Universidade Estadual de Maringá
Programa de Pós-graduação em Bioestatística

MÁRCIA LORENA ALVES DOS SANTOS

MODELOS HÍBRIDOS PARA A PREVISÃO MULTIESCALA DE SÉRIES TEMPORAIS

Maringá - Paraná
Março de 2017

MÁRCIA LORENA ALVES DOS SANTOS

MODELOS HÍBRIDOS PARA A PREVISÃO MULTIESCALA DE SÉRIES TEMPORAIS

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Bioestatística. Orientadora: Dra. Eniuce Menezes de Souza.

Maringá - Paraná

Março de 2017

Ao meu Deus, por sua infinita misericórdia em minha vida.
Aos meus amados pais, João e Márcia, pela parceria e amor indestrutíveis.
À minha querida avó, Aparecida, que sempre esteve disposta a me ajudar.
Aos meus estimados professores, sem eles, essa dedicatória jamais seria escrita.

Agradecimentos

A Deus, por ser meu ajudador nos momentos de dificuldade, minha força nos momentos de fraqueza, minha luz nos momentos de escuridão, meu protetor nos momentos de perigo, meu amigo nos momentos de solidão e minha salvação em meio a perdição.

Aos meus pais e toda minha família, por investirem grande parte de suas vidas em minha formação. Sou eternamente grata!

Ao Departamento de Estatística da Universidade Estadual de Maringá, pela criação do Programa de Pós-Graduação em Bioestatística (PBE), um imenso orgulho para todos os maringaenses.

À minha orientadora e coordenadora do PBE, Drª. Eniuce Menezes de Souza, pelo companheirismo durante esses dois anos e pelo exemplo profissional e pessoal que seguirei durante minha jornada acadêmica.

À vice-coordenadora do PBE, Drª. Isolde Previdelli, pela calorosa recepção, pelos sábios conselhos e por sempre acreditar no crescimento dos alunos, outro exemplo profissional e pessoal a ser seguido.

Aos companheiros do PBE, Tiago, José André, Omar, Beatriz, Rafaela, dentre outros que tiveram presentes nessa jornada. Em especial, ao Ricardo Puziol pela parceria e amizade desde a graduação e a Edilenia Queiroz por dividir a sala de estudos durante essa jornada.

Ao Grupo de Estudos em Séries Temporais e Espaciais (GESTE) pelas discussões e sugestões.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

Aos trabalhadores brasileiros, que mesmo com o difícil acesso à educação básica e universitária decorrente da corrupção generalizada, contribuem para um Brasil melhor.

E ainda que tivesse o dom de profecia, e conhecesse todos os mistérios e toda a ciência, e ainda que tivesse toda a fé de maneira tal que transportasse os montes, e não tivesse o amor, nada seria. (Aos Coríntios 13:02)

Resumo

Independente da área do conhecimento, é grande a necessidade de previsões acuradas tanto no curto como no longo prazo. Tal demanda torna-se mais emergente quando trata-se de séries temporais epidemiológicas. Nesse contexto, previsões acuradas poderão revelar instabilidades no crescimento do número de casos de determinada doença, denunciar seu comportamento, indicando o período de maior incidência ou expressar as regiões geográficas de uma localidade mais atingidas por determinada doença. Em síntese, no contexto epidemiológico, informações precoces acuradas implicam em estratégias de combate eficazes e, consequentemente, no possível controle de doenças. Assim, neste estudo, propõe-se um novo modelo híbrido de previsão, composto pela junção da metodologia *wavelet* aos modelos probabilísticos clássicos na análise de séries temporais. Tal modelo mostra-se capaz de fornecer previsões acuradas para o curto, médio e longo prazos, suprindo a necessidade de informações prévias e precisas presente na literatura. No primeiro momento, foram comparados o desempenho de previsão dos modelos clássicos SARIMA com quatro modelos híbridos de previsão baseados na transformada *wavelet* discreta. Dentre os modelo híbridos confrontados, três são apresentados por outros estudos e o quarto modelo consiste do novo modelo proposto. Para tal comparação, foram simuladas três séries temporais mediante sete cenários distintos, totalizando 21 séries simuladas, nas quais os modelos foram aplicados. Para a transformada *wavelet* discreta presente nos modelos híbridos, foram consideradas 25 *wavelets* mãe pertinentes a seis famílias. Além disso, os níveis de resolução da decomposição *wavelet* foram alternados entre 4 níveis, 5 níveis e 6 níveis. A raíz do erro quadrático médio e o erro absoluto médio foram responsáveis pela mensuração da acurácia preditiva dos modelos. O modelo proposto neste estudo, apresentou o melhor desempenho de previsão para as séries simuladas. Em seguida, tal modelo foi aplicado nas taxas de mortalidade infantil e nas taxas de internações por bronquiolite pertinente aos estados brasileiros e Distrito Federal. Para avaliar seu desempenho de previsão em séries temporais reais, em ambos os casos, o modelo híbrido proposto foi comparado ao modelo SARIMA, usualmente utilizados em séries epidemiológicas. Para as taxas de mortalidade infantil, o modelo híbrido proposto mostrou desempenho de previsão superior ao modelo SARIMA, apresentando uma redução no erro de previsão de até 63%. Para as taxas de internações de bronquiolite, os resultados foram ainda melhores, de modo que ao substituir a previsão do modelo SARIMA pela previsão do modelo híbrido proposto, a redução no erro de previsão chegou a 93.7%. Ademais, a automaticidade presente na aplicação do modelo híbrido proposto viabiliza a utilização deste por pesquisadores de qualquer área do conhecimento, mostrando o seu valor interdisciplinar.

Palavras-chave: Previsão, Modelo híbrido, *Wavelet*, Mortalidade infantil, Bronquiolite.

Abstract

Regardless of the area of knowledge, there is a great need for accurate forecasts of short or long term. Such demand becomes more emergent when it comes to epidemiological time series. In this context, accurate forecasts will reveal possible instabilities in the increasing number of cases of a particular disease, revealing their behavior, the period of greatest incidence or expressing the geographical regions most affected by a specific disease. In summary, in the epidemiological context, previous and accurate information implies in effective strategies of combat and, consequently, in the possible control of diseases. Thus, in this study, we propose a new hybrid forecasting model, composed by the combination of the wavelet methodology and the classical probabilistic models in the time series analysis. This model is able to provide accurate forecasts for the short, medium and long term, satisfying the need for previous and accurate information present in the literature. In the first moment, the forecasting performance of the classic SARIMA model was compared with four hybrid prediction models based on the discrete wavelet transform. Among the hybrid models confronted, three are presented by other studies and the fourth model consists of the new proposed model. For this comparison, three time series were simulated using seven different scenarios, totaling 21 simulated series, in which the models were applied. To apply the discrete wavelet transform in the hybrid models, 25 mother wavelets pertinent to six families were considered. In addition, resolution levels of the wavelet decomposition were alternated between 4 levels, 5 levels and 6 levels. The root mean square error and mean absolute error were responsible for the accuracy measure of the models. The model proposed in this study presented the best prediction performance for all simulated series. This model was then applied to the infant mortality rates and rates hospitalization for bronchiolitis pertinent to the Brazilian states and the Federal District. To evaluate their forecasting performance in real time series, in both cases, the proposed hybrid model was compared to the SARIMA model, usually used in epidemiological series. For infant mortality rates, the proposed hybrid model showed a predictive performance superior to the SARIMA model, presenting a reduction in prediction error of up to 63%. For the hospitalization rates by bronchiolitis, the results were even better, so that by replacing the SARIMA model prediction with the prediction of the proposed hybrid model, the reduction in prediction error reached 93.7%. In addition, the automaticity present in the application of the proposed hybrid model allows the use of this one by researchers of any area of knowledge, showing its interdisciplinary value.

Key-words: Forecasting, Hybrid model, *Wavelets*, Child mortality, Bronchiolitis.

Lista de ilustrações

Figura 1 – Ilustração da análise de multirresolução: decomposição da série em níveis de resolução ou multiescala. 24
Figura 2 – Etapas da análise *wavelet* (VIDAKOVIC; MUELLER, 1994). 28
Figura 3 – Comportamento das séries simuladas X_t^1, X_t^2 e X_t^3, respectivamente. . 34
Figura 4 – Série X_t^1 com X_t simulada a partir do sétimo cenário. 35
Figura 5 – Sistematização da metodologia utilizada para a comparação do desempenho preditivo entre os modelos. 36
Figura 6 – Modelo híbrido 1 para a série X_t^1 com X_t simulada a partir do cenário AR(1). 47
Figura 7 – Modelo híbrido 2 para a série X_t^1 com X_t simulada a partir do cenário AR(1). 48
Figura 8 – Modelo híbrido 3 para a série X_t^1 com X_t simulada a partir do cenário AR(1). 48
Figura 9 – Modelo híbrido 4 para a série X_t^1 com X_t simulada a partir do cenário AR(1). 49
Figura 10 – Abordagem clássica do modelo SARIMA. 49
Figura 11 – Modelagem abordada conforme com o estado. 55
Figura 12 – Modelos ajustados para as taxas de mortalidade infantil de Minas Gerais. 56
Figura 13 – Maiores reduções do erro de previsão, em porcentagem, fornecidas pelo modelo híbrido proposto com relação ao modelo SARIMA. 57
Figura 14 – Previsão de curto prazo. Séries dos estados do Goiás (GO) e Rondônia (RO), respectivamente. 62
Figura 15 – Previsão de médio prazo. Séries dos estados de Goiás (GO) e Piauí (PI), respectivamente. 63
Figura 16 – Previsão de longo prazo. Séries pertinentes aos estados do Acre (AC) e Rio Grande do Sul (RS), respectivamente. 63

Lista de tabelas

Tabela 1 – Modelos com menor RMSE de previsão à longo prazo (h=24) para a decomposição *wavelet* com 4 níveis perante todas as *wavelets* mãe. . . 39
Tabela 2 – Modelos com menor RMSE de previsão à longo prazo (h=24) para a decomposição *wavelet* com 5 níveis perante todas as *wavelets* mãe. . . 40
Tabela 3 – Modelos com menor RMSE de previsão à longo prazo (h=24) para a decomposição *wavelet* com 6 níveis perante todas as *wavelets* mãe. . . 40
Tabela 4 – Porcentagem do número de vezes que cada modelo apresentou o menor RMSE resultante da previsão de longo prazo (h=24) relativa as *wavelets* mãe. 41
Tabela 5 – *Ranking* das *wavelets* mãe segundo a porcentagem de vezes que o modelo híbrido 4 apresentou-se como melhor modelo de previsão relativo as 21 séries simuladas. 43
Tabela 6 – *Ranking* das melhores *wavelets* mãe respectivas a cada família *wavelet*, segundo os níveis de resolução. 44
Tabela 7 – Porcentagem média de redução do RSME de ajuste e previsão do modelo híbrido 4 em relação ao modelo clássico SARIMA. 45
Tabela 8 – Modelo ajustado em cada estado brasileiro e Distrito Federal. 55
Tabela 9 – Raiz do Erro Quadrático Médio de previsão. 56
Tabela 10 – Raiz do Erro Quadrático Médio segundo o prazo de previsão. 64

Sumário

1 Introdução **11**
1.1 Estado da arte 13
1.2 Justificativa 15
1.3 Objetivo 15
1.4 Estrutura do trabalho 16
2 Métodos e Simulação **17**
2.1 Métodos 17
2.1.1 Previsão baseada na esperança condicional 17
2.1.2 Modelos ARIMA 18
2.1.3 Modelo *Spline* 19
2.1.4 Modelo Harmônico 20
2.1.5 Análise Wavelets 20
2.1.5.1 O modelo *wavelet* 21
2.1.5.2 Análise de Multirresolução 22
2.1.5.3 *Wavelets* mãe 25
2.1.5.4 Transformada *Wavelet* Discreta 26
2.1.5.5 *Wavelet Denoising* 27
2.1.5.5.1 Global (Limiar Universal) 28
2.1.5.5.2 Global (Uma alternativa) 29
2.1.5.5.3 Adaptativo (*SureShrink*) 29
2.2 Metodologia 30
2.2.1 Os modelos híbridos 30
2.2.2 Séries Simuladas 33
2.2.3 Sistematização da metodologia 35
2.2.4 Avaliação dos modelos 36
2.2.4.1 Avaliação dos modelos segundo os níveis de resolução . . 37
2.2.4.2 Avaliação dos modelos segundo as *wavelets* mãe 37
2.2.4.3 Porcentagem média de redução do erro 38
2.3 Resultados e discussões 38
2.3.1 Desempenho de previsão à longo prazo dos modelos segundo os níveis de resolução e *wavelets* mãe 39

2.3.2 Desempenho de previsão à longo prazo do modelo híbrido 4 segundo a *wavelet* mãe . . . 42
2.3.3 Desempenho de previsão à longo prazo do modelo híbrido 4 em relação ao modelo clássico SARIMA . . . 44
2.3.4 Resultados dos modelos para uma das séries simuladas . . . 46
2.3.5 Desempenho computacional dos modelo confrontados . . . 50
2.4 Conclusão . . . 50
3 Modelo híbrido para previsão da taxa de mortalidade infantil no Brasil . . . 52
3.1 Introdução . . . 52
3.2 Metodologia . . . 53
3.3 Resultados e Discussões . . . 54
3.4 Conclusão . . . 58
4 Modelo híbrido proposto para previsão de hospitalizações por Bronquiolite no Brasil . . . 59
4.1 Introdução . . . 59
4.2 Metodologia . . . 60
4.3 Resultados e Discussões . . . 62
4.4 Conclusão . . . 65
5 Considerações finais e Trabalhos futuros . . . 66
Referências . . . 68

Capítulo 1

Introdução

A análise de séries temporais é uma importante ferramenta estatística para o estudo de séries epidemiológicas e essencial para modelagem de dados com dependência temporal em diversas áreas do conhecimento. Uma série temporal é qualquer conjunto de observações ordenadas no tempo (MORETTIN; TOLOI, 2006). O apogeu no âmbito de previsão de séries temporais ocorreu em meados da década de 1970. Os precursores desse desenvolvimento foram Box e Jenkins, que baseados em estudos anteriores, apresentaram a classe de modelos ARIMA, abreviação do termo *Autoregressive Integrated Moving Average*. Tais modelos tornaram-se clássicos na literatura e, consequentemente, estão entre os mais utilizados no estudo de séries temporais. Paralelo ao apogeu dos modelos ARIMA, ocorreu o desenvolvimento de outra metodologia para a análise de sinais e séries temporais, a análise *wavelet*. As *wavelets* são funções matemáticas que satisfazem determinados requisitos, necessários para garantir cálculo rápido e fácil da decomposição da série temporal nas suas partes constituintes (VIDAKOVIC; MUELLER, 1994).

A primeira referência ao que se conhece por *wavelet*, foi encontrada na tese do matemático Alfred Haar, em 1909. Durante algumas décadas, a base de Haar era a única base *wavelet* ortonormal conhecida, constituída por *wavelets* de suporte compacto e não diferenciáveis em todo seu intervalo. No final da década de 1970, o engenheiro geofísico Jean Morlet introduziu a ideia da transformada *wavelet*, sugerindo a compressão e expansão de funções trigonométricas para a análise de sinais no domínio do tempo e da frequência.

Anos após, o físico Alex Grossmann desenvolveu a fórmula de reconstrução da transformada *wavelet* sendo que, em 1983, Grossmann e Morlet apresentaram formalmente o conceito de transformada *wavelet*. Contudo, o auge do desenvolvimento da teoria *wavelet* ocorreu na segunda metade da década de 1980, quando ao ver a fórmula proposta por Grossmann e Morlet, o matemático Yves Meyer reconheceu sua gênese proveniente da análise harmônica, presente no trabalho do matemático Alberto Calderón divulgado

em 1960. Buscando provar as propriedades da transformada *wavelet*, Meyer desenvolveu uma nova base *wavelet* ortonormal continuamente diferenciável, mas precavida de suporte compacto. Em 1986, o especialista em análise de imagem Stephane Mallat, sugeriu a decomposição *wavelet* em camadas, na qual os termos correspondentes a cada camada são obtidos pela diferença entre duas aproximações sucessivas. Nesse mesmo ano, em uma visita a Chicago, Meyer e Mallat elaboraram toda a teoria matemática envolvida na análise de multirresolução, facilitando a construção de novas bases ortonormais (DAUBECHIES, 1996). A partir desses trabalhos, em 1988, a física e matemática Ingrid Daubechies, também aluna de Grossmann, desenvolveu bases *wavelets* ortonormais diferenciáveis em toda sua extensão e com suporte compacto, as quais constituem o alicerce das aplicações atuais (LIMA, 2002). Entretanto, a imensurável expansão das *wavelets* se deve ao desenvolvido por Mallat em 1998, possibilitando a grande eficiência computacional e aplicações, inclusive em tempo real. A transformada *wavelet* decimada, implementada a partir do algoritmo piramidal é muito utilizada até hoje em diversas áreas, mas em séries temporais, surgiram derivações tais como a transformada *wavelet* discreta não decimada (TWDND), a qual garante a invariância por translação e é essencial para análise de séries temporais (NASON; SAPATINAS; SAWCZENKO, 1997; PERCIVAL; WALDEN, 2006).

O conceito de análise de multirresolução proposto por Mallat, é uma eficaz ferramenta no procedimento preditivo. 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 temporal em estudo. É como se pudéssemos “olhar” os dados por um microscópio e ver seu comportamento em várias “magnificações” (MORETTIN, 1999). Tal decomposição dá-se pela aplicação da transformada *wavelet*, a qual irá expressar cada subsérie em termos de coeficientes *wavelets*. Espera-se que a modelagem individual dessas subséries, isto é, ajuste e previsão de modelos probabilísticos clássicos em cada subsérie, seguida pela combinação linear de previsões para a reconstrução da série temporal, otimizem a acurácia do procedimento preditivo.

Por analogia, assim como os carros híbridos são compostos por um motor a combustão auxiliado por outro elétrico, um modelo estatístico híbrido é composto pela combinação de diferentes metodologias estatísticas, que auxiliam-se na análise dos dados, assim como o motor elétrico auxilia o motor a combustão. Nessa lógica, a análise de multirresolução auxiliada pela abordagem multiescala de modelos clássicos da análise de séries temporais, ou seja, pelo tratamento individual das subséries de coeficientes *wavelets*, denomina-se modelo híbrido. Em síntese, os modelos híbridos de previsão aqui abordados resultam da combinação entre análise de multirresolução e previsão multiescala de modelos clássicos. Sendo assim, foi realizada uma investigação sobre a modelagem híbrida para séries temporais presente na literatura.

1.1 Estado da arte

As vantagens da aplicação de modelos híbridos baseados na decomposição *wavelet* para a previsão de séries temporais, vem sendo investigada e destacada em vários estudos. Na década de 90, Arino (1995) utilizou a decomposição *wavelet* para fragmentar a série observada em duas subséries, associadas às escalas de alta e baixa frequências, respectivamente. Feito isso, comparou e verificou a superioridade da previsão de modelos ARIMA nessas subséries seguida pela reconstrução (D8[1]), com relação a previsão de modelos ARIMA aplicados diretamente na série observada. Generalizações do trabalho de Arino deram origem a outros estudos. Nessa perspectiva, Homsy, Portugal e Araújo (2000) sugeriu o acréscimo do procedimento de *denoising* (Haar[2]) nas componentes de alta e baixa frequências, antes do ajuste de modelos ARIMA. Também, Wong et al. (2003) utilizou a decomposição *wavelet* para estimar as componentes de modelos estruturais de séries temporais (STSM), os quais descrevem uma série X_t como a combinação linear de três componentes: tendência T_t, sazonalidade S_t e ruido aleatório ε_t. Dados os estimadores $\hat{T}_t$ e $\hat{S}_t$, a previsão é feita pela extrapolação linear das funções ajustadas em cada componente. Segundo Wong et al. (2003), a abordagem *wavelet* baseado em STSM fornece previsões acuradas. Lima e Almeida (2004) combinou a decomposição *wavelet* (Haar) com a previsão de modelos de Redes Neurais nas subséries de alta e baixa frequências.

Em continuação, Conejo et al. (2005) chamou de modelo *Wavelet*-ARIMA a decomposição *wavelet* (D5[3]) em 3 níveis, seguida pelo ajuste e previsão multiescala de modelos ARIMA. Neste estudo, a transformada *wavelet* discreta (TWD) decimada foi aplicada para a obtenção dos níveis de resolução e suas respectivas subséries, onde foram realizadas previsões a partir de modelos ARIMA para 1, 2, . . ., 24 passos à frente. Outro modelo híbrido foi sugerido por Yousefi, Weinreich e Reinarz (2005). Tal modelo é composto pela TWD seguida pelo ajuste e extensão das subséries a partir de modelos harmônicos e função *spline*. Para ilustrar o poder preditivo dessa abordagem, o procedimento de previsão foi aplicado em dados de preços médios mensais de petróleo referente ao período de 02 de Janeiro de 1986 até 31 de Janeiro de 2003. A *wavelet* mãe Daubechies (D7[4]) foi adotada e os horizontes de previsão foram de 1, 2, 3 e 4 meses à frente. Em vista disso, os valores previstos pelos modelos foram comparados com a expectativa de mercado, representada pelos valores dos preços médios mensais futuros pertinentes ao período de previsão. Para avaliar a acurácia da previsão, comparou-se os coeficientes de correlação entre valores preditos e valores reais *versus* valores futuros e valores reais. De acordo com Yousefi, Weinreich e Reinarz (2005), em média, o procedimento de previsão baseado em *wavelet* mostrou-se superior ao cálculo da expectativa de mercado. Ao que parece, essa

[1] *Wavelet* mãe Daubechies 8 utilizada na análise de multirresolução
[2] *Wavelet* mãe Haar utilizada na análise de multirresolução
[3] *Wavelet* mãe Daubechies 5 utilizada na análise de multirresolução
[4] *Wavelet* mãe Daubechies 7 utilizada na análise de multirresolução

superioridade preditiva permanece mesmo quando o horizonte de previsão é prolongado.

Outra abordagem híbrida envolvendo a decomposição *wavelet* seguida pelo ajuste e previsão multiescala de modelos ARIMA, foi apresentada por Schlüter e Deuschle (2010). O objetivo foi verificar a relevância das previsões obtidas pela metodologia *wavelet* em relação aos modelos clássicos, dada sua complexidade. Para isso, foram propostos dois modelos de previsão baseados em *wavelets*, dados por: *wavelet denoising* seguido pelo ajuste e previsão de modelos clássicos e, decomposição *wavelet*, seguida pelo ajuste e previsão multiescala de modelos clássicos. *Wavelet denoising* foi realizado via método *SureShrink* [5] segundo as *wavelets* de Haar e Morlet. Por sua vez, a decomposição *wavelet* foi realizada via transformada *wavelet* contínua (TWC) junto a discretização de cada nível, mediante as *wavelets* de Haar e Daubechies (D4[6]). Os modelos clássicos de previsão utilizados foram: Census X-12 (utilizado para a implementação numérica do STSM), ARMA e ARIMA. Dois horizontes de previsão foram adotados, um dia à frente e uma semana à frente, para quatro séries temporais com características peculiares. Schlüter e Deuschle (2010) afirmam que não houve o melhor modelo de previsão para todas as séries, pois suas características e o horizonte de previsão influenciaram na escolha. Contudo, em todos os cenários avaliados foi possível encontrar um modelo baseado em *wavelets* com desempenho preditivo superior aos modelos clássicos. Assim, de forma geral, as melhores previsões foram geradas pelos modelos: *wavelet denoising* seguido pelo ajuste e previsão ARIMA e decomposição *wavelet* seguida pelo ajuste e previsão ARIMA multiescala. Bayer (2010) propôs o uso conjunto dos algoritmos *Holt Winters* com a decomposição *wavelet* via transformada *wavelet* decimada (D8[7]), para estimar as subséries de alta e baixa frequências da série observada.

