Modelos Machine Learning e modelos ARIMA na previsão do PSI20

Miguel Ângelo R. Fernandes

UNIVERSIDADE D COIMBRA

UNIVERSIDADE D
COIMBRA

Miguel Ângelo Rodrigues Fernandes

Modelos Machine Learning e modelos ARIMA na previsão do PSI20

Dissertação no âmbito do Mestrado em Economia Financeira orientada pelo Professor Doutor Pedro Miguel Avelino Bação e apresentada à Faculdade de Economia da Universidade de Coimbra

Fevereiro de 2019

UNIVERSIDADE D
COIMBRA

Miguel Ângelo Rodrigues Fernandes

MODELOS MACHINE LEARNING E ARIMA NA PREVISÃO DO PSI20

Trabalho de Projecto no âmbito do Mestrado em Economia Financeira orientado pelo Professor Doutor Pedro Miguel Avelino Bação e apresentado à Faculdade de Economia da Universidade de Coimbra

Fevereiro de 2019

Miguel Ângelo Rodrigues Fernandes

Modelos Machine Learning e ARIMA na previsão do PSI20

Trabalho de Projeto do Mestrado de Economia, com especialização em Economia Financeira, apresentado à Faculdade de Economia da Universidade de Coimbra para a obtenção do grau de mestre.

Orientado por: Professor Doutor Pedro Miguel Avelino Bação

Agradecimentos

Gostaria de agradecer a todos aqueles que me apoiaram durante todo o meu percurso escolar. Um agradecimento especial ao professor doutor Pedro Miguel Avelino Bação pela disponibilidade demonstrada. À faculdade de economia pela formação que me deram que permitiu fazer este trabalho e me preparou para o meu futuro profissional. A todos os meus amigos. À minha namorada pelo apoio incondicional na elaboração deste trabalho. À minha mãe por todo o amor e apoio ao longo de toda a minha vida.

Resumo

Este trabalho tem como objetivo comparar o desempenho de modelos ARIMA e modelos *machine learning*, mais especificamente o modelo SVM, na previsão da variação do índice PSI20. Para o efeito, foi recolhido o preço de fecho do PSI20, 21 de Novembro de 2006 a 28 de Novembro de 2018 (dias úteis apenas) e procedeu-se à sua transformação em logaritmos. As previsões foram testadas numa subamostra correspondente a 30% da amostra total.

Primeiramente, testou-se o modelo ARIMA. A especificação foi escolhida através dos critérios de informação AIC, BIC e HQC, não tendo os resultados sido consensuais quanto ao modelo ARIMA a utilizar. O teste AIC sugeriu o modelo ARIMA (3;0;3) e os testes BIC e HQC sugeriram o modelo ARIMA (0;0;1). Destes dois modelos, aquele que apresentou um menor *root mean square error* (RMSE) foi o modelo ARIMA (0;0;1).

De seguida, testou-se o modelo SVM, tendo o software utilizado (Gretl) selecionado o modelo ε-SVR. Os restantes elementos da especificação foram escolhidos de acordo com o seu desempenho, concluindo-se que o melhor modelo SVM é um modelo ε-SVR que utiliza uma função de kernel do tipo linear. Este é também o melhor modelo entre todos os estudados, embora o RMSE varie muito pouco.

Os modelos foram também usados para a previsão do sinal da variação da cotação do PSI20. Neste caso, os critérios de informação AIC, BIC e HQC foram concensuais quanto à especificação do modelo ARIMA, uma vez que todos os critérios apontam para a utilização do modelo ARIMA (0;0;1). O melhor modelo SVM com função de kernel linear permaneceu superior aos restantes modelos SVM, no entanto, o modelo ARIMA (0;0;1) conseguiu uma melhor taxa de acerto na previsão do sinal da variação da cotação do PSI20.

Palavras-Chave: PSI20; ARIMA; SVM; Forecasting

Classificação JEL: C52; C53; C58; G17

Abstract

This work compares the performance of ARIMA and machine learning models, specifically the SVM model, in predicting the evolution of the PSI20. For this purpose, the PSI20 closing price was collected between 21 November 2006 and 28 November 2018 (working days only) and transformed using logarithms. Both predictions were tested in a subsample corresponding to 30%.

Firstly, the ARIMA model was tested. The specification was chosen by means of the information criteria AIC, BIC and HQC. The information criteria did not select the same ARIMA model. The AIC test suggested the ARIMA (3; 0; 3) and the BIC and HQC suggested the ARIMA (0; 0; 1). Of these two models, the one with the lowest root mean square error (RMSE) is the ARIMA (0; 0; 1).

Then, the SVM model was tested. The software used (Gretl) selected an ε-SVR model. The remaining elements of the model were chosen according to the out-of-sample performance. The best SVM model is an ε-SVR model that uses a linear type kernel function. This is also the best among all models considered, but note that the RMSE varies very little across specifications.

The models were also used to predict the sign of the change of the PSI20. In this situation, the information criteria AIC, BIC and HQC were consensual regarding the specification of the ARIMA model, since all criteria point to the use of ARIMA (0; 0; 1). The SVM model with linear kernel function remained the best of the SVM models, however, the ARIMA (0; 0; 1) model achieved a better prediction rate of the PSI20 price change signal.

Key words: PSI20; ARIMA; SVM; Forecasting

JEL Classification: C52; C53; C58; G17

Índice

1.Introdução

Apesar da controvérsia que os rodeia, especialmente desde o início da crise financeira internacional (ver Alexandre, Martins, Andrade, Castro e Bação, 2009), os mercados financeiros são vistos por muitos agentes económicos como fundamentais para o desenvolvimento de uma economia (Hsu, Lessman, Shien-sung, Ma e Johnson, 2016). É neles que se faz a ligação entre os agentes económicos excedentários e os agentes económicos deficitários. Esta ligação promove uma afetação eficiente de recursos. No entanto, muitos agentes estão nos mercados financeiros com propósitos "especulativos", isto é, procurando obter ganhos com transações de muito curto prazo, possivelmente intra-diárias. Para estes, é de extrema importância perceber o momento exato para comprar ou vender um determinado ativo financeiro. Neste contexto, torna-se prioritário para os agentes tentar prever as flutuações nos mercados financeiros (Kambouroudis, McMillan e Tsakou, 2016).

Existem três grandes abordagens sobre a possibilidade de prever o comportamento dos mercados financeiros. A primeira, e a mais comum do ponto vista teórico, é a consideração da hipótese do passeio aleatório (*random walk*) que é compatível com a hipótese da eficiência dos mercados financeiros (Fama, 1965; Samuelson, 1965; Malkiel, 1973). A previsibilidade não significa que os mercados não são eficientes (Koijen e Van Nieuwerburgh, 2011), mas esta hipótese implica que a previsibilidade, a existir, seja muito limitada. A versão fraca da hipótese da eficiência dos mercados diz que as cotações refletem toda a informação passada disponível publicamente. A versão semiforte diz que, além disso, as cotações reagem imediatamente a nova informação. A versão forte acrescenta que as cotações refletem igualmente a informação que não está disponível publicamente. Deste modo, essencialmente, a hipótese de eficiência dos mercados financeiros implica que um investidor não consegue, em média, obter rentabilidades superiores aos restantes investidores se só usar informação disponível no mercado. A segunda abordagem sustenta a possibilidade de previsão na análise de indicadores "fundamentais" (macroeconómicos, sectoriais ou específicos à empresa), que possam ajudar a determinar o valor dos ativos, pressupondo que os restantes participantes no mercado não estão a fazer a mesma avaliação. A terceira abordagem baseia-se na "análise técnica", procurando extrair dos dados uma tendência (ainda que de muito curto prazo) de evolução do mercado, isto é, uma componente com alguma previsibilidade.

Neste trabalho serão usados dois modelos de previsão: o modelo *auto-regressive integrating moving average* (ARIMA) e o modelo *support vector machine* (SVM). O modelo ARIMA é um modelo importante por ser de uso comum e por o modelo ARIMA (1;0;0) corresponder ao *random walk*, ou seja, caso este seja o melhor modelo, não será possível obter de forma consistente rentabilidades acima do normal (mesmo ignorando custos de transação). O outro modelo em questão, o modelo SVM, é um modelo cuja utilização tem vindo a aumentar e a tornar-se bastante popular nesta área.

O trabalho começará por uma breve secção na qual estudos relacionados com ambos os modelos ou com o comportamento do PSI20 são apresentados de forma sintetizada. De seguida passa-se a uma secção de descrição dos modelos e do indicador de comparação das previsões. Na quarta secção são apresentados e comparados os resultados obtidos pelos modelos selecionados. A última secção resume as principais conclusões do trabalho.

2.Revisão da Literatura

Nesta secção serão apresentadas as perspetivas e as conclusões de vários autores relativamente à utilização de métodos ARIMA e *machine learning* na previsão de preços de ativos financeiros. Como foi referido anteriormente, a teoria económica conclui que as características próprias dos mercados financeiros tornam muito difícil a previsão da evolução do preço de um ativo financeiro. A possível descoberta de "anomalias" nas quais se possam basear estratégias de transação lucrativas é motivo de grande interesse por parte dos investidores. A questão é, então, saber quais os métodos de previsão que melhor conseguem prever a variação dos preços dos ativos financeiros, se é que algum consegue ser melhor do que o modelo do passeio aleatório.

