foi realizar uma análise bayesiana de modelos auto-regressivos de ordem p, AR(p), para dados em painel referentes às diferenças esperadas nas progênies (DEP) de touros da raça Nelore publicados de 2000 a 2006. Neste trabalho, adotou-se o modelo AR(2), indicado pela análise prévia da função de autocorrelação parcial. As comparações entre as prioris, realizadas por meio do Fator de Bayes e do Pseudo-Fator de Bayes, indicaram superioridade da priori independente t-Student multivariada – Gama inversa em relação à priori hierárquica Normal multivariada – Gama inversa e a priori de Jeffreys. Os resultados indicam a importância de se dividir os animais em grupos homogêneos de acordo com a acurácia. Constatou-se também que, em média, a eficiência de previsão dos valores de DEP para um ano futuro foi próxima de 80%.

Termos para indexação: algoritmos MCMC, dados em painel, modelo auto-regressivo.

Bayesian comparison of forecasting models to expected progenies difference in Nelore cattle genetic breeding

Abstract – The objective of this work was to accomplish a bayesian analysis of an autoregressive, AR(p), panel data model from Nelore sires' expected progenie difference (EPD) observed during 2000–2006. The AR(2) model was used due to the results of partial autocorrelation function analysis. The prior comparisons were performed through Bayes Factor and Pseudo-Bayes Factor, and the results showed the independent t-Student multivariate – inverse Gamma superiority in relation to the hierarchical multivariate Normal – inverse Gamma and Jeffreys prior. Results indicate the importance of sires grouping by accuracy values, and also show forecast efficiency around 80%.

Index terms: MCMC algorithm, panel data, autoregressive model.

Introdução

Dados em estrutura de painel são representados por observações longitudinais de um conjunto de indivíduos, e, geralmente, a análise estatística desses dados é realizada por meio de modelos de regressão que utilizam informações de todo o conjunto para estimar coeficientes individuais. Entre esses modelos, destaca-se o auto-regressivo, que se aplica a diversas situações práticas e apresenta boa qualidade, quando comparado com outros mais complexos em relação à previsão de dados futuros (Broemeling & Cook, 1993).

Uma possível aplicação de modelos auto-regressivos para dados em painel está relacionada com a modelagem e previsão de diferenças esperadas nas progênies (DEP). Essa aplicação é possível porque as DEPs são calculadas e publicadas anualmente em sumários de touros, caracterizando assim séries temporais que podem ser descritas pelo modelo em questão. O conjunto de dados em painel é então caracterizado pelos touros e suas observações de DEPs ao longo dos anos, e a análise desses dados possibilita fazer previsões de DEPs de cada touro para anos futuros, promovendo assim um avanço tecnológico na área de melhoramento genético animal, uma vez que os pecuaristas poderão descartar precocemente touros que não correspondem às suas expectativas de produção em um ano futuro.

A análise de modelos de séries temporais para dados em painel considera que a combinação de informações de todos os indivíduos seja utilizada para estimar os coeficientes individuais de cada série. Dessa maneira, também é importante estimar os parâmetros da

distribuição representativa da população de coeficientes, o que implica na utilização de modelos hierárquicos, os quais muitas vezes tornam-se inviáveis de serem analisados pela inferência freqüentista (Liu & Tiao, 1980).

Sob o ponto de vista da inferência bayesiana, esta hierarquia pode ser facilmente incorporada, devido ao fato de se considerar informações a priori a respeito dos parâmetros a serem estimados. A teoria bayesiana está fundamentada em uma distribuição conjunta dos dados amostrais, denominada função de verossimilhança, e nas distribuições a priori a respeito dos parâmetros. Esses componentes determinam a distribuição a posteriori, posteriori $\propto$ verossimilhança x priori, na qual se realiza a inferência (Silva et al., 2005).

Segundo Barreto & Andrade (2004), o objetivo central da análise bayesiana de séries temporais é fornecer modelos robustos de previsão de observações futuras. Uma observação futura é descrita, sob o ponto de vista bayesiano, por uma distribuição condicional aos dados passados, denominada distribuição preditiva.

Devido ao fato de os modelos auto-regressivos de ordem p apresentarem termos de defasagem temporal, a função de verossimilhança apresenta-se condicionada a essas p primeiras observações, e ao se tratar de dados em painel, tal condicionalidade é inerente a todos os indivíduos, o que caracteriza uma perda considerável de informações (Liu & Tiao, 1980). Uma possível solução é a utilização de uma função de verossimilhança exata.

Em relação às distribuições a priori, é importante que sejam escolhidas por meio de métodos apropriados, e não simplesmente assumidos (Barreto & Andrade, 2004). De acordo com Kass & Raftery (1995), a comparação de prioris pode ser efetuada pelo Fator de Bayes, que, de forma geral, corresponde a uma razão de chances a posteriori, seguindo um processo análogo ao de razão de verossimilhanças no enfoque Freqüentista. Segundo esses autores, quando se tem interesse em comparar prioris impróprias, como o caso da priori não informativa de Jeffreys, o Fator de Bayes não é indicado, por proporcionar valores que impedem sua interpretação. Como alternativa, Gelfand (1996) propôs o Pseudo-Fator de Bayes, que está fundamentado nas distribuições preditivas obtidas para dados futuros. De maneira geral, o Fator de Bayes é mais informativo, pois apresenta uma escala de valores que permite quantificar o grau de superioridade de um modelo em relação ao outro (Kass & Raftery, 1995).

