www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

ANÁLISE DA PRODUÇÃO DA CULTURA DO SOJA NO BRASIL ATRAVÉS DOS MODELOS ARIMA

ANALYSIS OF THE CULTURE OF SOYBEAN PRODUCTION IN BRAZIL THROUGH THE ARIMA MODELS

STEPHANIE RUSSO FABRIS[1], JONAS PEDRO FABRIS[2], ANGELA ISABEL DOS SANTOS DULLIUS[3]

[1]Graduanda em Design pela Universidade Federal de Sergipe, Brazil (sterussofabris@hotmail.com);
[2]Mestre em Engenharia Civil pela Universidade Federal de Santa Catarina, Brazil (jpfabris@hotmail.com);
[3] Professora do Departamento de Estatística da Universidade Federal de Santa Maria, Brasil (angeladullius@gmail.com)

Resumo

Neste estudo, mostraremos a aplicação da metodologia ARIMA na série representativa da produção da Cultura do Soja no Brasil no período de 1931 a 2009. A cultura da soja além de ser uma fonte abundante de aproveitamento alimentício geral e diversificado, o grão possui alto valor protéico (até 50% proteína), o que a torna uma das culturas que mais investimentos recebeu historicamente, sendo alvo de grande número de pesquisas visando melhorar sua qualidade e produtividade (DROS, 2004; MARION, 2004). Realizou-se, num primeiro instante, uma análise gráfica dos dados, observando-se o comportamento dos dados originais e da função de autocorrelação e a função de autocorrelação parcial. A obtenção do modelo mais adequado foi baseada na análise de gráficos e em testes estatísticos da própria metodologia, pelos quais foi possível determinar que o melhor modelo para a série da produção da Cultura do Soja no Brasil foi o modelo ARIMA(2,2,0) que apresentou um MAPE de 7,25%,obtido a partir do critério de validação

Palavras-chave: Soja; Modelos ARIMA, Previsão

Abstract

In this study, we show the application of the ARIMA methodology in the series representative of the production of Culture of Sugar Cane in Brazil in the period 1931 to 2009. The sugar cane in addition to being an abundant source of food utilization and diversified general, the grain has a high protein (up to 50% protein), which makes it one of the cultures that received more investments historically been the target of numerous research to improve quality and productivity (DROS, 2004, MARION, 2004). Was performed in a first moment, a graphical analysis of the data, observing the behavior of the original data and the autocorrelation function and partial autocorrelation function. Obtaining the most appropriate model was based on analyzing graphs and statistical tests of the methodology by which it was determined that the best model for the series production of Culture Soyben production in Brazil was the ARIMA (2,2,0) which had a MAPE equal to 7.25%, obtained from the validation criteria.

Key-words: Soybean; ARIMA Models; Forecasting

1. Introdução

A cultura da soja além de ser uma fonte abundante de aproveitamento alimentício geral e diversificado, o grão possui alto valor protéico (até 50% proteína), o que a torna uma das culturas que mais investimentos recebeu historicamente, sendo alvo de grande número de pesquisas visando melhorar sua qualidade e produtividade (DROS, 2004; MARION, 2004).

A partir da identificação dos modelos que melhor descrevem o comportamento da cultura a campo em uma determinada região, é possível inserir tais modelos em programas de simulação de produtividade, prever o impacto de mudanças climáticas sobre esta e, caso os eventos meteorológicos se comportem igual à média dos anos, indicar a melhor época de plantio para cada região. (ARAÚJO, 2008).

Entre os muitos exemplos de novos modelos quantitativos criados para simular a realidade e fazer previsões sobre o futuro destaca-se a metodologia ARIMA desenvolvida pelos professores Box e Jenkins, que serve para analisar o comportamento de variáveis através de series de tempo.

A análise de series de tempo, segundo a metodologia ARIMA (autoregressivos integrados médias móveis), tem como objetivo principal a realização de previsão. Essa metodologia permite que valores futuros de uma serie sejam previstos tomando por base apenas seus valores presentes e passados.

Este trabalho visa analisar através dos modelos ARIMA a produção anual da cultura do soja no Brasil no período de 1931 a 2009.

