Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Aplicação da Metodologia Box & Jenkins a dados de produção de leite cru do estado do Paraná (PR)

Eduardo Campana Barbosa[1][2]
Carlos Henrique Osório Silva[3]
Rômulo César Manuli[2]
Ricardo Gonçalves Tavares[4]
Tiago Bittencourt Nazaré[5]

1 Introdução

O aumento da produção de leite no Brasil pode ser explicado principalmente pela reestruturação do setor lácteo a partir da década de 1990, além da recente busca por mercados externos. Estes acontecimentos priorizaram o planejamento do processo produtivo do leite, enfatizando o quão importante é o estudo das tendências deste mercado, para atender sua demanda. Logo, este trabalho teve por objetivo a aplicação da metodologia Box & Jenkins aos dados de produção de leite cru (resfriado ou não) do estado do Paraná (PR). Foram ajustados modelos SARIMA (*Seasonal Auto Regressive Integrated Moving Average*) e estimados os valores de produção de leite para os 9 primeiros meses de 2013. Para a escolha do modelo de previsão utilizou-se o princípio da parcimônia, além os indicadores de erro RSME e MAPE.

2 Materiais e Métodos

As informações utilizadas são do banco de dados SIDRA do IBGE (http://www.ibge.gov.br/home/) e referem-se à série histórica de produção de leite cru (em litros) do estado do Paraná (PR). Foram obtidas observações mensais dos últimos 10 anos e 9 meses, sendo que as 9 últimas foram reservadas para comparar com as previsões.

A série foi decomposta nas componentes de tendência, sazonalidade e parte aleatória (não modelável). A presença de tendência e sazonalidade foi verificada, respectivamente, pelos testes de Cox-Stuart e G de Fisher, conforme em [5] e [2]. O modelo SARIMA

[1] Agradecimentos à Fapemig, Capes e CNPq pelo apoio financeiro.
[2] Mestrando em Estatística Aplicada e Biometria da UFV (DET-UFV); duducampana@hotmail.com; romulomanuli@ig.com.br
[3] Professor Associado III do Departamento de Estatística da UFV (DET-UFV); chos@ufv.br
[4] Aluno de Graduação em Ciências da Computação da UFV (DPI-UFV); ricardo.tavares@ufv.br
[5] Professor das Faculdades Integradas de Cataguases (FIC-UNIS); tiago@unis.edu.br

multiplicativo, conforme a equação (1), foi utilizado para modelar a autocorrelação periódica e inferior a um ano (sazonalidade).

$$\phi_p(B)\Phi_P(B^s)W_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (1)$$

$\phi_p(B)$ e $\theta_q(B)$ são os operadores auto-regressivos e de médias móveis e $\Phi_P(B^s)$ e $\Theta_Q(B^s)$ os operadores auto-regressivos e de médias móveis sazonais. W_t é a série Y_t após aplicação de "d" e/ou "D" diferenças de ordem 1 e s[6], para remover a tendência e sazonalidade dos dados, tornando a série estacionária, que segundo [4], é uma pressuposição para aplicar tais modelos. $\varepsilon_t \overset{i.i.d}{\sim} N(0, \sigma^2)$, ou um ruído branco associado ao tempo t.

A aplicação da metodologia Box & Jenkins consistiu em três etapas usuais: identificação dos parâmetros do modelo, estimação e diagnóstico. Neste trabalho, a identificação ocorreu selecionando os 4 modelos mais parcimoniosos pelo AIC (Akaike Information Criterion) [1]. A estimação dos parâmetros foi por máxima verossimilhança, supondo distribuição normal e satisfazendo as condições de invertibilidade e unicidade [7]. O diagnóstico consistiu em verificar se os resíduos estimados seriam um ruído branco, por meio do teste estatístico Ljung-Box (independência) [3] e Shapiro-Wilk (normalidade) [8].

Devido ao elevado valor das observações, a escolha do modelo de previsão ocorreu pelos indicadores Raiz Quadrada do Erro Quadrático Médio (RSME) e Erro Percentual Médio Absoluto (MAPE), calculados como em [5] utilizando os valores estimados e reais de produção de leite para os 9 primeiros meses de 2013. As análises estatísticas foram desenvolvidas no *software* livre R [6].

3 Resultados e Discussão

A Figura 1 apresenta a série em estudo e sua decomposição. As componentes de tendência e sazonalidade parecem estar bem definidas, como esperado, devido o crescimento do mercado lácteo desde 1990 e as variações sazonais causadas pelas estações do verão e inverno, respectivamente, maior e menor produção de leite.

As hipóteses de existência de tendência e sazonalidade foram testadas pelo teste não paramétrico de Cox-Stuart e G de Fisher. Pela aproximação normal bilateral $z_{\alpha/2}$ = 1,96, o valor da estatística t foi de 70,73 e como T (60) ≥ n – t (49,27) rejeita-se a hipótese nula a 5% de significância concluindo que a série apresenta tendência. Para o teste G de Fisher, o valor

[6] O termo s indica a periodicidade com que a sazonalidade ocorre

de G (0,7655) > Z (0,1132) o que implica que a sazonalidade com periodicidade de 12 meses é significante.

