Previsão Utilizando Processos SARFIMA com Estimação dos Parâmetros via MCMC

Cleber Bisognin[1] e Letícia Menegotto[1]

[1]Departamento de Estatística - Universidade Federal do Rio grande do Sul - UFRGS

RESUMO

Neste trabalho estamos interessados em analisar séries temporais com as características de longa dependência e sazonalidade e prever seus futuros valores. Para estudar tais séries utilizamos os modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$ que conseguem modelar séries com ambas as característica. A série temporal da média mensal da umidade relativa do ar, de Santa Maria, Rio grande do Sul possui ambas as características. Para estimação dos parâmetros do modelo SARFIMA$(p,d,q)\times(P,D,Q)_s$ utilizamos o estimador de Whittle (1951) calculado sob frequências que são obtidas utilizando o algoritmo de Metropolis-Hastings. Trata-se de uma nova metodologia para a estimação dos parâmetros dos modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$ com inovações gaussianas. Esta metodologia somente foi utilizada para modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$ com inovações α-estáveis. Além disso, para processos SARFIMA$(p,d,q)\times(P,D,Q)_s$ com inovações gaussianas, propomos a previsão de erro quadrático médio mínimo. Para avaliar as previsões propomos o intervalo de confiança de previsão a $100(1-\gamma)\%$ de confiança, baseado nas variâncias teórica e amostral do erro de previsão. A metodologia de previsão, incluindo o intervalo de confiança, é uma inovação para analisar séries temporais com longa dependência e sazonalidade. O modelo selecionado para a série temporal mostrou-se adequado dado que as predições mantiveram-se dentro do intervalo de confiança a 95% de confiança durante todo o período de setembro de 2012 a setembro de 2017. Por fim, realizamos previsões para o período de outubro de 2017 a setembro de 2018.

Palavras chave: Umidade relativa do ar, Longa dependência, Sazonalidade, Modelagem Estatística, Previsão.

1 INTRODUÇÃO

Para o desenvolvimento de diversos setores da atividade humana - tal como o setor agrícola ou econômico - é preciso entender o comportamento dos fenômenos climáticas da região em estudo. Muitos destes fenômenos, possuem a propriedade de longa dependência e, também a característica de repetir-se durante um certo período de tempo fixo, ou seja, apresentam sazonalidade.

Para estudar os fenômenos com estas características, Porter-Hudak (1990) propõem os processos SARFIMA$(p,d,q)\times(P,D,Q)_s$. Bisognin; Lopes (2009) demonstram várias propriedades dos processos SARFIMA$(p,d,q)\times(P,D,Q)_s$.

Um destes fenômenos meteorológicos que apresentam tais características é a umidade relativa do ar (ou simplesmente umidade relativa). A umidade relativa do ar é expressa como a razão entre a quantidade efetiva de vapor d'água no ar e a quantidade máxima de vapor d'água que a mesma quantidade de ar poderia conter, se estivesse saturada desta substância, em determinada temperatura (VIRGENS et al., 2009).

Séries temporais possibilitam analisar estes fenômenos, ao longo do tempo, e detectar possíveis mudanças que possam ocorrer em um futuro próximo. O estudo de séries temporais também possibilita, através do comportamento passado, ajustar um modelo matemático que nos permite realizar previsões.

Desta forma, o objetivo deste trabalho é ajustar um modelo SARFIMA$(p,d,q)\times(P,D,Q)_s$, onde a estimação dos parâmetros do modelo é feita via estimador de Whittle (1951), calculado utilizando o algoritmo de Metropolis-Hastings e calcular as previsões da média mensal da umidade relativa do ar, em Santa Maria, Estado do Rio Grande do Sul.

2 METODOLOGIA

Nesta seção são apresentados os modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$ usados para realizar a análise e previsão da série temporal da média mensal da umidade relativa do ar de Santa Maria, Rio Grande do Sul. Os dados são provenientes do BDMEP - Banco de Dados Meteorológicos para Ensino e Pesquisa mantido pelo INMET - Instituto Nacional de Meteorologia (`http://www.inmet.gov.br`).O BDMEP - Banco de Dados Meteorológicos para Ensino e Pesquisa, que é um banco de dados para apoiar as atividades de ensino e pesquisa e outras aplicações em meteorologia, hidrologia, recursos hídricos, saúde pública, meio ambiente, etc. Serão utilizadas as 566 observações mensais disponíveis, de janeiro de 1961 a setembro de 2017. Os dados foram acessados em 06/11/2017. O uso deste período se deve ao fato de o período de dados mais completo disponíveis na internet e que foram encontrados pelos autores.