2. Revisão de Literatura

2.1 Cultivo do Soja

A soja que hoje cultivamos é muito diferente dos seus ancestrais, que eram plantas rasteiras que se desenvolviam na costa leste da Ásia, principalmente ao longo do rio Yangtse, na China. Sua evolução começou com o aparecimento de plantas oriundas de cruzamentos naturais entre duas espécies de soja selvagem que foram domesticadas e melhoradas por cientistas da antiga China (EMBRAPA, 2011).

As primeiras citações do grão aparecem no período entre 2883 e 2838 AC, quando a soja era considerada um grão sagrado, ao lado do arroz, do trigo, da cevada e do milheto. Na segunda década do século XX, o teor de óleo e proteína do grão começa a despertar o interesse das indústrias mundiais. No entanto, as tentativas de introdução comercial do cultivo do grão na Rússia, Inglaterra e Alemanha fracassaram, provavelmente, devido às condições climáticas desfavoráveis (EMBRAPA, 2011).

Muito tempo depois os ocidentais passaram a considerar a soja como alimento funcional, aquele que, além das funções nutricionais básicas, produz efeitos benéficos à saúde, sendo seguro para o consumo sem supervisão médica (ARAÚJO, 2008).

Conforme dados da EMBRAPA (2011), a soja pertence à classe das dicotiledôneas, família leguminosa e subfamília Papilionoides. A espécie cultivada é a Glycine Max Merril. O sistema radicular é pivotante, com a raiz principal bem desenvolvida e raízes secundárias em grande número, ricas em nódulo de bactérias Fhisobium Japonicum fixadoras de nitrogênio atmosférico.

No Brasil, a soja parece ter sido primeiramente introduzida na Bahia, em 1882. Em 1908 foi introduzida em São Paulo, por imigrantes japoneses, e em 1914 foi introduzida no Rio Grande do Sul pelo professor Craig, da Universidade Federal do Rio Grande do Sul. Foi no Rio Grande do Sul que a soja começou a ser cultivada em larga escala. O município de Santa Rosa foi o pólo de disseminação da cultura, que inicialmente expandiu-se pela região

das missões. Até meados dos anos 30, esta era a região produtora de soja (ARAÚJO, 2008; MUNDSTOCK e THOMAS, 2005)

O Brasil é o segundo maior produtor mundial de soja atrás apenas dos EUA. Na safra 2010, a cultura ocupou uma área de 24,6 milhões de hectares, o que totalizou uma produção de 75 milhões de toneladas. A produtividade média da soja brasileira foi de 3.106 kg por hectares (EMBRAPA 2011).

2.2 Séries Temporais

Uma série temporal é um conjunto de observações ordenadas no tempo, geralmente em intervalos eqüidistantes. Se o processo estocástico que gerou a série de observações é invariante com respeito ao tempo, diz-se que o processo é estacionário. Se as características do processo se alteram no decorrer do tempo, é chamado de estacionário (RUSSO E CAMARGO, 2008)

A importância do conhecimento da série ser ou não estacionária reside no fato de que quando se trabalha com uma série estacionária, se está em presença de uma função amostral do processo que tem a mesma forma em todos os instantes do tempo $t \in N$, acarretando na facilidade de obtenção de estimativa das características do processo.

Ao consideramos a evolução temporal do processo, mede-se a magnitude do evento que ocorre em determinado instante de tempo. A análise no domínio do tempo é baseada em um modelo paramétrico, utilizando-se as funções de autocovariância e autocorrelação. A autocorrelação é a autocovariância padronizada, serve para medirmos a extensão de um processo para o qual o valor tomado no tempo *t*, depende daquele tomado no tempo *t-k*. Define-se a autocorrelação de ordem k como (RUSSO E CAMARGO, 2008):

$$\rho_k = \frac{\gamma_k}{\gamma_o} = \frac{Cov[Z_t, Z_{t+k}]}{\sqrt{Var(Z_t)}Var(Z_{t=k})}$$

onde: $Var(Z_t) = Var(Z_{t=k}) = \gamma_0$ = variância do processo

$\rho_o = 1$ e $\rho_k = \rho_{-k}$.

A autocorrelação pode ser estendida. Se medirmos a correlação entre duas observações seriais Z_t e Z_{t+k}, eliminando a dependência dos termos intermediários, Z_{t+1}, Z_{t+k-1}, temos a autocorrelação parcial, representada por:

$Cor(Z_t, Z_{t+k} \mid Z_{t+1}, ..., Z_{t+k-1})$.