Wadi, Hamarsheh e Alwadi (2013) apresentou o modelo híbrido dado pela decomposição da série observada via TWDND seguida pelo ajuste e previsão de modelos ARIMA em duas subséries de coeficientes *wavelets*. Neste estudo, foram comparadas a acurácia preditiva do modelo baseado na combinação entre TWD e modelo de previsão ARIMA, considerando duas *wavelets*, Haar e Daubechies e, também, o modelo baseado na combinação entre TWDND e previsão ARIMA. Além disso, foi feita a comparação com modelo ARIMA ajustado diretamente na série observada. A decomposição da série deu-se em 1 nível, formado pelas componentes de alta e baixa frequências. O objetivo foi eleger o modelo com melhor desempenho preditivo para dados bancários, por meio da aplicação em dados do retorno diário da bolsa de valores Amman na Jordânia, no período entre 1993 até 2009, onde o modelo híbrido dado pela TWDND combinado ao ajuste e previsão de modelos ARIMA nas subséries, apresentou o melhor desempenho preditivo.

[5] Stein's Unbiased Risk Estimator
[6] *Wavelet* mãe Daubechies 4 utilizada na análise de multirresolução
[7] *Wavelet* mãe Daubechies 8 utilizada na análise de multirresolução

1.2 Justificativa

Sabe-se que o procedimento usual de Box & Jenkins para a modelagem ARIMA é bastante dispendioso. Na modelagem clássica, este procedimento é realizado uma única vez para a série temporal observada. Com os modelos híbridos presentes na literatura, o procedimento de Box & Jenkins precisa ser repetido para cada subsérie decomposta, identificando o melhor modelo ARIMA para cada subsérie, que embora possibilite melhores resultados, torna-se bem mais dispendioso. Uma forma de automatização, de modo que não seja necessário que um especialista acompanhe todo o procedimento de modelagem e previsão, seria identificar um modelo ARIMA que minimize algum critério de informação, tal como AIC (*Akaike Information Criterion*), ou a modelagem *Holt Winters*. Em ambas opções, o dispêndio deixa de ser em nível de usuário especializado e passa a ser em nível computacional, devido ao esforço computacional exigido para encontrar o melhor modelo. Mais uma desvantagem deve ser ressaltada, pois sabe-se que modelos ARIMA identificados apenas pela minimização do AIC, por exemplo, possuem excelente ajuste, mas podem deixar a desejar quanto a previsão de longo prazo, resultado de maior interesse.

Embora o dispêndio computacional possa ser tolerado em algumas aplicações, inviabiliza aplicações *online* ou em tempo real. Assim, se fosse possível fixar a construção de um modelo simples probabilístico para a previsão em cada subsérie decomposta, seria excelente para evitar o dispêndio tanto em nível de usuário como em nível computacional. Nesse sentido, a proposta deste trabalho vem atender tal necessidade, permitindo uma modelagem e previsão automática. Isto se deve, também, a eficiência dos algoritmos de decomposição *wavelets*. Com a abordagem híbrida a partir de *wavelets*, previsões acuradas podem ser obtidas tanto para séries temporais estacionárias quanto não-estacionárias, além de horizontes de previsão curtos, médios ou longos. Frisando que, a estacionariedade da série temporal é um pressuposto exigido pela modelagem ARIMA. Uma série é dita estacionária se ela desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável (MORETTIN; TOLOI, 2006). Portanto, a possibilidade de uma modelagem híbrida e automática expande e flexibiliza muito a gama de aplicações que podem ser atendidas em muitas áreas, trazendo muitos benefícios diretos à comunidade usuária.

1.3 Objetivo

Motivado pelas considerações apresentadas, este estudo propõe um novo modelo híbrido de previsão, composto pela análise *wavelets* auxiliada pela previsão multiescala de modelos probabilísticos clássicos, e compara seu desempenho de previsão com os modelos clássicos ARIMA e outros três modelos híbridos. Dentre os modelos híbridos confronta-

dos, três constituem-se dos modelos apresentados por Schlüter e Deuschle (2010), Wadi, Hamarsheh e Alwadi (2013) e Yousefi, Weinreich e Reinarz (2005), somados a pequenas alterações, apresentadas na subseção 2.2.1. Pretende-se realizar a comparação dos modelos a partir de diversas séries temporais simuladas com diferentes características, tanto determinísticas como estocásticas. Outro ponto importante será a automatização dos modelos, permitindo que sejam de fácil aplicação, mas mantendo a eficiência computacional. Além disso, pretende-se aplicar a metodologia proposta em dados epidemiológicos, a fim de averiguar a suposição da eficácia preditiva do modelo híbrido de previsão proposto em dados reais.

1.4 Estrutura do trabalho

A sequência estrutural deste trabalho segue com o capítulo 2, fornecendo uma introdução aos conceitos envolvidos na construção dos modelos. Os detalhes do procedimento de previsão multiescala baseada na transformada *wavelet* e a descrição do estudo de simulação é abordada na seção 2.2. Os resultados e discussões são apresentados na seção 2.3, enquanto, a seção 2.4 aborda as conclusões sobre este capítulo, envolvendo as simulações. No capítulo 3, a metodologia proposta é aplicada as taxas mensais de mortalidade infantil pertinente aos estados brasileiros e Distrito Federal. No capítulo 4, a metodologia proposta é aplicada nas taxas mensais de internações por bronquiolite referentes aos estados brasileiros e Distrito Federal. Devido ao fato do modelo híbrido proposto ter apresentado os melhores resultados em relação aos demais modelos no capítulo 2, nas aplicações dos capítulos 3 e 4, o modelo proposto é comparado somente ao modelo SARIMA, o qual é comumente utilizado na área de saúde e epidemiologia para a análise de séries temporais.

Capítulo 2

Métodos e Simulação

2.1 Métodos

2.1.1 Previsão baseada na esperança condicional

Os procedimentos de previsão abordados neste trabalho, levam em consideração apenas as informações inerentes aos dados, sem a mediação de outras fontes de informação. Assim, o interesse é "prever os valores futuros" x_{t+h}, $h \geq 1$, admitindo que as observações $x_t, x_{t-1}, x_{t-2}, \ldots$ são conhecidas. A previsão do valor x_{t+h} com origem no instante t e horizonte h será denotada por $\hat{x}_t(h)$ e definida como a esperança condicional de x_{t+h} dados todos os valores passados,

$$\hat{x}_t(h) = E(x_{t+h}|x_t, x_{t-1}, x_{t-2}, \ldots). \tag{2.1}$$

A igualdade (2.1) é chamada de *função de previsão* de h unidades de tempo a partir da origem t. A previsão fornecida pela esperança condicional possui o menor MSE (VANDAELE, 1983). Além disso, a sentença "prever os valores futuros" pode parecer redundante ao leitor, porém é possível prever valores não observados dentro do período amostral. Dessa forma, essa expressão refere-se aos valores previstos fora do intervalo observado. Uma técnica usual para avaliar o desempenho preditivo de modelos, é realizar previsões dentro do período amostral e compará-las aos valores reais observados, sendo o período amostral um *training period* do modelo de previsão.

É importante esclarecer que a previsão não constitui um fim em si, mas apenas um meio de fornecer informações para uma consequente tomada de decisão (MORETTIN; TOLOI, 1981).

2.1.2 Modelos ARIMA

Sendo as observações vizinhas de uma série temporal dependentes, o interesse é modelar a autocorrelação entre as observações da série X_t em diferentes tempos t. Contudo, existem características pertinentes a série que obscurecem o acesso a essa autocorrelação e precisam ser tratadas antes da modelagem, como tendência e sazonalidade. Em outras palavras, é preciso eliminar toda forma de não-estacionariedade da série temporal para modelar a autocorrelação.

A classe dos modelos ARIMA é capaz de modelar satisfatoriamente séries temporais com tendência, que serão transformada em uma série estacionária por meio de *diferenças* sucessivas simples (MORETTIN; TOLOI, 2006). Para séries temporais com tendência e/ou sazonalidade, recorre-se a uma extensão da classe dos modelos ARIMA composta pelos modelos SARIMA, abreviação do termo *Seasonal Autoregressive Integrated Moving Average model*. Os modelos SARIMA são expressos por,

$$\phi(B)\Phi(B^s)W_t = \theta(B)\Theta(B^s)\varepsilon_t, \tag{2.2}$$

em que o *operador de retardo* B é definido por $B^j X_t = X_{t-j}$, para todo j. Também, ε_t representa o ruído branco gaussiano, formado por uma sequência de variáveis aleatórias $\{\varepsilon_t\}$ independentes e identicamente distribuídas $(i.i.d)$, com média zero e variância σ^2. A expressão (2.2) denomina-se modelo SARIMA multiplicativo de ordem $(p, d, q)\ (P, D, Q)_S$ (MORETTIN; TOLOI, 1981).

Os modelos SARIMA realizam d diferenças sucessivas simples para remover a tendência e D diferenças sucessivas sazonais para remover a sazonalidade da série não-estacionária homogênea X_t, transformando-a em uma série estacionária W_t, expressa por

$$W_t = \nabla^d \nabla_S^D X_t. \tag{2.3}$$

Os valores p e P expressam o número de componentes autorregressivas simples $\phi(B)$ e sazonal $\Phi(B^s)$, respectivamente. Analogamente, os valores q e Q expressam o número de componentes de médias móveis simples $\theta(B)$ e sazonal $\Theta(B^s)$, respectivamente. Tais componentes são representadas por,

$$\phi(B) = (1 - \alpha_1 B - \ldots - \alpha_p B^p) \tag{2.4}$$

$$\Phi(B^s) = (1 - \pi_S B^S - \ldots - \pi_P B^{P_S}) \tag{2.5}$$

$$\theta(B) = (1 + \beta_1 B + \ldots + \alpha_q B^q) \tag{2.6}$$

$$\Theta(B^s) = (1 + \theta_S B^S + \ldots + \theta_Q B^{Q_s}). \tag{2.7}$$

Portanto, o vetor de parâmetros $\boldsymbol{\xi} = (\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$ deve ser estimado, em que $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_P)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)$. O processo de modelagem dos modelos SARIMA foi sistematizado por Box e Jenkins (1970). Essa tarefa consiste basicamente em preparação dos dados, identificação do modelo, estimação dos parâmetros, verificação da adequação do modelo e, se for o caso, previsão do comportamento da série.

2.1.3 Modelo *Spline*

Uma possibilidade ao procedimento de previsão de séries temporais consiste na extrapolação da tendência da série, por meio do ajuste de funções lineares nos parâmetros. Os modelos usualmente empregados utilizam o ajuste de funções polinomiais (MORETTIN; TOLOI, 1981). Contudo, funções polinomiais pode levar a previsões com erros elevados, devido a presença de comportamentos locais e bruscos não descritos por um polinômio.

Alternativamente, pode-se estimar e extrapolar a tendência da série temporal por meio do ajuste de uma curva segmentada, conhecida por função *spline*. A ideia é fragmentar o período amostral da série em intervalos, de modo que em cada intervalo seja ajustado um segmento polinomial. A união desses segmentos polinomiais ocorrem nos pontos extremos dos intervalos, chamados de *nós de interpolação*. Tal união resultará em uma curva suave, a função *spline*, dada pela minimização da curvatura da série temporal. Para o modelo *spline* cúbico, em cada intervalo da curva segmentada será ajustado polinômio cúbico.

Para a definição formal de funções *splines*, é preciso introduzir uma partição do intervalo arbitrário $[a, b]$ onde pretense-se trabalhar. No contexto de previsão, esse intervalo representa o período amostral série temporal. Uma partição I é definida pelos pontos $x_0, x_1, \ldots, x_m$ de forma que

$$I : a = x_0 < x_1 < \ldots < x_{m-1} < x_m = b, \tag{2.8}$$

e em cada subintervalo (x_{i-1}, x_i), $i = 1, 2, \ldots, m$, as *splines* são funções polinomiais por partes (SOUZA, 2008). Portanto, uma função real $s(x)$ é chamada de *spline* de grau n, com sequencia de nós $\{x_0, x_1, \ldots, x_m\}$, se (MORETTIN, 1999; SOUZA, 2008)

(i) $s(x)$ é um polinômio de grau n em cada subintervalo (x_{i-1}, x_i);

(ii) $s(x)$ tem $n - 1$ derivadas contínuas em cada x_i e, portanto, em $[a, b]$;

(iii) a derivada de ordem $n-1$ de $s(x)$ é uma função constante por partes, com saltos nos nós x_i.

2.1.4 Modelo Harmônico

O procedimento de previsão empregando funções harmônicas é baseado na estimação e extrapolação da tendência harmônica (ciclos determinísticos) da série temporal X_t. Portanto, um modelo de regressão é definido, cujo as variáveis preditoras são compostas pelas funções seno e cosseno. Assim, o modelo harmônico de previsão adotado é expresso por,

$$\begin{aligned} x_i &= g(t_i) + \varepsilon_{t_i}, \\ &= \alpha_0 + \alpha_1 cos(\omega t_i) + \alpha_2 sin(\omega t_i) + \alpha_3 cos(2\omega t_i) + \alpha_4 sin(2\omega t_i) + \varepsilon_{t_i}, \end{aligned} \tag{2.9}$$

em que ω é a frequência angular conhecida, t_i é a localização temporal das observações e ε_t é ruído branco gaussiano, com média zero e variância σ^2. A frequência angular $\omega = 2\pi f$ expressa o número de ciclos completos em 2π unidades de tempo, isto é, a frequência de ciclos em radianos por unidade de tempo. Por sua vez, f representa a frequência de ciclos por unidade de tempo. É válido observar que, no caso do modelo harmônico, o período das curvas seno e cosseno é fixo e igual a $p = 2\pi/\omega$.

A função $g(t)$ representa a componente sistemática do modelo, na qual os parâmetros $\{\alpha_i\}$ deverão ser estimados. O ruído branco ε_t representa a componente aleatória do modelo, a qual deverá ser predita. Pode-se afirmar que o ruído branco ε_t descreve a variação da série observada X_t em torno da função $g(t)$. Assim, é perfeitamente coerente admitir-se que o ruído branco distribui-se ao redor de $g(t)$ com média zero, a fim de que toda a variação explicada de X_t fique sob responsabilidade de $g(t)$. Nessa condição, o modelo harmônico pode ser utilizado para previsão dos valores de X_t correspondentes aos valores $g(t)$ não observados.

2.1.5 Análise Wavelets

A construção dos modelos da classe ARIMA é baseada no domínio do tempo, isto é, tais modelos descrevem o comportamento da série temporal X_t ao longo do tempo. Outra forma de descrever o comportamento de X_t consiste em considerar o domínio da frequência. A análise de Fourier, por exemplo, expressa os efeitos periódicos que contribuem no comportamento da série X_t por meio de funções trigonométricas, senos e cossenos, em diferentes escalas de frequência. Dessa forma, ao sobrepor essas funções trigonométricas, obtém-se uma aproximação de X_t, dada pela uma combinação linear de funções harmônicas. Portanto, dada uma série temporal, seu comportamento pode ser descrito ao longo

do tempo, como faz os modelos ARIMA ou ao longo de escalas de frequência, como faz a análise de Fourier.

A grande ênfase da análise *wavelets* deve-se a junção de tais possibilidades, sendo capaz de descrever a estrutura comportamental de uma série X_t tanto no domínio do tempo como no domínio da frequência, simultaneamente. Para isso, X_t é decomposta em diferentes níveis de resolução hierarquicamente, permitindo que a série seja descrita em termos de uma forma grosseira, mais outra que apresenta detalhes que vão desde os suaves aos mais finos (LIMA, 2002). O objetivo é aproximar a série X_t por uma combinação linear de funções *wavelets*, as quais são localizadas no tempo, contrariamente ao que ocorre com as funções trigonométricas, que são localizadas apenas na frequência. A propriedade de localização temporal, torna as *wavelets* ideias para analisar séries não-estacionárias (MORETTIN, 1999).

Dificilmente, séries temporais reais apresentarão efeitos periódicos bem comportados, sendo frequente a presença de efeitos descontínuos ou com "pontas acentuadas", que exigem um grande número de combinações de senos e cossenos para serem representados razoavelmente (MORETTIN, 1999). Assim, a análise *wavelets* é vantajosa, devido sua capacidade de representar efeitos não estacionários com pequeno número de combinações de funções *wavelets*.

Semelhantemente a uma árvore genealógica, que revela as origens genéticas que contribuíram na formação de um indivíduo, a análise *wavelet* revela as origens comportamentais que contribuíram na formação de uma série temporal X_t. Assim, os ramos da árvore genealógica são caracterizados pelos níveis de resolução hierárquicos, nos quais a série temporal X_t é decomposta. Em síntese, o cerne da análise *wavelets* é construir a “árvore genealógica” da série temporal, permitindo o acesso aos “perfis” dos efeitos periódicos que contribuíram “geneticamente” em sua estrutura comportamental. Tal abordagem revela a presença e localização temporal de efeitos, que muitas vezes não são detectados em análises baseadas no domínio do tempo ou domínio da frequência, separadamente. Como exemplo, aqui é possível detectar pontos de mudança ou período sazonal, bem como sua localização temporal.

Por fim, a análise *wavelets* é composta pelas seguintes etapas: decomposição da série em coeficientes *wavelets* (subseções 2.1.5.2 e 2.1.5.4), limiarização (subseção 2.1.5.5) e reconstrução da série temporal (VIDAKOVIC; MUELLER, 1994).

2.1.5.1 O modelo *wavelet*

Considere o seguinte modelo de regressão,

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.10}$$

onde $\varepsilon_i \sim i.i.d.N(0, \sigma^2)$ e $t_i = (i-1)/n$ (MORETTIN, 1999). Suponha que, a função de regressão f pertença ao espaço de funções de quadrado integrável sobre o conjunto dos números reais $\mathbb{R}$, denotado por $L^2(\mathbb{R})$ e composto por funções cuja a integral de f^2 definida em $\mathbb{R}$ é finita. Qualquer função $f \in L^2(\mathbb{R})$ pode ser escrita como,

$$f(t) = \sum_k c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j \geq j_0} \sum_k d_{j,k}\psi_{j,k}(t), \tag{2.11}$$

em que

$$c_{j_0,k} =< f(t), \phi_{j_0,k}(t) >= \int_{-\infty}^{\infty} f(t)\phi_{j_0,k}(t)dt, \tag{2.12}$$

$$d_{j,k} =< f(t), \psi_{j,k}(t) >= \int_{-\infty}^{\infty} f(t)\psi_{j,k}(t)dt, \tag{2.13}$$

sendo $c_{j_0,k}$ e $d_{j,k}$ chamados de coeficientes *wavelets*, que codificam o comportamento de f em cada nível de resolução. Tal codificação é possível por meio da *wavelet* pai ϕ e *wavelet* mãe ψ, responsáveis pela construção dos níveis de resolução envolvidos na análise de multirresolução descrita na subseção 2.1.5.2. Os coeficientes *wavelets* $c_{j_0,k}$ são relacionados às escalas de baixa frequência e obtidos por translações inteiras da *wavelet* pai $\phi(t)$. Por sua vez, os coeficientes *wavelets* $d_{j,k}$ são relacionados às escalas de alta frequência e resultam de translações e dilatações inteiras da *wavelet* mãe $\psi(t)$.

Ademais, os coeficientes $c_{j_0,k}$, resultantes da *wavelet* pai, descrevem os comportamentos suaves, como tendência. Os coeficientes $d_{j,k}$, obtidos pela *wavelet* mãe, descrevem os comportamentos mais ruidosos, como influências sazonais e ruído. Portanto, o interesse da análise *wavelets* é estimar os coeficientes $c_{j_0,k}$ e $d_{j,k}$ a partir dos dados que constituem a série temporal X_t.

2.1.5.2 Análise de Multirresolução

Uma análise de multirresolução é uma sequência crescente, $\{V_j; j \in \mathbb{Z}\}$, de subespaços fechados de $L^2(\mathbb{R})$, representando os sucessivos níveis de decomposição, tais que eles satisfaçam às seguintes condições (MALLAT, 1989a):

MR1 $... \subset V_{-1} \subset V_0 \subset V_1 \subset ...$

MR2 $L^2(\mathbb{R}) = \overline{\bigcup_j V_j}$,

MR3 $\bigcap_j V_j = \lim_{j \longrightarrow -\infty} V_j = 0$,

MR4 $f(t) \in V_j \Leftrightarrow f(2t) \in V_{j+1}, \forall j$,

MR5 $V_{j+1} = V_j \oplus W_j, W_j \perp V_j$,

MR6 Existe $\phi \in L^2(\mathbb{R})$, denominada função escala, tal que $\{\phi(x-k); k \in \mathbb{Z}\}$ constitui uma base ortonormal para V_0.

A princípio, tem-se que os subespaços da análise de multirresolução são formados por funções que aproximam $f \in L^2(\mathbb{R})$. A melhor aproximação é obtida pela projeção ortogonal de f sobre cada subespaço V_j, a qual é denotada por (P_j) (MORETTIN, 1999). Portanto, sendo $g(t)$ uma função aproximante qualquer de f, então

$$\forall g(t) \in V_j, \quad ||g(t) - f(t)|| \geq ||P_j(t) - f(t)||. \tag{2.14}$$

Posto isso, a condição (MR1) exige que os subespaços fechados devem ser encaixados, de maneira que a cada nível de resolução sejam acrescentadas informações e não alteradas. O acréscimo de informações resulta na presença de novos "detalhes". Assim, conforme o valor de j avança, consequentemente, aumentamos a resolução e mais próxima a função aproximante será de f (MR2). Logo, a união desses subespaços deve fornecer cobertura total do espaço $L^2(\mathbb{R})$ que contém f, cuja a integral de f^2 definida em $\mathbb{R}$ é finita.

Inversamente, conforme o valor de j retrocede, diminui-se a resolução e a função aproximante perderá informações sobre f (MR3). Além disso, todos os subespaços estão relacionados por escala a um mesmo subespaço V_0, considerado o espaço de referência para os níveis de resolução. O subespaço V_0 é o nível mais suave, o qual representa a estrutura da série temporal observada X_t, dispensando os detalhes. Portanto, o acréscimo de informação a cada nível, iniciando-se em V_0, conduzirá de uma aproximação grosseira para uma aproximação com maior resolução da série temporal de interesse (LIMA, 2002). Dessa forma, detalhes que aparecem no subespaço V_j também devem estar presentes no subespaço V_{j+1}, justificando a condição (MR4) (OLIVEIRA, 2007).

Os subespaços V_j são gerados por translações inteiras $\phi(x-k)$, $k \in \mathbb{Z}$ da *wavelet*-pai ϕ, também conhecida por função escala. A função ϕ expressa por

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k), \tag{2.15}$$

fornece uma base ortonormal $\{\phi_{j,k}; j, k \in \mathbb{Z}\}$ para V_0 (MR6). A existência da função $\phi \in L(\mathbb{R})$ que fornece tal base ortonormal foi provada por (MALLAT, 1989a; MALLAT, 1989b). As bases ortonormais para os demais subespaços V_j são obtidas a partir de V_0, escalonando ϕ por 2^j. Por exemplo, $\{\sqrt{2}\phi(2t-k); j, k \in \mathbb{Z}\}$ será uma base ortonormal para o subespaço V_1 e assim por diante. Isso justifica o nome "multirresolução", significando que a resolução obtida em cada nível é múltipla da resolução inicial do nível V_0, sendo este um fator comum a todos os níveis de resolução.