A seguir definimos os modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$.

Definição 1. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico satisfazendo a equação

$$\phi(\mathcal{B})\Phi(\mathcal{B}^s)(1-\mathcal{B})^d(1-\mathcal{B}^s)^D(X_t-\mu)=\theta(\mathcal{B})\Theta(\mathcal{B}^s)\varepsilon_t, \tag{1}$$

onde μ é a média do processo, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é o processo ruído branco, $s\in\mathbb{N}$ é a sazonalidade, $\mathcal{B}$ é o operador de *defasagem ou* de *retardo*, isto é, $\mathcal{B}^j(X_t)=X_{t-j}$ e $\mathcal{B}^{sj}(X_t)=X_{t-sj}$, para $j\in\mathbb{N}$, ∇^d e ∇_s^D são os operadores, respectivamente, diferença e diferença sazonal, $\phi(\cdot)$ e $\theta(\cdot)$, $\Phi(\cdot)$ e $\Theta(\cdot)$ são os polinômios de ordem p, q, P e Q, respectivamente, definidos por

$$\phi(z) = \sum_{\ell=0}^{p}(-\phi_\ell)\,z^\ell, \quad \theta(z)=\sum_{m=0}^{q}(-\theta_m)\,z^m, \tag{2}$$

$$\Phi(z) = \sum_{r=0}^{P}(-\Phi_r)\,z^r, \quad \Theta(z)=\sum_{l=0}^{Q}(-\Theta_l)\,z^l, \tag{3}$$

onde ϕ_ℓ, $1\le\ell\le p$, θ_m, $1\le m\le q$, Φ_r, $1\le r\le P$, e Θ_l, $1\le l\le Q$, são constantes reais e $\phi_0=\Phi_0=-1=\theta_0=\Theta_0$. Então, $\{X_t\}_{t\in\mathbb{Z}}$ é um *processo sazonal auto-regressivo fracionariamente integrado de média móvel de ordem* $(p,d,q)\times(P,D,Q)_s$ com sazonalidade s, denotado por SARFIMA$(p,d,q)\times(P,D,Q)_s$, onde d e D são, respectivamente, *o grau de diferenciação e o grau de diferenciação sazonal*.

Para maiores detalhes sobre estes modelos, tais como a expressão da função densidade espectral, seu comportamento próximo às frequências sazonais, a estacionariedade, a dependência intermediária e longa e a função de autocovariância, ver Bisognin e Lopes (2009).

Para a estimação dos parâmetros dos processos SARFIMA$(p,d,q)\times(P,D,Q)_s$ utilizamos o estimador de verossimilhança aproximado de Whittle (1951) baseado em Cadeias de Markov. Considere $\boldsymbol{\eta}=(\sigma_\varepsilon^2,d,D,\boldsymbol{\Phi},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\Theta})$ o vetor de parâmetros a ser estimado. Seja C uma constante tal que $C=\int_{-\pi}^{\pi}I_n(\lambda)d\lambda$. Assim, assim temos que a equação

$$\sigma_n^2(\boldsymbol{\eta})=\int_{-\pi}^{\pi}\frac{I_n(\lambda)}{f_X(\lambda,\boldsymbol{\eta})}d\lambda=C\int_{-\pi}^{\pi}\frac{f(\lambda)}{f_X(\lambda,\boldsymbol{\eta})}d\lambda=C\,\mathbb{E}\left(\frac{1}{f_X(\lambda,\boldsymbol{\eta})}\right), \tag{4}$$

onde $f(\lambda)=\frac{1}{C}I_n(\lambda)$ é a função densidade em $[-\pi,\pi]$. O valor esperado em (4) pode ser aproximado pela média empírica

$$\overline{\sigma}_n^2(\boldsymbol{\eta})=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{f_X(\lambda_j,\boldsymbol{\eta})}, \tag{5}$$

onde N é suficientemente grande para satisfazer a lei dos grandes números. Aqui consideramos $N=15000$ e o algoritmo de Metropolis-Hastings para gerar a amostra $(\lambda_1,\cdots,\lambda_N)$. Denotamos este estimador por WMCMC.

