Análise de métodos de tratamento de outliers para predição dos retornos de índices de ações negociados em bolsa*

Cristiane Gea[1,2], Janio Lima [1], Eduardo Bezerra, Eduardo Ogasawara[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

[2]Universidade Federal Fluminense - UFF

{cristiane.gea, janio.lima}@aluno.cefet-rj.br, ebezerra@cefet-rj.br, eogasawara@ieee.org

Resumo. *Definidos como pontos extremos, os* outliers *são observações que não seguem um comportamento padrão em uma série temporal. Contudo, esta anomalia pode levar a especificação incorreta do modelo, estimativas viesadas dos parâmetros e previsões de baixa acurácia. Apesar de ser visto como um erro de medida, nas séries financeiras os* outliers *carregam informações relevantes sobre a dinâmica do mercado acionário e de fatores interrelacionados. Diante disso, este artigo propõe uma análise comparativa do desempenho de técnicas de tratamento de* outliers *nas previsões dos índices de ações. Os resultados apontam que métodos com suavização apresentam melhor desempenho, corroborando com a hipótese de que os* outliers *possuem conteúdo informacional relevante para a previsão da dinâmica dos índices de ações.*

Abstract. *Defined as extreme points, outliers are observations that do not follow a standard behavior in a time series. However, this anomaly can lead to incorrect model specifications, biased parameter estimates, and low accuracy predictions. Despite being seen as a measurement error, in financial series, outliers carry relevant information about the stock market dynamics and interrelated factors. Therefore, this article proposes a comparative analysis of performing outliers treatment techniques in stock index forecasts. The results show that methods with smoothing perform better, corroborating the hypothesis that outliers have relevant informational content for predicting the dynamics of stock prices.*

1. Introdução

Conceituados como pontos extremos existentes em uma série temporal, os *outliers* representam observações que se desviam acentuadamente do comportamento usual das demais observações em um conjunto de dados. Em séries financeiras, os *outliers* podem ser compreendidos como variações anormais nos preços de ações, indicando a existência de fatores que não podem ser modelados convencionalmente [Pimenta et al., 2018]. Considerado como uma anomalia existente nas séries temporais, os *outliers* podem levar a especificação incorreta de modelos, estimativas viesadas dos parâmetros e avaliação incorreta das previsões, pois interfere na relação entre as observações passadas e futuras. Visto que tais fatores estão associados à natureza altamente ruidosa, volátil e dinâmica das séries temporais financeiras, prever o comportamento futuro do mercado acionário torna-se uma tarefa desafiadora.

*Os autores agradecem à FAPERJ, à CAPES (código 001) e ao CNPq pelo financiamento do projeto.

Embora estudos anteriores demandaram esforços para desenvolver métodos de detecção de *outliers* e modelos preditivos robustos a *outliers*, poucos trabalhos focaram na aplicação de métodos para tratamento de *outliers*. Pimenta et al. [2018] empregam o método *locally weighted scatterplot smoothing* para detecção e remoção de *outliers* nas séries do preço de ações. Kumar and Patil [2018] aplicam o método de winsorização para a remoção de *outliers* nas séries de taxa de câmbio. Clewlow and Strickland [2000] propõem a aplicação de filtro recursivo para identificação de saltos nos preços, identificados na distribuição amostral dos retornos diários. Este filtro consiste em um procedimento iterativo repetido até que nenhum salto possa ser identificado.

A fim de preencher a lacuna existente na literatura, o presente trabalho visa realizar uma análise comparativa de nove métodos utilizados para o tratamento séries temporais a fim de averiguar se os *outliers* identificados nas séries de retornos de índices ações possuem conteúdo informacional relevante para o desempenho preditivo dos modelos de séries temporais. Para avaliar de modo sistemático o efeito das técnicas de *outliers* nas séries, faz-se uso do conceito de previsões *rolling origin* [Hyndman and Athanasopoulos, 2018]. As previsões *rolling origin* fornecem uma maior compreensão acerca do funcionamento dos modelos por meio da obtenção de diversas previsões para as séries temporais.

2. Métodos para tratamento de *outliers*

Os métodos de tratamento de *outliers* podem ser divididos em três grandes grupos: métodos explícitos de remoção de *outliers*, métodos de filtragem e decomposição de séries temporais e métodos de suavização de séries temporais.