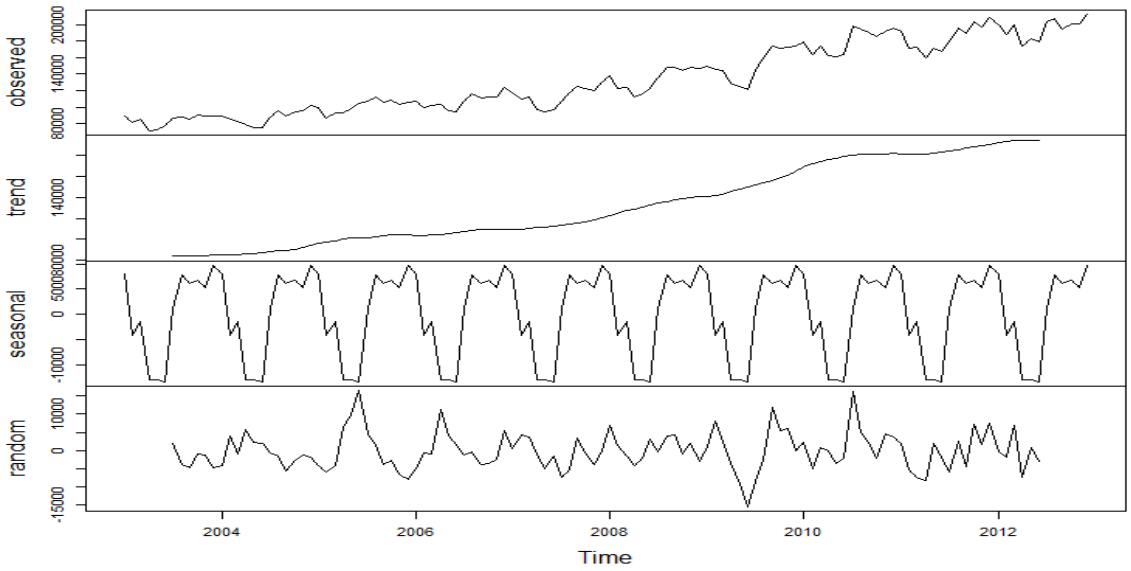

Figura 1: Decomposição da série temporal em tendência, sazonalidade e parte aleatória.

Logo, uma diferença de ordem 1 e ordem s = 12 foram aplicadas para eliminar estas componentes. Para verificar se tais procedimentos tornaram a série estacionária, a Figura 2 ilustra sua ACF e PACF. Os dados oscilam em torno de um valor médio e com variância aproximadamente constante. Além disso, a maioria dos coeficientes de autocorrelação são iguais a 0, o que é uma evidência de que a série é estacionária.

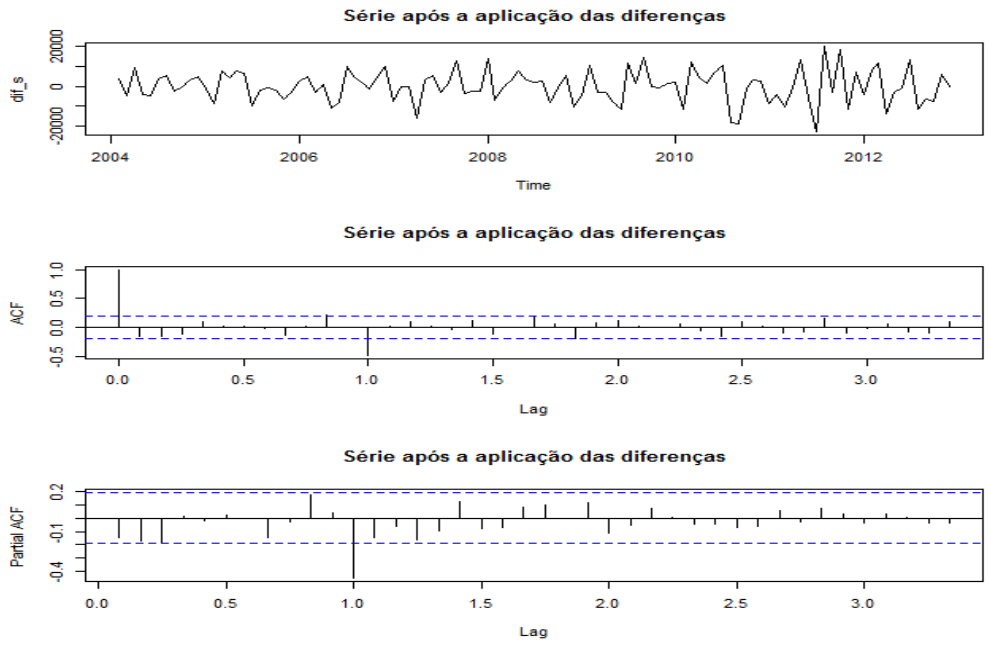

Figura 2: Série, ACF e PACF após aplicação das diferenças.