Ao longo deste trabalho o objetivo será tentar perceber qual dos dois métodos tem um desempenho melhor na previsão da variação do PSI20. O PSI20 é constituído por, no máximo, as 20 empresas com maior capitalização de mercado em Portugal na Euronext Lisboa, podendo, no entanto, apenas ter 18. Para poderem fazer parte do índice, as empresas devem apresentar níveis elevados de liquidez na bolsa e ter no mínimo 100 milhões de euros de capitalização. O índice pode ser ajustado de forma a garantir que o índice seja representativo do mercado. Para o efeito, é realizada uma revisão anual em Março para verificar se é ou não necessário esse ajustamento.

Hsu *et al*. (2016) fizeram uma comparação entre modelos *machine learning* e os melhores modelos econométricos na previsão do sinal da variação de índices de 34 mercados no período 2008-2014. Relativamente aos métodos *machine learning*, os modelos SVM mostraram-se superiores aos modelos *Artificial Neural Networks* (ANN) na previsão da evolução de séries temporais nos mercados financeiros. Por sua vez, relativamente aos modelos econométricos, os modelos ARMA, mais especificamente os modelos AR, são os que apresentam melhores resultados. No final é ainda feita uma comparação entre os modelos SVM e o modelo AR, tendo o modelo SVM apresentado o melhor desempenho.

Garcia (2017) comparou modelos GARCH e modelos ARMA na previsão de séries temporais, mais especificamente os retornos logaritmizados do preço de fecho do PSI20. O período amostral corresponde a registos diários (5 dias úteis por semana) do valor de fecho entre 1 de Novembro de 2004 e 18 de Agosto de 2016. Nesse estudo, verificou-se que existia evidências de efeitos ARCH, ou seja, a variância não era constante, sendo o modelo mais adequado o modelo GARCH. No entanto, apesar de o modelo GARCH ser aparentemente o mais apropriado para a série, um modelo ARMA (em particular o modelo MA (1)) apresentou melhores resultados.

Mullainathan e Spiess (2017) apresentam uma introdução aos métodos *machine learning* no contexto dos métodos habitualmente usados na área da econometria. O tipo de métodos *machine learning* mais utilizado por economistas é o *Least Absolute Shrinkage and Selection Operator* (LASSO) que corresponde à utilização de uma função de perda quadrática, uma função linear para relacionar *inputs* e *output* e um regularizador que penaliza a soma dos valores absolutos dos coeficientes. Os modelos *machine learning*, mais especificamente os modelos SVM, podem ser utilizados para além da sua principal função (separação de dados, ou classificador de dados) como um modelo de previsão econométrico. A grande vantagem deste tipo de modelos face aos modelos econométricos tradicionais é a sua capacidade de descobrir padrões generalizáveis numa série temporal e dessa forma encontrar funções que consigam uma boa previsão para a evolução dessas séries. A grande desvantagem é a inexistência de parâmetros interpretáveis à luz de alguma teoria.

Shen, Jiang e Zhang (2012) estudaram a utilização dos métodos *machine learning* (SVM; ANN e *Random Forest*) na previsão da direção da variação dos índices *Bombay Stock Exchange* (BSE) e CNX Nifty através da utilização destes métodos em duas etapas (sendo a primeira etapa sempre composta pelo modelo SVM). No fundo, correspondia à

aplicação e comparação dos modelos híbridos SVM-SVM, SVM-ANN e SVM-RF. A amostra utilizada tem em conta um horizonte temporal de 10 anos. Este estudo concluiu que o melhor modelo era o modelo hibrido SVM-ANN. No entanto, é também possível verificar que o modelo SVM é o que apresenta melhores resultados nos modelos de uma etapa.

Lahmiri (2011) estudou a utilização de modelos Probabilistic Neural Network (PNN) e modelos SVM na previsão do índice S&P500 através da utilização de informação económica e informação técnica. Para o efeito, foi utilizada uma amostra constituída pelos retornos diários mais as variáveis económicas que o autor considerou estarem relacionadas (taxas de juro e de câmbio) para um horizonte temporal de 11 de Janeiro de 2000 a 31 de Janeiro de 2008. A principal conclusão deste estudo é que a capacidade de previsão melhora quando o modelo SVM utiliza a informação económica adicional.

Pascoal e Monteiro (2014) estudaram a eficiência de mercado, a irregularidade e a existência de efeitos de muito longo prazo do índice PSI20 e concluíram que os retornos da série do PSI20 são imprevisíveis, a série é extremamente irregular (mais irregular que uma série com distribuição normal) e não apresenta efeitos de muito longo prazo, apesar da volatilidade persistente. Os autores apontam estas características como sendo próprias de um mercado financeiro eficiente. Este estudo tem em conta um horizonte temporal entre 2000 e 2013 e para o efeito foi utilizado o preço de fecho.

3.Metodologia

O preço de fecho do PSI20 foi obtido no *website* da Euronext e diz respeito aos dias úteis entre 21 de Novembro de 2006 e 28 de Novembro de 2018. Para reduzir a possibilidade de *overfitting*, será deixada de fora da amostra usada na estimação dos parâmetros uma proporção de aproximadamente 30%, respeitante ao período de 23 de Fevereiro de 2015 a 28 de Novembro de 2018. O desempenho dos modelos será comparado nesta parte da amostra. A percentagem da amostra a deixar de parte de forma a testar os modelos foi escolhida de acordo com o que tem sido feito em estudos prévios sobre o modelo SVM (Madge e Bhatt, 2015). A variável que será objeto de análise é a primeira diferença do logaritmo do preço de fecho de cada sessão diária, que é uma medida aproximada (pois não inclui os dividendos) do retorno do investimento nos títulos constituintes do índice, esta é a variável normalmente estudada em trabalhos semelhantes a este.

Como indicador representativo da qualidade da previsão e base de comparação entre os modelos será utilizado o *root mean square error* (RMSE), ou seja, a raiz quadrada do erro quadrático médio.

O RMSE é calculado da seguinte forma:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (1)$$

Na equação 1, n é o número de previsões, $\hat{y}_i$ é a i-ésima previsão, y_i é o valor efetivamente observado. O melhor modelo será aquele para o qual o RMSE é menor. Como se verá à frente, para explorar devidamente as características do modelo SVM, fazer-se-à também fazer previsões do sinal da variação do PSI20. Neste caso, a medida da qualidade das previsões será a taxa de acerto (a percentagem de previsões corretas).

3.1. Modelo ARIMA

Os modelos ARIMA são muito utilizados na previsão. Combinam a facilidade de utilização (dada a disponibilidade de *software*) com a capacidade para representarem aproximadamente o comportamento de qualquer série estacionária ou integrada. A sigla ARIMA significa *auto-regressive integrating moving average*. O modelo ARIMA (p,d,q) é um modelo ARMA (p,q) para a d-ésima diferença da variável dependente. Um modelo ARMA (p,q) contém *p* desfasamentos da variável dependente (eventualmente com coeficientes nulos, exceto para o *p*-ésimo desfasamento) e *q* desfasamentos (eventualmente com coeficientes nulos, exceto para o *q*-ésimo desfasamento) do ruído branco que introduz variação aleatória no comportamento da variável dependente.

ARMA(p;q):

$$y_t = c + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (2)$$

Os valores de *p* e *q* serão determinados de acordo com os critérios de informação *Akaike information criterion* (AIC), *Bayesian information criterion* (BIC) e *Hannan–Quinn information criterion* (HQC). O número máximo de desfasamentos foi fixado em quatro, tanto para a componente autorregressiva como para a componente média móvel. Mais do que quatro desfasamentos conduzia frequentemente a problemas na estimação do modelo.

3.2. Modelo SVM

Machine learning é o nome dado a algoritmos computacionais que usam informação anterior sobre o seu desempenho para irem melhorando esse desempenho. Entre os métodos de *machine learning* encontram-se os modelos *support vector machine* (SVM). Segundo Madge (2015), devido às características dos mercados financeiros, a utilização de *machine learning* na tentativa de prever a evolução das séries financeiras está a aumentar exponencialmente nos últimos tempos, sendo ANN e SVM os dois modelos mais utilizados para o efeito. A versão tradicional do modelo SVM é adequada para problemas de classificação de dados, ou seja, quando a variável a prever é discreta. Será este o caso quando se fizer a previsão do sinal da variação do PSI20. A ideia fundamental do algoritmo é encontrar hiperplanos que separarem as regiões do espaço em que a variável dependente toma os diferentes valores que pode apresentar. Esses hiperplanos estarão enquadrados por certos vetores de observações das variáveis explicativas (os *support vectors*).