O objetivo deste trabalho foi utilizar a inferência bayesiana para ajustar modelo auto-regressivo de segunda ordem, para dados em painel referentes às diferenças esperadas nas progênies de touros da raça Nelore publicados de 2000 a 2006.

Material e Métodos

O modelo auto-regressivo para dados em painel é dado por:

$Y_{it} = \phi_{i1}Y_{i(t-1)} + \phi_{i2}Y_{i(t-2)} + \ldots + \phi_{mp}Y_{m(t-p)} + e_{it}$, ou ainda:

$$Y_{it} = \sum_{j=1}^{p} \phi_{ij} Y_{i(t-j)} + e_{it},$$

em que: i = 1,2,...m; j = 1,2,...p e t = 1,2,...,n_i. De acordo com essa notação, têm-se m indivíduos, com n_i observações longitudinais cada um, indicando que cada indivíduo i pode apresentar um número diferente de observações. O modelo contempla p parâmetros por indivíduo. No modelo em questão, Y_{it} é o valor atual de um processo estocástico de um indivíduo i, cujos valores já assumidos no passado são dados por $Y_{i(t-1)}$, $Y_{i(t-2)}$, ... , $Y_{i(t-p)}$; ϕ_{i1}, ϕ_{i2}, ..., ϕ_{ip}, são os parâmetros de auto-regressão e e_{it} é o resíduo, também denominado de ruído branco, $e_{it} \overset{iid}{\sim} N(0,\sigma_e^2)$.

No que se segue, foi considerado $n = n_1 = n_2 = ... = n_m$, ou seja, o mesmo número de observações longitudinais para cada indivíduo i, i = 1,2,...,m. Respeitando o modelo apresentado, pode-se escrever a função de verossimilhança exata como:

$$L(\mathbf{Y} \mid \Phi, \sigma_e^2) = \Psi(\Phi, \sigma_e^2 \mid \mathbf{Y}_p)\sigma_e^{2-\left(\frac{m(n-p)}{2}\right)} \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{m}\sum_{t=p+1}^{n}(Y_{it} - \sum_{j=1}^{p}\phi_{ij}Y_{i(t-j)})^2\right\},$$

em que:

$$\Psi(\Phi, \sigma_e^2 \mid \mathbf{Y}_p) = \sigma_e^{2-\left(\frac{mp}{2}\right)} \left|\mathbf{V}_p\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_e^2}\mathbf{Y}_p'\mathbf{V}_p\mathbf{Y}_p\right\}.$$

A matriz V_P, obtida diretamente pelo método de Yule-Walker (Morettin & Toloi, 2004), assume uma estrutura bloco diagonal, a qual é dada, respectivamente, para modelos AR(1) e AR(2), por:

$$\mathbf{V}_p = \begin{bmatrix} 1-\phi_{11}^2 & 0 & 0 & 0 \\ 0 & 1-\phi_{21}^2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 1-\phi_{m1}^2 \end{bmatrix}_{mm} \quad \text{e}$$

$$\mathbf{V}_p = \begin{bmatrix} 1-\phi_{12}^2 & -\phi_{11}(1+\phi_{12}) & 0 & 0 & 0 & 0 & 0 \\ -\phi_{11}(1+\phi_{12}) & 1-\phi_{12}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-\phi_{22}^2 & -\phi_{21}(1+\phi_{22}) & 0 & 0 & 0 \\ 0 & 0 & -\phi_{21}(1+\phi_{22}) & 1-\phi_{22}^2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1-\phi_{m2}^2 & -\phi_{m1}(1+\phi_{m2}) \\ 0 & 0 & 0 & 0 & 0 & -\phi_{m1}(1+\phi_{m2}) & 1-\phi_{m2}^2 \end{bmatrix}$$

Reescrevendo a função de verossimilhança para todos os indivíduos em forma matricial, têm-se:

$$L(\mathbf{Y} \mid \Phi, \sigma_e^2) \propto \Psi(\Phi, \sigma_e^2 \mid \mathbf{Y}_p)\sigma_e^{2^{-\left(\frac{m(n-p)}{2}\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{Y}_1 - \mathbf{X}\Phi)'(\mathbf{Y}_1 - \mathbf{X}\Phi)\right\},$$

em que:

$\mathbf{Y}_p = [y_{11}, y_{12},..., y_{1p}, y_{21}, y_{22},..., y_{2p},..., y_{m1}, y_{m2},..., y_{mp}]'$,

$\mathbf{Y}_1 = [y_{1p+1}, y_{1p+2},..., y_{1n}, y_{2p+1}, y_{2p+2},..., y_{2n},..., y_{mp+1},..., y_{mn}]'$,

$\Phi = [\phi_{11}, \phi_{12},..., \phi_{1p}, \phi_{21}, \phi_{22},..., \phi_{2p},..., \phi_{m1}, \phi_{m2},..., \phi_{mp}]' \in R^{mp}$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & 0 & 0 \\ 0 & \mathbf{X}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{X}_m \end{bmatrix} \quad \text{e} \quad \mathbf{X}_i = \begin{bmatrix} y_{ip} & \cdots & y_{i1} \\ y_{ip+1} & \cdots & y_{i2} \\ \vdots & \ddots & \vdots \\ y_{in-1} & \cdots & y_{in-p} \end{bmatrix}_{(n-p)\times p}.$$

Os vetores Y1 e Φ, considerados nas expressões, apresentam, respectivamente, as dimensões $m(n-p) \times 1$ e $mp \times 1$.