Métodos explícitos de remoção de *outliers*. Os métodos explícitos de remoção de *outliers* possuem como fundamentação a definição de valores de corte e a substituição dos valores extremos por estes valores de corte. Na regra de alcance interquartil os valores de corte são $Q1 - 1.5 \times (Q3 - Q1)$ e $Q3 + 1.5 \times (Q3 - Q1)$, onde $Q1$ e $Q3$ correspondem ao primeiro e terceiro quartis, respectivamente. Por outro lado, na winsorização os valores de corte são P_k e P_{100-k}, onde k refere-se ao percentual de observações da série temporal que são considerados como valores extremos [Bali et al., 2016].

Métodos de filtragem e decomposição de séries temporais. Os métodos de decomposição de séries temporais possuem como premissa a separação da série em componentes de tendência e sazonal. Partindo do conceito de que uma série temporal (y) é formada aditivamente pela combinação de tendência (T), ciclo (C), sazonalidade (S) e erro (E), Shumway and Stoffer [2017] conceituam o ajustamento sazonal como a remoção do componente sazonal de uma série ($y - S$).

Visando obter estimativa suavizada do componente de tendência de uma série temporal, o filtro Hodrick-Prescott (HP) realiza a separação da tendência de longo prazo a partir de flutuações de curto prazo. Conforme, Hodrick and Prescott [1997] a obtenção desta estimativa suavizada é realizada através da minimização da variância da série temporal (y_t) em torno de seu componente de crescimento (g_t), conforme a Equação 1, onde $y_t = c_t - g_t$, c_t é o componente cíclico de y_t e λ é um parâmetro positivo que penaliza a

variabilidade no componente de crescimento da série temporal.

$$\min_{\{g_t\}_{t=-1}^{T}} \left\{ \sum_{t=1}^{T} c_t^2 + \lambda \sum_{t=1}^{T} \left[(g_t - g_{t-1}) - (g_{t-1} - g_{t-2}) \right]^2 \right\} \tag{1}$$

Com o objetivo de detectar *outliers* na distribuição amostral dos retornos diários, o filtro recursivo (ou filtro autorregressivo) é um procedimento iterativo que se repete até que nenhum *outlier* seja identificado. Por meio de uma implementação recursiva, $y_t = y_t + \sum_{i=1}^{p} (f_i \times y_{t-i})$, o filtro f_i busca remover a autocorrelação entre as observações. Assim, quando *outliers* são detectados, o filtro é recalculado, de modo que os resíduos sejam definidos para zero [Lee et al., 2017].

Métodos de suavização de séries temporais. Os métodos de suavização são utilizados para lidar com irregularidades e flutuações aleatórias encontradas nas séries temporais [Ao, 2010]. Na classe dos métodos de suavização exponencial, as observações são vistas como combinações ponderadas das observações passadas, onde as observações mais recentes recebem pesos maiores do que as observações mais antigas. Segundo Hyndman and Athanasopoulos [2018], a suavização exponencial simples (SES) é adequada para os casos em que não há padrões claros de tendência ou sazonalidade. A série suavizada $\hat{y}_t$ de y_t pode ser representada pela seguinte expressão: $\hat{y}_t = \alpha y_t + (1 - \alpha)\hat{y}_{t-1}$, onde a constante α, $0 \leq \alpha \leq 1$, é o fator de suavização. A SES pode ser interpretada como uma média ponderada entre a observação mais recente e a suavização mais recente. Por outro lado, caso as séries temporais apresentem tendência linear, o método mais adequado é a suavização exponencial quadrática[1], que pode ser interpretado como uma extensão da SES com duas constantes de suavização e, portanto, visto como uma aplicação dupla do método SES.

Segundo Ao [2010], o método de suavização por médias móveis é aplicado visando a redução de flutuações aleatórias e o ajustamento dos componentes cíclico e sazonal de uma série temporal. Nesse contexto, a estatística da suavização por médias móveis (m_t) pode ser obtida por meio da média das últimas k observações, $m_t = (1/k) \sum_{n=0}^{k-1} y_{t-n}$. Por fim, *lowess smoothing* é um método de suavização de séries temporais usualmente composto por mínimos quadrados ponderados e um modelo polinomial de segunda ordem [Liu et al., 2019]. Segundo Härdle [1990], com base no conceito de ajustamento de mínimos quadrados polinomial local, com vizinhos do tipo k-NN, o método tem início a partir de uma estimativa piloto dos vizinhos, definindo iterativamente os pesos de robustez e realizando a suavização diversas vezes.