Neste estudo, será utilizado o modelo *support vector regression* (SVR) do género ε-SVR, ou seja, um SVM em que a variável a prever pode ser contínua e em que se estima uma função que relaciona essa variável com outras (a "regressão" a que o nome do método alude). O algoritmo tem três elementos que importa destacar. Em primeiro lugar, pretende-se que a função seja o mais "suave" possível, procurando-se minimizar a norma dos seus coeficientes. Em segundo lugar, pretende-se que os erros estejam, em valor absoluto, abaixo de um certo limiar (ε), não interessando o seu valor exato desde que aquela condição seja cumprida. Caso tal não seja possível, serão admitidos excessos relativamente a ε, atribuindo-se-lhes uma penalização linear. Em terceiro lugar, a função subjacente pode não ser linear. A não linearidade é modelada através de uma função *kernel* apropriada, isto é, que possa representar um produto interno entre dois vetores resultantes duma certa transformação não linear dos valores observados das variáveis independentes e do vetor correspondente ao ponto no qual se quer avaliar a função. Os valores previstos pelo modelo serão combinações dos valores que a função *kernel* (ou o produto interno, no caso independentes.

No caso do Gretl, a função de kernel pode corresponder a uma função linear, polinomial, *radial basis function* (RBF) ou sigmoid. No presente caso, a função de kernel a utilizar será aquela que melhor se ajustar aos dados da amostra.

A função de kernel do tipo linear corresponde ao produto interno:

$$K(u;v) = u'v \quad (3)$$

A função de kernel do tipo polinomial acrescenta termos de grau superior (interações):

$$K(u;v) = (\gamma u'v + c)^d \quad (4)$$

Na equação 4, $\gamma > 0$, d é o grau do polinómio e c é outro parâmetro da função kernel.

A função RBF tem efeito de filtro, reduzindo a presença de ruídos de alta frequência.

$$RBF(u;v) = e^{-\gamma\|u-v\|^2}, \text{com } \gamma > 0 \quad (5)$$

A função sigmoid não é uma função de kernel apropriada (para todos os possíveis valores dos parâmetros), mas é utilizada com sucesso na prática. É dada pela expressão:

$$K(u;v) = \tanh(\gamma\, u'v + c) \quad (6)$$

Na equação 6, *tanh* é a tangente hiperbólica, γ e c são parâmetros da função kernel.

É de destacar que os parâmetros a utilizar foram determinados experiencialmente tendo sempre como objetivo a minimização do RMSE.

4. Resultados

Nesta secção serão apresentados os resultados da utilização dos modelos acima referidos para prever os retornos do PSI20 e o seu sinal. A amostra utilizada neste trabalho pode ser observada no Anexo 1. No Anexo 2 encontra-se o resultado do teste de raiz unitária ADF, que rejeita a hipótese nula de a série ter uma raiz unitária. A utilização de modelos ARIMA com zero diferenças (ou seja, modelos ARMA) encontra justificação nesse resultado.

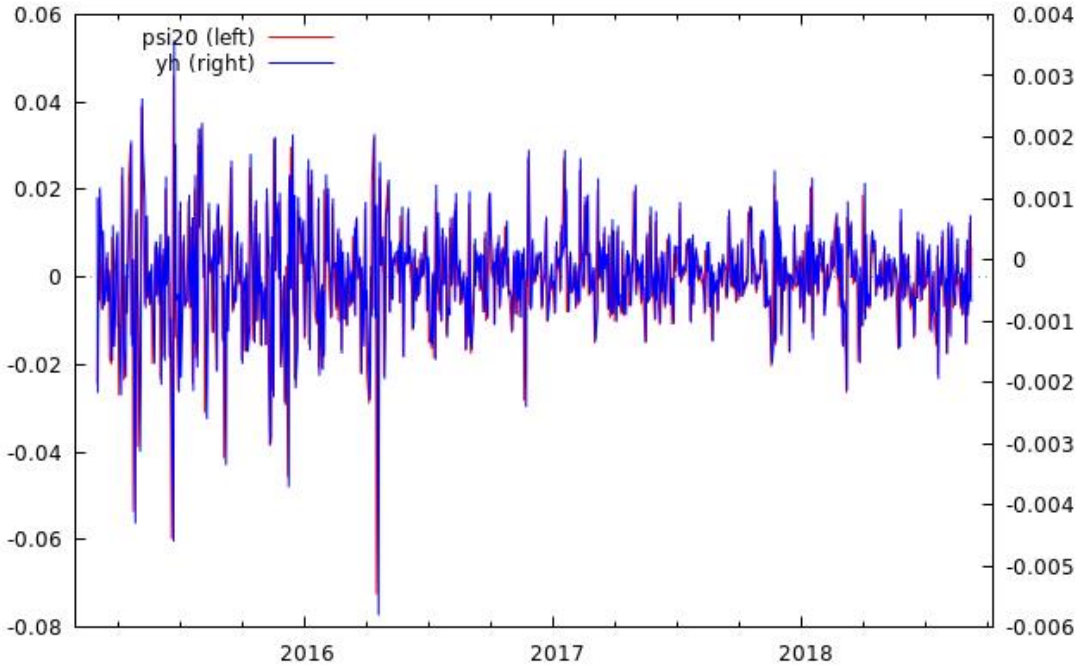
4.1. Previsão da taxa de variação da cotação do PSI20

O modelo ARIMA que melhor se ajusta à série em questão é o modelo ARIMA (0;0;1) de acordo com dos critérios BIC e HQC, tal como é visível no *Anexo 3*. O critério de informação AIC sugeriu o modelo ARIMA (3;0;3). Note-se que os valores no Anexo 3

mostram que as diferenças entre os modelos são muito pequenas, o que pode indiciar a existência de grande incerteza/dificuldades na estimação.

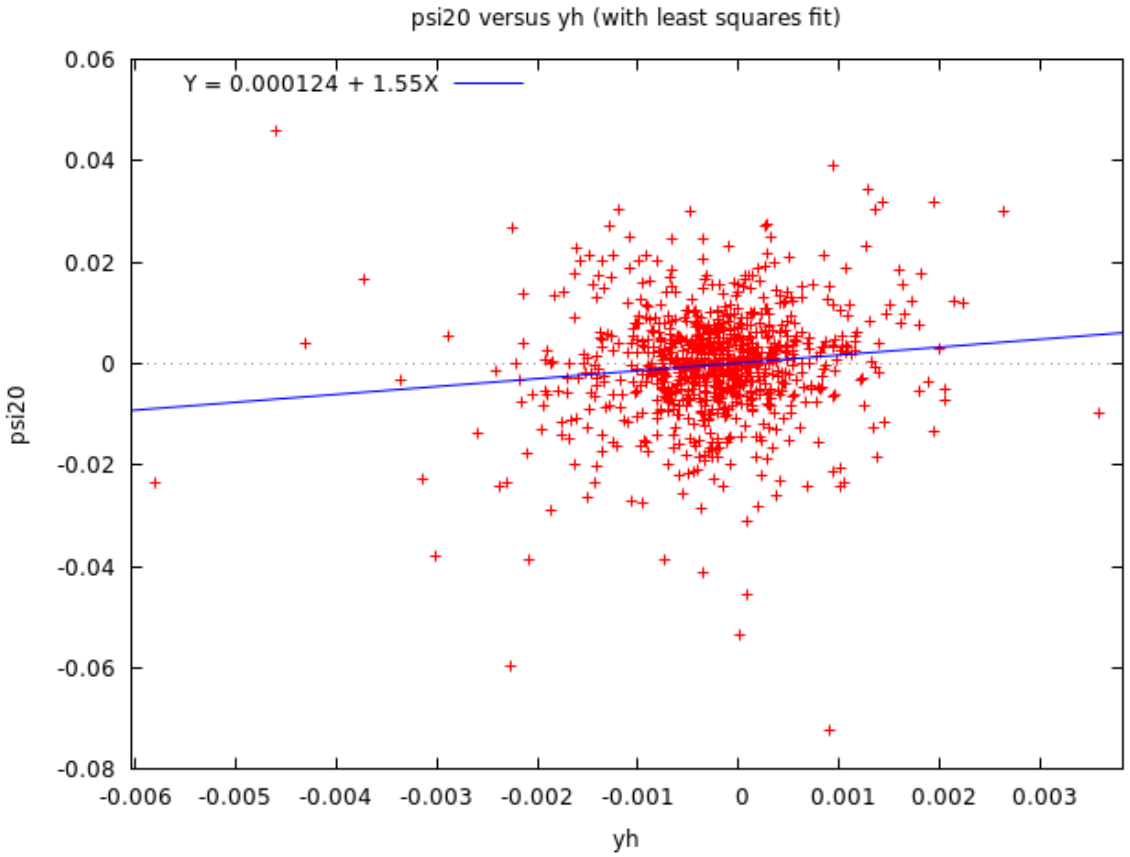
As previsões (*yh*) *out-of-sample* produzidas pelo modelo ARIMA (0;0;1) estão na Figura 1. Os coeficientes estimados podem ser vistos no Anexo 4.

