



Universidade de Brasília
Instituto de Exatas
Departamento de Estatística

O modelo ARMA multivariado, cointegração e suas aplicações.

Alfredo Rossi Saldanha Cunha - 14/0015302
Gabriel Lobato Ramos - 13/0111309

Orientador:
Raul Matsushita

Brasília
9 de dezembro de 2017

Agradecimentos

Agradecemos ao nosso orientador, Raul Matsushita, pelas correções e incentivos durante a elaboração do trabalho. Agradecemos ao professor Jhames Sampaio por nos introduzir ao tema de Séries Temporais e finalmente ao professores George von Borries e Matria Teresa Costa pelas diversas matérias lecionadas durante o curso de graduação.

Resumo

Séries temporais multivariadas constituem uma forma de organização de dados estatísticos bastante comum. Este trabalho se propõe a efetuar um breve estudo acerca do modelo autoregressivo e de médias móveis (ARMA) multivariado (MARMA), como extensão do modelo univariado de Box e Jenkins. Em primeiro lugar, são introduzidos conceitos elementares como a função de autocorrelação serial, a correlação contemporânea, a função de autocorrelação cruzada e a estacionariedade. Em segundo lugar, o projeto trata de aspectos pertinentes ao caso multivariado, como a formulação do modelo ARMA na forma matricial e a introdução do conceito de cointegração. Uma breve revisão dos pacotes R que implementam os modelos MARMA também é apresentado no texto. Finalmente, para fins ilustrativos, séries históricas financeiras da Bolsa de valores de Nova Iorque (NYSE) são analisadas. Os resultados sugerem que a cointegração, na forma apresentada em livros textos, é uma construção (ou conceito) sensível à exposição de fontes não lineares de não estacionariedade, como, por exemplo, quebras estruturais.

Palavras-chave: modelo ARMA multivariado, séries multivariadas e cointegração.

Sumário

1	Introdução	1
2	Metodologia	2
2.1	Caso bivariado	2
2.1.1	Exemplo Bivariado com Dados	4
2.1.2	Exemplo de Cointegração	6
2.2	Propriedades de segunda ordem de séries temporais multivariadas	8
2.3	Estimação das funções de média e covariância	10
2.4	Modelo ARMA multivariado	10
2.4.1	Processo linear	10
2.4.2	Modelo ARMA multivariado	11
2.5	Ajustamento do modelo e previsão	12
2.6	Cointegração	13
2.6.1	Conceito	13
2.6.2	Estimação da Cointegração	14
2.6.3	Modelo de correção de erros	15
2.6.4	Testes para estacionariedade	15
2.7	Pacotes do software R	16
3	Resultados	17
3.1	Banco de Dados	17
3.2	Apresentação	17
3.2.1	Gráficos	18
3.2.2	Estacionariedade e tendencia	20
3.3	Análise Multivariada	21
3.3.1	Setor: Finanças	21
3.3.2	Setor: Commodities	22
3.3.3	Setor: Saúde	23
3.3.4	Setor: Serviços	24
3.3.5	Setor: Tecnologia	25
3.3.6	Conclusão da análise multivariada	26
3.4	Cointegração	26
3.4.1	Geral	28
3.4.2	Finanças	30
3.4.3	Commodities	31
3.4.4	Saúde	32
3.4.5	Serviços	33
3.4.6	Tecnologia	34
4	Conclusão	35

1 Introdução

Grosso modo, uma série temporal é um conjunto de observações em ordem cronológica. No caso univariado, tais observações são valores escalares que constituem a variável quantitativa de interesse. Na extensão para o caso multivariado, a série temporal é representada por uma sequência cronológica de vetores X_1, \dots, X_t, \dots , em que cada vetor X_t é composto por duas ou mais variáveis de interesse. O estudo de modelos multivariados para a análise de séries temporais proporciona um leque bastante amplo de aplicações em diversas áreas. Por exemplo, em climatologia, Privalsky (2015) utiliza modelos autoregressivos multivariados para avaliar possíveis relações entre duas séries temporais oceânicas observadas em regiões distintas. Em sismologia, Kakizawa, et al. (1998) consideram o problema de discriminar observações temporais produzidas por terremotos daquelas que são produzidas por explosões nucleares. Em engenharia, Baldeon, A. e Roberto P. (2004) usaram modelos de séries temporais multivariados para identificar sistemas mecânicos. Em economia, Holland, M. e Palaia, D. (2010) analisam a taxa de câmbio e a paridade do poder de compra no Brasil com o auxílio de séries temporais multivariadas, e em finanças, De Paula (2006) estudou mercados financeiros emergentes com modelos temporais multivariados. Portanto, o estudo dessa temática proporciona uma importante agregação de conhecimentos para o profissional da área de estatística.

Existe uma diversidade de abordagens e modelos para a análise de séries temporais multivariadas. Por exemplo, em perspectiva Bayesiana, modelos lineares dinâmicos multivariados são tratados por West e Harrison (1997). Sob a ótica de modelos estruturais, uma abordagem via modelos em espaço de estados é apresentada por Harvey (1989). Como extensão natural do modelo (univariado) de Box e Jenkins, os autores Wei (2006) e Brockwell e Davis (1996) consideram uma classe de filtros lineares multivariados conhecidos como MARMA (ARMA multivariado) ou VARMA (ARMA vetorial). Há também modelos multivariados que proporcionam uma análise no domínio da frequência também abordado por Wei (2006).

Por não ser possível contemplar essa diversidade de modelos neste trabalho, nosso foco se concentrará no modelo MARMA (ou VARMA). O outro foco deste projeto, além da modelagem com o modelo MARMA, é o estudo de cointegração entre as séries. Sendo assim, se duas séries são cointegradas ao longo do tempo, o comportamento de uma permite entender o movimento da outra. Como a cointegração é um equilíbrio ao longo do tempo, é possível usar várias séries para verificar sua cointegração em determinado período de tempo. Não necessariamente todas as séries devem estar no mesmo instante de tempo. Assim, podemos estudar quais fatores ou períodos podem possuir essa relação de cointegração e quais não possuem. Portanto, são muitas as possibilidades para análise. Contudo para o desenvolvimento deste projeto serão utilizadas todas as séries no mesmo instante de tempo e o comprimento do período será alterado para ver se o equilíbrio para cada setor é atingido nos mesmos tamanhos ou se existem distinções.

Para a realização deste projeto definem-se os seguintes objetivos específicos: estudar as propriedades estocásticas de segunda ordem processos lineares multivariados, os estimadores pertinentes aos modelos MARMA, a análise de cointegração e suas variações. Além disso, usar as implementações computacionais já constituídas no R e aplicar os métodos estudados em um conjunto real de dados, gerando conclusões acerca da metodologia estudada e da aplicação em si.

2 Metodologia

Este capítulo será dedicado a apresentar e explicar a metodologia desse modelo. Dado esse objetivo, o capítulo foi dividido em 6 seções. Na Seção 2.1 se apresenta um exemplo de uma série bivariada para entender o comportamento e alguns conceitos iniciais. Na Seção 2.2 são apresentadas as propriedades de segunda ordem e estacionariedade enquanto que a Seção 2.3 mostra a estimação dos conceitos abordados na seção anterior. A Seção 2.4 apresenta o modelo ARMA multivariado e suas implicações. Logo após, a Seção 2.5 aborda sobre o ajustamento do modelo e previsão. Para finalizar, a Seção 2.6 aborda sobre cointegração mostrando desde sua identificação até sua estimação e correção e finalmente a Seção 2.7 apresenta as ferramentas computacionais utilizadas.

No desenvolvimento do capítulo são consideradas m séries temporais $X_{it}, t = 0, \pm 1, \pm 2, \dots$ em que i representa a i -ésima série ($i=1, \dots, m$) e t o instante da série. Será considerado que os segundos momentos das séries sejam finitos, ou seja, $EX_{it}^2 < \infty$.

2.1 Caso bivariado

Como ilustração, considere os preços das ações dos bancos JP Morgan Chase & Co. e Morgan Stanley na bolsa de valores de Nova York, entre 18 de Maio de 2012 e 12 de Maio de 2017, totalizando 1254 preços em cada série. Os preços são os valores dados no momento de fechamento da Bolsa (ver Figura 1). Na Figura 2 estão representados graficamente os preços do banco Jp Morgan Chase & Co. contra os preços do Morgan Stanley. O ponto vermelho representa os preços do primeiro dia e o ponto azul do ultimo dia.



Figura 1: Preços das ações dos bancos JP Morgan Chase & Co. e Morgan Stanley.

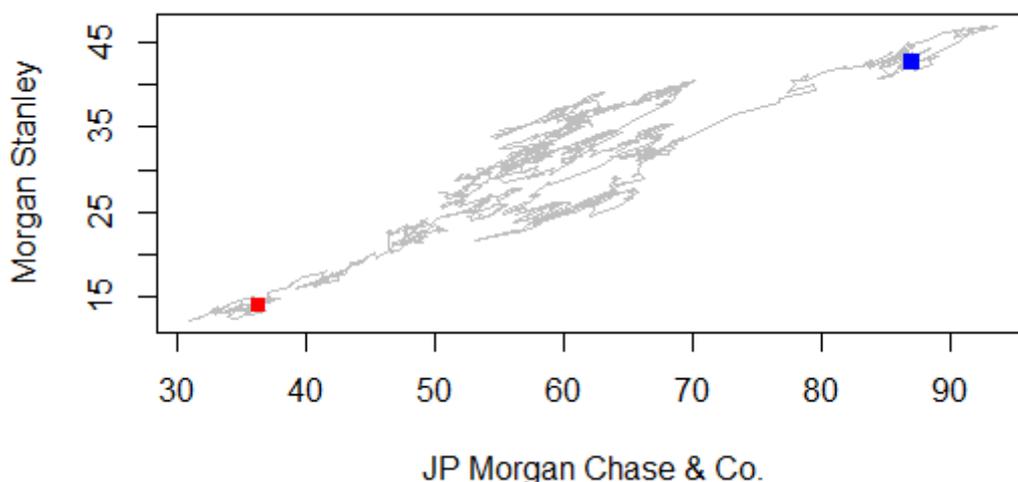


Figura 2: Mapa dos preços das ações dos bancos JP Morgan Chase & Co. e Morgan Stanley.

Assim consideramos o vetor $\mathbf{X}_t = (X_{t1}, X_{t2})'$ no qual X_{t1} , é a série correspondente os valores das ações do banco JP Morgan Chase & Co no tempo t e X_{t2} a série correspondente os valores das ações no tempo t do banco Morgan Stanley. O tempo t corresponde a data (dia, mês e ano). Neste primeiro momento, as definições básicas, define-se o vetor de médias

$$\mu_t = E\mathbf{X}_t = \begin{bmatrix} EX_{t1} \\ EX_{t2} \end{bmatrix},$$

e a matriz de covariância

$$\Gamma(t+h, t) = Cov(\mathbf{X}_{t+h}, \mathbf{X}_t) = \begin{bmatrix} Cov(X_{t+h,1}, X_{t1}) & Cov(X_{t+h,1}, X_{t2}) \\ Cov(X_{t+h,2}, X_{t1}) & Cov(X_{t+h,2}, X_{t2}) \end{bmatrix}.$$

A série X_t é fracamente estacionária se a media μ_t e $\Gamma(t+h, t)$ forem independentes do tempo t , para todo h . Então no caso estacionário se tem

$$\mu = E\mathbf{X}_t = E \begin{bmatrix} \mu_1 \\ \mu_2 \\ , \end{bmatrix}$$

em que $E(X_{t1}) = \mu_1$ e $E(X_{t2}) = \mu_2$ e a matriz de covariância

$$\Gamma(h) = Cov(\mathbf{X}_{t+h}, \mathbf{X}_t) = \begin{bmatrix} \gamma_{11}(h) & \gamma_{12}(h) \\ \gamma_{21}(h) & \gamma_{22}(h) \end{bmatrix}.$$

Os elementos da diagonal são as autocovariâncias de X_{t1} e X_{t2} e os outros elementos as covariâncias entre $X_{t+h,i}$ e X_{tj} em que $i \neq j$. Nesse caso se tem $\gamma_{12}(h) = \gamma_{12}(-h)$. Como se sabe o estimador natural para o vetor de médias μ , das observações X_1, \dots, X_n é

$$\bar{\mathbf{X}} = (1/n) \sum_{t=1}^n \mathbf{X}_t,$$

e o estimador de $\Gamma(h)$ é

$$\hat{\Gamma}(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-h} (\mathbf{X}_{t+h} - \bar{\mathbf{X}}_n)(\mathbf{X}_t - \bar{\mathbf{X}}_n)' & , 0 \leq h \leq n-1, \\ \hat{\Gamma}(-h)' & , -n+1 \leq h < 0. \end{cases}$$

No qual $\bar{\mathbf{X}}_{\mathbf{n}}$ é $\frac{1}{n} \sum_{t=1}^n \mathbf{X}_t$. Logo a correlação $\rho_{ij}(h)$ entre $X_{t+h,i}$ e X_{tj} é estimada por

$$\hat{\rho}_{ij}(h) = \hat{\gamma}_{ij}(h) / [\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0)]^{(1/2)}.$$

Intuitivamente se $j = i$, então $\hat{\rho}_{ij}(h)$ é a autocorrelação estimada da i -ésima série.

A seguir apresentaremos um exemplo para estudar o caso estacionário bivariado.

