

PREVISÃO DOS VALORES DO CUSTO UNITÁRIO BÁSICO ATRAVÉS DOS MODELOS ARIMA E GARCH

MOREIRA, Séphora Fernandes¹

Estudante de Engenharia Civil,
Universidade Federal do Cariri

sephora.fernandes@aluno.ufca.edu.br

FIRMINO, Paulo Renato Alves²

Centro de Ciência e Tecnologia,
Universidade Federal do Cariri

paulo.firmino@ufca.edu.br

Resumo

A estatística-computacional vem aprimorando o campo da modelagem e previsão de séries temporais. Um dos métodos atuais mais utilizados para melhorar os resultados de preditores é a otimização de modelos individuais e suas combinações. Este trabalho busca encontrar a parcimônia, no sentido de desempenho e sofisticação, de modelos autoregressivos e de médias móveis integrados (ARIMA), estudar o desempenho do modelo heteroscedástico condicional autoregressivo generalizado (GARCH) para séries temporais. Além disso, a combinação baseada na média simples é considerada. O trabalho assume, como caso de estudo, o custo unitário básico da construção civil do Brasil.

Palavras-chave: Séries Temporais. ARIMA. GARCH.

1 INTRODUÇÃO

Uma série temporal pode ser retratada como um conjunto de dados que representam uma variável ao longo de um período, em intervalos de tempo regulares, de forma que se possa observar uma inter-relação entre os pontos da mesma. Vale salientar que os dados que compõem a série são ordenados a partir de um período fixo, ou seja, demonstrando regularidade nos períodos observados (PICHILIANI, 2012)

Atualmente, os campos de pesquisa e estudo da engenharia vêm buscando demonstrar que a estatística computacional pode ser usada para resolver problemas, bem como ajudar no planejamento de novos produtos e sistemas (BRUM,2006), o que pode acarretar a melhoria de projetos que já existem ou que serão implantados, planejando, desenvolvendo e otimizando os processos de produção. A partir dessa visão interdisciplinar, criou-se uma medida de estimativa de Custo Unitário Básico (CUB), o qual afere valores de construção nas regiões do Brasil, medido mensalmente.

Em estudos que envolvem séries temporais, como o CUB, fatores como sazonalidade e tendência, são componentes de interesse, uma vez que os mesmos interferem de forma direta nos resultados finais do sistema subjacente, podendo alterar de forma significativa a visão que o engenheiro civil terá sobre a área de construção em uma determinada cidade, em um determinado período.

Quando se estuda a modelagem de séries temporais, a combinação de modelos pode melhorar significativamente a qualidade das estimativas (CLEMEN, 1989). Como ferramentas de implementação, à linguagem R tem ganhado destaque, envolvendo pacotes e funções que possibilitam a obtenção de modelos de acordo com algoritmos de

1 Bolsista PIBIC/CNPq no Programa Institucional de Iniciação Científica e Tecnológica da UFCA.

2 Docente Associado da UFCA. Orientador do bolsista.

otimização, tais como o *forecast* e *acf*, entre outros (CASTAÑEDA, 2017).

Entre os modelos de previsão podem-se ainda destacar o ARIMA (de *Autoregressive Integrated Moving Average*) e GARCH (*modelo heteroscedástico condicional autoregressivo generalizado*). O ARIMA foi inicialmente estudado em 1976 por George Box e Gwilym Jenkins, sendo usualmente utilizado para séries temporais que não dão sinais de estacionariedade. GARCH, por sua vez, possibilitam a modelagem de séries temporais financeiras, uma vez que ele é especialista em dados com volatilidade.

2 DESENVOLVIMENTO

2.1. O CUB é usado pelo campo da engenharia civil para auxiliar a construção de casas, comércio e imóveis populares, uma vez que ele mostra dados acerca de valores e taxas dos terrenos em localidades do Brasil. O mesmo tem embasamento legal, descrito na lei federal 4.591 de 16 de dezembro de 1964, bem como embasamento técnico, com normas (como a ABNT NBR 12721:2006) que auxiliam os cálculos e a apresentação dados obtidos para apuração do CUB. Especificamente, a como a ABNT NBR 12721:2006 preconiza que todos os dados, como mão de obra, materiais e despesas diversas, devem ser postos em consideração, pois só assim, o CUB poderá refletir o preço real de uma obra. Os dados são retirados da população na forma de pequenas amostras e utilizam a tabela Student para cálculo mensal dos resultados finais do indicador.

