

Econometria de séries financeiras: da modelagem clássica às ondaletas

Aluno: Matheus Carneiro Nogueira

1 Introdução

Foram estudados os principais modelos lineares de econometria de séries financeiras, Autoregressive Model (AR) e Moving Average (MA), para entender os aspectos fundamentais do modelo ARIMA, Autoregressive Integrated Moving Average. Para o completo entendimento do modelo, foram revisados conceitos básicos de estatística e matemática financeira, como retorno de ativos, propriedades de suas distribuições, linearidade de séries temporais, estacionariedade e auto correlação. Então, o ferramental estudado foi aplicado à séries financeiras obtidas no YahooFinance por meio da impleméticação de rotinas na linguagem R, com o intuito de ajustar as séries escolhidas dentro de modelos ARIMA.

2 Objetivos

Estudar e entender os aspectos teóricos fundamentais da econometria de séries financeiras, com intuito de produzir modelos do tipo ARIMA utilizando o software livre R. Para tal, primeiro objetiva-se conhecer os conceitos básicos que, somados, dão origem a esse modelo. Por fim, estaremos aptos a produzir e calibrar um modelo ARIMA que nos permite fazer previsões acerca da série financeira analisada. Com isso, tornamo-nos capazes de criar um algoritmo em R para ajustar as séries financeiras a um modelo ARIMA que permita previsões satisfatórias.

3 Metodologia

3.1 Teoria de Econometria de Séries Financeira

Primeiramente, é essencial estudar e compreender o funcionamento de dois modelos clássicos da econometria de séries financeiras: *Auto Regressive Model - AR* e *Moving Average Model - MA* [1]. O primeiro baseia-se no fato de que o valor atual de uma série temporal está linearmente relacionado com seus valores passados, ou seja, há uma auto regressão dos valores da série, enquanto o segundo supõe que o erro associado a cada novo valor da série está linearmente associado aos valores médios dos erros passados. A fim de ajustar tais modelos, é importante definir quais valores passados (*lags*) são relevantes para estimar o valor atual da série. Essa informação especifica a ordem dos modelos *AR* e *MA*, representados pela letras *p* e *q* em *AR(p)* e *MA(q)*. Enfim, as equações que definem, respectivamente, esse modelos em suas ordens generalizadas, são

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \dots + \phi_p r_{t-p} + a_t \quad (1)$$

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \quad (2)$$

onde r_t é o valor atual do *log-return* da série, ϕ_i e θ_j são os coeficientes dos lags, a_t é uma *white noise serie* e c_0 uma constante. Ver [1].

Antes de seguir com a apresentação teórica dos modelos, vale um comentário sobre as séries financeiras estudadas. Como visto nas equações 1 e 2, não estamos interessados nos valores em si das séries. Isto quer dizer que não serão estudados nem analisados os preços dos ativos de empresas, mas sim o *log-return* desses preços, calculado como:

$$r_t = \ln(1 + R_t) = p_t - p_{t-1} \quad (3)$$

onde R_t é o retorno simples da série. De forma simplificada, o *log-return* de uma série é a diferença dos logaritmos naturais do valor atual da série e seu valor anterior.

O segundo passo natural para estudar econometria de séries temporais é entender como os *lags* de uma série influenciam seu valor atual. É esta análise, em última instância, que nos permite definir, ou ao menos estimar com certa precisão, as ordens p e q dos modelos $AR(p)$ e $MA(q)$. Para tal análise, introduzimos duas funções: *AutoCorrelation Function - ACF* e *Partial AutoCorrelation Function - PACF*. A rigor, a PACF é função da ACF, mas podemos abstrair essa informação. A equação que define a ACF é

$$\rho_l = \frac{Cov(r_t, r_{t-l})}{\sqrt{Var(r_t) - Var(r_{t-l})}} = \frac{\gamma_l}{\gamma_0} \quad (4)$$

onde ρ_l é a autocorrelação do *lag-l*, $Cov(x, y)$ é a covariância entre x e y e $Var(x)$ é a variância de x . Para definições mais aprofundadas ver [1]. Especificamente, utilizamos a *PACF* para definir a ordem de um modelo $AR(p)$ e a *ACF* para definir a ordem de um modelo $MA(q)$.

Feito o estudo desse dois modelos básicos, podemos compor o modelo $ARMA(p, q)$, isto é, *Autorregressive Moving Average*. Este modelo se origina das propriedades combinadas dos modelos $AR(p)$ e $MA(q)$, apresentando-se como uma ferramenta mais ampla e generalizada para a análise de séries financeiras. Para calibrar os parâmetros p e q deste modelo, as funções ACF e PACF não são, necessariamente, suficientes. É preciso, então, generalizar essas funções para a *Extended Autocorrelation Function - EACF*, proposta em [2]. A análise desta função baseia-se em seu retorno como uma tabela bidimensional que indica os valores adequados de p e q para calibrar o modelo $ARMA(p, q)$ com os atrasos significativos. A equação que define um modelo $ARMA(p, q)$ de ordem generalizada é, segundo [1],

$$r_t = \phi_0 + \sum_{i=1}^p \phi_i r_{t-i} + a_t - \sum_{j=1}^q \theta_j a_{t-j} \quad (5)$$

onde os termos acompanhados por ϕ provém do modelo $AR(p)$ e os acompanhados por θ provém do modelo $MA(q)$. A equação 5 torna fácil perceber que um modelo $ARMA(p, q)$ é, justamente, uma composição dos modelos $AR(p)$ e $MA(q)$ das equações 1 e 2, respectivamente.

O modelo ARMA, no entanto, ainda possui uma especificidade indesejada: considera-se que as séries financeiras analisadas possuem caráter estacionário. Para fins deste estudo, estamos interessados apenas no aspecto fracamente estacionário, que se traduz em uma média constante e covariância independente do tempo. Com o intuito de ampliarmos o escopo de nossa análise de séries financeiras, pode ser usada uma ideia simples em séries não estacionárias para obter uma série estacionária equivalente. Para tal, basta

utilizarmos, ao invés dos valores da série, a diferença entre seus valores e seus valores prévios. Esta técnica de diferenciação dá origem ao modelo que é objetivo inicial desta pesquisa, $ARIMA(p, d, q)$, onde o parâmetro d é, justamente, o número de diferenças necessário para obter-se uma série estacionária capaz de ser analisada pelo ferramental comentado. Percebe-se, desse modo, que um modelo $ARIMA(p, d, q)$ é equivalente a um modelo $ARMA(p, q)$ após d diferenciações. A estacionaridade de um série é, usualmente, de fácil percepção ao analisarmos seu gráfico. Entretanto, há a necessidade de um método menos subjetivo que a interpretação de um gráfico, como, por exemplo, o *Unit Root Test* apresentado em [1]. Esse teste consiste em procurar, no polinômio característico de uma série, alguma raiz unitária. Caso exista uma raiz de valor 1, isso significa que a série não é estacionária, necessitando do processo de diferenciação comentado. Em 3.2 será apresentado um teste que utiliza o *Unit Root Test* para verificar a probabilidade de existir uma raiz unitária em uma dada série.

Uma vez ajustado um modelo $ARIMA(p, d, q)$, abre-se a possibilidade de serem feitas previsões do comportamento da série analisada. Este ferramental é extremamente interessante, pois nos torna capazes de tomar decisões futuras com algum grau de confiança. Tome por exemplo um investidor que precisa decidir entre comprar ou vender ações de determinada empresa, ou realizar movimentações de câmbio. Uma vez armado com um algoritmo capaz de modelar a série financeira dos retornos das ações dessa empresa, ou a série financeira da evolução do câmbio em questão, esse investidor possuirá algum grau de segurança em suas operações, pois tal modelagem o permitirá analisar as previsões das séries desejadas. De certo modo, é intuitivo aceitar que um modelo $ARIMA(p, d, q)$ é capaz de fazer previsões, tendo em vista o fato destes modelos se basearem nos valores passados da série para calcular o valor atual. Sendo assim, é natural estender esta ideia para valores futuros da série. Para tal, de acordo com [1], a previsão de l -passos à frente, denotando a origem de previsão como h é

$$\hat{r}_h(l) = \phi_0 + \sum_{i=1}^p \phi_i \hat{r}_h(l-i) r_{h+l-i} - \sum_{j=1}^q \theta_j a_h(l-j) \quad (6)$$

É interessante, uma vez feita a previsão de l -passos à frente, utilizar alguma medida de erro para comparar os valores previstos com os valores reais, caso existam. Essa abordagem será apresentada na seção 3.2. Somado a isso, é importante comentar que a previsão de séries temporais é incapaz de prever eventos externos que influenciem os dados analisados, uma vez que esses eventos não respeitam a evolução natural da série. Essa análise será aprofundada na seção 3.3, tendo em vista a crise financeira de 2008 e a crise da Covid-19.

3.2 Implementação em R

Com o intuito de implementar uma rotina de ajuste e previsão de modelos $ARIMA(p, d, q)$, foi utilizado o software livre *R*, por ser uma linguagem de programação amplamente utilizada para tratamento de dados econômicos, além de existir amplas referências de estudo de fácil acesso. Outra motivação é a facilidade de importar dados financeiros em diferentes formatos. Neste projeto, os dados utilizados foram obtidos na plataforma *Yahoo Finance* e consistem em arquivos *csv*. A explicação aprofundada dos dados utilizados encontra-se na subseção 3.3, enquanto esta atém-se em apresentar o ferramental de *R* utilizado.

Primeiramente, é necessária a utilização de pacotes de funções específicas para o tratamento de econometria de séries financeiras. Os pacotes utilizados são *tseries* e *Forecast*,

ver [3]. O primeiro oferece métodos de tratamento de séries temporais, enquanto o outro oferece métodos para ajuste dos modelos *ARIMA*. Vale comentar que, a priori, foi utilizado o pacote *FitARMA* para os ajustes dos modelos *ARIMA*. No entanto, embora excelente para o ajuste dos modelos, este pacote não oferece métodos de previsão. Desse modo, por esta ser uma parte crucial deste trabalho, optou-se por substituir o pacote *FitARMA* pelo pacote *Forecast*. O segundo passo é, naturalmente, importar os dados dos arquivos *csv* obtidos e realizar uma breve rotina de tratamento. Muitas tabelas de dados históricos possuem informações faltantes, representados como entradas nulas (*null*), o que impossibilita ou, em menor escala, "suja" as informações. Junto a esse tratamento, tornam-se essenciais algumas conversões de tipo de dados como, por exemplo, converter um *data-frame* proveniente do arquivo *csv* para um vetor numérico. Alguns métodos úteis para essas manipulações são *as.numeric()* para converter os dados em um vetor numérico, *!is.na()* e *!is.na()* para limpar entradas *NA* e *Nan*, as quais, em *R* representam valores nulos ou infinitos. Neste ponto pode-se perceber uma dificuldade inerente ao processo de análise de dados: muitas vezes as informações obtidas possuem características indesejadas que precisam de correção e tratamento, o que pode tornar-se surpreendentemente trabalhoso dependendo da qualidade dos dados encontrados.