2.1.1 Exemplo Bivariado com Dados

Os preços de ações são dependentes do tempo t e suas covariâncias e correlações não dependem apenas de uma diferença entre tempos (h), portanto não são estacionários (a Figura 1 mostra tendências ao longo do tempo). Será usado o retorno dos preços das ações, ou seja, $\Delta X_t = X_t - X_{t-1}$, que por sua vez é suficiente para deixar os dados estacionários, ao aplicar um teste de estacionariedade que será discutido posteriormente. Os mesmos gráficos para os retornos são apresentados nas Figuras a seguir:

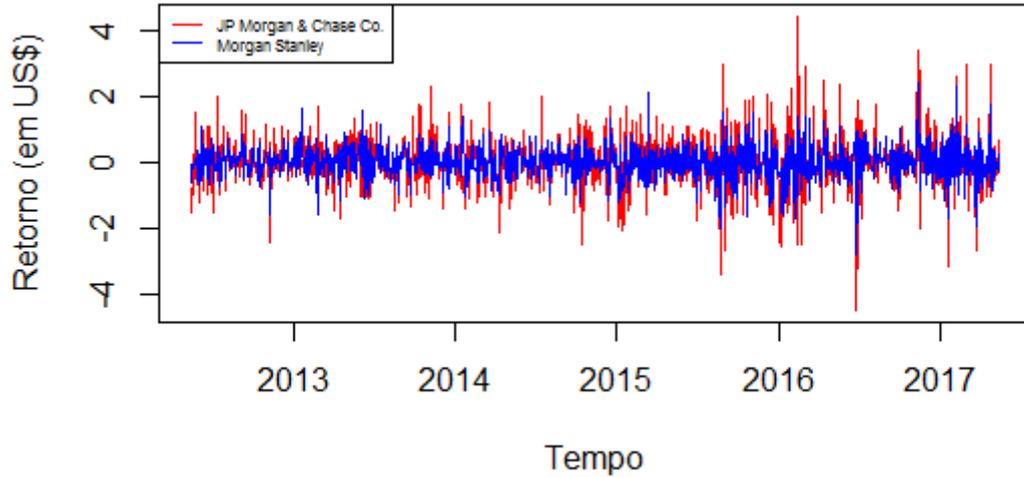


Figura 3: Retorno das ações dos bancos JP Morgan Chase & Co. e Morgan Stanley.

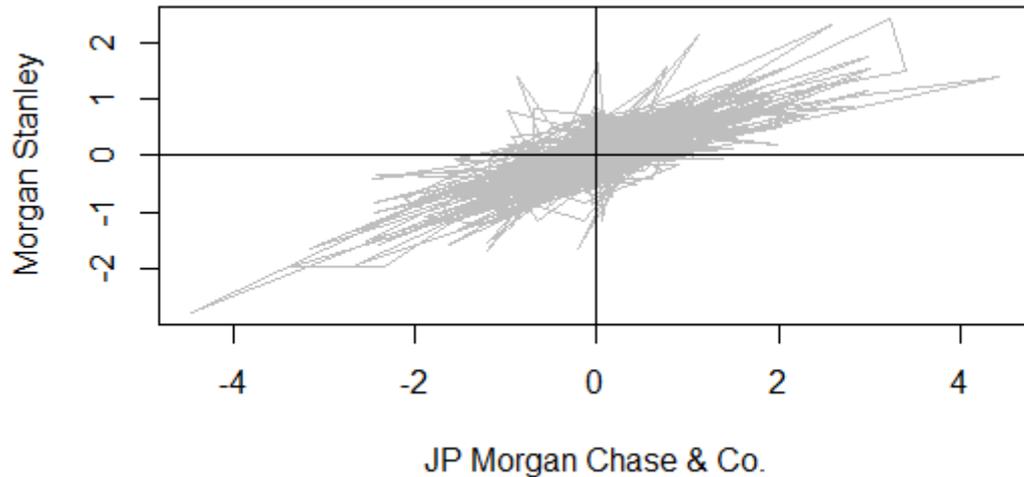


Figura 4: Mapa do retorno das ações dos bancos JP Morgan Chase & Co. e Morgan Stanley.

Na figura 4 é observado o comportamento dos retornos das duas ações usadas e percebe-se que a maioria dos valores fica entorno de zero. Pelo mesmo gráfico pode-se notar uma correlação positiva entre os retornos, a medida que o retorno aumenta se um aumenta, o outro retorno também tende a aumentar.

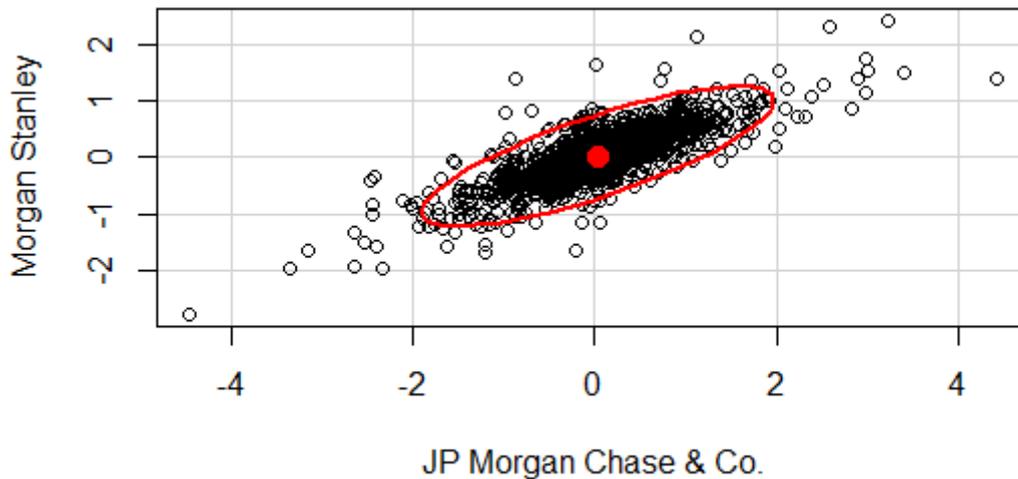


Figura 5: Mapa do retorno das ações dos bancos JP Morgan Chase & Co. e Morgan Stanley. com o limite de 95% dos retornos.

A Figura 5 mostra o mapa dos retornos com uma elipse que indica 95% dos retornos, ou seja, dentro da elipse estão 95% dos valores observados. Portanto, dentro da elipse se concentram a maioria dos pontos mostrando uma relação, enquanto que os pontos fora da elipse possuem uma magnitude mais elevada.

Abaixo estão as estimativas amostrais do vetor de médias para a série bivariada, seguidas das matrizes de covariância e correlação. Vetor de médias

$$\bar{\mathbf{X}} = E\mathbf{X}_t = \begin{bmatrix} 0,040 \\ 0,023 \end{bmatrix},$$

matrizes de covariância, no lag 0 e 5 (para mostrar como a covariância converge para 0 conforme aumenta o número de diferença entre tempos, lag)

$$\hat{\Gamma}(0) = \begin{bmatrix} 0,620 & 0,319 \\ 0,319 & 0,256 \end{bmatrix}, \hat{\Gamma}(-5) = \begin{bmatrix} -0,011 & -0,015 \\ -0,009 & -0,010 \end{bmatrix}, \hat{\Gamma}(5) = \begin{bmatrix} -0,011 & -0,009 \\ -0,015 & -0,010 \end{bmatrix},$$

e matrizes de correlação

$$\hat{\rho}(0) = \begin{bmatrix} 1 & 0,799 \\ 0,799 & 1 \end{bmatrix}, \hat{\rho}(-5) = \begin{bmatrix} -0,017 & -0,038 \\ -0,023 & -0,041 \end{bmatrix}, \hat{\rho}(5) = \begin{bmatrix} -0,017 & -0,023 \\ -0,038 & -0,041 \end{bmatrix}.$$

Após esta breve introdução de séries temporais usando um exemplo bivariado, introduziremos a ideia de cointegração por meio de um exemplo.

2.1.2 Exemplo de Cointegração

Consideram-se duas séries temporais não estacionárias X_{t1} e X_{t2} , aparentemente correlacionadas, como é o caso no exemplo da seção 3.1.1. É possível encontrar um parâmetro β_t que, multiplicado por uma das duas séries, torna a distância entre as séries relativamente constante ao longo do tempo. Usando a relação $X_{t2} - \beta X_{t1} = u_t$, a variável u_t será um processo que expressa a relação entre as duas séries. Quando a serie u_t for de fato estacionária, pode-se concluir que X_{t1} e X_{t2} são cointegradas de ordem 1. Em termos práticos significa que existe de fato uma correlação entre as duas séries.

Uma série é dita cointegrada de ordem p , escrito da forma $I(p)$ caso seja preciso fazer p diferenças até que a série fique estacionária. No caso de cointegração, pode-se dizer que X_{t1} e X_{t2} são cointegradas de ordem $CI(d, p)$, se as ambas as séries forem integradas de ordem d . Se existir um β para o qual $X_{t1} - \beta X_{t2} = u_t$, então as duas séries são cointegradas de ordem $d - p$. Concluindo-se que se $CI(1, 1)$ então u_t será estacionária, pois $d - p = 1 - 1 = 0$, são necessárias zero diferenças para que u_t seja estacionária.

O β é estimado por meio de mínimos quadrados, ou seja a partir da fórmula de uma regressão linear simples $X_{t2} = \beta_0 + \beta_1 X_{t1} + e_t$. Então após um simples manipulação algébrica tem-se $X_{t2} - \beta_1 X_{t1} = \beta_0 + e_t$. Fazendo com que $u_t = \beta_0 + e_t$. Então se as duas variáveis forem correlacionadas sabe-se que a regressão de X_{t1} e X_{t2} resultará em β_0 e resíduos e_t constantes com distribuição Normal e média zero. Fazendo com que u_t seja estacionário.

Usando as duas séries da seção no qual X_{t1} são os valores das ações dos preços do Banco JP Morgan Chase & Co e X_{t2} os valores do Banco Morgan Stanley. Na Figura 6 pode-se perceber que as duas séries são correlacionadas. Porém há uma "nuvem" de pontos no meio do gráfico de dispersão, indicando que diversos valores do preço da ação Morgan Stanley para o mesmo preço da ação da JP Morgan. Isso segue que, devido a esses pontos as séries não sejam linearmente cointegráveis. Fazendo uma regressão linear simples de X_{t1} sobre X_{t2} , tem-se que o resíduos (série cointegrada) não se comportam como série estacionária, como é visto na Figura 7. Após a manipulação algébrica tem-se a série cointegrada não estacionária:

$$X_{t2i} - 0.5753 \times X_{t1} = e_t - 4.0311 = u_t$$

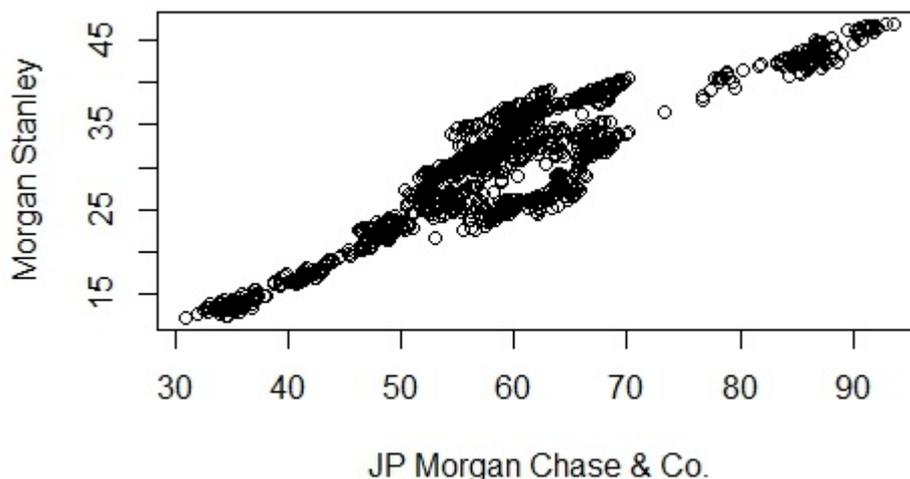


Figura 6: Gráfico de dispersão de preços das ações dos bancos JP Morgan Chase & Co. e Morgan Stanley.

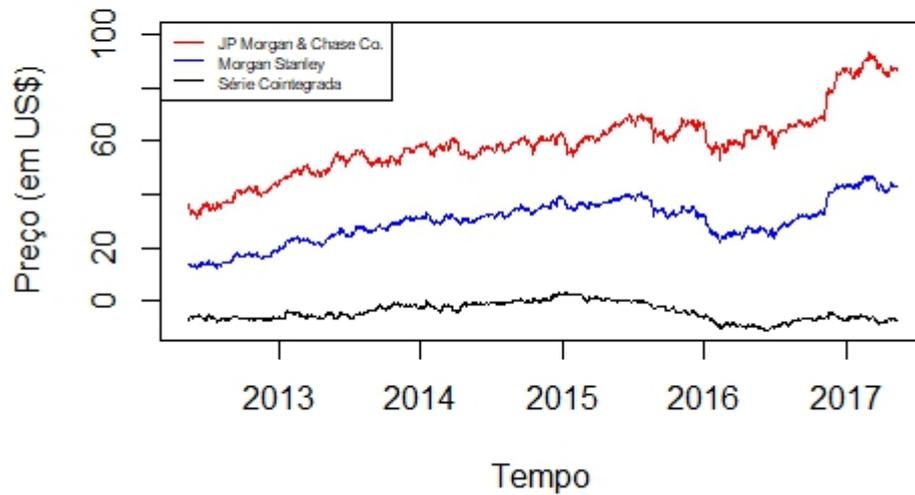


Figura 7: Gráfico séries dos preços da ações dos bancos JP Morgan Chase & Co. e Morgan Stanley e de sua série conintegrada.

Agora será feito o mesmo procedimento para as duas séries porém apenas do dia 18 de maio de 2015 até dia 31 de dezembro de 2014, Figura 8. A série cointegrada é representada pela equação:

$$X_{t2} - 0.8399 \times X_{t1} = e_t - 16.9306 = u_t$$

Após realizado o teste para estacionariedade de Dickey-Fuller que será explicado nas seções seguintes, a série cointegrada mostra-se estacionária com nível de significância de 10%.