2.2. MODELOS DE PREVISÃO (ARIMA, GARCH)

Uma série temporal segue padrões temporais, em ordem cronológica, podendo ser discretas ou contínuas (MORETTIN, 2006). A Eq. 1 sintetiza o formato de série temporal abordada neste trabalho:

$$\mathbf{y}_t = \{\mathbf{y}_t \in \mathbb{R} \mid \mathbf{t} = \mathbf{1}, \mathbf{2}, \mathbf{3}, \dots, \mathbf{n}\}, \mathbf{n} \in \mathbb{N} \quad (1)$$

em que t é o índice cronológico e n o número de observações na amostra analisada.

Para o estudo de séries temporais, é preciso identificar os seus padrões, tais como aleatoriedade, estacionariedade, tendência e sazonalidade. Após esse reconhecimento inicial, é preciso identificar o modelo que mais se encaixa com as particularidades da série e em seguida quais serão os as estimativas dos seus parâmetros e a equação matemática que opera sobre elas. De fato, algumas séries podem não ser estacionárias, ou seja, tendo tendência nos dados e fazendo com que a mesma não fique ao redor de uma linha horizontal ao longo do período de observações (WERNER, 2003). Contudo, a não-estacionariedade pode ser contornada com sucessivas diferenças da série original (MORETTIN, 2006). Nesses casos pode ser usado o formalismo ARIMA (p, d, q).

O ARIMA tem como função principal analisar o potencial do modelo e depois prever os valores para a série temporal estudada (PINTO, 2008). Ele utilizada observações do passado, para tentar prever os valores futuros da variável estudada (ZHANG, 2003). Ele trata de uma abordagem linear do problema de previsão.

O modelo autoregressivo de médias móveis integrado pode ser estudado a partir dos seus três parâmetros de análise (p, d, q), sendo p o número de parâmetros autoregressivos (AR), d o número de diferenciações para série estacionarizar (I) e q o número de parâmetros de médias móveis (MA). A partir das Equações 1, 2 e 3, sendo respectivamente o modelo autoregressivo (AR), de médias móveis (MA) e autoregressivo de médias móveis (ARMA),

foi possível chegar à Eq. 3, do ARIMA.

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \epsilon_t \quad (1)$$

$$Y_t = \mu + \epsilon - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \quad (2)$$

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \epsilon - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \quad (3)$$

em que $\{\phi_i\}_{i=1}^p$ e $\{\theta_j\}_{j=1}^q$ são, respectivamente, os parâmetros das componentes AR(p) e MA(q), e ϵ é o erro aleatório intrínseco ao modelo e μ é o intercepto do modelo. Quanto ao ARIMA, tem-se

$$W_t = \phi_1 W_{t-1} + \dots + \phi_p W_{t-p} + \epsilon - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \quad (4)$$

em que $W_t = Y_t - Y_{t-1}$ representa a primeira diferença (I=1) no tempo t da série. Mais de uma diferenciação é também possível.

O GARCH (p, q), proposto por Bollerslev (1986), também chamado de modelo heteroscedástico condicional autoregressivo generalizado, trata-se da generalização do ARCH (q) (ENGLE, 1982), ele pode ser utilizado para retornar à volatilidade da série temporal com menos parâmetros do que o modelo ARCH (q). Este modelo é normalmente utilizado para a modelagem de séries temporais financeiras, uma vez que ele é especialista em dados com volatilidade, podendo utilizar dados passados para aprender sobre os futuros. O seu diferencial é capacidade de incorporar características das séries financeiras às suas estruturas, como a acomodação de caudas pesadas ou até mesmo a autocorrelação entre os quadrados de retorno dos dados (FIORUCI, 2012). O GARCH (p, q) é definido como:

$$\sigma_{t|t-1}^2 = \omega + \beta_1 \sigma_{t-1|t-2}^2 + \dots + \beta_p \sigma_{t-p|t-p-1}^2 + \alpha_1 r_{t-1}^2 + \alpha_2 r_{t-2}^2 + \dots + \alpha_q r_{t-q}^2 \quad (5)$$

em que $\omega > 0$, $\alpha_i \geq 0$, $\beta_i \geq 0$. As constantes do modelo, α_i e β_i são calculadas utilizando Regressão Linear.