Uma vez realizado esse tratamento inicial, uma excelente estratégia é visualizar os dados, por meio do comando *plot()*, ou similares, para estudar o comportamento dos preços das ações e procurar informações interessantes. Na subseção 3.3 essa estratégia será utilizada para perceber a influência da crise de 2008 e da Covid-19 nas séries obtidas. Como comentado em 3.1, estamos interessados no *log-return* e não nos preços em si das ações. Para calcular esse retorno basta utilizar, em conjunto, os métodos *diff()* e *log()* da seguinte forma:

$$\text{logreturn} = \text{diff}(\log(\text{precos}), \text{lag} = 1)$$

A função *diff()* retorna a diferença entre os valores r_t e r_{t-lag} . Ao ser aplicada junto ao método *log()*, que retorna o logaritmo natural da entrada, obtém-se a definição da equação 3.

Com o *log-return* em mãos, novamente são usados os métodos *!is.na()* e *!is.na()* para limpar entradas *NA* e *Nan* que podem surgir das operações anteriores. Em seguida, é necessário converter nossos dados armazenados em um vetor numérico para uma série temporal. Para tal, usa-se a função *ts()* do pacote *tseries* e, mais uma vez, aplica-se uma rotina de tratamento de dados, via *tsclean(série)* [3]. Esta função é extremamente interessante, pois ela usa uma interpolação linear para substituir valores faltantes ou *outliers*.

A última etapa antes da modelagem propriamente dita consiste em separar parte da série obtida para ajustar um modelo *ARIMA* e parte para comparar com a previsão realizada. Esta é uma abordagem clássica utilizada em praticamente todas as áreas de modelagem de dados. A parcela da série separada para ajuste e validação é arbitrária, mas dois cuidados devem ser tomados. Considera-se intuitiva a necessidade de evitar um *underfitting*, isto é, separar poucos dados para o ajuste do modelo e muitos para a validação. Por exemplo, uma divisão 20/80 não faz sentido pois o ajuste teria uma precisão muito abaixo do ideal, devido a pouca quantidade de dados para fundamentar a previsão. Por outro lado, não é interessante haver um *overfitting*, ocasião na qual o modelo é excessivamente ajustado para aquela situação, de forma a perder graus de generalização. Desse modo, e tendo em vista a frequência diária dos dados, para as séries de aproximadamente 2500 observações foram deixados cerca de 150 a 250 dados para validação de previsão, o que corresponde, a uma frequência diária, uma dois terços de um

ano. Tal divisão é trivial de ser realizada em R, bastando criar duas séries que herdem parte dos dados da série original.

Enfim tratados os dados e definidas as séries, a rotina a ser implementada é, praticamente, uma repetição do processo abordado na subseção 3.1 com alguns detalhes a mais. Começamos aplicando um método de decomposição sazonal do pacote *tseries* chamado *decompose()*, ver [3]. No entanto, como esperado pelo perfil da série de *log-returns* de preços de ações, não há sazonalidade aparente e esta função não retorna nada. Comentários acerca da análise de sazonalidade serão realizados na seção 3.3. Em seguida, como comentado na subseção anterior, é utilizada uma função deste mesmo pacote cujo objetivo é verificar a existência de uma raiz unitária no polinômio característico da série. A sintaxe desta função encontra-se a seguir:

$$adf.test(série, alternative = "stationary")$$

Esta função executa o chamado *Augmented Dickey-Fuller Test*, que se baseia no teste da raiz unitária explicado anteriormente. Ver [1]. Define-se a hipótese nula como a existência de uma raiz unitária, que deve ser negada e, consequentemente, uma hipótese alternativa que deve ser confirmada. A interpretação deste teste baseia-se no retorno *p-value*, cujo valor expressa a probabilidade de haver uma raiz unitária. Logo, desejamos o menor valor possível para confirmar a hipótese alternativa e verificar que a série é *fracamente estacionária*. Vale lembrar que uma série estacionária será ajustada em um modelo *ARIMA*($p, 0, q$), onde $d = 0$ significa que não é necessário diferenciar a série.

Uma vez confirmada o caráter estacionário da série, o próximo passo é analisar as funções de correlação apresentadas, *ACF* e *PACF*. Como é de se esperar, o software *R* dispõe, dentro do pacote *forecast* dois métodos que calculam essas duas funções. São eles:

$$Acf = (\text{série})$$

$$Pacf = (\text{série})$$

onde o único argumento obrigatório é uma variável de classe *ts* - *timeseries*, ver [3]. Existem argumentos facultativos, como o número máximo de *lags*, mas foi optado por manter as definições *default* das funções. O retorno de ambos os métodos é extremamente similar, bem como sua interpretação. Ele consiste em um gráfico que exibe o grau de correlação de cada *lag* com o valor atual da série. O ideal é que os valores do gráfico tornem-se, abruptamente, estatisticamente nulos a partir de determinado *lag l*, de modo que todos os *lags* anteriores a *l* são considerados para o ajuste do modelo *ARIMA*(p, d, q), definindo, como expresso em 3.1, as ordens *p* e *q*. Não é de se surpreender que em séries reais esse comportamento pode não ser percebido de forma tão clara, o que será evidenciado na subseção 3.3. É importante ressaltar que não foi usada a função *EACF* [2], simplesmente por não existir, dentro dos pacotes usuais, um método que a implemente. Dito isso, a melhor estratégia será ajustar diferentes modelos *ARIMA* para avaliar quais ordens *p* e *q* são as mais adequadas.

Carregado de possíveis candidatos para as ordens *p* e *q*, estamos aptos a ajustar um modelo. Novamente, o pacote *forecast* possui um método adequado que implementa um modelo *ARIMA*(p, d, q), cuja sintaxe é

$$arima(serie, order = c(p, d, q))$$

Os argumentos necessários são apenas a série modelada (uma variável da classe *ts*) e a ordem do modelo. Esse método retorna um objeto da classe *arima*, que pode ser visualizado por meio das funções *coef(arimaObject)*, que exibe os coeficientes θ_i e ϕ_i do modelo

e `print(arimaObject)`, que exibe informações relevantes para a análise e comparação de diferentes modelos, além de seus coeficientes. Esses detalhes serão abordados na subseção 3.3.

Além das informações obtidas pela função `print(arimaObject)`, é interessante analisar os resíduos dos modelos ajustados, cujo comportamento ideal é caracterizado por não existência de correlação, visualizado via *ACF*, média constante e igual a zero, covariância independente do tempo e histograma similar à distribuição normal. A função do pacote `forecast` que calcula todas essas informações é a `checkresiduals(arimaObject)` [3].

Após analisar e comparar os diferentes modelos *ARIMA* ajustados, e escolhidos os mais adequados, temos todo o ferramental para fazer previsões do comportamento de séries temporais. Com esse intuito, o pacote `forecast` possui a função `forecast()` que utiliza o método descrito pela equação 6 para prever n passos a frente o valor da série e cuja sintaxe é descrita a seguir e definida em [3].

$$\text{forecast}(\text{object}, h = \dots, \text{level} = c(X, Y))$$

onde `object` é uma variável da classe *arima*, tipicamente o retorno da função `arima()`, h é o número de passos à frente para serem calculados a partir da origem de previsão e `level` é o intervalo $[X, Y]$, em porcentagem de confiança. Há outros argumentos opcionais que não foram usados neste estudo.

Por fim, após ser feita a previsão dos modelos *ARIMA*, além de visualizar por meio de um `plot()` a série resultante, é extremamente importante averiguar a precisão das previsões realizadas. Para tal, foi separada parte da série original justamente com o intuito de comparar os resultados obtidos com os valores reais da série. Com isso em mente, algumas maneiras distintas são possíveis de serem implementadas. Primeiramente, com algumas manipulações e conversões de tipos de dados, é possível simplesmente calcular medidas de erro e dispersão, como o desvio padrão entre as duas séries. De forma alternativa, existe um método em *R* denominado `accuracy()`, dentro do pacote `forecast` [3] que recebe como argumento as duas séries a serem comparadas e retorna informações relevantes sobre o *training set*, o modelo a ser previsto e o *test set*, série de controle e comparação. Essas medidas são: *Mean Error (ME)*, *Root Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, *Mean Percentage Error (MPE)*, *Mean Absolute Percentage Error (MAPE)*, *Mean Absolute Scaled Error (MASE)* e *AutoCorrelation Function (ACF)*.

Com todo este ferramental de métodos e funções em *R*, somos capazes de tratar, limpar, ajustar e prever modelos *ARIMA*(p, d, q). Todo o código produzido para esta análise encontra-se em [4].

3.3 Discussão de Resultados

Foram estudados os preços ajustados das ações de 10 companhias aéreas relevantes no cenário mundial e distribuídas em vários continentes. São elas, com seus respectivos códigos nas bolsas de valores: *American Airlines (AAL)*, *Delta Airlines (DAL)*, *United Airlines (UAL)*, *Latam (LTMAQ)*, *Lufthansa (LHA.DE)*, *Air France - KLM (AF.PA)*, *IAG (IAG.L)*, *China Southern Airlines (1055.HK)*, *China Eastern Airlines (0670.HK)* e *Japan Airlines (9201.T)*. Todos os dados referentes a essas empresas foram retirados da plataforma *Yahho Finance*. Para fins de organização, estão seção está dividida em duas partes. A primeira possui o objetivo e estudar os diferentes impactos das crises de 2008 e da Covid-19, enquanto a segunda irá apresentar, de fato, a aplicação das técnicas discutidas nas seções 3.1 e 3.2 para as ações das empresas citadas anteriormente.

3.3.1 Impacto de crises em séries financeiras

Todo o estudo da econometria de séries financeiras [1] assume que os dados gerados possuem alguns aspectos conhecidos. Sejam médias constantes, covariâncias independente do tempo, sazonalidades ou *trends*, todo o ferramental apresentado é usado para séries "bem comportadas". Não é difícil compreender, portanto, que eventos externos ao andamento usual de uma série financeira influenciam muito o comportamento e, mais importante, a previsibilidade dos dados. Dito isso, analisaremos a resposta dos preços das ações e seus *log-returns* de algumas das companhias aéreas supracitadas a momentos de crise e instabilidade econômica. A justificativa de excluir algumas companhias desta análise é o fato de apresentarem um volume de dados pré 2007 insuficiente para uma interpretação interessante. Infelizmente, todas as companhias americanas estudadas enquadram-se nesse aspecto, mas isso não compromete a validade dos comentários desta seção.

Utilizando os métodos descritos na seção 3.2, foram produzidas as imagens abaixo que apresentam, em primeira instância, três gráficos distintos: a evolução dos preços ajustados das ações, o *log-return* destes preços e a série temporal deste *log-return*. Vale relembrar que diversos métodos de tratamento foram utilizados em diferentes momentos para limpar os dados de informações indesejadas, o que justifica possíveis diferenças entre os gráficos. Em segundo lugar, a figura 1 possibilita a comparação destes gráficos tendo em vista cinco companhia aéreas, *AirFrance-KLM*, *Lufthansa*, *Latam*, *China Southern* e *China Eastern*, respectivamente. O que podemos esperar são quedas abruptas nos preços das ações destas empresas, mas o efeito no *log-return* dos dados pode ser um pouco menos intuitivo. As linhas azuis demarcam, primeiramente, o período de 2008 a 2009 e, ao final dos gráficos, o início de 2020. Uma diferença clara na análise do impacto destas duas crises nas séries financeiras é o fato de a crise da Covid-19 ser muito recente, ainda vivenciada, e a crise de 2008 já ter sido, de algum modo, superada. Isso se traduz em não sabermos ainda como será o comportamento dessas séries pós 2020, mas não impossibilita outras análises importantes.