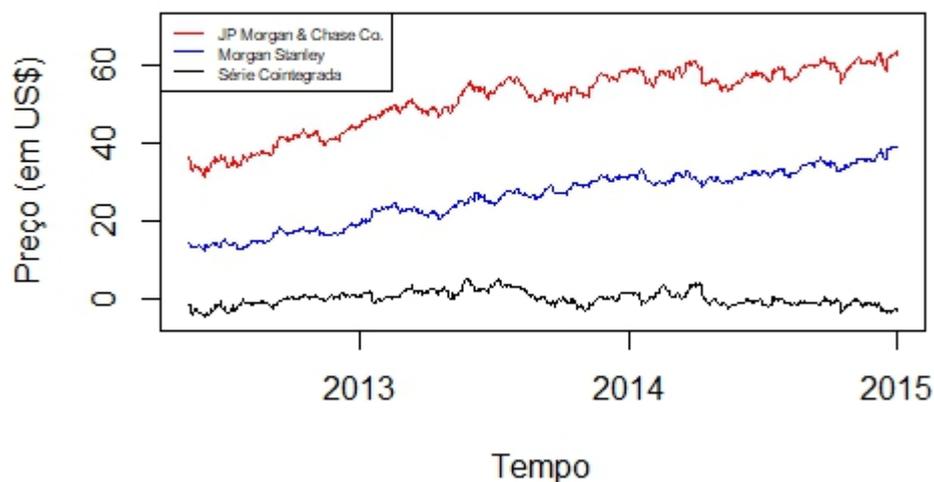


Figura 8: Gráfico das séries dos preços da ações dos bancos JP Morgan Chase & Co. e Morgan Stanley e de suas séries conintegradas de 18/05/2012 a 31/12/2014.

2.2 Propriedades de segunda ordem de séries temporais multivariadas

Nesta seção serão discutidas propriedades básicas de séries multivariadas (sendo ou não estacionárias). No caso m -variado as médias são dadas pelo vetor:

$$\mu_{\mathbf{t}} = E\mathbf{X}_{\mathbf{t}} = \begin{bmatrix} \mu_{t1} \\ \mu_{t2} \\ \vdots \\ \mu_{tm} \end{bmatrix},$$

e as covariâncias são dadas pela matriz

$$\Gamma(t+h, t) = \begin{bmatrix} \gamma_{11}(t+h, t) & \dots & \gamma_{1m}(t+h, t) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(t+h, t) & \dots & \gamma_{mm}(t+h, t) \end{bmatrix}.$$

na qual a covariância entre a série i com a série j é

$$\gamma_{ij}(t+h, t) = Cov(\mathbf{X}_{\mathbf{t+h,i}}, \mathbf{X}_{\mathbf{t,j}}).$$

A matriz de covariâncias fornece uma medida de dependência não somente dentro da série mas também entre as séries e em tempos distintos.

Além dessas características de segunda ordem, existem definições e conceitos relevantes para que seja compreendido de melhor maneira o modelo ARMA multivariado. Entre essas definições se encontram a de estacionariedade, ruído branco (R.B.) e ruído independente identicamente distribuído (i.i.d.).

Definição 2.2.1: Um processo é estacionário caso suas características em determinado instante de tempo sejam as mesmas que em outro tempo posterior (ou anterior). Existem dois tipos de estacionariedade, fracamente estacionário e estritamente estacionário. Neste trabalho são vistos processos multivariados fracamente estacionários. Portanto ao se falar de estacionariedade, se estará fazendo referência ao conceito de fracamente estacionário, em que são atendidas as seguintes suposições:

- $\mu_x(t)$ independente de t ;
- $\gamma_x(t+h, t)$ independente de t para todo h .

Elas representam uma extensão da estacionariedade no caso das séries univariadas.

A partir das exigências necessárias é visto que, para um processo multivariado estacionário, a média é constante ao longo do tempo, e a covariância depende apenas da diferença entre tempo h . Portanto, para séries estacionárias, o vetor de médias se reduz a:

$$\mu = E\mathbf{X}_{\mathbf{t}} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix},$$

e a matriz de covariâncias

$$\Gamma(h) = \begin{bmatrix} \gamma_{11}(h) & \dots & \gamma_{1m}(h) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(h) & \dots & \gamma_{mm}(h) \end{bmatrix}.$$

Para uma série multivariada estacionária, a função $\gamma_{ij}(h)$ para qualquer h é chamada de função de covariância cruzada entre as séries X_{ti} e X_{tj} , para i diferente de j .

Existem quatro propriedades básicas necessárias e suficientes para ela ser função de covariância cruzada de um processo estacionário:

- $\gamma_{ij}(h) = \gamma'_{ij}(-h)$;
- $|\gamma_{ij}(h)| < [\gamma_{ii}(0)\gamma_{jj}(0)]^{(1/2)}$;
- $\gamma_{ij}(\cdot)$ é a função de autocovariância (FAC); e
- $\sum_{j,k=1}^n a'_j \Gamma(j-k) a_k \geq 0$ para todo $n \in \{1, 2, \dots\}$ e $a_1, \dots, a_n \in R^m$.

Define-se a matriz de correlação como

$$R(h) = \begin{bmatrix} \rho_{11}(h) & \dots & \rho_{1m}(h) \\ \vdots & \ddots & \vdots \\ \rho_{m1}(h) & \dots & \rho_{mm}(h) \end{bmatrix},$$

em que

$$\rho_{ij}(h) = \gamma_{ij}(h) / [\gamma_{ii}(0)\gamma_{jj}(0)]^{(1/2)}.$$

Definição 2.2.2: Uma série multivariada Z_t é chamada de ruído i.i.d. se as observações forem i.i.d. com média 0 e matriz de covariância Σ :

$$Z_t \sim i.i.d.(0, \Sigma),$$

na qual

$$\Gamma(h) = \begin{cases} \Sigma, & h = 0, \\ 0, & c.c. \end{cases}$$

Como a condição de ser i.i.d. é uma suposição forte, porque muitas vezes não se conhece a distribuição geradora das observações da série, define-se o ruído branco multivariado da seguinte forma.

Definição 2.2.3: Uma série multivariada Z_t chamada de ruído branco (R.B.) com média 0 e matriz de covariância Σ , é representado como

$$Z_t \sim RB(0, \Sigma),$$

em que, se Z_t for estacionário

$$\Gamma(h) = \begin{cases} \Sigma, & h = 0, \\ 0, & c.c. \end{cases}$$

Nota-se que o R.B. não faz suposição sobre a distribuição de Z_t , tornando-o mais flexível de se utilizar na prática, dado que ele não requer suposições acerca de distribuição do processo.

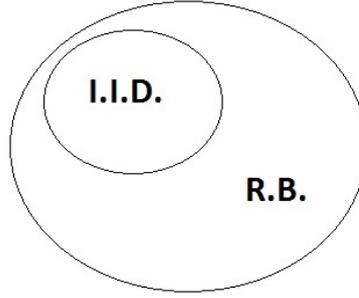


Figura 9: Diagrama das definições de série I.I.D. e Ruído Branco

2.3 Estimação das funções de média e covariância

Agora serão discutidos estimadores da média, covariância e correlação para uma série multivariada estacionária.

Um estimador não viesado para o vetor de médias baseado nas observações de uma amostra de tamanho n para cada série é a média univariada de X_i . Então a média para a i -ésima série é dada por

$$\bar{X}_i = \frac{\sum_{t=1}^n X_{ti}}{n}.$$

O estimador para a matriz de covariância $\Gamma(h)$, é apenas a forma matricial do caso univariado.

$$\hat{\Gamma}(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-h} (\mathbf{X}_{t+h} - \bar{\mathbf{X}}_n)(\mathbf{X}_t - \bar{\mathbf{X}}_n)' & , 0 \leq h \leq n-1, \\ \hat{\Gamma}(-h)' & , -n+1 \leq h < 0. \end{cases}$$

A partir dos estimadores acima, o estimador da correlação entre a i -ésima e a j -ésima série é

$$\hat{\rho}_{ij}(h) = \frac{\hat{\gamma}_{ij}(h)}{\sqrt{\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0)}}.$$

2.4 Modelo ARMA multivariado

2.4.1 Processo linear

Antes de abordar o modelo ARMA multivariado será definido um processo linear multivariado, que tem a seguinte definição:

Definição 2.4.1.1: O processo multivariado X_t é um processo linear se tiver a representação

$$X_t = \sum_{j=-\infty}^{\infty} C_j Z_{t-j}, \quad Z_t \sim RB(0, \Sigma),$$

em que C_j é uma sequência somável de matrizes $m \times m$, deve ser somável a sequência de coeficientes. É importante ressaltar que quando $j = 0$ temos $C_0 Z_t$ na qual C_0 é a matriz identidade $m \times m$ e dessa forma X_t se reduz a Z_t . No caso estacionário, o processo linear com média 0 a função de covariância fica em função apenas dos coeficientes C_j e da matriz de covariâncias do ruído branco Z_t , da seguinte forma

$$\Gamma(h) = \sum_{j=-\infty}^{\infty} C_{j+h} \Sigma C_j', \quad h = 0, 1, \dots$$

2.4.2 Modelo ARMA multivariado

Como extensão imediata da classe de modelos autoregressivos de médias móveis, define-se o modelo ARMA multivariado (MARMA) como:

$$\Phi_p(B)X_t = \Theta_q(B)Z_t,$$

em que

$$\Phi_p(B) = \Phi_0 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_p B^p,$$

e

$$\Theta_q(B) = \Theta_0 - \Theta_1 B - \Theta_2 B^2 - \dots - \Theta_q B^q$$

são os polinômios autorregressivos e de médias móveis de ordem p e q respectivamente e o operador B indica a defasagem no tempo, portanto B^p indica p passos anteriores e não a potência de B . Como Φ_0 e Θ_0 devem ser matrizes não singulares de dimensão $m \times m$, considera-se aqui que $\Phi_0 = \Theta_0 = I$ e $Z_t \sim RB(0, \Sigma)$.

Quando $p = 0$ o modelo se reduz ao modelo de média móvel multivariado (MMA) de ordem q , MMA(q),

$$X_t = Z_t - \Theta_1 Z_{t-1} - \dots - \Theta_q Z_{t-q},$$

e quando $q = 0$ o modelo se reduz um modelo autorregressivo (MAR) de ordem p , MAR(p),

$$X_t = Z_t + \Phi_1 X_{t-1} + \dots + \Phi_p X_{t-p}.$$

O processo é estacionário se o módulo das raízes do polinômio $\Phi_p(B)$ estiverem fora do círculo unitário. Nesse caso, temos

$$X_t = \Psi_p(B)Z_t,$$

de forma que

$$\begin{aligned} \Psi_p(B) &= [\Phi_p(B)]^{-1} \Theta_q(B) \\ &= \sum_{s=0}^{\infty} \Psi_s B^s, \end{aligned}$$

de tal forma que $\sum_{s=0}^{\infty} \Psi_s^2 < \infty$. O processo é invertível se as raízes do polinômio $\Theta_q(B)$ estiverem fora do círculo unitário, tendo assim

$$\Pi(B)X_t = Z_t,$$

em que

$$\Pi(B) = [\Theta_q(B)]^{-1} \Phi_p(B) = I - \sum_{s=1}^{\infty} \Pi_s B^s,$$

no qual $\sum_{s=1}^{\infty} \Pi_s < \infty$.

Para compreender o que são processos invertíveis e causais, seguem as próximas definições.

Definição 3.4.2.1: Um processo X_t é invertível se $\sum_{s=1}^{\infty} \Pi_s < \infty$ (componentes absolutamente somáveis) e, $Z_t = \sum_{j=1}^{\infty} \Pi_s X_{t-j}, \forall t$. Invertibilidade é equivalente a condição, $\det \Theta(Z) \neq 0$ para todo $z \in C$ de maneira que $|z| \leq 1$. Portanto, enquanto que caso multivariado a matriz Π_j de dimensão $m \times m$ seja somável, o caso univariado apenas os escalares π precisam ser.

Definição 3.4.2.2: O processo X_t é causal se existir uma matriz ψ_j , $m \times m$, tal para que o processo possa ser escrito na forma linear

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad \text{para todo } t,$$

no qual, os componentes de ψ_j devem ser somáveis. A condição de causalidade é equivalente a condição

$$\det\Phi(Z) \neq 0 \text{ para todo } z \in C \text{ de tal forma que } |Z| \leq 1.$$

As matrizes ψ_j são provenientes das equações

$$\psi_j = \Theta_j + \sum_{k=1}^{\infty} \Phi_k \psi_{j-k}, \quad j = 0, 1, \dots$$

na qual $\Theta_0 = I$, $\Theta_j = 0$ se, $j > q$, $\Phi_j = 0$ se, $j > p$ e $\Psi_j = 0$ se, $j < 0$.

A matriz de covariância de um processo ARMA multivariado causal pode ser encontrada como

$$\Gamma(h) = \sum_{j=0}^{\infty} \psi_{h+j} \Sigma \psi_j', \quad h = 0, 1, \dots$$

2.5 Ajustamento do modelo e previsão

Para o processo de estimação, o logaritmo da função de verossimilhança para um processo ARMA(p,q) multivariado gaussiano e média zero é dado por

$$\ln L(\Phi, \Theta, \Sigma | X) = -\frac{nm}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \text{tr} \Sigma^{-1} S(\Phi, \Theta)$$

na qual

$$S(\Phi, \Theta) = \sum_{t=1}^n Z_t Z_t',$$

e

$$Z_t = X_t - \Phi_1 X_{t-1} - \dots - \Phi_p X_{t-p} + \Theta_1 Z_{t-1} + \dots + \Theta_q Z_{t-q}.$$

Como consequência de não existir observações antes de $t = 0$ (X_t não é observado) é necessário um ajuste. Para evitar esse problema, é trabalhado com o logaritmo da função de verossimilhança condicional (L_*) expresso por

$$\ln L_*(\Phi, \Theta, \Sigma | X) = -\frac{n-p}{2} \ln |(2\pi)^m \Sigma| - \frac{1}{2} \text{tr} \Sigma^{-1} S_*(\Phi, \Theta),$$

e

$$S_*(\Phi, \Theta) = \sum_{t=p+1}^n Z_t Z_t'.$$

em que se assume que os termos Z_{p+1-q}, \dots, Z_p sejam iguais a 0.