3. Avaliação Experimental

Para o presente estudo, utilizamos as séries temporais dos retornos diários dos principais índices de ações mundial[2]. Foram utilizados os seguintes índices: Ibovespa (BVSP), MOEX Russia Index (IMOEX), BSE Sensex (BSESN), S&P IPSA (IPSA), IPC Mexico (MXX), Hang Seng Index (HSCE), S&P/TSX Index (GSPTSE), CAC 40 (FCHI), German Stock Index (GDAXI), Nikkei 225 (N225), FTSE 100 (FTSE) e S&P 500 (GSPC). O período analisado foi de 03 de janeiro de 2000 até 28 de maio de 2021.

[1] Também conhecido como método suavização exponencial dupla ou método linear de Holt.

[2] Calculadas a partir dos preços de fechamento dos índices de ações extraídos do *website* Yahoo Finance.

3.1. Protocolo experimental

Visando uma avaliação robusta para a influência das diferentes técnicas de tratamento de *outliers* nas previsões, foi adotada a abordagem de *rolling-origin*. Tal abordagem decompõe a série em janelas móveis de treino e teste, caracterizadas pela atualização sucessiva da origem da previsão e pela produção de previsões para cada nova origem [Hyndman and Athanasopoulos, 2018]. Partindo do conceito de janelas móveis, a cada iteração (atualização da origem da previsão) uma observação é adicionada no final das séries e uma observação é removida no início da série, de modo a permanecer constante o tamanho do conjunto de treinamento. O tamanho da janela de teste é de 15 passos à frente, enquanto a janela de treino contém um ano de observações prévias. As previsões *rolling-origin* fornecem uma melhor compreensão sobre o funcionamento dos modelos pela obtenção de diversas previsões para as séries temporais, viabilizando análise da centralidade e dispersão dos erros de previsões.

Por fim, os efeitos da aplicação dos métodos para o tratamento de *outliers* nas previsões *rolling origin* são avaliados por meio do modelo ARIMA, cuja base reside na hipótese de que as séries temporais são geradas a partir de processo autorregressivo e procuram modelar o processo das séries a fim de prever os seus valores futuros [Kumar and Patil, 2018] [3]. Nos modelos ARIMA o valor corrente de uma série temporal x_t pode ser descrita como uma função de seus p valores passados, $x_{t-1}, \ldots, x_{t-p}$ e de seus q valores de ruído branco passados, $\omega_{t-1}, \ldots, \omega_{t-q}$, de modo que sua forma geral pode ser representada como $\phi(B)(1-B)^d x_t = \theta(B)\omega_t$, onde B é o operador de defasagem, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ é o operador autorregressivo (AR) e $\theta(B) = 1 + \theta_1 B + \cdots + \theta_p B^q$ é o operador de médias móveis (MA) [Shumway and Stoffer, 2017].

Como medida de acurácia, foi utilizado o Erro Médio Absoluto Relativo (rMAE), que permite comparações das predições dos modelos candidatos contra predições de modelos de referência. O rMAE pode ser descrito como rMAE $= (\text{MAE}_a/\text{MAE}_b)$, onde MAE_a é o erro médio absoluto do modelo analisado e MAE_b corresponde ao erro médio absoluto do modelo de referência [Davydenko and Fildes, 2013]. No escopo deste trabalho, o modelo de referência foi gerado a partir das séries temporais sem aplicação de método para tratamento de *outliers*. Assim, quanto mais próximo de zero estiver o rMAE, melhor será o desempenho preditivo do modelo, indicando que o mesmo erra menos em suas previsões do que o próprio modelo de referência. Por outro lado, modelos com rMAE ≥ 1 apresentam piores desempenhos preditivos.

3.2. Resultados

Foram aplicados os diferentes métodos de tratamento de *outliers* ao conjunto de dados. Ao final de cada tratamento, foram analisadas a remanescência de *outliers* nas séries por meio da análise por boxplot. Pela Tabela 1 é possível observar que a regra de alcance interquartil e a winsorização removeram a totalidade dos *outliers* em todas as séries de retorno. A suavização por médias móveis realizou a remoção completa na maioria das séries analisadas (com exceção de BSESN e FTSE).