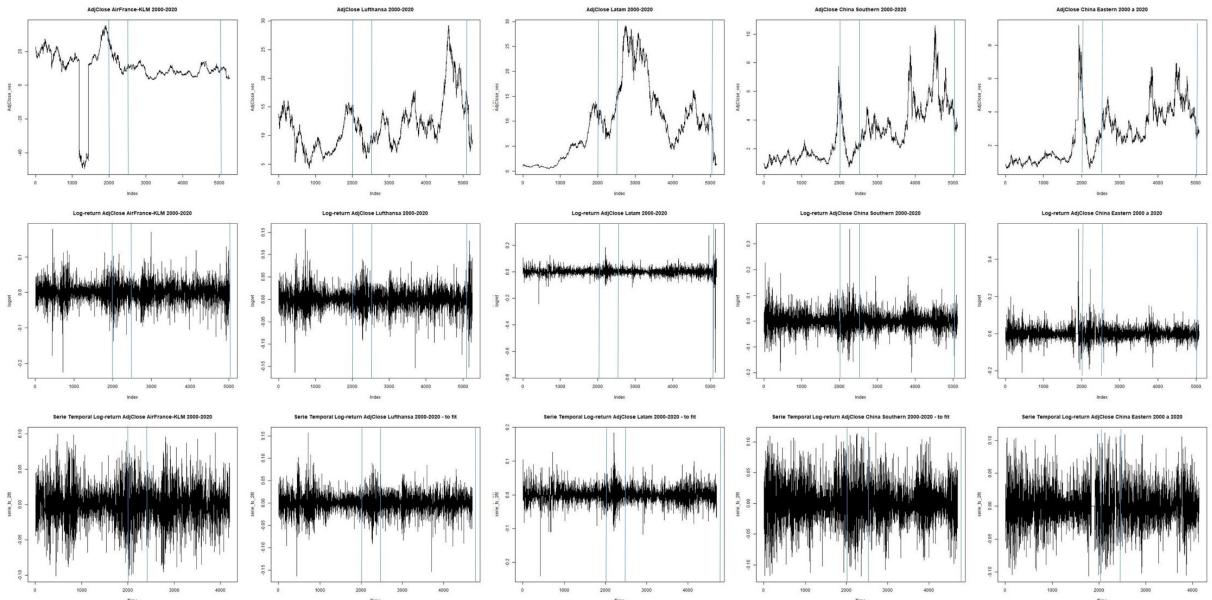


Figura 1: Gráficos Comparativos - Preços Ajustados, Log-Return, Série Temporal do Log-Return

Em todas as companhias existe a queda abrupta no primeiro gráfico, como previsto. Consequentemente, deve haver algum efeito observável nas séries do *log-return*, exibidas na figura 1 nos segundos e terceiros gráficos. De fato, o efeito produzido é percebido em forma de picos, tanto positivos quanto negativos. Repare que entre as linhas azuis, isto é, entre 2008 e 2009, há um grande grau de volatilidade nessas séries, haja vista esse comportamento mais acentuado. Tal mudança drástica se dá devido à natureza do *log-return*, pois, como expresso na equação 3, esse retorno computa as diferenças entre os valores das ações no tempo t e $t - 1$. Sendo assim, é trivial o fato de quedas abruptas nos preços terem como consequência picos acentuados no *log-return*. Apenas no caso da *AirFrance-KLM* este fato não é tão perceptível. Para fins de ajuste de modelos *ARIMA*, esses picos de volatilidade são muito indesejáveis, uma vez que não correspondem à evolução natural da série. É por esse motivo que choques externos são imprevisíveis ao olharmos apenas para a série estudada e, além disso, tornam o desenvolvimento dela também de difícil previsão.

Ao mesmo tempo em que a queda de preços foi abrupta, a recuperação não tardou muito a ocorrer. A principal estratégia para contornar uma crise como a de 2008 é tentar reanimar a economia, ou seja, criar maneiras de movimentar e reiniciar as transações econômicas em suas diferentes escalas. Para tal, uma primeira ação clássica é aplicar uma redução de juros para injetar dinheiro na economia e incentivar o consumo e investimento que, por sua vez, incentivam a produção. Caso este método se mostre insuficiente, uma solução é, assim como foi descoberto em 2008, realizar um socorro público às empresas de caráter privado. Nesse sentido, o governo, ao invés de passar pelo intermédio dos bancos, que em uma crise podem reter recursos com receio de quebra, injeta capital de forma direta nas empresas. Em maiores detalhes, foi esta a estratégia, denominada *Quantitative Easing - QE* adotada pelos Bancos Centrais em 2008, que passaram a comprar títulos de dívida de empresas privadas, aumentando o volume de capital em circulação e garantindo liquidez para o mercado que estava congelado.

Feita esta análise, olhemos para a terceira linha azul nos gráficos da figura 1, que indica o início do ano de 2020. Novamente é possível perceber a queda nos preços das ações das companhias aéreas que, da mesma forma visualizada em 2008, traduz-se em picos de volatilidade nas séries do *log-return*. Por outro lado, existem algumas diferenças consideráveis entre a crise anterior e a gerada pela Covid-19. Primeiramente, enquanto aquela teve como origem o próprio mercado financeiro, esta configura-se como um choque inteiramente externo ao mercado, tornando-o ainda mais invisível aos modelos de econometria. Em segundo lugar, enquanto em 2008 o mercado congelou suas atividades de forma espontânea, em 2020, novamente, a parada de diversas atividades econômicas ocorreu como consequência de ordens externas. A principal diferença, no entanto, é que ainda vivemos a crise da Covid-19, tornando-nos incapazes de analisar com precisão seus reais efeitos a médio e longo prazo. É certo que o *Quantitative Easing* está sendo utilizado, fruto de seu bom funcionamento no passado, tornando-se mais uma estratégia importante no combate à instabilidade econômica. O *U.S Federal Reserve*, por exemplo anunciou um programa de *QE* de mais de 700 bilhões de dólares. Há aqueles que consideram que esta estratégia, aprendida em 2008, causará uma recuperação mais rápida para os efeitos negativos da crise de 2020, no entanto, possíveis segundas ondas de contaminação, possibilidade de novas medidas de contenção e um enorme medo por parte da sociedade tornam, ao meu ver, essa previsão pouco confiável.

3.3.2 Ajuste e previsão de modelos ARIMA

Nesta seção serão apresentados os resultados obtidos da aplicação dos métodos discutidos na seção 3.2, tendo como dados de entrada os preços ajustados das ações de dez empresas aéreas, enumeradas na seção 3.3.1. Dito isso, foi feito o ajuste para modelos *ARIMA* de todas as empresas no período de 2010 a 2019, reservando uma parte da série para validação e comparação dos modelos com os valores reais. As imagens 2 a 6 exibem, simultaneamente, os preços ajustados das ações das empresas aéreas e a série temporal do *log-return* destes preços, após os devidos tratamentos já apresentados.