A função de autocorrelação pode ser obtida considerando-se um modelos de regressão para um processo estacionário com média zero. A variável dependente Z_{t+k} depende das variáveis Z_{t+k-1}, Z_{t+k-2}, ...

Assim, $Z_{t+k} = \phi_{k1} Z_{t+k-1} + \phi_{k2} Z_{t+k-2} + ... + \phi_{kk} Z_t + a_{t-k}$

onde: ϕ_{ki} - i-ésimo parâmetro da regressão

a_{t+k} - é o tremo de erro descorrelatado com Z_{t+k-j} para $j \geq 1$.

Modelos ARIMA

O processo gerador de uma série é identificado através do método da comparação. São apresentadas a seguir, duas representações do modelo linear (BOX ET AL 1994):

- *Representação em média móvel (MA):* admite-se que a observação atual de uma variável possa ser explicada através de uma soma ponderada de ruídos anteriores e de um ruído atual. Assim, o modelo linear geral é (MATOS, 200; RUSSO, 2006):

$$\tilde{Z}_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + ... \qquad \tilde{Z}_t = \mu + \psi(B) a_t$$

$$\tilde{Z}_t = a_t + \psi_1 B a_t + \psi_2 B^2 a_t + ...$$
$$\tilde{Z}_t = (1 + \psi_1 B + \psi_2 B^2 + ...) a_t$$
$$\tilde{Z}_t = \psi(B) a_t$$

onde $\psi(B) = \sum_{j=0}^{\infty} \psi_j{}^j$, com $\psi_0 = 1$, é a série de pesos usados na explicação da variável.

- *Representação auto-regressiva (AR):* supõe-se que a observação presente da variável possa ser explicada por uma soma ponderada das observações anteriores da mesma variável e de um erro atual, obtendo-se assim, a representação auto-regressiva do Modelo Linear Geral:

$$\tilde{Z}_t = a_t + \pi_1 \tilde{Z}_{t-1} + \pi_2 \tilde{Z}_{t-2} + ...$$
$$\tilde{Z}_t = a_t + \pi_1 B \tilde{Z}_t + \pi_2 B^2 \tilde{Z} + ...$$
$$\tilde{Z}_t = (1 - \pi_1 B - \pi_2 B^2 - ...)$$
$$\tilde{Z}_t = a_t \pi(B) \tilde{Z}_t$$

onde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$ é a série dos pesos usados na explicação da variável.

Quando se tem uma série não estacionária, deve-se antes de ajustar um modelo estacionário (MA, AR, ARMA) eliminar sua tendência, para isto, pode-se usar a diferenciação discreta.

Substituindo-se $\vec{Z}_t$ por $w_t = \nabla^d Z_t$ obtém-se um modelo capaz de descrever certos tipos de séries não estacionárias. Tais modelos são chamados de "integrados" porque o modelo estacionário ajustado a série $\{ w_t \}$ tem que ser somado para ajustar-se aos dados estacionários $\{ \tilde{Z}_T \}$. Como $w_t = \nabla^d Z_t$ onde $\nabla \tilde{Z}_t = \tilde{Z}_t - \tilde{Z}_{t-1}$ o modelo $w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + ... + \phi_p w_{t-p} + a_t$ é dito modelo auto-regressivo integrado ARI(p,d).

O modelo $w_t = a_t + \phi_1 w_{t-1} + ... + \phi_q w_{t-q}$ é chamado de modelo integrado de média móvel IMA(d,q).

E o modelo

$$\phi(B) \nabla^d Z_t = \theta(B) a_t$$
$$\phi(B)(1-b)^d Z_t = \theta(B) a_t$$
$$\Omega(B) Z_t = \theta(B) a_t$$

é chamado de modelo auto-regressivo integrado de média móvel ARIMA(p,d,q).

Esta filtragem é repetida quantas vezes forem necessárias, até que se obtenha na saída um processo com as características necessárias para representar o processo não-estacionário homogêneo.

Se $w_t = \nabla_s{}^d Z_t$ é estacionária, pode-se representar w_t por um modelo ARMA(p,q), ou seja, $\phi(B) w_t = \theta(B) a_t$ (1)

Se w_t é uma diferença de Z_t, então Z_t é uma integral (soma) de w_t, daí diz-se que Z_t segue um modelo Auto-Regressivo-Integrado-Médias Móveis, ou modelo ARIMA(p,d,q), assim: $\phi(B) = \nabla^d Z_t = \theta(B)\, a_t$ de ordem (p,d,q) (2)

e escrevemos ARIMA(p,d,q) se p e q são as ordens de $\phi(B)$ e $\theta(B)$, respectivamente, no modelo (1) todas as raízes de (B) estão fora do círculo unitário.