A partir da função de maximização condicional anterior são encontradas propriedades relevantes discutidas por Nicholls (1966,1977) e Anderson (1980). Considerando η o vetor de parâmetros e $\hat{\eta}$ o seu estimador de máxima verossimilhança, foi mostrado que $\hat{\eta}$ é assintoticamente não viesado, consistente e

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} N(0, \Sigma_{\hat{\eta}}),$$

no qual é indicado convergência em distribuição para a distribuição normal com média 0 e $\Sigma_{\hat{\eta}}$ igual a

$$\Sigma_{\hat{\eta}} = [-E(\frac{\partial^2 \ln L_*(\eta)}{\partial \eta \partial \eta'})]^{-1}.$$

Se o processo \mathbf{X}_t não possui um vetor de média igual a zero, tem-se $\hat{\mu} = \bar{\mathbf{X}}$. Assim é feita a subtração da média amostral pela populacional e mostrado que

$$\sqrt{n}(\bar{\mathbf{X}} - \mu) \xrightarrow{d} N(0, \Sigma_{\bar{\mathbf{X}}}),$$

no qual

$$\Sigma_{\bar{\mathbf{X}}} = [\Phi(1)]^{-1} \Theta(1) \Sigma \Theta(1)' [\Phi'(1)]^{-1} \text{ e } \hat{\mu} = \bar{X}$$

é assintoticamente independente de \bar{X} . Este resultado é equivalente à estimação conjunta de μ com os demais parâmetros por máxima verossimilhança.

Depois de estimado o modelo, é necessária uma análise de resíduos para verificar a adequação do ajuste.

$$\hat{Z}_t = X_t - \hat{\Phi}_1 X_{t-1} + \dots + \hat{\Phi}_p X_{t-p} + \hat{\Theta}_1 \hat{Z}_{t-1} + \dots + \hat{\Theta}_q \hat{Z}_{t-q}.$$

Para o modelo ser adequado, os resíduos devem ser ruído branco, as matrizes de autocorrelação dos resíduos devem ser insignificantes e não apresentar algum padrão de comportamento.

A previsão para o modelo ARMA multivariado para l observações futuras, assim como no caso univariado, é recursivamente dado por

$$\hat{X}_n(l) = \Phi_1 \hat{X}_n(l-1) + \dots + \Phi_p \hat{X}_n(l-p) - \Theta_1 E(Z_{n+l-1}) - \dots - \Theta_q E(Z_{n+l-q}),$$

Sendo que $\hat{X}_n(j) = X_{n+j}$ para $j \leq 0$, $E(Z_{n+j}) = Z_{n+j}$ para $j \leq 0$; e $E(Z_{n+j}) = 0$ para $j > 0$.

Ou seja, para valores futuros o termo de erro é considerado 0 e as previsões sempre são baseadas nos valores anteriores.

A matriz de variância-covariância do resíduo é dada por

$$\text{var}(e_n(l)) = \sum_{j=0}^{l-1} \Psi_j \Sigma \Psi_j'$$

e Ψ_0 é matriz identidade e e a diferença entre o valor do modelo ajustado e o valor real da série. Além disso, l indica a quantidade de passos a frente para a previsão. Para realizar intervalos de confiança para a previsão, são baseados nas estimativas pontuais e na variância calculada para cada observação futura. Naturalmente é esperado um aumento desse intervalo conforme aumenta o número de observações previstas, em vista que os erros vão se acumulando ao longo do tempo.

A seguir será discutido com detalhes a parte de cointegração que foi exemplificada anteriormente.

2.6 Cointegração

2.6.1 Conceito

Uma série não estacionária X_t é dita cointegrada de ordem d , escrito de forma $I(d)$, se a $(d-1)$ -ésima diferença na série for não estacionária, mas a (d) -ésima diferença é estacionária. Seguindo essa notação uma série $I(0)$ é estacionária.

Um vetor de séries temporais \mathbf{X}_t de dimensões $(m \times 1)$ é cointegrado de ordem (d, b) . Escrito da forma $CI(d, b)$ no qual $0 < b \leq d$, se cada componente do vetor é $I(d)$ mas combinações lineares das séries $\beta \mathbf{X}_t$ forem $I(d-b)$ para um vetor não nulo e constante β , de dimensão $(m \times 1)$. Nesse caso o vetor β é o vetor cointegrador. O vetor β não é unico pois também será um vetor cointegrador $c\beta$ para qualquer constante não nula c . O caso mais comum é o qual $d = b = 1$.

Para esclarecer tomaremos o vetor bidimensional $AR(1)$:

$$\begin{bmatrix} X_{t,1} \\ X_{t,2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\phi & 0 \end{bmatrix} \begin{bmatrix} X_{t-1,1} \\ X_{t-1,2} \end{bmatrix} + \begin{bmatrix} Z_{t,1} \\ Z_{t,2} \end{bmatrix}$$

Nota-se que $X_{t,1}$ é um passeio aleatório,

$$(1 - B)X_{t,1} = Z_{t,1}.$$

$X_{t,2} = -\phi X_{t-1,1} + Z_{t,2}$ e depois de diferenciar tem-se:

$$(1 - B)X_{t,2} = -\phi(1 - B)X_{t-1,1} + (1 - B)Z_{t,2},$$

$$(1 - B)X_{t,2} = -\phi Z_{t-1,1} + Z_{t,2} - Z_{t-1,2} (*).$$

Assim (*) é $MA(1)$. Em que cada componente da série é um processo $I(1)$. Tomando $\beta' = \begin{bmatrix} -\phi & 1 \end{bmatrix}$

Mostrando assim o que ocorre no exemplo visto na seção 3.1.2, no qual a combinação linear de duas séries não estacionárias é uma série estacionária.

Se o vetor de séries X_t de dimensão $(m \times 1)$ conter mais que apenas duas séries, cada uma sendo $I(1)$, então podem existir $k (< m)$ vetores $(m \times 1)$ linearmente independentes $\beta_1, \beta_2, \beta_3, \dots, \beta_k$, de tal forma que $\beta \mathbf{X}_t$, seja um vetor de séries, de dimensão $(k \times 1)$ estacionárias. No qual β é

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},$$

chamado de vetor cointegrador.

2.6.2 Estimação da Cointegração

Para estimar o vetor cointegrador de maneira que o vetor de séries não estacionárias (\mathbf{X}_t) seja estacionário ($\beta \mathbf{X}_t$) é utilizada a ideia de regressão linear sendo estimada por mínimos quadrados. O método basicamente consiste em gerar uma regressão na qual cada série é uma variável, estimar os coeficientes do modelo e verificar se os resíduos do modelo são $I(0)$ ou $I(1)$.

Para realizar a estimação, é usualmente utilizado que o primeiro elemento do vetor de cointegração β_1 será o denominador comum para dividir todo o vetor de cointegração de maneira que o primeiro elemento do vetor seja 1 e os demais $\beta_2/\beta_1, \beta_3/\beta_1, \beta_4/\beta_1, \dots$. Dessa forma, é possível escrever o seguinte modelo de regressão para $X_{t,1}$

$$X_{t,1} = \phi_1 X_{t,2} + \dots + \phi_{k-1} X_{t,k} + u_t,$$

em que cada ϕ_i é a razão entre seu respectivo coeficiente e o primeiro (β_{i+1}/β_1), $i = 1, \dots, k$. O termo u_t é o erro do modelo. Após estimar o valor de cada coeficiente, é verificado se os resíduos estimados possuem raiz unitária ou não. Se possuir raiz unitária ($I(1)$), então X_t não é cointegrada. Além disso, $\hat{\phi}_i$ não possui distribuição assintótica t-student caso os resíduos não sejam estacionários. Entretanto, se \hat{u}_t for $I(0)$ as séries são cointegradas e o vetor cointegrador é dado por $\hat{\beta}' = [1, \hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_k]$ e não acarretará em problemas de estimação em cada ϕ_i , porque pode-se considerar a existência de distribuição assintótica t-student. Um dos problemas do método de regressão para estimação do vetor de cointegração é que não

necessariamente ao realizar a regressão $X_{t,1} = \phi_1 X_{t,2} + \dots + \phi_{k-1} X_{t,k} + u_t$, serão obtidas as mesmas conclusões ao se realizar a regressão $X_{t,2} = \phi_1 X_{t,1} + \dots + \phi_{k-1} X_{t,k} + u_t$, por exemplo. Porém, usualmente os resultados são consistentes e devido a facilidade da metodologia é muito utilizado para estimação de cointegração. Outro problema ocorre se as séries X_t são cointegradas. Nesse caso, é necessário fazer a estimação do modelo de correção de erros, porque podem existir perturbações que precisam ser corrigidas. A próxima subseção explicará sobre o modelo de correção de erros.

2.6.3 Modelo de correção de erros

O modelo de correção de erros (MCE), da mesma maneira que na estimação da cointegração, é uma regressão de uma variável sobre outras, usando o método dos mínimos quadrados para estimar os coeficientes, em que o “erro de equilíbrio” refere-se aos resíduos estimados da equação cointegrada. O MCE é uma forma de conciliar o comportamento de curto e longo prazo de determinada variável. A regressão utilizada na seção anterior é uma equação cointegrante e os ε_t são os resíduos estimados dessa equação, ou seja, a equação representa a relação de longo prazo entre as variáveis. Para representar o equilíbrio de curto e longo prazo, é utilizada a seguinte equação.

$$\Delta X_{t,1} = \phi_0 + \phi_1 \Delta X_{t,2} + \dots + \phi_{k-1} \Delta X_{t,k} + \phi_k u_{t-1} + \varepsilon_t,$$

no qual Δ é o operador de diferenças, u_{t-1} é o valor defasado do termo de erro da regressão, ε_t é o termo de erro aleatório da regressão de correção de erro.

Dessa forma, $\Delta X_{t,1}$ depende das diferenças das outras séries e também do termo de erro de equilíbrio. Como todas as variáveis do MCE são estacionárias (cada série é $I(1)$), o modelo pode ser eficientemente estimado por mínimos quadrados e o teste de hipóteses padrão é válido. Se as variáveis forem cointegradas, a estimativa de ϕ_k será estatisticamente significativa.

2.6.4 Testes para estacionariedade

Existem diversas formas de verificar se uma série temporal é estacionária. Uma das formas é por meio de gráfico, sendo o mais utilizado o gráfico de função de autocorrelação que mostra a autocorrelação da série para vários lags. Outra forma é mediante análise das condições de estacionariedade vistas nas seções anteriores (Definição 3.2.1), com base em testes de hipóteses como o de Dickey-Fuller.

O teste de Dickey-Fuller determina se u_t é estacionária ou não, ou seja, se existe raiz unitária ou não,

$$\begin{aligned} H_0 : \Phi &= 1, \\ H_1 : \Phi &< 1. \end{aligned}$$

No caso deste trabalho, o teste será utilizado para testar se os resíduos de um modelo de regressão linear simples são estacionários. Assim considerando a forma

$$e_t = \Phi e_{t-1} + \sum_{j=1}^{p-1} \Phi \Delta e_{t-j} + a_t,$$

que pode ser reescrita como

$$\Delta e_t = \lambda e_{t-1} + \sum_{j=1}^{p-1} \Phi \Delta e_{t-j} + a_t,$$

desejando-se testar $H_0 : \lambda = 0$ vs $H_1 : \lambda < 0$. No qual a_t é a componente aleatória do modelo dos resíduos.

Existem valores críticos tabelados desenvolvidos por Engle e Granger (1987), para a estatística $T = \lambda/S_\lambda$. Deve-se rejeitar a hipótese se os valores obtidos em T forem menores que o valores críticos. Se H_0 é rejeitado, λ seja menor do que zero, conclui-se que as séries primitivas podem ser cointegradas.

2.7 Pacotes do software R

Para realizar as análises, tanto para cointegração quanto para o ARMA multivariado, são necessários de pacotes com funções previamente criadas para a análise de dados. Para esse trabalho foram utilizados os seguintes pacotes: readxl, MTS, tseries.

O primeiro pacote tem como principal propósito a importação dos dados. Como são muitas planilhas, esse pacote utiliza a própria extensão xlsx nativa dos arquivos para importá-los.

O segundo (MTS), pelo próprio nome em ingles (Multivariate Time Series) é utilizado para análise multivariada do modelo ARMA. Portanto, ele possui funções para a geração dos gráficos (MTSplot), matriz de correlação cruzada (ccm), estimar a ordem de um modelo autoregressivo (VARorder) ou de médias móveis multivariado (VMAorder), estimar o modelo (VARMA) e verificar o ajuste (MTSdiag).

O último pacote é muito completo e possui muitas funções construídas para os mais diversos problemas, aqui ele será utilizado para realizar o teste aumentado de Dickey- Fuller (adf.test) para análise inicial das séries e na metodologia da cointegração.

Foi desenvolvido um algoritmo para otimizar o processo de análise de dados, de forma que os gráficos fossem gerados automaticamente.

3 Resultados

Este capítulo será dedicado a explicar os dados e utilizar todas as técnicas mencionadas no capítulo anterior no conjunto de dados obtidos, de forma que complemente a explicação teórica por meio da aplicação em um conjunto real de informação. Dessa forma, o capítulo foi dividido em 4 seções. Na Seção 3.1 se apresenta a explicação sobre o conjunto de dados, sua coleta, variáveis disponíveis e período de tempo que será feita a análise. Na Seção 3.2 são as séries e seu respectivos comportamentos, na Seção seguinte (3.3) é feita a análise multivariada das séries e finalmente na última Seção (3.4) é feita a cointegração entre as séries. Para todos os testes realizados. Foi considerado o nível de significância de 10%.

3.1 Banco de Dados

Os Estados Unidos são uma das maiores economias do planeta porque além de sua influência política e econômica mundial, variações na sua economia alteram o comportamento e valor de várias empresas nacionais e internacionais em várias áreas de atuação. Para analisar o comportamento do mercado empresarial de capital aberto americano, foram obtidas as séries em 5 áreas: Finanças, Commodities, Saúde, Serviços e Tecnologia. Dentro de cada área foram escolhidas 2 empresas.

Tabela 1: Séries por setor.