A informação acrescentada ao "subir" de nível é representada pelo subespaço W_j. Tal acréscimo de detalhes é realizado pela *wavelet* mãe ψ. Assim, o subespaço W_j armazena as informações acrescidas ao mudar do nível V_j para V_{j+1}, de modo que W_j é o complemento ortogonal de V_j em V_{j+1} (MR5), gerado por translações e dilatações (ou compressões) inteiras da *wavelet* mãe. A função ψ expressa por

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \tag{2.16}$$

fornece uma base ortonormal $\{\psi_{j,k}; j, k \in \mathbb{Z}\}$ para W_0. As bases ortonormais para os demais subespaços W_j são obtidas a partir de W_0, escalonando ψ por 2^j. Por exemplo, $\{\sqrt{2}\psi(2t-k); j, k \in \mathbb{Z}\}$ será uma base ortonormal para o subespaço W_1, $\{2\psi(2^2t-k); j, k \in \mathbb{Z}\}$ será uma base ortonormal para o subespaço W_2, $\{\sqrt{2^3}\psi(2^3t - k); j, k \in \mathbb{Z}\}$ será uma base ortonormal para o subespaço W_3 e assim por diante.

Os movimentos de translação e dilatação fornecem vantagens a análise *wavelets* sobre a análise de Fourier, para séries não estacionárias. As dilatações, propiciam a capacidade das *wavelets* distinguir características locais de uma série em diferentes níveis de resolução, e as translações permitem cobrir todo o espaço na qual a série é estudada (LIMA, 2002).

Cada nível de resolução V_j pertinente a decomposição *wavelet* é resultado de um "casamento" entre *wavelet*-pai ϕ e *wavelet* mãe ψ. Dado que, os subespaços V_j e W_j são ortogonais, o subespaço V_{j+1} é fruto do "casamento" representado pela soma direta $V_{j+1} = V_j \oplus W_j$, em que V_j representa o comportamento suave e W_j os detalhes da série observada X_t no nível de decomposição j (MR5). A Figura 1 ilustra a decomposição *wavelet* em subespaços de detalhes W_j e subespaços suaves V_j.

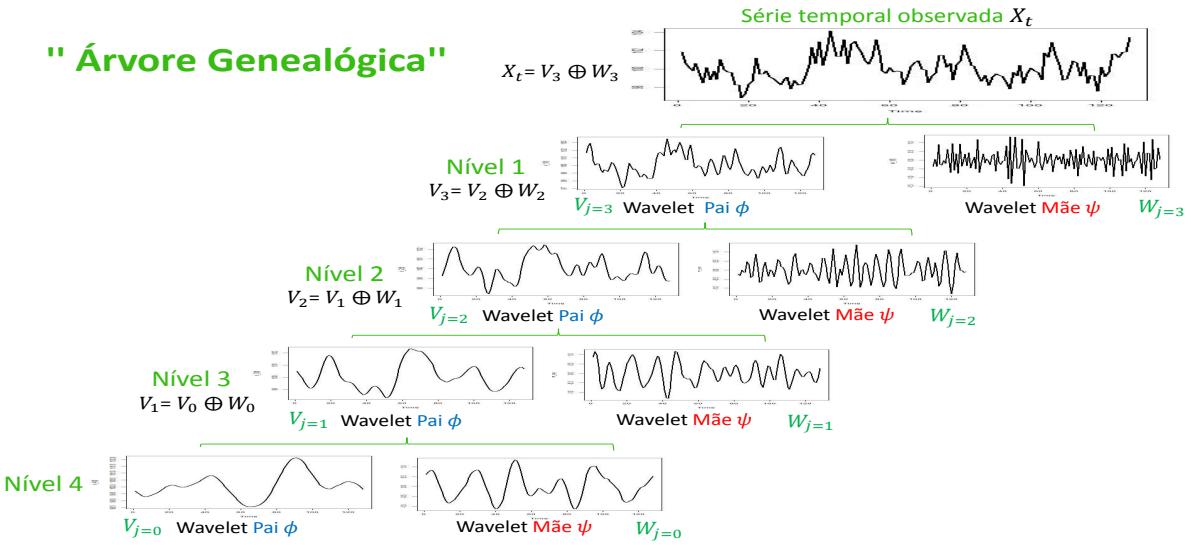

Figura 1 – Ilustração da análise de multirresolução: decomposição da série em níveis de resolução ou multiescala.

2.1.5.3 *Wavelets* mãe

As *wavelets* podem ser suaves ou não, ter suporte compacto ou não, ser simétricas ou não e podem ter expressões matemáticas simples ou não (MORETTIN, 1999). A suavidade da *wavelet* contribui para uma reconstrução perfeita da série temporal X_t e está relacionada ao número de momento nulos da *wavelet*. Uma *wavelet* possui P momentos nulos se, e somente se,

$$M_k = \int_{-\infty}^{+\infty} t^k \psi(t) dt = 0, \quad 0 \leq k < P, \tag{2.17}$$

em que M_k representa o momento de ordem k da *wavelet* (MALLAT, 2008). Quanto maior o número de momentos nulos P, maior será a suavidade da *wavelet*. Ainda mais, se a *wavelet* possui P momentos nulos, então os coeficientes *wavelets* de qualquer polinômio $f \in L^2(\mathbb{R})$ com grau menor ou igual a $P-1$ serão nulos. Isso se deve ao fato da *wavelet* ser ortogonal ao polinômio com grau $P-1$ (MALLAT, 2008). Assim, o produto interno, que resulta nos coeficientes *wavelets*, será $< f(t), \psi(t) >= 0$. Portanto, P indica o maior grau do polinômio interpolador de f que pode ser representado, cujos coeficientes *wavelets* não serão nulos. Geralmente, as *wavelets* mãe são classificadas em famílias de acordo com o número de momentos nulos (OLIVEIRA, 2007).

Enquanto que, as funções senos e cossenos são ondas periódicas que continuam infinitamente, as funções *wavelets* são ondas que "viajam" por um ou mais períodos e são definidas em um intervalo finito, ou seja, possuem suporte compacto. Uma função ψ tem suporte compacto, se existe um intervalo fechado e limitado, fora do qual $\psi(t) = 0$. Além disso, o acréscimo de momentos nulos P aumenta o tamanho do suporte das *wavelets*. Deste modo, se ψ possui $L = 2P$ coeficientes não-nulos, então ψ terá suporte em $[-P+1, P]$ (MORETTIN, 1999).

Nesse momento, o leitor pode questionar sobre a ausência da *wavelet* pai ao tratar da escolha da *wavelet*. Na realidade, selecionada a *wavelet* mãe, esta vem acompanhada por seu "cônjuge", a *wavelet* pai. Melhor dizendo, a *wavelet* mãe ψ é escrita em função da *wavelet* pai ϕ. Para melhor compreender esse relacionamento, note que a função ϕ pode ser representada por,

$$\phi(t) = \sqrt{2} \sum_k l_k \phi(2t - k), \tag{2.18}$$

gerando a família ortonormal $\{\phi_{j,k}; j, k \in \mathbb{Z}\}$ em $L^2(\mathbb{R})$ expressa em (2.15). Então, a função ψ é obtida a partir da função ϕ da seguinte forma,

$$\psi(t) = \sqrt{2} \sum_k h_k \phi(2t - k), \tag{2.19}$$

em que $h_k = (-1)^k l_{1-k}$ é chamado de "relação do filtro de quadratura" (*quadrature mirror filter relation*). Também, h_k representa o filtro passa-alta e l_k expressa o filtro passa-baixa, ambos conhecidos como "filtros de quadratura" (*quadrature mirror filters*), usados para calcular a transformada *wavelet* discreta (MORETTIN, 1999). As equações (2.18) e (2.19) são chamadas "Equações de Dilatação".

2.1.5.4 Transformada *Wavelet* Discreta

A transformada *wavelet* fornece as subséries de coeficientes *wavelets* que codificam o comportamento da série temporal X_t em cada nível de resolução da análise de multirresolução. Seja X_t uma série temporal composta por $n = 2^J, J > 0$ observações. A transformada *wavelet* discreta de X_t segundo a *wavelet* mãe ψ pode ser expressa por,

$$d_{j,k}^{(\psi)} = \sum_{t=0}^{n-1} X_t \psi_{j,k}(t/n), \tag{2.20}$$

sendo que a equação (2.20) seria calculada para $j = 0, 1, \ldots, J$ e $k = 0, 1, \ldots, 2^j$, com $J \in \mathbb{Z}$ representando o nível de detalhes mais fino (MORETTIN, 1999). Entretanto, na realidade, os coeficientes *wavelets* são estimados pelo algoritmo piramidal, no qual os coeficientes de cada nível são obtidos dos coeficientes do nível anterior (MALLAT, 2008). Utilizando a equação da função escala ϕ dada em (2.18) e a definição da *wavelet* ψ expressa em (2.19) tem-se

$$d_{j,k} = \langle f, \psi_{j,k} \rangle = \sum_{n \in \mathbb{Z}} h_{n-2k} \langle f, \phi_{j+1,n} \rangle, \tag{2.21}$$

$$d_{j,k} = \sum_{n \in \mathbb{Z}} h_{n-2k} c_{j+1,n}, \tag{2.22}$$

$$c_{j,k} = \langle f, \phi_{j,k} \rangle = \sum_{n \in \mathbb{Z}} l_{n-2k} \langle f, \phi_{j+1,n} \rangle, \quad e \tag{2.23}$$

$$c_{j,k} = \sum_{n \in \mathbb{Z}} l_{n-2k} c_{j+1,n}. \tag{2.24}$$

As quatro equações anteriores representam a base do algoritmo piramidal de Laplace, também chamado de algoritmo "piramidal" por Mallat (1989) ou "em cascata", por Daubechies (1992) (MALLAT, 2008; SOUZA, 2008; MORETTIN, 1999). Observe que, o algoritmo calcula os coeficientes $c_{j,k}$ e $d_{j,k}$ do *j-ésimo* nível a partir dos coeficientes suaves do nível $j + 1$, expressos por $c_{j+1,k}$. Considerando a notação $L_j t = \sum_{n \in \mathbb{Z}} l_{n-2k} t_n$ e $H_j t = \sum_{n \in \mathbb{Z}} h_{n-2k} t_n$, os coeficientes *wavelets* podem ser escritos como,

$$c_{j,k} = L_j c_{j+1,k} \quad e \quad d_{j,k} = H_j c_{j+1,k}, \tag{2.25}$$

em que L_j pode ser interpretado como o filtro passa-baixa e H_j um filtro passa-alta (SOUZA, 2008). Dentre as transformadas discretas, tem-se: a transformada *wavelet* discreta decimada (TWD) e a transformada *wavelet* discreta não decimada (TWDND).

Na transformada *wavelet* discreta decimada, eliminam-se as amostras pares ou ímpares, dependendo da escolha da "origem", de maneira que no nível j possui a metade dos coeficientes do nível $j+1$, justificando o nome "piramidal" (MALLAT, 1989b). Esse processo de filtragem de informações é conhecido por decimação diádica. Nas expressões (2.22) e (2.24) nota-se que os coeficientes $c_{j,k}$ e $d_{j,k}$ são obtidos de $c_{j+1,n}$ por médias móveis amostradas nos inteiros pares, que é a decimação (MORETTIN, 1999). Para a reconstrução da série, a transformada *wavelet* discreta inversa (TWDI) utiliza os mesmos coeficientes em ordem reversa e ao invés da decimação, as componentes da série são alongadas pela interpolação de zeros entre os coeficientes da transformada. Sendo assim, a transformada decimada depende da escolha da origem, visto que os coeficientes *wavelets* fornecidos pelos elementos pares resulta em uma base ortogonal distinta da base obtida ao considerar os elementos ímpares (NASON; SILVERMAN, 1995).

Uma maneira de contornar essa vulnerabilidade ao ponto inicial, consiste em dispensar a decimação diádica, mantendo os n coeficientes em cada nível de resolução. A transformada *wavelet* discreta não-decimada, mantém os elementos pares e ímpares da amostra, evitando a perda de informações. Nessa perspectiva, a expressão da *wavelet* mãe não-decimada, equivalente à igualdade (2.16), é dada por,

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j(t-k)), \quad j,k \in \mathbb{Z}. \tag{2.26}$$

Pode-se implementar a TWDND, aplicando o algoritmo piramidal duas vezes na série transladada de um elemento. Naturalmente, o esforço computacional exigido para o cálculo da TWDND é maior em relação a TWD. Contudo, o algoritmo para calcular a TWDND ainda é considerado rápido (NASON, 2010). Os coeficientes suaves $c_{j,k}$ e os coeficientes de detalhes $d_{j,k}$ são obtidos de c_{j+1}, mas elimina-se a decimação e modificam-se os filtros L_j e H_j a cada nível pela inserção de zeros ((NASON; SILVERMAN, 1995)).

2.1.5.5 *Wavelet Denoising*

A *wavelet denosing* ou limiarização permite descrever o comportamento da série temporal com o menor número de coeficientes *wavelets* possível, garantindo a parcimônia do modelo sem prejudicar a representação da série. O objetivo é reduzir ou remover o ruído presente na série X_t, diminuindo ou zerando a magnitude dos coeficientes *wavelets*, de modo que a reconstrução expresse claramente as características essenciais da série observada (MORETTIN, 1999). A limiarização pode ser aplicada em todos os níveis de decomposição ou apenas em níveis associados às escalas de alta frequência, onde

concentram-se as contaminações por ruídos. Por sua vez, os ruídos podem ser vistos como os detalhes minuciosos que não contribuem na formação da estrutura comportamental da série, e por isso podem ser eliminados.

A etapa de *denoising* ou limiarização, ocorre após a etapa da decomposição *wavelet* pela TWD, quando são obtidos os coeficientes *wavelets* $d_{j,k}^{(\psi)}$ contaminados por ruídos. Assim, elimina-se os coeficientes *wavelets* menores do que um determinado valor, chamado de limiar ou *threshold* λ, obtendo os coeficientes *wavelets* isentos de ruído. Por fim, a série é reconstruída pela TWDI, fornecendo uma aproximação "limpa" de ruído $\hat{X}_t$ da série temporal X_t. O procedimento de limiarização é ilustrado no fluxograma a seguir.

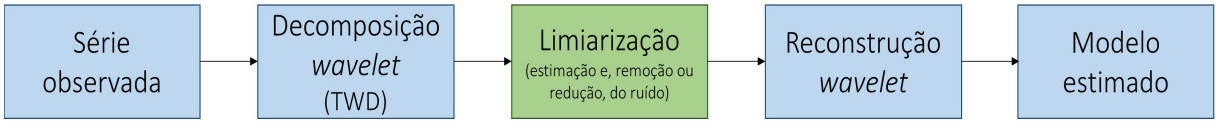

Figura 2 – Etapas da análise *wavelet* (VIDAKOVIC; MUELLER, 1994).

A princípio, é necessário estimar o limiar de corte λ, que irá estabelecer o valor mínimo para os coeficientes *wavelets* permanecerem no modelo. Dito em outras palavras, o valor do limiar irá determinar quais coeficientes *wavelets* são considerados ruídos ou não. Entretanto, existem duas vertentes de limiares, a global e a adaptativa (KOZAKEVICIUS; BAYER, 2014). O limiar global independe do nível j, isto é, o valor de λ permanece constante para todos os níveis de decomposição j. Na perspectiva adaptativa, o limiar habitua-se ao nível j no qual será aplicado, sendo mais flexível e refinado.

2.1.5.5.1 Global (Limiar Universal)

O limiar *universal*, proposto por (DONOHO; JOHNSTONE, 1994), permanece inalterável independente do nível de decomposição j a qual será aplicado. Sua expressão é dada por,

$$\lambda_n = \lambda_{j,n} = \hat{\sigma}\sqrt{2\log n}, \tag{2.27}$$

em que $\hat{\sigma}$ é a estimativa do desvio-padrão do ruído, o qual deve ser estimado a partir dos dados, e n é o número de observações da série observada. Para estimar o nível de ruído $\hat{\sigma}$, (DONOHO; JOHNSTONE, 1994) propuseram o seguinte estimador,

$$\hat{\sigma} = mediana\{|d_{J-1,k}| : 0 \leq k \leq n/2\}/0.6745, \tag{2.28}$$

em que $J-1$ é a escala mais fina e o fator 0.6745 é determinado estatisticamente em (DONOHO et al., 1995). A tendência do limiar universal é reduzir muitos coeficientes, especialmente em níveis associados às escalas de alta frequência (MORETTIN, 1999).

2.1.5.5.2 Global (Uma alternativa)

O acréscimo de detalhes a cada nível é realizado pela *wavelet* mãe ψ e armazenados nos subespaços W_j, gerados por translações e dilatações inteiras da função ψ. Assim, o nível mais alto W_j expressa grande parte do ruído presente em uma série. Sendo assim, outra alternativa de limiar, consiste em zerar os coeficientes *wavelets* do subespaço W_j associado às escalas de alta frequência. O subespaço W_j representa o último nível de detalhes da decomposição *wavelet*, abrigando uma considerável parte dos ruídos presentes na série temporal.

Para exemplificar, na Figura 1, o último nível de detalhes será o subespaço $W_{j=3}$, o qual expressa os detalhes mais minuciosos da série temporal X_t. Dessa forma, o limiar aqui considerado, consiste em zerar os coeficientes *wavelets* desse nível, eliminado o subespaço $W_{j=3}$, do procedimento de reconstrução da série. Portanto, a série $\hat{X}_t$ será uma representação esparsa da série observada X_t, possuindo menor número de coeficientes *wavelets* estimados.

2.1.5.5.3 Adaptativo (*SureShrink*)

Também proposto por (DONOHO; JOHNSTONE, 1995), o procedimento *SureShrink* baseia-se na minimização de um estimador não-viesado do risco de Stein em cada nível de decomposição. Dado o nível j com n_j coeficientes *wavelets* e parâmetro de dispersão σ_j, o limiar é expresso por

$$\lambda_j = \text{argmin}_{0 \leq t \leq \sqrt{2 \log n_j}} \text{SURE}(d_j, t), \tag{2.29}$$

dado que

$$\text{SURE}(d_j, t) = n_j - 2 \sum_{k=1}^{n_j} I_{\{|d_{j,k}| \leq t\sigma_j\}} + \sum_{k=1}^{n_j} \min \left\{ \left(\frac{d_{j,k}}{\sigma_j} \right)^2, t^2 \right\}, \tag{2.30}$$

em que n é o comprimento da série observada e $d_{j,k}$ são os coeficientes *wavelets*. Dessa forma, a cada nível de decomposição tem-se um valor λ_j específico, o qual é estimado a partir dos coeficientes *wavelets* d_j pertinentes ao nível j (MORETTIN, 1999).

2.2 Metodologia

Neste estudo, foi comparado o desempenho preditivo dos modelos SARIMA com quatro modelos híbridos, dados pela combinação entre TWD e modelos clássicos. A escolha dos modelos SARIMA se deve ao fato destes serem largamente utilizados para a previsão de séries temporais nas diversas áreas do conhecimento. Com relação aos quatro modelos híbridos, três foram eleitos como melhores modelos de previsão em (YOUSEFI; WEINREICH; REINARZ, 2005; SCHLÜTER; DEUSCHLE, 2010; WADI; HAMARSHEH; ALWADI, 2013). Contudo, tais modelos sofreram pequenas adaptações ao contexto deste estudo, as quais são descritas na subseção 2.2.1. O quarto modelo híbrido a ser comparado é desconhecido pela literatura e refere-se a proposta deste trabalho.

2.2.1 Os modelos híbridos

O primeiro modelo híbrido (ou modelo híbrido 1), é composto pelo *wavelet denoising* segundo o método *SureShrink* seguido pelo ajuste e previsão do modelo ARIMA na série reconstruída, e foi apresentado por (SCHLÜTER; DEUSCHLE, 2010). A proposta original executa a etapa de limiarização via transformada *wavelet* contínua (TWC) e algoritmo *a tróus*. Para a aplicação no presente estudo, a TWC e algoritmo *a tróus* foram substituídos pela TWD e algoritmo piramidal. O motivo de tal adaptação deve-se a restrição deste trabalho ao emprego da transformada *wavelet* discreta. Como alternativa, poderia substituir a TWC pela TWDND, também abordada neste estudo. Porém, o procedimento preditivo ocorre após a reconstrução da série e não em multiescala, sendo dispensável o uso da TWDND. Além disso, vale ressaltar que a TWD é a mais adequada para realizar o procedimento de *denoising*, o qual dispensa transformadas redundantes como a TWC ou a TWDND. O termo redundante refere-se ao número de coeficientes excedentes fornecidos pela transformada, comprometendo sua parcimônia. Além disso, dois tipos de *denoising* foram acrescentados ao método *SureShrink*, estes são: limiar universal e o limiar alternativo, ambos descritos na subseção 2.1.5.5. O acréscimo de tais procedimentos tem o intuito de comparar a contribuição dos mesmos no desempenho de previsão do modelo híbrido 1.

Duas adaptações foram necessárias ao segundo modelo híbrido (ou modelo híbrido 2), proposto por (WADI; HAMARSHEH; ALWADI, 2013). Tal modelo consiste da TWDND seguida pelo ajuste e previsão de modelos ARIMA nas subséries de coeficientes *wavelets* de alta e baixa frequências. A primeira adaptação refere-se ao número de níveis de resolução: a proposta original aborda a TWDND com 1 nível de resolução, neste estudo a mesma é abordada com 4, 5, e 6 níveis de resolução. A segunda adaptação concerne a substituição dos modelos da classe ARIMA por modelos da classe SARIMA.

Visando a automatização do modelo híbrido 2, o ajuste multiescala dos modelos

SARIMA é realizado pela função *auto.arima* do *software* R, a qual retorna o modelo que minimiza os critérios de informação AIC, AIC_c e BIC. Por *default*, a seleção do modelo é baseada no procedimento *stepwise*, o qual propõe quatro possíveis modelos para iniciar o algoritmo. Essa proposta inicial restringe e torna finito o espaço de modelos. Assim, a função *auto.arima* retornará o "melhor modelo" pertinente a um espaço de modelos finito, que não será necessariamente o melhor modelo de todos. É possível realizar uma busca completa, ou seja, considerar um espaço de modelos mais amplo, basta desconsiderar o procedimento *stepwise*. Contudo, tal opção pode ser muito lenta, principalmente para modelos sazonais, fugindo do escopo desse trabalho justificado na subseção 1.2. Detalhes sobre essa e outras restrições da função *auto.arima* são esclarecidas em (HYNDMAN; KHANDAKAR, 2007).

O terceiro modelo híbrido (ou modelo híbrido 3) consiste da TWD seguida pelo ajuste e extensão multiescala de funções harmônicas e função *spline*, e foi sugerido por (YOUSEFI; WEINREICH; REINARZ, 2005). A proposta original realiza a TWD em 5 níveis de resolução, de modo que as funções harmônicas são aplicadas nas subséries de coeficientes *wavelets* pertinentes aos níveis $\{W_i\}_{i=1}^{4}$ e a função *spline* nos níveis W_5 e V_5. No presente trabalho, é utilizado funções harmônicas em todas as subséries de alta frequência $\{W_i\}_{i=1}^{5}$ e função *spline* cúbica apenas na subsérie de baixa frequência pertinente ao nível V_5. Apesar da TWD atender a restrição deste estudo, é vantajoso substituí-la pela TWDND, devido ao procedimento de previsão multiescala. O interesse do procedimento de previsão multiescala na TWDND, remete a estabilidade do número de coeficientes *wavelets* em cada nível de resolução, fornecendo o maior o número de informações.

