

Econometria de séries financeiras: da modelagem clássica às ondaletas

Aluno: Matheus Carneiro Nogueira

Orientador: Edgard Pimentel

1 Introdução

Séries Temporais são objetos de extrema relevância nas mais diversas áreas de conhecimento. Informações como evolução de gastos energéticos, retornos de ativos financeiros e evolução populacional de um território são apenas algumas das aplicações destes objetos. Sendo assim, o estudo de modelos de econometria que fornecem descrições precisas de séries temporais é extremamente relevante. Com isso em mente, focando em análise de modelos econométricos para séries financeiras, foi desenvolvida esta pesquisa comparativa de 3 diferentes modelos. Primeiramente, foi estudado e implementado o modelo ARIMA para a previsão de séries temporais dos retornos de ações de 10 companhias aéreas. Os resultados desta primeira etapa já foram apresentados no relatório anual de 2020 e podem ser acessados em [8]. Em seguida, foram estudados os modelos VAR e um modelo híbrido que combina a teoria das Ondaletas (Wavelets) com a modelagem ARIMA já apresentada utilizando as mesmas 10 séries temporais.

2 Objetivos

Após estudado o modelo ARIMA, uma ferramenta clássica para a análise de séries temporais, e constatado que ele se mostra pouco confiável para a previsão de retornos financeiros, o objetivo passou a ser estudar e entender dois novos modelos com o intuito de implementá-los e comparar a qualidade de suas previsões. Para isso, foram utilizadas as mesmas 10 séries temporais de preços de ações de companhias aéreas. Primeiramente, o objetivo é produzir um modelo VAR com o intuito de buscar, por meio do ferramental disponível por este modelo, interconexões das séries utilizadas de tal modo a melhorar a qualidade da previsão. Por fim, o objetivo passa a ser estudar a teoria de Wavelets a fim de criar um modelo híbrido que une esta teoria ao modelo ARIMA, já conhecido, de modo a utilizar a análise simultânea dos domínios do tempo e frequência como ferramenta para aumentar a qualidade de previsão.

3 Metodologia

Para fins didáticos, a metodologia utilizada no decorrer da pesquisa será dividida em duas grandes partes. Primeiramente serão apresentados os fundamentos teóricos do modelo

VAR com a discussão de seus resultados. Em seguida, os fundamentos teóricos da teoria de Wavelets com a implementação do modelo híbrido e, também, a discussão de seus resultados.

3.1 Modelo VAR

Antes de analisar o modelo VAR (Vector Autoregressive Model), vale relembrar o modelo AR (Autoregressive Model) apresentado em [1]. A equação que define este modelo é

$$r_t = \phi_0 + \phi_1 r_{t-1} + \dots + \phi_p r_{t-p} + a_t \quad (1)$$

onde os valores ϕ_i são os coeficientes a serem estimados, r_t é o valor presente da série, r_{t-i} são os valores defasados e a_t é um ruído branco de média zero e variância σ_a^2 .

Este modelo foi utilizado para compor o modelo ARIMA e prever o comportamento de 10 séries temporais estacionárias e univariáveis de forma independente, cujos resultados estão já foram apresentados e estão disponíveis em [8].

O modelo VAR, como o nome sugere, também é um modelo auto regressivo mas, ao invés de aceitar apenas séries univariáveis, isto é, valores escalares para as variáveis r_t , r_{t-i} e para os coeficientes ϕ_i , é capaz de modelar séries multivariáveis. Sendo assim, podemos apresentar a equação que define o modelo VAR definida em [3]. Seja $Y_t = (y_{1t}, y_{2t}, \dots, y_{nt})$ um vetor de dimensão $(n \times 1)$ com os valores atuais de n séries temporais. Sendo assim, o modelo VAR com um lag-p de valores defasados é

$$Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \dots + \Pi_p Y_{t-p} + \epsilon_t, \text{ onde } t = 1, \dots, T \quad (2)$$

Note que esta equação é praticamente idêntica à equação 1, uma vez que ambas definem modelos auto regressivos. A diferença está no fato de Π_i não serem mais coeficientes escalares, mas sim matrizes de coeficientes de dimensão $(n \times n)$ e ϵ_t ser um vetor $(n \times 1)$ de ruídos brancos.

Embora este modelo possa ser usado para séries multivariáveis, é mais adequado, dentro de nosso escopo de modelagem de 10 séries temporais, enxergá-lo como o encapsulamento de n séries temporais univariáveis utilizando notação matricial. O que pode facilitar esta maneira distinta de enxergá-lo é o exemplo de reescrita da equação 2 do modelo para um exemplo de um VAR(2) de duas variáveis.

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \pi_{11}^1 & \pi_{12}^1 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} \pi_{11}^2 & \pi_{12}^2 \\ \pi_{21}^2 & \pi_{22}^2 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

Note que podemos ler cada linha das matrizes e vetores da equação cima como modelos AR(2) definidos de acordo com a equação 1.

Uma vez apresentado o modelo VAR(p) e a sua relação com o modelo AR(p) é natural que surja a seguinte pergunta: qual a vantagem de utilizarmos um modelo vetorial para a análise de 10 séries temporais se podemos realizá-las individualmente?

O modelo vetorial, dentro do escopo desta pesquisa, apresenta duas vantagens interessantes para a modelagem de nossas 10 séries temporais.

1. Ao encapsular as séries dentro de um vetor, o modelo enxerga um só bloco de informação e busca correlações e interdependências entre as séries. Caso essas relações existam, o modelo pode fornecer previsões mais precisas.

2. O modelo vetorial fornece ferramentas estatísticas para analisar a interconexão das séries temporais utilizada.

A primeira vantagem é obviamente interessante, uma vez que o grande objetivo é, justamente, obter modelos de previsão de séries temporais mais precisos. A segunda, por sua vez, também é importante uma vez que as 10 séries utilizadas, por serem de um mesmo nicho - todas são séries dos preços/retornos de ações de companhias aéreas - podem revelar análises de interconexão interessantes. A apresentação dessa ferramentas está descrita na seção 3.1.1.

Uma vez compreendida a equação que define o modelo VAR e as suas características possivelmente vantajosas, podemos definir a metodologia utilizada para a estimação de seus parâmetros, para a definição do lag-p e o seu método de previsão.

Com o intuito de estimar os parâmetros necessários para o modelo, utiliza-se o fato dele poder ser enxergado como o encapsulamento de n séries a serem modeladas via AR(p). Desse modo, é possível estimar os parâmetros da cada uma das equações separadamente sem perda de generalidade [3]. Sendo assim, é utilizado o método de mínimos quadrados para realizar esta estimativa de coeficientes.

Novamente, uma vez que o modelo VAR é um modelo auto regressivo assim como o AR, o lag-p, isto é, o número de valores passados das séries que serão utilizados para estimar o modelo e, consequentemente, prever os valores futuros, é estimado a partir da minimização de alguns critérios. Os critérios mais comuns a serem minimizados são Akaike (AIC), Schwarz-Bayesian (BIC) e Hannan-Quinn (HQ) [3]. Esses critérios encontram-se definidos abaixo.

$$\begin{aligned} \text{AIC}(p) &= \ln \left| \sum(p) \right| + \frac{2}{T}pn^2 \\ \text{BIC}(p) &= \ln \left| \sum(p) \right| + \frac{\ln T}{T}pn^2 \\ \text{HQ}(p) &= \ln \left| \sum(p) \right| + \frac{2 \ln \ln T}{T}pn^2 \end{aligned}$$

Cada um desses critérios possui características próprias. O AIC assintoticamente superestima a ordem do lag-p com probabilidade positiva, enquanto os critérios BIC e HQ tende a estimar de forma consistente a ordem do lag-p se a ordem verdadeira p for menor ou igual a uma ordem p-max pré estabelecida. Após a estimativa dos coeficientes, a fim de verificar a qualidade do modelo obtido analisa-se os resíduos (erros) tanto das séries quanto das ACF e PACF, que são funções de autocorrelação cuja apresentação foi feita anteriormente [2].

A etapa mais importante desta pesquisa é a **previsão**. Uma vez assumido que os coeficientes são conhecidos e que não existe nenhum termo exógeno à série, o método de previsão linear utilizado para a previsão de um passo a frente é:

$$Y_{T+1|T} = c + \Pi_1 Y_T + \dots + \Pi_p Y_{T-p+1} \quad (3)$$

Esse método baseia-se em todos os valores passados da série até o tempo $T - P + 1$, sendo T o índice do último valor conhecido da série. Sendo assim, imagine que nossa série possui entradas mensais e que o último valor conhecido é de Dezembro de 2020. Se nosso lag-p foi definido como 2, podemos rescrever a equação 3 de forma mais didática.

$$Y_{Jan2021|Dez2020} = c + \Pi_1 Y_{Dez2020} + \Pi_2 Y_{Nov2020}$$

O que essa equação nos diz é: o valor da série em Janeiro de 2021 (a ser previsto), conhecendo todos os valores até Dezembro de 2020, é função dos valores de Novembro de 2020 e Dezembro de 2020, uma vez que nosso lag-p é igual a 2.

A previsão de h-passos a frente pode ser obtida utilizando a regra da cadeia de previsão [3] como se segue:

$$Y_{T+h|T} = c + \Pi_1 Y_{T+h-1|T} + \dots + \Pi_p Y_{T+h-p|T} \quad (4)$$

onde, para $j \leq 0$, isto é, para valores conhecidos ($t \leq T$), $Y_{T+j|T} = Y_{T+j}$.

Se os coeficientes não são conhecidos, mas são frutos de estimativas tais quais as apresentadas anteriormente, a melhor maneira de prever o comportamento de Y_{T+h} é

$$\hat{Y}_{T+h|T} = c + \hat{\Pi}_1 \hat{Y}_{T+h-1|T} + \dots + \hat{\Pi}_p \hat{Y}_{T+h-p|T} \quad (5)$$

onde $\hat{\Pi}_i$ são as matrizes de coeficientes estimados. A diferença é apenas de notação, de certo modo. A existência do símbolo circunflexo denota o fato daquele parâmetro ter sido estimado e não conhecido a priori. O tamanho do horizonte de previsão, isto é, a quantidade de passos futuros a serem previstos, deve ser definida com cautela, visto que o erro das previsões tende a aumentar com o aumento do horizonte. Veremos que, para horizontes grandes, o valor da previsão parece convergir para um valor constante.