Figura 1: Série temporal e previsão ARIMA (0;0;1)



Fonte: cálculos do autor

Figura 2: Dispersão entre valores observados e previsões ARIMA (0;0;1)

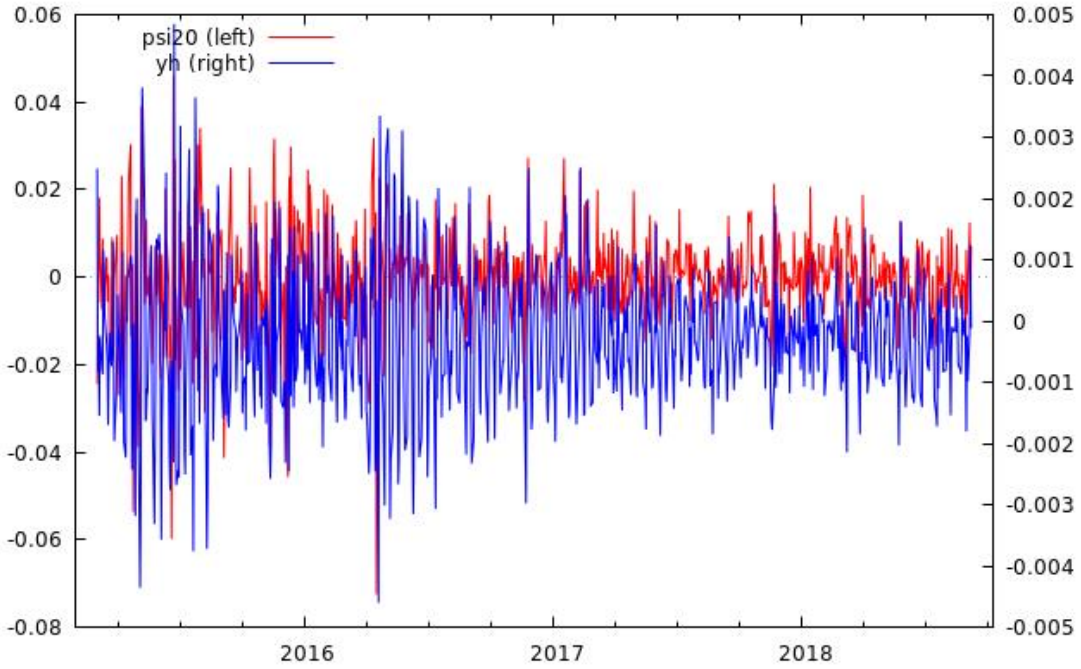


Fonte: cálculos do autor

A Figura 1sugere que o modelo não é bem-sucedido na previsão dos retornos. Note-se que as escalas dos valores observados e das previsões são muito diferentes. Para avaliar melhor o desempenho do modelo, a Figura 2 mostra um diagrama de dispersão com as mesmas duas séries, que inclui igualmente a linha de regressão entre essas séries. Apesar de significativa, a relação entre as previsões e os valores observados é fraca. O modelo apresentado tem um RMSE de 0,0108786 na previsão *out-of-sample*, inferior ao RMSE na amostra usada na estimação (0,0137527). Note-se que o desvio padrão da variável dependente é 0,013795 na amostra usada na estimação e 0,010947 *out-of-sample*.

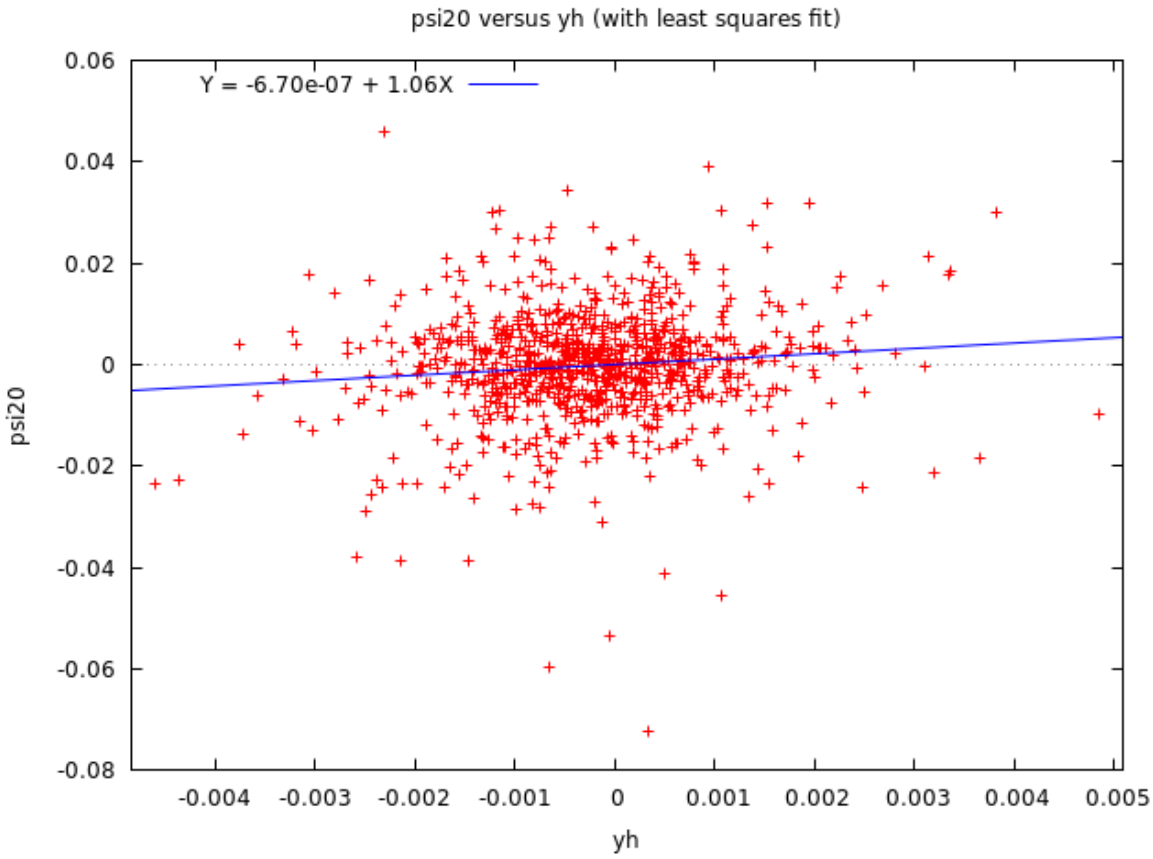
O modelo sugerido pelo critério de informação AIC, como já tinha sido referido, é o modelo ARIMA (3;0;3). A Figura 3 apresenta as previsões obtidas *out-of-sample* com este modelo e o Anexo 5 contém os coeficientes estimados para este modelo.

Figura 3: Série temporal e previsão ARIMA (3;0;3)



Fonte: cálculos do autor

Figura 4: Dispersão entre valores observados e previsões ARIMA (3;0;3)



Fonte: cálculos do autor

Tal como no caso do ARIMA (0;0;1), a escala dos valores observados é bastante diferente da escala dos valores previstos pelo modelo ARIMA (3;0;3). A Figura 4 mostra que há uma relação significa, e sem enviesamentos, entre as previsões e os valores observados. O modelo ARIMA (3;0;3) apresenta um RMSE de 0,0108782 na previsão *out-of-sample*, inferior ao RMSE na amostra usada na estimação (0,0137035), mas tal refletirá essencialmente a menor volatilidade dos retornos.

No caso do modelo SVM, experimentámos as quatro funções *kernel* (linear, polinomial, RBF e sigmoid) que o software Gretl disponibiliza (na versão 2018c). O procedimento depende de certos parâmetros, alguns dos quais dependem da função *kernel* escolhida (recorde-se a descrição feita na secção 3.2). Para a função *kernel* RBF, o Gretl procede automaticamente à otimização dos parâmetros ε (tolerância aos erros), C (penalização dos excessos dos erros relativamente à tolerância) e γ (parâmetro da própria função *kernel*). Esta otimização é feita na amostra de treino através de *cross-validation*. Por outras palavras, dado um certo conjunto de valores para aqueles parâmetros, o modelo SVM é treinado em subamostras da amostra de treino, calculando-se previsões para a parte restante da amostra de treino. Os valores escolhidos para aqueles parâmetros são os que minimizam o RMSE das previsões.

Fez-se uso desta funcionalidade do Gretl para o *kernel* RBF. Para as restantes funções foi adotado um procedimento de procura em grelha (com um número reduzido de pontos, pois o procedimento é bastante demorado), minimizando o RMSE na amostra de treino. A vantagem da *cross-validation* relativamente a este procedimento reside no facto de este procedimento favorecer a ocorrência de *overfitting*, ou seja, os parâmetros são otimizados para a amostra de treino, mas não são válidos fora dessa amostra, produzindo más previsões. No caso do *kernel* linear, otimizou-se os parâmetros ε e C. No *kernel* polinomial otimizou-se os parâmetros ε, C, γ e d (o grau do polinómio). O parâmetro c não foi optimizado, tomando o valor zero (o *default* do Gretl). No *kernel* sigmoid optimizou-se os parâmetros ε, C e γ. Novamente, o parâmetro c tomou o valor zero. Os valores otimizados, bem como o RMSE na amostra de treino e na amostra de teste (*out-of-sample*) estão na Tabela 1. Note-se que, para cada função *kernel*, o procedimento foi feito quatro vezes, usando um, dois, três e quatro desfasamentos do retorno como variáveis independentes.

