Modelagem de séries temporais financeiras: uma abordagem estatística para a identificação de modelos de média condicional

Guilherme Keiel* e Fernando Augusto Bender†

Resumo

A análise de séries temporais financeiras lida com a avaliação de ativos no decorrer do tempo, possibilitando o entendimento do seu comportamento dinâmico e construir modelos capazes de prever valores futuros da série. No primeiro estágio do procedimento para construção de modelos é necessária a seleção da subclasse adequada baseada nas observações do processo, para então realizar a estimação de seus parâmetros. Neste artigo é investigada uma metodologia para determinação de uma subclasse de modelos pertencentes à classe *ARIMAX* para a descrição de processos geradores de séries financeiras. São apresentadas condições para a determinação da subclasse adequada e do número mínimo de parâmetros em cada estrutura. Para ilustrar o método, analisou-se os dados da série diária do índice da Bolsa de Valores de São Paulo entre os anos 2000 à 2014, atestando para o uso das estruturas de modelos *ARIMAX*$(5, 5, 1, 2)$ e *ARIMAX*$(6, 6, 1, 2)$.

Palavras-chave

Séries temporais financeiras. Identificação de modelo. *ARIMAX*. Ibovespa.

Financial time series modeling: a statistical approach for a conditional mean model identification

Guilherme Keiel* e Fernando Augusto Bender†

Abstract

The analysis of financial time series concerns with assets valuation over time, allowing the understanding of their dynamic behavior and to build models capable of predicting future values of the series. In the first stage of model building procedure, it is necessary to select an appropriate subclass of models based on process observations, then to carry out its parameters estimation. In this article a methodology for determination of a subclass of models belonging to the ARIMAX class is investigated for description of the financial series generating processes. Conditions for determining the appropriate subclass and the minimum number of parameters in each structure are presented. In order to illustrate the method, data from the daily series of the São Paulo Stock Exchange index between years 2000 and 2014 are analyzed, attesting for the use of ARIMAX$(5, 5, 1, 2)$ and ARIMAX$(6, 6, 1, 2)$ model estructures.

Keywords

Finantial time series. Model identification. ARIMAX. Ibovespa.

I. Introdução

A análise de séries temporais financeiras lida com a avaliação do preço ou do retorno de ativos no decorrer do tempo. As séries de ativos, por sua vez, geralmente possuem certas propriedades como baixa correlação serial e agrupamento de volatilidade, conforme observa Morettin [1], dificultando a identificação de modelos para os processos. A análise e classificação dessas séries quanto as suas principais características é de suma importância pois possibilita entender o comportamento dinâmico do ativo ou mesmo de um mercado de ações a fim de usá-la para diversas finalidades, entre as quais a de realizar previsões futuras da série.

Na década de 70, o livro *Time series analysis: forecasting and control* dos autores Box e Jenkins [2] sumarizou as metodologias conhecidas até então para a formulação de séries temporais através de equações matemáticas, popularizando o uso do modelo autorregressivo integrador de média móvel (do inglês *autoregressive integrated moving-average, ARIMA*), o

* Departamento de Engenharia Elétrica, Universidade Federal do Rio Grande do Sul, UFRGS, RS, Brasil. e-mail: guilherme.keiel@ufrgs.br

† Centro de Ciências Exatas e Tecnologia, Universidade de Caxias do Sul, RS, Brasil. e-mail: fabender@ucs.br

Data de envio: 27/10/2017
Data de aceite: 07/02/2018

http://dx.doi.org/10.18226/23185279.v6iss1p22

qual foi empregado para análise macroeconômica por Schwert [3] e Dhrymes e Peristiani em [4], por exemplo. Avanços nesta área culminaram no uso do modelo autoregressivo condicional heteroscedástico generalizado (do inglês *generalized autoregressive conditional heteroscedasticity, GARCH*), introduzido por Bollerslev em [5], capaz de descrever também séries com volatilidade variante no tempo, característica frequente em séries de retornos. Estes modelos tornaram-se os mais difundidos na área da econometria, resultando nos trabalhos de Lamoureux e Lastrapes [6] e de Antoniou e Holmes [7], para constar. Atualmente, diversos autores utilizam ainda estes modelos em estruturas multivariadas, capazes de modelar simultaneamente mais de uma variável endógena do processo e sua interdependência, conforme aplicado por Öller em [8]. Somente uma parcela menor dos trabalhos leva em consideração influências exógenas no modelo do processo, como usado por Bos et al. [9], através do modelo autorregressivo integrador de média móvel com entrada exógena (do inglês *autoregressive integrated moving-average, ARIMAX*). Este último propicia, entre os modelos de média condicional, representações mais fidedignas da dinâmica do preço/retorno de ativos em decorrência de fatores externos, como valores de mercados internacionais e, portanto, também será objeto de estudo neste trabalho.