Por fim, com o intuito de verificar a qualidade da previsão, foi utilizado o RMSE, Root Mean Square Error, entre os valores originais das séries separados como grupo de teste e os valores previstos para esse intervalo separado. O RMSE é calculado como se segue:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^h (\hat{Y}_t - Y_t)^2}{h}} \quad (6)$$

onde h é tanto o número de passos a frente e a quantidade de valores comparados.

3.1.1 Análise de Interconexões

Após a rápida explicação teórica por detrás do modelo VAR, haja vista a sua enorme similaridade com o modelo AR já estudado com base em [1] e apresentado no relatório disponível em [8], podemos analisar o fundamento teórico de três ferramentas estatísticas disponíveis pelo modelo vetorial que permitem a análise e procura de possíveis interconexões ou interdependência entre as séries modeladas. São elas: Granger Causality, Impulse Response Function e Forecast Error Variance Decomposition (FEVD).

A **Causalidade de Granger** consiste em uma métrica que busca descobrir se uma das séries modelada pelo VAR é causa das demais séries modeladas. Em outras palavras, esta ferramenta busca responder a pergunta: será que os valores de cada uma das séries do modelo é uma consequência dos valores de alguma das outras séries? É importante ressaltar que essa análise se limita a estudar a causalidade em termos de previsão.

Para entender melhor essa análise, tome o exemplo de duas séries temporais arbitrárias: a série que mede a quantidade de chuva semanal no estado do Rio de Janeiro e a série que mede o nível de água semanal do rio Paraíba do Sul. É natural pensarmos que, se em uma dada semana a quantidade de chuva aumenta em relação a semana anterior, o nível de água do rio também aumentará. Essas duas séries, mais do que correlacionadas, estão causalmente relacionadas, ou seja, a evolução ou os valores de uma das séries (quantidade de chuvas) é influencia de maneira causal a evolução ou os valores da outra (nível de água no rio Paraíba do Sul). Se essas séries forem modeladas via VAR e submetidas ao teste

de causalidade de Granger, espera-se que ele chegue neste mesmo resultado obtido pelo senso comum.

A maneira de determinar se uma série granger-cause outra é analisar os coeficientes das matrizes Π_i da equação 2. Com o intuito de visualizar essa análise, é apresentado o exemplo para um modelo VAR de 2 variáveis. A série y_2 **falha** em causar a série y_1 , se todos os coeficientes π_{12}^i forem zero, ou seja, se as matrizes Π_i forem triangulares inferiores, como se segue para o caso de um VAR(2).

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \pi_{11}^1 & 0 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} \pi_{11}^2 & 0 \\ \pi_{21}^2 & \pi_{22}^2 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

A **função de resposta impulsional**, por sua vez, é uma análise que busca responder a seguinte pergunta: como as séries do modelo vão reagir a um choque, a uma mudança abrupta ou a um impulso gerado em uma das séries? Essa análise é interessante por trazer uma ferramenta capaz de prever o comportamento e a reação de séries temporais em relação às outras. Imagine, por exemplo, um cenário de plantação agrícola em um ambiente de pouca chuva anual cujas séries relevantes são a produção semanal de trigo e a quantidade semanal de chuva. Novamente, é de bom senso imaginar que uma semana de tempestades atípicas, que serão percebidas como um choque abrupto na série de quantidade de chuva, vai alterar o comportamento da série que mede a quantidade de trigo colhido na semana seguinte, por exemplo. Sendo assim, conhecer a resposta impulsional da série de quantidade semanal de trigo em função da série de quantidade semanal de chuva pode ser uma informação útil para prevenção de desestres.

A equação que melhor define a função de resposta impulsional é

$$\frac{\partial y_{i,t+s}}{\delta \eta_{j,t}} = \frac{\partial y_{i,t}}{\delta \eta_{j,t-s}} = \theta_{i,j}^s$$

onde $i, j = 1, \dots, n$ e $s > 0$. O gráfico de $\theta_{i,j}^s \times s$ é, justamente, a função de resposta impulsional. Embora o entendimento teórico não seja trivial, a interpretação visual dos resultados apresentada na seção 3.2.2 é simples.

A última das ferramentas de análise de interconexão estudadas é a **Decomposição de Variância de Erro de Previsão**, ou **FEVD**, em inglês. Essa análise, assim como apresentada em [3], responde à seguinte pergunta: quanto da variância do erro de previsão de $y_{i,T+h}$ foi causado por causa do choque estrutural η_j . O resultado desta análise é uma matriz de n colunas, uma para cada série modelada via VAR, e n^2 linhas, que são, justamente, os valores calculados pela decomposição. Quanto mais próximos de zero os valores de uma determinada coluna, menos os choques da série desta coluna interferiram na variância do erro da previsão da série analisada. Sendo assim, é de se esperar que, ao analisar uma série Y , por exemplo, na coluna da matriz referente a própria série Y , todos os valores sejam próximos do valor máximo, 1.

A equação que define a FEVD é

$$\text{FEVD}_{i,j}(h) = \frac{\sigma_{\eta_j}^2 \sum_{s=0}^{h-1} (\theta_{ij}^s)^2}{\sigma_{\eta_1}^2 \sum_{s=0}^{h-1} (\theta_{i1}^s)^2 + \dots + \sigma_{\eta_n}^2 \sum_{s=0}^{h-1} (\theta_{in}^s)^2}, \quad i, j = 1, \dots, n$$

Com isso, encerramos a análise teórica metodológica necessária para entender os resultados obtidos com o modelo VAR ao modelar 10 séries temporais de preços de ações de companhias aéreas.

3.2 Resultados VAR

3.2.1 Resultados Previsão

Assim como para a implementação dos modelos ARIMA já apresentados no relatório passado e dos modelos híbridos Wavelets com ARIMA, foi usada a linguagem R por ser uma linguagem popular e adequada para estudos econométricos. Os pacotes utilizados para ambos os modelos são ggplot2 para a criação de gráficos, Metrics para o cálculo do RMSE, tseries para tratamento de séries temporais e vars para a implementação do modelo VAR desde sua estimativa até previsão. Todos os pacotes estão disponíveis em [7].

Com o intuito de fornecer uma comparação justa, foram usadas para o modelo VAR as mesmas 10 séries de companhias aéreas já utilizadas para os modelos ARIMA. As companhias selecionadas são: IAG, Japan Airlines, China Southern, China Eastern, Latam, Delta, United, American, Lufthansa e AirFrance-KLM. Os dados coletados vão de 2010 a 2019. Essas companhias foram escolhidas por serem 10 das maiores do mundo e representarem diferentes regiões do globo, a fim de obter uma análise mais geral ao invés de uma focada em algum país ou continente.

Para não estender desnecessariamente este relatório, serão apresentadas as imagens de apenas 2 das 10 companhias. Todo o material produzido para esta pesquisa, desde o código em R até as imagens geradas, estão disponíveis em um repositório público [8]. Ao final desta seção estarão as tabelas com os erros comparados das 10 companhias aéreas em relação aos modelos ARIMA.

A metodologia apresentada na seção anterior foi fielmente seguida ao implementar o código em R. Por ser um processo muito similar àquele realizado para os modelos ARIMA, aqui será feita uma breve recapitulação apenas. Para explicações mais detalhadas daquilo que não for abordado com profundidade nesta seção, ver o relatório do ano anterior disponível em [8].

O primeiro passo após a importação dos dados presentes em arquivos csv é selecionar apenas os valores de interesse, que são os preços fechados ajustados das ações das companhias aéreas. Com esses valores em mãos, podemos inseri-los em vetores numéricos para procurar e substituir valores faltantes, que são comuns em bases de dados utilizadas para obter as séries. Feita essa procura, as séries podem ser visualizadas.

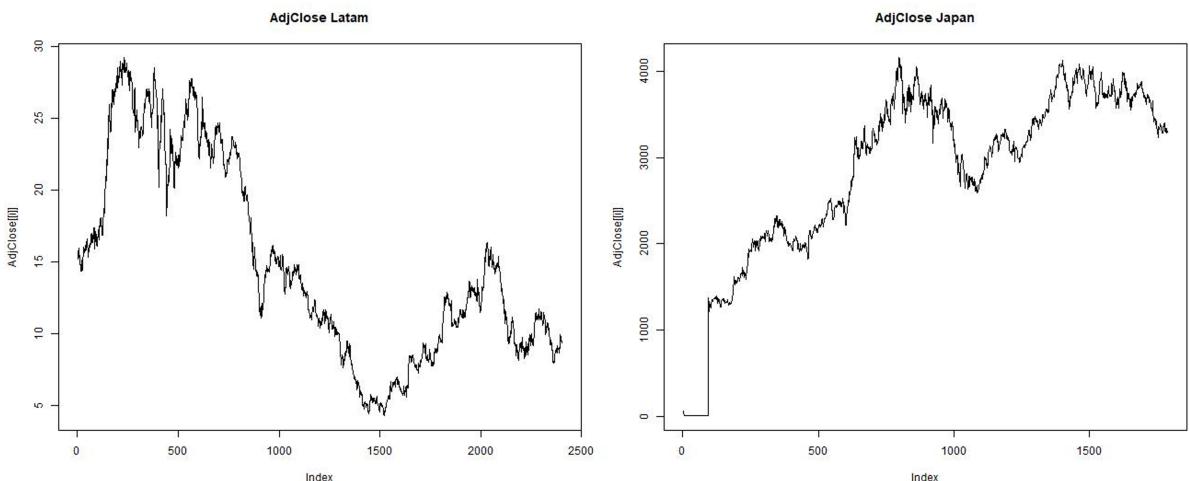


Figura 1: Adjusted Closed Price para Latam e Japan Airlines

Neste momento poderia ser feito um teste de estacionariedade, como o Augmented Dickey-Fuller Teste, assim como foi feito para os modelos ARIMA e será feito para os modelos híbridos. No entanto, optou-se por não fazê-lo e obter de uma vez os retornos-log desses preços, haja vista que esta é a variável de interesse da pesquisa. Para tal, basta calcular a diferença dos logs das entradas das séries. Um cuidado importante a ser tomado é, após obter o retorno, submeter as séries a um novo tratamento de busca por valores faltantes ou inválidos, uma vez que as operações realizadas para obter os retornos podem inserir tais valores nas séries. Para isso, utiliza-se a função `tsclean` do pacote `tseries`. As imagens abaixo exibem os retornos-log para as mesmas duas séries já apresentadas.