Por sua vez, a média simples trata-se de um artifício estatístico que busca entender a tendência central da série temporal. Ela é a média das previsões dos modelos individuais, para dado instante de tempo.

2.3 MÉTRICAS DE DESEMPENHO

São formas de avaliar a qualidade do modelo, refletindo seu desempenho ao prever valores da série temporal sob estudo (MAKRIDAKIS, 1993). As mesmas são definidas a partir do erro de previsão, ou seja, o valor real da variável subtraído (ou mesmo dividindo) do valor da previsão do modelo. A Eq. 5 ilustra o caso da diferença.

$$\epsilon_t = Y_t - \hat{Y}_t \quad (6)$$

em que Y_t é o valor correto da série no instante t e \hat{Y}_t sua previsão. Dentre as principais métricas de desempenho, destacam-se o MSE (acrônimo para *Mean Square Error*), o MAPE (sigla para *Mean Absolute Percentual Error*) e o BIC (*Bayesian information critirium*). Matematicamente, tem-se

$$MSE = \frac{\sum_{t=1}^N (Y_t - \hat{Y}_t)^2}{N} \quad (7)$$

em que quanto menor o resultado do MSE, melhor será o modelo. Por sua vez, o MAPE é dado pela Eq. 7:

$$MAPE = \frac{100}{N} \sum_{t=1}^N \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right| \quad (8)$$

em que, assim como o MSE, quanto menor for seu valor, melhor será o modelo de previsão.

Por fim, o BIC, alternativa Bayesiana, é dado pela Eq.8:

$$BIC = -2 \ln l + (\ln N * k) \quad (9)$$

em que o l é o valor assumido pela função de verossimilhança diante dos dados e das estimativas adotadas para os parâmetros do modelo, N é o tamanho da amostra adotado na modelagem e o k o número de parâmetros do modelo. Assim como ocorre nas métricas de desempenho citadas anteriormente, quando menor for o valor desse critério de análise, melhor o modelo se adequará aos dados da série.

METODOLOGIA

O estudo sobre os modelos estatísticos foi baseado em livros relacionados ao assunto, como “*The Nature of Statistical Learni*” (VLADIMIR VAPNIK, 2000), de “Modelo para séries temporais” (MORETIN, 1981), “*Time Series Analysis With Applications In R*” (Cryer And Chan, 2008), além da leitura de artigos científicos e dissertações.

Para avaliação dos resultados, foram implementados algoritmos no Rstudio, uma interface para programação em R, para cada um dos modelos, e GARCH, considerando combinação dos modelos baseada em médias simples.

Os dados utilizados para análise dos modelos temporais são referentes ao CUB, entre os anos de 2012 e 2020, sendo observações mensais, totalizando 96 observações. Esses dados foram retirados do site oficial do SIDUCON- BR, o qual realizou o levantamento de dados a partir de Normas da ABNT e da Lei Federal 4. 591. As 64 primeiras observações foram destinadas ao treinamento dos modelos, enquanto que as restantes foram usadas para testar o desempenho dos modelos diante de situações inéditas a eles.

Para selecionar o melhor modelo de ARIMA foi realizada uma otimização no código através da implementação da função *model.arima* do pacote *forecast* do Rstudio, a qual trouxe como resultado o ARIMA (0,1,0), sendo necessário apenas trabalhar com o argumento “d”, diferenciação para estacionarizar a série temporal. Vale salientar que o mesmo foi selecionado de acordo com os parâmetros do BIC.