Tabela 1: parâmetros otimizados e RMSE dos modelos SVM

Kernel	Desfasamentos	ε	*C*	γ	*d*	RMSE treino	RMSE previsão
Linear	1	0,00917633	1,25	-	-	0,0137529	0,0108769
	2	0,00917633	4,25	-	-	0,0137526	0,0108761
	3	0,00917633	2,75	-	-	0,0137522	0,0108764
	4	0,00968612	1,5	-	-	0,0137527	0,0108825
Polinomial	1	0,00764694	4,5	3,75	2	0,0137824	0,0109404
	2	0,00662735	2,25	0,5	4	0,0137392	0,0109926
	3	0,0101959	5	4,75	4	0,0135994	0,0110616
	4	0,0101959	4,5	4	4	0,0133086	0,0113552
RBF	1	0,1	0,03125	0,0000305176	-	0,0137939	0,0109435
	2	0,1	0,03125	0,0000305176	-	0,0137939	0,0109435
	3	0,1	0,03125	0,0000305176	-	0,0137939	0,0109435
	4	0,1	2	0,0000305176	-	0,0137946	0,010943
Sigmoid	1	0,00917633	0,25	0,25	-	0,0137541	0,0108778
	2	0,00968612	0,25	0,25	-	0,0137563	0,0108783
	3	0,0101959	0,25	0,25	-	0,0137717	0,0108744
	4	0,0101959	0,25	0,25	-	0,0137816	0,0109407

Fonte: cálculos do autor.

Os resultados apresentados na Tabela 1 sugerem que o procedimento originou de facto *overfitting*, pelo menos no caso do *kernel* polinomial, caso em que se observa uma deterioração do desempenho *out-of-sample* em simultâneo com uma melhoria do desempenho na amostra de treino. Para o *kernel* sigmoid também se observa que a versão com melhor desempenho *out-of-sample* apresenta o pior desempenho *in-sample*. O mesmo modelo é, aliás, o que tem menor RMSE *out-of-sample* (embora seja de notar que as diferenças são extremamente pequenas, refletindo o facto de esta série de retornos ser, essencialmente, imprevisível, pelo menos com os modelos aqui utilizados). No caso do *kernel* RBF, os três primeiros desfasamentos produzem exatamente os mesmos resultados. Os gráficos relativos a estas previsões encontram-se no Anexo 6.

A Tabela 2 reúne, para maior facilidade de análise, os RMSEs obtidos na amostra de treino e *out-of-sample* pelos melhores modelos (no caso dos modelos ARIMA, em termos dos critérios de informação; no caso dos modelos SVM, em termos do RMSE na amostra de treino).

Tabela 2: RMSE das previsões

kernel	RMSE treino	RMSE previsão
ARIMA (0,0,1)	0,0137527	0,0108786
ARIMA (3,0,3)	0,0137035	0,0108782
Linear (3)	0,0137522	0,0108764
Polinomial (4)	0,0133086	0,0113552
RBF (1-3)	0,0137939	0,0109435
Sigmoid (1)	0,0137541	0,0108778

Fonte: cálculos do autor (entre parênteses: o número de desfasamentos)

Entre os modelos presentes na Tabela 2, o modelo que apresenta o menor RMSE é o modelo SVM com um *kernel* linear. Contudo, é importante voltar a referir que as diferenças são muito pequenas (na Tabela 2, apenas o modelo com o *kernel* polinomial tem um desempenho claramente inferior ao dos restantes).

4.2. Previsão do Sinal da taxa de variação da cotação do PSI20

Os modelos SVM também têm sido muito usados na previsão do sinal da variação da cotação, ou seja, para prever se a cotação aumenta ou diminui. Esta análise permite-nos verificar se os melhores modelos a prever a taxa de variação são também os melhores a

prever o sinal da variação da cotação do ativo. Relativamente ao tratamento dos dados, é atribuído o valor 1 a uma variação positiva e 0 a uma variação negativa ou nula.

Na previsão do sinal da variação da cotação do ativo, continua-se a usar os mesmos modelos ARIMA, transformando o valor previsto em 1 se for positivo e em 0 no caso contrário. Quanto aos modelos SVM, continua-se a usar as quatro funções *kernel*, mas o número de desfasamentos do sinal da variação não foi otimizado (são sempre quatro desfasamentos). Também não foram optimizados os restantes parâmetros, com a exceção de quando se trata do *kernel* RBF, uma vez que neste caso se pode fazer uso da funcionalidade de otimização dos parâmetros oferecida pelo Gretl.

Tabela 3: Previsão do sinal da taxa de variação do PSI20

Modelo	Observado	Previsto		Taxa de acerto	
ARIMA(0,0,1)		0	1	treino	teste
	0	319	134	0,526927	0,54902
	1	280	185		
ARIMA(3,0,3)		0	1	treino	teste
	0	290	163	0,522748	0,528322
	1	270	195		
Linear		0	1	treino	teste
	0	239	214	0,529248	0,533769
	1	214	251		
Polinomial		0	1	treino	teste
	0	246	207	0,535283	0,526144
	1	228	237		
RBF		0	1	treino	teste
	0	114	339	0,545497	0,528322
	1	94	371		
Sigmoid		0	1	treino	teste
	0	221	232	0,500929	0,476035
	1	249	216		

Fonte: cálculos do autor (0 corresponde a descida e 1 a subida da cotação).

É possível verificar que o modelo SVM com *kernel* RBF tem o melhor desempenho na amostra de treino, possivelmente em resultado da otimização dos parâmetros. Porém,

parece ter havido *overfitting*, pois fora da amostra fica abaixo do desempenho de outros modelos, nomeadamente do ARIMA (0,0,1), que é o melhor modelo nesta comparação. Note-se que o modelo com *kernel* sigmoid tem o pior desempenho, tanto na amostra de treino como *out-of-sample*, não conseguindo neste caso nem sequer atingir o nível de uma regra simples que faça sempre a mesma previsão (que teria uma taxa de acerto de 50,7% se a previsão for sempre de subida, e de 49,3% se for sempre de descida).

5. Conclusão

Prever a variação dos preços de ativos financeiros é algo bastante difícil. A incerteza e a variabilidade elevada própria dos mercados financeiros tornam esta tarefa extremamente complicada. No entanto, tentar compreender o que pode vir a acontecer nos mercados financeiros é de extrema importância para aqueles que planeiam investimentos, tentam alocar os recursos de forma eficiente e têm como função gerir o risco dos seus investimentos.

De forma a conseguir resolver esse problema, têm sido utilizadas técnicas cada vez mais complexas e mais elaboradas. No entanto, resta ainda saber de que forma é que estas novas técnicas conseguem melhorar o desempenho da previsão dos preços dos ativos financeiros.

Este trabalho tem como objetivo fazer uma comparação entre o modelo ARIMA (o modelo mais comum) e um modelo de *machine learning* – o modelo SVM – na previsão dos retornos do índice PSI20. O melhor modelo na previsão da variação do PSI20 para o período considerado foi o modelo SVM com um *kernel* linear, mas por uma margem muito pequena.

É possível que resultados melhores possam ser alcançados através de uma otimização mais fina dos parâmetros dos modelos. Outro tema a abordar em trabalhos futuros poderá ser a utilização de modelos híbridos, combinando ambos os modelos aqui utilizados, o modelo ARIMA e o modelo SVM. O modelo ARIMA é um modelo muito conhecido pelas capacidades de forecasting em séries temporais lineares; em contrapartida, o modelo SVM destaca-se pela sua capacidade de conseguir realizar forecasting em séries financeiras não lineares. Seria interessante comparar os resultados deste modelo híbrido com os modelos ARIMA e SVM separadamente.

Relativamente à previsão do sinal da taxa de variação da cotação do PSI20, o modelo ARIMA (0;0;1) gerou as melhores previsões, conseguindo acertar em 54,9% da amostra. Novamente, o modelo com *kernel* linear foi o melhor entre os modelos SVM.

Referências Bibliográficas

Alexandre, F., Martins, I.G., Andrade, J.S., Castro, P.R., & Bação, P. (2009). *A Crise Financeira Internacional*. Imprensa da Universidade de Coimbra.

Fama, E.F. (1965). Random walks in stock market prices. *Financial Analysts Journal*, 21 (5), 55-59.

Garcia, E. M. A. (2017). *Previsão de séries temporais financeiras: o caso PSI 20* . Dissertação de Mestrado. Faculdade de Ciências da Universidade de Lisboa. Disponível em http://repositorio.ul.pt/bitstream/10451/30762/1/ulfc121594_tm_Elsa_Garcia.pdf

Hsu, M. W., Lessmann, S., Sung, M. C., Ma, T., & Johnson, J. E. (2016). Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, *61*, 215-234.

Kambouroudis, D. S., McMillan, D. G., & Tsakou, K. (2016). Forecasting stock return volatility: a comparison of GARCH, Implied volatility, and realized volatility models. *Journal of Futures Markets*, *36*(12), 1127-1163.

Koijen, R.S.J., & Van Nieuwerburgh, S. (2011). Predictability of Returns and Cash Flows. *Annual Review of Financial Economics*, 3 (1), 467-491.

Lahmiri, S. (2011). A comparison of PNN and SVM for stock market trend prediction using economic and technical information. *International Journal of Computer Applications*, *29*(3), 24-30.