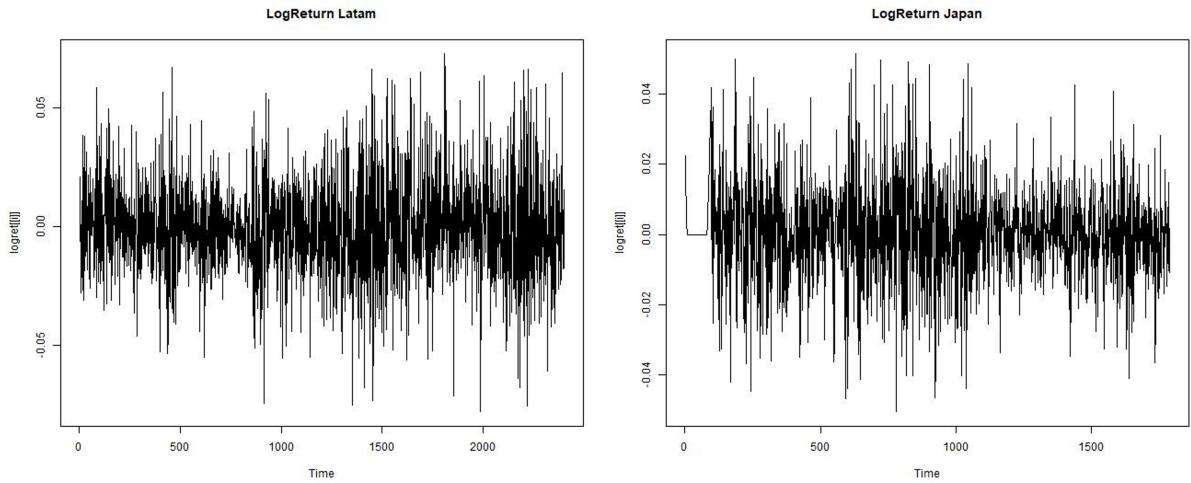


Figura 2: Log-return para Latam e Japan Airlines

Uma etapa que não existia para os modelos ARIMA mas que deve ser realizada para o modelo VAR é reduzir a quantidade de entradas das séries de modo a garantir que todas possuam o mesmo tamanho, igual ao tamanho da menor série utilizada. Essa alteração, embora signifique a perda de informações de algumas séries, é necessária pois o modelo VAR supõe que todas as séries serão encapsuladas em um vetor e, para isso ser possível, todas devem possuir o mesmo número de entradas. Sendo assim, ao invés de criar novos valores para séries menores, opta-se por reduzir, ou seja, deletar entradas, de séries maiores.

Feita esta redução, as séries podem ser divididas em grupos de treino e teste. O grupo de treino é aquele que será usado para estimar os coeficientes do modelo VAR e, com os coeficientes estimados, realizar a previsão. O grupo de teste, por sua vez, sempre é guardado para comparar as previsões realizadas com os valores originais jamais vistos pelo modelo. Dito isso, foram realizadas duas separações. Primeiro foram separados 90% das séries para treino e 10% para teste, em prol de avaliar a previsão para grandes horizontes de tempo, visto que 10% das séries corresponde a 178 entradas. Em seguida, os grupos de teste foram formados com apenas os 20 últimos valores das séries, com o intuito de avaliar a previsão para horizontes mais curtos.

Para estimar os valores do lag, via ACF e PACF, diferentemente do que foi feito anteriormente, optou-se por utilizar a função `VARselect` do pacote `vars` [7] para definir o valor do lag automaticamente, ao invés da análise individual e visual utilizada para os modelos ARIMA. A única necessidade foi definir um lag máximo de 15, escolhido por ser um valor pouco maior que o maior lag definido para os modelos ARIMA implementados

anteriormente. Com o valor do lag definido como 9, foi utilizada a função VAR para estimar o modelo. O único parâmetro necessário é um booleano para sazonalidade que foi definido como falso. O motivo para ele ser falso é o fato de nenhuma das séries utilizadas apresentar perfil sazonal, assim como foi mostrado anteriormente. Novamente, caso seja do interesse do leitor acessar esses resultados passados, todo o material da pesquisa encontra-se em [8].

As imagens abaixo exibem as estimativas para as duas mesmas companhias.

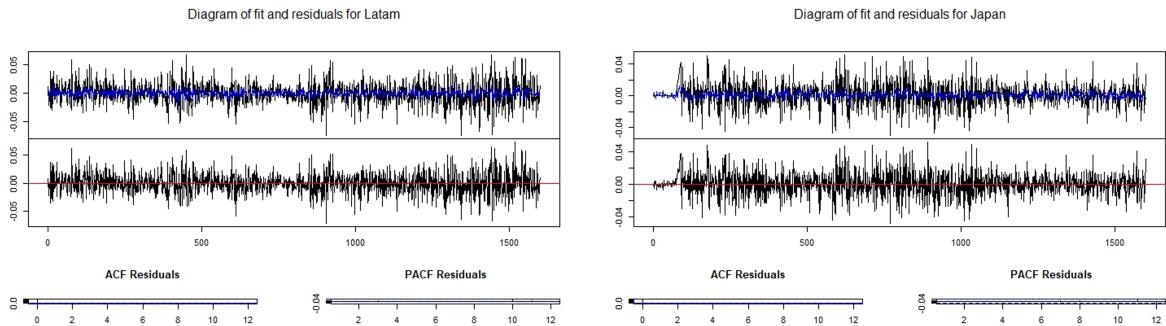


Figura 3: Estimativa e resíduos para o modelo VAR

Como pode ser percebido, os resíduos, assim como desejado, parecem ser independentes, visto que as funções de autocorrelação e autocorrelação parcial estão praticamente zeradas, o que torna os resíduos próximos de um ruído branco.

Com isso, estamos aptos a, enfim, prever valores futuros das séries via função predict do mesmo pacote. Os resultados dessas previsões podem ser observados nas figuras abaixo, que exibem, além do valor previsto, o valor original das séries separados como grupo de teste. Como comentado, foram realizadas estimativas e previsões para duas divisões distintas.

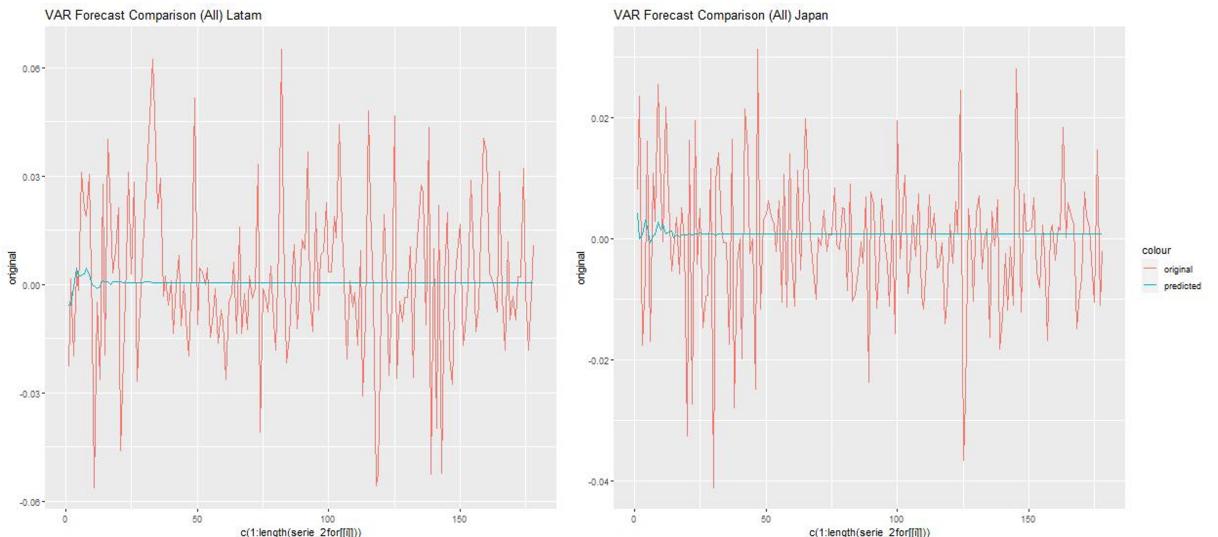


Figura 4: Previsão do modelo VAR para os 10% finais das série de Latam e Japan Airlines

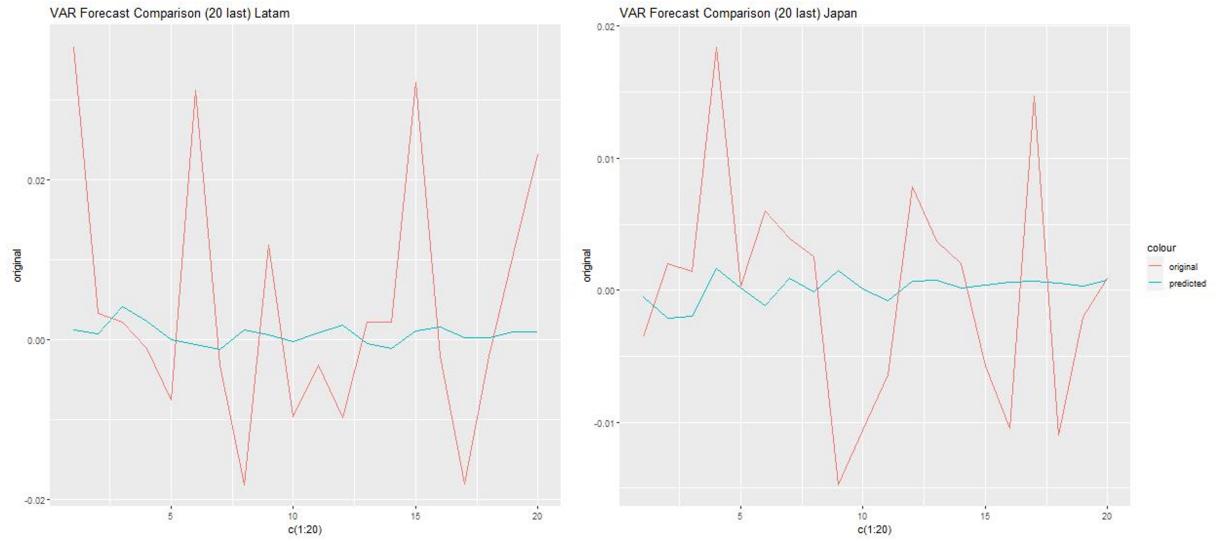


Figura 5: Previsão do modelo VAR para os 20 valores finais das séries de Latam e Japan Airlines

Por fim, foram calculados os erros RMSE, como definidos na equação 6 e construídas as seguintes tabelas que comparam esses valores para os modelos ARIMA e modelo VAR, tanto para a divisão 90/10 quanto para as 20 entradas finais das séries.