A metodologia para construção de modelos, estabelecida por Box e Jenkins [2], consiste em três estágios: (i) identificação; (ii) estimação de parâmetros e (iii) checagem diagnóstica. Na etapa de identificação, subclasses de modelos são selecionadas de uma classe geral preestabelecida, com base na análise dos dados e/ou conhecimento do processo. Depois, são estimados os seus parâmetros buscando a minimização de algum critério de desempenho. Por fim, os modelos estimados são avaliados quanto a capacidade de produzirem saídas adequadas à série. Desta forma, atesta-se a importância da correta avaliação estatística da série no estágio de identificação de subclasses de modelos, fundamental para a construção de modelos adequados para a representação do processo e, consequentemente, potencializar o resultado das predições dos mesmos.

Este trabalho aborda uma metodologia para a determinação de subclasses de modelos pertencentes à classe *ARIMAX* para a descrição matemática de processos geradores de séries temporais financeiras. São revisadas condições necessárias para a determinação de uma subclasse de modelo adequada e o número mínimo de parâmetros de cada uma de suas estruturas. Para ilustrar este método analisou-se os dados da série diária do índice da Bolsa de Valores de São Paulo (Ibovespa) no período Janeiro de 2000 à 2014, utilizando como entrada exógena o valor do dólar no mesmo período.

Este artigo é organizado como segue: na Seção 2 são apresentados os referenciais teóricos fundamentais para a caracterização de processos geradores de séries quanto a sua estacionariedade, correlação e estrutura de modelo. A Seção 3 descreve a metodologia proposta para a análise estatística e classificação destas séries em subclasses de modelos adequadas. A Seção 4 apresenta a avaliação e resultados para o exemplo numérico do Ibovespa. Por fim, a Seção 5 apresenta as considerações finais e aponta perspectivas de trabalhos futuros.

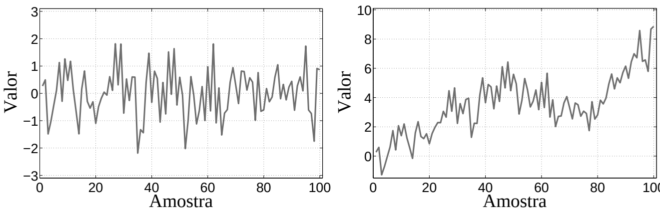
II. Referencial Teórico

Nesta seção serão revisados os referenciais teóricos necessários para a caracterização empírica de processos geradores de séries temporais financeiras e apresentadas as estruturas de modelos utilizadas para descrevê-los matematicamente.

A. Estacionariedade

A classificação de processos quanto à estacionariedade é a base na análise de séries temporais, tal que uma série estacionária pode ser inteiramente descrita pela sua média, variância e correlação. Primeiramente, Box e Jenkins [2] lembram que uma série temporal $\{y_k\}_{k=1}^{N_t}$, onde k é o número e N_t o total de amostras, é uma realização de um processo estocástico $\{\mathbf{Y}_t\}$, que por sua vez, é composto por uma sequência de variáveis aleatórias $\mathbf{Y}$.

Um processo estocástico é estritamente estacionário se sua distribuição de probabilidade é invariante sob deslocamento no tempo. Já este processo é dito por Tsay [10] ser estacionário no sentido amplo (ou fraco) se sua média e covariância são invariantes no tempo. Portanto, a média da série se mantém constante com a adição de novas amostras enquanto que em uma série não-estacionária esta apresentará tendências, conforme verifica-se nos exemplos da Figura 1.



(a) estacionária (b) não-estacionária

Fig. 1: Exemplo de séries temporais.

A análise da estacionariedade é imprescindível para a determinação de modelos pertencentes à classe *ARIMAX*, portanto é o primeiro ensaio a ser realizado na etapa de identificação de uma estrutura de modelo.

B. Correlação

O coeficiente de correlação mede a dependência linear entre duas variáveis aleatórias $\mathbf{U}$ e $\mathbf{Y}$, conforme observa Tsay [10], sendo definido como

$$\rho_{\mathbf{U},\mathbf{Y}} = \frac{cov(\mathbf{U},\mathbf{Y})}{\sigma_{\mathbf{U}}\sigma_{\mathbf{Y}}} \tag{1}$$

no intervalo [-1, 1]. Este coeficiente de correlação pode ser estimado, por meio da amostragem do processo, pela equação

$$\rho_{y,u_l} = \frac{\sum_{k=k_0+l}^{k_0+N}(y_k-\mu_y)(u_{k-l}-\mu_u)}{\sqrt{\sum_{k=k_0}^{k_0+N}(y_k-\mu_y)^2\sum_{k=k_0+l}^{k_0+N}(u_{k-l}-\mu_u)^2}} \tag{2}$$

onde μ_u e μ_y são as médias das subséries $\{u_k\}_{k=k_0}^{k_0+N}$ e $\{y_k\}_{k=k_0}^{k_0+N}$, respectivamente, N o seu tamanho e $l \in \mathbb{Z}^+$ inserindo um atraso entre as subséries analisadas. Se este segmento for a própria série em sua totalidade, tem-se que $k_0 = 1$ e $N_t = N - 1$.