O quarto modelo híbrido (ou modelo híbrido 4), proposto neste estudo, constitui-se da análise *wavelet* seguida pelo ajuste e previsão multiescala de modelos probabilísticos clássicos, os quais possuem propriedades teóricas capazes de descrever e prever satisfatoriamente subséries de coeficientes *wavelets*, independente da série temporal decomposta. Outra característica relevante do modelo híbrido 4 refere-se a sua automaticidade. Os modelos a serem ajustados em cada subsérie de coeficientes *wavelets* são predeterminados e fixos, evitando a busca por modelos que forneçam um bom ajuste e previsão para cada subsérie, automatizando sua aplicação.

O desempenho preditivo dos modelos híbridos também foram analisados segundo 6 famílias *wavelets*, com diferentes números de coeficientes não nulos $L = 2P$, totalizando 25 *wavelets* mãe. Assim, as *wavelets* adotadas neste estudo, foram: *Best Localized* 18 (BL18), *Best Localized* 20 (BL20), Coiflet 6 (C6), Coiflet 12 (C12), Coiflet 18 (C18), Coiflet 24 (C24), Daubechies 2 (D2), Daubechies 4 (D4), Daubechies 6 (D6), Daubechies 8 (D8), Daubechies 12 (D12), Daubechies 16 (D16), Daubechies 18 (D18), Daubechies 20 (D20), Fejer-Korovkin 4 (FK4), Fejer-Korovkin 6 (FK6), Fejer-Korovkin 8 (FK8), Fejer-Korovkin 22 (FK22), *Least Asymetric* 8 (LA8), *Least Asymetric* 12 (LA12), *Least Asymetric* 18 (LA18),

Least Asymetric 20 (LA20), *Minimum-Bandwidth* 4 (MB4), *Minimum-Bandwidth* 8 (MB8) e *Minimum-Bandwidth* (MB24). Tais *wavelets* encontram-se disponíveis nos pacotes *wavelets* e *waveslim* do *software* R. Maiores detalhes sobre as *wavelets* são encontrados em (MALLAT, 2008; DAUBECHIES et al., 1992; VIDAKOVIC, 1999).

A seguir, tem-se uma descrição sucinta da abordagem dos modelos híbridos e clássico SARIMA:

Modelo híbrido 1 - Denoising + SARIMA: inicia-se com procedimento de *denoising* via TWD, a partir de três limiares distintos: *SureSrhink*, limiar universal e limiar alternativo (ver subseção 2.1.5.5). Todos os limiares foram aplicados apenas no nível mais alto W_j associado às escalas de alta frequências. Em seguida, realiza-se a reconstrução da série via TWDI, fornecendo uma aproximação da série temporal "limpa" de ruído. Por fim, efetua-se o ajuste e previsão do modelo SARIMA na série reconstruída. A TWD presente na composição do modelo híbrido 1, é abordada com 4, 5 e 6 níveis de resolução segundo diferentes *wavelets* mãe.

Modelo híbrido 2 - TWDND + SARIMA multiescala: inicia-se com a TWDND, obtendo níveis de resolução compostos por subséries de coeficientes *wavelets* com mesmo tamanho. Após, a modelagem individual de cada subsérie é realizada por meio do ajuste e previsão de modelos clássicos SARIMA. Por fim, a TWDND inversa é aplicada para a reconstrução da série aproximada junto a previsão efetuada em multiescala. A TWDND presente no modelo híbrido 2 é abordada com 4, 5 e 6 níveis de resolução a partir de diferentes *wavelets* mãe.

Modelo híbrido 3 - TWDND + Harmônico e *Spline* multiescala: incia-se com a TWDND, obtendo níveis de resolução compostos por subséries de coeficientes *wavelets* com mesmo tamanho. Após, a modelagem individual de cada subsérie é realizada por meio do ajuste e previsão de modelos Harmônicos para as subséries de alta frequência W_j (ver subseção 2.1.4). Para a subsérie pertinente ao nível mais suave V_j de baixa frequência, realiza-se o ajuste e previsão de modelos *Spline* (ver subseção 2.1.3). Por fim, a TWDND inversa é aplicada para a reconstrução da série aproximada junto a previsão efetuada em multiescala. A TWDND presente no modelo híbrido 3 é abordada com 4, 5 e 6 níveis de resolução a partir de diferentes *wavelets* mãe.

Modelo híbrido 4 - TW + modelos probabilísticos clássicos multiescala: inicia-se com a transformada *wavelet* (TW), obtendo níveis de resolução compostos por subséries de coeficientes *wavelets*. Após, a modelagem individual de cada subsérie é realizada por meio do ajuste e previsão de modelos probabilísticos clássicos na análise de séries temporais. Por fim, a transformada *wavelet* inversa (TWI) é aplicada para a reconstrução da série aproximada junto a previsão efetuada em multiescala. A TW

presente no modelo híbrido 4 é abordada com 4, 5 e 6 níveis de resolução a partir de diferentes *wavelets* mãe.

Modelo 5 - clássico SARIMA: ajuste do modelo clássico SARIMA na série temporal de acordo com as etapas da metodologia de Box & Jenkins.

2.2.2 Séries Simuladas

A avaliação do desempenho preditivo dos modelos descritos, tanto clássico como híbridos, deu-se pela aplicação destes em 21 séries temporais simuladas a partir de 3 séries com características particulares, denotadas por X_t^1, X_t^2 e X_t^3, respectivamente. Por sua vez, as expressões dessas 3 séries são escritas em função de uma série temporal de referência X_t, como segue:

$$X_t^1 = 1.15X_t + \frac{t}{200} + \sin\left(\frac{3\pi(t-1)}{180}\right) + 100, \tag{2.31}$$

$$X_t^2 = 0.01(X_t)^2 + 0.16X_t + \frac{t^2}{500000} + \sin\left(\frac{3\pi(t-1)}{180}\right), \tag{2.32}$$

$$X_t^3 = 0.05X_t + \sin\left(\frac{3\pi(t-1)}{10}\right). \tag{2.33}$$

Assim, X_t^1 é uma série temporal com tendência linear e periodicidade, X_t^2 é uma série temporal com componente não linear, tendência e periodicidade, e por fim, X_t^3 uma série com componente sazonal. A série de referência X_t foi utilizada como pivô para a investigação do desempenho preditivo dos modelos em diferentes cenários de simulação. Quer dizer, em cada cenário foi simulada a série X_t e, em seguida, esta foi substituída nas expressões (2.31), (2.32) e (2.33), obtendo as 3 séries temporais não-estacionárias. Por exemplo, após a série de referência X_t ser simulada a partir do processo AR(1) com coeficiente autorregressivo $\alpha_1 = 0.5$, esta foi substituída nas expressões (2.31), (2.32) e (2.33), obtendo as três séries temporais X_t^1, X_t^2 e X_t^3 com características determinísticas. O mesmo foi realizado para os demais cenários de simulação. Dessa forma, mais que investigar o desempenho preditivo para três tipos de séries, foi investigado o desempenho para tais séries segundo 7 cenários distintos, totalizando 21 séries simuladas. Os cenários, nos quais a série de referência X_t foi simulada, são:

1. Primeiro cenário: AR(1) com coeficiente autorregressivo $\alpha_1 = 0.5$, sendo a expressão de X_t dada por $(1 - 0.5B)X_t = \varepsilon_t$;

2. Segundo cenário: AR(2) com coeficientes autorregressivos $\alpha_1 = 0.4$ e $\alpha_2 = 0.2$, sendo a expressão de X_t dada por $(1 - 0.4B - 0.2B^2)X_t = \varepsilon_t$;

3. Terceiro cenário: MA(1) com coeficiente de médias móveis $\beta_1 = 0.39$, sendo a expressão de X_t dada por $X_t = (1 + 0.39B)\varepsilon_t$;

4. Quarto cenário: MA(2) com coeficientes de médias móveis $\beta_1 = 0.39$ e $\beta_2 = 0.6$, sendo a expressão de X_t dada por $X_t = (1 + 0.39B + 0.6B^2)\varepsilon_t$;

5. Quinto cenário: ARMA(1,1) com coeficientes $\alpha_1 = 0.5$ e $\beta_1 = 0.69$, sendo a expressão de X_t dada por $(1 - 0.5B)X_t = (1 + 0.69B)\varepsilon_t$;

6. Sexto cenário: ARIMA(1,1,1) com coeficientes $\alpha_1 = 0.7$ e $\beta_1 = 0.39$, sendo a expressão de X_t dada por $(1 - 0.7B)(1 - B)X_t = (1 + 0.39B)\varepsilon_t$;

7. Sétimo cenário: SARIMA(1,1,1)(1,1,1) com coeficientes $\alpha_1 = 0.5$, $\beta_1 = 0.6$, $\pi_1 = 0.8$, $\theta_1 = 0.3$, sendo a expressão de X_t dada por $(1 - 0.5B)(1 - 0.8B^{12})(1 - B)(1 - B^{12})X_t = (1 + 0.6B)(1 + 0.3B^{12})\varepsilon_t$.

É válido observar que, ε_i é um ruído branco gaussiano, ou seja, um processo estocástico de variáveis aleatórias não correlacionadas com média zero e variância constante, tal que $\varepsilon_i \sim i.i.d.N(0, \sigma^2)$. As séries simuladas possuem comprimento igual a 128 observações, nas quais 104 observações foram destinadas ao ajuste dos modelos e 24 observações foram reservadas para avaliar a acurácia das previsões, dado que o horizonte de previsão adotado foi 1, 2, 3, . . ., 24 passos à frente. O comprimento estabelecido, $128 = 2^7$, tem o propósito de favorecer a filtragem de informações por decimação diádica realizada pela TWD (ver subseção 2.1.5.4), abordada pelo modelo híbrido 1. O comportamento das séries X_t^1, X_t^2 e X_t^3, com a série de referência X_t simulada a partir do segundo cenário, é exibido a seguir.

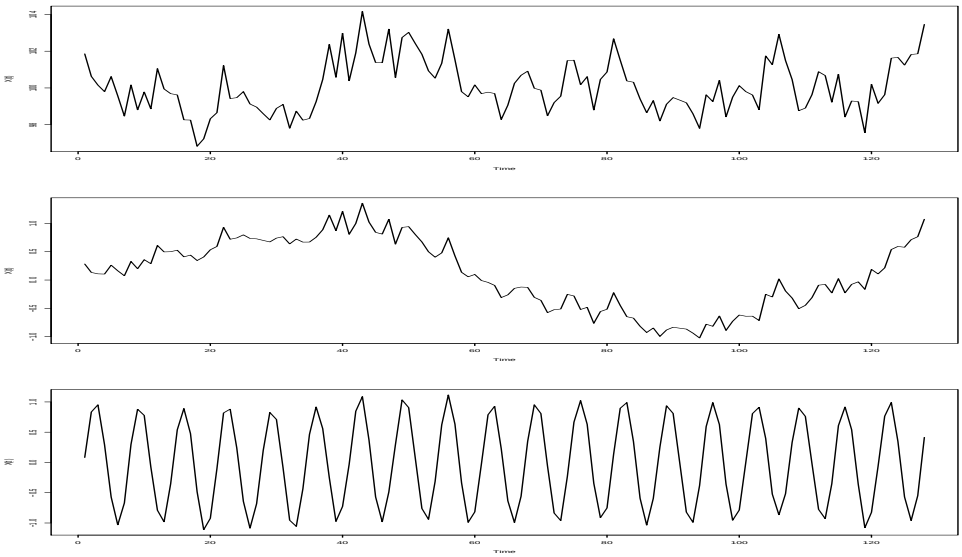

Figura 3 – Comportamento das séries simuladas X_t^1, X_t^2 e X_t^3, respectivamente.

É considerável ressaltar que, após simular a série de referência X_t segundo o sétimo cenário, ou seja, segundo o processo SARIMA(1,1,1)(1,1,1), e em seguida, substituí-la na expressão (2.31), a série resultante apresenta variância não constante, como mostra a Figura 4. Tal comportamento inviabiliza o ajuste dos modelos clássicos SARIMA. Para estabilizar a variância, fez-se necessário o uso da transformação Box Cox, em particular, a logarítmica.

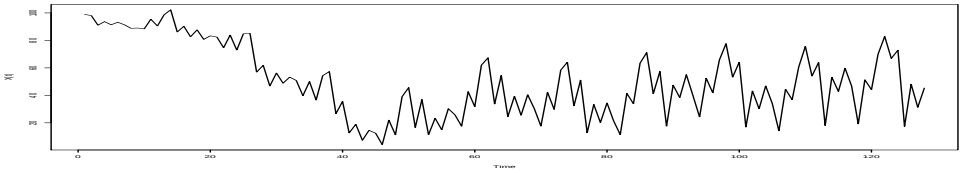

Figura 4 – Série X_t^1 com X_t simulada a partir do sétimo cenário.

A simulação descrita nesta subseção e toda implementação dos modelos, foram realizadas com o auxílio do *software* R (R Core Team, 2016), conforme a semente especificada *set.seed(1234)*. Além disso, as expressões das séries simuladas X_t^1, X_t^2 e X_t^3 com base na série de referência X_t, são semelhantes ao estudo de simulação apresentado em (MAHESWARAN; KHOSA, 2012).

2.2.3 Sistematização da metodologia

O desempenho preditivo dos modelos híbridos foi avaliado por intermédio de quatro vertentes, dadas por: número de níveis de resolução, *wavelet* mãe, tipos de séries e cenários de simulação. Quanto aos níveis de resolução, as possibilidades são 4, 5, ou 6 níveis pertinentes a transformada *wavelet*, conforme descritos na subseção 2.2.1. Em relação a *wavelet* mãe, tem-se 25 *wavelets* distintas, apresentadas também na subseção 2.2.1. Os três tipos de séries referem-se ao comportamento particular apresentado por X_t^1, X_t^2 e X_t^3, relatado na subseção 2.2.2. Por fim, os sete cenários de simulação da série de referência X_t, também são descritos na subseção 2.2.2. Para os modelos clássicos SARIMA, as vertentes são apenas: os tipos de séries e os cenários de simulação.

No entanto, para uma comparação adequada, é necessário que exista homogeneidade. Nesse sentido, é essencial que a comparação do desempenho preditivo entre os cinco modelos ocorra considerando: a mesma quantidade de níveis de resolução, a mesma *wavelet* mãe, o mesmo tipo de série e o mesmo cenário de simulação. A Figura 5 descreve a sequência do raciocínio utilizado para a comparação dos modelos em estudo.

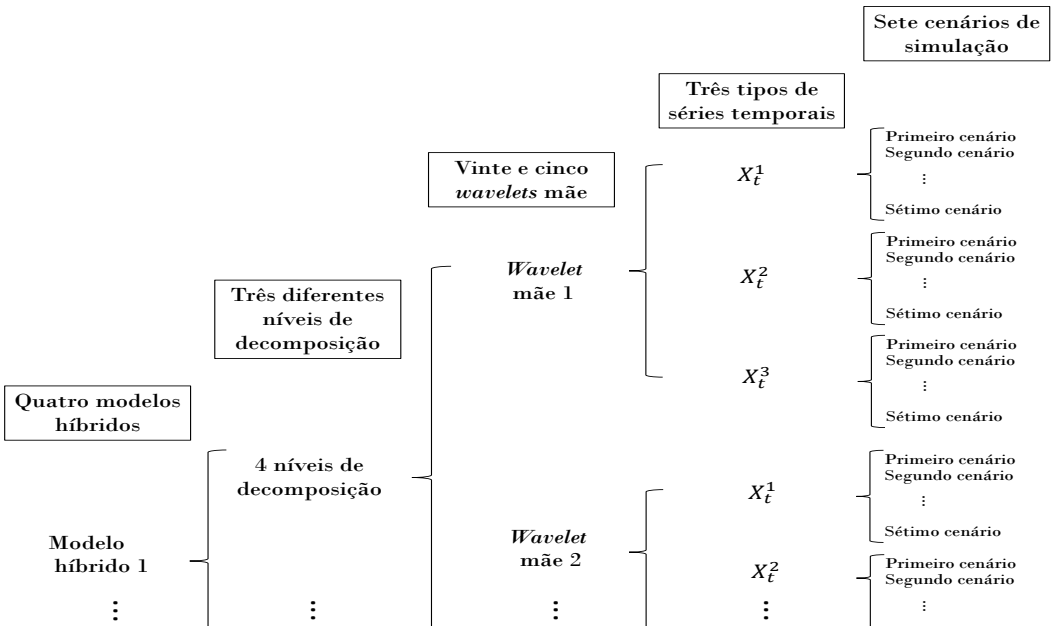

Figura 5 – Sistematização da metodologia utilizada para a comparação do desempenho preditivo entre os modelos.

Para exemplificar, a primeira comparação do desempenho preditivo entre os modelos híbridos 1, 2, 3, 4 e clássico SARIMA foi conduzida mediante a transformada *wavelet* com 4 de níveis de resolução ou decomposição a partir da *wavelet* mãe D2, sendo cada modelo aplicado na série X_t^1, com X_t simulada conforme o primeiro cenário. Considerando todas as vertentes de avaliação, os cinco modelos foram confrontados 1.575 vezes.

2.2.4 Avaliação dos modelos

As métricas adotadas para mensurar a acurácia de previsão dos modelos confrontados, foram *Raiz do Erro Quadrático Médio* (RMSE) e *Erro Absoluto Médio* (MAE), frequentemente utilizadas na análise de séries temporais. A definição do MSE consiste basicamente na diferença entre o verdadeiro valor x_{t+h} e o valor previsto $\hat{x}_t(h)$ segundo o modelo ajustado, cuja expressão é dada por,

$$MSE(\hat{x}_t(h)) = \frac{\sum(x_{t+h} - \hat{x}_t(h))^2}{n}. \tag{2.34}$$

Analogamente, o MAE mensura a média absoluta das “distâncias” entre os valores previstos e observados, cuja expressão é dada por,

$$MAE(\hat{x}_t(h)) = \frac{\sum|(x_{t+h} - \hat{x}_t(h))|}{n}. \tag{2.35}$$

Em vista disso, o valor mínimo do RMSE foi considerado indicador do melhor desempenho preditivo. Portanto, ao mencionar que determinado modelo é o melhor modelo

de previsão, significa que o mesmo apresentou o menor RMSE associado a previsão, dentre os cinco modelos confrontados. A técnica para avaliar o desempenho preditivo dos modelos, consiste em realizar previsões dentro do período amostral e compará-las aos valores observados. Para tanto, foram simuladas séries com 128 observações, nas quais 104 observações foram destinadas ao ajuste dos modelos e 24 observações foram reservadas para avaliar a acurácia das previsões, dado que o horizonte de previsão adotado foi 1, 2, 3, . . ., 24 passos à frente. O MAE foi utilizado para comparar o desempenho de ajuste do melhor modelo híbrido selecionado com o modelo clássico SARIMA.

É natural acreditar que, conforme o horizonte de previsão h aumenta, menor será a acurácia dos valores previstos. Porém, tal crença não é uma verdade incontestável, como mostra (YOUSEFI; WEINREICH; REINARZ, 2005), verificando que a eficácia do procedimento preditivo baseado na análise *wavelet* não é desgastada pelo aumento do horizonte de previsão.

2.2.4.1 Avaliação dos modelos segundo os níveis de resolução

O propósito é investigar uma possível interferência do número de níveis de resolução da transformada *wavelet* no desempenho preditivo dos modelos híbridos. Visto que, a TWD, presente no modelo híbrido 1, reduz pela metade o número de coeficientes *wavelets* a cada nível de resolução e observando que as séries simuladas possuem $128 = 2^7$ observações, obtêm-se o número máximo de 6 níveis de resolução. Por essa razão, a quantidade de níveis de resolução da transformada *wavelet* relativa aos quatro modelos híbridos, variou entre 4, 5 e 6 níveis.

Nessa perspectiva, para a avaliação dos modelos, fixou-se a quantidade de níveis de resolução e calculou-se a porcentagem de vezes que cada modelo confrontado foi eleito o melhor modelo de previsão para cada tipo de série X_t^1, X_t^2 e X_t^3, com relação as 25 *wavelets* mãe.

2.2.4.2 Avaliação dos modelos segundo as *wavelets* mãe

Sendo possível escolher a *wavelet* que melhor se adapta aos dados, o objetivo é investigar a influência da escolha da *wavelet* mãe, utilizada na transformada *wavelet*, no desempenho preditivo dos modelos híbridos. Neste trabalho, foram adotadas 6 famílias *wavelets*, dentre elas: Daubechies, *Coiflets*, *Symmlets*, *Best Localized*, *Minimum Bandwidth* e Fejer-Korovkin. Tais *wavelets* são representadas em termos do número de coeficientes não nulos L, dado por $L = 2P$, em que P é o número de momentos nulos. Os diferentes números de coeficientes L abordados em cada família, constam na subseção 2.2.1.

A eleição do modelo com melhor desempenho preditivo para cada *wavelet* mãe, baseia-se no número de vezes que o modelo apresentou o menor RSME de previsão com

relação as 21 séries simuladas, dadas pelos 3 tipos de séries X_t^1, X_t^2 e X_t^3 simuladas a partir de 7 cenários distintos.

2.2.4.3 Porcentagem média de redução do erro

O intuito é comparar o desempenho preditivo do melhor modelo híbrido de previsão, a ser selecionado neste estudo, em relação ao modelo clássico SARIMA. Dito em outras palavras, quer-se responder a seguinte questão: Em quanto o erro de previsão do modelo clássico SARIMA foi reduzido ao substituí-lo pelo modelo híbrido de previsão selecionado? O leitor pode estranhar tal comparação, pois metodologias diferentes estão sendo confrontadas, uma clássica e outra híbrida. Ainda mais, sabe-se que, em geral, a modelagem híbrida é superior a modelagem clássica. Entretanto, tal comparação baseia-se na ampla utilização dos modelos SARIMA para a previsão de séries temporais em diversas áreas, principalmente, em epidemiologia. A proporção de erro do modelo híbrido de previsão selecionado em relação ao modelo clássico SARIMA para cada série simulada i é dada por,

$$Prop_{[i]} = \left(\frac{\textit{RMSE do modelo híbrido selecionado}_{[i]}}{\textit{RMSE do modelo clássico SARIMA}_{[i]}} \right), \tag{2.36}$$

em que $i = 1, \ldots, 21$. O complementar de $Prop_{[i]}$ resultará na proporção de redução do erro, como segue,

$$\text{Redução}_{[i]} = 1 - Prop_{[i]}, \tag{2.37}$$

em que $i = 1, \ldots, 21$. Portanto, ao calcular a média da $Redução_{[i]}$, tem-se a *porcentagem média de redução do erro de previsão* proporcionada pelo modelo híbrido selecionado em relação ao erro produzido pelo modelo clássico SARIMA. A expressão da porcentagem média de redução é dada por,

$$\text{Porcentagem média} = \left(\frac{\sum_{i=1}^{21} \text{Redução}_{[i]}}{21} \right) 100. \tag{2.38}$$

O mesmo cálculo foi realizado em termos do RMSE de ajuste e em termos do MAE para ajuste e previsão.

2.3 Resultados e discussões

Nesta seção, são apresentados os resultados da comparação entre os cinco modelos. Na subseção 2.3.1, serão feitas avaliações para a identificação do melhor modelo de previsão de acordo com a quantidade de níveis de resolução e *wavelet* mãe. Na subseção 2.3.2, é investigado a melhor *wavelet* mãe para o melhor modelo de previsão identificado na subseção 2.3.1. Na subseção 2.3.3, o desempenho do melhor modelo, considerando

as melhores *wavelets*, é comparado ao modelo clássico SARIMA. Na subseção 2.3.4, os resultados são ilustrados para uma das séries simuladas.

