



Modelo híbrido de previsão multiescala para séries da taxa mensal de mortalidade infantil

Márcia Lorena Alves dos Santos¹ e Eniuce Menezes de Souza²

¹Programa de Pós Graduação em Bioestatística - UEM

²Professora Adjunta do Departamento de Estatística - UEM

RESUMO

Embora os modelos SARIMA tenham sido largamente utilizados na previsão de séries temporais, às vezes a precisão da previsão é baixa ou não tão boa como desejada. Além disso, alguns pressupostos precisam ser validados para garantir uma modelagem da classe SARIMA adequada e apenas previsões de curto alcance são recomendadas. Neste trabalho, o objetivo é apresentar um modelo híbrido fixo que associa os modelos probabilísticos clássicos com uma análise *wavelet* multiescala para melhorar a qualidade da previsão. Após a decomposição *wavelet* multiescala, modelos probabilísticos clássicos são ajustados aos coeficientes *wavelet* em cada escala e é realizada a previsão, em seguida, a transformada inversa *wavelet* é aplicada, de modo a obter as previsões no domínio do tempo. Os métodos de previsão clássicos e híbridos são aplicados e comparados em dados reais referentes a taxa mensal de mortalidade infantil a cada 1.000 crianças pertinentes aos estados brasileiros e Distrito Federal. Os resultados mostraram a melhoria da previsão fornecida pelo modelo híbrido, em termos da raiz do erro quadrático médio, mantendo a qualidade preditiva mesmo com horizonte de previsão longo.

Palavras chave: Previsão, SARIMA, *Wavelets*, Mortalidade infantil.

1 INTRODUÇÃO

A análise de séries temporais é um importante instrumento para modelagem de dados com dependência temporal em diversas áreas do conhecimento. Os precursores

no âmbito de previsão de séries temporais foram Box e Jenkins, que baseados em estudos anteriores, apresentaram a classe de modelos ARIMA, abreviação do termo *Autoregressive Integrated Moving Average* [1]. Tais modelos tornaram-se clássicos na literatura e, conseqüentemente, estão entre os mais utilizados no estudo de séries temporais. Em paralelo, ocorreu o desenvolvimento de outra metodologia para a análise de séries, a análise *wavelet* [7].

Dentro da metodologia *wavelets*, o conceito de análise de multirresolução proposto por Mallat, é uma eficaz ferramenta no procedimento preditivo [2]. Consiste na decomposição da série em “camadas” de resolução, representadas por subséries com diferentes frequências, as quais favorecem a identificação de comportamentos não observados diretamente na série observada [6]. Tal decomposição dá-se pela aplicação da transformada *wavelet*, a qual irá expressar cada subsérie em termos de coeficientes *wavelets* [5]. Espera-se que a abordagem individual dessas subséries, seguida pela combinação linear de previsões otimizem a acurácia do procedimento preditivo [4]. Compostos pela combinação de diferentes metodologias, os modelos híbridos são capazes de associar a análise de multirresolução com a previsão multi-escala de modelos probabilísticos clássicos da análise de séries temporais.

Dessa forma, os modelos híbridos tornam-se mais dispendiosos em relação ao procedimento usual da modelagem ARIMA. A análise de multirresolução irá decompor a série observada em várias subséries de coeficientes *wavelets*, de modo que em cada subsérie seja ajustado um modelo probabilístico clássico. Isto é, ao invés de ajustar um único modelo clássico na série observada seguindo o procedimento usual, na abordagem híbrida é necessário ajustar um modelo clássico para cada subsérie decomposta. Logo, o dispêndio é multiplicado pelo número de subséries fornecidas pela análise de multirresolução. À vista disso, se fosse possível fixar um modelo clássico capaz de modelar qualquer subsérie de coeficientes *wavelets*, tal dispêndio seria evitado. Por conseqüência, o modelo híbrido dispensaria o acompanhamento de um especialista no procedimento da modelagem híbrida, sendo acessível por pesquisadores de outras áreas.

Além disso, sabe-se que a classe de modelos ARIMA é adequada para séries estacionárias [1]. Assim, para ajustar tais modelos a uma série temporal é necessário remover as fontes de variação não estacionárias, tais como tendência e sazonalidade, de modo que seja possível o acesso a autocorrelação da série observada. Contudo, mesmo com a condição de estacionariedade em que o ARIMA conduz a um excelente ajuste, a previsão dessa classe de modelos é recomendada apenas para o curto prazo, ou seja, horizontes de previsão curtos. Em termos práticos, isso significa que as previsões de curto prazo são mais confiáveis ou mais acuradas do que as previsões de longo prazo. Tal limitação é superada pela abordagem híbrida a partir de *wavelets*, na qual previsões acuradas podem ser obtidas tanto para séries temporais estacionárias quanto não-estacionárias, além de horizontes de previsão curtos ou longos.

Motivado pelas considerações apresentadas anteriormente, o presente trabalho irá comparar o desempenho de previsão de modelos clássicos ARIMA com um modelo híbrido composto pela combinação entre *wavelets* e modelos probabilísticos clássicos. O modelo híbrido abordado é fixo, ou seja, foi identificado de acordo com as características teóricas de tais modelos quanto dos coeficientes *wavelets*. Fixado

modelos probabilísticos clássicos para modelar as subséries de coeficientes *wavelets*, dispensando a etapa de busca por modelos que conduz a um bom ajuste e uma boa previsão para cada subséries.

2 METODOLOGIA

Os dados em estudo foram obtidos por meio do site do DATASUS e destes calculou-se as taxas mensal de mortalidade infantil a cada 1.000 crianças referentes aos estados brasileiros e Distrito Federal. As taxas foram obtidas a partir da razão entre o número de óbitos infantil e o número de nascidos vivos pertinente a cada estado e Distrito Federal, no período entre Janeiro de 1996 a Dezembro de 2014, totalizando 228 observações.