O coeficiente de autocorrelação ρ_l, segundo Tsay [10], é uma particularização do coeficiente de correlação medindo a dependência linear entre amostras atuais e atrasadas de uma mesma série. É estimado, em um dado atraso l, por meio da equação

$$\rho_l = \frac{\sum_{k=k_0+l}^{k_0+N} (y_k - \mu_y)(y_{k-l} - \mu_y)}{\sum_{k=k_0}^{k_0+N} (y_k - \mu_y)^2}. \tag{3}$$

Estendendo o conceito, o coeficiente de autocorrelação parcial ρ_{c_l}, segundo Tsay [10], é um coeficiente de autocorrelação condicional que mede a dependência linear entre amostras atuais e atrasadas da série após eliminado o efeito de um destes atrasos na sua composição. Uma forma de obtê-lo consiste na estimação, pelo método dos mínimos quadrados, de uma estrutura *AR*(l) para a subsérie de ordem tal qual for o atraso considerado. A estimativa do último parâmetro regressor desta estrutura é equivalente à estimativa do coeficiente de autocorrelação parcial para aquele atraso, logo $\rho_{c_l} = \hat{a}_l$.

A avaliação de ambos os coeficientes de autocorrelação e de correlação é muito útil na etapa de determinação de uma subclasse de modelos e do seu número mínimo de parâmetros, conforme estabelecido por Box e Jenkins [2], portanto ambos serão considerados nos testes aplicados neste artigo.

C. Modelos polinomiais de entrada e saída

O comportamento dinâmico de uma série temporal pode ser descrito através de uma equação matemática, a qual dá-se o nome de modelo. Entre os distintos modelos existentes destacam-se os modelos polinomiais de entrada e saída, os quais expressam a relação entre as entradas, saídas e ruído através de uma equação algébrica.

O modelo polinomial de entrada e saída utiliza a seguinte estrutura geral:

$$A(z)y_k = \frac{B(z)}{F(z)}u_k + \frac{C(z)}{D(z)}e_k \tag{4}$$

onde $A(z), B(z), C(z), D(z)$ e $F(z)$ são polinômios de parâmetros expressados em termo do operador de atraso z^{-1}, y_k a saída, u_k a entrada e e_k o resíduo de equação no instante k.

A partir da estrutura geral derivam-se algumas classes de modelos, dentre elas o modelo autorregressivo de média móvel com entrada exógena (do inglês *autoregressive moving average with exogenous input, ARMAX*), que expressa a média condicional da saída y no instante k como uma função das observações passadas da saída $y_{k-1}, \ldots y_{k-p}$, dos resíduos $e_{k-1}, \ldots e_{k-q}$ e das entradas $u_k, \ldots u_{k-r}$, segundo Ljung [11], através da equação

$$A(z)y_k = B(z)u_k + C(z)e_k \tag{5}$$

onde $p, q, r \in \mathbb{Z}^+$ definem a ordem dos polinômios autorregressivos $A(z) = 1 - a_1 z^{-1} \ldots - a_p z^{-p}$, de média móvel $C(z) = 1 - c_1 z^{-1} \ldots - c_q z^{-q}$ e da entrada exógena $B(z) = -b_0 - b_1 z^{-1} \ldots - b_r z^{-r}$, respectivamente.

Estendendo esse modelo para que o polinômio autorregressivo possua raízes unitárias, uma característica de processos não-estacionários, tem-se a classe *ARIMAX*, descrita por

$$A(z)y_k = B(z)u_k + \frac{C(z)}{1 - z^{-d}}e_k \tag{6}$$

expressando y_k como uma função das saídas, resíduos e entradas passadas, além de incluir um integrador (somador) no resíduo, onde $d \in \mathbb{Z}^+$ define a ordem da diferenciação aplicada na série afim de torná-la estacionária.

Os processos autorregressivos *AR* têm como vantagem a dependência linear ponderada de amostras passadas da própria série, facilitando a estimação dos parâmetros, obtidos pelo processo de regressão linear. Os processos de médias móveis *MA* apresentam dependência linear ponderada das amostras prévias do resíduo da equação, porém não são diretamente mensuráveis, dificultando a etapa de estimação desses parâmetros. Nos casos em que a saída da série depende linearmente de uma variável explicativa, dá-se o nome de processos com entrada exógena *X*.