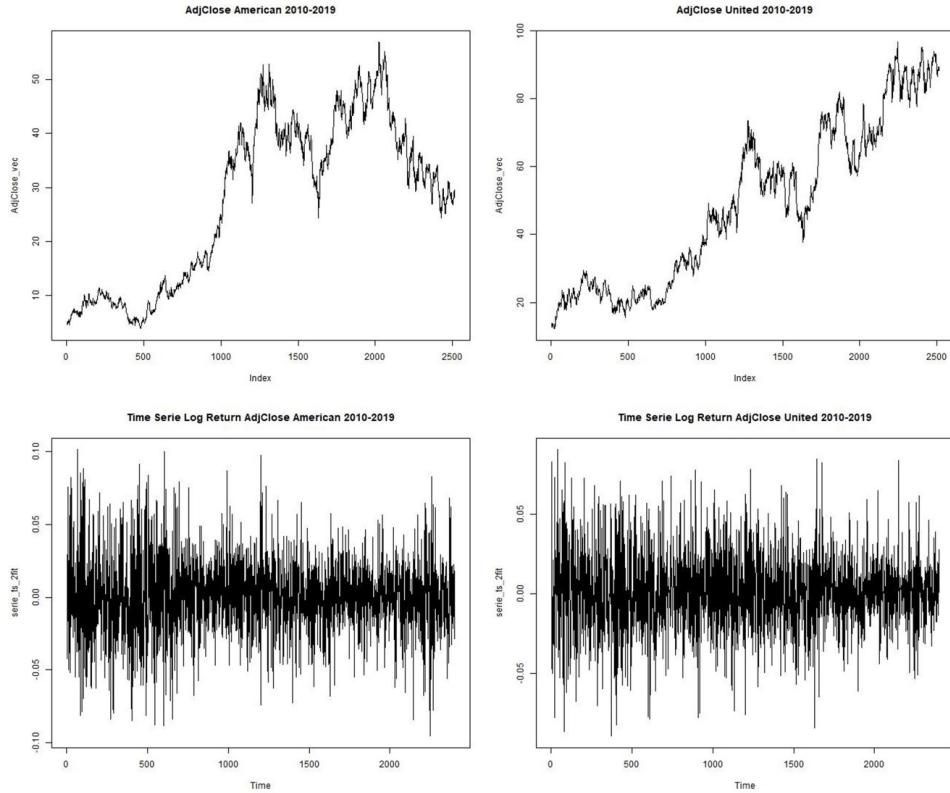


Figura 2: Preços ajustados e Log Return - American e United

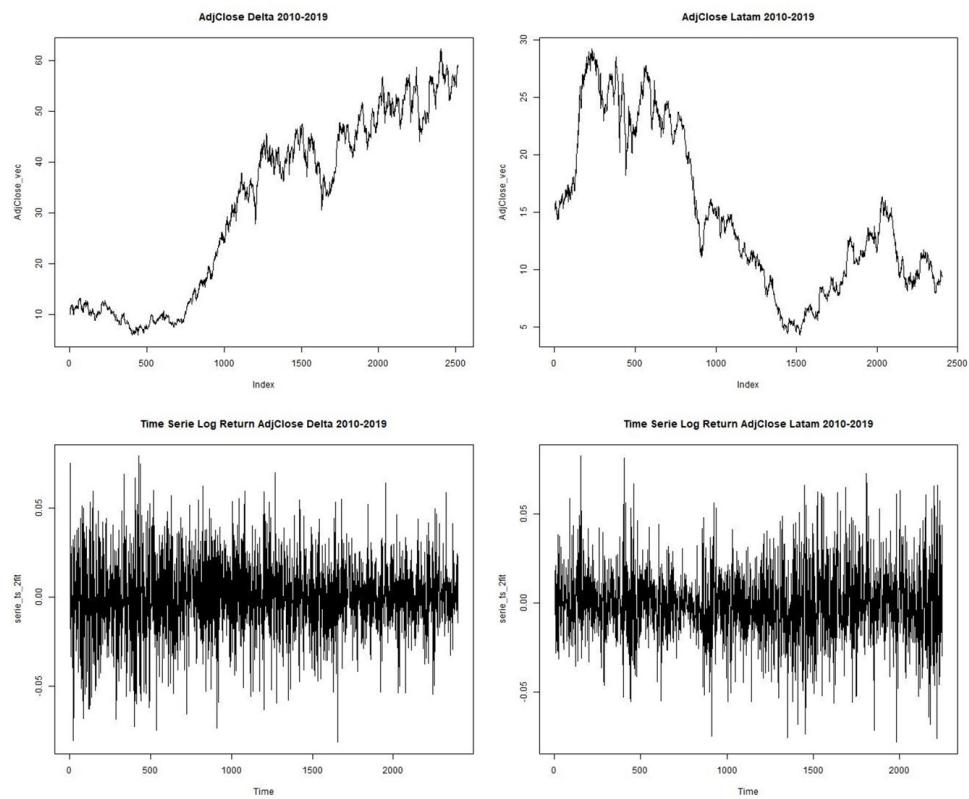


Figura 3: Preços ajustados e Log Return - Delta e Latam

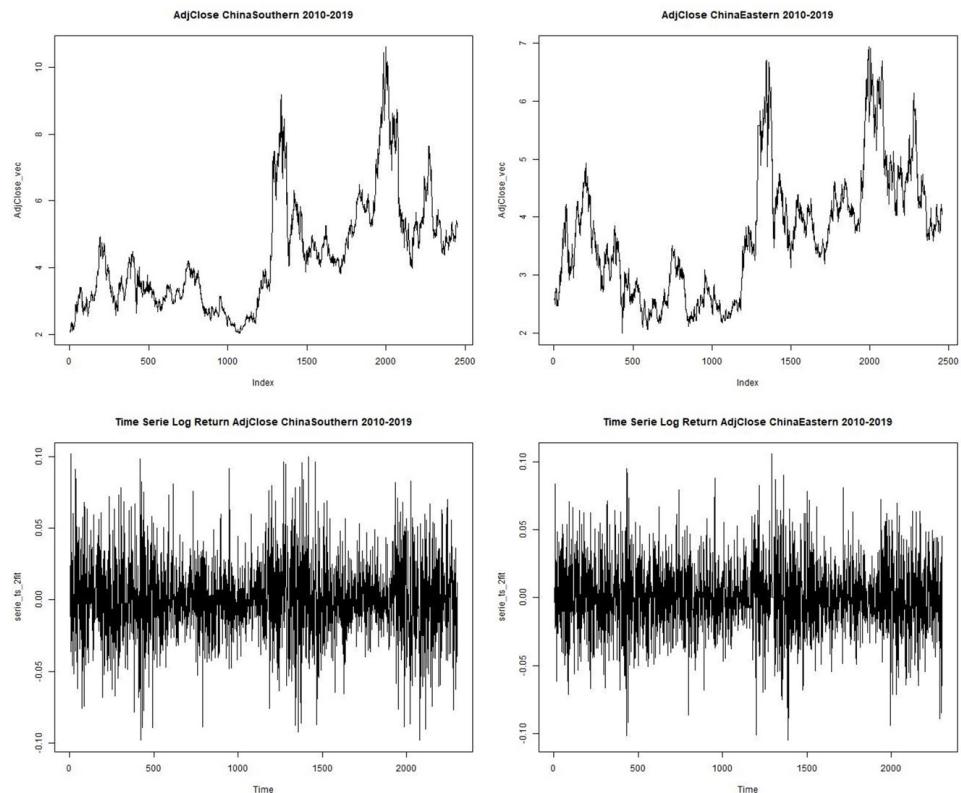


Figura 4: Preços ajustados e Lo Return - China Southern e China Eastern

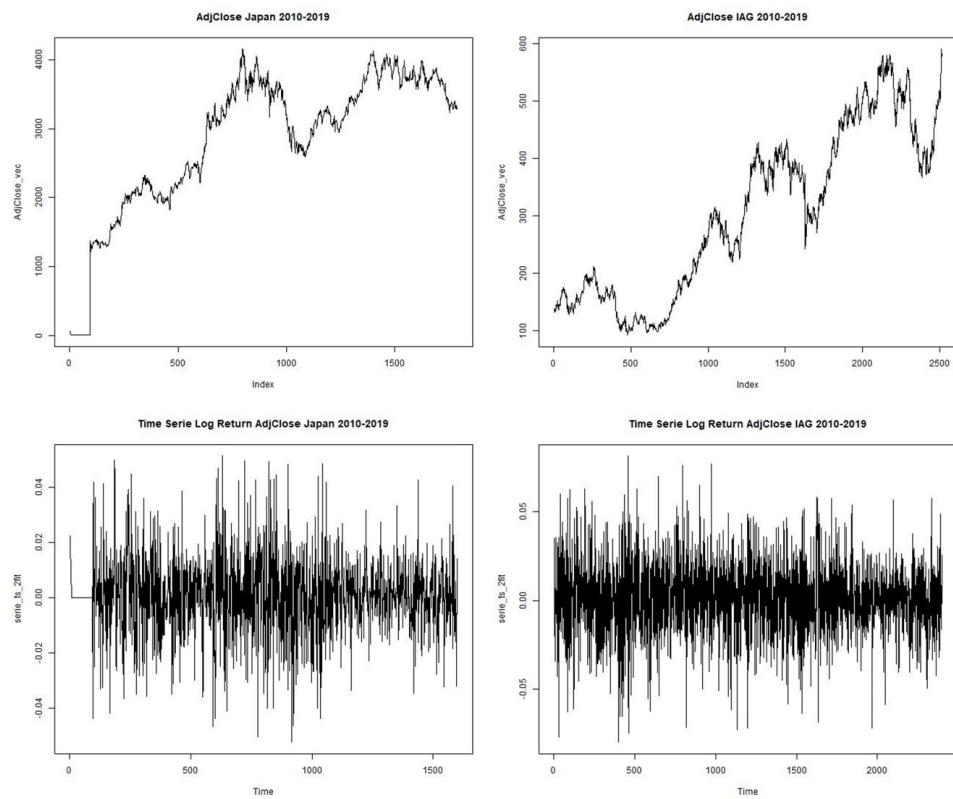


Figura 5: Preços ajustados e Log Return - Japan e IAG

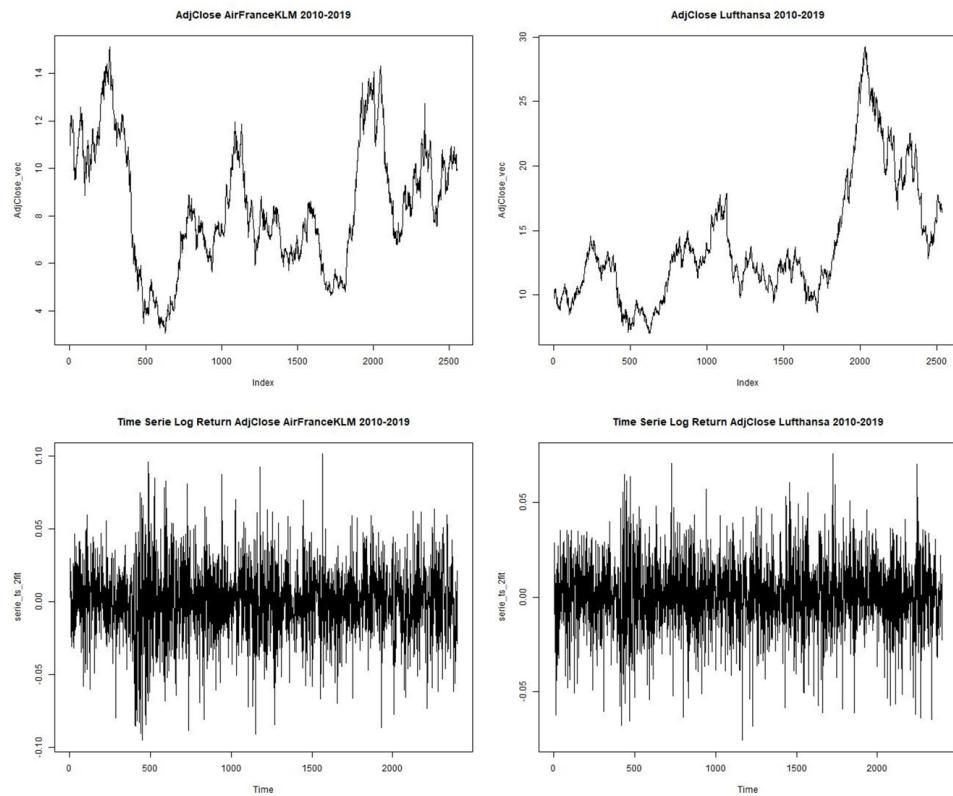


Figura 6: Preços ajustados e Log Return - AirFrance KLM e Lufthansa

Uma vez visualizadas as séries, foram executados ambos *Augmented Dickey-Fuller Test* e o teste para sazonalidade, via métodos *adf.test()* e *decompose()*. O *output* da execução do primeiro encontra-se abaixo:

```
AugmentedDickey - FullerTest
data : serie.ts.2fit
Dickey - Fuller = -12.924, LagOrder = 13, p - value = 0.01
alternative hypothesis : stationary
```

Para todas as séries analisadas, este método retornou *p-values* iguais ou menores que 0.01, o que demonstra, no máximo, 1% de probabilidade de existir uma raiz unitária no polinômio característico das séries. Devido a isso, omite-se o resultado das demais séries, que são encontrados em [4]. Consequentemente, todos os *log-returns* estudados apresentam aspecto estacionário. O método *decompose()* também apresentou retornos similares para todas as séries, sendo este a seguinte mensagem de erro: *Error in decompose(serieTs2fit): série temporal não tem período, ou tem menos de 2*. Isto quer dizer que os retornos das ações das companhias aéreas não possuem perfil sazonal, o que pode ser contra intuitivo, haja vista o claro comportamento sazonal de viagens aéreas, mais intensas em períodos de férias escolares, por exemplo. Desse modo, a perda de sazonalidade entre as viagens em si e os retornos dos ativos indica um grau de maturidade do mercado de ações dessas companhias, sendo capaz de suavizar seu perfil sazonal. É interessante, no entanto, suspeitar que uma mudança na escala temporal da análise pode ter como consequência o aparecimento de ciclos sazonais, algo que será objetivo de análise futura, após introduzido o ferramental das *ondaletras*.

Uma vez confirmado o caráter estacionário e não sazonal das séries a serem ajustadas, tornamo-nos aptos a analisar as funções de correlação já conhecidas, *ACF* e *PACF*. Utilizando os métodos apresentados na seção 3.2, foram produzidos os seguintes correogramas.