Para selecionar os parâmetros do Modelo GARCH, foi utilizado o pacote ‘fGarch’, sendo analisado todos os parâmetros, afim de identificar a melhor forma de construir o mesmo, garantido que posteriormente ele pudesse se adaptar a qualquer série temporal que venha a ser analisada.

3 RESULTADOS E DISCUSSÃO

Para que fosse viável a análise dos dados do modelo foi construído um código computacional no Rstudio com os modelos individuais GARCH e ARIMA, em seguida foi realizada uma combinação simples por Média. Abaixo é possível analisar o gráfico e suas medidas de otimização, o MSE e o MAPE, tanto para o treino como para o teste.

Analisando o gráfico, é possível observar que visivelmente todos os modelos conseguem acompanhar os dados (cor preta), mesmo o GARCH perdendo as duas primeiras observações. Isso acontece por causa de seu modelo GARCH (2,2), uma vez que ele precisa de 2 passos anteriores a observação (t) para aprender e conseguir prevê o dado

(t).

Vale salientar ainda que todos os modelos foram construídos com o mesmo número de dados para treinamento e teste, de forma que todos pudessem trabalhar de forma igualitária, mas para a construção do GARCH foi necessário diferenciar a série, para que o modelo pudesse ser otimizado, e por este motivo uma observação foi perdida, além das duas do modelo. Para contornar o problema do GARCH foi realizado o teste com as duas últimas observações do treinamento, além disso, foi preciso ainda transformar os valores obtidos no modelo e na previsão, para que pudessem ser analisados sem a diferenciação.

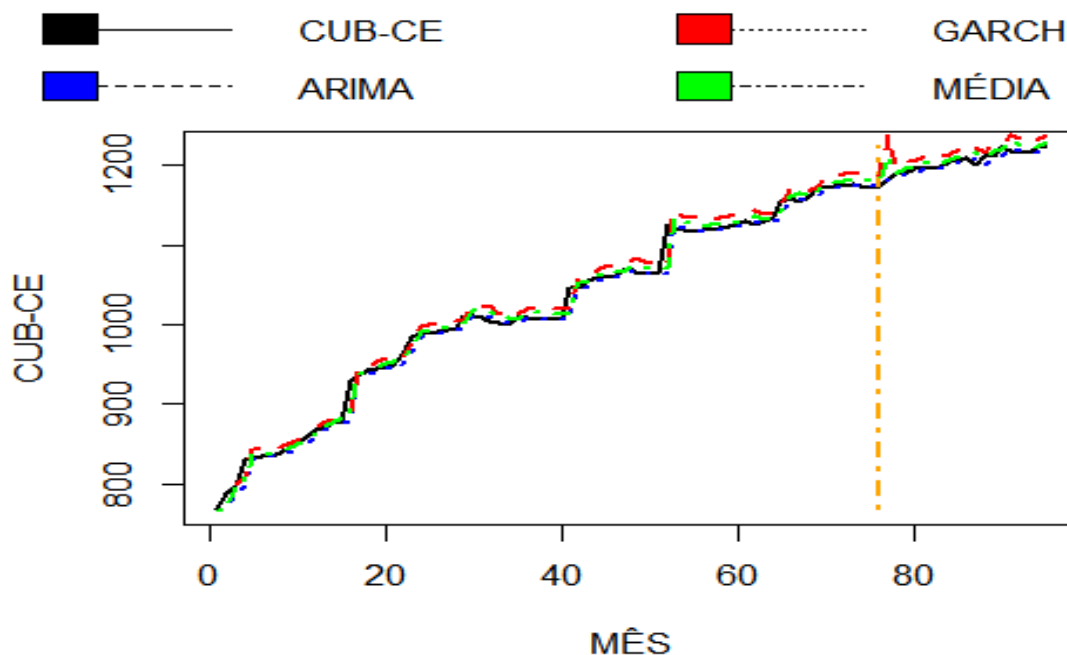


IMAGEM 1: GRÁFICO DE RESULTADO

Tabela 1 – RESULTADOS MAPE E MSE PARA TREINAMENTO E TESTE

MSE			
	GARCH	ARIMA	MÉDIA
TREINO	195.8027	153.7721	133.7244
TESTE	342.3076	26.4026	87.5880
MAPE			
TREINO	0.0116	0.0064	0.0073
TESTE	0.0125	0.0032	0.0059

Fonte: Autores.