Diferentes subclasses de modelo podem ser obtidas considerando um ou mais processos pertencentes à classe de modelos *ARIMAX*(p, d, q, r), onde a notação p, d, q, r adotada por Box e Jenkins [2] define a ordem dos polinômios de cada uma destas estruturas na composição de um modelo. A metodologia apresentada busca, através da constatação dos diferentes processos que a compõem, determinar subclasses de modelos adequadas.

III. METODOLOGIA

Nesta seção será descrita a metodologia para a análise estatística e classificação de uma determinada série temporal financeira em uma subclasse de modelo adequada. Nela serão apresentados os testes para a classificação da série quanto a sua estacionariedade, dependência autorregressiva, dependência de médias móveis e, por fim, dependência de entrada exógena.

A. Teste da estacionariedade

O ensaio para a verificação da estacionariedade do processo leva em consideração janelas de N amostras de uma série $\{y_k\}_{k=1}^{N_t}$, portanto subséries, com amostras iniciais k_0 e finais $k_0 + N$. A divisão em subséries visa construir modelos baseados na dinâmica apenas daquele segmento, útil quando a série apresentar características muito específicas por regiões ou sazonalidade.

A metodologia empregada consiste no teste Dickey-Fuller, proposto por Dickey e Fuller em [12], no qual uma estrutura de modelo *AR*(1) é utilizada para a regressão da subsérie. O parâmetro dessa estrutura deve ser estimado e depois avaliado quanto às hipóteses nula, que estabelece $a_1 \geq 1$, ou alternativa, onde $a_1 < 1$, no chamado teste de raiz unitária. Realizando a estimava de a_1 por mínimos quadrados, o coeficiente DF_τ é calculado pela estatística

$$\mathrm{DF}_\tau = \frac{\hat{a}_1 - 1}{\sigma_{\hat{a}_1}} \tag{7}$$

sendo $\sigma_{\hat{a}_1}$ o desvio padrão do parâmetro estimado, que pode ser avaliado por

$$\sigma_{\hat{a}_1} = \sqrt{\frac{\sum_{k=k_0}^{k_0+N} (y_k - a_1 y_{k-1})^2}{(N-1)\sum_{k=k_0}^{k_0+N} (y_{k-1})^2}}. \tag{8}$$

Para rejeitar a hipótese nula e o segmento da série ser considerado estacionário, o parâmetro DF_τ não deve exceder o valor crítico DF_{crit}, conforme definido na desigualdade (9). Este valor crítico foi tabelado por Dickey e Fuller em [13] para tamanhos de amostras e níveis de significância distintos.

$$\mathrm{DF}_\tau < \mathrm{DF}_{crit} \tag{9}$$

Desta forma, um segmento de série que não atender à desigualdade é considerado não-estacionário e atestará para o uso de modelos com a estrutura integradora *I* na sua composição. Caso isto ocorra é preciso trabalhar com uma nova subsérie, produzida tomando-se até d diferenças da subsérie em questão, até que a desigualdade (9) seja satisfeita.

Para a análise completa da série, o ensaio de estacionariedade deve ser repetido para os demais segmentos desejados, o qual atestará individualmente para o uso, ou não, da estrutura *I* e sua ordem d nos modelos de cada uma das subséries que a compõem.

B. Teste de dependência autorregressiva

No próximo ensaio, para a verificação da dependência linear de uma subsérie com suas saídas passadas, avalia-se o coeficiente de autocorrelação da subsérie. Se seu módulo for maior do que um limiar preestabelecido, então a região é dita suscetível à dependência do próprio processo, caso contrário o nível de dependência é considerado não-significativo.

Primeiramente, são definidos o vetor Y com o segmento de série $\{y_k\}_{k=k_0}^{k_0+N}$ e o vetor Y_l com esse próprio segmento deslocado l amostras para a esquerda. Após isso, determina-se o coeficiente de autocorrelação ρ_l da subsérie, calculado através da equação (3). Para constatar a significância do coeficiente de autocorrelação naquele atraso, verifica-se se o seu módulo excede o limiar definido na desigualdade

$$|\rho_l| > \delta_\rho. \tag{10}$$

O limite δ_ρ é calculado pela equação (11) e depende do desvio padrão (DP) tolerável na estimativa, que por sua vez é consequência do intervalo de confiança estipulado pelo pesquisador. Este intervalo normalmente é definido entre 95% e 99%, como observa-se em Tsay [10].

$$\delta_\rho = \frac{\mathrm{DP}}{\sqrt{N}} \tag{11}$$

Desta forma, um segmento de série em que o coeficiente de autocorrelação atender à desigualdade (10) será considerado dependente da própria variável e atestará para o uso de modelos com a estrutura autorregressiva *AR* na sua composição. No entanto, apenas indicará que esta dependência é significativa no atraso l entre as amostras, logo uma avaliação nos demais atrasos se faz necessária. O cálculo do coeficiente de autocorrelação é repetido nos $l = 1, \cdots, N-1$ atrasos, até que se verifique não haver mais coeficientes com nível significativo. A posição do último coeficiente significativo definirá a ordem mínima p da estrutura *AR*.