A Tabela 2 mostra a média e o desvio padrão do rMAE para as previsões realizadas por meio do método ARIMA para as séries de retorno de índices de ações, influenciada

[3] São utilizadas funções que selecionam automaticamente um modelo ARIMA otimizado [Hyndman and Khandakar, 2008].

Tabela 1. Desempenho de cada método na remoção de *outliers*

	BVSP	IMOEX	BSESN	IPSA	MXX	HSCE	GSPTSE	FCHI	GDAXI	N225	FTSE	GSPC
Regra de alcance interquartil	X	X	X	X	X	X	X	X	X	X	X	X
Winsorização	X	X	X	X	X	X	X	X	X	X	X	X
Suavização Exponencial Simples												
Suavização Exponencial Quadrática												
Suavização por médias móveis	X	X		X	X	X	X	X	X	X		X
Lowess Smoothing	X											
Filtro HP												
Filtro Recursivo												
Ajustamento Sazonal												

Nota: Cada lacuna preenchida com X indica que houve a remoção completa dos *outliers*

Tabela 2. Avaliações do rMAE (média e desvio padrão) por tratamento de *outliers*

	BVSP		IMOEX		BSESN		IPSA		MXX		HSCE	
	mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
Regra de alcance interquartil	0.944	2.261	0.932	1.420	0.897	1.710	0.926	1.160	0.911	1.529	0.916	2.281
Winsorização	0.871	2.036	0.863	1.284	0.842	1.576	0.858	1.052	0.853	1.406	0.853	2.085
Suavização Exponencial Simples	0.155	0.423	0.174	0.291	0.169	0.340	0.186	0.242	0.168	0.334	0.163	0.423
Suavização Exponencial Quadrática	0.091	0.277	0.090	0.147	0.093	0.189	0.099	0.126	0.097	0.225	0.091	0.250
Suavização por Médias Móveis	0.002	0.005	0.004	0.007	0.002	0.005	0.003	0.003	0.002	0.004	0.002	0.005
Lowess Smoothing	0.008	0.019	0.028	0.042	0.009	0.019	0.015	0.019	0.008	0.015	0.009	0.022
Filtro HP	1.000	2.562	1.001	1.663	1.002	2.075	1.003	1.354	1.002	1.798	1.001	2.673
Filtro Recursivo	1.117	2.833	1.188	1.979	1.184	2.420	1.252	1.690	1.182	2.123	1.170	3.097
Ajustamento Sazonal	0.982	2.500	0.980	1.581	0.994	2.020	0.986	1.315	0.985	1.740	0.989	2.604
	GSPTSE		**FCHI**		**GDAXI**		**N225**		**FTSE**		**GSPC**	
	mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
Regra de alcance interquartil	0.868	1.216	0.901	1.680	0.903	1.771	0.927	1.827	0.0012	1.687	0.863	1.382
Winsorização	0.820	1.129	0.849	1.558	0.857	1.657	0.868	1.683	0.900	1.388	0.825	1.302
Suavização Exponencial Simples	0.152	0.238	0.149	0.320	0.153	0.328	0.155	0.319	0.850	1.288	0.143	0.268
Suavização Exponencial Quadrática	0.080	0.126	0.082	0.194	0.083	0.183	0.082	0.176	0.150	0.248	0.080	0.165
Suavização por Médias Móveis	0.001	0.002	0.001	0.002	0.001	0.003	0.001	0.003	0.079	0.135	0.001	0.002
Lowess Smoothing	0.009	0.012	0.007	0.016	0.007	0.015	0.009	0.017	0.006	0.011	0.007	0.011
Filtro HP	1.002	1.621	1.001	2.023	1.001	2.112	1.000	2.100	1.001	1.688	1.002	1.822
Filtro Recursivo	1.147	1.809	1.123	2.261	1.134	2.399	1.130	2.373	1.128	1.889	1.095	1.951
Ajustamento Sazonal	1.008	1.587	0.991	1.971	0.990	2.051	0.986	2.049	0.989	1.645	1.001	1.775

pelos diferentes tipos de tratamento de *outliers*. Os resultados mostram que, dentre os métodos que conseguiram remover a totalidade dos *outliers*, a suavização por médias móveis apresentou rMAE próximo de zero e menor dispersão na série de erro de previsão (com exceção de FTSE). Por outro lado, a regra de alcance interquartil apresentou o pior desempenho preditivo.