Acerca da previsão, será comparada a qualidade desta a partir de dois modelos:

- ARIMA: o modelo clássico ARIMA de acordo com as etapas de identificação, estimação, validação e previsão da metodologia de Box-Jenkins;
- Modelo híbrido: consiste na decomposição da série observada via transformada *wavelet*, seguida pelo ajuste e previsão de modelos probabilísticos clássicos em cada uma das subséries de coeficientes *wavelet* e, por fim, a transformada inversa é aplicada para obter as previsões no domínio do tempo. A decomposição multiescala foi feita em 4 níveis a partir das *wavelets*-Mãe Daubechies 2, Daubechies 18, Daubechies 20 e Best Localized 18.

Para avaliar a acurácia, as previsões foram feitas dentro do período amostral e comparadas com os valores observados. O período de previsão foi entre Janeiro de 2013 e Dezembro de 2014, sendo o horizonte de previsão igual a 24 meses. A mensuração do erro associado aos modelos de previsão foi realizada por meio da raiz do Erro Quadrático Médio. O melhor modelo foi aquele com menor erro de previsão. A análise foi desenvolvida com o auxílio do software R (R Core Team, 2016) [3].

A maioria das séries observadas apresentaram variância não constante, inviabilizando o ajuste de modelos SARIMA. Para estabilizar a variância, fez-se necessário o uso da transformação Box Cox, em particular, a logarítmica. Apenas as séries pertinentes aos estados do Piauí, Amapá, Rondônia e Goiás não precisaram de transformação.

3 RESULTADOS

Os modelos SARIMA foram ajustados nas séries das taxas mensais de mortalidade infantil no período entre Janeiro de 1996 e Dezembro de 2013. Contudo, em 11 séries observadas não foi identificado modelos SARIMA capaz de modelá-las, de maneira que os resíduos fossem estacionários. Estas são: Roraima, Pará, Amapá, Ceará, Rio Grande do Norte, São Paulo, Paraná, Santa Catarina, Rio Grande do Sul, Mato Grosso e Distrito Federal. Nesses casos, considerou-se somente a modelagem híbrida. As séries das taxas mensais de mortalidade infantil pertinentes aos demais estados

brasileiros foram modeladas pelo modelo SARIMA e pelo modelo híbrido e, após, comparadas a qualidade de previsão de ambos modelos.

A Tabela 1 apresenta os erros de previsão dos modelos. Pode-se verificar que a metodologia clássica SARIMA possui maiores erros de previsão para quase todas as séries observadas, evidenciando a superioridade dos modelos híbridos de previsão. Entretanto, para a série referente ao estado de Minas Gerais, o modelo SARIMA apresentou melhor desempenho preditivo com relação ao modelo híbrido.

Tabela 1: Raiz do Erro Quadrático Médio de previsão.

Estados	SARIMA	Híbrido	Estados	SARIMA	Híbrido
1.Rondônia	3.45	1.31 (D20)	15.Sergipe	2.96	1.39 (D18)
2.Acre	3.16	1.71 (D18)	16.Bahia	0.74	0.48 (D20)
3.Amazonas	2.67	1.21 (D20)	17.Minas Gerais	1.15	1.41 (D20)
4.Roraima	—	2.02 (D20)	18.Espírito Santo	2.33	1.27 (BL18)
5.Pará	—	0.51 (D20)	19.Rio de Janeiro	1.13	0.52 (D20)
6.Amapá	—	2.89 (D20)	20.São Paulo	—	8.04 (D2)
7.Tocantins	2.38	1.30 (D20)	21.Paraná	—	0.46 (D20)
8.Maranhão	1.98	0.78 (D20)	22.Santa Catarina	—	0.51 (D20)
9.Piauí	2.64	1.60 (D18)	23.Rio Grande do Sul	—	0.57 (D20)
10.Ceará	—	0.93 (D20)	24.Mato Grosso do Sul	2.37	0.87 (D20)
11.Rio Grande do Norte	—	0.70 (D20)	25.Mato Grosso	—	1.23 (D20)
12.Paraíba	1.91	1.13 (BL18)	26.Goiás	1.27	0.52 (D20)
13.Pernambuco	1.32	1.01 (D20)	27.Distrito Federal	—	0.82 (D20)
14.Alagoas	1.56	0.95 (D20)			

4 CONCLUSÃO

De acordo com os resultados obtidos, considerando os dados em questão, é confirmada a hipótese de superioridade do modelo híbrido de previsão de longo prazo quando comparado ao modelo SARIMA.

Referências

- [1] BOX, G. E.; JENKINS, G. Time series analysis: forecasting and control. Holden-Day, 1970.
- [2] MALLAT, S. A *wavelet* tour of signal processing. Academic Press, 2008.
- [3] NASON, G. *Wavelet* Methods in Statistics with R. Springer New York, 2010.
- [4] ARINO, M. A. Time series forecasts via wavelets: an application to car sales in the Spanish market. Institute of Statistics & Decision Sciences, Duke University, 1995.
- [5] VIDAKOVIC, B.; PETER, M. “*Wavelets* for kids.” Instituto de Estadística, Universidad de Duke (1994).
- [6] MORETTIN, P. A. Ondas e Ondaletas Vol. 23. Edusp, 1999.
- [7] DAUBECHIES, I. et al. Ten lectures on wavelets. SIAM, 1992.
- [8] R: A Language and Environment for Statistical Computing, R Core Team: R Foundation for Statistical Computing, Vienna, Austria, 2016.