Para a análise completa da série, este ensaio deve ser repetido para os demais segmentos, o qual atestará individualmente para o uso, ou não, da estrutura *AR* e sua ordem mínima p em cada uma das subséries que a compõem.

C. Teste de dependência de médias móveis

No próximo ensaio, para a verificação da dependência linear de uma subsérie com os resíduos passados, avalia-se o coeficiente de autocorrelação parcial da subsérie. Se seu módulo for maior do que um limiar preestabelecido, então a região é dita suscetível à dependência de médias móveis, caso contrário o nível de dependência é considerado não-significativo.

Primeiramente, assim como no teste de dependência autorregressiva, são definidos o vetor Y com o segmento de série $\{y_k\}_{k=k_0}^{k_0+N}$ e o vetor Y_l com esse próprio segmento deslocado l amostras para a esquerda. Após isso, determina-se o coeficiente de autocorrelação parcial ρ_{c_l} da subsérie, estimando uma estrutura *AR*(l). Para constatar a significância do coeficiente de autocorrelação parcial naquele atraso, verifica-se se o seu módulo excede o limiar definido na desigualdade (12), assumindo um determinado intervalo de confiança.

$$|\rho_{c_l}| > \delta_{\rho_c} \tag{12}$$

Desta forma, um segmento de série em que o coeficiente de autocorrelação parcial atender à desigualdade (10) será considerado dependente dos resíduos e atestará para o uso de modelos com a estrutura de médias móveis *MA* na sua composição. No entanto, apenas indicará que esta dependência é significativa no atraso l entre as amostras, logo uma avaliação nos demais atrasos se faz necessária. O cálculo do coeficiente de autocorrelação parcial é repetido nos $l = 1, \cdots, N-1$ atrasos, até que se verifique não haver mais coeficientes com nível significativo. A posição do último coeficiente significativo definirá a ordem mínima q da estrutura *MA*.

Para a análise completa da série, o ensaio deve ser repetido para os demais segmentos, o qual atestará individualmente para o uso, ou não, da estrutura *MA* e sua ordem mínima q em cada uma das subséries que a compõem.

D. Teste de dependência exógena

No ensaio seguinte, para a verificação da dependência linear de um segmento de série a qual deseja-se modelar o processo e um segmento gerado por um processo distinto, avaliam-se os coeficientes de correlação entre ambas as subséries. Se seu módulo for maior do que um limiar preestabelecido, então a região é dita suscetível à dependência exógena daquele processo, caso contrário o nível de dependência é considerado não-significativo.

Primeiramente, são definidos o vetor Y com o segmento de série $\{y_k\}_{k=k_0}^{k_0+N}$ e o vetor U_l com o segmento $\{u_k\}_{k=k_0+l}^{k_0+N}$ próprio segmento deslocado l amostras para a esquerda. Após isso, determina-se o coeficiente de correlação ρ_{y,u_l} entre as subséries, calculado através da equação (2). Para constatar a significância do coeficiente de correlação naquele atraso, verifica-se se o seu módulo excede o limiar definido na desigualdade (13), assumindo um determinado intervalo de confiança.

$$|\rho_{y,u_l}| > \delta_{\rho_{y,u}} \tag{13}$$

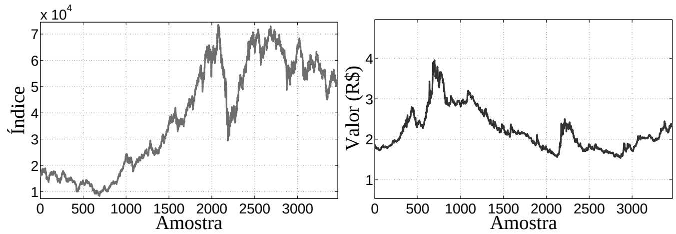
Desta forma, um segmento de série em que o coeficiente de correlação com outra subsérie atender à desigualdade (13) será considerado dependente de variável exógena e atestará para o uso de modelos com a estrutura de entrada exógena *X* na sua composição. No entanto, apenas indicará que esta dependência é significativa entre subséries defasadas em l atrasos entre as amostras, logo uma avaliação nos demais atrasos se faz necessária. O cálculo do coeficiente de correlação pode ser repetido nos $l = 1, \cdots, N-1$ atrasos, até que se verifique não haver mais coeficientes com nível significativo. A posição do último coeficiente ainda significativo definirá a ordem mínima b da estrutura *X*.

Para a análise completa da série, o ensaio deve ser repetido para os demais segmentos, o qual atestará individualmente para o uso, ou não, da estrutura *X* e sua ordem mínima r em cada uma das subséries que a compõem.