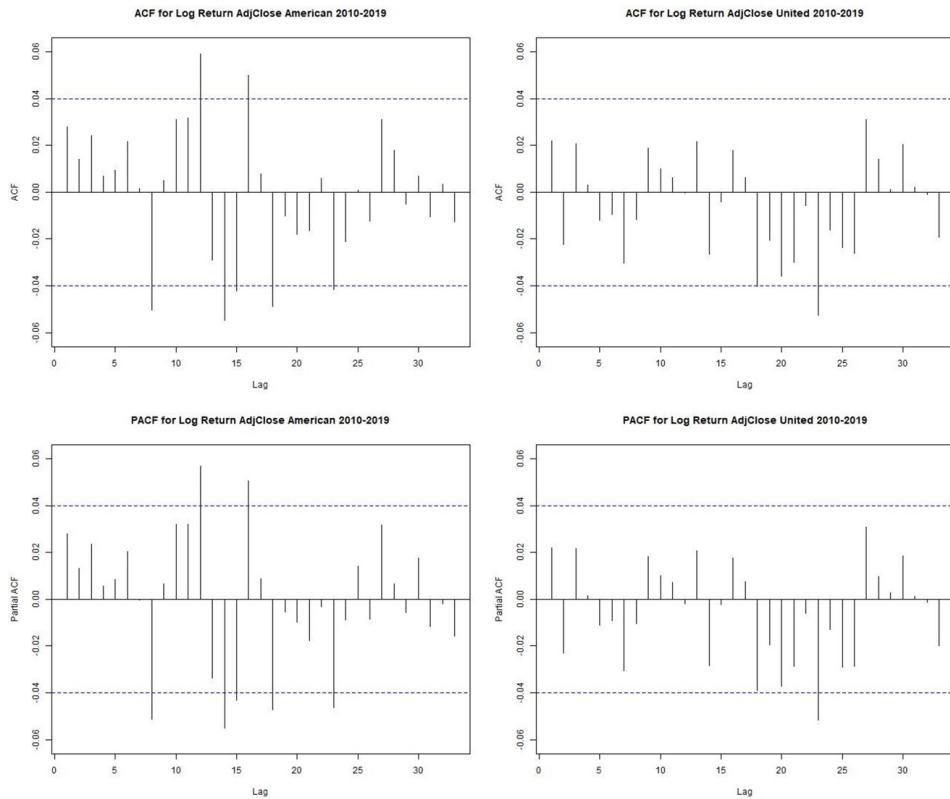


Figura 7: ACF e PACF - American e United

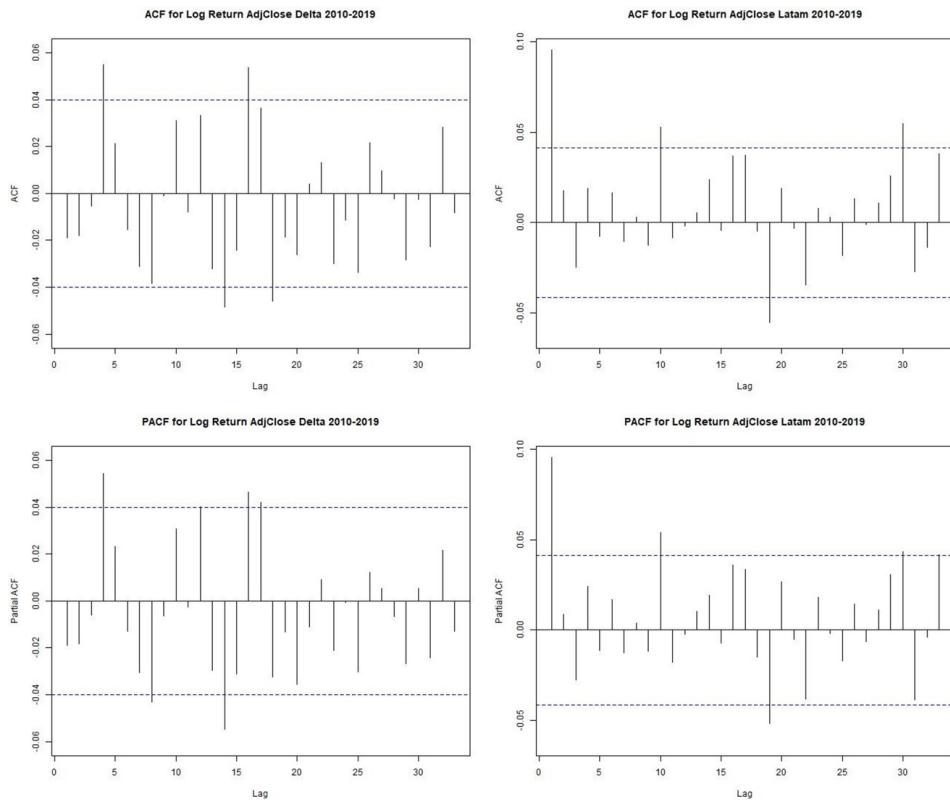


Figura 8: ACF e PACF - Delta e Latam

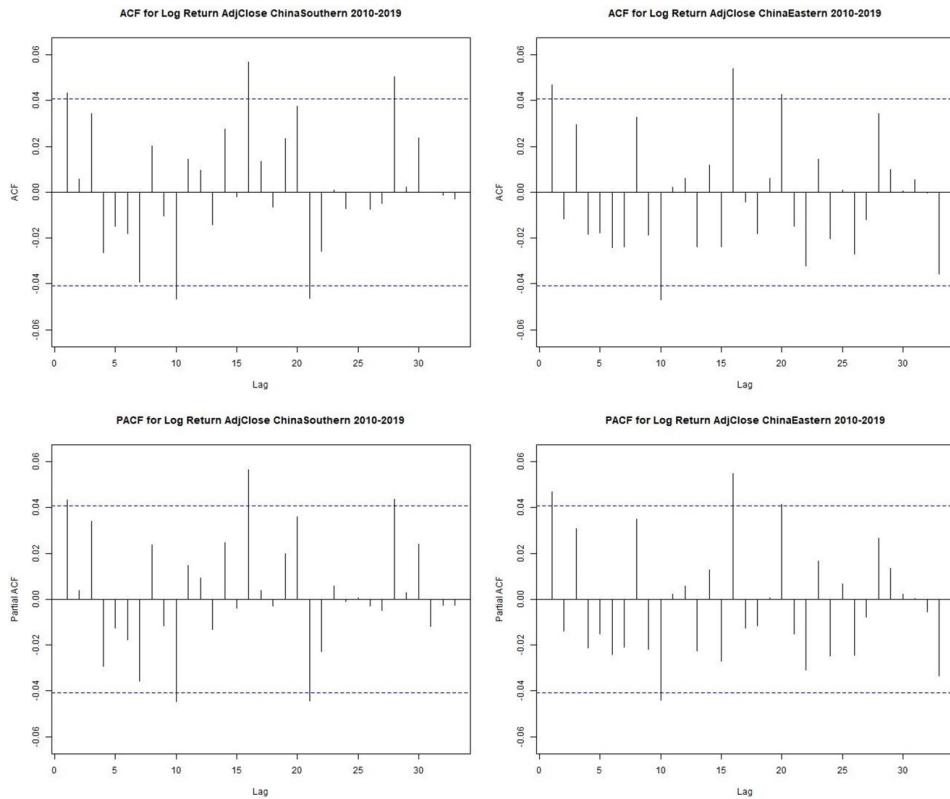


Figura 9: ACF e PACF - China Southern e China Eastern

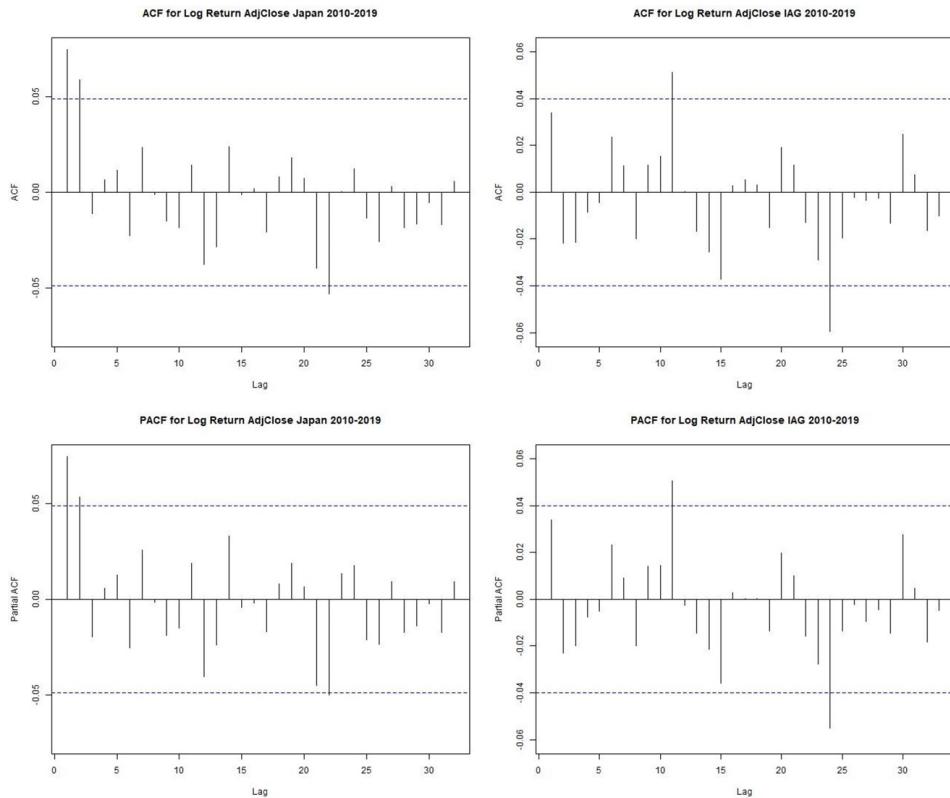


Figura 10: ACF e PACF - Japan e IAG

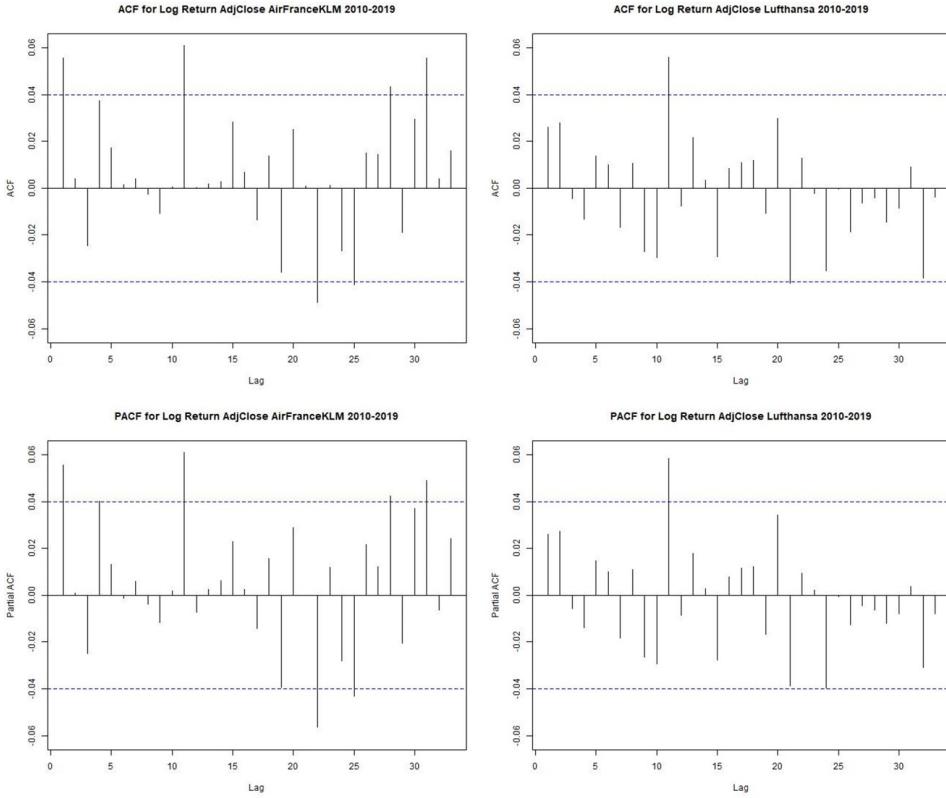


Figura 11: ACF e PACF - AriFrance-KLM e Lufthansa

A interpretação dos gráficos presentes nas figuras 7 a 11 é, de certo modo, trivial. Como apresentado na seção 3.2, procuramos os *lags* estatisticamente não nulos, isto é, acima da linha pontilhada superior ou abaixo da inferior dos gráficos. Dessa forma, foram ajustados os seguintes modelos $ARIMA(p, d, q)$ para cada companhia aérea: **American** - $(8, 0, 0)$, $(0, 0, 8)$, $(12, 0, 0)$, $(0, 0, 12)$, $(14, 0, 0)$, $(0, 0, 14)$, $(8, 0, 8)$, $(12, 0, 8)$ e $(14, 0, 8)$; **United** - $(23, 0, 0)$, $(0, 0, 23)$; **Delta** - $(4, 0, 0)$, $(0, 0, 4)$, $(0, 0, 8)$, $(4, 0, 4)$; **Latam** - $(1, 0, 0)$, $(0, 0, 1)$, $(10, 0, 0)$, $(0, 0, 10)$, $(19, 0, 0)$, $(0, 0, 19)$, $(1, 0, 1)$, $(1, 0, 10)$, $(1, 0, 19)$, $(10, 0, 1)$; **China Southern** - $(1, 0, 0)$, $(0, 0, 1)$, $(10, 0, 0)$, $(0, 0, 10)$, $(16, 0, 0)$, $(0, 0, 16)$, $(1, 0, 1)$, $(1, 0, 10)$, $(16, 0, 1)$; **China Eastern** - $(1, 0, 0)$, $(0, 0, 1)$, $(10, 0, 0)$, $(0, 0, 10)$, $(16, 0, 0)$, $(0, 0, 16)$, $(1, 0, 1)$, $(1, 0, 10)$, $(1, 0, 16)$, $(10, 0, 1)$, $(10, 0, 16)$, $(16, 0, 1)$; **Japan** - $(1, 0, 0)$, $(0, 0, 1)$, $(2, 0, 0)$, $(0, 0, 2)$, $(1, 0, 1)$, $(1, 0, 2)$, $(2, 0, 1)$, $(2, 0, 2)$; **IAG** - $(1, 0, 0)$, $(0, 0, 1)$, $(11, 0, 0)$, $(0, 0, 11)$, $(24, 0, 0)$, $(0, 0, 24)$, $(1, 0, 1)$, $(1, 0, 11)$, $(1, 0, 24)$, $(11, 0, 1)$, $(11, 0, 24)$, $(24, 0, 1)$; **AirFrance KLM** - $(1, 0, 1)$, $(0, 0, 1)$, $(11, 0, 0)$, $(0, 0, 11)$, $(0, 0, 21)$, $(1, 0, 1)$, $(1, 0, 11)$, $(1, 0, 21)$, $(11, 0, 1)$, $(11, 0, 11)$, $(21, 0, 1)$ e **Lufthansa** - $(11, 0, 0)$, $(0, 0, 11)$, $(21, 0, 0)$.

Uma vez ajustados todos os inúmeros modelos, torna-se essencial compará-los a fim de selecionar os mais adequados. Para tal, é feita uma análise envolvendo duas medidas estatísticas, o *log likelihood* e o *aic*. Suas interpretações se dão de modo que o *log likelihood* deve ser o maior possível, enquanto o *aic* o menor possível. Com base nessas medidas, foram selecionados de dois modelos $ARIMA$ de cada companhia aérea para serem calculadas as previsões dos retornos de seus ativos. Os modelos selecionados foram: **American** - $(12, 0, 8)$ e $(14, 0, 8)$; **United** - $(23, 0, 0)$ e $(23, 0, 23)$; **Delta** - $(0, 0, 4)$ e $(4, 0, 4)$; **Latam** - $(1, 0, 0)$ e $(1, 0, 19)$; **China Southern** - $(1, 0, 0)$ e $(16, 0, 1)$; **China Eastern** - $(10, 0, 16)$ e $(0, 0, 1)$; **Japan** - $(0, 0, 2)$ e $(2, 0, 2)$; **IAG** - $(1, 0, 0)$ e $(11, 0, 24)$; **AirFrance KLM** - $(1, 0, 0)$ e $(11, 0, 11)$ e **Lufthansa** - $(11, 0, 0)$ e $(21, 0, 0)$. Com o intuito de não poluir este

relatório, serão exibidos a seguir os coeficientes e outras medidas importantes apenas dos melhores modelos de cada companhia, com base em seus *log likelihood* e *aic*. No entanto, todas as informações produzidas que não estão exibidas encontram-se em [4].

ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	ar9	ar10
0.0875	-0.2647	-0.0537	0.3572	-0.1591	-0.0099	-0.0805	-0.643	0.0241	0.0176
ar11	ar12	ma1	ma2	ma3	ma4	ma5	ma6	ma7	ma8
0.0326	0.0872	-0.0641	0.2853	0.0864	-0.3499	0.1658	0.0334	0.0947	0.6223
loglikelihood = 5349.78									
aic = -10655.56									

Tabela 1: Coeficientes American - ARIMA(12,0,8)

ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	ar9	ar10
0.1518	0.022	-0.0951	0.4539	-0.0262	-0.2079	-0.0159	-0.6647	0.0316	0.0259
ar11	ar12	ar13	ar14	ma1	ma2	ma3	ma4	ma5	ma6
0.0216	0.0826	-0.0182	-0.0535	-0.1288	-0.0027	-0.1207	-0.4531	-0.0188	0.2193
ma7	ma8								
0.0207	0.6281								
loglikelihood = 5350.57									
aic = -10653.15									

Tabela 2: Coeficientes American - ARIMA(14,0,8)

ma1	ma2	ma3	ma4
-0.0219	-0.0146	0.0001	0.0607
Log Likelihood = 5844.54			
aic = -11667.07			

Tabela 3: Coeficientes Delta - ARIMA(0,0,4)

ar1	ar2	ar3	ar3	ma1	ma2	ma3	ma4
0.400	-0.225	-0.5555	0.1122	-0.4251	0.2224	0.5638	-0.0684
Log Likelihood = 5854							
aic = -11688.01							

Tabela 4: Coeficientes Delta - ARIMA(4,0,4)

ar1
0.0957
Log Likelihood = 5599.05
aic = -11192.11

Tabela 5: Coeficientes Latam - ARIMA(1,0,0)

ar1	ma1	ma2	ma3	ma4	ma5	ma6	ma7	ma8	ma9
-0.1615	0.2599	0.0402	-0.0268	0.0102	-0.0003	0.0126	-0.009	0.0064	-0.0128
ma10	ma11	ma12	ma13	ma14	ma15	ma16	ma17	ma18	ma19
0.0484	0.0078	-0.0034	-0.007	0.0272	0.0019	0.0247	0.0427	0.0016	-0.0648
loglikelihood = 5612.51									
aic = -11181.01									

Tabela 6: Coeficientes Latam - ARIMA(1,0,19)

ar1
0.0434
Log Likelihood = 5030.81
aic = -10055.62

Tabela 7: Coeficientes China Southern - ARIMA(1,0,0)

ar1	ar2	ar3	ar4	ar5	ar6	at7	ar8	ar9	ar10
0.1138	-0.0011	0.0325	-0.0291	-0.0113	-0.0147	-0.032	0.0252	-0.0095	-0.0430
ar11	ar12	ar13	ar14	ar15	ar16	ma1			
0.0176	0.0106	-0.0174	0.0264	-0.0088	0.0573	-0.0693			
Log Likelihood = 5043.31									
aic = -10048.61									

Tabela 8: Coeficientes China Southern - ARIMA(16,0,1)

ma1
0.0484
Log Likelihood = 5113.95
aic = -10221.91

Tabela 9: Coeficientes China Eastern - ARIMA(0,0,1)

ma1	ma2
0.0727	0.0604
Log Likelihood = 4549.41	
aic = -9090.83	

Tabela 10: Coeficientes Japan - ARIMA(0,0,2)

ma1	ar2	ma1	ma2
-0.2099	-0.0255	0.282	0.1005
Log Likelihood = 4549.6			
aic = -9087.19			

Tabela 11: Coeficientes Japan - ARIMA(2,0,2)

ar1
0.0341
Log Likelihood = 5931.49
aic = -11856.97

Tabela 12: Coeficientes IAG - ARIMA(1,0,0)

ar1
0.0557
Log Likelihood = 5478.34
aic = -10950.67

Tabela 13: Coeficientes KLM - ARIMA(1,0,0)

Enfim ajustados e decididos os melhores modelos *ARIMA*, o método *forecast()* torna-se o próximo passo natural. Como comentado na seção 3.2, um de seus parâmetros obrigatórios é o intervalo de confiança e o outro é o número de passos de previsão, além, é claro, da série a ser prevista. Para este estudo, definiu-se este intervalo como 80% a 95%, o que é percebido pelas faixas de tons cinzas das imagens a seguir. A quantidade de passos, por outro lado, é diferente para cada modelo, mas todos entre 150 e 250 previsões. Novamente, todas as imagens produzidas encontram-se em [4], para aqui estarem dispostas apenas as mais relevantes, uma de cada companhia.