Neste trabalho, optou-se por comparar três diferentes distribuições a priori para os parâmetros Φ e σ_e^2. Foram consideradas duas prioris informativas, priori hierárquica Normal multivariada – Gama Inversa (Modelo 1) e priori independente t-Student multivariada – Gama Inversa (Modelo2); e uma não informativa, a priori de Jeffreys (Modelo 3).

Para o Modelo 1,

$P(\Phi, \sigma_e^2) = P(\Phi|\sigma_e^2)P(\sigma_e^2)$, sendo $(\boldsymbol{\Phi}|\sigma_e^2) \sim N(\boldsymbol{\mu}, \sigma_e^{-2}\mathbf{P})$ e $\sigma_e^2 \sim GI(\alpha,\beta)$ portanto:

$$P(\Phi, \sigma_e^2) \propto \sigma_e^{2^{-\left(\frac{mp+2\alpha}{2}+1\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[2\beta + (\Phi - \mu)'\mathbf{P}^{-1}(\Phi - \mu)\right]\right\}.$$

Os componentes μ, P, α e β são denominados hiperparâmetros, e representam os parâmetros das distribuições dos parâmetros do modelo considerado.

Para o Modelo 2, optou-se pela utilização de prioris independentes, as quais são dadas por: Φ~t – Student (μ, P) com v graus de liberdade e $\sigma_e^2 \sim GI(\alpha, \beta)$, ou seja:

$$P(\Phi, \sigma_e^2) \propto \left[1 + (\Phi - \mu)'\mathbf{P}^{-1}(\Phi - \mu)\right]^{-\frac{v+mp}{2}} (\sigma_e^2)^{-(\alpha+1)} \exp\left\{-\frac{\beta}{\sigma_e^2}\right\}.$$

O Modelo 3 representa a situação em que não se tem informações definidas a respeito dos parâmetros, portanto utilizou-se a abordagem descrita por Broemeling & Cook (1993) da priori de Jeffreys para modelos auto-regressivos, a qual é dada por: $P(\Phi, \sigma_e^2) \propto 1/\sigma_e^2$.

De acordo com a teoria bayesiana, posteriori ∝ verossimilhança x priori para cada distribuição a priori especificada, deve-se obter uma distribuição a posteriori, uma vez que no presente trabalho função de verossimilhança é única. Sendo assim, obtiveram-se as seguintes distribuições conjuntas a posteriori:

Modelo 1:

$$P(\Phi, \sigma_e^2 \mid \mathbf{Y}) \propto \Psi(\Phi, \sigma_e^2 \mid \mathbf{Y}_p)\sigma_e^{2^{\frac{-m(n-p)}{2}}} \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{Y}_1 - \mathbf{X}\Phi)'(\mathbf{Y}_1 - \mathbf{X}\Phi)\right\}$$

$$\times\sigma_e^{2^{-\left(\frac{mp+2\alpha}{2}+1\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[2\beta + (\Phi - \mu)'\mathbf{P}^{-1}(\Phi - \mu)\right]\right\},$$

em que:

$$D = \beta + [(\mathbf{Y}_1'\mathbf{Y}_1 + \boldsymbol{\mu}'\mathbf{P}^{-1}\boldsymbol{\mu}) - (\mathbf{X}'\mathbf{Y}_1 + \mathbf{P}^{-1}\boldsymbol{\mu})'(\mathbf{X}'\mathbf{X} + \mathbf{P}^{-1})^{-1}$$

$$(\mathbf{X}'\mathbf{Y}_1 + \mathbf{P}^{-1}\boldsymbol{\mu}]/2, \boldsymbol{\Sigma} = \mathbf{X}'\mathbf{X} + \mathbf{P}^{-1},$$

$$\text{e } \hat{\Phi}_B = (\mathbf{X}'\mathbf{X} + \mathbf{P}^{-1})^{-1}(\mathbf{X}'\mathbf{Y}_1 + \mathbf{P}^{-1}\boldsymbol{\mu}).$$

Modelo 2:

$$P(\Phi, \sigma_e^2 \mid \mathbf{Y}) \propto \Psi(\Phi, \sigma_e^2 \mid \mathbf{Y}_p)\sigma_e^{2^{-\left[\frac{m(n-p)+2\alpha}{2}+1\right]}}$$

$$\times \exp\left\{-\frac{1}{\sigma_e^2}\left[\left(\frac{(\Phi - \hat{\Phi})'(\mathbf{X}'\mathbf{X})(\Phi - \hat{\Phi}) + (\mathbf{Y}_1 - \hat{\mathbf{Y}}_1)'(\mathbf{Y}_1 - \hat{\mathbf{Y}}_1)}{2}\right) + \beta\right]\right\}$$

$$\times\left[1 + (\Phi - \mu)'\boldsymbol{P}^{-1}(\Phi - \mu)\right]^{-\left(\frac{v+p}{2}\right)}, \text{em que:}$$

$$\hat{\Phi} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}_1) \text{ e } \hat{\mathbf{Y}}_1 = \mathbf{X}\hat{\Phi} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}_1).$$

Modelo 3:

$$P(\Phi, \sigma_e^2 \mid \mathbf{Y}) \propto \sigma_e^{2^{-\left(\frac{mn}{2}+1\right)}} \left|\mathbf{V}_p\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma_e^2}\left[\mathbf{Y}_p'\mathbf{V}_p^{-1}\mathbf{Y}_p + \right.\right.$$

$$+ (\boldsymbol{\Phi} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1)'(\mathbf{X}'\mathbf{X})(\boldsymbol{\Phi} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1) +$$

$$\left.\left. + (\mathbf{X}'\mathbf{Y}_1)'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1 + \mathbf{Y}_1'\mathbf{Y}_1\right]\right\}.$$