Analisando os resultados do MSE e do MAPE, tanto para treinamento, quanto para teste, é possível chegar à conclusão que o modelo individual ARIMA se sai melhor em todas as etapas, o GARCH consegue acompanhar bem o treino, mas não se sai tão bem no teste e a média, por utilizar os dados do GARCH acaba sendo prejudicada no teste, porém

acompanha de forma satisfatória o treinamento, inclusive se destaca treino se for observar apenas o MSE.

Sobre os critérios de parcimônia, o ARIMA utiliza o critério de informação Bayesiano (BIC), enquanto isso, o GARCH utiliza a máxima verossimilhança.

4 CONSIDERAÇÕES FINAIS

A partir do estudo sobre este trabalho, foi possível identificar a eficiência da previsão de séries temporais e dos estudos realizados a partir de seus resultados, uma vez que eles podem ser decisivos nas tomadas de decisões em várias áreas da construção civil, e da área financeira. Para que se tivesse uma maior confiabilidade diante da comparação dos modelos, foi realizada uma avaliação diante dos seus parâmetros de desempenho e a possibilidade de uma combinação dos modelos.

Este trabalho abre a possibilidade de um estudo mais aprofundado sobre a combinação dos modelos, e o desenvolvimento de outros projetos em prol da otimização da tecnologia frente ao assunto abordado.

AGRADECIMENTOS

Agradeço a CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, por me proporcionar a experiência acadêmica e o auxílio financeiro que me possibilitou estudar e escrever este trabalho.

REFERÊNCIAS

- BEN-HUR, A.; WESTON, J. A user's guide to support vector Machines, 2008.
- BRUM, Aline Loreto. Análise da Complexidade Computacional de Problemas de Estatística Descritiva com Entradas Intervalares, 2006.
- CASTAÑEDA, Daniel Francisco Neyra, Séries temporais com R, V. 1, P- 1,2, 2017.
- CLEMEN, Robert T, Combining forecasts: A review and annotated bibliography. International journal of forecasting V. 5, n. 4, p. 559-583, 1989.
- CRYER, J. D.; CHAN, K. S. Time Series Analysis: With Applications in R. 2 ed. New York. Springer Science & Business Media, 2008.
- GENEST, C.; FRAVE, A.-C. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. Journal of hydrologic engineering, v. 12, issue 4, p. 347-368, 2007.
- LORENA, Ana Carolina; CARVALHO, André C. P. L. F.. Uma Introdução às Support Vector Machines, 2007. (??)
- MORETIN, Pedro Alberto, Modelos para previsão de séries temporais, 1981.
- MORETTIN, P. A. Análise de séries temporais. Edgard Blucher, 2006.
- PICHILINI, Mauro Carlos, Conversando sobre banco de dados, V. único, p –210, 2012.
- PINTO, P. L. A.; MAIA, S. F. . Previsão do preço das commodities agrícolas brasileiras: uma abordagem por modelos univariados. In: XLVI Congresso da Sociedade Brasileira de Economia e Sociologia Rural - SOBER Amazônia mudanças globais e agronegócio: O desenvolvimento em questão. 2008,
- T. MITCHLL. Machine Learning. McGraw Hill, 1997.
- VAPNIK, V. N. The nature of Statistical learning theory. Springer-Verlag, New York, 1995.
- WANG, Lipo, Support Vector Machines: Theory and applications, 2005.
- WERNER, Liane; Ribeiro, José Luis Duarte. Previsão de demanda: uma aplicação dos modelos Box-Jenkins na área de assistência técnica de computadores pessoais. Gestão & Produção (UFSCAR. Impresso), São Carlos - SP, v. 10, n.1, p. 47-68, 2003.
- G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," Neurocomputing, vol. 50, pp. 159-175, 2003.

S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, no. 4, pp. 527-529, 1993.

FAVERO, Luiz Paulo; BELFIORE, Patrícia. *Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel*, 2017.