IV. Exemplo Numérico

Nesta seção, a metodologia descrita na Seção 3 será aplicada em um exemplo numérico real, o qual consiste na série histórica do índice diário da Bolsa de Valores de São Paulo, o Ibovespa. A análise estatística da série é realizada com o auxílio do *software* Matlab.

A série utilizada no trabalho, ilustrada na Figura 2 (a), retrata o fechamento diário do Ibovespa no período de 03/01/2000 à 02/01/2014, contabilizando apenas dias úteis, em um total de 3469 amostras e a mesma encontra-se disponível na página da BM&F Bovespa [14]. O Ibovespa, tido como principal indicador do desempenho médio do mercado de ações brasileiro, é o índice de uma carteira hipotética de ativos composta pelas ações de maior volume negociado na Bovespa no último quadrimestre. Por outro lado, os dados utilizados como variável explicativa do processo compreendem a série com o valor do dólar, em Reais, naquele mesmo período, apresentada na Figura 2 (b).



(a) Ibovespa (b) dólar

Fig. 2: Séries históricas do período de Janeiro de 2000 à 2014.

A. Resultados

Primeiramente, selecionou-se dois segmentos da série financeira para demonstrar a análise. Para fins práticos, as subséries utilizadas possuem tamanhos iguais, contendo 500 amostras cada. A Figura 3 apresenta os dois segmentos em destaque, iniciados a partir das amostras 1500 e 2500, compreendendo o período anterior e posterior à recuperação de uma forte queda do índice no final do ano de 2008, respectivamente.

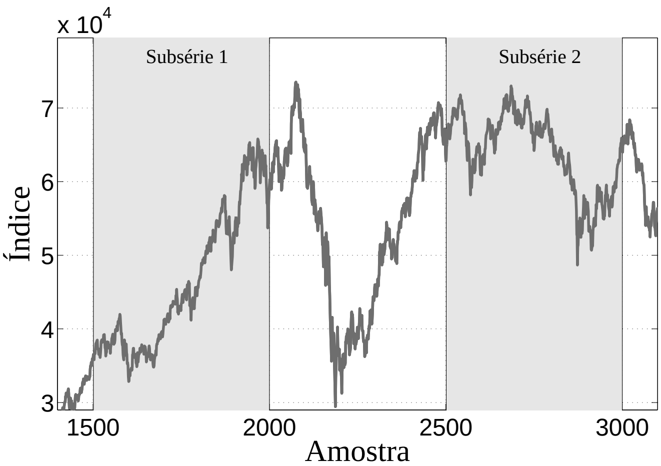


Fig. 3: Segmentos da série selecionados para a análise.

Na sequência, as duas subséries são avaliadas quanto à sua estacionariedade, por meio do coeficiente DF$_\tau$. Para a hipótese nula do teste ser rejeitada e a subsérie ser considerada estacionária esta deve satisfazer a desigualdade (9), tendo como valor crítico tabelado DF$_{crit} = -2,58$ uma vez que $N = 500$ e o nível de significância desejado, de 1%. Desta forma, de acordo com os resultados dos coeficientes descritos na Tabela I, constatou-se a não-estacionariedade de ambas as subséries, sendo uma diferenciação destes dados o suficiente para atender à desigualdade (9). Logo, atesta-se para o uso da estrutura *I* de ordem $d = 1$ na composição da subclasse de modelo para a descrição destes segmentos.

TABLE I: Valores do teste Dickey-Fuller.

Diferença (d)	DF$_\tau$ Subsérie 1	DF$_\tau$ Subsérie 2
0	1,0608	-0,0229
1	-23,7105	-22,4326

Na etapa seguinte, avaliou-se a dependência linear das subséries com suas saídas passadas, por meio do coeficiente de autocorrelação ρ_l, adotando um intervalo de confiança de 95% (DP=1.96). Logo, o coeficiente deve exceder o limiar $\delta_\rho = 0,0877$ para que a desigualdade (10) seja satisfeita. No primeiro segmento da série obteve-se autocorrelação significativa apenas no quinto atraso avaliado, atestando para a utilização da estrutura *AR* de ordem $p = 5$ na composição do modelo, como pode ser verificado na Tabela II. Da mesma forma, avaliando a dependência do segundo segmento obteve-se autocorrelação significativa no primeiro e sexto atraso, indicando para o uso de uma estrutura com $p = 6$.

TABLE II: Autocorrelação das subséries.