2.3.1 Desempenho de previsão à longo prazo dos modelos segundo os níveis de resolução e *wavelets* mãe

O desempenho preditivo à longo prazo dos modelos híbridos foi avaliado com respeito ao número de níveis de resolução pertinentes a transformada *wavelet*. Quanto ao modelo clássico SARIMA, seu desempenho preditivo é imune a alteração dos níveis de resolução, dado que sua metodologia compreende unicamente o domínio do tempo. Para prosseguir com a exposição dos resultados, é preciso recordar a sequência das vertentes de avaliação sistematizadas na subseção 2.2.3. Para cada possibilidade de níveis de resolução da transformada *wavelet*, foram contempladas 25 *wavelets* mãe. Por sua vez, para cada *wavelet* mãe, o desempenho preditivo dos modelos foram examinados em 21 séries simuladas.

De modo geral, dentre os cinco modelos confrontados, dois sobressaíram-se por seu bom desempenho preditivo, estes são: modelo híbrido 2 e modelo híbrido 4. Tal resultado é contemplado pelas Tabelas 1, 2 e 3, as quais exibem o melhor modelo de previsão para cada tipo de série simulada X_t^1, X_t^2 e X_t^3, segundo as 25 *wavelets* mãe empregadas na transformada *wavelet* com 4 níveis, 5 níveis e 6 níveis de resolução, respectivamente.

Tabela 1 – Modelos com menor RMSE de previsão à longo prazo (h=24) para a decomposição *wavelet* com 4 níveis perante todas as *wavelets* mãe.

Wavelet	Melhor modelo			
	X_t^1 (7 cenários)	X_t^2 (7 cenários)	X_t^3 (7 cenários)	Geral (21 séries)
D2	Hibr 2	Hibr 2	Hibr 1	Hibr 2
D4	Hibr 2	Hibr 2	Hibr 1	Hibr 2
D6	Hibr 4 e Hibr 2	Hibr 4	Hibr 4	Hibr 4
D8	Hibr 2	Hibr 2	Hibr 4	Hibr 4 e Hibr 2
D12	Hibr 2	Hibr 2	Hibr 4	Hibr 2
D16	Hibr 2	Hibr 4	Hibr 4	Hibr 4
D18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
D20	Hibr 4	Hibr 4	Hibr 4	Hibr 4
LA8	Hibr 4	Hibr 4	Hibr 4	Hibr 4
LA12	Hibr 2	Hibr 2	Hibr 4	Hibr 4
LA18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
LA20	Hibr 4 e Hibr 2	Hibr 2	Hibr 4	Hibr 4
BL18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
BL20	Hibr 4	Hibr 4	Hibr 4	Hibr 4
C6	Hibr 4	Hibr 2	Hibr 4	Hibr 4
C12	Hibr 2	Hibr 2	Hibr 4	Hibr 4
C18	Hibr 2	Hibr 2	Hibr 4	Hibr 4
C24	Hibr 4	Hibr 2	Hibr 4	Hibr 4
MB4	Hibr 2	Hibr 2	Hibr 1	Hibr 2
MB8	Hibr 4	Hibr 4	Hibr 4	Hibr 4
MB24	Hibr 2	Hibr 2	Hibr 4	Hibr 2
FK4	Hibr 2	Hibr 2	Hibr 1	Hibr 2
FK6	Hibr 2	Hibr 2	Hibr 4	Hibr 4
FK8	Hibr 2	Hibr 4	Hibr 4	Hibr 4
FK22	Hibr 2	Hibr 2	Hibr 4	Hibr 4

Tabela 2 – Modelos com menor RMSE de previsão à longo prazo (h=24) para a decomposição *wavelet* com 5 níveis perante todas as *wavelets* mãe.

Wavelet	Melhor modelo			
	X_t^1 (7 cenários)	X_t^2 (7 cenários)	X_t^3 (7 cenários)	Geral (21 séries)
D2	Hibr 2	Hibr 2	Hibr 1	Hibr 2
D4	Hibr 4	Hibr 2	Hibr 4	Hibr 4
D6	Hibr 2	Hibr 2	Hibr 4	Hibr 2
D8	Hibr 2	Hibr 2	Hibr 4	Hibr 4
D12	Hibr 2	Hibr 2	Hibr 4	Hibr 2
D16	Hibr 4	Hibr 4	Hibr 4	Hibr 4
D18	Hibr 2	Hibr 2	Hibr 4	Hibr 4
D20	Hibr 4	Hibr 2	Hibr 4	Hibr 4
LA8	Hibr 2	Hibr 4	Hibr 4	Hibr 4
LA12	Hibr 4	Hibr 4	Hibr 4	Hibr 4
LA18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
LA20	Hibr 4 e Hibr 2	Hibr 4	Hibr 4	Hibr 4
BL18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
BL20	Hibr 2	Hibr 4	Hibr 4	Hibr 4
C6	Hibr 4	Hibr 2	Hibr 4	Hibr 4
C12	Hibr 4	Hibr 2	Hibr 4	Hibr 4
C18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
C24	Hibr 4	Hibr 4	Hibr 4	Hibr 4
MB4	Hibr 4	Hibr 2	Hibr 4	Hibr 4
MB8	Hibr 2	Hibr 2	Hibr 4	Hibr 4
MB24	Hibr 2	Hibr 2	Hibr 4	Hibr 4
FK4	Hibr 2	Hibr 4	Hibr 4	Hibr 4
FK6	Hibr 2	Hibr 2	Hibr 4	Hibr 4
FK8	Hibr 2	Hibr 2	Hibr 4	Hibr 4
FK22	Hibr 2	Hibr 4	Hibr 4	Hibr 4

Tabela 3 – Modelos com menor RMSE de previsão à longo prazo (h=24) para a decomposição *wavelet* com 6 níveis perante todas as *wavelets* mãe.

Wavelet	Melhor modelo			
	X_t^1 (7 cenários)	X_t^2 (7 cenários)	X_t^3 (7 cenários)	Geral (21 séries)
D2	Hibr 2	Hibr 2	Hibr 1	Hibr 2
D4	Hibr 2	Hibr 4	Hibr 4	Hibr 4
D6	Hibr 4	Hibr 4	Hibr 4	Hibr 4
D8	Hibr 2	Hibr 2	Hibr 4	Hibr 4
D12	Hibr 2	Hibr 2	Hibr 4	Hibr 4 e Hibr 2
D16	Hibr 2	Hibr 4	Hibr 4	Hibr 4
D18	Hibr 2	Hibr 4	Hibr 4	Hibr 4
D20	Hibr 2	Hibr 4	Hibr 4	Hibr 4
LA8	Hibr 2	Hibr 4	Hibr 4	Hibr 4
LA12	Hibr 4	Hibr 2	Hibr 4	Hibr 4
LA18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
LA20	Hibr 4 e Hibr 2	Hibr 2	Hibr 4	Hibr 4
BL18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
BL20	Hibr 2	Hibr 2	Hibr 4	Hibr 2
C6	Hibr 2	Hibr 2	Hibr 4	Hibr 4
C12	Hibr 2	Hibr 2	Hibr 4	Hibr 4
C18	Hibr 4	Hibr 4	Hibr 4	Hibr 4
C24	Hibr 4	Hibr 4	Hibr 4	Hibr 4
MB4	Hibr 2	Hibr 4	Hibr 4	Hibr 4
MB8	Hibr 4 e Hibr 2	Hibr 2	Hibr 4	Hibr 4
MB24	Hibr 2	Hibr 4	Hibr 4	Hibr 4
FK4	Hibr 2	Hibr 4	Hibr 4	Hibr 4
FK6	Hibr 2	Hibr 4	Hibr 4	Hibr 4
FK8	Hibr 2	Hibr 2	Hibr 4	Hibr 4
FK22	Hibr 2	Hibr 4	Hibr 4	Hibr 4

O modelo híbrido 1 aparece poucas vezes nas Tabelas 1, 2 e 3, mostrando que este apresentou o menor RMSE de previsão num pequeno número de vezes. Portanto, o desempenho preditivo do modelo híbrido 1 foi ineficaz, independente do limiar escolhido (ver subseção 2.2.1). Mesmo assim, o limiar adaptativo *SureShrink* destacou-se perante os limiares universal e alternativo.

A notação "Hibr 4 e Hibr 2" presente nas tabelas, diz respeito a equivalência do desempenho preditivo apresentado pelos modelos híbridos 4 e 2. Além disso, a coluna "Geral" retrata o melhor modelo de previsão para cada *wavelet* mãe, considerando os três tipos de séries simuladas conjuntamente. Melhor dizendo, a coluna "Geral" leva em consideração a porcentagem do número de vezes em que determinado modelo apresentou-se como melhor modelo de previsão dentre as 21 séries simuladas. Ao observá-la, é possível constatar o predomínio do modelo híbrido 4 em relação aos demais modelos confrontados, independente da quantidade de níveis de resolução. Para reforçar tal superioridade, a Tabela 4 vem sumarizar os resultados expressos nas Tabelas 1, 2 e 3, mostrando a porcentagem de vezes que cada modelo apresentou o melhor desempenho preditivo com relação as 25 *wavelets* mãe.

Tabela 4 – Porcentagem do número de vezes que cada modelo apresentou o menor RMSE resultante da previsão de longo prazo (h=24) relativa as *wavelets* mãe.

		Modelos					
		clássico SARIMA	Híbrido 1	Híbrido 2	Híbrido 3	Híbrido 4	Total (25 *wavelets*)
4 níveis	X_t^1	0	0	56%	0	44%	100%
	X_t^2	0	0	60%	0	40%	100%
	X_t^3	0	16%	0	0	84%	100%
	Geral	0	0	24%	0	76%	100%
5 níveis	X_t^1	0	0	52%	0	48%	100%
	X_t^2	0	0	56%	0	44%	100%
	X_t^3	0	4%	0	0	96%	100%
	Geral	0	0	12%	0	88%	100%
6 níveis	X_t^1	0	0	68%	0	32%	100%
	X_t^2	0	0	40%	0	60%	100%
	X_t^3	0	4%	0	0	96%	100%
	Geral	0	0	8%	0	92%	100%

Conforme os níveis de resolução são acrescidos, o êxito correlato ao desempenho preditivo do modelo híbrido 4, abrange uma porcentagem maior de *wavelets* mãe. Segundo a Tabela 4, para a transformada *wavelet* com 4 níveis, 5 níveis e 6 níveis de resolução, o modelo híbrido 4 apresentou-se como melhor modelo de previsão em 76%, 88% e 92% das 25 *wavelets* mãe, respectivamente. Nessa lógica, aplicar a transformada *wavelet* com

6 níveis de resolução, poderá potencializar o desempenho preditivo do modelo híbrido 4. Quanto ao modelo híbrido 2, seu desempenho preditivo enfraquece à medida que os níveis de resolução são acrescidos, mantendo-se sempre inferior ao modelo híbrido 4.

No que concerne ao modelo híbrido 1, seu desempenho preditivo foi superior aos demais modelos em uma porcentagem pequena de vezes, especialmente, para a série X_t^3. Os modelos híbrido 3 e clássico SARIMA, apresentou desempenho preditivo inferior aos demais modelos em todas as circunstâncias de avaliação, apresentadas na subseção 2.2.3, justificando a ausência de porcentagens relativa aos mesmos na Tabela 4.

2.3.2 Desempenho de previsão à longo prazo do modelo híbrido 4 segundo a *wavelet* mãe

De acordo com os resultados apresentados na subseção 2.3.1, o modelo híbrido 4, proposto nesse estudo, apresentou o melhor desempenho preditivo (ver subseção 2.2.4), independente do número de níveis de resolução da transformada *wavelet*. Também, tem-se evidências de que a transformada *wavelet* com 6 níveis de resolução, potencializa o desempenho preditivo do modelo híbrido 4. Mediante tais considerações, o próximo objetivo é identificar as *wavelets* mãe que otimizam a performance do modelo híbrido 4. Para isso, fez-se necessário uma análise minuciosa sobre o desempenho preditivo do modelo híbrido 4 mediante as diferentes *wavelets* mãe, apresentadas na subseção 2.2.1. Os resultados dessa análise são descritos a seguir.

Obteve-se evidências de que a "natureza" da *wavelet* mãe, quer dizer, o número de coeficientes não nulos L e a família a qual pertence, interfere no desempenho preditivo do modelo híbrido 4. Observou-se que, dependendo da *wavelet* mãe adotada na transformada *wavelet*, o modelo híbrido 4 foi eficaz para os três tipos de séries simuladas, sendo eleito o melhor modelo de previsão em virtude de uma considerável vantagem em relação ao modelo híbrido 2. Em outras vezes, o modelo híbrido 4 mostrou-se eficaz apenas para um tipo de série simulada, sendo eleito o melhor modelo de previsão para determinada *wavelet* mãe, devido uma pequena vantagem em relação ao modelo híbrido 2.

Para exemplificar o argumento da parágrafo anterior, basta considerar as *wavelets* LA18 e C12 expressas na Tabela 3. Em ambas, o modelo híbrido 4 foi eleito o melhor modelo de previsão. Contudo, para a *wavelet* LA18, tal modelo apresentou o melhor desempenho de previsão para 90.5% das 21 séries simuladas *contra* 9.5% pertinente ao modelo híbrido 2. Nessa circunstância, o desempenho de previsão do modelo híbrido 4 foi notavelmente superior ao modelo híbrido 2. Por outro lado, para a *wavelet* C12, o modelo híbrido 4 apresentou o melhor desempenho preditivo para 52.4% das 21 séries simuladas contra 47. 6% referente ao modelo híbrido 2. Nesse caso, o desempenho de previsão do modelo híbrido 4 foi razoavelmente superior ao desempenho do modelo híbrido 2, sendo

eficaz apenas para a série simulada X_t^3.

Com o propósito de exibir as evidências que sugerem a relação entre "natureza" da *wavelet* mãe e desempenho preditivo do modelo híbrido 4, realizou-se o *ranking* das *wavelets* mãe para as três possibilidades de níveis de resolução. A classificação das *wavelets* mãe, baseou-se na porcentagem de vezes que o modelo híbrido 4 apresentou o menor RMSE de previsão com relação as 21 séries simuladas (ver subseção 2.2.4.2).

Tabela 5 – *Ranking* das *wavelets* mãe segundo a porcentagem de vezes que o modelo híbrido 4 apresentou-se como melhor modelo de previsão relativo as 21 séries simuladas.

Colocação	Modelo híbrido 4					
	4 Níveis		5 Níveis		6 Níveis	
	Wavelet Mãe	(%)	Wavelet Mãe	(%)	Wavelet Mãe	(%)
1º lugar	D20	90.5	BL18	90.5	LA18	90.5
2º lugar	D18	80.9	LA12	90.5	BL18	85.7
3º lugar	BL18	80.9	LA18	85.7	C24	80.9
4º lugar	BL20	76.2	D16	80.9	C18	76.2
5º lugar	LA18	76.2	C24	80.9	D16	76.2
6º lugar	C24	71.4	LA20	66.7	D20	66.7

Ao examinar a Tabela 5, é notável a preminência de *wavelets* mãe com elevado número de coeficientes não nulos $L = 2P$ ou elevado número de momentos nulos P, para as três variações de níveis de resolução. Essa característica evidencia a possível implicação: se o modelo híbrido 4 possui o melhor desempenho preditivo para determinada *wavelet* mãe, então tal *wavelet* possui um número elevado de momentos nulos P. Todavia, a recíproca não é válida! Como contra exemplo, considere a *wavelet* MB24 expressa na Tabela 1. Apesar de possuir elevado número de coeficientes não nulos $L = 24$ e, consequentemente, elevado número de momentos nulos $P = 12$, o modelo com melhor desempenho preditivo para a *wavelet* MB24 é o modelo híbrido 2.

Com a finalidade de expressar o desempenho de previsão do modelo híbrido 4 em relação as famílias *wavelets*, foi realizado o *ranking* entre as 6 melhores *wavelets* mãe respectivas as 6 famílias *wavelets* adotadas neste estudo, como mostra a Tabela 6. A melhor *wavelet* mãe de cada família foi eleita segundo a porcentagem de vezes em que o modelo híbrido 4 apresentou o melhor desempenho preditivo para as 21 séries simuladas. Dessa forma, as melhores *wavelets* mãe pertinentes as famílias *Best Localized*, Coiflet, Daubechies, *Least Asymetric*, Fejer Korovkin e *Minimum Bandwidth* são BL18, C24, D20, LA18, FK22 e MB24, respectivamente. É válido destacar que, todas as *wavelets* eleitas possuem elevado número de coeficientes não nulos L.

Os resultados da Tabela 6, comprovam o enfraquecimento proporcionado pelas famílias Fejer Korovkin e *Minimum Bandwidth* ao desempenho do modelo híbrido 4, as quais ocupam as últimas colocações do *ranking*, independente da quantidade de níveis de resolução. Por outro lado, as famílias *Best Localized* (BL), Coiflet (C), Daubechies (D) e *Least Asymetric* (LA) potencializam o desempenho preditivo do modelo híbrido 4.

Tabela 6 – *Ranking* das melhores *wavelets* mãe respectivas a cada família *wavelet*, segundo os níveis de resolução.

Colocação	Modelo híbrido 4					
	4 Níveis		5 Níveis		6 Níveis	
	Wavelet Mãe	(%)	Wavelet Mãe	(%)	Wavelet Mãe	(%)
1º lugar	D20	90.5	BL18	90.5	LA18	90.5
2º lugar	BL18	80.9	LA12	90.5	BL18	85.7
3º lugar	LA18	76.2	D16	80.9	C24	80.9
4º lugar	C24	71.4	C24	80.9	D16	76.2
5º lugar	MB8	71.4	FK22	66.7	FK22	66.7
6º lugar	FK22	52.4	MB24	52.4	MB24	61.9

Em síntese, três *wavelets* mãe destacam-se por otimizar o desempenho preditivo do modelo híbrido 4, estas são: Daubechies 20 (D20), *Best Localized* 18 (BL18) e *Least Asymetric* 18 (LA18). Tais *wavelets* ocupam as primeiras colocações, para as três possibilidades de níveis de resolução, tanto na Tabela 5 como na Tabela 6. Em particular, realizando a transformada *wavelet* segundo essas *wavelets* mãe, o modelo híbrido 4 apresentou o melhor desempenho de previsão para 90.5% das 21 séries simuladas, como mostra a Tabela 5.

2.3.3 Desempenho de previsão à longo prazo do modelo híbrido 4 em relação ao modelo clássico SARIMA

Agora, o interesse consiste em responder a seguinte questão, proposta na subseção 2.2.4.3: Em quanto o erro de previsão do modelo clássico SARIMA foi reduzido ao substituí-lo pelo modelo híbrido 4? Essa indagação visa ressaltar o ganho preditivo do pesquisador ao adotar o modelo híbrido 4, que pode ser facilmente aplicado devido sua automaticidade, em vez do modelo clássico SARIMA, comumente utilizado para previsão de séries temporais. Lembre que, o erro de previsão refere-se ao valor do RMSE (ver subseção 2.2.4).

Para tal fim, calculou-se a porcentagem média de redução do erro de previsão proporcionada pelo modelo híbrido 4 em relação ao erro produzido pelo modelo clássico SARIMA. Também, calculou-se a porcentagem de redução em termos do ajuste. Além do

mais, conforme a subseção 2.3.2, as *wavelets* mãe D20, BL18 e LA18 otimizam o desempenho preditivo do modelo híbrido 4. Por esse motivo, tais *wavelets* foram empregadas na transformada *wavelet* presente no modelo híbrido 4, para o mesmo ser comparado ao modelo SARIMA. Também, três horizontes de previsão foram considerados para a comparação, estes são: 6 passos à frente (curto prazo), 12 passos à frente (médio prazo) e 24 passos à frente (longo prazo). Apesar deste capítulo dar ênfase a previsão de longo prazo, nesse momento, é válido considerar outros prazos, mostrando que a superioridade do desempenho preditivo relativo ao modelo híbrido 4 é válida para as três possibilidades de horizontes de previsão.

De acordo com a Tabela 7, mesmo para previsões de curto prazo, para as quais o modelo SARIMA é recomendado, o modelo híbrido 4 forneceu reduções médias no erro de previsão que variam entre 55.9% até 80.4%, como segue.

Tabela 7 – Porcentagem média de redução do RSME de ajuste e previsão do modelo híbrido 4 em relação ao modelo clássico SARIMA.

Níveis de resolução	*Wavelet*		D20		BL18		LA18	
			MAE	RMSE	MAE	RMSE	MAE	RMSE
4 níveis	Ajuste		-18.5	-32.8	18.5	7.4	45.8	38.8
	Previsão	curto prazo	82.2	80.4	72.2	72.7	53.4	55.9
		médio prazo	87.2	85.6	74.8	73.8	58.3	60.9
		longo prazo	82.4	81.5	71.1	71.2	61.2	62.6
5 níveis	Ajuste		-1.2	-0.3	6.5	-7.7	37.3	30.3
	Previsão	curto prazo	70.2	71.1	72.1	72.7	69.1	70
		médio prazo	78.1	78.8	72.7	72.7	71.9	71.8
		longo prazo	76.8	76.7	69.1	69	67.3	68.4
6 níveis	Ajuste		-0.3	-0.5	27.6	12.5	41.6	35
	Previsão	curto prazo	71.1	71.7	72.9	72.9	68.9	69.7
		médio prazo	78.3	78.9	73.6	72.9	72.6	72.3
		longo prazo	76.8	76.7	69.1	68.9	70.1	70.7

As maiores reduções do erro de previsão obtidas pela substituição do modelo SARIMA pelo modelo híbrido 4, deram-se no caso da transformada *wavelet* com 4 níveis de resolução e *wavelet* mãe D20. Nessas circunstâncias, o valor do RMSE de previsão a curto prazo, em média, teve uma redução de 80.4%. Para a previsão de médio prazo, o valor do RMSE, em média, foi 85.6% menor. Por fim, o valor do RMSE de previsão a longo prazo relativo ao modelo híbrido 4, em média, foi 81.5% menor quando comparado ao RMSE de previsão produzido pelo modelo clássico SARIMA.

Ainda que o erro de previsão é reduzido ao optar pelo modelo híbrido 4 ao invés do modelo SARIMA, nem sempre o mesmo ocorre com o erro de ajuste. Considerando a transformada *wavelet* com 4 níveis de resolução e *wavelet* mãe D20, o valor do RMSE associado ao ajuste, apresentou o aumento de 32.8% ao substituir o modelo SARIMA pelo modelo híbrido 4. Esse resultado reforça o fato de que não necessariamente um modelo com um excelente ajuste implica em melhor qualidade preditiva e vice-versa. Ainda, tal resultado destaca que, mesmo com um ajuste ruim, o modelo híbrido 4 fornece boas previsões.

Entretanto, mesmo que a essência deste estudo esteja voltada ao erro de previsão, é ideal que o modelo apresente a capacidade de ajustar e prever o comportamento de uma série temporal, com os menores erros possíveis. Dessa forma, entre as três possibilidades de níveis de resolução, ao considerar a transformada *wavelet* com 6 níveis, obteve-se o mais sensato quadro de equilíbrio entre os valores do RMSE de ajuste e previsão. Nessa condição, as *wavelets* mãe BL18 e LA18 forneceram reduções ao erro de ajuste de 12.5% e 35%, respectivamente. Quanto a redução do erro de previsão fornecida por tais *wavelets*, esta variou entre 68.9% até 72.9%, de acordo com o prazo de previsão. Ou seja, ao realizar a transformada *wavelet* que compõe o modelo híbrido 4, com 6 níveis de resolução a partir da *wavelet* mãe BL18 ou LA18, além do ganho na previsão, tem-se o ganho no ajuste com relação ao modelo SARIMA.