O objetivo deste trabalho é analisar séries temporais com a característica de longa dependência e sazonalidade e prever os seus futuros valores utilizando os modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$. Para tanto, apresentamos a seguir, a previsão de erro quadrático médio mínimo. Para avaliarmos as previões obtidas, utilizamos o intervalo de confiança de previsão a $100(1-\gamma)\%$ de confiança o qual é baseado na distribuição e nas variâncias teórica e amostral dos erros de previsão. Tais resultados são apresentados a seguir.

Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo *SARFIMA*$(p,d,q)\times(P,D,Q)_s$ causal e inversível definido na equação (1), com média igual a zero e sazonalidade $s\in\mathbb{N}$. Então, para todo $h\geqslant 1$, a previsão de erro quadrático médio mínimo é dada por

$$\widehat{X}_n(h) = -\sum_{k\in\mathbb{N}} \pi_k \ \widehat{X}_n(h-k), \tag{6}$$

onde os coeficientes $\{\pi_k\}_{k\in\mathbb{Z}_{\geqslant}}$ são os coeficientes da representação autoregressiva infinita.

As variâncias teórica e amostral do erro de previsão são dadas, respectivamente, por

$$Var(e_n(h)) = \sigma_\varepsilon^2 \sum_{k=0}^{h-1} \psi_k^2, \qquad \widehat{Var}(e_n(h)) = \widehat{\sigma}_\varepsilon^2 \sum_{k=0}^{h-1} \widehat{\psi}_k^2, \tag{7}$$

onde os coeficientes $\{\psi_k\}_{k=0}^{h-1}$ são os coeficientes da representação média móvel infinita. Os coeficientes $\{\widehat{\psi}_k\}_{k=0}^{h-1}$ são obtidos quando substituímos os parâmetros teóricos no modelo pelos seus respectivos valores estimados e σ_ε^2 e $\widehat{\sigma}_\varepsilon^2$ são, respectivamente, a variância e a variância estimada do processo $\{\varepsilon_t\}_{t\in\mathbb{Z}}$.

Agora suponha que o processo $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ seja Gaussiano com $\mathbb{E}(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma_\varepsilon^2$ e $\mathbb{E}(\varepsilon_t\varepsilon_k) = 0$, para todo $t \neq k$. O intervalo de previsão a $100(1-\gamma)\%$ de confiança para X_{n+h} é dado por

$$\widehat{X}_n(h) - z_{\frac{\gamma}{2}}\widehat{\sigma}_\varepsilon \left(\sum_{k=0}^{h-1}\widehat{\psi}_k^2\right)^{\frac{1}{2}} \leq X_{n+h} \leq \widehat{X}_n(h) + z_{\frac{\gamma}{2}}\widehat{\sigma}_\varepsilon \left(\sum_{k=0}^{h-1}\widehat{\psi}_k^2\right)^{\frac{1}{2}}, \tag{8}$$

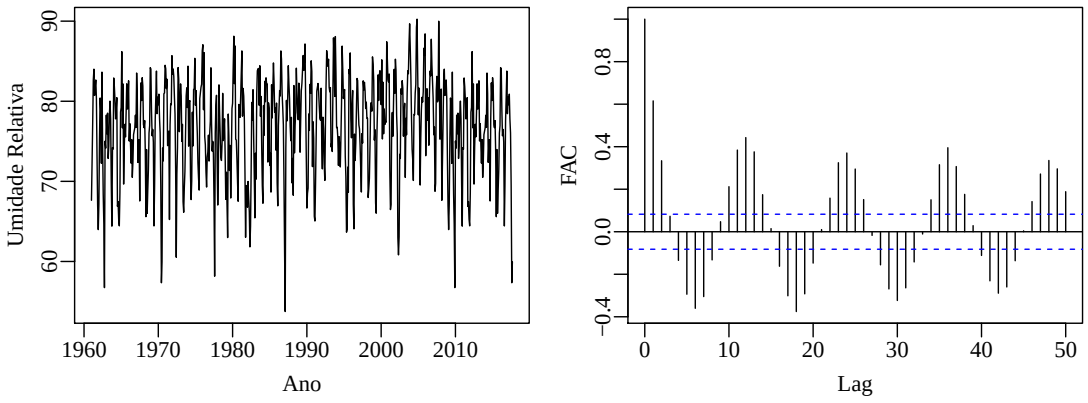
onde $z_{\frac{\gamma}{2}}$ é o valor tal que $\mathbb{P}(Z \geqslant z_{\frac{\gamma}{2}}) = \frac{\gamma}{2}$, com $Z \sim \mathcal{N}(0,1)$.