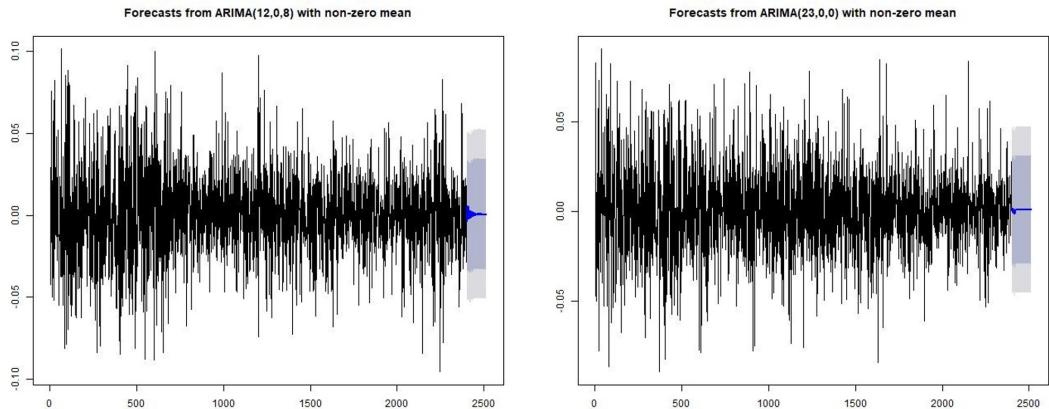


Figura 12: Forecast American e United

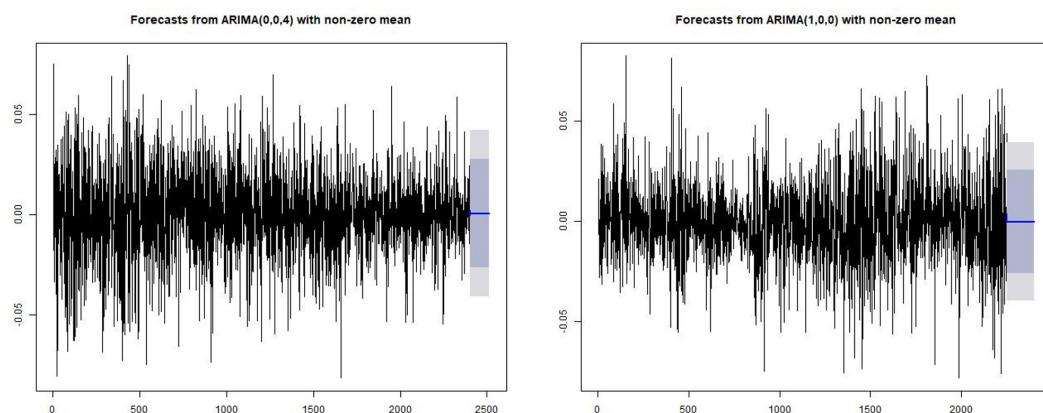


Figura 13: Forecast Delta e Latam

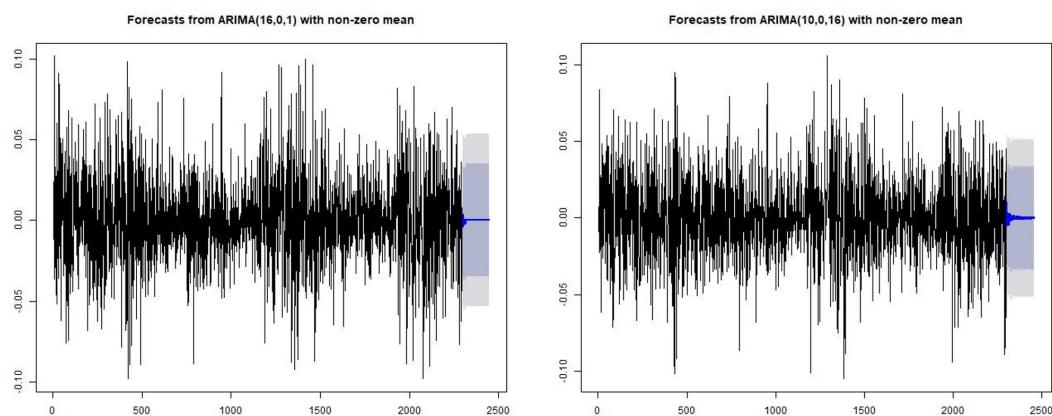


Figura 14: Forecast China Southern e China Eastern

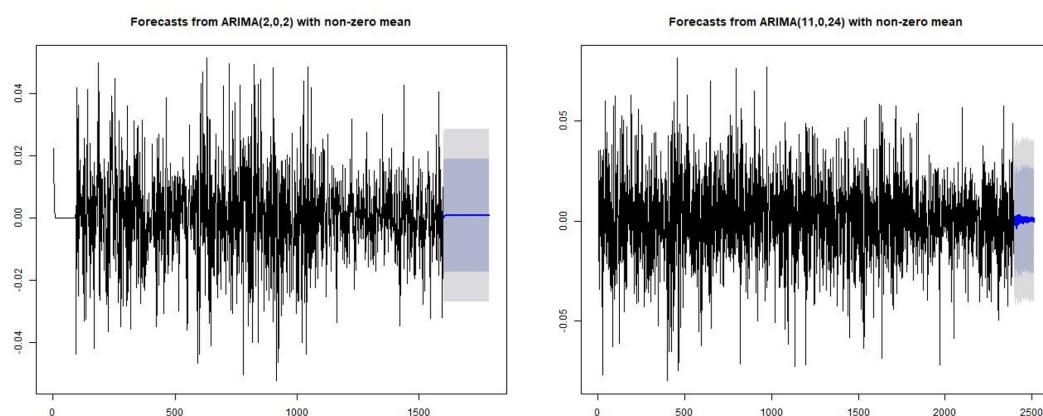


Figura 15: Forecast Japan e IAG

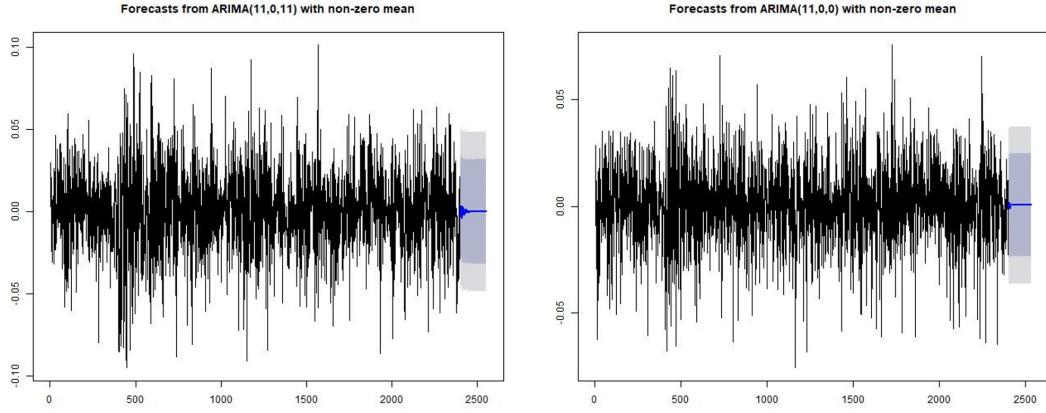


Figura 16: Forecast AirFrance-KLM e Lufthansa

Como critério de comparação, a fim de selecionar, enfim, o modelo cuja previsão apresenta a maior acurácia, foram utilizados os métodos apresentados na seção 3.2: *accuracy()* e *checkresiduals()*. Com esse objetivo, dentre as várias medidas de erro fornecidas pela função *accuracy()*, optou-se por ter a *RMSE* - *Root Mean Square Error* como medida de referência. Isso se justifica por seu caráter mais preciso, devido ao fato de utilizar a raiz quadrada do quadrado dos erros. Sendo assim, ao analisar o retorno desta função, procuramos o modelo cuja previsão (*trainning set*) possui o *RMSE* menor e mais próximo possível da série de controle (*testing set*). O retorno da função *checkresiduals()*, por sua vez, deve ser analisado buscando-se a série residual menos auto-correlacionada, com média mais próxima de zero, maior perfil de estacionaridade e histograma similar à uma distribuição normal. As figuras e tabelas a seguir exibem esses resultados.