Nessa perspectiva de equilíbrio entre os erros de ajuste e previsão, a *wavelet* mãe LA18 é sugerida como uma escolha atrativa, mesmo não apresentando as maiores reduções do erro de previsão. Independente do número de níveis de resolução, quando aplicada a *wavelet* mãe LA18 na transformada *wavelet* que compõe o modelo híbrido 4, este forneceu ganhos tanto em termos de ajuste como em termos de previsão, em relação ao modelo SARIMA. Agora, visando apenas a redução do erro de previsão, a melhor alternativa é a *wavelet* mãe D20, independente do número de níveis de resolução e do prazo de previsão.

2.3.4 Resultados dos modelos para uma das séries simuladas

Para ilustrar a abordagem dos modelos confrontados e os resultados obtidos, a série X_t^1, com a série de referência X_t simulada a partir do cenário AR(1) foi considerada (ver subseção 2.2.2). Também, a transformada *wavelet* foi realizada com 6 níveis de resolução, segundo a *wavelet* mãe LA18.

A Figura 6 ilustra a abordagem do modelo híbrido 1. Observe que, dentre os três tipos de limiares, o limiar *universal* descartou todos os coeficientes do nível W_1, assim como o limiar alternativo. Todavia, o limiar universal nem sempre apresentará tal comportamento. Por outro lado, o limiar adaptativo *SureShrink* conservou alguns coeficientes *wavelets* no nível W_1, considerados importantes para a reconstrução da série.

Quanto a previsão, a função *auto.arima* do *software* R (ver subseção 2.2.1), retornou o modelo ARIMA(5,0,0) com intercepto igual a 99.91 e erro padrão igual a $\hat{EP} = 0.34$, para as séries reconstruídas a partir dos limiares universal e alternativo, respectivamente. Em relação a série reconstruída a partir do limiar *SureShrink*, o modelo ajustado pela função *auto.arima* foi ARIMA(0,0,2) com intercepto igual a 99.97 ($\hat{EP} = 0.19$). Visto que, o intercepto $c \neq 0$ e $d = 0$ nos três modelos, é esperado que o comportamento da previsão destes modelos tenderá para a média dos dados, como representado na Figura 6. Apesar de fornecer um bom ajuste (primeiras 104 observações), a previsão de logo prazo (h=24) dos modelos ARIMA selecionados foi insatisfatória, com $RMSE = 1.28$ e $MAE = 1.01$ para os limiares globais e $RMSE = 1.27$ e $MAE = 0.98$ para o limiar adaptativo.

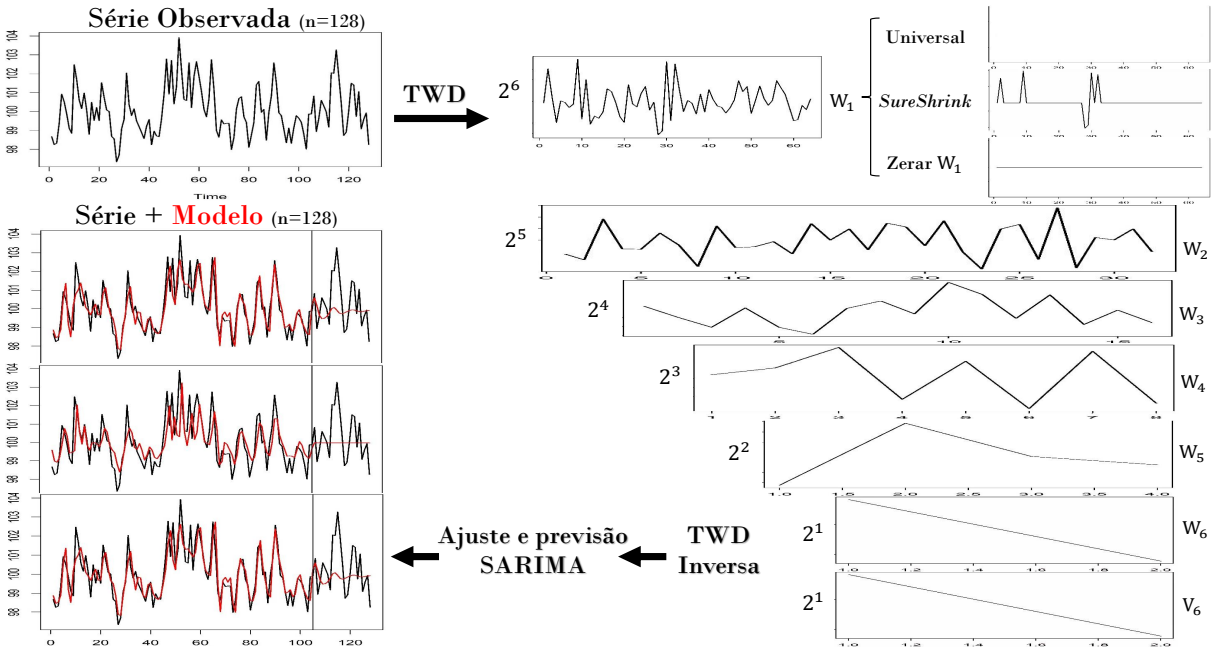

Figura 6 – Modelo híbrido 1 para a série X_t^1 com X_t simulada a partir do cenário AR(1).

A Figura 7 ilustra a abordagem do modelo híbrido 2. Os modelos ARIMA ajustados em cada um dos 6 níveis de resolução pela função *auto.arima*, foram: ARIMA(5,0,0), ARIMA(5,0,0), ARIMA(4,0,0), ARIMA(1,0,0), ARIMA(1,0,0), ARIMA(0,0,0) com intercepto igual a 0.04 e ARIMA(1,2,4), respectivamente. Em particular, para o nível W_6, o modelo ajustado ARIMA(0,0,0) não possui componentes autorregressivas e nem componentes de médias móveis, consequentemente, a previsão irá para a média dos coeficientes *wavelets* em cada nível. Tal resultado é esperado e deve-se a restrição do espaço de modelos decorrente do procedimento *stepwise* utilizado pela função *auto.arima* para encontrar o "melhor modelo". Mesmo com o espaço de modelos restrito, o desempenho preditivo do modelo híbrido 2 foi satisfatório, com $RMSE = 0.7$ e $MAE = 0.61$.

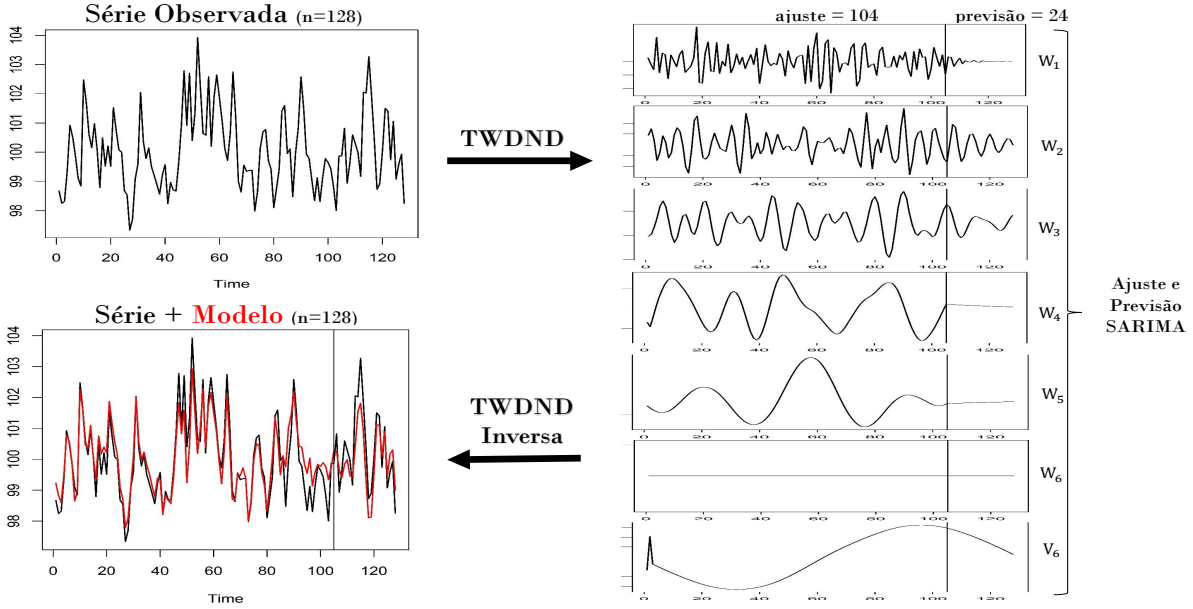

Figura 7 – Modelo híbrido 2 para a série X_t^1 com X_t simulada a partir do cenário AR(1).

A Figura 8 ilustra a abordagem do modelo híbrido 3. Nela são representadas a estimação e extrapolação do modelo harmônico nas subséries de coeficientes *wavelets* associadas às escalas de alta frequência e modelo *spline* na subsérie de coeficientes *wavelets* relacionada ao nível V_6. O desempenho preditivo do modelo híbrido 3 deixou a desejar, apresentando $RMSE = 1.12$ e $MAE = 0.88$.

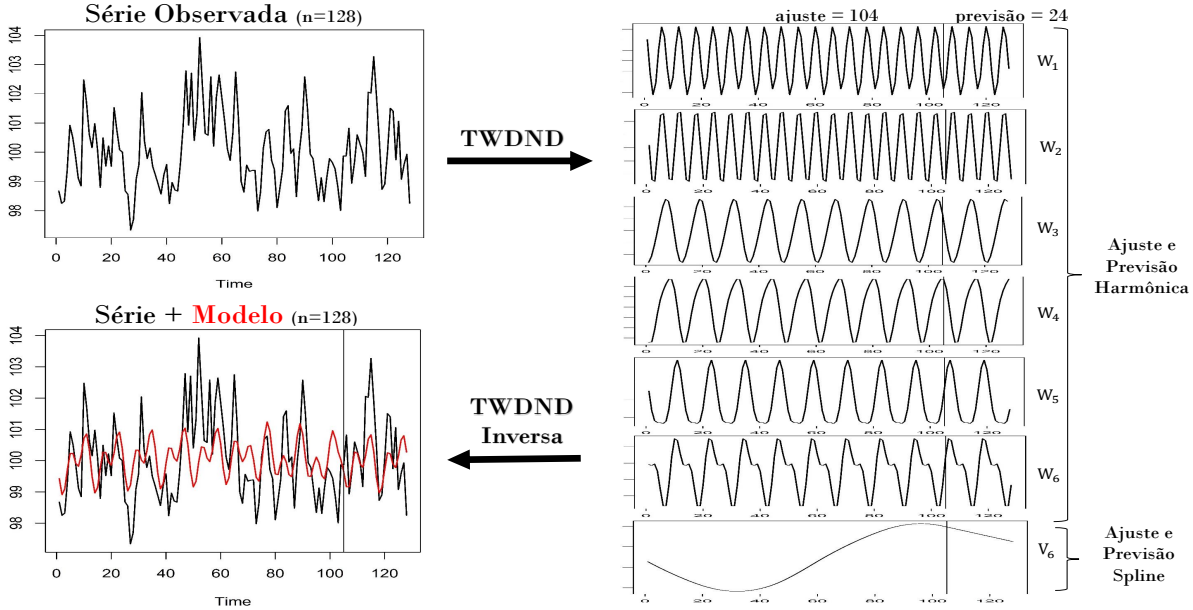

Figura 8 – Modelo híbrido 3 para a série X_t^1 com X_t simulada a partir do cenário AR(1).

O desempenho preditivo do modelo híbrido 4 foi superior ao demais modelos, com $RMSE = 0.50$ e $MAE = 0.42$ para a série representada na Figura 9.

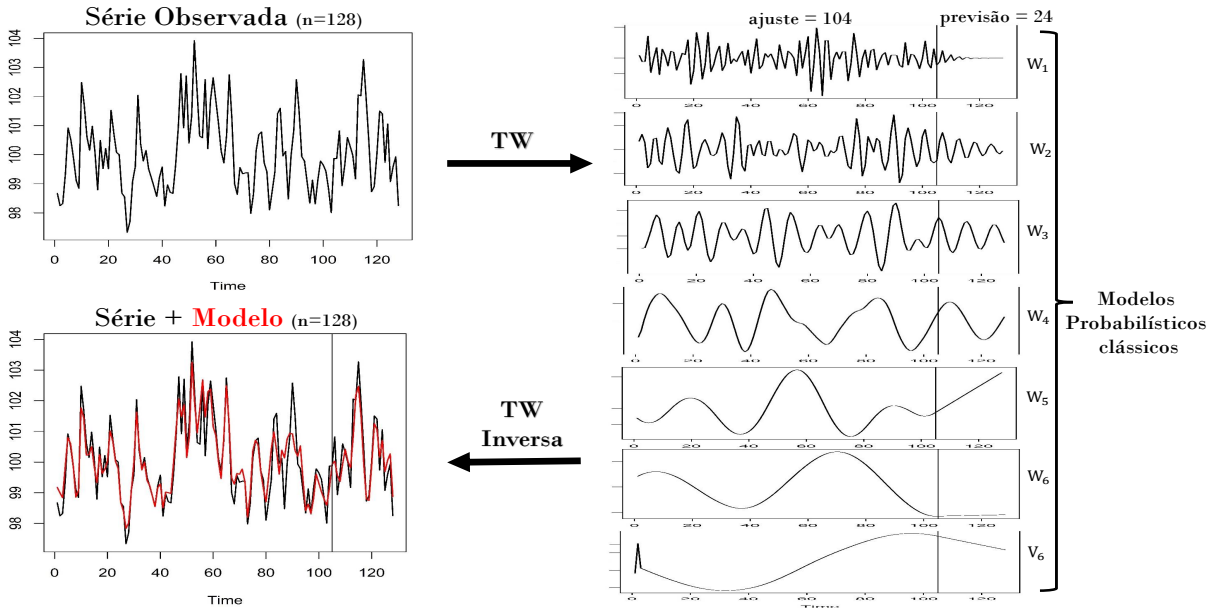

Figura 9 – Modelo híbrido 4 para a série X_t^1 com X_t simulada a partir do cenário AR(1).

A Figura 10 ilustra o ajuste e previsão de modelo clássico SARIMA diretamente na série X_t^1. A função *auto.arima* foi utilizada, retornando o modelo ARIMA(1,0,0) com intercepto igual a 99.97 ($\hat{ep} = 0.25$). Como o intercepto $c \neq 0$ e $d = 0$, espera-se que o comportamento da previsão tenderá para a média das observações. Portanto, o desempenho preditivo foi ineficaz, com $RMSE = 1.27$ e $MAE = 0.98$.

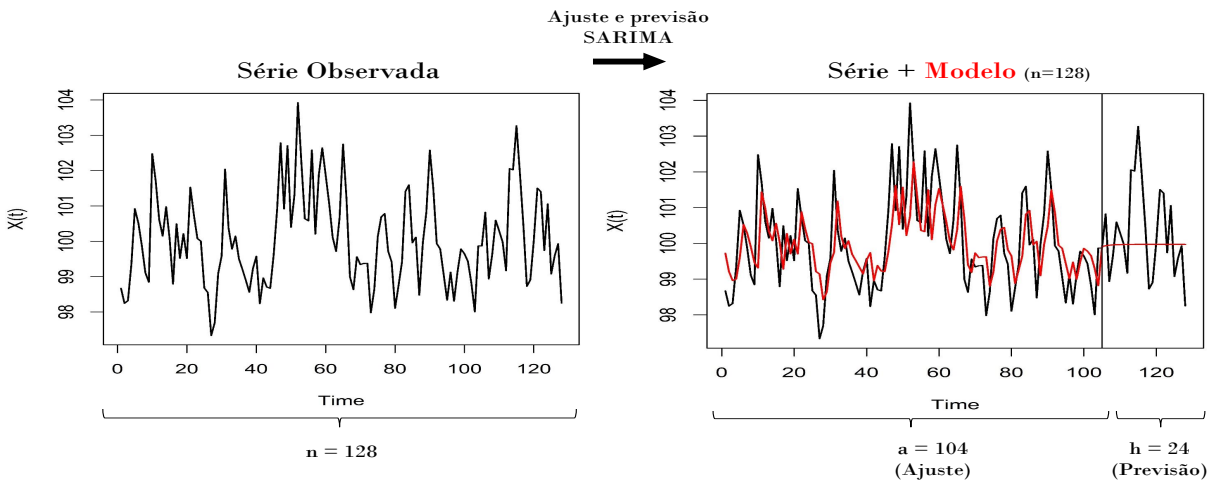

Figura 10 – Abordagem clássica do modelo SARIMA.

2.3.5 Desempenho computacional dos modelo confrontados

Neste estudo, o termo desempenho computacional diz respeito a mensuração, em segundos, do tempo gasto por cada modelo confrontado para fornecer o ajuste e previsão de uma única série temporal simulada. Nessa perspectiva, quanto menor for o tempo utilizado pelo modelo confrontado, melhor será seu desempenho computacional ou maior será sua eficiência computacional.

O modelo híbrido 4 apresentou o melhor desempenho computacional dentre os cinco modelos confrontados. Para cada série simulada, em média, teve a duração de 0.51 segundos, com desvio padrão ($\hat{\sigma}$) de 0.06 segundos. A execução do modelo híbrido 1 teve a duração média de 5.20 segundos por série temporal e $\hat{\sigma} = 1.39$. O modelo híbrido 2 apresentou duração média de 22.26 segundos, com $\hat{\sigma} = 1.40$. O modelo híbrido 3 teve duração média de 1.53 segundos por série temporal, com $\hat{\sigma} = 0.29$. Por fim, o modelo clássico SARIMA apresentou duração média de 3.45 segundos por série temporal, com $\hat{\sigma} = 0.29$.

É válido ressaltar que, a mensuração do tempo gasto pela modelagem SARIMA presente nos modelos híbridos 1 e 2, e também no modelo clássico SARIMA, não está considerando o tempo relativo aos passos iterativos da metodologia de Box & Jenkins, cuja identificação do modelo pode demorar de minutos a horas, dependendo do usuário e série temporal. A mensuração do desempenho computacional da modelagem SARIMA é baseada no tempo utilizado pela função *auto.arima* do *software* R (ver subseção 2.2.1, a qual identifica os modelos SARIMA por meio da minimização de critérios de informação. Contudo, modelos SARIMA identificados pela minimização de critérios de informação, em geral, deixam a desejar no que refere-se a acurácia preditiva.

2.4 Conclusão

O modelo híbrido 4, proposto neste estudo, apresentou o melhor desempenho de previsão à longo prazo, 24 passos à frente, para os três tipos de séries simuladas X_t^1, X_t^2 e X_t^3, dentre os cinco modelos confrontados. Além disso, seu desempenho preditivo foi potencializado ao abordar *wavelets* mãe com elevado número de coeficientes não nulos L na execução da transformada *wavelet*. Nessa perspectiva, em particular, as *wavelets* mãe pertinentes as famílias *Best Localized*, Daubechies e *Least Asymetric* otimizaram as previsões fornecidas pelo modelo híbrido 4. Quanto ao número de níveis de resolução, verificou-se o equilíbrio entre erro de ajuste e previsão ao considerar a decomposição *wavelet* com 6 níveis de resolução. Obviamente, o principal interesse deste estudo é a redução do erro de previsão, contudo, não se recusa o ganho em termos do ajuste. Além disso, uma quantidade inferior a 4 níveis de resolução, não forneceu benefícios ao procedimento preditivo dos modelo híbridos.

É válido comentar que, dentre os três tipos de séries simuladas X_t^1, X_t^2 e X_t^3, o modelo híbrido 4 apresentou desempenho preditivo particularmente excelente para a série X_t^3, a qual representa uma série com componente sazonal. Tal resultado pode ser associado a propriedade peculiar dos modelos probabilísticos clássicos ajustados em cada subsérie de coeficientes *wavelets* concernentes a análise *wavelet*.

Portanto, para as séries simuladas X_t^1, X_t^2 e X_t^3, o modelo híbrido 4 atingiu sua melhor performance ao considerar a decomposição *wavelet* com 6 níveis de resolução e utilizar *wavelets* mãe com elevado número de coeficientes e pertinentes às famílias *Best Localized*, Daubechies e *Least Asymetric*, para a construção das subséries de coeficientes *wavelets* dadas pela transformada *wavelet*.

Além de apresentar o melhor desempenho preditivo, o modelo híbrido 4 também apresentou o melhor desempenho computacional, utilizando, em média, 0.51 segundos ($\hat{\sigma} = 0.06$) para fornecer o ajuste e previsão de uma série temporal, sendo este o menor tempo dentre os cinco modelos confrontados.

Capítulo 3

Modelo híbrido para previsão da taxa de mortalidade infantil no Brasil

3.1 Introdução

A taxa de mortalidade infantil não resume-se somente a proporção de crianças que morrem antes do primeiro ano de vida para cada mil nascidos vivos, mas traduz as condições de vida e assistência de saúde fornecidas por um país. Nessa perspectiva, a taxa de mortalidade infantil é usada como indicador de desenvolvimento humano, a qual apresenta números alarmantes quando refere-se a países pobres. Em nível mundial, os piores resultados pertencem as nações africanas e asiáticas. De acordo com Organização Mundial da Saúde (OMS), a África Subsaariana, composta pela região do continente africano ao sul do Deserto do Saara, tem a maior taxa de mortalidade infantil no mundo, de modo que 1 em cada 12 crianças morre nos primeiros cinco anos de vida (YOU et al., 2011). Segundo o *The World Factbook*, publicada anualmente pela *Central Intelligence Agency* (CIA) dos Estados Unidos, em 2016, a taxa de mortalidade infantil no Afeganistão atingiu 112 mortes a cada mil nascidos vivos (CIA, 2017).

Conforme a classificação internacional da OMS, a partir de 50 mortes a cada mil nascidos vivos, são consideradas taxas altas, entre 20 e 49 mortes a cada mil nascidos vivos, são taxas médias e abaixo de 20 mortes a cada mil nascidos vivos, são consideradas taxas baixas. A taxa de mortalidade infantil brasileira é classificada como média. Entretanto, uma comparação feita pelo Fundo das Nações Unidas para a Infância (UNICEF) em 2005, apontou o Brasil como a terceira maior taxa da América do Sul, perdendo apenas para a Bolívia e Guiana (UNICEF et al., 2005). As piores taxas de mortalidade infantil no território brasileiro concentram-se na região do Semiárido, a qual engloba os estados de Alagoas, Bahia, Ceará, Minas Gerais, Paraíba, Pernambuco, Piauí, Rio Grande do Norte e Sergipe. Os valores chegam a superar as taxas de mortalidade de alguns países africanos.

Segundo o Ministério da Saúde, entre os anos de 2001 e 2003, alguns municípios ultrapassaram a taxa de mortalidade infantil do Sudão, dada por 63 mortes a cada mil nascidos vivos (UNICEF et al., 2005).

Visando a indicação das regiões com altas taxas de mortalidade infantil, este estudo tem por objetivo aplicar o modelo híbrido proposto de previsão nas taxas de mortalidade infantil pertinentes aos estados brasileiros e Distrito Federal. Dessa forma, o modelo híbrido proposto poderá nortear a atenção das políticas públicas para os estados brasileiros que apresentarão as maiores taxas de mortalidade infantil nos próximos 24 meses, tendo em vista a redução e equidade dessas taxas dentro do território nacional. Apesar de seus benefícios, nesse momento, os dados serão utilizados para fim metodológico, sendo a previsão realizada dentro do período amostral, a fim de comparar o desempenho preditivo entre o modelo híbrido proposto e modelos clássicos SARIMA. A aplicação do modelo híbrido proposto para previsões fora do período amostral ficará como propostas futuras.