As rotina de estimação dos parâmetros e previsão dos modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$ foram implementadas pelos autores no *software* R.

Uma vez que os parâmetros dos modelos são estimados, os resíduos do modelo são analisados, ou seja, foi aplicado o teste Ljung-Box para examinar se os resíduos são não correlacionados. Para isto, foi utilizada a rotina *Box.test*. A verificação da acurácia da previsão do modelo foi realizada através do MPE (erro percentual médio) e pelo MAPE (média dos erros percentuais absolutos). As medidas foram calculadas utilizando-se a rotina *accuracy* do pacote *forecast*.

3 RESULTADOS E DISCUSSÕES

A Figura 1 apresenta o gráfico das séries temporais. Podemos perceber, pelo gráfico da série e da função de autocorrelação amostral, que a série temporal apresenta sazonalidade $s = 12$. Foi aplicado o teste de Hylleberg, Engle, Granger and Yoo (HEGY),para verificar se a série apresenta raiz unitária sazonal (hipótese nula). O teste apresentou p-valor = 0.01. Para este teste, utilizamos a rotina *hegy.test*, do pacote *uroot* do *software* R. Também foi aplicado o teste de Dickey-Fuller, cuja hipótese nula é de que a séries temporal analisada é não estacionária. O p-valor resultante foi 0.01. Para este teste, utilizamos a rotina *adf.test*, do pacote *tseries* do *software* R.



(a) Série Temporal (b) FAC

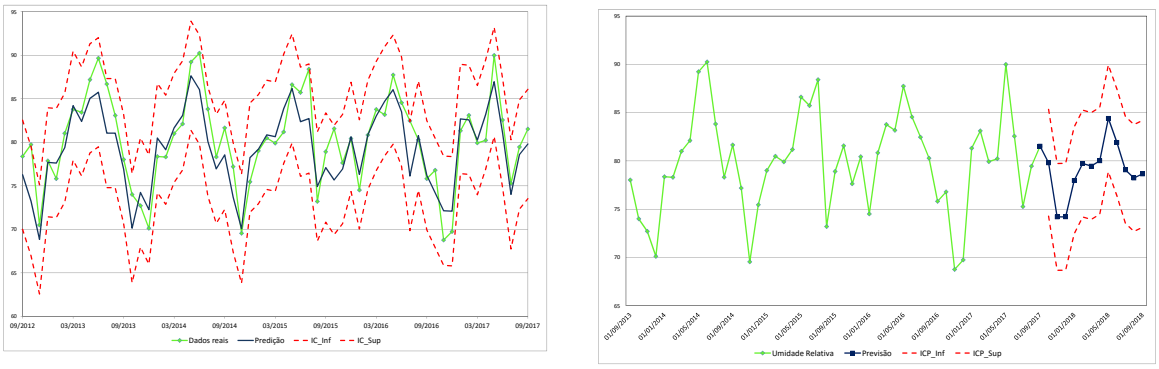
Figura 1: Série Temporal da média mensal da umidade relativa do ar de Santa Maria, Rio Grande do Sul: (a) Série Teporal e (b) função de autocorrelação amostral.
Fonte: Autores.