Madge, S., & Bhatt, S. (2015). Predicting Stock Price Direction using Support Vector Machines. *Independent Work Report Spring*.

Malkiel, B.G. (1973). *A Random Walk Down Wall Street*. W.W. Norton & Company.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31* (2), 87-106.

Pascoal, R., & Monteiro, A. M. (2014). Market efficiency, roughness and long memory in PSI20 index returns: Wavelet and entropy analysis. *Entropy*, *16* (5), 2768-2788.

Euronext (2018), *Index Rule Book, PSI 20 Index*, Disponível em: www.indices.euronext.com

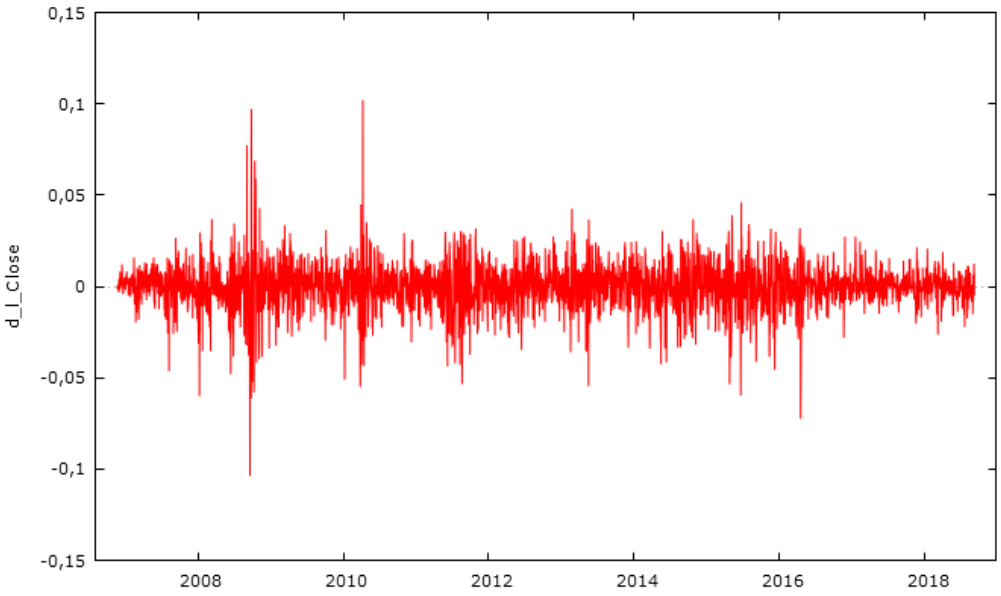
Samuelson, P. (1965). Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review*, 6 (2), 41-49.

Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, 1-5.

Anexos

Anexo 1

Figura 5: Série do preço de fecho do PSI20 em diferença de logaritmos



Fonte: cálculos do autor

Anexo 2

Teste ADF para os retornos do PSI20

Teste Aumentado de Dickey-Fuller para psi20
testar para baixo a partir de 28 desfasamentos, critério AIC
dimensão de amostragem 3060
hipótese nula de raiz unitária: a = 1

teste com constante
incluindo 15 desfasamentos de (1-L)psi20
modelo: (1-L)y = b0 + (a-1)*y(-1) + ... + e
valor estimado de (a - 1): -0,950587
estatística de teste: tau_c(1) = -13,1799
valor p assimptótico 1,597e-029
coeficiente de 1ª-ordem para e: -0,001
diferenças desfasadas: F(15, 3043) = 2,338 [0,0025]

Anexo 3

Modelos selecionados pelos critérios de informação:

ARIMA	AIC	BIC	HQC
(0,0,1)	-12348	-12331	-12341
(3,0,3)	-12353	-12307	-12336

Fonte: cálculos do autor

Anexo 4

Coeficientes do modelo ARIMA (0;0;1)

Variável	**Coeficiente**	**Erro Padrão**	**Valor p**
Const	−0,000249312	0,000318740	0,4341
theta_1	0,0756883	0,0214399	0,0004

Fonte: cálculos do autor

Anexo 5

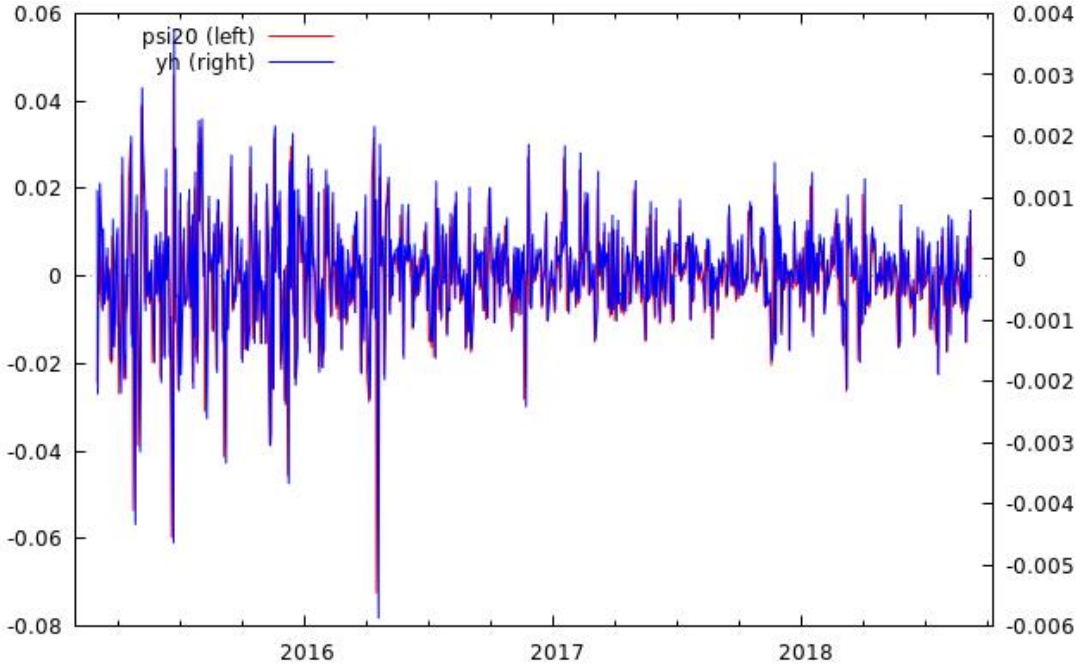
Coeficientes do modelo ARIMA (3;0;3):

Variável	**Coeficiente**	**erro padrão**	**valor p**
Const	−0.000250196	0.000312685	0.4236
phi_1	1.40323	0.00654595	0.0000
phi_2	−1.01265	0.00585847	0.0000
phi_3	0.0142305	0.00597718	0.0173
theta_1	−1.33237	0.0227363	0.0000
theta_2	0.913303	0.0311107	1.97e-189
theta_3	0.0493934	0.0225847	0.0287