3.2 Metodologia

Os dados em estudo foram obtidos por meio do site do Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS) e destes calculou-se as taxas mensais 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 mensais infantil e o número de nascidos vivos, no período entre Janeiro de 1996 a Dezembro de 2014, totalizando 228 observações. A seguir, tem-se a fórmula utilizada para o cálculo das taxas,

$$Taxa_{i,j} = \frac{\text{Número de Óbitos Infantil}_{i,j}}{\text{Número de Nascidos Vivos}_{j}} * 1.000, \tag{3.1}$$

em que, $i = 1, ..., 12$ (mês) e $j = 1996..., 2014$ (ano). O óbito infantil diz respeito ao óbito de crianças com idade inferior a 1 ano, enquanto que, os nascidos vivos referem-se as crianças que, depois da separação do corpo materno, respire ou dê qualquer outro sinal de vida. Acerca da previsão, será comparada a qualidade desta a partir de dois modelos:

- SARIMA clássico: o modelo SARIMA de acordo com as etapas de identificação, estimação, validação e previsão da metodologia de Box & Jenkins;
- Modelo híbrido: consiste da 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 (D2), Daubechies 18 (D18), Daubechies 20 (D20) e *Best Localized* 18 (BL18).

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 (RMSE). O melhor modelo de previsão foi aquele que apresentou o menor RMSE associado a previsão (ver subseção 2.2.4 no capítulo 2). Além disso, para comparar o desempenho preditivo do modelo híbrido proposto ao desempenho do modelo clássico SARIMA, calculou-se a porcentagem de redução do erro de previsão do modelo clássico SARIMA ao substituí-lo pelo modelo híbrido proposto. A expressão da proporção de redução é dada por,

$$Prop_{[i]} = \left(\frac{\textit{RMSE do modelo híbrido}_{[i]}}{\textit{RMSE do modelo clássico SARIMA}_{[i]}} \right), \tag{3.2}$$

em que $i = 1, \ldots, 27$ representa as vinte e sete séries temporais referentes aos estados brasileiros e Distrito Federal. O complementar de $Prop_{[i]}$ resultará na proporção de redução do erro, como segue,

$$\text{Redução}_{[i]} = 1 - Prop_{[i]}, \tag{3.3}$$

em que $i = 1, \ldots, 27$. Multiplicando (3.3) por 100 tem-se a porcentagem de redução do erro de previsão. A análise foi desenvolvida com o auxílio do *software* R (R Core Team, 2016).

3.3 Resultados e Discussões

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, Rondônia e Goiás não precisaram de transformação.

Os modelos SARIMA foram ajustados nas séries das taxas mensais de mortalidade infantil no período entre Janeiro de 1996 e Dezembro de 2013. Contudo, para 11 séries temporais não foi identificado um modelo SARIMA capaz de modelá-las, de maneira que os resíduos fossem estacionários, isto é, $\varepsilon_i \sim i.i.d.N(0, \sigma^2)$. 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 pelos modelos SARIMA e híbrido proposto. Após, foram

comparadas a acurácia preditiva de ambos modelos. A Figura 11 ilustra os estados e suas respectivas abordagens de modelagem.

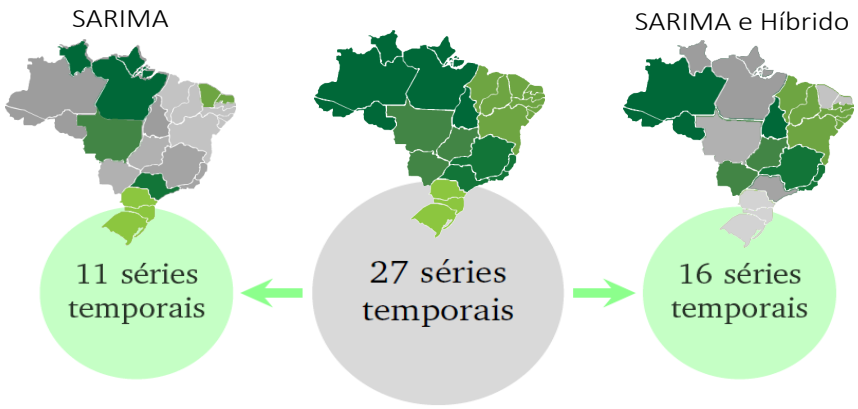

Figura 11 – Modelagem abordada conforme com o estado.

A Tabela 8 apresenta os modelos SARIMA ajustados em 16 séries das taxas de mortalidade infantil, seguindo as etapas da metodologia de Box & Jenkins e destaca para quais séries foram abordadas as modelagens clássica e híbrida ou somente a modelagem híbrida.

Tabela 8 – Modelo ajustado em cada estado brasileiro e Distrito Federal.

Estados	Modelo	Estados	Modelo
1.Rondônia	*SARIMA(3,1,0)(0,0,0)*/híbrido	15.Sergipe	*SARIMA(5,1,0)(0,0,0)*/híbrido
2.Acre	*SARIMA(1,1,2)(0,0,0)*/híbrido	16.Bahia	*SARIMA(0,1,3)(2,1,1)*/híbrido
3.Amazonas	*SARIMA(1,1,1)(0,0,0)*/híbrido	17.Minas Gerais	*SARIMA(3,1,2)(0,0,0)*/híbrido
4.Roraima	híbrido	18.Espírito Santo	*SARIMA(4,1,0)(0,0,0)*/híbrido
5.Pará	híbrido	19.Rio de Janeiro	*SARIMA(5,1,0)(0,0,0)*/híbrido
6.Amapá	híbrido	20.São Paulo	híbrido
7.Tocantins	*SARIMA(2,1,0)(0,0,0)*/híbrido	21.Paraná	híbrido
8.Maranhão	*SARIMA(1,1,1)(0,0,0)*/híbrido	22.Santa Catarina	híbrido
9.Piauí	*SARIMA(1,1,2)(0,0,0)*/híbrido	23.Rio Grande do Sul	híbrido
10.Ceará	híbrido	24.Mato Grosso do Sul	*SARIMA(4,1,0)(0,0,0)*/híbrido
11.Rio Grande do Norte	híbrido	25.Mato Grosso	híbrido
12.Paraíba	*SARIMA(0,1,3)(0,0,0)*/híbrido	26.Goiás	*SARIMA(4,1,0)(0,0,0)*/híbrido
13.Pernambuco	*SARIMA(2,1,3)(0,0,0)*/híbrido	27.Distrito Federal	híbrido
14.Alagoas	*SARIMA(1,1,1)(0,0,0)*/híbrido		

A Tabela 9 apresenta os valores do RMSE associado a previsão, tanto do modelo SARIMA como para o modelo híbrido proposto. Além disso, indica a *wavelet* mãe em-

pregada na transformada *wavelet* presente na modelagem híbrida. Pode-se verificar que a metodologia clássica SARIMA possui maiores erros de previsão para quase todas as séries temporais, evidenciando a superioridade do modelo híbrido proposto de previsão.

Tabela 9 – 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)			

Entretanto, para a série da taxa de mortalidade infantil do estado de Minas Gerais, o modelo SARIMA apresentou melhor desempenho preditivo com relação ao modelo híbrido proposto. Para compreender o motivo desse resultado, a Figura 12 ilustra o comportamento da previsão tanto do modelo SARIMA como do modelo híbrido proposto.

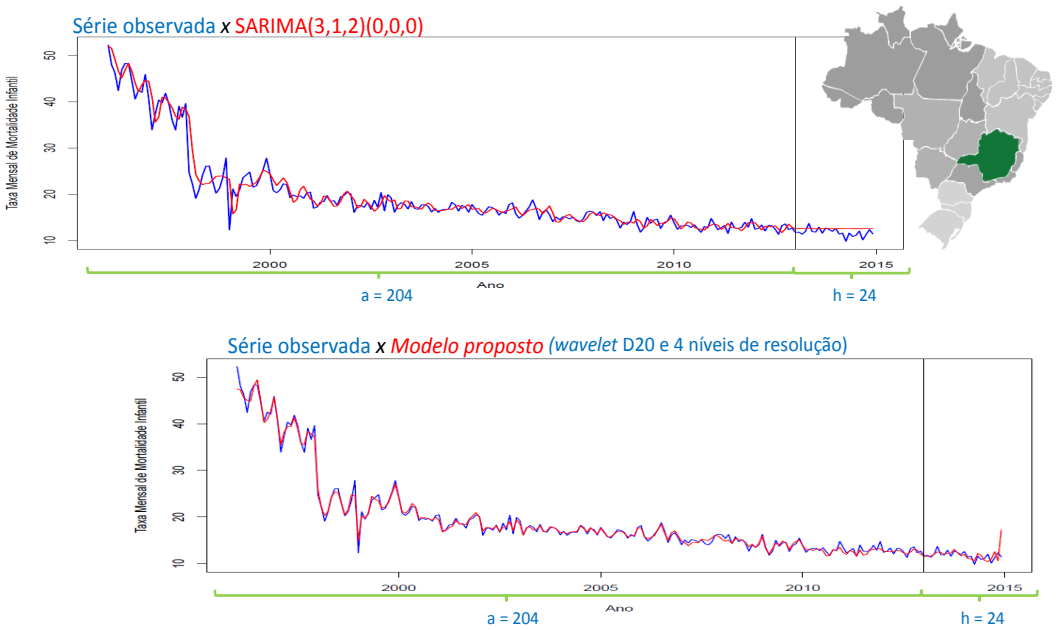

Figura 12 – Modelos ajustados para as taxas de mortalidade infantil de Minas Gerais.

Observe que, ao considerar as previsões de curto prazo (h=6) e médio prazo (h=12), os valores previstos pelo modelo híbrido proposto estão visivelmente mais próximo dos valores reais em comparação aos valores previstos pelo modelo SARIMA ajustado. No entanto, o comportamento destoante dos últimos valores previsto pelo modelo híbrido proposto, acarretou na mensuração elevada do RMSE em relação ao modelo SARIMA. Portanto, a representação gráfica mostra que, de forma geral, o modelo híbrido apresentou desempenho preditivo superior ao modelo SARIMA. Aqui, é interessante salientar a importância da análise gráfica junto aos valores das estatísticas.

A Figura 13 mostra as séries e seus respectivos estados que apresentaram as maiores reduções, em porcentagem, do RMSE de previsão ao substituir o modelo SARIMA pelo modelo híbrido proposto. Em azul, tem-se a série da taxa de mortalidade infantil observada e, em vermelho, o ajuste e previsão do modelo híbrido proposto. As porcentagens indicadas no canto superior dos gráficos, em verde, representam as reduções no erro. A linha vertical, presente nestes gráficos, marca o início do período de previsão de 24 meses.

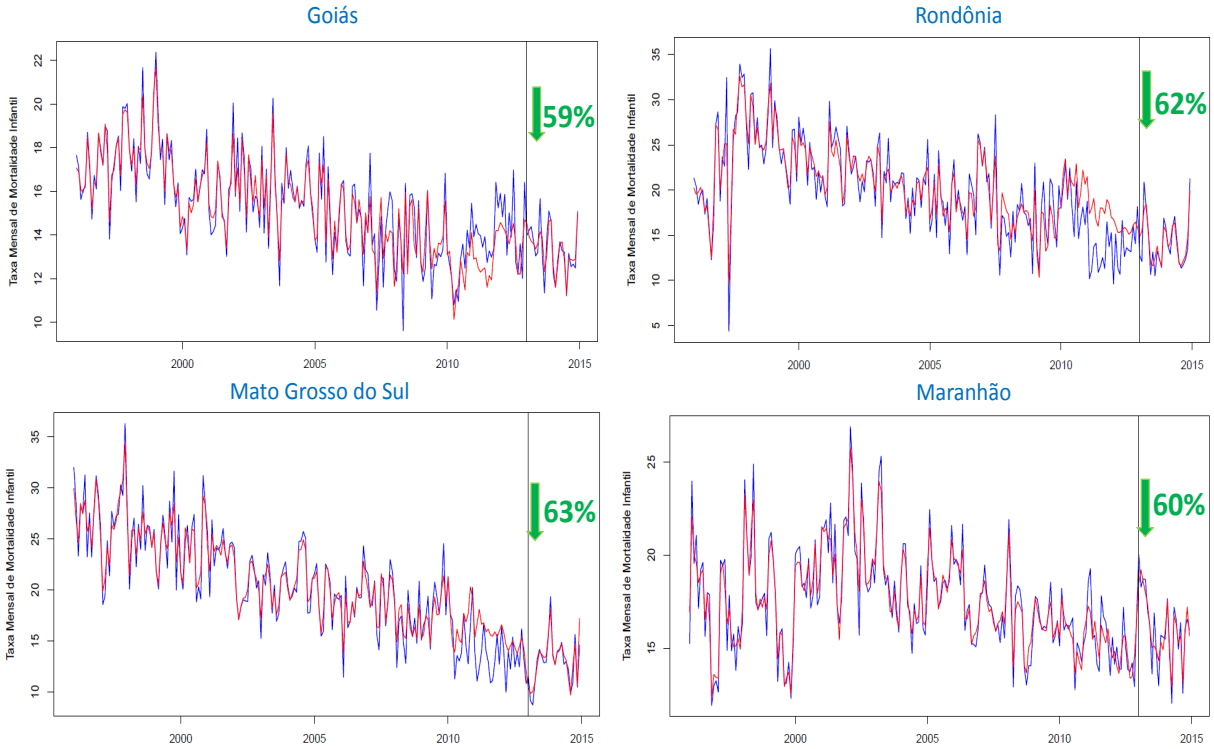

Figura 13 – Maiores reduções do erro de previsão, em porcentagem, fornecidas pelo modelo híbrido proposto com relação ao modelo SARIMA.

Em relação ao desempenho computacional do modelo híbrido, em média, o tempo gasto para realizar o ajuste e previsão para as séries das taxas mensais de mortalidade infantil foi igual a 0.52 segundos por série temporal modelada, com desvio padrão $\hat{\sigma} = 0.14$. A mensuração do desempenho computacional dos modelos clássicos SARIMA não foi re-

alizada, pois a identificação desses modelos segundo os passos iterativos da metodologia de Box & Jenkins pode demorar de minutos a horas, dependendo do conhecimento do usuário e do comportamento da série temporal.

3.4 Conclusão

O modelo híbrido proposto, neste estudo, apresentou o melhor desempenho preditivo de longo prazo em comparação ao modelo SARIMA para as 16 séries das taxas de mortalidade infantil, em que foram possíveis a modelagem clássica SARIMA e a modelagem híbrida. Portanto, o modelo híbrido proposto é uma eficaz ferramenta para a obtenção de informações prévias e precisas do comportamento das taxas de mortalidade infantil nas diferentes regiões brasileiras. Sendo, agora, recomendado para a previsão fora do período amostral.

Capítulo 4

Modelo híbrido proposto para previsão de hospitalizações por Bronquiolite no Brasil

4.1 Introdução

A bronquiolite viral aguda (BVA) é uma doença respiratória que promove a infecção dos bronquíolos, ramificações menores dos brônquios. Tal infecção é ocasionada por vírus respiratórios, dentre os quais os mais prevalentes são: vírus sincicial respiratório (VSR), adenovírus (ADV), vírus influenza tipo A e B, e o vírus parainfluenza (PIV) tipo 1, 2 e 3 (MONTEIRO et al., 2014). Contudo, o principal responsável pelo diagnóstico de BVA em internações de crianças com idade inferior à 12 meses é o vírus sincicial respiratório (VSR) (LOURENÇÃO et al., 2005). Estima-se que cerca de 40% a 60% das crianças são infectadas pelo VSR nessa faixa etária, sendo este vírus um dos causadores da mortalidade de crianças no primeiro ano de vida no Brasil (ALBERNAZ et al., 2003).

A incidência de infecções virais possui comportamento sazonal conforme as condições climáticas de cada região. Períodos com baixas temperaturas proporcionam o aumento de infecções ocasionadas pelo VSR (CALEGARI et al., 2005). Nessa perspectiva, estudos mostram que o pico sazonal da incidência de infecção pelo VSR na cidade do Rio de Janeiro, inicia-se em Março e encerra-se em Maio (NASCIMENTO et al., 1991). Entre os meses de Abril/Maio a Julho/Agosto ocorre o pico sazonal na cidade de São Paulo (VIEIRA et al., 2001). Em Salvador, tal período dá-se nos meses de Maio a Julho (MOURA et al., 2003), porém em Porto Alegre, o pico sazonal de infecções por VSR ocorre durante os meses de Junho a Setembro (STRALIOTTO et al., 2004).

Dada a inexistência de vacinas, outros meios de prevenção são utilizados no combate ao VSR, como o anticorpo humanizado Palivizumabe, um medicamento que neutraliza e inibe a fusão contra o VSR. Tal medicação é administrada em cinco doses mensais du-

rante os meses de maior incidência do vírus. Visando otimizar o efeito preventivo, é ideal que a primeira dose seja administrada um mês antes do início do pico sazonal. Entretanto, devido ao alto custo, a disponibilização do Palivizumabe é restrita ao grupo de risco, composto por: prematuros com menos de 32 semanas, presença de cardiopatia congênita em menores de 2 anos e crianças portadoras de displasia broncopulmonar sintomática (PEDIATRIA, 2011).

Mediante essas considerações, é inquestionável a importância de estudos que buscam identificar o pico sazonal da incidência de infecções por VSR nas regiões brasileiras, de modo que a administração do Palivizumabe possa ser iniciada a tempo de prevenir a transmissão, viabilizando um investimento financeiro eficiente no combate ao VSR. Este capítulo tem por objetivo aplicar o modelo híbrido de previsão, proposto neste trabalho, nas taxas de internações por bronquiolite.

O modelo híbrido proposto de previsão vem suprir a necessidade de informações prévias e precisas sobre o comportamento das taxas de internações por bronquiolite nas diferentes regiões brasileiras, norteando estratégias de administração do Palivizumabe. É válido esclarecer que, as previsões serão realizadas dentro do período amostral, de modo a comparar o desempenho de previsão do modelo híbrido proposto com modelos usuais em epidemiologia. Assim, nesse capítulo, os dados são de importância metodológica e não biológica. A utilização do modelo híbrido para a previsão fora do período amostral ficará como propostas futuras.

4.2 Metodologia

Os dados em estudo foram obtidos por meio do site do Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS) e destes calculou-se as taxas mensais de internações por bronquiolite a cada 10.000 crianças referentes aos estados brasileiros e Distrito Federal. As taxas foram obtidas a partir da razão entre o número mensal de internações por bronquiolite e o número anual de nascidos vivos pertinente a cada estado e Distrito Federal, no período entre Janeiro de 2000 a Setembro de 2016, totalizando 201 observações. Isto é,

$$Taxa_{i,j} = \frac{\text{Número de internações}_{i,j}}{\text{Número de Nascidos Vivos}_j} * 10.000, \tag{4.1}$$

em que, $i = 1, ..., 12$ (mês) e $j = 2000, ..., 2016$ (ano).

É válido esclarecer que, o DATASUS disponibiliza o número de nascidos vivos até 2014. Porém, o conjunto de dados pertinente as internações por bronquiolite possuem informações mensais até 2016. Para obter os dados de nascidos vivos referente aos anos de 2015 e 2016, ajustou-se uma curva polinomial cúbica na série do número de nascidos

vivos, de forma que seu comportamento foi projetado fornecendo os últimos dois valores faltantes. Em seguida, obteve-se a média entre os valores da projeção da curva polinomial cúbica e os valores obtidos pela projeção ingênua, obtendo o número de nascidos vivos para os anos de 2015 e 2016.

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

- SARIMA clássico: o modelo SARIMA, comumente utilizado na análise de séries temporais epidemiológicas, cuja modelagem seguem as etapas de identificação, estimação, validação e previsão, sistematizadas pela metodologia de Box & Jenkins;
- Modelo híbrido proposto:
    1. Inicia-se com a decomposição da série temporal via transformada *wavelet*, obtendo o modelo apresentado na subseção 2.1.5.1 do capítulo 2. Assim, obtém-se as subséries de coeficientes *wavelets* representadas nos níveis de resolução. A decomposição multiescala foi feita em 4 níveis a partir da *wavelet* mãe Daubechies 20 (D20). Outras *wavelets* mãe poderiam ser utilizadas, como *Least Asymetric*, Coiflets e *Best Localized*;
    2. Para cada uma das 4 subséries de coeficientes *wavelets* obtidas no passo 1, um modelo probabilístico clássico é ajustado para obter as previsões de acordo com o prazo de previsão (curto, médio e longo);
    3. A transformada *wavelet* inversa é aplicada para obter as previsões no domínio do tempo.

Para avaliar sua acurácia, as previsões foram feitas dentro do período amostral e comparadas com os valores observados. Foram analisadas previsões de curto prazo (6 meses), médio prazo (12 meses) e longo prazo (24 meses). O período de previsão de curto prazo, foi entre Abril de 2016 e Setembro de 2016. O período de previsão de médio prazo, foi entre Outubro de 2015 e Setembro de 2016. O período de previsão de longo prazo, foi entre Outubro de 2014 e Setembro de 2016. A mensuração do erro de previsão foi realizada por meio da Raiz do Erro Quadrático Médio (RMSE). O melhor modelo de previsão foi aquele que apresentou o menor RMSE associado a previsão (ver subseção 2.2.4 no capítulo 2).

Além disso, para comparar o desempenho preditivo do modelo híbrido proposto em relação ao desempenho do modelo clássico SARIMA, calculou-se a porcentagem de redução do erro de previsão do modelo SARIMA ao substituí-lo pelo modelo híbrido proposto. A expressão da proporção de redução é dada por,

$$Prop_{[i]} = \left(\frac{\textit{RMSE do modelo híbrido}_{[i]}}{\textit{RMSE do modelo clássico SARIMA}_{[i]}} \right), \tag{4.2}$$

em que $i = 1, \ldots, 27$ representa as vinte e sete séries temporais referentes aos estados brasileiros e Distrito Federal. O complementar de $Prop_{[i]}$ resultará na proporção de redução do erro, como segue,

$$\text{Redução}_{[i]} = 1 - Prop_{[i]}, \tag{4.3}$$

em que $i = 1, \ldots, 27$. Multiplicando a expressão (4.3) por 100 tem-se a porcentagem de redução do erro de previsão. A previsão foi baseada puramente nas observações da série temporal, sem a interferência de informações adicionais. A análise foi desenvolvida com o auxílio do *software* R (R Core Team, 2016).

O desenvolvimento desse estudo ocorreu conforme recomendado pela Resolução 466/2012 do Conselho Nacional de Saúde. O projeto foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Estadual de Maringá (Parecer 739.422/2014) e o Termo de Consentimento Livre e Esclarecido não foi utilizado visto que os dados foram obtidos de fontes secundárias.

4.3 Resultados e Discussões

As Figuras 14 , 15 e 16 ilustram os valores observados (azul), os valores ajustados e previstos pelo modelo SARIMA (verde) e modelo híbrido (vermelho) para as séries que apresentaram a maior e a menor redução, em porcentagem, do erro de previsão conforme o prazo de previsão. A linha vertical, presente nestas figuras, representa o início do período de previsão.