Com relação aos métodos que não conseguiram remover totalmente os *outliers*, *lowess smoothing* apresentou o melhor desempenho preditivo, com menor rMAE e menor dispersão na série de erro de previsão, em detrimento do filtro recursivo, que apresentou o pior desempenho. Por outro lado, em todos os casos analisados, os métodos de suavização de séries temporais apresentam previsões com desempenho superior às dos métodos de decomposição e filtragem de séries temporais, com rMAE ≥ 1, indicando que tais modelos erraram mais em suas previsões do que os modelos de referência. Diante destes resultados, os *outliers* existentes nas séries financeiras não podem ser considerados erros de medida e que o método utilizado para os tratar interfere na capacidade do modelo de prever a trajetória futura dos retornos de índices de ações.

4. Conclusão

A partir deste artigo, analisou-se o desempenho de nove métodos para tratamento de *outliers* e previsão dos retornos de ações, utilizando o modelo ARIMA e o conceito de previsões *rolling origin*. Os achados do artigo podem ser resumidos como segue. Primeiro, dos métodos analisados somente três foram eficientes na remoção dos *outliers*, que são a regra de alcance interquartil, a winsorização e a suavização por médias móveis (exceto BSESN e FTSE). Segundo, no geral, os métodos de suavização apresentaram melhores desempenhos, com erro médio absoluto relativo próximo de zero, enquanto o método de filtro recursivo apresentou pior performance. A partir deste cenário, é possível inferir que a forma de tratamento das séries temporais impacta significativamente no desempenho preditivo dos modelos, visto que os métodos que realizaram remoção de *outliers* ou de qualquer componente de uma série apresentaram desempenhos inferiores. Por fim, este cenário corrobora a hipótese de que os *outliers* possuem conteúdo informacional relevante para a predição do desempenho futuro do mercado acionário, em vez de serem resumidos a simples erro de medição, necessitando de técnicas adequadas para o seu tratamento.

Referências

Sio-Iong Ao. *Applied Time Series Analysis and Innovative Computing*. Springer Science & Business Media, April 2010. ISBN 978-90-481-8768-3.

Turan G. Bali, Robert F. Engle, and Scott Murray. *Empirical Asset Pricing: The Cross Section of Stock Returns*. John Wiley & Sons, April 2016. ISBN 978-1-118-09504-1.

Les Clewlow and Chris Strickland. *Energy Derivatives: Pricing and Risk Management*. Lacima Publications, 2000. ISBN 978-0-9538896-0-0.

A. Davydenko and R. Fildes. Measuring Forecasting Accuracy: The Case Of Judgmental Adjustments To Sku-Level Demand Forecasts. *International Journal of Forecasting*, 29(3):510–522, 2013.

R.J. Hodrick and E.C. Prescott. Postwar U.S business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, 29(1):1–16, 1997.

R.J. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3):1–22, 2008.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, May 2018. ISBN 978-0-9875071-1-2.

Wolfgang Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1990. ISBN 978-0-521-42950-4.

Hemanth P. Kumar and Basavaraj S. Patil. Forecasting volatility trend of INR USD currency pair with deep learning LSTM techniques. In *Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018*, pages 91–97, 2018.

E.H. Lee, C. Wickham, P.A. Beedlow, R.S. Waschmann, and D.T. Tingey. A likelihood-based time series modeling approach for application in dendrochronology to examine the growth-climate relations and forest disturbance history. *Dendrochronologia*, 45:132–144, 2017.

J. Liu, Q. Li, W. Chen, Y. Yan, Y. Qiu, and T. Cao. Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks. *International Journal of Hydrogen Energy*, pages 5470–5480, 2019.

A. Pimenta, C.A.L. Nametala, F.G. Guimarães, and E.G. Carrano. An Automated Investing Method for Stock Market Based on Multiobjective Genetic Programming. *Computational Economics*, 52(1):125–144, 2018.

Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, April 2017. ISBN 978-3-319-52452-8.