Erros das previsões para os 10% finais das séries		
Série	RMSE ARIMA	RMSE VAR
American	0.02271	0.01974
United	0.01319	0.02186
Delta	0.01307	0.01764
Latam	0.02030	0.02167
ChinaSouth	0.02005	0.01925
ChinaEast	0.02116	0.01864
Japan	0.01166	0.01163
IAG	0.01997	0.02219
Lufthansa	0.01680	0.02041
AirFrance-KLM	0.01983	0.02078

Tabela 1: RMSE para divisão 90/10

Erros das previsões para os 20 valores finais das séries		
Série	RMSE ARIMA	RMSE VAR
American	0.01924	0.02265
United	0.01090	0.01780
Delta	0.01166	0.01782
Latam	0.02063	0.01584
ChinaSouth	0.01589	0.02144
ChinaEast	0.01773	0.02041
Japan	0.00824	0.00823
IAG	0.01424	0.01459
Lufthansa	0.01061	0.01967
AirFrance-KLM	0.01510	0.01931

Tabela 2: RMSE para 20 valores finais

3.2.2 Resultados Interconexões

Dados os resultados obtidos na seção 3.2.1, podemos partir para os resultados das análises de interconexão discutidas na seção 3.1.1. Como comentado, uma das possíveis vantagens de se utilizar um modelo autoregressivo vetorial ao invés de vários modelos AR simples é o fato de podermos buscar interconexões ou relações de dependência entre as séries modeladas.

Ao executar a função causality do pacote vars, que implementa o algoritmo de Granger Causality apresentado na seção 3.1.1, são obtidas respostas idênticas para todas as séries. A resposta do teste de causalidade para as séries das companhias Latam e Japan Airlines, por exemplo, estão exibidas abaixo.

<pre>\$Granger</pre> <pre>Granger causality H0: Latam do not Granger-cause IAG Japan ChinaSouth ChinaEast Delta United American Lufthansa AirFranceKLM</pre> <pre>data: VAR object VAR_est F-Test = 0.78586, df1 = 81, df2 = 15090, p-value = 0.9219</pre> <pre>\$Instant</pre> <pre>H0: No instantaneous causality between: Latam and IAG Japan ChinaSouth ChinaEast Delta United American Lufthansa AirFranceKLM</pre> <pre>data: VAR object VAR_est Chi-squared = 140.49, df = 9, p-value < 2.2e-16</pre>	<pre>\$Granger</pre> <pre>Granger causality H0: Japan do not Granger-cause IAG ChinaSouth ChinaEast Latam Delta United American Lufthansa AirFranceKLM</pre> <pre>data: VAR object VAR_est F-Test = 0.69749, df1 = 81, df2 = 15090, p-value = 0.9823</pre> <pre>\$Instant</pre> <pre>H0: No instantaneous causality between: Japan and IAG ChinaSouth ChinaEast Latam Delta United American Lufthansa AirFranceKLM</pre> <pre>data: VAR object VAR_est Chi-squared = 12.535, df = 9, p-value = 0.1848</pre>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figura 6: Granger Causality para Latam e Japan Airlines

Dado o output dessa função, a interpretação do resultado é clara: nenhuma das séries, embora todas serem de um mesmo nicho de mercado, apresenta relação de causalidade com as demais.

Partindo para a análise da função de resposta impulsional IRF definida na seção 3.1.1, foi utilizada uma função do mesmo pacote com o nome de irf cujo retorno pode ser

expressado de maneira visual como a imagem abaixo para a série da Japan Airlines.

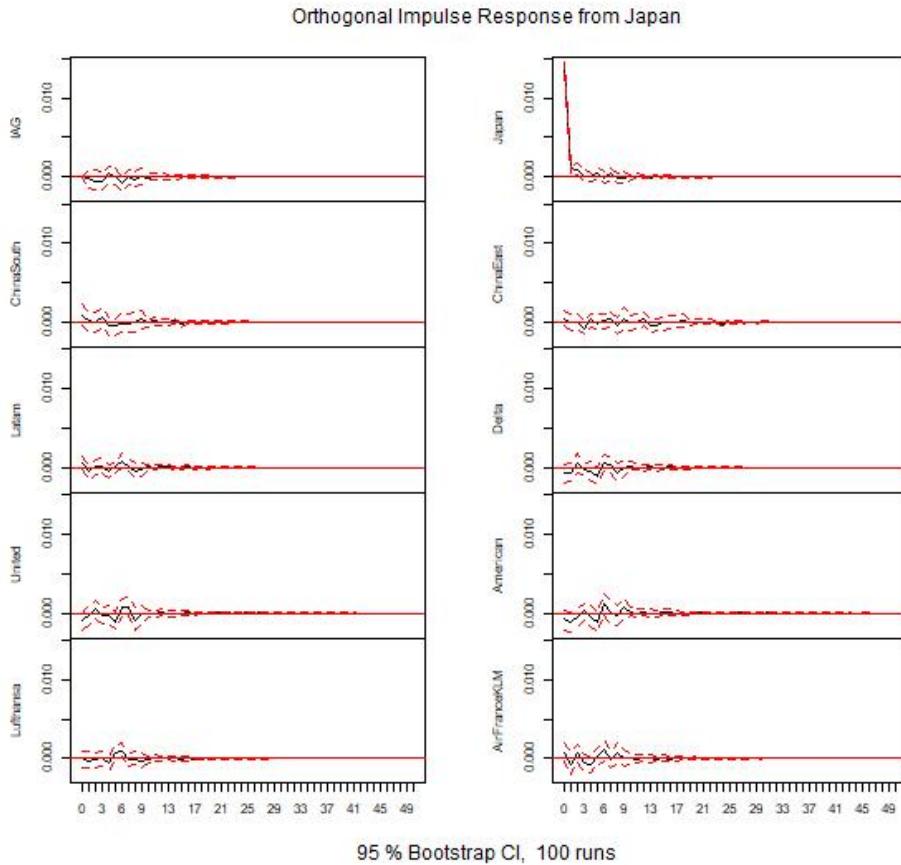


Figura 7: Impulse Response Function para a série da Japan Airlines

Para entender este resultado, que também é muito similar entre todas as séries temporais das companhias aéreas, vale lembrar que a resposta impulsional procura medir o comportamento das demais séries quando uma delas sofre um choque abrupto em seus valores. Note, portante, que esse choque na figura 7 está justamente na companhia Japan. Além disso, note que as demais companhias apresentam algumas oscilações em suas evoluções naturais mas nenhuma com amplitude significativa e todas com curta duração, visto que as oscilações praticamente desaparecem após 15 passos. Desse modo, podemos interpretar este resultado como uma garantia de que, além de não estarem causalmente relacionadas, as séries também possuem grau de independência relevante, uma vez que não respondem aos choques umas das outras.

Por fim, a Decomposição de Vriância do Erro de Previsão foi computada por meio da função fevd também do pacote vars [7]. Novamente, para facilitar a interpretação do resultado, vale lembrar que essa análise busca verificar o grau de importância dos choques de uma série na hora de prever o comportamento de outra. A FEVD constrói uma matriz cujas colunas são as séries e os valores de casa linha explicitam esse grau de importância para a série corrente. A figura 8 exibe parte da matriz para o caso da companhia japonesa.

\$Japan	IAG	Japan	ChinaSouth	ChinaEast	Latam	Delta	United	American	Lufthansa	AirFranceKLM
[1,]	0.001752217	0.9982478	0.000000000	0.000000000	0.000000000	0.000000000	0.000000e+00	0.000000000	0.000000000	0.000000000
[2,]	0.002696987	0.9896156	0.001291409	0.0001687736	0.0009321967	0.002710594	0.0002898222	1.121502e-06	0.003256548	0.0001991737
[3,]	0.004255036	0.9847144	0.00487208332	0.0001914268	0.0012395753	0.003652964	0.0005690053	1.014825e-03	0.003439729	0.0002021589
[4,]	0.004226050	0.9777375	0.0010580227	0.0033243788	0.0015135011	0.003630845	0.0007381799	3.210777e-03	0.004089070	0.0004717205
[5,]	0.004397701	0.9706490	0.0037001952	0.0033390512	0.0018409966	0.003717152	0.0028226345	4.392617e-03	0.004059386	0.0010812210
[6,]	0.004515128	0.9672089	0.0039882382	0.0033487888	0.0027580033	0.004645151	0.0028672026	5.546828e-03	0.004042769	0.0010789693
[7,]	0.004974882	0.9628030	0.0040887630	0.0048282712	0.0028934987	0.004760841	0.0029549598	6.687002e-03	0.004931437	0.0010773191
[8,]	0.005384026	0.9589769	0.0042584925	0.0051763849	0.0028890677	0.004776114	0.0033346985	7.562207e-03	0.006162025	0.0015601115
[9,]	0.005677425	0.9567050	0.0045010428	0.0051647106	0.0029939645	0.005196589	0.0039067709	7.632703e-03	0.006215211	0.0020066237
[10,]	0.005876566	0.9518601	0.0073306011	0.0052346894	0.0029793606	0.005199641	0.0040600051	7.615700e-03	0.006202928	0.0036403908
[11,]	0.005925844	0.9508959	0.0077596588	0.0052484691	0.0029798470	0.005198266	0.0040568051	7.615202e-03	0.006413620	0.0039063425
[12,]	0.005922017	0.9502562	0.0081212475	0.0052452937	0.0029964258	0.005229426	0.0041156246	7.650942e-03	0.006540083	0.0039227667
[13,]	0.005945804	0.9476620	0.0103412569	0.0052363948	0.0030810882	0.005217338	0.0041774171	7.734631e-03	0.006689829	0.0039142475
[14,]	0.005953003	0.9474459	0.0103386226	0.0052402684	0.0031765732	0.005216861	0.0042394253	7.748148e-03	0.006705973	0.0039352361
[15,]	0.005989640	0.9471455	0.0104325880	0.0052492070	0.0031839130	0.005216926	0.0042442072	7.825806e-03	0.006746342	0.0039658523
[16,]	0.005992024	0.9457762	0.0117761515	0.0052658092	0.0031793375	0.005236303	0.0042416239	7.815845e-03	0.006750851	0.0039658793
[17,]	0.006005900	0.9455497	0.0119497023	0.0052699950	0.0031786552	0.005244946	0.0042561744	7.821449e-03	0.006749549	0.0039739483
[18,]	0.006034902	0.9454951	0.0119492898	0.0052754578	0.0031791767	0.005245810	0.0042566475	7.822978e-03	0.006762364	0.0039782988
[19,]	0.006067991	0.9453424	0.0120441738	0.0052748643	0.0031823219	0.005249153	0.0042678081	7.827317e-03	0.006763506	0.0039805109
[20,]	0.006077425	0.9452594	0.0120581791	0.0052884595	0.0031837464	0.005251301	0.0042676145	7.830472e-03	0.006782155	0.0040012139
[21,]	0.006086714	0.9452276	0.0120614738	0.0052960578	0.0031839476	0.005251720	0.0042744425	7.834204e-03	0.006782411	0.0040014087
[22,]	0.006087518	0.9451567	0.0120921995	0.0052966274	0.0031847128	0.005256958	0.0042791888	7.834955e-03	0.006805539	0.0040056007
[23,]	0.006089711	0.9451361	0.0120935241	0.0052980167	0.0031847122	0.005257985	0.0042849608	7.834959e-03	0.006812050	0.0040079635
[24,]	0.006089633	0.9451139	0.0121019681	0.0053002936	0.0031867763	0.005265805	0.0042859713	7.834775e-03	0.006812317	0.0040085407
[25,]	0.006091329	0.9450879	0.0121179405	0.0053001888	0.0031870492	0.005265936	0.0042869575	7.838526e-03	0.006815683	0.0040085273