Para obter as distribuições marginais a posteriori, e assim realizar as inferências, é necessário integrar a distribuição conjunta a posteriori em relação a cada parâmetro, Φ e σ_e^2. Neste trabalho, verificou-se que essas integrais não apresentam soluções analíticas, e por isso utilizou-se os algoritmos MCMC, Metropolis-Hastings e Gibbs Sampler (Gamerman & Lopes, 2006), respectivamente, para obter amostras das marginas a posteriori de Φ e σ_e^2. Portanto, para que esses algoritmos possam ser computacionalmente implementados, é necessário obter as distribuições condicionais completas a posteriori, sendo essas as seguintes:

Modelo 1:

$$\Phi \mid \sigma_e^2, \mathbf{Y} \sim \Psi(\Phi, \sigma_e^2 \mid \mathbf{Y}_p) \text{Normal multivariada}(\hat{\Phi}_B, \sigma_e^{-2}\Sigma^{-1})$$

$$\sigma_e^2 \mid \Phi, \mathbf{Y} \sim \text{Gama Inv.}\left(\frac{mp+mn+2\alpha}{2}, \frac{1}{2}(\mathbf{Y}_p'\mathbf{V}_p\mathbf{Y}_p) + \right.$$

$$\left. + D + \frac{1}{2}(\Phi - \hat{\Phi}_B)'\Sigma(\Phi - \hat{\Phi}_B)\right).$$

Modelo 2:

$\Phi \mid \sigma_e^2$, $Y \sim \psi(\Phi, \sigma_e^2 \mid \mathbf{Y}_p)$ x Normal multivariada

$(\hat{\Phi}, (\mathbf{X}'\mathbf{X})^{-1})$ x t-Student multivariada $(\boldsymbol{\mu}, \mathbf{P}^{-1})$

$$\sigma_e^2 \mid \Phi, \mathbf{Y} \sim \text{Gama Inv.}\left(\frac{mn+2\alpha}{2}, \frac{1}{2}(\mathbf{Y}_p'\mathbf{V}_p\mathbf{Y}_p) + \right.$$

$$\left. +\beta + \frac{(\Phi-\hat{\Phi})'(\mathbf{X}'\mathbf{X})(\Phi-\hat{\Phi}) + (\mathbf{Y}_1-\hat{\mathbf{Y}}_1)'(\mathbf{Y}_1-\hat{\mathbf{Y}}_1)}{2}\right)$$

Modelo 3:

$\Phi \mid \sigma_e^2$, $Y \sim$ Normal multivariada $(0, V^{-1}{}_p)$ x

$$\sigma_e^{2-\left(\frac{mn}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma_e^2}\left[(\Phi-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1)'(\mathbf{X}'\mathbf{X})(\Phi-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1)\right]\right\} \text{ e}$$

$$\sigma_e^2 \mid \Phi, \mathbf{Y} \sim \text{Gama Inversa}\left(\frac{mn}{2}, \frac{1}{2}\left[\mathbf{Y}_p'\mathbf{V}_p\mathbf{Y}_p + (\Phi-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1)'(\mathbf{X}'\mathbf{X})(\Phi-\right.\right.$$

$$\left.\left. -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1) + (\mathbf{X}'\mathbf{Y}_1)'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1 + \mathbf{Y}_1'\mathbf{Y}_1\right]\right).$$

A distribuição condicional completa para o parâmetro σ_e^2 é dada por uma distribuição Gama-Inversa, ou seja, ela apresenta uma forma conhecida, portanto, utilizou-se o algoritmo Gibbs Sampler. O mesmo não acontece para a distribuição condicional do parâmetro Φ, a qual, para todas as prioris especificadas, não apresenta uma forma definida, devendo-se então utilizar, nessa situação, o algoritmo Metropolis-Hastings.

Os algoritmos Gibbs Sampler e Metropolis-Hastings foram implementados matricialmente no software estatístico R (R Development Core Team, 2006). A função mnormt ("multivariate Normal and t-Student distributions") foi utilizada para a geração dos números aleatórios implícita nos algoritmos em questão. O tamanho final das cadeias de cada parâmetro, bem como a especificação do "burn-in", foram determinadas conforme o protocolo de Nogueira et al. (2004), que sugeriram, respectivamente, 20.000 e 10.000 iterações. A constatação final da convergência foi dada pelo critério de Gelman & Rubin (1992). Ambos os critérios foram avaliados mediante o pacote BOA ("Bayesian Output Analysis") do software R. Os códigos do R encontram-se no seguinte endereço eletrônico: http://www.dex.ufla.br/~muniz/Downloads/downloads.html.