Fonte: cálculos do autor

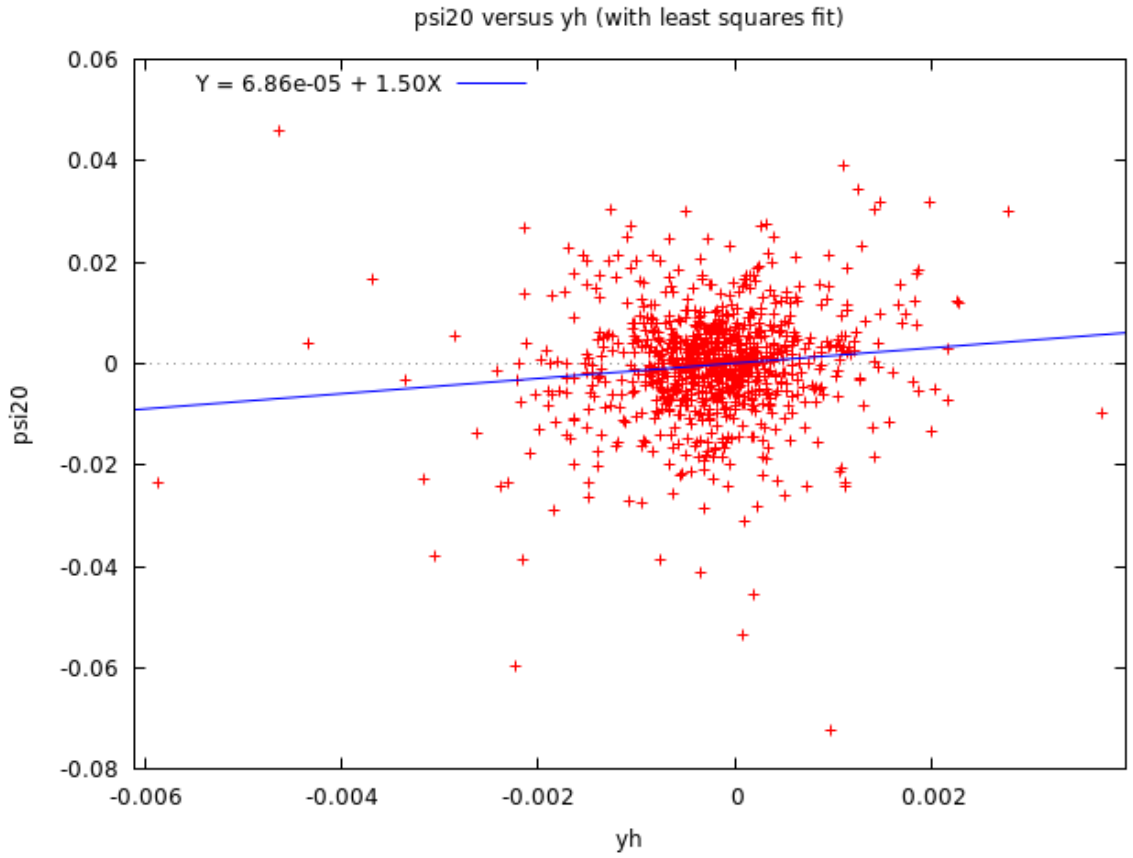
Anexo 6

Figura 6: Série temporal e previsão SVM kernel Linear



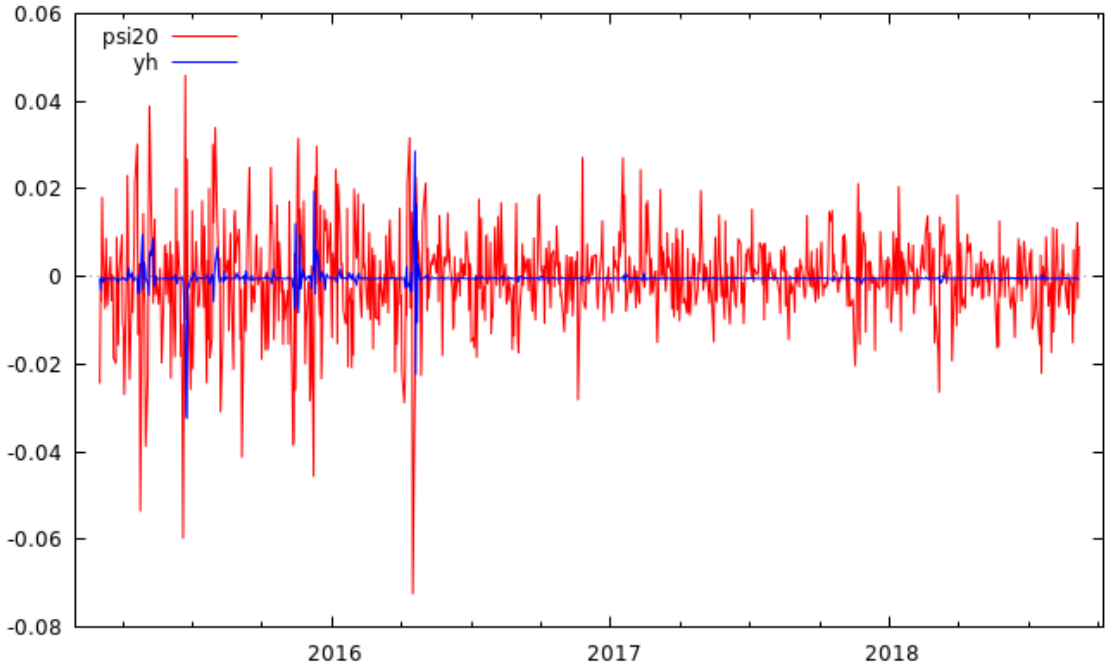
Fonte: cálculos do autor

Figura 7: Dispersão entre valores observados e previsões SVM kernel linear



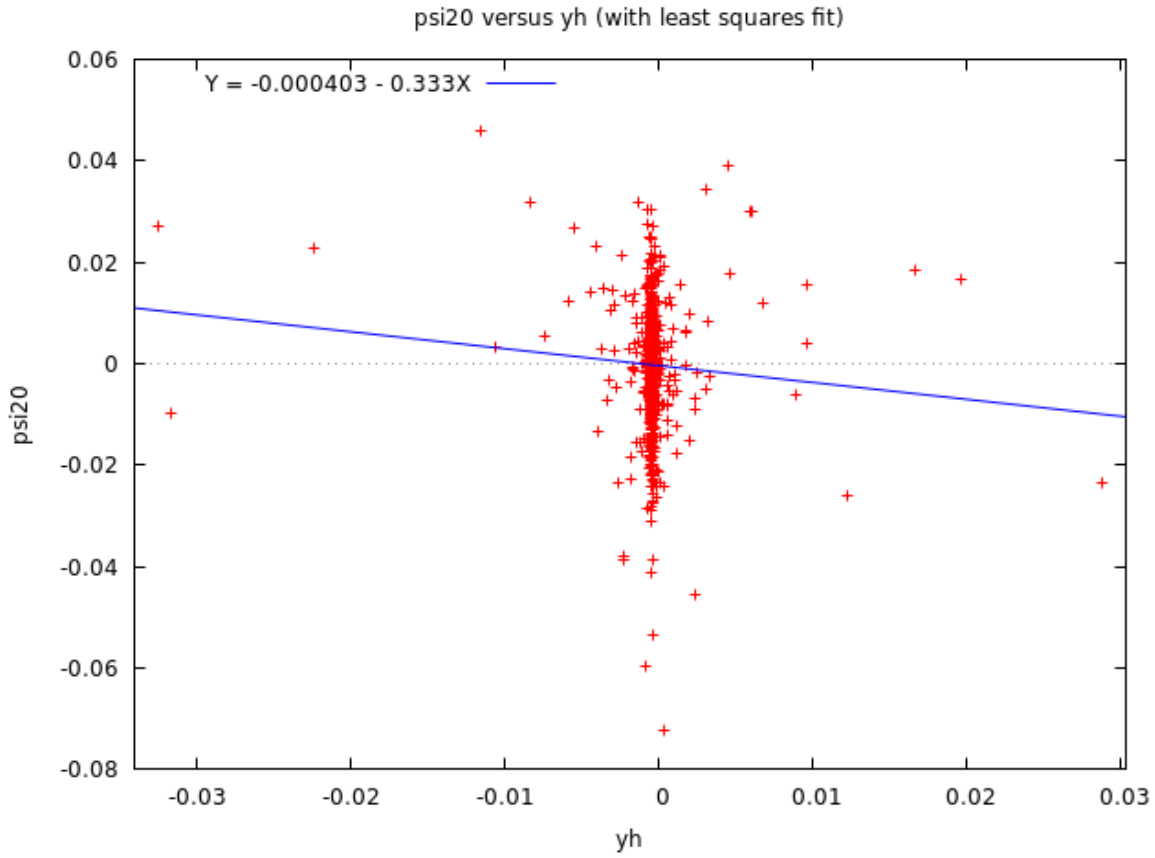
Fonte: cálculos do autor

Figura 8: Série temporal e previsão SVM kernel polinomial



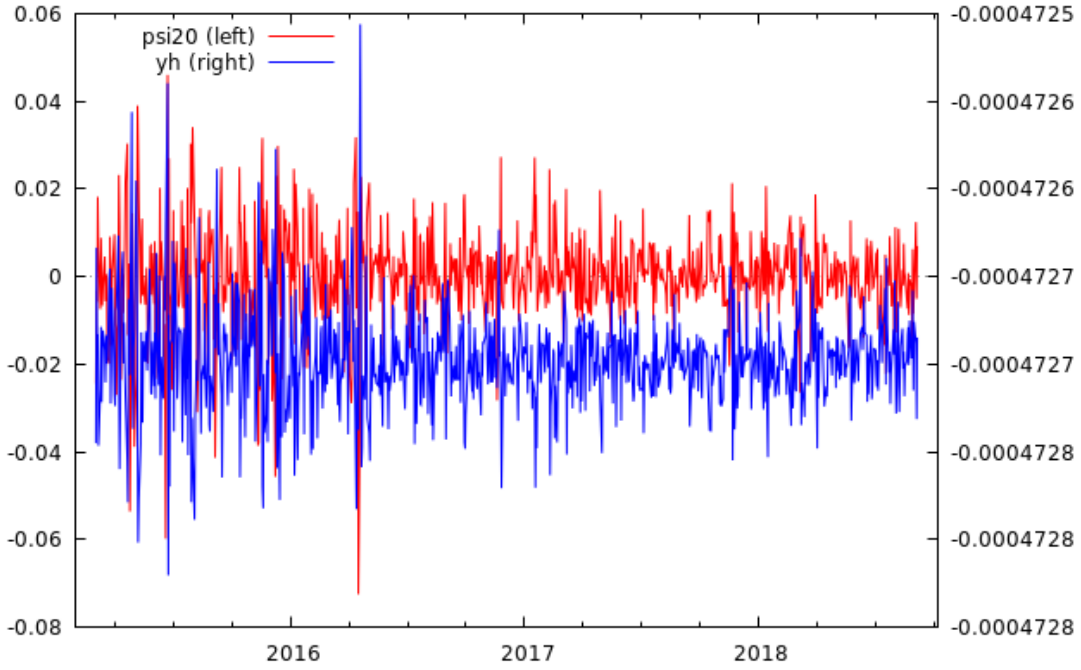
Fonte: cálculos do autor

Figura 9: Dispersão entre valores observados e previsões SVM kernel polinomial