Figura 8: Matriz resultante da FEVD para Japan Airlines

Para compreender a matriz da figura 8, note que a segunda coluna, que se refere à companhia Japan, está com todos os valores muito próximos de 1. Isso, assim como explicado na seção 3.1.1, é óbvio pois os choque em uma série X devem ser os principais contribuintes para os erros na própria série X. Somado a isso, note que todas as demais entradas para todas as demais colunas são próximas de 0 o que, novamente, faz sentido, haja vista a figura 8 que exibe as respostas impulsionais para o caso japonês. Perceba, por fim, que, embora ainda muito pequenos, os valores da terceira coluna, referentes à companhia China Southern são os maiores quando comparados com todas as demais e ao olhar para a resposta impulsional desta companhia quando a Japan sofre um choque também notamos uma amplitude um pouco maior que das demais logo após o choque. Dito isso, podemos interpretar esse resultado de forma a, mais uma vez, fortalecer a tese de que as 10 séries utilizadas são razoavelmente independentes.

As conclusões mais detalhadas a cerca de todos os resultados apresentados encontram-se na seção 4.

3.3 Modelo Híbrido: Wavelets + ARIMA

Com o intuito de entender a Teoria de Wavelets, é importante trazer uma brevíssima contextualização histórica que explique não só a origem dessa teoria, mas também suas vantagens em relação às teorias anteriores.

A análise de sinais no domínio da frequência tem como grande expoente o desenvolvimento da Análise de Fourier. Inicialmente, J. Fourier constatou que qualquer função periódica de período 2π pode ser representada como uma soma de senos e cossenos, gerando a Série de Fourier da função [6]:

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

Em seguida, esta teoria foi estendida para englobar funções não periódicas com o desenvolvimento da Transformada de Fourier. Uma transformada pode ser entendida como uma "função de funções", isto é, uma relação que leva uma função de um domínio

X , como o domínio do tempo, por exemplo, para um contradomínio Y , como o domínio da frequência, de modo que um sinal qualquer, medido em função do tempo, possa ser analisado sob a perspectiva da frequência. A equação que define a Transformada de Fourier, $F(\xi)$, de um sinal $f(t)$ é

$$F(\xi) = \int_{-\infty}^{\infty} f(t)e^{2\pi i \xi t} dt$$

A partir dessa transformada, podemos obter informações referentes à frequência de um sinal que antes apenas fornecia informações temporais. No entanto, não podemos obter ambos os tipos de informação ao mesmo tempo. Ao obter informações acerca da frequência, não sabemos em que momento do tempo ela foi obtida. Esse é um dos pontos em que a Teoria de Wavelets fornecerá vantagens.

A etapa seguinte, e final, da evolução histórica necessária para apresentar o conceito de Wavelets é a Short Time Fourier Transform, ou, como será denominada, STFT. A ideia por detrás dessa transformada é, ao invés de olhar para o sinal como um todo, dividi-lo em segmentos pequenos o suficiente para que sejam estacionários [5]. Para tal, é escolhida uma função de janela de suporte finito e constante que, inicialmente, é colocada no início do sinal e é transladada por toda a sua duração. Essa é uma das ideias utilizadas para o desenvolvimento da teoria de Wavelets. A equação que define a STFT para um sinal $x(t)$ e janela ω é

$$\text{STFT}_x^{(\omega)}(t, f) = \int_t [x(t)\omega^*(t - t')]e^{2\pi ift} dt$$

Uma vez apresentados de forma breve, haja vista que a Análise de Fourier não é o foco desta pesquisa, podemos, enfim, enumerar alguns desafios que essa teoria não consegue superar e, em seguida, introduzir a teoria de Transformada Wavelet Contínua, ou CWT, que é capaz de suprir essas necessidades.

1. Falta de localização temporal: não é possível saber em qual momento do tempo cada frequência existiu
2. Janela de tamanho fixo: cria uma limitação da análise

Dito isso, a Transformada Wavelet Contínua possui dois parâmetros em sua definição, justamente para superar os desafios enumerados. Um parâmetro é para a captura da informação da frequência, chamado s (scale) que é definido como o inverso da frequência f , enquanto o outro, τ (translation) é responsável com transladar a Wavelet-mãe pelo sinal. A equação que define a CWT é

$$\text{CWT}_x^\psi(\tau, s) = \Psi_x^\psi(\tau, s) = \frac{1}{\sqrt{s}} \int x(t)\psi^*\left(\frac{t - \tau}{s}\right) dt \quad (7)$$

onde, s e τ são os parâmetros de escala e translação, $x(t)$ é o sinal original no domínio do tempo, $\psi^*(t)$ é o complexo conjugado da Wavelet mãe.

Existem diversas funções Wavelet que podem ser utilizadas como Wavelet mãe. É esta função que dará o perfil da curva a ser multiplicada dentro da integral pela função do sinal de entrada. Alguns dos exemplos mais importantes da bibliografia de Wavelets são as famílias das Wavelets de Daubechies, de Coiflet, de Haar e de Symmlet [6], exibidas na figura abaixo.

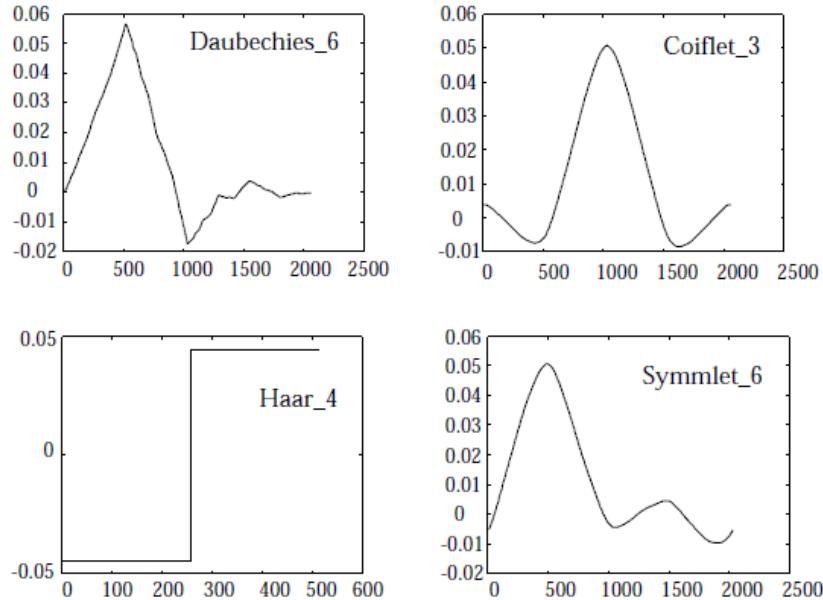


Figura 9: Famílias importantes de Wavelets

Para fins didáticos, vale enumerar os passos utilizados ao computar uma CWT para um sinal qualquer $x(t)$.

1. Posiciona-se a Wavelet mãe ψ no início do sinal $x(t)$, ou seja $\tau = 0$. Define-se $s = 1$, que é a versão mais comprimida da Wavelet.
2. A Wavelet na escala 1 e $\tau = 0$ é multiplicada pelo sinal e, em seguida, integrada e normalizada.
3. Translada-se a Wavelet alterando o valor de τ e repete-se o passo anterior até alcançar ela alcançar o final do sinal.
4. Incrementa-se o valor de s , dilatando o sinal e repetem-se os passos anteriores para todos os valores de s desejados.
5. A computação para cada valor de s preenche uma linha do plano tempo X frequência (apresentado a seguir)
6. CWT é obtida quando todos os s são calculados.

O resultado final desse procedimento, ou seja, a própria CWT, pode ser visualizado por meio de um plano tempo X frequência como o da figura abaixo.

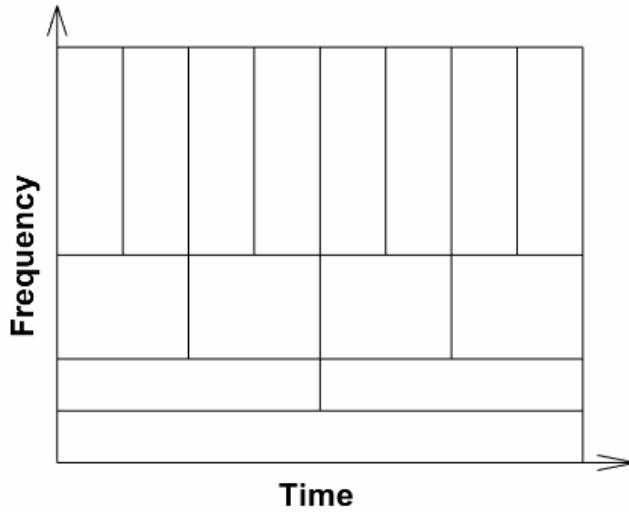


Figura 10: Plano Tempo X Frequênciа

Uma observação importante que pode ser feita com base na análise deste plano e do conhecimento do Princípio de Incerteza de Heisenberg [5] é a de que, quanto melhor a definição da frequência, pior é a definição do tempo, e vice-versa. Isso quer dizer que, quanto mais para baixo no plano, melhor será a resolução no domínio da frequência e pior no domínio do tempo (note que a altura do retângulo é pequena, quando comparada com a largura). De modo análogo, quanto mais para cima no plano, pior a resolução da frequência e melhor no tempo (note que, agora, a altura dos retângulos é maior que a largura).