A distribuição preditiva de determinado dado futuro (Barreto & Andrade, 2004) é dada por:

$$P(\mathbf{Y}_{(n+1)} \mid \mathbf{Y}) \propto \iint_{\sigma_e^2\,\Phi} (\sigma_e^2)^{-\frac{m}{2}} exp\left\{-\frac{1}{2\sigma_e^2}\left[\left(\mathbf{Y}_{(n+1)}-\mathbf{X}\Phi\right)'\left(\mathbf{Y}_{(n+1)}-\right.\right.\right.$$

$$\left.\left.\left. -\mathbf{X}\Phi\right)\right]\right\} P(\Phi, \sigma_e^2 \mid \boldsymbol{Y})\, d_\Phi d_{\sigma_e^2},$$

em que:

$P(\Phi, \sigma_e^2 \mid \mathbf{Y})$ é a distribuição a posteriori e

$(\sigma_e^2)^{-\frac{m}{2}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[\left(\mathbf{Y}_{(n+1)}-\mathbf{X}\Phi\right)'\left(\mathbf{Y}_{(n+1)}-\mathbf{X}\Phi\right)\right]\right\}$ é a função de verossimilhança desenvolvida a partir do seguinte modelo:

$$\mathbf{Y}_{i(n+1)} = \phi_{i1}\mathbf{Y}_{in} + \phi_{i2}\mathbf{Y}_{i(n-1)} + \phi_{i3}\mathbf{Y}_{i(n-2)} + ... + \phi_{ip}\mathbf{Y}_{i(n+1-p)} + e_{i(n+1)}.$$

A integral descrita não apresenta solução analítica, porém conforme a teoria descrita por Heckman & Leamer (2001), é possível demonstrar que mediante a utilização da técnica MCMC, tem-se:

$$\mathbf{Y}_{(n+1)}^{(q)} \mid \mathbf{Y} \sim N\left(\mathbf{X}\Phi^{(q)}, \sigma_e^{2(q)}\mathbf{I}\right),$$

em que: I é uma matriz de ordem mp × mp.

O conjunto de valores gerados para essa distribuição Normal multivariada, provenientes de cada q iteração dos algoritmos MCMC, constitui a distribuição preditiva para um dado futuro, cuja estimativa, $\hat{P}(\mathbf{Y}_{(n+1)} \mid \mathbf{Y})$, é representada pela média desta distribuição. Caso seja de interesse, pode-se generalizar esse método para a predição de k dados futuros, $Y_{(n+k)}$.

A fim de comparar os Modelos 1 (priori hierárquica Normal multivariada - Gama Inversa) e 2 (priori independente t-Student multivariada - Gama Inversa), utilizou-se o Fator de Bayes (FB) sob o enfoque apresentado por Barreto & Andrade (2004). Esse utiliza valores gerados pelos métodos MCMC para obter as estimativas do fator de normalização, $P(Y \mid M_p)$ também denominado de Verossimilhança Marginal, o qual compõe a expressão do Fator de Bayes:

$$FB_{ij} = \frac{\hat{P}(\mathbf{Y}\mid M_i)}{\hat{P}(\mathbf{Y}\mid M_j)} = \frac{\frac{1}{Q}\sum_{q=1}^{Q} L(\mathbf{Y}\mid\theta^{(q)}, M_i)}{\frac{1}{Q}\sum_{q=1}^{Q} L(\mathbf{Y}\mid\theta^{(q)}, M_j)}.$$

O termo $\theta^{(q)}$ indica os valores gerados para os parâmetros na q-ésima iteração (q = 1,2,...,Q) para cada

um dos modelos comparados. Assim, $L(Y \mid \theta^{(q)}, M_p)$ corresponde a valores da função de verossimilhança obtidos pela substituição dos valores atuais dos parâmetros. Usando a função de verossimilhança adotada neste trabalho, tem-se a seguinte estimativa da Verossimilhança Marginal de um modelo p:

$$\hat{P}(\mathbf{Y}|M_p) = \frac{1}{Q}\sum_{q=1}^{Q} \Psi(\Phi^{(q)}, \sigma_e^{2(q)} \mid \boldsymbol{Y}_p)\sigma_e^{2(q)^{-m\left(\frac{n-p}{2}\right)}}$$

$$\exp\left\{-\frac{1}{2\sigma_e^{2(q)}}\left[\mathbf{Y}_1 - \mathbf{X}\Phi^{(q)}\right]'\left[\mathbf{Y}_1 - \mathbf{X}\Phi^{(q)}\right]\right\}$$

Em relação à interpretação do FB, segundo Kass & Rafftery (1995), se FB>1, a indicação é de que o modelo do numerador é o melhor, se FB<1, o modelo do denominador é o preferido, e se FB = 1, a qualidade dos dois modelos é a mesma.

Para comparar os Modelos 1 e 2 com o Modelo 3 (Priori não informativa de Jeffreys), utilizou-se o Pseudo-Fator de Bayes, uma vez que o Modelo 3 considera uma priori imprópria. Gelfand (1996) apresenta a seguinte definição para este critério de comparação:

$$PsdFB_{ij} = \ln\left(\frac{\prod_{i=1}^{k}\hat{P}(Y_{(n+k)} \mid Y, M_i)}{\prod_{i=1}^{k}\hat{P}(Y_{(n+k)} \mid Y, M_j)}\right),$$

em que: $\prod_{i=1}^{k}\hat{P}(\mathbf{Y}_{(n+k)} \mid \mathbf{Y}, M\_)$ é o produtório de k valores futuros gerados pelas distribuições preditivas de cada modelo. Se $PsdFB_{ij}>0$, seleciona-se o modelo i, caso contrário o modelo j. Neste trabalho, adotou-se k = 1, e por se tratar de dados em painel, generalizou-se a expressão de acordo com Ansari et al. (2002), obtendo-se:

$$PsdFB_{ij} = \ln\left(\frac{\prod_{i=1}^{m}\hat{P}(Y_{i(n+l)} \mid Y, M_i)}{\prod_{i=1}^{m}\hat{P}(Y_{i(n+l)} \mid Y, M_j)}\right),$$

a qual considera o produto das estimativas de uma observação futura de cada indivíduo i.