De acordo com os critérios de informação de Akaike (AIC) e Bayesiano (BIC) e pela log-verossimilhança, foi selecionado o modelo SARFIMA$(p,d,q)\times(P,D,Q)_s$, com $P = 1 = p$, $q = 0 = Q$, $\widehat{d} = 0,000$, $\widehat{D} = 0,3130$, $\widehat{\phi}_1 = 0,5335$, $\widehat{\Phi}_1 = -0,1166$, $\widehat{\sigma}_\varepsilon^2 = 20,55$. Realizando o teste de resíduos de Box-Pierce,

p-valor = 0.6108, e analisando as funções de autocorrelação e autocorrelação parcial amostrais, verificados a adequabilidade do modelo. Também foi realizado o teste de Kolmogorov-Smirnov, para verificar a normalidade dos resíduos. O p-valor resultante foi de $0,2202$.

Após o cálculo das predições do modelo, formam calculadas as medidas de acurácia, cujos valores são: RMSE (2,8508), MAE (2,2507), MPE (0,0014) e MAPE (2,9732).

A Figura 2 apresenta o gráfico da predição e previsão da Série Temporal da média mensal da umidade relativa do ar de Santa Maria, Rio Grande do Sul, com os intervalos de confiança a 95%. Analisando a predição (Figura 2(a)), a predição da média mensal da umidade relativa do ar, de Santa Maria, para o período de setembro de 2012 a setembro de 2017, manteve-se dentro do intervalo de confiança a 95% de confiança, o que podemos concluir que a adequabilidade do modelo. Na Figura 2(b) apresentamos a previsão para os dados durante o período de outubro de 2017 a setembro de 2018 com os intervalos de confiança de 95% de confiança.



(a) (b)

Figura 2: Série Temporal da média mensal da umidade relativa do ar de Santa Maria, Rio Grande do Sul: (a) dados reais, predição e intervalos de confiança a 95% para a previsão no período de setembro de 2012 a setembro de 2017 e (b) dados reais, Previsão e intervalos de confiança a 95% para a previsão (período: outubro de 2017 a setembro de 2018.

Fonte: Autores.

4 CONCLUSÃO

Neste trabalho propomos as previsões de erro quadrático médio mínimo para os modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$ cujos parâmetros foram estimados utilizando o estimador de WMCMC. Ao concluirmos nossa análise, podemos verificar que o modelo SARFIMA$(p,d,q)\times(P,D,Q)_s$, com $P=1=p$, $q=0=Q$, $\widehat{d}=0,000$, $\widehat{D}=0,3130$, $\widehat{\phi}_1=0,5335$, $\widehat{\Phi}_1=-0,1166$, $\widehat{\sigma}^2_\varepsilon=20,55$, mostrou-se adequado para previsão da média mensal da umidade relativa do ar de Santa Maria, Rio Grande do Sul. Tal conclusão está baseada na análise de resíduo, no cálculo das medidas de acurácia utilizando as predições do modelo e através do intervalo de confiança de previsão. Na Figura 2(a) é possível constatar a boa predição do modelo. Assim, os modelos SARFIMA$(p,d,q)\times(P,D,Q)_s$ apresentam-se como uma boa alternativa para estudo de séries temporais com as características de longa dependência e sazonalidade.

Referências

[1] BISOGNIN, C.; LOPES, S. R. C. Properties of seasonal long memory processes. **Mathematical and Computer Modelling**, v. 49, n. 9, p. 1837-1851, 2009.

[2] HYLLEBERG, S.; ENGLE, R.; GRANGER, C.; YOO, B. Seasonal integration and cointegration. **Journal of econometrics**, v. 44, n. 1, p. 215-238, 1990.

[3] PORTER-HUDAK, S. An application of the seasonal fractionally differenced model to the monetary aggregates. **Journal of the American Statistical Association**, v. 85, n. 410, p. 338-344, 1990.

[4] VIRGENS Fº, J. S.; LEITE, M. L.; FRANCO, J. R.; KORELO, M. Modelo computacional estocástico para simulação de séries climáticas diárias de umidade relativa do ar, baseado na parametrização dinâmica das distribuições de probabilidade decorrente da retroalimentação de dados. **Revista Brasileira de Climatologia**, v. 5, p. 133-151.

[5] WHITTLE, P. **Hypothesis Testing in Time Series Analysis**. New York, Hafner, 1951.