A figura abaixo ilustra uma parte do algoritmo descrito acima, onde a curva azul é uma representação da Wavelet mãe e a curva amarela é o sinal original. Note que, da esquerda para a direita, a curva azul é transladada sempre com uma mesma largura (s) e, só após passar por todo o sinal, tem sua largura alterada.

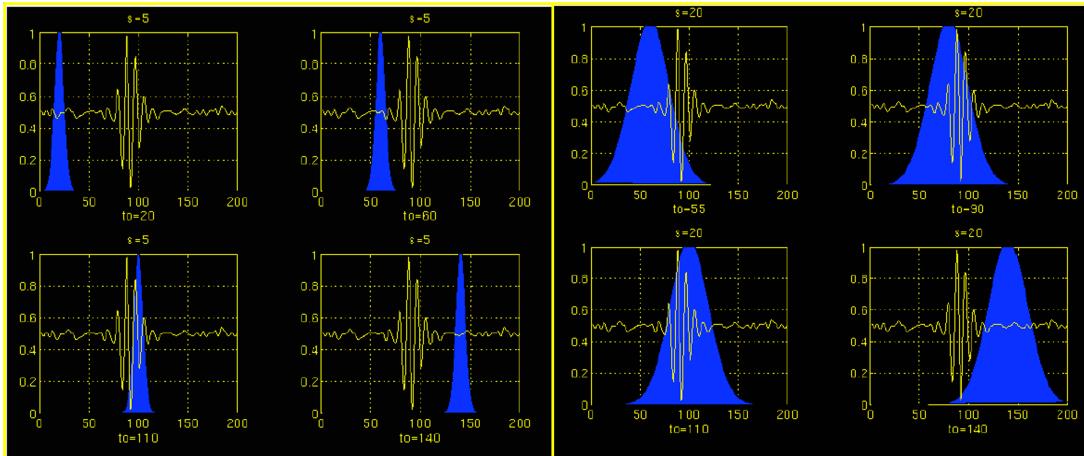


Figura 11: Ilustração do algoritmo de obtenção da CWT

Embora a transformada contínua seja de extrema importância para diversas áreas do conhecimento e tenha sido a primeira a ser desenvolvida, é natural que, com o objetivo

de implementar rotinas computacionais para a previsão de séries temporais que utilizam a teoria de Wavelets, seja apresentada a Transformada Wavelet Discreta, ou DWT. É esta versão da transformada implementada pelos algoritmos computacionais, visto que o mundo digital é intrinsecamente discreto.

A DWT, diferentemente da sua irmã contínua, não será apresentada como uma equação, mas sim como um algoritmo iterativo de filtros. De certo modo, essa abordagem é mais trivial de ser compreendida do que uma equação matemática pura, uma vez conhecido o conceito de filtro. Dito isso, podemos definir um filtro, de maneira superficial, mas suficiente, como um operador que divide o sinal de entrada em frequências maiores ou menores que uma frequência de corte. Um filtro passa-baixas é aquele que filtra, ou que impede a passagem, das frequências mais altas que a frequência de corte e deixa passar as mais baixas. Um filtro passa altas, por sua vez, faz exatamente o contrário.

Para facilitar o entendimento da implementação do algoritmo da DWT, foi ilustrado um exemplo de uma série $x[n]$ de 512 observações que será a entrada do algoritmo. Na figura, Filtro PB significa passa-baixas, enquanto Filtro PA, passa altas.

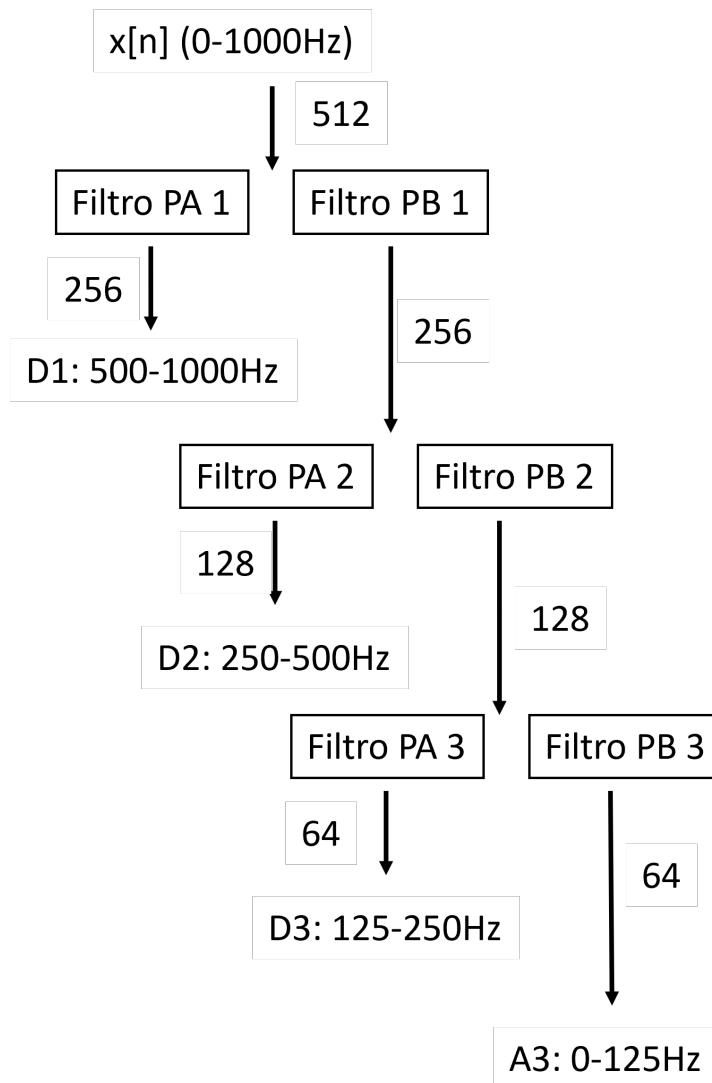


Figura 12: Implementação DWT via filtros

O algoritmo da DWT, para o exemplo da figura 12, funciona da seguinte forma. Uma

série de 512 entradas passa por dois filtros: um passa-baixas e outro passa-altas, cuja frequência de corte é 500Hz. A saída de cada filtro vai ter exatamente a metade do tamanho da série de entrada, logo 256 valores. A saída do filtro passa-altas, cujas frequências vão de 500 a 1000 Hz, são os coeficientes da decomposição nesse nível. A saída do passa-baixas, também com 256 valores mas com frequências entre 0 e 500 Hz será a entrada dos filtros do próximo nível de decomposição. Essa iteração deve ser feita até atingir o nível de decomposição desejado ou até a série acabar, visto que, a cada iteração, perde-se metade do tamanho da série.

Ao final do algoritmo, os coeficientes que formam a DWT são as saídas dos filtros passa-altas em cada nível de decomposição e a saída do passa-baixas do último nível.

Note que, a cada iteração, perde-se metade da informação que existia no nível anterior. Isso é um problema significativo, uma vez que esses coeficientes serão tratados como novas séries a serem modelados via ARIMA. Sendo assim, em níveis mais profundos da decomposição, responsáveis por fornecer informação sobre as baixas freqüências, teremos poucas entradas para a modelagem ARIMA, o que pode gerar deficiências na estimativa do modelo e, consequentemente, na qualidade da previsão. A fim de contornar esse problema, existe uma outra transformada Wavelet denominada Maximal Overlap Discrete Wavelet Transform, ou MODWT, que impede a perda de metade da informação a cada nível da decomposição, ou seja, garante que todas as saídas dos filtros possuam uma quantidade constante de entradas [4]. Esta é a transformada utilizada pelo pacote de R que foi utilizado para a obtenção dos resultados a serem apresentados em 3.4.

Uma vez apresentados os fundamentos teóricos da teoria de Wavelets e conhecida a modelagem ARIMA, estudada na primeira parte desta pesquisa cujo relatório está disponível em [8], pode ser definido o algoritmo, em alto nível, do modelo híbrido de previsão que combina ARIMA com Wavelets. A imagem a seguir ilustra esse algoritmo, inspirado por [4].

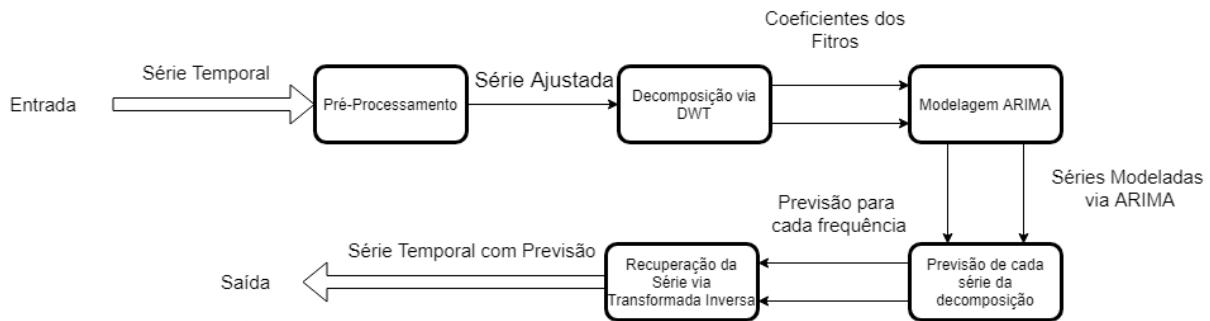


Figura 13: Diagrama do algoritmo do modelo híbrido ARIMA + Wavelets

A imagem 13 mostra que o algoritmo funciona da seguinte maneira. Primeiramente, a série bruta é pré-processada, o que engloba análise de estacionariedade (Augmented Dickey-Fuller Test) [1], completar valores faltantes e obter o log-return a partir da série original. Uma vez ajustada, a série será decomposta via DWT, ou, melhor dizendo, MODWT. Com o resultado da composição em mãos, cada uma das novas séries obtidas em freqüências distintas será modelada via ARIMA e, em seguida, prevista. Por fim, a série original é recuperada via DWT inversa.

O grande objetivo desta pesquisa é verificar se, ao prever as séries decompostas e filtradas em freqüências específicas, será obtida uma maior precisão. Ao realizar a previsão via ARIMA para cada uma das decomposições, espera-se que o algoritmo seja capaz de

perceber padrões que, na série original, eram imperceptíveis. Por exemplo, uma série pode exibir um perfil oscilatório bem comportado ao longo de anos, ou seja, para frequências menores, e comportamentos semanais, frequências maiores, também muito bem definidos. Embora seja difícil, muitas vezes, para o modelo ARIMA perceber essas diferentes características de uma série, espera-se que, ao decompor os dados originais em séries de frequências distintas, possamos ser capaz de fazê-lo.