Foram utilizados dados cedidos pelo Grupo de Melhoramento Animal da Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo (Pirassununga, SP), os quais correspondem aos sumários de touros Nelore compreendidos entre os anos de 2000 e 2006. Constam, no arquivo, dados de DEPs para a característica ganho de peso entre a desmama (205 dias de idade) e o sobreano (550 dias de idade) de 117 touros, divididos em três grupos, conforme o valor da acurácia referente à última observação, isto é, aos dados de 2006. Esses grupos foram divididos como seguem: baixa (0 a 40%), contendo 31 touros, média (41 a 60%), contendo 63 touros e alta (acima de 60%), com a freqüência de 23 touros. Também foi considerada a análise de todos os animais, sem a separação em grupos de acordo com a acurácia. Foram utilizados, portanto, quatro arquivos de dados, e todos eles submetidos a análises considerando os três modelos adotados (modelo 1, 2 e 3).

Para avaliar a capacidade de previsão de cada modelo, em todos os grupos, a última observação foi suprimida, ou seja, o valor da DEP referente ao ano de 2006, visando assim à comparação direta entre os valores estimados por meio da distribuição preditiva e os verdadeiros valores observados.

Todo o desenvolvimento teórico foi realizado para um modelo auto-regressivo de ordem p, AR(p), mas, ao se aplicar o método, considerou-se apenas o modelo AR(2), uma vez que as análises das funções de autocorrelação parcial indicaram essa ordem como a mais adequada para todos os conjuntos de dados.

Os valores iniciais para os parâmetros foram obtidos de um ajuste médio, ou seja, obteve-se uma série cujos valores são as médias de DEPs em cada ano, e a esta ajustou-se, separadamente, para cada estrutura de dados, um modelo AR(2) por meio do método da máxima verossimilhança. Esse ajuste foi realizado mediante o pacote ar.mle ("Autoregressive Maximum Likelihood Estimate") do software R. A média e a variância dos valores iniciais de cada parâmetro foram usadas como hiperparâmetros das prioris informativas.

Resultados e Discussão

Os avaliadores da qualidade de ajuste, Fator de Bayes e Pseudo-Fator de Bayes, foram calculados separadamente dentro de cada estrutura de dados especificadas de acordo com a acurácia (Tabela 1). A distribuição t-Student multivariada proporcionou melhor qualidade de ajuste em relação às outras duas prioris utilizadas em todas as estruturas de dados consideradas, embora a superioridade em relação à distribuição Normal multivariada seja de pequena magnitude, se avaliada mediante o critério de interpretação do Fator de Bayes mediante escala

apresentada por Kass & Rafftery (1995). Resultados semelhantes foram obtidos por Barreto & Andrade (2004), os quais concluíram que a priori t-Student multivariada foi a mais adequada para modelos auto-regressivos usados na previsão de rompimento e barragens. Huerta & West (1999) também relatam a superioridade dessa mesma distribuição em relação a outras prioris informativas, simétricas e assimétricas, tendo em vista a capacidade preditiva como critério de comparação.

O pior desempenho foi apresentado pela priori de Jeffreys (Tabela 1), indicando que mesmo quando não se têm informações a priori, prioris informativas podem ser consideradas visando à comparação de sua eficiência em relação a prioris não informativas (Lambert et al., 2005).

Os fatos abordados nos dois parágrafos anteriores são menos evidentes na estrutura de baixa acurácia, uma vez que os valores obtidos para o fator de Bayes e para o Pseudo-Fator de Bayes são de menores magnitudes. Isso talvez seja devido a uma maior heterogeneidade no comportamento das séries referentes a esta estrutura, o que dificulta a seleção de um melhor modelo.

É possível inferir que as séries referentes à população de indivíduos com baixa acurácia apresentam um comportamento diferenciado das

Tabela 1. Valores obtidos para o Fator de Bayes e para o Pseudo-Fator de Bayes em cada conjunto de dados considerado.

Estrutura dos dados	Critério
Acurácia baixa	$FB_{21} = \frac{105.542,540}{52.502,635} = 2,0102$
	$PsdFB_{23} = \ln\left(\frac{-0,0875}{-0,0035}\right) = 3,2188$
	$PsdFB_{13} = \ln\left(\frac{-0,0541}{-0,0035}\right) = 2,7380$
Acurácia média	$FB_{21} = \frac{206.664,210}{62.402,354} = 3,3118$
	$PsdFB_{23} = \ln\left(\frac{-0,1255}{-0,0011}\right) = 4,7369$
	$PsdFB_{13} = \ln\left(\frac{-0,0902}{-0,0011}\right) = 4,4067$
Acurácia alta	$FB_{21} = \frac{176.001,435}{58.213,034} = 3,0234$
	$PsdFB_{23} = \ln\left(\frac{-0,0987}{-0,0023}\right) = 3,7591$
	$PsdFB_{13} = \ln\left(\frac{-0,0857}{-0,0023}\right) = 3,6179$
Geral	$FB_{21} = \frac{112.040,331}{48.111,012} = 2,3287$
	$PsdFB_{23} = \ln\left(\frac{-0,0898}{-0,0021}\right) = 3,7556$
	$PsdFB_{13} = \ln\left(\frac{-0,0741}{-0,0021}\right) = 3,5634$

demais, pois a percentagem de significância dos parâmetros foi menor (Tabela 2). O parâmetro ϕ_2 apresentou percentagem de significância um pouco menor que aquela obtida para o parâmetro ϕ_1 em duas estruturas de dados, média e geral, indicando que as séries de DEPs de alguns indivíduos podem não apresentar comportamento auto-regressivo de segunda ordem, podendo essas serem descritas por uma ordem inferior, AR(1), ou ainda se apresentarem como um processo estacionário, que indica a ausência de autocorrelação.