Setor	Séries
Finanças	Morgan Stanley (<i>ms</i>)
	J.P Morgan & Chase Co. (<i>jmp</i>)
Commodities	Exxon Mobil Co. (<i>xom</i>)
	Chevron Co. (<i>cvx</i>)
Saúde	Johnson Unsp BDR (<i>jnj</i>)
	UnitedHealth Group Inc. (<i>unh</i>)
Serviços	Amazon Inc. (<i>amzn</i>)
	Walt Disney Co. (<i>dis</i>)
Tecnologia	Microsoft Co. (<i>msft</i>)
	Facebook Inc. (<i>fb</i>)

As séries foram obtidas do site do Yahoo Finance no período de 18 de Maio de 2012 até 12 de Maio de 2015. No banco de dados existem as variáveis data, preço de abertura, preço de fechamento, máximo diário, mínimo diário e volume. Dessa forma, será trabalhado com o preço de fechamento diário das empresas e para realizar os testes de cointegração foram determinados intervalos de tempo de 180 dias onde é testado a cointegração e esse intervalo de tempo é acrescido os 20 posteriores e retirado os 20 dias anteriores, mantendo sempre o intervalo de 180 dias e fazendo a progressão no tempo da data inicial até o último intervalo. A ideia de se utilizar esse intervalo de tempo previamente estabelecido é para minimizar o impacto na resposta da cointegração para diferentes níveis de preço de duas empresas ao longo de um período muito maior (equilíbrio de curto e longo prazo).

3.2 Apresentação

Primeiro serão os gráficos das séries e logo em seguida uma tabela mostrando se as séries são estacionárias e quais tendências seguem.

3.2.1 Gráficos

Os gráficos serão apresentados por setor. A maioria das séries possuem tendencia crescente e não possuem quebras estruturais ou mudanças bruscas nos valores. No setor saúde é um caso no qual aproximadamente na metade do período de tempo total a séries se cruzam a Johnson Unsp BDR passa a ter menores valores que a UnitedHealth Group Inc. A duas séries de empresas no ramo da tecnologia também possuem alguns cruzamentos mas nada significativo. O setor de Commodities não apresenta uma tendencia clara.

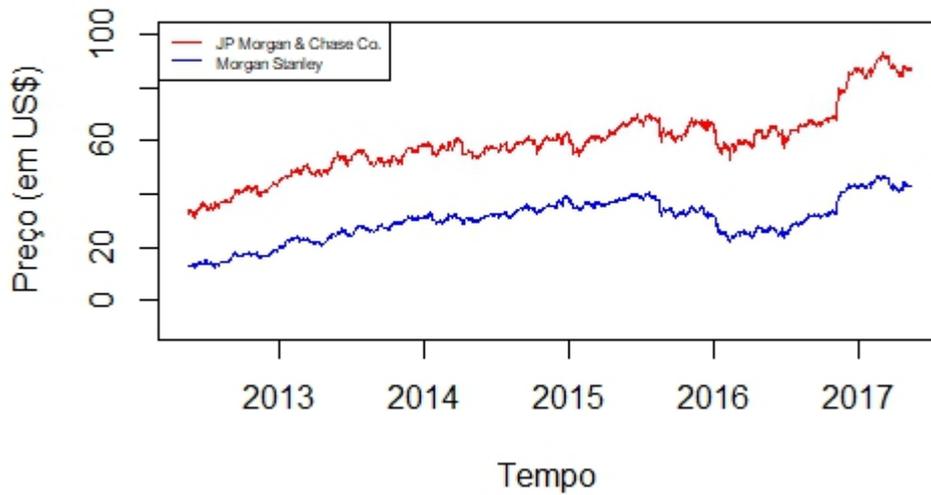


Figura 10: Gráfico da séries JP Morgan Chase & Co. e Morgan Stanley..

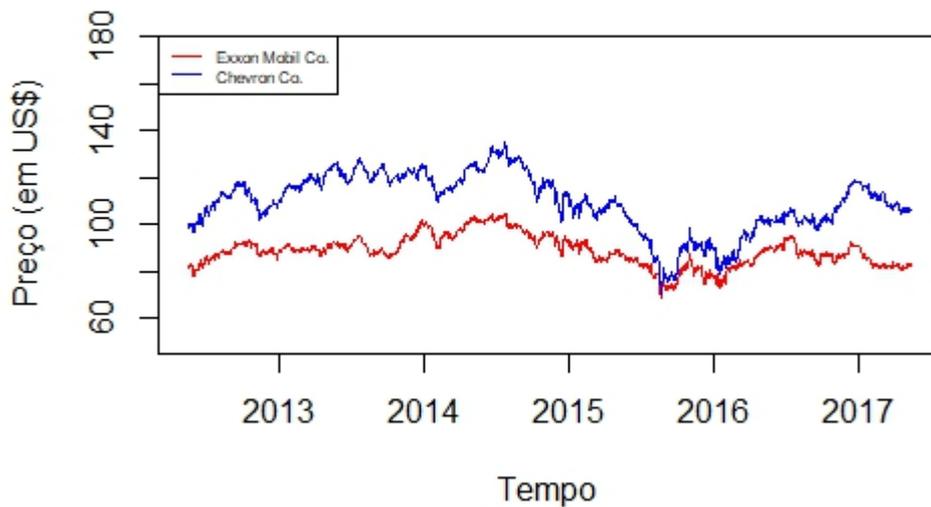


Figura 11: Gráfico da séries Exxon Mobil Co. e Chevron Co..

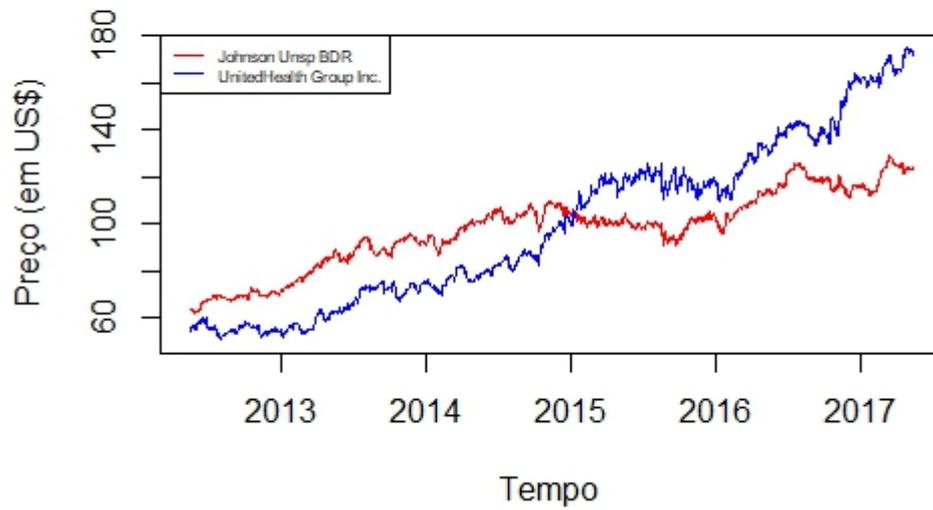


Figura 12: Gráfico da séries Johnson Unsp BDR e UnitedHealth Group Inc..

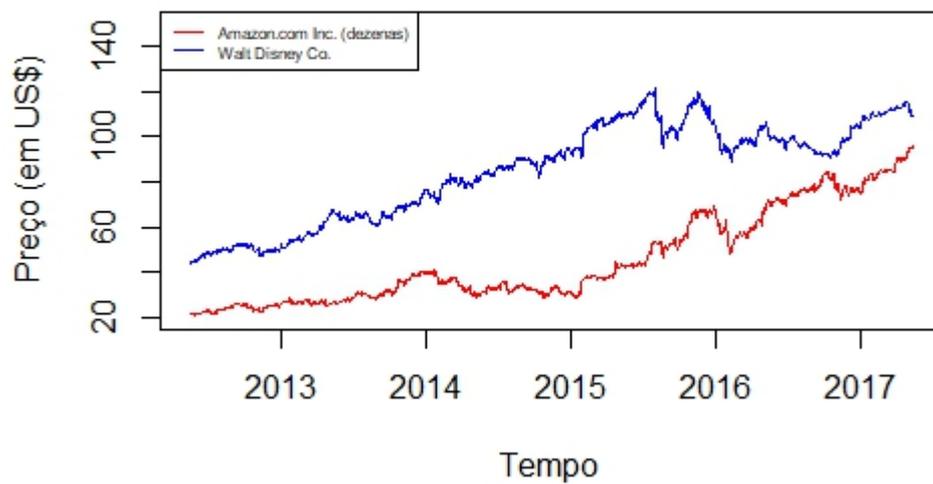


Figura 13: Gráfico da séries Amazon Inc. e Walt Disney Co..

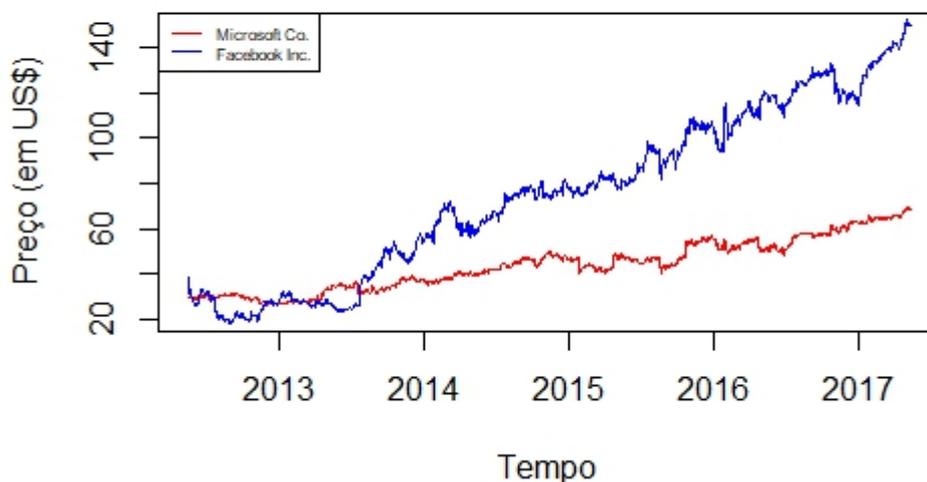


Figura 14: Gráfico da séries Microsoft Co. e Facebook Inc..

3.2.2 Estacionariedade e tendencia

A tabela abaixo resume o que é observado nos gráficos. Após aplicar o teste de Dickey-Fuller nas séries nota-se que nove das dez séries é naturalmente estacionária mas todas as séries são $I(1)$ ou seja, elas são estacionárias após uma diferença. A série da empresa Exxon Mobil apresenta um P-Valor que indica estacionariedade mas para alguns intervalos a mesma série é não estacionária e assim ela poderá ser cointegrada.

Tabela 2: Teste de Dickey-Fuller.

Série	P-Valor	
	Sem diferenças	1 Diferença
J.P Morgan Chase & Co.	0,817	<0,01
Morgan Stanley	0,984	<0,01
Exxon Mobil Co.	0,010*	<0,01
Chevron Co.	0,782	<0,01
Johnson Unsp BDR	>0,99	<0,01
United Health Group Inc.	0,091	<0,01
Amazon.com Inc.	>0,99	<0,01
Walt Disney Co.	>0,99	<0,01
Microsoft Co.	0,304	<0,01
Facebook Inc.	0,308	<0,01

* Existem intervalos nos quais a série é não estacionária.

Tabela 3: Estacionariedade e tendência das séries.

Série	Estacionária		Tendnecia
	Sem diferenças	1 Diferença	
J.P Morgan Chase & Co.	Não	Sim	Crescente
Morgan Stanley	Não	Sim	Crescente
Exxon Mobil Co.	Sim	Sim	Estacionária
Chevron Co.	Não	Sim	Decrescente
Johnson Unsp BDR	Não	Sim	Crescente
United Health Group Inc.	Não	Sim	Crescente
Amazon.com Inc.	Não	Sim	Crescente
Walt Disney Co.	Não	Sim	Crescente
Microsoft Co.	Não	Sim	Crescente
Facebook Inc.	Não	Sim	Crescente

3.3 Análise Multivariada

Nesta seção será feita a análise multivariada das séries, após uma diferença, ou seja, séries estacionárias. As séries serão analisadas por setor. Os modelos escolhidos foram os modelos multivariados AR(1). Esses tipo de modelo é suficiente para avaliar o tipo de série estudada, modelos com mais parâmetros não elevam a qualidade da modelagem de maneira significativa (AIC e BIC pouco se alteram) de modo a compensar o esforço algébrico e computacional requerido para lidar com mais parâmetros.

As matrizes de autocorrelação dos erros do modelo foram insignificante e não apresenta nenhum padrão de comportamento.

O teste T para identificar se o parâmetro é significativo, tido nível de significância de 10% foi usado. Os parâmetros não significativos serão apresentados iguais a zero na matriz Φ .

Os gráficos de correlação entres as séries tem como base o tempo $t = 0$ pois as séries não representam nenhuma quebra ou mudança de tendencia que levaria a correlação e autocorrelação das séries a mudar de maneira significativa se escolhido outro tempo. O *lag* escolhido foi o de 120 que indica o período de 6 meses.

3.3.1 Setor: Finanças

A tabela 4 apresentada as estimativas para o modelo VAR(1) do setor de financeiro. Nesse caso os parâmetros constantes foram os únicos significativos.