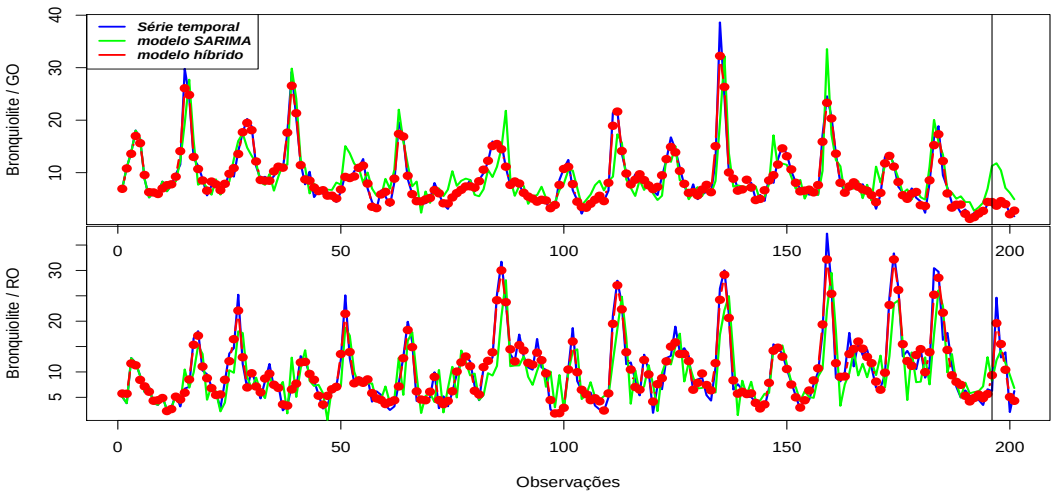

Figura 14 – Previsão de curto prazo. Séries dos estados do Goiás (GO) e Rondônia (RO), respectivamente.

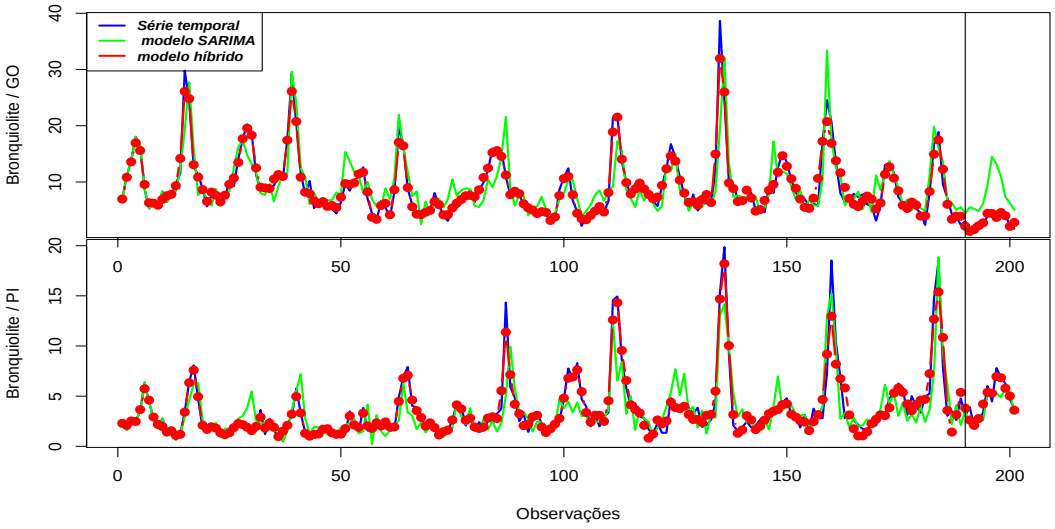

Figura 15 – Previsão de médio prazo. Séries dos estados de Goiás (GO) e Piauí (PI), respectivamente.

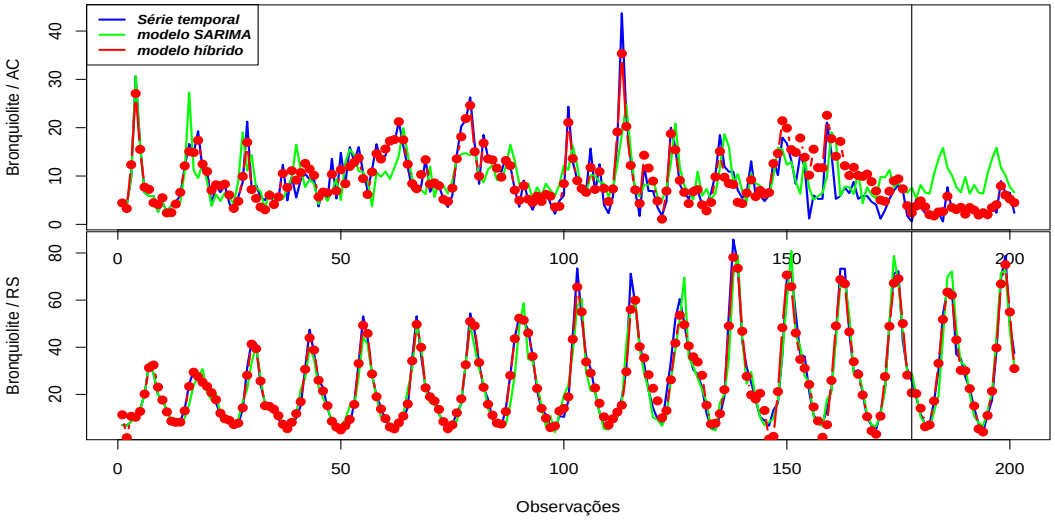

Figura 16 – Previsão de longo prazo. Séries pertinentes aos estados do Acre (AC) e Rio Grande do Sul (RS), respectivamente.

A Tabela 10 apresenta os erros de previsão de curto, médio e longo prazo apresentados pelos modelos SARIMA e híbrido proposto, além da redução, em porcentagem, do erro de previsão ao optar pelo modelo híbrido proposto ao invés do modelo SARIMA.

Tabela 10 – Raiz do Erro Quadrático Médio segundo o prazo de previsão.

Previsão	curto prazo			médio prazo			longo prazo		
Estados brasileiros	SARIMA	Híbrido	Redução (%)	SARIMA	Híbrido	Redução (%)	SARIMA	Híbrido	Redução (%)
1.Rondônia	6.00	3.33	**44.48**	6.25	2.47	60.38	6.63	2.93	52.90
2.Acre	4.95	1.32	73.33	6.08	0.98	83.74	6.96	1.10	**84.18**
3.Amazonas	5.57	0.96	82.68	4.17	0.79	81.04	3.09	0.98	68.35
4.Roraima	3.70	1.89	48.91	3.19	1.60	49.69	4.01	1.40	65.15
5.Pará	6.78	0.90	86.68	5.02	0.88	82.30	3.46	1.55	55.03
6.Amapá	6.20	1.88	69.57	4.49	1.36	69.68	4.43	1.55	64.89
7.Tocantins	0.97	0.33	65.48	2.28	0.51	77.38	3.45	1.25	63.76
8.Maranhão	0.70	0.26	62.82	2.01	0.21	89.10	1.81	0.30	82.98
9.Piauí	1.47	0.54	63.13	1.20	0.65	**45.44**	2.35	1.30	44.50
10.Ceará	4.30	0.80	81.22	5.14	0.76	85.12	5.02	1.27	74.64
11.Rio Grande do Norte	3.66	0.60	83.52	2.57	0.56	77.90	4.34	0.77	82.06
12.Paraíba	4.15	0.66	84.03	2.96	0.55	81.41	2.97	0.68	77.14
13.Pernambuco	3.98	1.12	71.73	4.55	0.89	80.42	3.79	1.02	72.93
14.Alagoas	1.97	0.49	74.68	1.46	0.36	75.05	1.28	0.35	72.67
15.Sergipe	6.69	1.39	80.02	4.97	1.15	76.79	4.00	1.11	72.19
16.Bahia	3.85	0.61	84.10	2.23	0.48	78.02	2.14	0.82	61.53
17.Minas Gerais	6.77	0.88	86.90	5.67	0.96	82.97	4.04	1.45	64.00
18.Espírito Santo	10.50	2.57	75.50	7.68	1.98	74.10	6.07	2.06	65.97
19.Rio de Janeiro	3.67	0.73	80.06	3.25	0.83	74.47	3.36	0.94	71.89
20.São Paulo	3.69	1.38	62.64	2.68	1.07	60.07	2.61	1.10	57.85
21.Paraná	3.69	0.82	77.57	2.66	0.69	74.05	1.88	0.72	61.36
22.Santa Catarina	2.74	0.97	64.35	2.02	0.72	64.37	2.15	0.72	66.21
23.Rio Grande do Sul	7.99	4.35	45.60	6.67	3.23	51.48	5.08	3.15	**37.95**
24.Mato Grosso do Sul	5.98	1.16	80.55	4.58	1.01	77.94	4.56	1.06	79.69
25.Mato Grosso	1.93	0.61	68.09	1.97	0.53	72.71	1.81	0.73	59.71
26.Goiás	5,54	0.53	**90.33**	5.43	0.58	**89.32**	4.96	0.99	80.02
27.Distrito Federal	13.37	2.16	83.81	12.92	2.06	84.01	13.52	5.08	62.42

Em comparação ao modelo SARIMA, o modelo híbrido proposto foi capaz de reduzir o erro de previsão para as 27 séries temporais observadas, independente do comprimento do prazo de previsão. As maiores reduções do erro de previsão fornecidas pelo modelo híbrido para as previsões de curto, médio e longo prazos foram: 90% para a série do estado de Goiás, 89% para a série do estado de Goiás e 84% para a série do Acre, respectivamente. Em contrapartida, as menores reduções do erro de previsão fornecidas pelo modelo híbrido para as previsões de curto, médio e longo prazos foram: 44% para a série do estado de Rondônia, 45% para a série do estado do Piauí e 37% para a série do estado do Rio Grande do Sul, respectivamente.

Ainda, sob outra perspectiva, o erro de previsão pode ser interpretado em termos do número de internações. Para exemplificar, considere a previsão de longo prazo (h=24). Assim, no estado do Acre foram observadas, em média, 33.800 internações por bronquiolite entre o período de Janeiro de 2015 e Dezembro de 2016. Segundo o modelo SARIMA, a média de internações por bronquiolite previstas para o mesmo período é igual a 93.000

internações. Enquanto que, conforme o modelo híbrido proposto, a média de internações por bronquiolite previstas para este período é igual a 35.600 internações. Portanto, o erro de previsão do modelo SARIMA em termos do número de internações foi igual a 59.200 internações acima do valor real observado. Por outro lado, o erro de previsão do modelo híbrido foi igual a 1.800 internações acima do valor real observado. Dessa forma, o modelo híbrido reduziu o erro de previsão do modelo SARIMA em 57.400 internações.

O parágrafo anterior registra a situação na qual o modelo SARIMA fornece previsões muito além dos verdadeiros valores de internações observados. Entretanto, para o estado do Rio Grande do Norte, ocorreu o oposto. Em média, o número real de internações observadas no período entre de Janeiro de 2015 e Dezembro de 2016 foi igual a 63.400 internações. Segundo o modelo SARIMA, o valor médio previsto para este período é igual a 34.800 internações. De acordo com o modelo híbrido proposto, o valor médio previsto é igual a 63.000 internações. Nesse caso, o erro de previsão do modelo SARIMA em termos do número de internações foi igual a 28.600 internações abaixo do valor real observado. Por outro lado, o erro de previsão do modelo híbrido foi igual a 400 internações abaixo do valor real observado. Sendo assim, o modelo híbrido reduziu o erro de previsão do modelo SARIMA em 28.200 internações.

Quanto ao desempenho computacional do modelo híbrido, em média, o tempo gasto para realizar o ajuste e previsão para as séries das taxas mensais de internações por bronquiolite foi igual a 0.46 segundos por série temporal modelada, com desvio padrão $\hat{\sigma} = 0.01$. A mensuração do desempenho computacional dos modelos clássicos SARIMA não realizada, pois a identificação desses modelos segundo os passos iterativos da metodologia de Box & Jenkins pode demorar de minutos a horas, dependendo do conhecimento do usuário e do comportamento da série temporal.

4.4 Conclusão

O modelo híbrido, proposto neste estudo, apresentou o melhor desempenho preditivo de curto, médio e longo prazos para as 27 séries das taxas mensais de internações por bronquiolite referentes aos estados brasileiros e Distrito Federal. Isso significa que, o modelo híbrido superou o desempenho preditivo do modelo SARIMA, comumente utilizado, em todas as circunstâncias. Assim, o modelo híbrido proposto, é uma promissora e eficaz ferramenta para a obtenção de informações prévias e precisas sobre o comportamento das taxas de internações por bronquiolite nas diferentes regiões brasileiras.

Capítulo 5

Considerações finais e Trabalhos futuros

O conhecimento prévio e preciso fornece instruções para uma preparação antecipada, a qual sempre trará benefícios. Em epidemiologia, tais benefícios traduzem-se no controle de doenças. Nesse contexto, previsões acuradas poderão revelar instabilidades futuras no crescimento do número de casos de determinada doença, norteando a tomada de decisão de políticas públicas para a prevenção de surtos. Ainda, previsões acuradas poderão denunciar o comportamento futuro de uma doença, indicando o período de maior incidência e direcionando o melhor momento para iniciar uma campanha de imunização ou a distribuição de medicamentos. Além disso, previsões acuradas poderão expressar as regiões geográficas mais atingidas por uma doença, voltando a atenção das autoridades públicas para a urgência dessas regiões. Em síntese, no contexto epidemiológico, informações precoces acuradas implicam em estratégias eficazes de combate a doenças, promovendo o desenvolvimento humano.

O modelo híbrido proposto neste trabalho, vem otimizar as análises de séries temporais epidemiológicas, devido ao seu notável desempenho preditivo em relação aos modelos SARIMA comumente utilizados. Através de séries simuladas, no capítulo 2, verificou-se a superioridade de previsões à longo prazo da metodologia proposta. Tal resultado foi confirmado no capítulo 3, ao realizar previsões de 24 passos à frente para as taxas mensais de mortalidade infantil referente aos estados brasileiros e Distrito Federal. Além da capacidade de previsão à longo prazo, no capítulo 4, o modelo híbrido proposto também apresentou bons resultados para previsão de curto e médio prazos, ou seja, previsões de 6 passos e 12 passos à frente, respectivamente. Portanto, independente do horizonte de previsão, o modelo híbrido proposto forneceu previsões acuradas.

Além disso, sabe-se que a classe de modelos ARIMA é adequada para séries estacionárias (BOX; JENKINS, 1970), logo é necessário remover fontes de variação não estacionárias, tais como tendência e sazonalidade (ver subseção 2.1.2). Contudo, mesmo com

a condição de estacionariedade, a previsão de modelos ARIMA é recomendada apenas para o curto prazo. Tal limitação é superada pela modelagem híbrida a partir de *wavelets*, na qual previsões acuradas podem ser obtidas tanto para séries estacionárias quanto não-estacionárias, além de horizontes de previsão curtos, médios ou longos.

Por envolver a análise *wavelets*, a carência dos conhecimentos estatísticos necessários para a desenvolvimento do modelo híbrido proposto pode desmotivar sua aplicação, principalmente em pesquisadores de outras áreas, como a área da saúde. Contudo, outra grande vantagem desse modelo é a automaticidade presente em sua estrutura, uma vez implementado. Os modelos clássicos probabilísticos a serem ajustados em cada nível de resolução da análise *wavelet* são fixos e simples. Logo, o dispêndio do modelo híbrido proposto é eliminado tanto em nível de usuário, pois seu mecanismo independe de conhecimento prévios, como em nível computacional, devido a simplicidade dos modelos clássicos ajustados em cada subsérie e a eficiência dos algoritmos de decomposição *wavelets*. Portanto, mesmo envolvendo conceitos da análise *wavelet*, o modelo híbrido proposto pode ser facilmente utilizado em aplicações de todas as áreas do conhecimento, uma vez implementado.

Nesse sentido, como trabalho futuro, propõe-se o desenvolvimento de um possível *software* para a previsão de séries temporais, baseado no mecanismo do modelo híbrido proposto, de modo a viabilizar sua aplicação interdisciplinar. Além disso, melhorias técnicas do modelo são propostas, como o tratamento de possíveis problemas de borda no ajuste. Nesse trabalho, não foi considerado nenhum tratamento de borda, o que poderia melhorar ainda mais o ajuste do modelo híbrido proposto. Também, propõe-se como trabalhos futuros a seguinte questão: O quão curta pode ser a série temporal, de modo que a excelência preditiva do modelo híbrido proposto seja mantida? Tal questão, vem atender a necessidade de previsões acuradas para séries temporais compostas por poucas observações.

Referências

ALBERNAZ, E. P. et al. Fatores de risco associados à hospitalização por bronquiolite aguda no período pós-neonatal. *Revista de Saúde Pública*, SciELO Public Health, v. 37, n. 4, p. 485–493, 2003.

ARINO, M. A. *Time series forecasts via wavelets: an application to car sales in the Spanish market*. [S.l.]: Institute of Statistics & Decision Sciences, Duke University, 1995.

BAYER, F. M. Wavelets e modelos tradicionais de previsao: Um estu do comparativo. *Rev. Bras. Biom*, v. 28, n. 2, p. 40–61, 2010.

BOX, G. E.; JENKINS, G. *Time series analysis: forecasting and control*. [S.l.]: Holden-Day, 1970.

CALEGARI, T. et al. Clinical-epidemiological evaluation of respiratory syncytial virus infection in children attended in a public hospital in midwestern brazil. *Brazilian Journal of Infectious Diseases*, SciELO Brasil, v. 9, n. 2, p. 156–161, 2005.

CIA. *The Wold Factbook*. 2017. Disponível em: <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2091rank.htm>.

CONEJO, A. et al. Day-ahead electricity price forecasting using the wavelet transform and arima models. *IEEE Transactions on Power Systems*, IEEE, v. 20, 05 2005.

DAUBECHIES, I. Where do wavelets come from? a personal point of view. *Proceedings of the IEEE*, IEEE, v. 84, n. 4, p. 510–513, 1996.

DAUBECHIES, I. et al. *Ten lectures on wavelets*. [S.l.]: SIAM, 1992. v. 61.

DONOHO, D. L.; JOHNSTONE, I. M. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, JSTOR, p. 425–455, 1994.

DONOHO, D. L.; JOHNSTONE, I. M. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, Taylor & Francis Group, v. 90, n. 432, p. 1200–1224, 1995.

DONOHO, D. L. et al. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 301–369, 1995.

HOMSY, G. V.; PORTUGAL, M. S.; ARAÚJO, J. P. *Wavelets e previsões de séries de tempo: uma análise empírica*. [S.l.]: Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Economia, 2000.

HYNDMAN, R.; KHANDAKAR, Y. Automatic time series forecasting: The forecast package for r 7. 2008. 2007.

KOZAKEVICIUS, A. de J.; BAYER, F. M. Filtragem de sinais via limiarização de coeficientes wavelet. 2014.

LIMA, F. G.; ALMEIDA, F. Previsão de séries temporais financeiras com o uso das wavelets. *XXXVIII EnANPAD*, 2004.

LIMA, P. C. de. Wavelets: uma introdução. *Matemática Universitária*, v. 33, p. 13–44, 2002.

LOURENÇÃO, L. G. et al. Infecções pelo vírus sincicial respiratório em crianças. *Pulmão RJ*, v. 14, n. 1, p. 59–68, 2005.

MAHESWARAN, R.; KHOSA, R. Comparative study of different wavelets for hydrologic forecasting. *Computers & Geosciences*, Elsevier, v. 46, p. 284–295, 2012.

MALLAT, S. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, v. 11, p. 674–693, 1989b.

MALLAT, S. A wavelet tour of signal processing. *Academic Press,*, 2008.

MALLAT, S. G. Multiresolution approximations and wavelet orthonormal bases of l2(r). *Transactions of the American mathematical society*, v. 315, 1989a.

MONTEIRO, A. et al. Infecções respiratórias em crianças menores de dois anos de idade submetidas a profilaxia com palivizumabe. *Rev Paul Pediatr*, v. 32, n. 2, p. 152–8, 2014.

MORETTIN, P. A. *Ondas e Ondaletas Vol. 23*. [S.l.]: Edusp, 1999.

MORETTIN, P. A.; TOLOI, C. *Análise de séries temporais*. [S.l.]: Blucher, 2006.

MORETTIN, P. A.; TOLOI, C. M. de C. *Modelos para previsão de séries temporais*. [S.l.]: Instituto de matematica pura e aplicada, 1981. v. 1.

MOURA, F. E. A. et al. Estudo de infecções respiratórias agudas virais em crianças atendidas em um centro pediátrico em salvador (ba). *Jornal Brasileiro de Patología e Medicina Laboratorial*, SciELO Brasil, v. 39, n. 4, p. 275–282, 2003.

NASCIMENTO, J. P. et al. Longitudinal study of acute respiratory diseases in rio de janeiro: occurrence of respiratory viruses during four consecutive years. *Revista do Instituto de Medicina Tropical de São Paulo*, SciELO Brasil, v. 33, n. 4, p. 287–296, 1991.

NASON, G. *Wavelet Methods in Statistics with R*. [S.l.]: Springer New York, 2010. (Use R!).

NASON, G.; SAPATINAS, T.; SAWCZENKO, A. Statistical modelling of time series using non-decimated wavelet representations. *submitted for publication*, 1997.

NASON, G. P.; SILVERMAN, B. W. The stationary wavelet transform and some statistical applications. In: *Wavelets and statistics*. [S.l.]: Springer, 1995. p. 281–299.

OLIVEIRA, H. de. *Análise de Fourier e Wavelets*. [S.l.]: Hélio Magalhaes de Oliveira, 2007.

PEDIATRIA, S. B. de. *Diretrizes para o Manejo da Infecção Causada pelo Vírus Sincicial Respiratório (VSR)*. 2011. Disponível em: <http://www.sbp.com.br/pdfs/diretrizes_manejo_infec_vsr_versao_final1.pdf>.

PERCIVAL, D. B.; WALDEN, A. T. *Wavelet methods for time series analysis*. [S.l.]: Cambridge university press, 2006. v. 4.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>.

SCHLÜTER, S.; DEUSCHLE, C. *Using wavelets for time series forecasting: Does it pay off?* [S.l.], 2010.

SOUZA, E. M. d. Análise de wavelets para detecção e correção do multicaminho no posicionamento relativo gnss estático e cinemático. Universidade Estadual Paulista (UNESP), 2008.

STRALIOTTO, S. M. et al. Respiratory viruses in the pediatric intensive care unit: prevalence and clinical aspects. *Memórias do Instituto Oswaldo Cruz*, SciELO Brasil, v. 99, n. 8, p. 883–887, 2004.

UNICEF et al. *Situação da infância brasileira 2006: crianças de até 6 anos: o direito à sobrevivência e ao desenvolvimento*. [S.l.]: Unicef, 2005.

VANDAELE, W. *Applied time series and Box-Jenkins models*. [S.l.: s.n.], 1983.

VIDAKOVIC, B. Statistical modeling by wavelets. Wiley, 1999.

VIDAKOVIC, B.; MUELLER, P. Wavelets for kids. *Instituto de Estadística, Universidad de Duke*, 1994.

VIEIRA, S. E. et al. Clinical patterns and seasonal trends in respiratory syncytial virus hospitalizations in são paulo, brazil. *Revista do Instituto de Medicina Tropical de Sao Paulo*, SciELO Brasil, v. 43, n. 3, p. 125–131, 2001.

WADI, S. A.; HAMARSHEH, A.; ALWADI, H. Maximum overlapping discrete wavelet transform in forecasting banking sector. *Applied Mathematical Sciences*, v. 7, n. 80, p. 3995–4002, 2013.

WONG, H. et al. Modelling and forecasting by wavelets, and the application to exchange rates. *Journal of Applied Statistics*, Taylor & Francis, v. 30, n. 5, p. 537–553, 2003.

YOU, D. et al. Levels and trends in child mortality. report 2015. estimates developed by the un inter-agency group for child mortality estimation. New York New York UNICEF 2015., 2011.

YOUSEFI, S.; WEINREICH, I.; REINARZ, D. Wavelet-based prediction of oil prices. *Chaos, Solitons & Fractals*, Elsevier, v. 25, n. 2, p. 265–275, 2005.