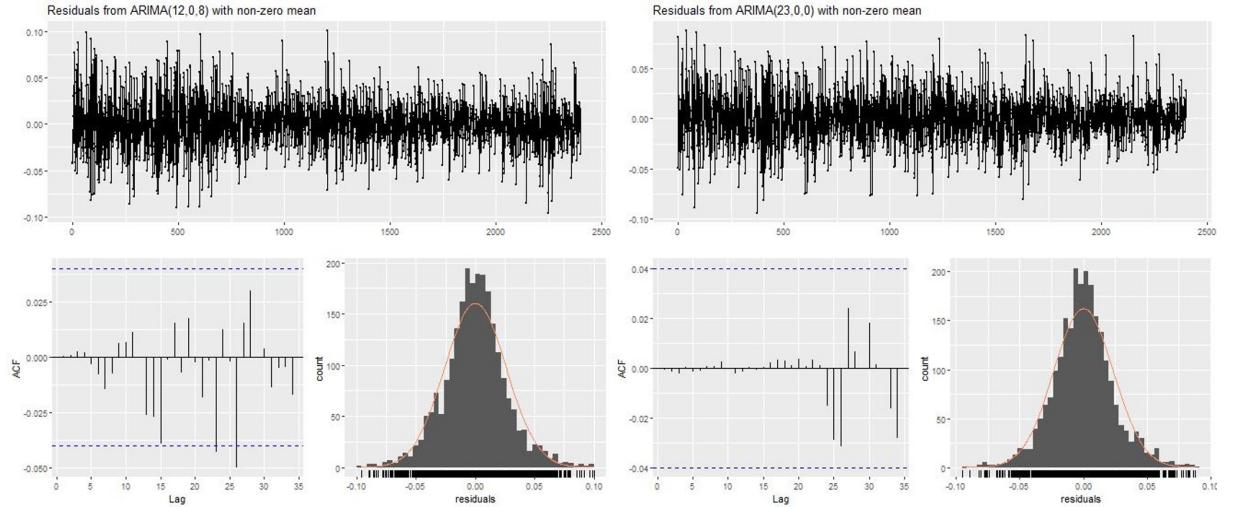


Figura 17: Checkresiduals American e United

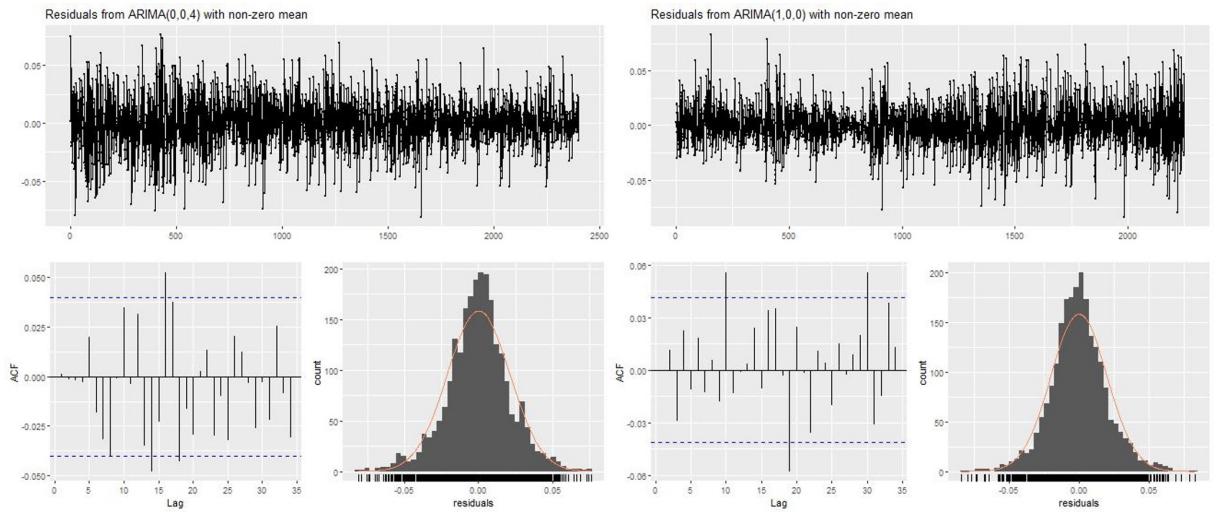


Figura 18: Checkresiduals Delta e Latam

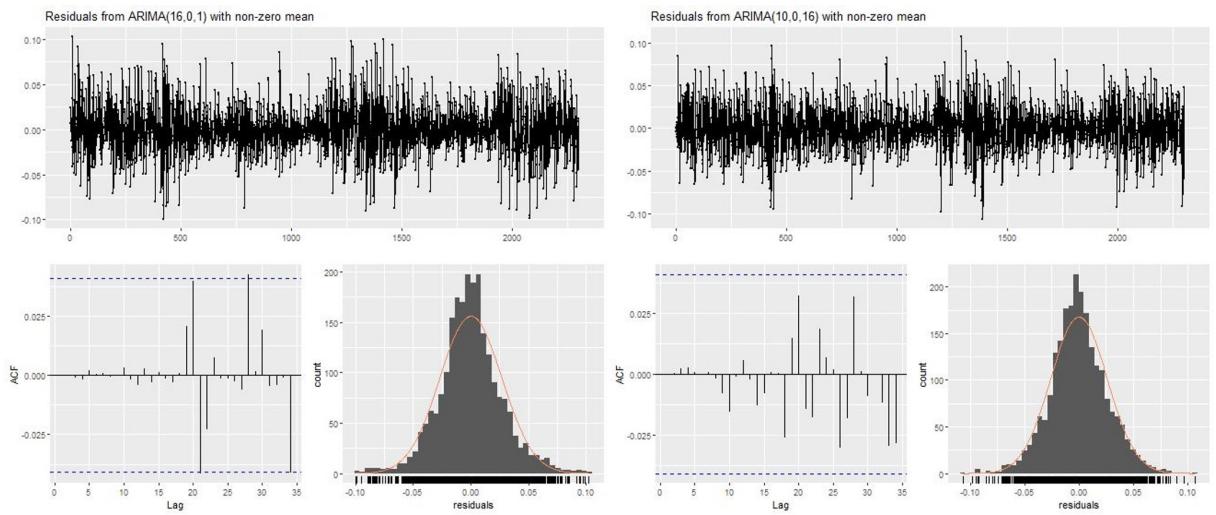


Figura 19: Checkresiduals China Southern e China Eastern

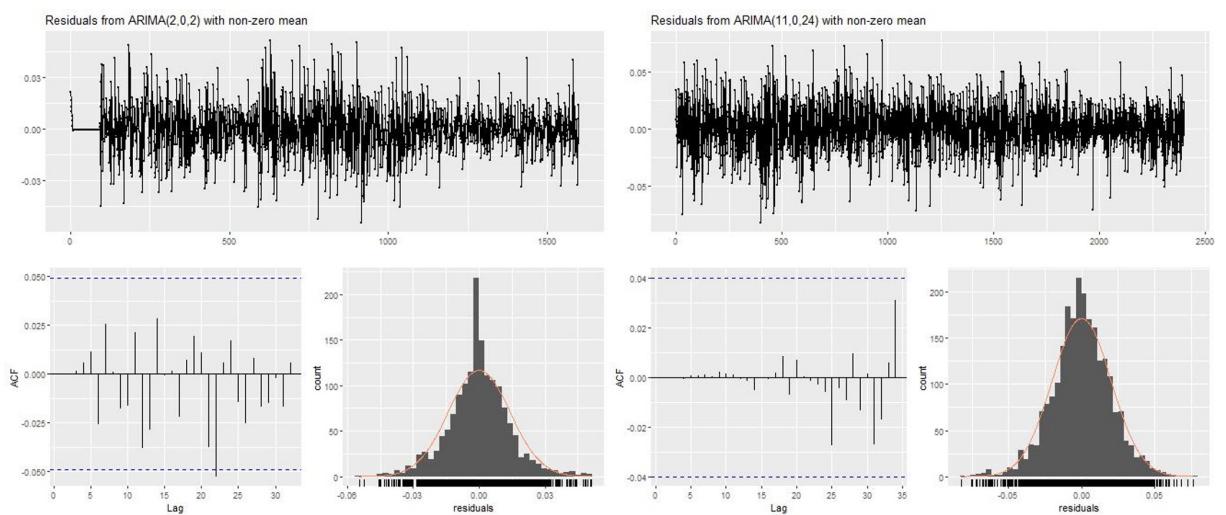


Figura 20: Checkresiduals Japan e IAG

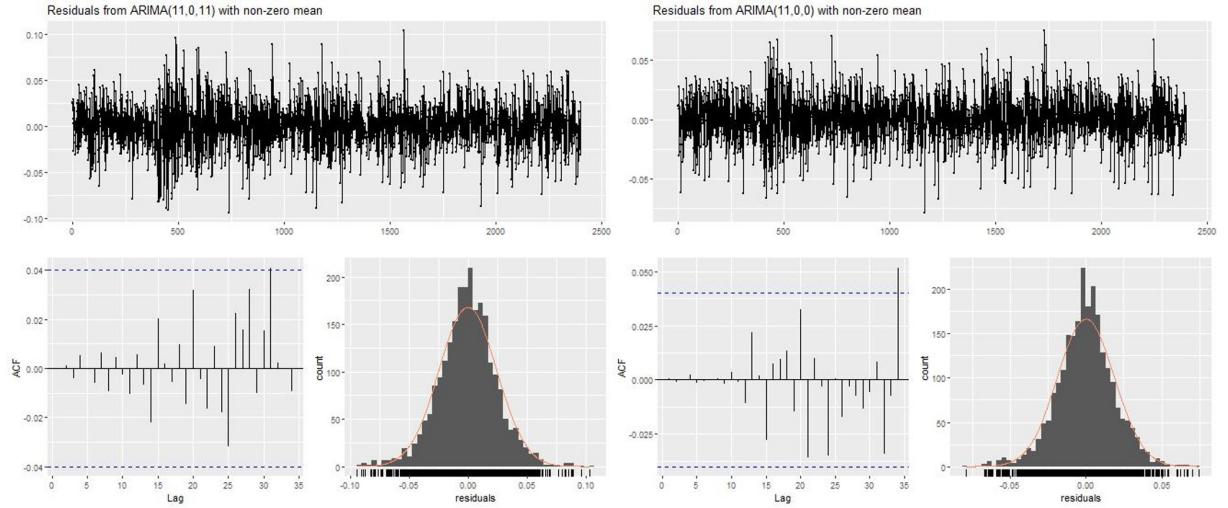


Figura 21: Checkresiduals AirFrance-KLM e Lufthansa

American (12,0,8)	RMSE
Training Set	0.02604072
Test Set	0.02271122

United (23,0,0)	RMSE
Training Set	0.02342442
Test Set	0.01319110

Tabela 14: RMSE American e United

Delta (0,0,4)	RMSE
Training Set	0.02119129
Test Set	0.01307335

Latam (1,0,0)	RMSE
Training Set	0.02009249
Test Set	0.02029929

Tabela 15: RMSE Delta e Latam

China Southern (16,0,1)	RMSE
Training Set	0.02700575
Test Set	0.02004664

China Eastern (10,0,16)	RMSE
Training Set	0.02601087
Test Set	0.02116064

Tabela 16: RMSE China Southern e China Eastern

Japan (2,0,2)	RMSE
Training Set	0.01408789
Test Set	0.01165704

IAG (11,0,24)	RMSE
Training Set	0.02021170
Test Set	0.01997025

Tabela 17: RMSE Japan e IAG

AirFrance-KLM (11,0,11)	RMSE
Training Set	0.02450562
Test Set	0.01983093

Lufthansa (11,0,0)	RMSE
Training Set	0.01878714
Test Set	0.01680258

Tabela 18: RMSE AirFrance-KLM e Lufthansa

Enfim, chegamos ao melhor modelo de cada companhia aérea, tendo como base a comparação dos diversos resultados comentados e exibidos anteriormente. São eles: **American** - (12,0,8); **United** - (23,0,0); **Delta** - (0,0,4); **Latam** - (1,0,0); **China Southern** - (16,0,1); **China Eastern** - (10,0,16); **Japan** - (2,0,2); **IAG** - (11,0,24); **AirFranceKLM** - (11,0,11); e **Lufthansa** - (11,0,0). Após toda esta análise e aplicação dos conceitos teóricos estudados na seção 3.1, podemos observar alguns fatos interessantes sobre a modelagem econométrica clássica, começando por dois grandes comentários acerca das figuras 12 a 16. Primeiramente, para o intervalo de confiança escolhido, as previsões se mostram um pouco frouxas, no sentido de que os retornos previstos pela modelagem *ARIMA* dentro de 80-95% de confiança podem assumir, praticamente, qualquer valor dentro da amplitude da série de retornos, o que é evidenciado pelos limites inferior e superior das faixas cinzas. Somado a isso, nota-se que a qualidade da precisão decai consideravelmente com o aumento do número de passos futuros previstos, o que é percebido pela tendência da curva azul, isto é, dos valores previstos, aproximarem-se de zero em todos os modelos ajustados. Tal caráter poderia ser fruto de modelos mal ajustados. Contudo, ao analisarmos os resíduos exibidos nas figuras 17 a 21, essa suposição parece ser negada, pois diversas características desejáveis são percebidas nas imagens, sejam funções *ACF* sem *lags* significativos, distribuição próxima à normal ou médias próximas de zero. Desse modo, tal fato parece decorrer das próprias limitações dos modelos *ARIMA*.

4 Conclusões

Como discutido na seção anterior, a modelagem *ARIMA* apresenta-se como uma ferramenta extremamente poderosa para o estudo e compreensão de séries temporais aplicadas no contexto financeiro. Isso se justifica por todos os conceitos teóricos cujo entendimento é necessário para a plena aplicação e compreensão deste modelo. Além de todo o amplo conhecimento pré-requisito para a formulação deste trabalho, cujo objetivo é a previsão de séries temporais, a modelagem *ARIMA* mostra-se útil em alguns contextos plausíveis da realidade. Tome por exemplo um investidor que pretende escolher entre algumas empresas aquela mais confiável de se tornar fim de seu investimento. Caso esta pessoa preocupe-se em especial com a evolução de curto prazo dos retornos dos ativos, utilizar uma aplicação que ajuste modelos *ARIMA* para dados históricos das empresas e preveja, dentro de um intervalo de confiança confortável, esse desenvolvimento, seria uma estratégia muito interessante. Por outro lado, para fins de análise a longo prazo, a modelagem *ARIMA* revela-se pouco adequada, justificado pelas conclusões discutidas na seção 3.3.2. Desse modo, outras ferramentas da econometria de séries financeiras devem mostrar-se mais confiáveis. O estudo da teoria das *ondaletas*, objetivo futuro e passo seguinte desta pesquisa, talvez apresente-se, de fato, como uma evolução mais poderosa e generalizada da modelagem clássica aqui abordada.

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] CRAN R Project, <https://cran.r-project.org/>
- [4] Github Repository, <https://github.com/MathNog/UndergraduateResearchProject.git>