Modelo multivariado AR(1) para o setor de Finanças:

$$\begin{bmatrix} \Delta X_{t,jpm} \\ \Delta X_{t,ms} \end{bmatrix} = \begin{bmatrix} 0,0438 \\ 0,0235 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta X_{t-1,jpm} \\ \Delta X_{t-1,ms} \end{bmatrix}.$$

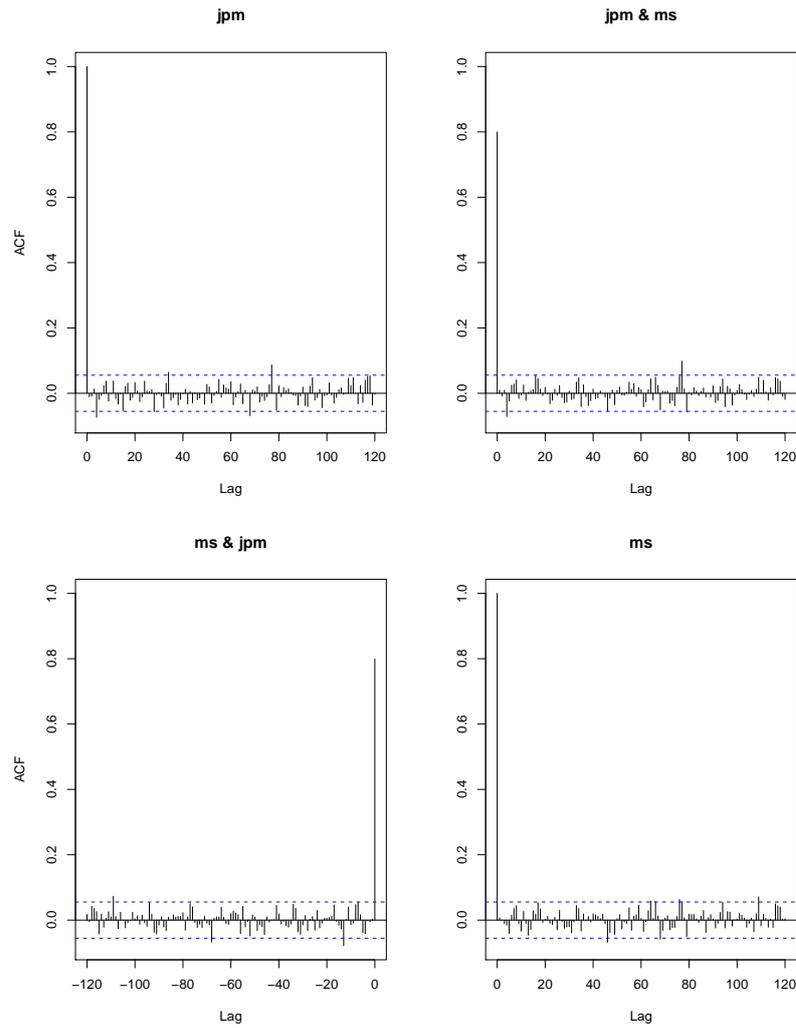


Figura 15: Matriz de correlação do setor de Finanças.

3.3.2 Setor: Commodities

Diferentemente do setor anterior, neste caso tem-se que nenhum parâmetro é significativo, ou seja, o modelo AR(1) para essa extensão temporal é insignificativo.

Modelo multivariado AR(1) para o setor de Commodities:

$$\begin{bmatrix} \Delta X_{t,xom} \\ \Delta X_{t,cvx} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta X_{t-1,xom} \\ \Delta X_{t-1,cvx} \end{bmatrix}.$$

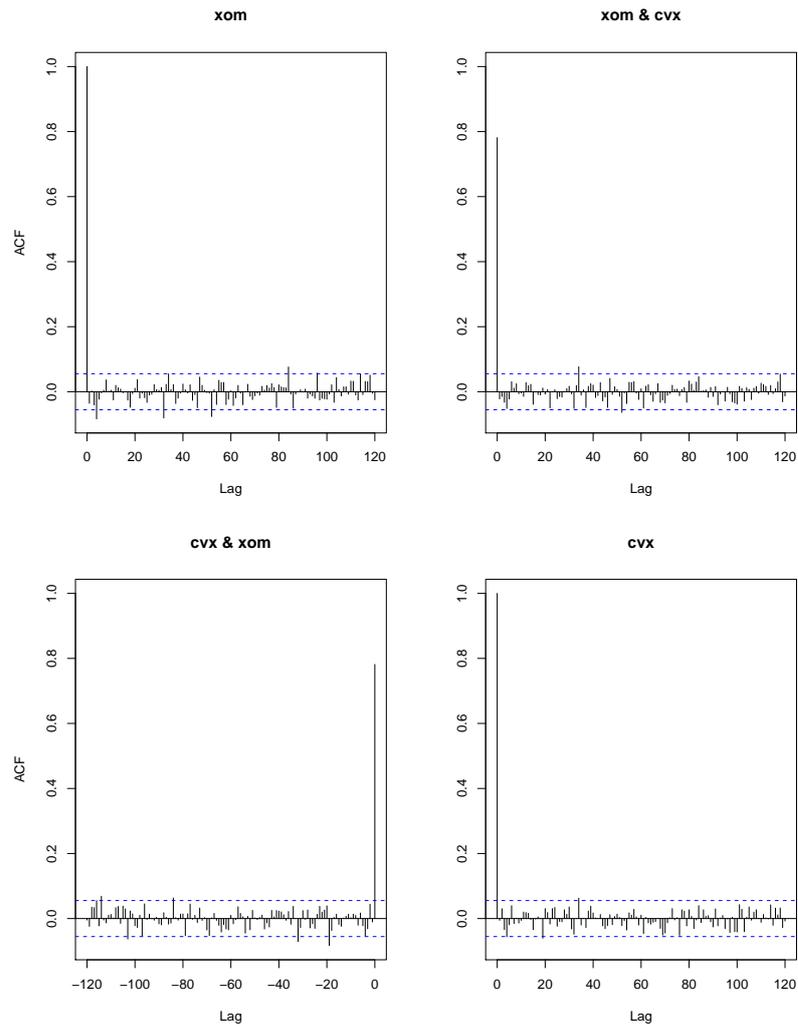


Figura 16: Matriz de correlação do setor de Commodities.

3.3.3 Setor: Saúde

Neste caso as séries apresentam parâmetro constante significativo.

Modelo multivariado AR(1) para o setor de Saúde:

$$\begin{bmatrix} \Delta X_{t,jnj} \\ \Delta X_{t,unh} \end{bmatrix} = \begin{bmatrix} 0,053 \\ 0,096 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta X_{t-1,jnj} \\ \Delta X_{t-1,unh} \end{bmatrix}.$$

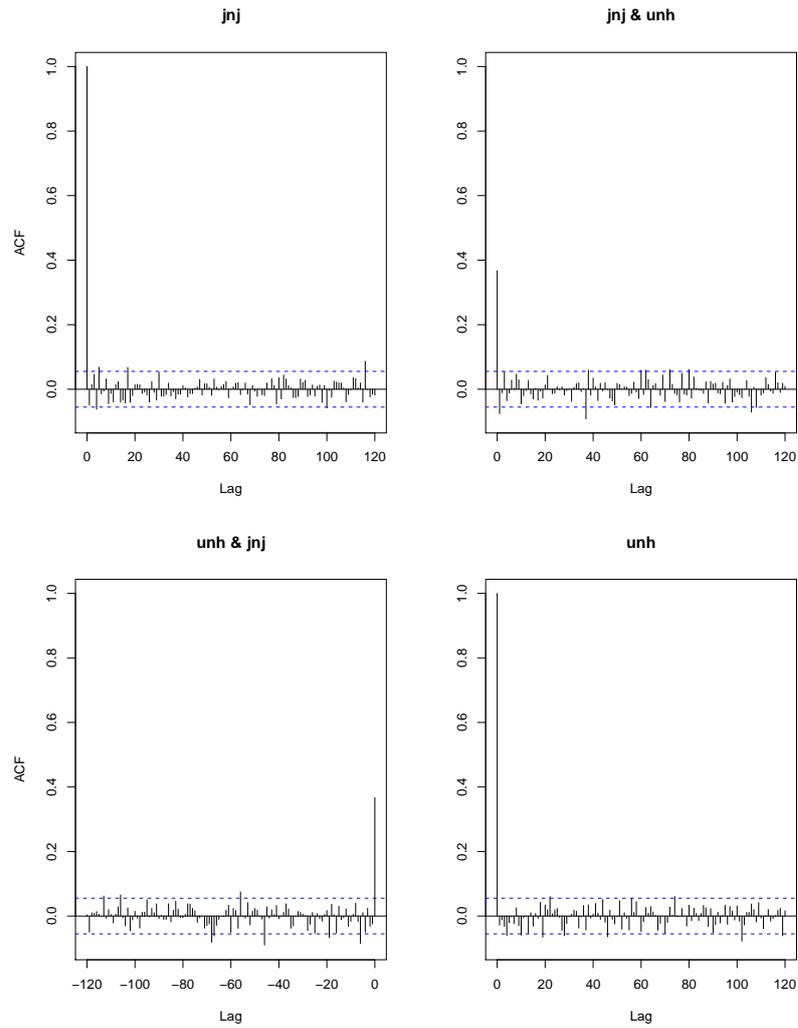


Figura 17: Matriz de correlação do setor de Saúde.

3.3.4 Setor: Serviços

Outro caso aonde apenas os parâmetros constantes são significativos.

Modelo multivariado AR(1) para o setor de Serviços:

$$\begin{bmatrix} \Delta X_{t,amzn} \\ \Delta X_{t,dis} \end{bmatrix} = \begin{bmatrix} 0,588 \\ 0,051 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta X_{t-1,amzn} \\ \Delta X_{t-1,dis} \end{bmatrix}.$$

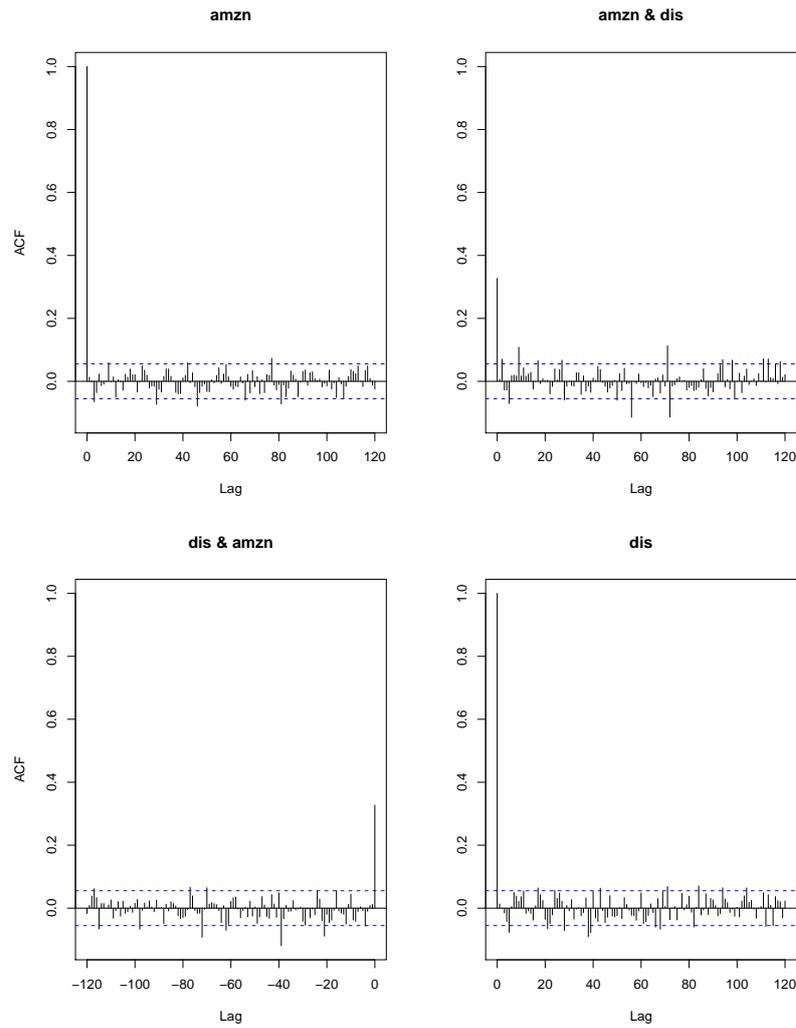


Figura 18: Matriz de correlação do setor de Serviços.

3.3.5 Setor: Tecnologia

Finalmente, no setor de tecnologia existe um elemento significativo na matriz Φ . Esse parâmetro mostra que de fato existe uma relação intra temporal na série.

Modelo multivariado AR(1) para o setor de Tecnologia:

$$\begin{bmatrix} \Delta X_{t,fb} \\ \Delta X_{t,msft} \end{bmatrix} = \begin{bmatrix} 0,091 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0,0358 & 0 \end{bmatrix} \begin{bmatrix} \Delta X_{t-1,fb} \\ \Delta X_{t-1,msft} \end{bmatrix}.$$

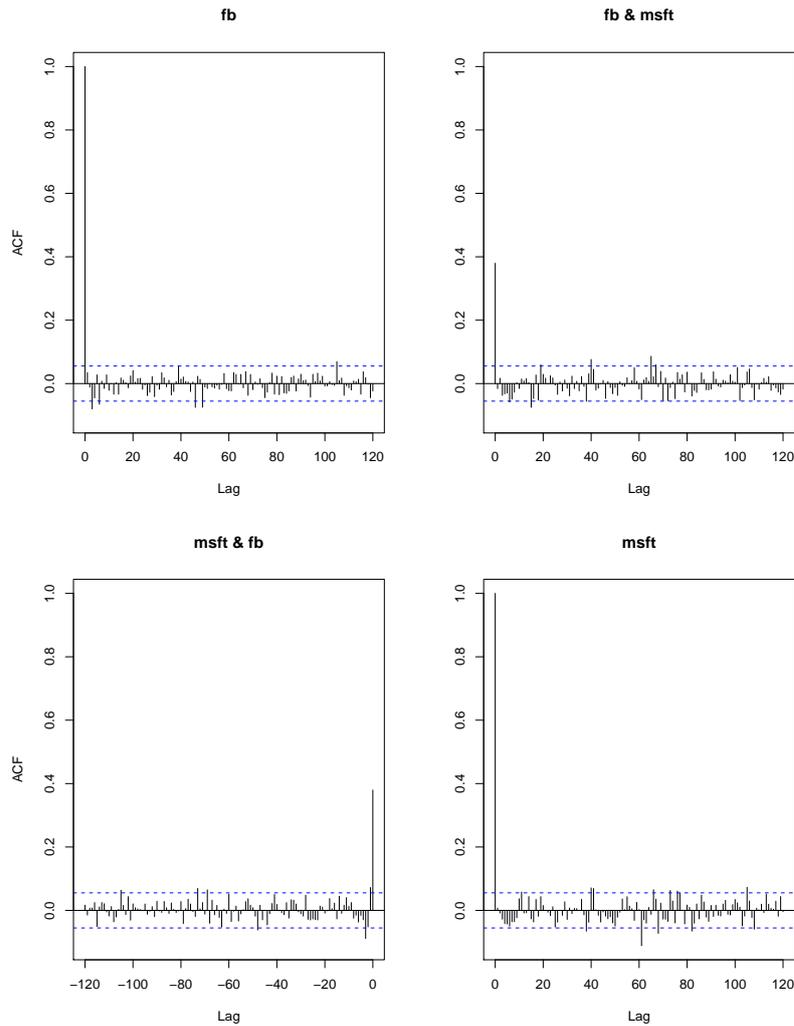


Figura 19: Matriz de correlação do setor de Tecnologia.