Escrever (2) é equivalente a escrever:

$\xi(B)\, Z_t = \theta(B)\, a_t$, (3)

onde $\phi(B)$ é o operador auto-regressivo não- estacionário de ordem p + d, com d raízes iguais a um (sobre o círculo unitário) e as restantes p fora do círculo unitário ou seja:

$\xi(B) = \phi(B)\, \nabla^d = \phi(B)\,(1 - B)^d$ (4)

Portanto o modelo (4) supõem que a d-ésima diferença da série Z_t pode ser representada por um modelo ARMA(p,q), estacionário e inversível. Na maioria dos casos usuais, d = 1 ou d = 2 que correspondem a dois casos comuns de séries não-estacionárias homogêneas.

3. Resultados e Discussões

Os dados analisados referem-se a produção da cultura do soja no Brasil no período de 1931 a 2009. Realizou-se, num primeiro instante, uma análise gráfica dos dados, onde suspeita-se que as observações não são independentes, e a figura 2 apóia essa convicção.

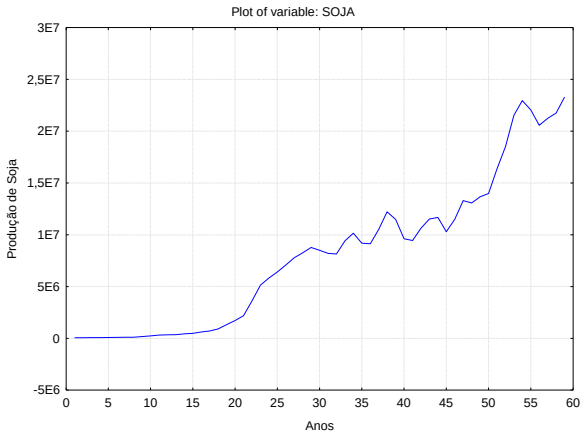

Figura 2 - Leitura diária dos dados

A Figura 2 mostra a grande variabilidade dos dados, pressupondo que a série não seja estacionária.

Identificação da estrutura do modelo

Analisando os valores da série (Figura 3 e 4) observou-se que se trata de uma série não estacionária, precisando diferenciá-la para torná-la estacionária.

Obtenção dos valores da ordem (p,d,q)

Para se obter a ordem (p,d,q) , a análise foi efetuada através das funções de autocorrelação (ACF) e autocorrelação parcial (PACF) da série.

A seguir apresentamos os gráficos da função de autocorrelação e da função de autocorrelação parcial dos dados referente produção da cultura do soja.

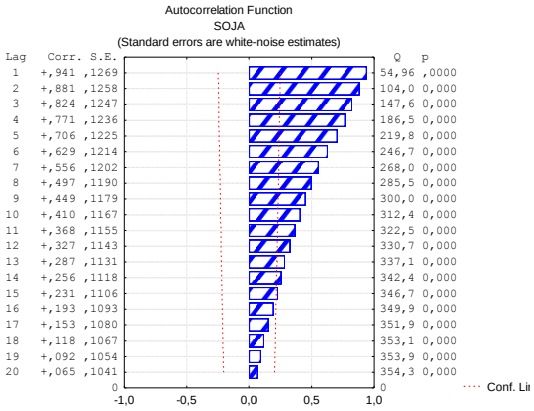

Figura 3 - Coeficientes da função de autocorrelação

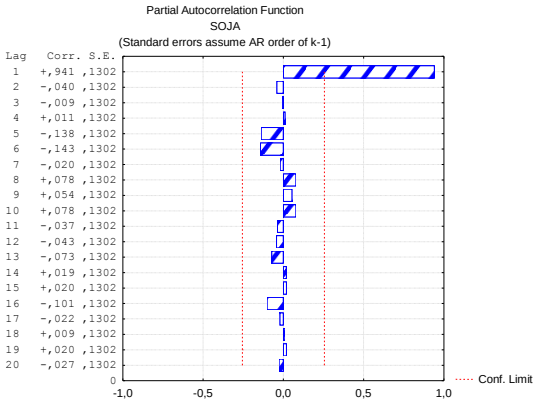

Figura 4 - Coeficientes da função de autocorrelação parcial

Nota-se pelas Figuras 3 e 4 que os dados são altamente correlacionados. Os coeficientes de autocorrelação excedem os dois erros padrões. A ACF da série não apresenta nenhuma queda para zero, confirmando a inexistência da componente sazonal. Antes de ajustar-se os dados devemos remover a autocorrelação dos dados.