Atraso (l)	δ_ρ	Subsérie 1	Subsérie 2
1	0,0877	-0,0656	-0,0895
2	0,0877	0,0484	0,0275
3	0,0877	-0,0043	-0,0322
4	0,0877	-0,0082	0,0201
5	0,0877	0,1222	-0.0104
6	0,0877	-0,0147	-0,0994

Seguindo a metodologia apresentada, avaliou-se a dependência linear das subséries com seus resíduos passados, por meio dos coeficientes de autocorrelação parcial ρ_{c_l}, com um intervalo de confiança de 95%. Logo, $\delta_\rho = 0,0877$ na desigualdade (12). Obteve-se autocorrelação parcial significativa apenas no quinto atraso avaliado, atestando para a utilização da estrutura *MA* com ordem $p = 5$ na composição do modelo, como pode ser verificado na Tabela III. Da mesma forma, no segundo segmento obteve-se uma autocorrelação significativa no sexto atraso, indicando para o uso de uma estrutura com $p = 6$.

TABLE III: Autocorrelação parcial das subséries.

Atraso (l)	δ_{ρ_c}	Subsérie 1	Subsérie 2
1	0,0877	-0,0657	-0,0029
2	0,0877	0,0443	0,0275
3	0,0877	0,0018	-0,0320
4	0,0877	-0,0107	0,0193
5	0,0877	-0,1319	-0.0089
6	0,0877	-0,0279	-0,1018

Por fim, avaliou-se a dependência linear das subséries com a subsérie do dólar, por meio do coeficiente de correlação ρ_{u,y_l}, com um intervalo de confiança de 95%. Logo, $\delta_{\rho_{u,y}} = 0,0877$ na desigualdade (13). No primeiro segmento da série obteve-se uma correlação significativa em até dois atrasos, atestando para a utilização da estrutura *X* de ordem $r = 2$ na composição do modelo, como pode ser verificado na Tabela IV. Da mesma forma, avaliando-se a dependência do segundo segmento com a respectiva subsérie obteve-se uma correlação significativa em dois atrasos, também indicando para o uso da estrutura de entrada exógena.

TABLE IV: Correlação com subsérie do dólar.

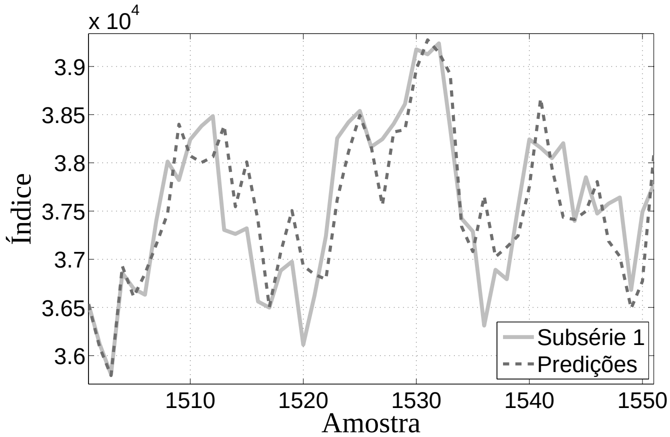
Atraso (l)	$\delta_{\rho_{u,y}}$	Subsérie 1	Subsérie 2
0	0,0877	-0,5008	-0,4201
1	0,0877	-0,1929	-0,2169
2	0,0877	0,0092	-0,0218
3	0,0877	0,0255	0,0254
4	0,0877	-0,0257	0,0787
5	0,0877	0,0031	0,0160

Com base no resultado de cada um dos ensaios é possível atestar para o uso de modelos com as estruturas *ARIMAX*$(5,5,1,2)$ e *ARIMAX*$(6,6,1,2)$ para a descrição dos processos geradores das subséries 1 e 2 do Ibovespa, respectivamente. Esse resultado possibilita avançar para a próxima etapa no processo de construção de modelos, a qual consiste na estimação dos parâmetros dessas estruturas. Apenas como uma forma de avaliação da qualidade dos modelos pertencentes à subclasse selecionada, estimou-se os parâmetros do modelo *ARIMAX*$(5,5,1,2)$ referente à subsérie 1 através do método de minimização do somatório dos quadrados dos erros de predição, conforme metodologia de Ljung [11], comparando-o frente à modelos de outras subclasses. Os resultados estão descritos na Tabela V.

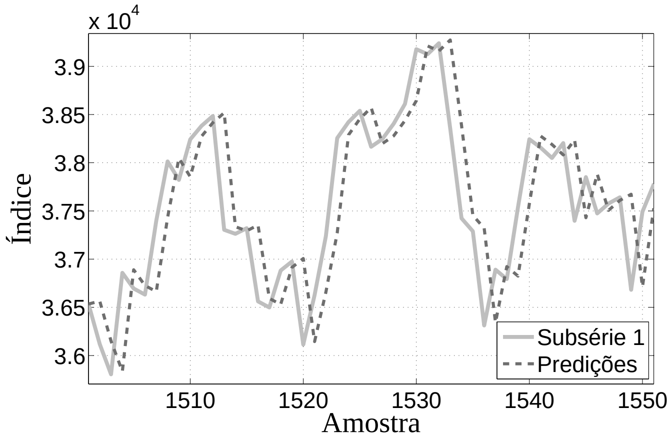
TABLE V: Erros de predição dos modelos.