3.3.6 Conclusão da análise multivariada

Após a análise multivariada e a modelagem das séries, nota-se que apesar dos gráficos apresentarem uma alta correlação entre as séries. Os modelos feitos na amostra de tamanho de 1254 observações não conseguiu captar tal correlação.

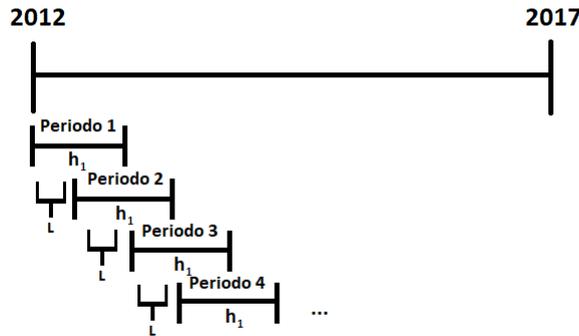
A seguir a Seção de Cointegração, será feito um método iterativo para detectar em qual janela de tempo as séries são mais "cointegráveis", ou seja, mais correlacionadas. Dessa forma uma modelagem futura poderá ser feita na janela de tempo mais apropriada para obtenção de modelos mais significativos.

3.4 Cointegração

Para a análise de cointegração serão analisadas em dois momentos as séries históricas. Primeiramente, todas as séries serão analisadas conjuntamente, variando o período analisado e o comprimento do período. E, logo após, a análise será para cada setor individualmente. O modelo de cointegração utilizado em todas as análises é o modelo de regressão linear simples, ou seja, usando apenas uma série para explicar a série que está sendo considerada como resposta. Além disso, não foi trabalhado com defasagens no tempo, tanto a série que seria a resposta quanto a

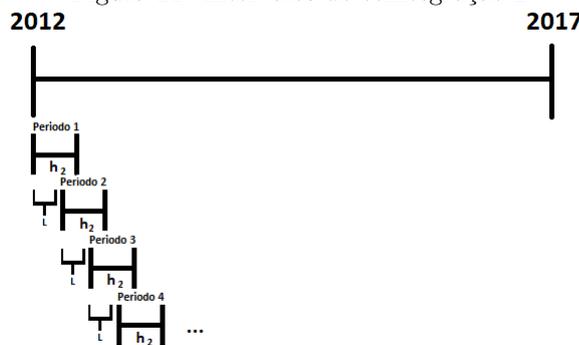
explicativa estão no mesmo período de tempo t . Como séries financeiras são muito voláteis e sofrem quebras estruturais ao longo dos anos, é muito difícil que exista cointegração entre séries se utilizando períodos muito longos. Portanto, para tentar manter o equilíbrio e ver as reais relações lineares entre as séries é necessário fragmentar todo o período de análise em intervalos menores. Além disso, entre um período e outro existem vários comprimentos de tempo que vão sair e que vão entrar. Essas variações que serão trabalhadas podem ser melhor visualizadas na seguinte imagem.

Figura 20: Intervalos de cointegração 1.



Pela imagem, a construção das relações ao longo de todo o período de análise fica mais intuitiva. A análise será feita de 2012 até 2017, entretanto, os testes de cointegração não serão realizados para esse período completo, eles serão realizados ao longo do tempo em cada período menor e com igual comprimento (representado pelos períodos de intervalo menor começando em 2012 e terminando em 2017). O comprimento desses intervalos chamados de h são variáveis, no caso h_1 é um intervalo de tempo específico, podendo variar de 20 em 20 dias, ou seja, comprimentos de 20 dias, 40 dias, 60 dias e assim sucessivamente. Esses comprimentos serão aumentados e diminuídos de 20 dias, o que seria o equivalente a 1 mês na bolsa. Além do comprimento do período utilizado para teste, existe o comprimento do passo entre cada um dos períodos chamado de L . Esse comprimento de passo define quantas observações vão ser retiradas do período atual e quantas vão ser adicionadas para construir o novo período. A quantidade de observações que saem e entram é a mesma (L) e portanto existe uma sobreposição de trechos entre os períodos. Nesse trabalho esse comprimento de passo usualmente será de 20 observações por fornecer a ideia de um passo mensal, mas também será trabalhado com o passo de 1 dia, aumentando significativamente o número de testes realizados. A imagem abaixo mostra essa variação do comprimento dos intervalos em contraste com a figura 21.

Figura 21: Intervalos de cointegração 2.



Dada as considerações feitas sobre a variação do comprimento do período e do comprimento do passo, serão analisadas de maneira geral todas as 10 séries mostrando especificamente os resultados para os comprimentos de período de 20, 40 e 60 dias (1, 2 e 3 meses) e os passos de 1 e 20 dias. Também será analisado as variações dos comprimentos por cada setor, tanto para o passo de 1 dia quanto de 20 dias. Em todos os gráficos de comprimento do período estão em uma escala de 1 a 60, entretanto representam a quantidade de meses de 20 dias, ou seja, quando está no eixo o valor 1, o comprimento é de 20 dias, o valor 2 é para 40 dias e assim sucessivamente até o 60, representando 1200 dias. Foram criadas duas tabelas para a análise geral das 10 séries, cada tabela para um valor de L . Elas foram coloridas na tonalidade azul onde o mais claro indica uma proporção de testes cointegrados entre 0% e 9,99%, o azul moderado indica a proporção de 10% até 19,99% e o azul escuro para proporções iguais ou superiores a 20%.

3.4.1 Geral

- $L=1$

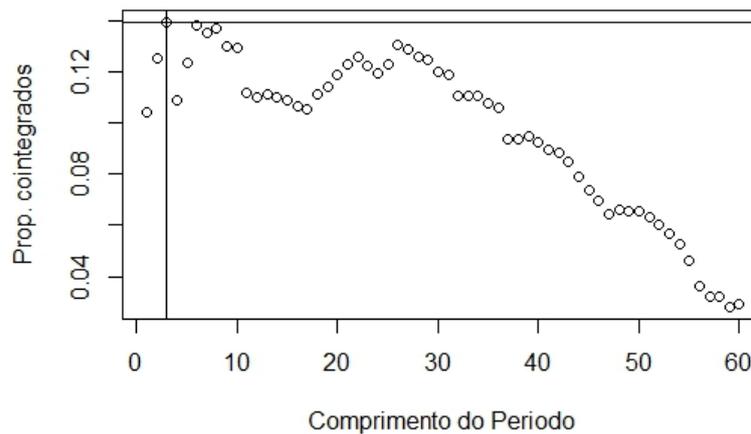


Figura 22: Proporção de casos cointegrados quando $L=1$.

Tabela 4: Proporção de casos cointegrados quando $L=1$

	jpm	ms	jnj	unh	msft	fb	amzn	dis	xom	cvx
jpm	-	12,65%	16,42%	17,76%	8,88%	7,79%	10,72%	12,65%	7,96%	16,25%
ms	13,40%	-	10,22%	11,22%	13,82%	7,37%	11,14%	11,14%	8,38%	11,22%
jnj	13,57%	13,90%	-	14,74%	15,66%	17,00%	16,67%	16,58%	10,22%	11,14%
unh	16,16%	12,40%	16,67%	-	14,24%	13,90%	11,22%	12,23%	14,49%	18,01%
msft	19,43%	23,03%	24,79%	20,18%	-	20,18%	20,52%	16,08%	12,90%	14,66%
fb	11,73%	11,73%	15,33%	12,31%	17,34%	-	12,81%	16,25%	16,08%	12,56%
amzn	16,33%	11,81%	9,13%	12,65%	16,58%	14,99%	-	11,56%	12,73%	11,31%
dis	13,90%	14,24%	12,65%	11,56%	11,98%	15,66%	14,24%	-	18,43%	20,10%
xom	9,13%	10,22%	9,46%	15,49%	9,55%	15,41%	14,07%	17,09%	-	28,14%
cvx	13,90%	8,96%	6,28%	12,90%	8,38%	9,46%	8,04%	17,09%	24,79%	-

- $L=20$

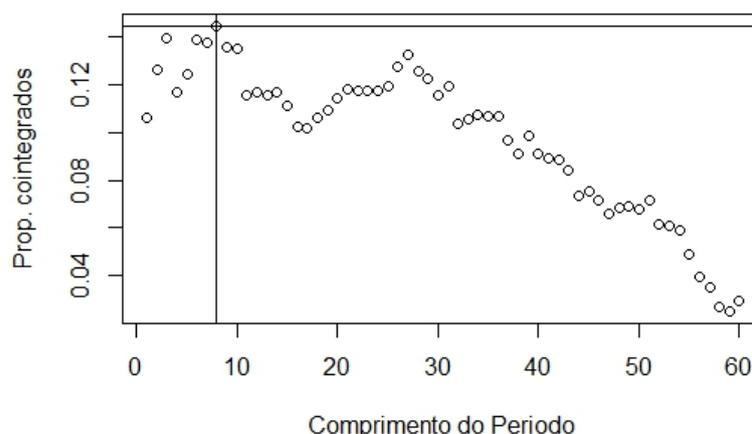


Figura 23: Proporção de casos cointegrados quando $L=20$.

Tabela 5: Proporção de casos cointegrados quando $L=20$.

	jpm	ms	jnj	unh	msft	fb	amzn	dis	xom	cvx
jpm	-	10,00%	23,33%	20,00%	10,00%	6,67%	5,00%	10,00%	10,00%	15,00%
ms	13,33%	-	15,00%	11,67%	16,67%	6,67%	8,33%	15,00%	13,33%	15,00%
jnj	16,67%	15,00%	-	15,00%	18,33%	16,67%	16,67%	18,33%	11,67%	13,33%
unh	20,00%	8,33%	16,67%	-	13,33%	13,33%	3,33%	8,33%	8,33%	13,33%
msft	16,67%	16,67%	26,67%	18,33%	-	21,67%	18,33%	13,33%	11,67%	11,67%
fb	13,33%	13,33%	23,33%	10,00%	21,67%	-	11,67%	16,67%	23,33%	15,00%
amzn	16,67%	10,00%	6,67%	10,00%	16,67%	15,00%	-	10,00%	10,00%	8,33%
dis	15,00%	13,33%	11,67%	5,00%	13,33%	16,67%	13,33%	-	16,67%	18,33%
xom	10,00%	5,00%	11,67%	15,00%	11,67%	20,00%	13,33%	18,33%	-	30,00%
cvx	13,33%	11,67%	8,33%	10,00%	8,33%	15,00%	8,33%	16,67%	26,67%	-

Nas imagens que mostram o comprimento do período e a proporção de testes cointegrados existem tendências e uma dispersão um pouco maior para quando $L = 20$. O comprimento que aumenta mais a cointegração é de 60 dias para $L = 1$ e 160 dias para $L = 20$, entretanto o comprimento de período de 60 dias para o $L = 20$ é o segunda que possui a maior proporção, por isso, as duas tabelas foram feitas considerando esse comprimento. Analisando as tabelas, é notório que a maioria dos setores possui alguma cointegração, seja dentro dele ou com outro setor, isso se deve ao fato do comprimento do período ser o que maximiza os resultados dos testes para ambos tamanhos de L . As relações que merecem destaque, dentro do próprio setor é o setor de saúde, tecnologia e commodity. Já entre os setores, as associações que se diferenciaram foram entre tecnologia e saúde, finanças e saúde e um pouco menos entre o setor de serviços e commodities.

Como dito anteriormente, essas proporções são relações pontuais para o comprimento do período igual a 60 dias, portanto é possível que as séries possam ter mais cointegração em períodos mais longos ou mais curtos, exatamente porque cada área possui seu ponto de equilíbrio da relação entre o curto e longo prazo. Nas próximas seções serão analisados os setores de maneira individual para captar essa relação e também a relação entre o tamanho do passo.

3.4.2 Finanças

- $L=1$

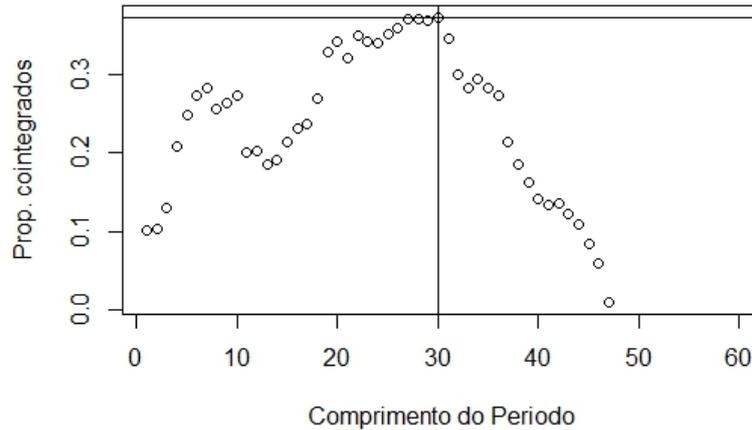


Figura 24: Proporção de casos cointegrados quando no setor de Finanças quando $L=1$.

- $L=20$

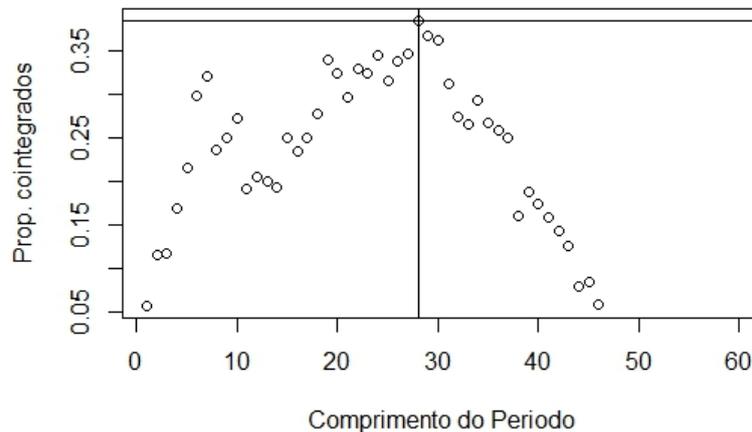


Figura 25: Proporção de casos cointegrados quando no setor de Finanças quando $L=20$.