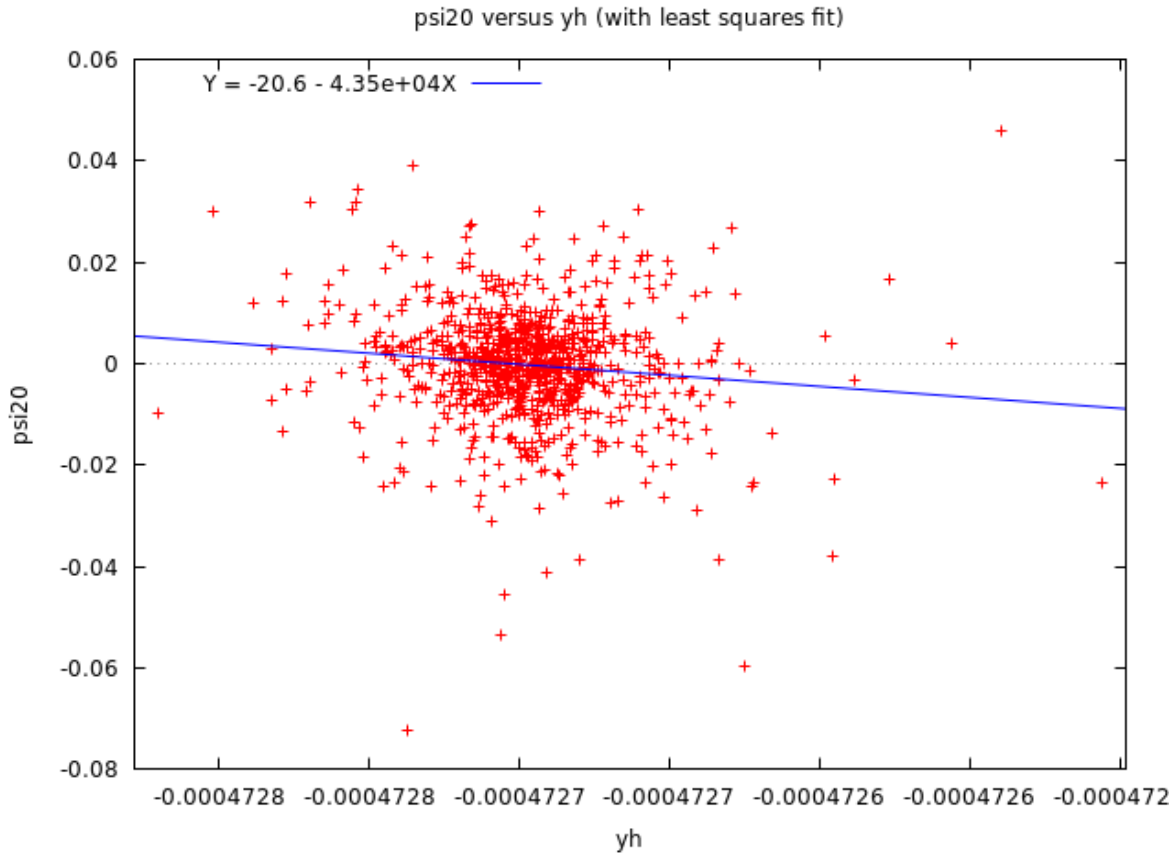
Fonte: cálculos do autor

Figura 10: Série temporal e previsão SVM kernel RBF



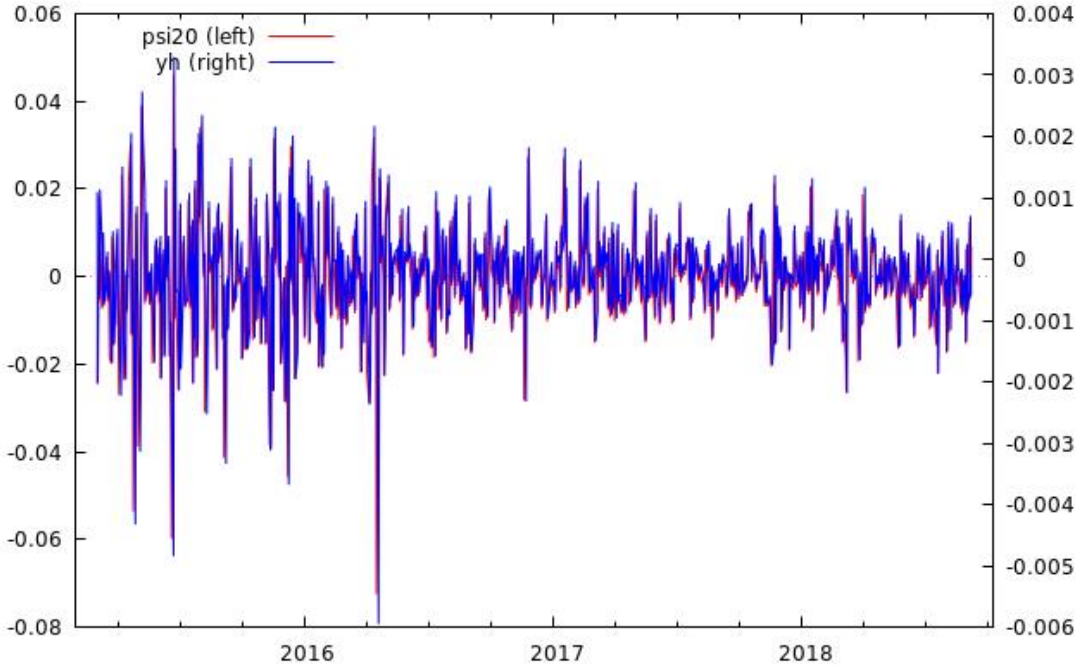
Fonte: cálculos do autor

Figura 11: Dispersão entre valores observados e previsões SVM kernel RBF



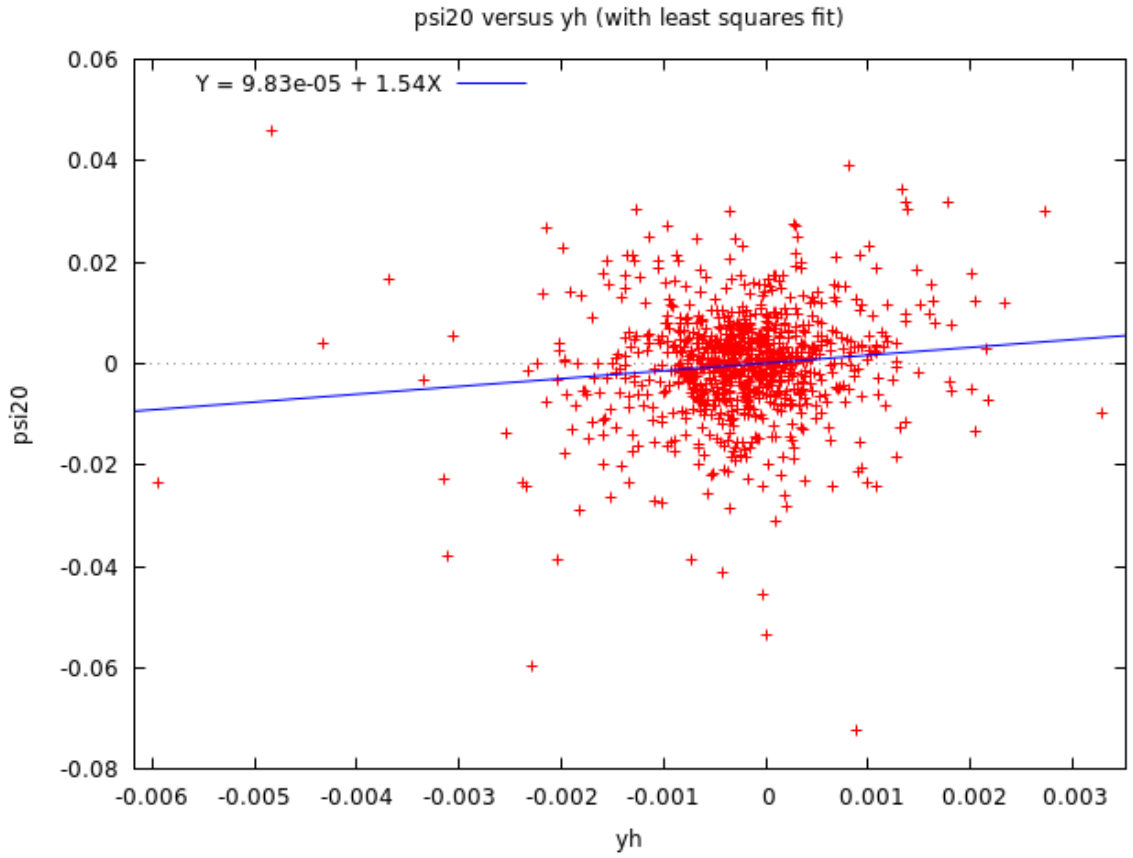
Fonte: cálculos do autor

Figura 12: Série temporal e previsão SVM kernel sigmoid



Fonte: cálculos do autor

Figura 13: Dispersão entre valores observados e previsões SVM kernel sigmoid



Fonte: cálculos do autor