Para encontrar um conjunto de dados independentes, normalmente distribuídos, deve-se modelar a estrutura da série e depois deve-se fazer a previsão dos dados. Com os dados da produção da Soja foi necessário fazer uma diferenciação (*d=1*).

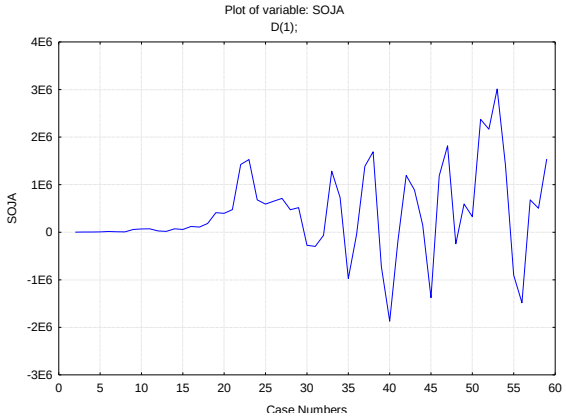

Figura 5 – Gráfico da diferenciação (*d*=1)

Estimação dos parâmetros do modelo

Os parâmetros do melhor modelo para a série da Produção do Soja no Brasil foi utilizado os modelos ARIMA. Vários modelos foram modelados e os melhores modelos encontram-se na Tabela 1. A análise conjunta das ACF e PACF permitiu estabelecer o seguinte modelo: ARIMA (2,1,0). Após várias tentativas, este foi o melhor modelo encontrado, conforme mostra a Tabela 1.

Foi verificado que todos os modelos são significantes, mas pelo critério de validação de ajuste MAPE, escolheu-se o modelo ARIMA(2,1,0), o qual possui o menor MAPE igual a 4,66%.

Tabela 1 – Sumário dos Parâmetros dos Modelos

Modelo	Parâmetro	Erro Padrão	teste *t*	p-value	Mape (%)
(1,1,0)	0,4992	0,1191	4,1907	0,0000	11,53
(2,1,0)	0,655953	0,130494	5,02670	0,000005	4,66
	-0,323205	0,130803	-2,47093	0,016545	

Análise da autocorrelação dos resíduos

Para validar que a autocorrelação tenha sido removida dos dados, os coeficientes de autocorrelação foram definidos pelo modelo ARIMA (2,1,0). Os resultados são mostrados nas Figuras 6 e 7.

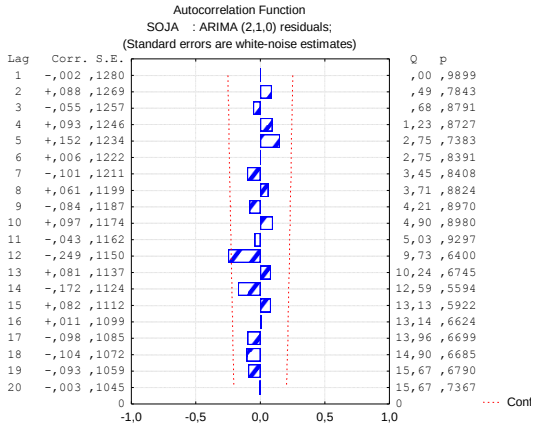

Figura 6 - Coeficientes da função de autocorrelação dos dados ajustados

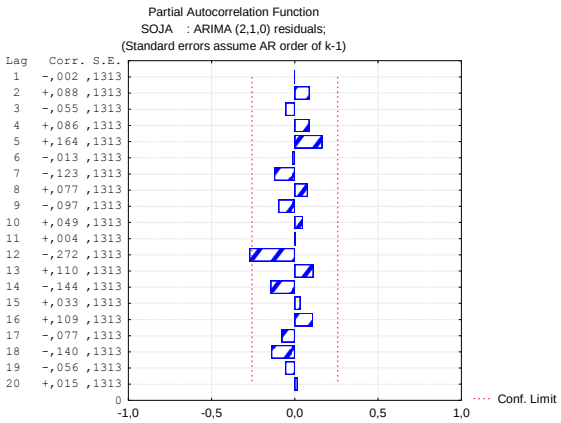

Figura 7 - Coeficientes de função de autocorrelação parcial dos dados ajustados

As Figura 6 e 7 mostra que os dados são independentes de observação para observação, pois pode-se observar que todos os lags estão dentro dos limites de confiança.