Por meio dos dois gráficos é possível notar a semelhança entre as sequências de proporção de cointegração. Nas duas imagens existem tendências de alta e de baixa, mostrando exatamente os comprimentos que favorecem a cointegração (encontrando o equilíbrio) e os que desfavorecem (não encontrando o equilíbrio). Além disso, o comprimento que aumenta mais a cointegração é muito próximo para ambos L , que é o período de 560 dias (na Figura 25 é o valor 28) para $L = 20$ e de 600 dias (na Figura 24 é o valor 30) para $L = 1$. Em ambos comprimentos de passo, a proporção de cointegração é alta e supera, em alguns comprimentos, 35% de cointegração.

3.4.3 Commodities

- $L=1$



Figura 26: Proporção de casos cointegrados quando no setor de Commodities quando $L=1$.

- $L=20$

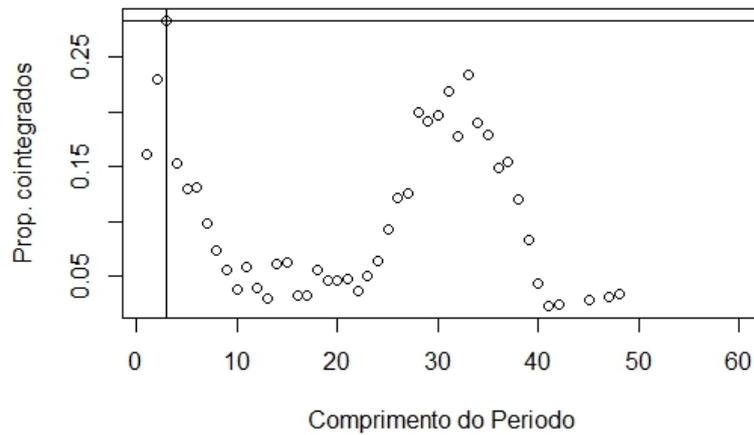


Figura 27: Proporção de casos cointegrados quando no setor de Commodities quando $L=20$.

Novamente, as duas imagens possuem tendências de alta e de baixa. Os comprimentos que favorecem a cointegração estão no início, com pontos sem uma clareza de sequência e depois entre os valores de 30 e 40 no gráfico. O comprimento que aumenta mais a proporção de testes que cointegram é o mesmo para ambos L , que é o período de 60 dias (nas Figuras 26 e 27 é o valor 3). Mantendo a mesma ideia dos outros setores, o L igual a 20 é o que tem as proporções um pouco maiores do que para o L igual a 1. Esse setor também é um dos que apresentaram os maiores valores de proporção.

3.4.4 Saúde

- $L=1$



Figura 28: Proporção de casos cointegrados quando no setor de Saúde quando $L=1$.

- $L=20$

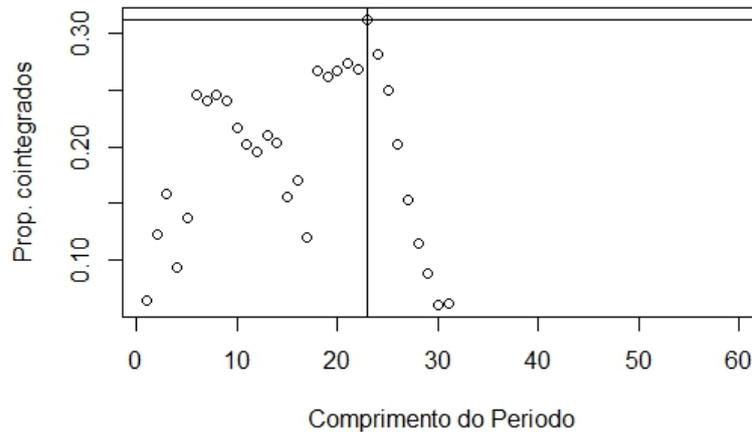


Figura 29: Proporção de casos cointegrados quando no setor de Saúde quando $L=20$.

Para o setor de saúde acontece algo diferente, comprimentos de períodos muito grandes não possuem nenhuma cointegração. Nos dois gráficos mostram um comportamento semelhante mas com grandes distâncias das proporções cointegradas em pequenas alterações de comprimento do H , ou seja, são séries muito sensíveis ao comprimento utilizado, indo em alguns trechos com cointegração 0 até outros com mais de 25% de testes cointegrados. Nesse setor, o comprimento que maximiza as cointegrações, para $L = 20$ e $L = 1$ é o mesmo período que é de 460 dias (valor 23 nas Figuras 28 e 29).

3.4.5 Serviços

- $L=1$

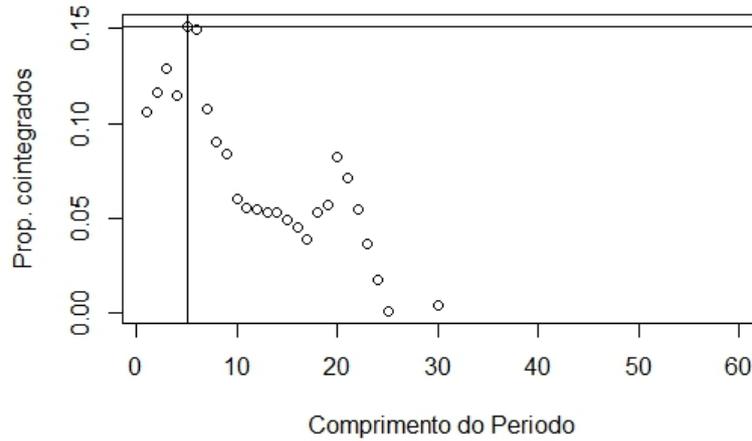


Figura 30: Proporção de casos cointegrados quando no setor de Serviços quando $L=1$.

- $L=20$

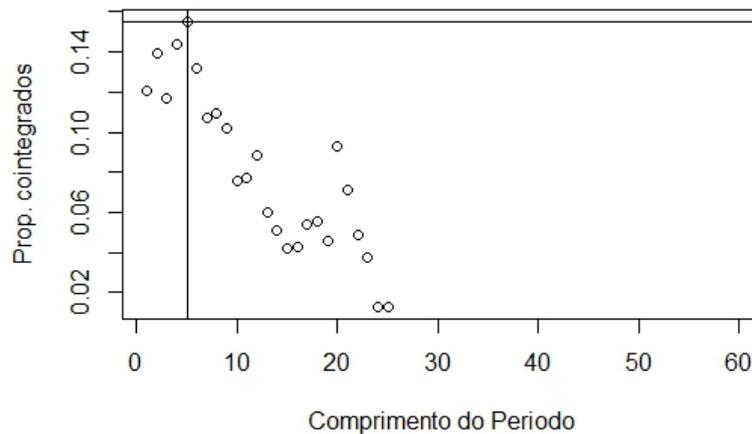


Figura 31: Proporção de casos cointegrados quando no setor de Serviços quando $L=20$.

O setor de serviços é o segundo setor que tem menos proporções de cointegração, o melhor comprimento e melhor passo cointegram não mais que 15%. Assim como no setor de saúde, comprimentos de períodos muito grandes (a partir de 480 dias) não possuem nenhuma cointegração. Tanto para o $L = 1$ e $L = 20$, o comprimento que maximiza é 100 dias (valor 5 nas Figuras 30 e 31). Nesse setor, o passo de 20 teve uma proporção máxima inferior que o de 1 (para 100 dias, um pouco mais de 15% de testes cointegrados para $L = 1$ e um pouco menos de 15% para $L = 20$).

3.4.6 Tecnologia

- $L=1$

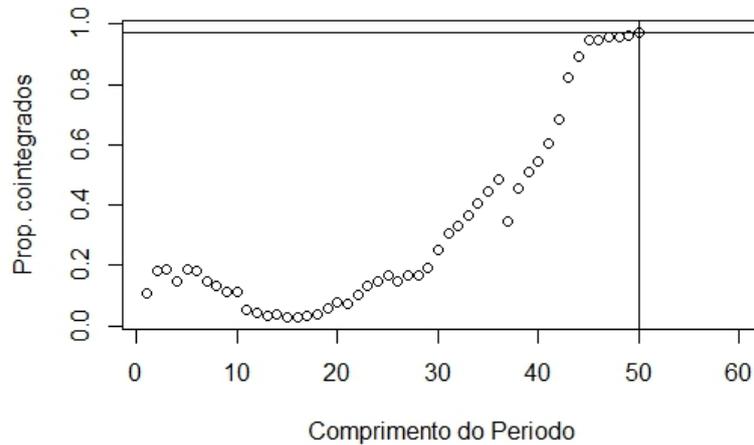


Figura 32: Proporção de casos cointegrados quando no setor de Tecnologia quando $L=1$.

- $L=20$

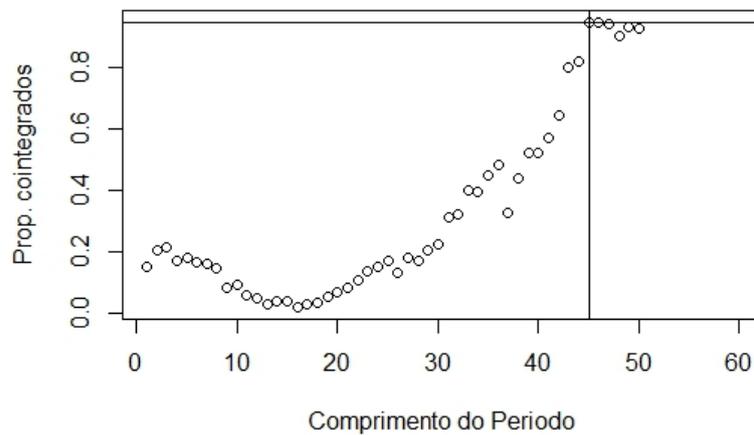


Figura 33: Proporção de casos cointegrados quando no setor de Tecnologia quando $L=20$.

Para a área de tecnologia, os resultados foram mais modestos e diferentes. É o setor que apresentou pouca cointegração em períodos menores e médios, as maiores taxas de cointegração foram obtidas quando se usou períodos de comprimento igual ou superior a 1020 dias (em ambos tamanhos de L), se destacando dos demais setores. Ainda atingindo seu melhor resultado com períodos maiores, para $L = 1$ de 1000 dias e $L = 20$ foi de 920 dias, os resultados ficaram inferiores a 10% de proporção de testes cointegrados.

4 Conclusão

Este trabalho apresentou um breve estudo sobre o modelo ARMA multivariado e a cointegração, sendo esta última um tema natural da área.

Para fins ilustrativos, efetuou-se uma análise de dados financeiros da Bolsa de valores de Nova Iorque. Esses dados foram constituídos por 10 ativos observados diariamente de 18/5/2012 a 12/5/2017. Todos os códigos utilizados para elaboração das análises podem ser acessados por meio do link de acesso:

<https://www.dropbox.com/sh/bb4bg7xe37dzy2n/AAB1hqNI405IbL05osHhkrdRa?dl=0>.

Como as séries não são estacionárias, diferenças foram efetuadas antes da aplicação dos modelos MARMA. Nesse trabalho restringiu-se ao caso bivariado. Como dito anteriormente a modelagem feita sobre toda a amostra não foi significativa devido a extensão de observações. A análise de cointegração indicou os períodos de tempo para os quais as séries conseguem ser mais relacionadas e comportar-se no tempo de maneira mais conjunta. Então conclui-se que para a modelagem de séries temporais deve levar-se em conta que quanto maior for o número de não necessariamente se obterá um modelo mais eficiente.

Para a cointegração notou-se que as séries se cointegram em intervalos de tempo nos quais não existem quebras na tendência. Então duas séries A e B são cointegradas se muito similares e apenas separadas por um valor constante. Os setores que mais se destacaram internamente com a cointegração foram de finanças e saúde, enquanto que o de tecnologia e serviços apresentaram resultados bem mais modestos. Entre a análise de todos os setores, um equilíbrio maior é obtido em comprimentos de tempo menores, entretanto também é baixo sua relação. Verifica-se que os resultados com comprimento de passo de 1 ou 20 dias não possuíam muita diferença, mas usualmente o de 20 dias ficava ligeiramente melhor. Juntando esse fator com o esforço computacional para análise passo a passo se existe cointegração, o comprimento de passo mais recomendado seria de 20 dias.

Referências

- [1] Baldeon, A. e Roberto, P. (2004). **Modelos multivariados de séries temporais na identificação de sistemas mecânicos**. Tese de Doutorado. Universidade Estadual de Campinas, Faculdade de Engenharia Mecânica.
- [2] Brockwell, P. J. e Davis, R. A. (1996). **Introduction to time series and forecasting**. 2. ed. Springer.
- [3] De Paula, J. (2006). **Contagio em mercados financeiros emergentes**. Tese de Doutorado. Universidade Estadual de Campinas . Instituto de Matemática, Estatística e Computação Científica.
- [4] Harvey, A. C. (1989). **Forecasting, structural time series models and the Kalman filter**. Cambridge University Press.
- [5] Holland, M. e Palaia, D. (2010). **Taxa de câmbio e paridade de poder de compra no Brasil: análise econométrica com quebra estrutural**. Economia. Aplicada. vol. 14.
- [6] Kakizawa, Y., Shumway, R. H. e Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. **Journal of the American Statistics Association**. vol. 93.
- [7] Morettin, P.A e Toloi, C.M.C. (2006) **Análise de séries temporais**. 2. ed. Blucher.
- [8] Privalsky, V. (2015). On studying relations between time series in climatology. **Earth System. Dynaics**. vol. 6.
- [9] Wei, W. W. S. (2006). **Time series analysis:Univariate and Multivariate Methods**. 2. ed. Pearson Education Inc.
- [10] West, M. e Harrison, M. (1997). **Bayesian Forecasting and Dynamic Models**. 2 ed. Springer.