Tabela 2. Percentagem de significância dos parâmetros do modelo AR(2) ajustado às séries individuais de diferenças esperadas nas progênies.

Estrutura dos dados	Significância (%)	
	Parâmetro ϕ_1	Parâmetro ϕ_2
Acurácia baixa	43,47	43,47
Acurácia média	73,01	65,07
Acurácia alta	65,21	65,21
Geral	61,53	57,26

Os resultados da Tabela 2 também estão relacionados com a seleção de indivíduos, cujas densidades preditivas serão apresentadas, pois não há razão em discutir a previsão de valores futuros de séries individuais que não apresentam um processo auto-regressivo de primeira ou de segunda ordem, uma vez que a densidade preditiva está fundamentada no modelo AR(2) para dados em painel.

O método utilizado para realizar previsões de dados futuros individuais com base na obtenção das distribuições preditivas foi eficiente, uma vez que a percentagem de intervalos de credibilidade que continham os verdadeiros valores das DEPs referentes ao ano de 2005 variou entre 77,78 e 85,71% (Tabelas 3, 4 e 5). Em relação a essas percentagens,

Tabela 3. Verdadeiros valores da última observação (y_6), suas estimativas e intervalos de credibilidade (95%), considerando a estrutura de acurácia baixa.

Touro	y_6	$\hat{y}_6$	LI	LS	Touro	y_6	$\hat{y}_6$	LI	LS
6	8,78	9,60	7,11	12,02	20	8,41	9,40	5,72	12,18
8	8,71	10,85	5,34	15,35	21	9,05	8,44	6,82	12,32
9	8,48	10,89	4,88	16,09	22	8,40	8,97	3,21	12,25
10	3,28	9,39	4,42	10,57	23	9,15	12,77	10,07	16,28

Tabela 4. Verdadeiros valores da última observação (y_6), suas estimativas e intervalos de credibilidade (95%), considerando a estrutura de acurácia média.

Touro	y_6	$\hat{y}_6$	LI	LS	Touro	y_6	$\hat{y}_6$	LI	LS
2	8,49	9,87	-0,78	11,02	41	12,66	11,29	7,83	16,15
5	6,55	5,25	1,09	17,74	42	8,95	8,67	2,72	14,00
6	9,11	9,57	6,93	12,89	43	10,23	10,73	6,06	15,96
8	12,32	10,04	8,79	17,72	44	6,84	5,52	2,79	13,54
10	9,06	12,75	9,28	15,24	45	8,80	8,05	2,26	17,54
11	9,77	7,18	2,28	15,86	46	6,43	9,29	7,06	12,07
16	10,99	14,20	11,03	17,84	47	9,90	7,95	2,03	18,55
17	7,21	5,20	3,81	8,75	48	9,18	13,12	10,81	17,39
21	7,66	7,05	3,29	9,59	51	11,25	14,32	2,09	19,32
22	8,92	9,42	3,96	18,98	52	5,28	5,82	2,65	9,44
24	4,94	8,58	5,26	17,31	53	12,49	13,22	7,74	18,56
25	8,22	8,63	3,05	12,12	54	8,10	13,76	3,02	18,02
26	9,39	11,35	7,48	14,68	55	8,70	6,71	3,21	11,33
28	8,55	8,55	4,96	12,94	56	9,29	7,17	3,86	15,51
30	9,67	7,86	-0,07	13,36	57	9,61	8,61	5,68	12,93
31	12,87	12,78	5,52	18,96	58	9,91	11,51	6,13	14,66
33	10,06	15,86	11,44	20,13	59	9,00	10,50	4,88	16,46
35	14,95	12,01	6,47	18,32	60	9,65	10,79	2,92	18,97
37	8,36	10,21	2,66	17,91	61	10,73	10,31	6,20	15,72
38	8,58	7,01	2,21	18,01	62	5,22	7,19	0,02	12,68
40	10,86	10,44	6,10	15,40	63	9,59	15,15	11,04	21,07

Tabela 5. Verdadeiros valores da última observação (y_6), suas estimativas e intervalos de credibilidade (95%), considerando a estrutura de acurácia alta.

Touro	y_6	$\hat{y}_6$	LI	LS	Touro	y_6	$\hat{y}_6$	LI	LS
1	9,20	10,38	3,10	12,01	12	14,10	14,24	2,39	23,83
3	9,06	10,94	7,12	12,08	14	10,84	14,06	6,40	16,90
6	10,55	17,46	12,30	23,07	15	8,38	6,69	1,40	13,95
7	10,03	17,62	11,99	25,09	17	9,71	10,82	4,12	15,62
8	12,55	11,25	2,22	25,13	19	13,91	13,64	4,05	23,69
9	9,12	7,51	3,19	16,31	20	11,41	12,55	2,98	19,16
10	10,85	12,93	4,90	18,65	22	11,79	12,91	4,06	17,92

Tabela 6. Estimativas (médias a posteriori), intervalos de credibilidade (95%), fator de Gelman-Rubin, e "burn-in" para cada estrutura de dados em relação à acurácia.