A fim de comparar o resultado do algoritmo, foi utilizado o RMSE (Root Mean Square Error) entre os valores previstos e os valores originais separados em um grupo de teste. A definição de RMSE, já apresentada na equação 6, encontra-se novamente abaixo.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^h (\hat{Y}_t - Y_t)^2}{h}}$$

3.4 Resultados Wavelets + ARIMA

Com o intuito de implementar o algoritmo de modelagem e previsão do modelo híbrido apresentado na seção anterior, foi utilizada, novamente, a linguagem R pelos mesmos motivos já enumerados. Durante a pesquisa, foi descoberto um pacote dedicado a combinar Wavelets com o modelo ARIMA denominado WaveletARIMA e disponível em [7]. Este pacote fornece apenas duas funções: WaveletFitting e WaveletFittingarma. A segunda função foi aquela utilizada nesta pesquisa e ela implementa o algoritmo da MODWT para a decomposição Wavelet.

Essa função recebe como parâmetros a série temporal, o nível de decomposição, o horizonte de previsão e o lag máximo para as ordens AR e MA do modelo ARIMA (ver [1] ou relatório anterior em [8]). Como retorno, ela devolve um objeto de classe WaveletFittingarma que é, basicamente, uma lista de listas com todos os coeficientes estimados e valores futuros previstos, além de algumas outras métricas menos importantes para esta pesquisa. Note, portanto, que esta função é responsável todo o processo de decomposição via MODWT e utilizando o filtro Haar, estimativa do modelo ARIMA e previsão, restando apenas o pré processamento e o cálculo do erro da previsão.

Dito isso, o pré-processamento foi feito de forma idêntica ao já apresentado para o modelo VAR e também realizado para o modelo ARIMA da primeira etapa desta pesquisa. De mesmo modo, foram utilizadas exatamente as mesmas séries de todo o restante da pesquisa a fim de obter uma comparação justa dos diferentes modelos. Sendo assim, será feita apenas uma recapitulação breve dos passos do pré-processamento.

Uma vez importada a série e selecionado apenas o preço ajustado fechado AdjClosed do arquivo csv, em todas as séries foi realizada uma busca e substituição de valores faltantes. As imagens abaixo exibem duas das séries temporais utilizadas.

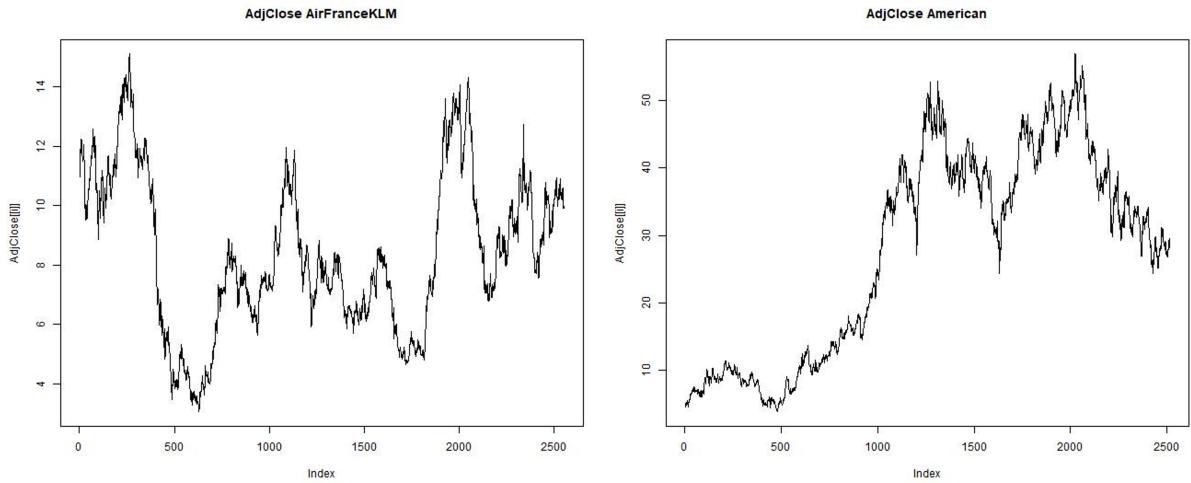


Figura 14: Preço fechado ajustado para AirFrance-KLM e American Airlines

Em seguida, foi realizado um teste de estacionariedade, o Augmented Dickey-Fuller Test. As imagens acima já levantam a suspeita de que ambas as séries são não-estacionárias, principalmente a série da American Airlines. Para não dependermos apenas do aspecto visual, o resultado do ADF Test de ambas encontra-se abaixo.

```
[1] "AirFranceKLM"
[1] "American"

Augmented Dickey-Fuller Test                               Augmented Dickey-Fuller Test
data: series[[i]]                                         data: series[[i]]
Dickey-Fuller = -2.4936, Lag order = 13, p-value = 0.3694  Dickey-Fuller = -1.1553, Lag order = 13, p-value = 0.9137
alternative hypothesis: stationary                         alternative hypothesis: stationary
```

Figura 15: Resultado do ADF Test para AirFrance-KLM e American Airlines

O valor que nos interessa é o p-value, que mede a probabilidade da série possuir uma raiz unitária e, consequentemente, ser não estacionária. Percebe-se que a série da American, como suspeitado pela análise visual, é menos estacionária que a da AirFrance-KLM. Uma vez feita essa análise, o próximo passo é obter o log-return dessas séries, uma vez que não estamos interessados no preço da ação em si, mas no seu retorno. É importante, após obter o retorno, executar novamente uma rotina de limpeza, realizada via função tsClean, para procurar e substituir valores faltantes ou inválidos.

As imagens abaixo exibem o retorno-log das mesmas duas séries.

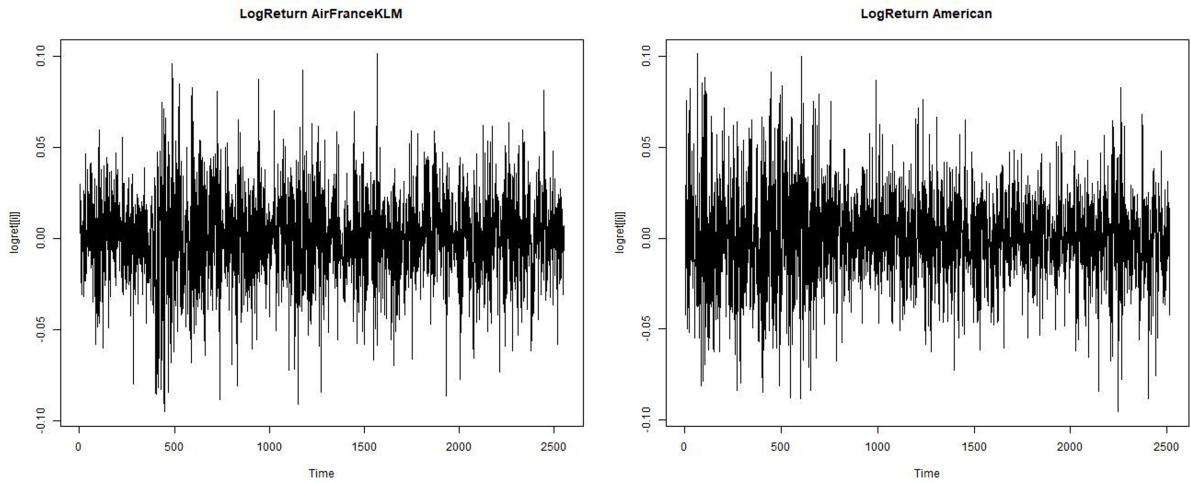


Figura 16: Log-return das séries AirFrance-KLM e American Airlines

Como essas são, de fato, as séries que nos interessam, foi feito novamente um teste de estacionariedade para verificar se a série precisa ou não ser diferenciada antes de ser decomposta via Wavelets e modelada via ARIMA [1]. Ambas as séries de retorno possuem aparência de estacionárias, o que é confirmado pelo ADF-Test exibido abaixo, onde podemos ver que ambos os p-values são menores de 0.1, concluindo-se que elas são estacionárias e não precisam de diferenciação.

<pre>[1] "AirFranceKLM" Augmented Dickey-Fuller Test data: logret[[i]] Dickey-Fuller = -12.639, Lag order = 13, p-value = 0.01 alternative hypothesis: stationary 1: In adf.test(logret[[i]], alternative = "stationary") : p-value smaller than printed p-value</pre>	<pre>[1] "American" Augmented Dickey-Fuller Test data: logret[[i]] Dickey-Fuller = -13.177, Lag order = 13, p-value = 0.01 alternative hypothesis: stationary 1: In adf.test(logret[[i]], alternative = "stationary") : p-value smaller than printed p-value</pre>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figura 17: Resultado do ADF Test para log-retorno de AirFrance-KLM e American Airlines

Enfim, as séries foram divididas em grupos de treino e teste que possuem os mesmos objetivos de sempre: o grupo de treino é usado para estimar o modelo e fazer a previsão enquanto o grupo de teste não é apresentado ao modelo e serve para comparar os valores previstos com os originais.

Novamente são feitas duas divisões: a primeira separa 90% da séries para treino, deixando os 10% finais para teste enquanto a segunda separa apenas os 20 últimos valores para teste. Essas duas divisões são feitas para comparar a qualidade de previsão para horizontes longos e curtos, respectivamente. O nível máximo de decomposição de cada série foi escolhido a partir do menor número inteiro mais próximo do log do tamanho da série. Além disso, os valores máximos do lag para ambos os valores de AR e MA foram definidos como 15 dado que, ao utilizar o modelo ARIMA, todas as séries apresentaram lags menores que este valor.

Ao fim de cada estimativa, foi construído um gráfico comparando os valores previstos e os valores separados como grupo de teste. As imagens abaixo exibem esses resultados

para as duas mesmas companhias aéreas e para ambas as divisões de treino/teste.