	ARIMAX$(5,5,1,2)$	*ARIMA*$(5,5,1)$	*ARI*$(5,1)$	*AR*(1)
MSE	474815,70	666444,60	674106,91	690458,37
MAE	514,50	597,88	597,91	600,18
ME	2677,55	3668,74	3674,77	3845,99

A estimação dos parâmetros do modelo *ARIMAX*$(5,5,1,2)$ resultou em um custo do erro quadrático médio (em inglês *mean squared error, MSE*) das predições de 474815,70 e que chega a ser 45% menor do que o custo obtido com a estimação do modelo *AR*(1), por exemplo. O erro absoluto médio (em inglês *mean absolute error, MAE*) e erro absoluto máximo (em inglês *maximum absolute error, ME*) também ficaram menores em relação aos demais modelos, justificando a estrutura de modelo obtida previamente na etapa de identificação. A Figura 4 compara as predições de um passo adiante das 50 primeiras amostras da subsérie 1, realizadas com ambos os modelos *ARIMAX*$(5,5,1,2)$ e *AR*(1), em relação à subsérie real. Observa-se que o modelo usado em (b) não descreve adequadamente a subsérie, uma vez que as predições são praticamente uma repetição das suas amostras reais passadas (pois $\hat{a}_1 = 1,0008$), enquanto que o modelo (a) faz predições significativas em alguns momentos.



(a) modelo *ARIMAX*$(5,5,1,2)$



(b) modelo *AR*(1)

Fig. 4: Predições de uma amostra adiante da subsérie.

V. CONCLUSÃO

Este artigo investigou uma metodologia para determinação de subclasses de modelos pertinentes em meio à classe *ARIMAX* para a descrição de séries temporais financeiras, através de avaliações da estacionariedade e dos coeficientes de correlação da série. Aplicou-se essa metodologia para a identificação do processo gerador da série diária do Ibovespa no período entre 2000 e 2014. Verificou-se um comportamento predominantemente não-estacionário, com dependência autorregressiva, de médias móveis do resíduo e correlação com a variável explicativa do dólar em ambos os segmentos, apontando para a utilização das estruturas $ARIMAX(5,5,1,2)$ e $ARIMAX(6,6,1,2)$ na próxima etapa da construção do modelo. Para trabalhos futuros, cita-se o uso de modelos condicionais da variância ou multivariados para o exemplo abordado.

VI. Referências

[1] P.A. Morettin, *Econometria financeira: um curso em séries temporais financeiras*, Edgard Blücher, 2008.

[2] G.E.P. Box and G.M. Jenkins, *Time series analysis: forecasting and control*, Holden-Day series in time series analysis. Holden-Day, 1970.

[3] G.William Schwert, "Effects of model specification on tests for unit roots in macroeconomic data," *Journal of Monetary Economics*, vol. 20, no. 1, pp. 73 – 103, 1987.

[4] Phoebus J. Dhrymes and Stavros C. Peristiani, "A comparison of the forecasting performance of wefa and arima time series methods," *International Journal of Forecasting*, vol. 4, no. 1, pp. 81 – 101, 1988.

[5] Tim Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307 – 327, 1986.

[6] Christopher G. Lamoureux and William D. Lastrapes, "Heteroskedasticity in stock return data: Volume versus garch effects," *The Journal of Finance*, vol. 45, no. 1, pp. 221–229, 1990.

[7] Antonios Antoniou and Phil Holmes, "Futures trading, information and spot price volatility: evidence for the ftse-100 stock index futures contract using garch," *Journal of Banking & Finance*, vol. 19, no. 1, pp. 117 – 129, 1995.

[8] Lars-Erik Öller, "Macroeconomic forecasting with a vector arima model: A case study of the finnish economy," *International Journal of Forecasting*, vol. 1, no. 2, pp. 143 – 150, 1985.

[9] Charles S. Bos, Philip Hans Franses, and Marius Ooms, "Inflation, forecast intervals and long memory regression models," *International Journal of Forecasting*, vol. 18, no. 2, pp. 243 – 264, 2002, Forecasting Long Memory Processes.

[10] R.S. Tsay, *Analysis of Financial Time Series*, Wiley Series in Probability and Statistics. Wiley, 2005.

[11] L. Ljung, *System Identification: Theory for the User*, vol. 1, Prentice Hall, New Jersey, NJ, 1999.

[12] D Dickey and Wayne A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," vol. 74, 06 1979.

[13] W.A. Fuller, *Introduction to statistical time series*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 1976.

[14] BM&FBOVESPA, ***Séries históricas**. Disponível em: http://www.bmfbovespa.com.br*, Acesso em: 14 abr. 2016.