Como 1 *lag* múltiplo de 12 em ambos os correlogramas ainda foi significativo, parece existir sazonalidade estocástica. Portanto, simulou-se diversos modelos SARI $A\,(p,d,q)\,x\,(P,D,Q)_s$ com os valores de p e q variando de 0 a 3, e P e Q de 0 a 2. Devido às diferenças aplicadas, tem-se que d=1 e D=1. Os 4 modelos mais parcimoniosos foram selecionados e os resultados podem ser visualizados na Tabela 1.

Tabela 1: Resultados e Comparação dos modelos de previsão

N	SARIMA	AIC	p-valor (L.B)	p-valor (S.W)	RMSE	MAPE (%)
1	(1,1,1) x (0,1,1)	2190,64	0,8037	0,1451	11010,38	3,82
2	(0,1,3) x (0,1,1)	2190,65	0,8749	0,1824	11075,60	3,84
3	(0,1,3) x (2,1,1)	2190,84	0,9687	0,0898	13172,71	5,02
4	(0,1,2) x (0,1,1)	2192,13	0,7426	0,0701	11920,38	4,32

Os 4 modelos estimam resíduos que são ruído branco, visto que o p-valor para o teste Ljung-Box e Shapiro-Wilk foram superiores a 0,05. Nota-se que os modelos 1 e 2 são superiores e apresentam AIC, RMSE e MAPE bem próximos. O SARIMA $(1,1,1)\ x\ (0,1,1)_{12}$ possui uma pequena vantagem e além disso, não apresenta nenhum coeficiente não significativo, o que corrobora sua escolha. A Tabela 2 a estimativa dos parâmetros do modelo e a Tabela 3 a previsão para os 9 meses de 2013.

Tabela 2: Estatísticas para os coeficientes do SARIMA $(1,1,1)\ x\ (0,1,1)_{12}$

Parâmetros	Estimativa	Erro Padrão	t	p-valor
ϕ_1	0,5040	0,2205	2,2857	0,0223*
θ_1	-0,7728	0,1704	-4,5352	0,0000*
Θ_1	-0,6666	0,1064	-6,2650	0,0000*

* Significativo a 5% de probabilidade

Tabela 3: Previsão, Intervalos de Confiança (95%) e Resíduos Estimados.

Mês	L.I (5%)	Previsto	L.S (95%)	Real	Resíduo
Jan/2013	196.860	209.182	221.505	236.101	26.919
Fev/2013	180.560	195.825	211.090	201.189	5.364
Mar/2013	184.926	201.864	218.802	208.950	7.086
Abr/2013	167.484	185.626	203.768	192.805	7.179
Mai/2013	171.161	190.292	209.424	185.957	-4.335
Jun/2013	168.895	188.904	208.914	191.257	2.353

Jul/2013	188.257	209.077	229.896	210.503	1.426
Ago/2013	195.414	217.000	238.585	225.176	8.176
Set/2013	189.616	211.934	234.252	223.880	11.946

4 Conclusões

O modelo SARIMA $(1,1,1)\ x\ (0,1,1)_{12}$ obteve boas propriedades estatísticas e ajuste satisfatório aos dados de produção de leite cru do estado do Paraná (PR). Nota-se que apenas a previsão para o mês de Janeiro de 2013 excedeu os limites de confiança de 95%, o que não inviabiliza a capacidade preditiva do modelo. Logo, a metodologia Box & Jenkins mostrou-se apropriada para modelagem de dados de produção de leite, podendo ser utilizada para obtenção de previsões futuras e como ferramenta de auxilio para o desenvolvimento de planejamentos e tomadas de decisões, principalmente para produtores de leite.

Referências

[1] AKAIKE, H. **A new look ate the statistical model identification**, IEEE Transactions on Automatic Control, 1974.

[2] JENKINS, G. M.; WATTS, D. G. **Spectral analysis and its applications**. San Francisco: Holden-Day, 1968. 525p.

[3] LJUNG, G. M.; BOX, G. E. P. On a measure of lack of fit in time series models. **Biometrika**, v.65, p.297-303, 1978.

[4] MAKRIDAKIS, S.; WHEELWRIGHT, S.; HYNDMAN, R.J. **Forecasting: Methods and Applications**. John Wiley & Sons. 3a Edição. New York, 1998.

[5] MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. Associação Brasileira de Estatística. São Paulo: Edgard Blücher, 2.ed., 2006. 538p.

[6] **R DEVELOPMENT CORE TEAM. R: a language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing. 2012.

[7] REZENDE, José Luiz Pereira de.; COELHO JUNIOR, Luiz Moreira.; OLIVEIRA, Antônio Donizette de ; SÁFADI, Thelma. Análise dos preços de carvão vegetal em quatro regiões no estado de Minas Gerais. **CERNE (UFLA), Lavras, v. 11, n.3, p. 237-252, 2005.**

[8] TORMAN, V. B. L.; BIRCK, Alan Rodrigues.; RIBOLDI, João. Comparação dos Testes de Aderência à Normalidade Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises e Shapiro-Wilk por simulação. **In: 11 Anais do 50ª Simpósio de Estatística Aplicada à Experimentação Agronômica (SEAGRO), 2005.**