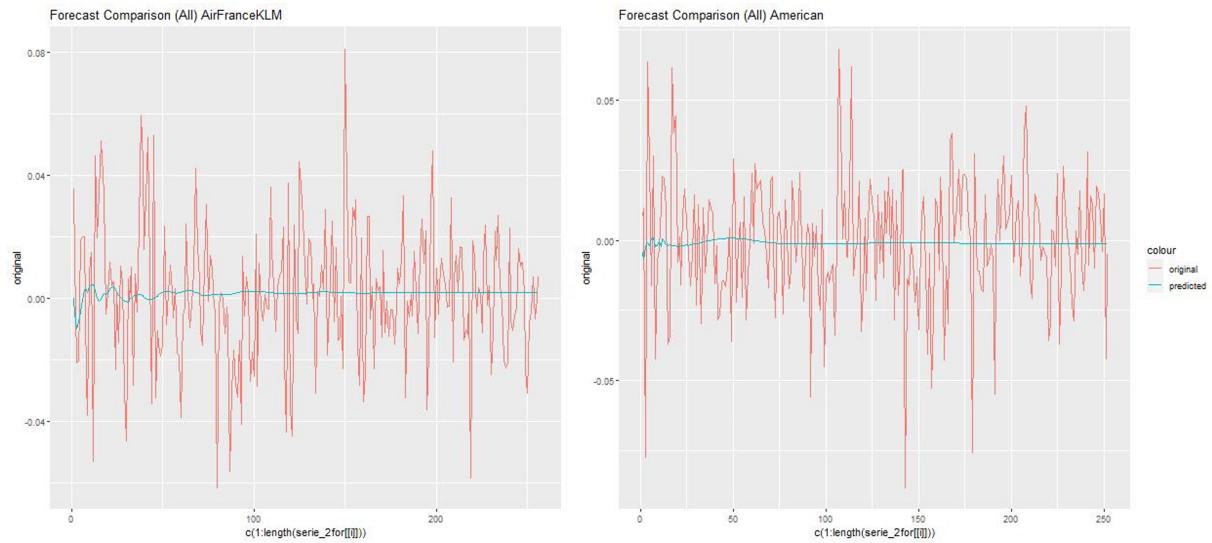


Figura 18: Previsão para 10% finais da AirFrance-KLM e American. Vermelho Original e Azul Previsto

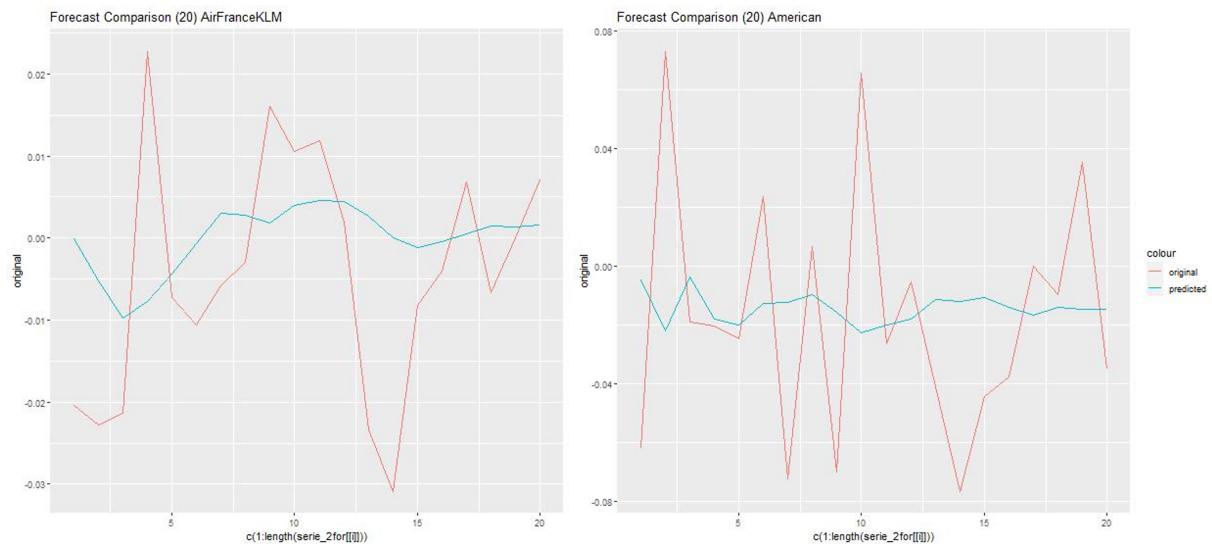


Figura 19: Previsão para 20 valores finais da AirFrance-KLM e American. Vermelho Original e Azul Previsto

Por fim, foi calculado o RMSE entre os valores previstos e os grupos de teste de todas as 10 séries para ambas as divisões implementadas. Esses erros estão exibidos nas tabelas abaixo em comparação com os erros obtidos anteriormente com os modelos ARIMA.

Erros das previsões para os 10% finais das séries		
Série	RMSE ARIMA	RMSE Wavelets/ARIMA
American	0.02271	0.02244
United	0.01319	0.01445
Delta	0.01307	0.01342
Latam	0.02030	0.02411
ChinaSouth	0.02005	0.02307
ChinaEast	0.02116	0.02344
Japan	0.01166	0.01158
IAG	0.01997	0.01914
Lufthansa	0.01680	0.01653
AirFrance-KLM	0.01983	0.02191

Tabela 3: RMSE para divisão 90/10

Erros das previsões para os 20 últimos valores das séries		
Série	RMSE ARIMA	RMSE Wavelets+ ARIMA
American	0.01924	0.01784
United	0.01090	0.01037
Delta	0.01166	0.01147
Latam	0.02063	0.02220
ChinaSouth	0.01589	0.01531
ChinaEast	0.01773	0.01857
Japan	0.00824	0.00814
IAG	0.01424	0.01400
Lufthansa	0.1061	0.01147
AirFrance-KLM	0.01510	0.01438

Tabela 4: RMSE para 20 últimos valores

As conclusões mais detalhadas a cerca de todos os resultados apresentados encontram-se na seção 4.

4 Conclusão

Primeiramente, pode-se notar algo interessante em comum para ambos os modelos VAR e híbrido, que combina Wavelets com o modelo ARIMA. As figuras 4, e 18 deixam claro que nenhum dos dois modelos é capaz de prever com confiança horizontes muito longos. Em ambos os casos as previsões parecem estabilizar na média após, no máximo, 50 passos à frente. O modelo híbrido, provavelmente devido à decomposição Wavelet que deve encontrar padrões em baixas frequências, ainda exibe alguma oscilação ao longo de todo o horizonte de previsão. No entanto, é uma oscilação extremamente sutil que não pode levar à conclusão de que o modelo é capaz de prever com confiança séries de retornos financeiros por um longo período de tempo.

Dessa forma, partindo para as conclusões sobre o modelo VAR, ao analisar as tabelas 1 e 2 que exibem a comparação dos RMSE para os modelos ARIMA e VAR, percebe-se que houve pouca melhora ao se utilizar o modelo vetorial. Esse resultado, embora a

priori possa parecer contra intuitivo, pode ser explicado via resultados apresentados na seção 3.2.2. Ao analisar as possíveis interconexões entre as 10 séries modeladas com VAR, todos os algoritmos deram indícios de que as 10 séries são razoavelmente independentes. Sendo assim, como o modelo não possui informações de relações entre as séries, podemos enxergar o modelo VAR como um simples conjunto de modelos AR que foram computados de forma independente. Esse fato traz uma conclusão plausível para os resultados do RMSE exibidos: para séries pouco relacionadas ou independentes, o modelo VAR não apresenta melhorias na previsão visto que não existe vantagem em analisar em conjunto séries que são independentes. Além disso, percebe-se que em muitos casos o resultado do VAR foi pior do que aquele obtido pelo modelo ARIMA. Um possível motivo para isso é o fato de que, para o modelo vetorial, é definido apenas 1 valor de lag único para todas as séries, enquanto que, ao utilizar o modelo ARIMA, foram definidos valores individuais para cada série, de modo a garantir maior similaridade da estimativa do modelo com o comportamento real da série. Seria interessante, por fim, com o intuito de aumentar a certeza da conclusão em relação à qualidade do modelo VAR para séries independentes, utilizá-lo para modelar e prever séries que são intrinsecamente dependentes ou, no mínimo, assumidamente relacionadas.

Um comentário que deve ser feito é o fato interessante de que, mesmo as 10 séries sendo retornos de ações de empresas de um mesmo nicho - todas são companhias aéreas - a evolução das séries mostra-se independente. Pode-se comentar que o andamento das bolsas em que as ações de cada companhia estão anunciadas sejam distintos uma das outras, mas é, ao meu ver, curioso que nenhum grau de dependência mais significativo do que os resultados apresentados tenha sido obtido. Eventualmente, realizar essa análise para empresas de nichos relacionados e de uma mesma bolsa de valores seria uma análise muito interessante para complementar estes resultados.

Partindo para os resultados do modelo híbrido que combinou a teoria da decomposição de Wavelets com o modelo ARIMA, a principal conclusão, além daquela em relação à previsão de longos horizontes, advém do fato de que, para 20 passos à frente, 7 das 10 séries estudadas apresentaram menor erro em suas previsões. Conclui-se, portanto, que o ferramental poderoso de decomposição em diferentes frequências para a previsão de séries temporais fornece bons resultados quando comparado com modelos clássicos, como os modelos ARIMA e VAR. Como possível continuação ou temas futuros de pesquisa, seria proveitoso testar outros filtros de Wavelet, uma vez que o pacote utilizado apenas implementava para Wavelets Haar e, somado a isso, avaliar esse modelo híbrido para outros tipos de séries temporais, além de comparar seus resultados para séries de retornos financeiros com o modelo GARCH, que é comumente utilizado para a previsão desse tipo de série.

Referências

- [1] R. S. Tsay, **Analysis of financial time series**, Third edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ. (2010).
- [2] R. S. Tsay and G. C. Tiao, **Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models**, J. Amer. Statist. Assoc. 79, p.84-96 (1984).
- [3] Zivot, Eric and Wang, Jiahui. Vector autoregressive models for multivariate time series. **Modeling Financial Time Series with S-Plus** ®, p. 385-429, 2006

- [4] Bailey, Ken. **A combined wavelet and ARIMA approach to predicting financial time series.** 2017. Tese de Doutorado. Dublin City University.
- [5] Polikar, Robi et al. **The wavelet tutorial.** 1996. UC San Diego, Jacobs School of Engineering.
- [6] Graps, Amara. **An introduction to wavelets.** IEEE computational science and engineering, v. 2, n. 2, p. 50-61, 1995.
- [7] CRAN R Project, <https://cran.r-project.org/>
- [8] Github Repository, <https://github.com/MathNog/UndergraduateResearchProject.git>