Estrutura dos dados	$\hat{\sigma}_e^2$	LI	LS	$\sqrt{R}$	Burn-in
Acurácia baixa	84,32	53,52	128,32	0,9918	562
Acurácia média	31,00	19,25	47,33	0,9994	320
Acurácia alta	37,99	26,35	58,97	0,9999	128

Hay & Pettit (2001) e Brandt & Willians (2007), que também utilizaram modelos auto-regressivos para dados em painel, obtiveram valores entre 58 e 88%. Discussões relacionadas com as classes de acurácia devem ser consideradas, pois a capacidade preditiva do modelo é um pouco menor, quando a acurácia é baixa, 77,78%, e nas acurácias médias e altas, respectivamente, 83,33 e 85,71%. Esse fato também é explicado mediante estimativas obtidas para a variância do erro (σ_e^2) em cada uma das estruturas de dados consideradas (Tabela 6). A variância do erro foi maior na estrutura de dados relacionada com animais de baixa acurácia, e esse resultado confirma discussões anteriores relacionadas com a maior variabilidade no comportamento das séries dos indivíduos desse arquivo. Resultados semelhantes, que ressaltam a importância da avaliação da acurácia na avaliação genética animal, são apresentados por Garnero et al. (2002).

Conclusões

1. A análise bayesiana de dados de diferenças esperadas nas progênies (DEPs) de touros Nelore, considerando anos sucessivos, resulta em boas estimativas para o valor de um dado futuro, obtidas a partir da distribuição preditiva.

2. O método usado possibilita a detecção de diferenças nos comportamentos de séries referentes a indivíduos pertencentes a grupos distintos de acurácia, indicando que esse fator deve ser levado em consideração ao se fazer as previsões.

3. As diferenças entre as acurácias não impõem barreiras para a adoção desse método na presença de dados provenientes de animais com baixa acurácia para DEPs, visto que a percentagem de acerto na previsão de um dado futuro foi próxima a 80%.

Referências

ANSARI, A.; JEDIDI, K.; DUBE, L. Heterogeneous factor analysis models: a Bayesian approach. **Psychometrika**, v.67, p.49-78, 2002.

BARRETO, G.; ANDRADE, M.G. Robust Bayesian approach for AR(p) models applied to stream flow forecasting. **Journal of Applied Statistical Science**, v.12, p.269-292, 2004.

BRANDT, P.T.; WILLIANS, J.T. **Modeling multiple time series**. Beverly Hills: Sage University, 2007. 148p.

BROEMELING, L.D.; COOK, P. Bayesian estimation of the mean of an autoregressive process. **Journal of Applied Statistics**, v.20, p.25-38, 1993.

GAMERMAN, D.; LOPES, H.F. **Markov chain Monte Carlo**: stochastic simulation for Bayesian inference. 2.ed. New York: Chapman & Hall, 2006. 324p.

GARNERO, A.V.; FERNANDES, M.B.; FIGUEIREDO, L.F.C.; LÔBO, R.B. Influência da incorporação de dados de progênies na classificação de touros da raça Nelore. **Revista Brasileira de Zootecnia**, v.31, p.918-923, 2002.

GELFAND, A.E. Model determination using sampling based methods. In: GILKS, W.R.; RICHARDSON, S.; SPIEGELHALTER, D.J. (Ed.). **Markov chain Monte Carlo in practice**. London, UK: Chapman and Hall, 1996. p.145-162.

GELMAN, A.; RUBIN, D.B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v.7, p.457-511, 1992.

KASS, R.E.; RAFTERY, A.E. Bayes factors. **Journal of the American Statistical Association**, v.90, p.773-795, 1995.

HAY, J.L.; PETTITT, A.N. Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. **Biostatistics**, v.2, p.433-444, 2001.

HECKMAN, J.; LEAMER, E. **Handbook of Econometrics**: v.5. Amsterdam: Elsevier Science, 2001. 744p.

HUERTA, G.; WEST, M. Priors and component structures in autoregressive time series models. **Journal of the Royal Statistical Society**: Serie B (Statistical Methodology), v.61, p.881-899, 1999.

LAMBERT, P.C.; SUTTON, A.J.; BURTON, P.R.; ABRAMS, K.R.; JONES, D.R. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. **Statistics in Medicine**, v.24, 2401-2428, 2005.

LIU, L.M.; TIAO, G.C. Random coefficient first-order autoregressive models. **Journal of Econometrics**, v.13, p.305-325, 1980.

MORETTIN, P.A.; TOLOI, C.M.C. **Análise de séries temporais**. São Paulo: Edgard Blucher, 2004. 537p.

NOGUEIRA, D.A.; SÁFADI, T.; FERREIRA, D.F. Avaliação de critérios de convergência para o método de Monte Carlo via Cadeias de Markov. **Revista Brasileira de Estatística**, v.65, p.59-88, 2004.

R Development Core Team. **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2006. Disponível em: http://www.R-project.org. Acesso em: 29 nov. 2007.

SILVA, F.F.; MUNIZ, J.A.; AQUINO, L.H.; SÁFADI, T. Abordagem Bayesiana da curva de lactação de cabras Saanen de primeira e segunda ordem de parto. **Pesquisa Agropecuária Brasileira**, v.40, p.27-33, 2005.

Recebido em 30 de maio de 2007 e aprovado em 3 de dezembro de 2007