O gráfico da probabilidade half normal dos resíduos representados pela Figura 8 mostra, claramente, que os resíduos formam um processo de ruído branco.

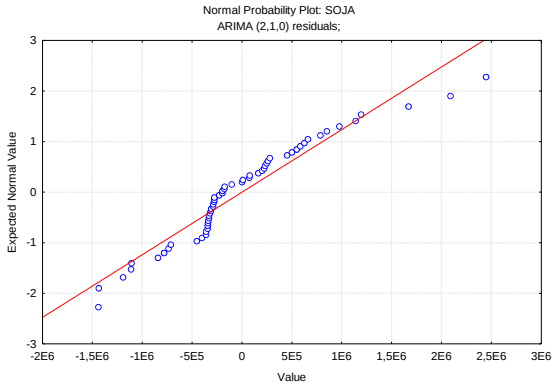

Figura 8 – Gráfico da Half - Normal

Previsão

Previsões são extrapolações obtidas através das funções de previsão, para além do

período no qual o modelo foi estimado. A previsão exposta na Tabela 2 estabelece os valores futuros da série fazendo uma relação dos valores observados e previstos.

Na Tabela 2, vemos os valores previstos para os próximos 5 anos para a produção da Cultura do Soja no Brasil.

Tabela 2. Tabela dos valores previstos

Previsto	Observado	Resíduo/Observado
22899480	22047349	0,0387
22411399	20565279	0,0898
22107205	21246302	0,0405
22065418	21750468	0,0145
22136325	23290696	0,0496

5. CONCLUSÃO

Após a análise do comportamento dos dados reais e das funções de autocorrelação, o melhor modelo encontrado para a série produção da cultura do soja no Brasil foi o modelo ARIMA(2,1,0). É importante ressaltar que para a escolha do melhor modelo e para fins de previsão utilizou-se o MAPE através do qual obteve-se um erro de 4,66%.

Contudo o que foi estudado pode-se afirmar que o modelo encontrado é adequado para a série, pois mantém os dados dentro dos limites de controle. Portanto, é possível constatar que a metodologia ARIMA satisfez os requisitos para a escolha do melhor modelo para a série produção da Cultura do Soja no Brasil.

6. REFERÊNCIAS

BOX, G. E. P ; JENKINS, G.M.; REINSEL,G. C. **Times series analysis: forecasting and control**, 3ª Ed. San Francisco: Holden –Day, 1994.

DROS, J. M. **Administrando os avanços da soja**: dois cenários de expansão do cultivo de soja na América do Sul. Amsterdã: AIDEnvironment, 2004. 71p.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA (EMBRAPA). **Soja.** 2011 Acessado em: 20/12/2011. Disponível em: www.embrapa.gov.br.

MATOS, O. C. **Econometria Básica**. 3ª ed. São Paulo: Atlas. 2000.

MARION, E. **Parâmetros hídricos para estimativa do rendimento de grãos de soja**. Florianópolis. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina. 2004

MUNDSTOCK, C. M.; THOMAS, A. L. **Soja: fatores que afetam o crescimento e o rendimento de grãos**. Porto Alegre: Departamento de Plantas de Lavoura da Universidade Federal do Rio Grande do Sul, 2005.

RUSSO, S. L.; CAMARGO, M. E. **Função de transferência: uma técnica complementar na aplicação de gráficos de controle.** Revista Eletrônica Produção & Engenharia, v. 1, n. 1, p. 95-106, set./dez. 2008.

RUSSO, S. L.; JARDIM, I.; KLIMAN, P.; KLIDZIO, R. **Estudo comparativo do fluxo de caminhões nos portos de Uruguaiana e Foz do Iguaçu.** Bauru: XIII SIMPEP. 2006.