Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Previsão de Volatilidade Realizada via Modelos *HAR* e Métodos de *Machine Learning*

Andressa de Oliveira Dorneles

25 de abril de 2023

Andressa de Oliveira Dorneles

Previsão de Volatilidade Realizada via Modelos *HAR* e Métodos de *Machine Learning*

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Prof. Dr. Flávio Augusto Ziegelmann

Porto Alegre
Abril de 2023

Andressa de Oliveira Dorneles

Previsão de Volatilidade Realizada via Modelos *HAR* e Métodos de *Machine Learning*

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a):_____
Prof. Dr. Flávio Augusto Ziegelmann, UFRGS
Doutor pela University Of Kent At Canterbury, UKC, Grã-Bretanha

Banca Examinadora:

Prof. Dr. Hudson da Silva Torrent, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, RS

Porto Alegre
Abril de 2023

"Precisamos estar dispostos a nos livrar da vida que planejamos, para podermos viver a vida que nos espera." (Joseph Campbell)

Agradecimentos

De todos os caminhos que percorri durante minha vida, talvez, o mais tortuoso tenha sido o que me levou ao curso de estatística. Da improbabilidade de cursar estatística até a conclusão desse curso, edificada no presente texto, tem-se uma jornada única, de aprendizagem, sabedoria e autoconhecimento.

Agradeço, primeiramente a Deus, se hoje eu existo, é por causa d'Ele. Agradeço aos meus pais, Janete e Ricardo, por todo o amor do mundo, por todo o auxílio e compreeenção, por me amarem do jeito que eu sou. Agradeço aos meus avós, por tudo que fizeram por mim. E agradeço também ao resto da minha família pelo apoio.

Tenho enorme gratidão a todos que me permitiram desfrutar de um campo de estudos tão diverso quanto o da estatística, em especial a todos os professores do departamento de estatística da UFRGS que tive contato. Agradeço à Prof. Luciana Nunes por seu contagiante entusiasmo em ensinar, à Prof. Suzi Camey por ter ministrado a melhor matéria do curso de forma esplêndida (*Processos estocásticos*). Agradeço também à Prof. Vanessa Leotti por todo conhecimento prático transmitido, ao Prof. Eduardo Horta pelas discussões filosóficas e por ter me auxiliado a conseguir minha primeira bolsa de iniciação científica, ao Prof. Hudson Torrent pelas disciplinas ministradas e por aceitar fazer parte da banca. Por fim, realizo um agradecimento especial ao meu orientador Prof. Flávio Ziegelmann por todo conhecimento transmitido nas três disciplinas que ministrou enquanto fui aluna, pela paciência em me ensinar, levantando questionamentos que me fizeram sair da minha zona de conforto. Agradeço também por todas as oportunidades que ele me proporcionou enquanto aluna de graduação, inclusive minha segunda bolsa de iniciação científica. Além disso, sei, que se eu precisar, poderei sempre contar com seu auxílio.

Ainda, agradeço imensamente a todos que trilharam essa jornada comigo. Em especial, agradeço às minhas duas grandes amigas de curso, Fernanda e Raquel, por terem trilhado esse caminho comigo, me auxiliando e tecendo uma amizade duradoura. Agradeço também a todos os meus amigos, os que cursam estatística e os demais que cultivei ao longo da vida.

Nada descreve a explosão de felicidade de fechar um ciclo no âmbito profissional. Tal qual eternizado pelo poema de Robert Frost, digo, houve um momento em minha vida que precisei escolher entre dois possíveis caminhos, eu optei pelo caminho menos percorrido, e isso tem feito toda a diferença.

Resumo

Este estudo propõe realizar previsões para volatilidade realizada diária e comparar as previsões obtidas por diferentes métodos de *machine learning* e por um modelo *benchmark*. Os métodos utilizados são os seguintes: regressão Ridge, *Least absolute shrinkage and selection operator* (LASSO), *adaptive* LASSO (AdaLASSO), *Elastic Net* e *Random forest*. Já o modelo *benchmark* utilizado é o *Heterogeneous Autoregressive Model* (HAR), com estimação via Mínimos Quadrados. O objetivo principal do estudo é entender quais propostas implicam em melhores resultados dependendo do período analisado. Os períodos considerados foram pré-pandemia de COVID-19, ínicio da pandemia, durante a sua ocorrência e também foi analisado o período total (sem distinção do cenário pandêmico). Para entender quais propostas se destacaram nos respectivos períodos, foi usado o *Model confidence set* (MCS). Com as análises do MCS, pôde-se concluir que o método *Random forest* foi o que mais se destacou.

Palavras-Chave: Volatilidade, *Forecasting*, *Machine Learning*, *Heterogeneous Autoregressive Model*.

Abstract

This study proposes to perform forecasts for daily realized volatility and compare the forecasts obtained by different machine learning methods and by a benchmark model. The methods used are as follows: Ridge regression, *Least absolute shrinkage and selection operator* (LASSO), *adaptive* LASSO (AdaLASSO), *Elastic Net* and *Random forest*. The benchmark model used is the Heterogeneous Autoregressive Model (HAR), with Least Squares estimation. The main objective of the study is to understand which approaches result in better results depending on the analyzed period. For that, the periods considered were pre-pandemic of COVID-19, beginning of the pandemic, during its occurrence and the total period was also analyzed (without distinction of the pandemic scenario). To understand which approaches stood out in the respective periods, the Model confidence set (MCS) was used. With the MCS analyses, it could be concluded that the Random forest method was the one that stood out the most.

Keywords: Realized volatility, Forecasting, HAR, Machine Learning.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A previsão de volatilidade de ativos financeiros tem vital importância para o mercado financeiro, sendo um objeto de estudo relevante em finanças. Tem-se que a volatilidade é uma medida estatística que indica a intensidade das oscilações no preço de um ativo ou derivativo em um determinado período de tempo. Em outros termos, analisando a volatilidade é possível dimensionar os riscos de um investimento, considerando o quanto se quer arriscar ao investir, criando uma estratégia de investimento. Posto isto, há a proposta de realizar previsões de volatilidade realizada e posteriores comparações entre as previsões obtidas por diferentes propostas.

A variância realizada é normalmente calculada como a soma dos quadrados dos retornos intradiários ao longo de um dia. Neste estudo, as previsões serão referentes a volatilidade realizada e assume-se que volatilidade realizada é a raiz quadrada da variância realizada, tal qual em Corsi (2008). Um modelo comumente usado em previsões de volatilidade realizada é o modelo HAR, *Heterogeneous Autoregressive Model*. O HAR é um modelo que além de autoregressivo também é considerado heterogêneo, pois acrescenta volatilidades em períodos diferentes (diários, semanais, mensais) (Corsi, 2008).

Entre várias outras opções para estimação de volatilidade realizada, há os métodos de *machine learning*, que são alternativas importantes para realizar regressões com muitas covariáveis. Os métodos de regularização são mais recentes na literatura e se diferenciam do tradicional método de mínimos quadrados ordinários (MQO), pois adicionam penalização para a realização das regressões (James, 2013). Um método que será usado no estudo é a regressão Ridge, em que há a presença de um parâmetro de penalização, que controla a magnitude da penalidade, sendo que essa penalidade é efetivada nos quadrados dos coeficientes. Outro método que será usado é o *Least absolute shrinkage and selection operator*, chamado LASSO, o qual atribui uma penalidade no valor absoluto dos coeficientes, de forma que essa penalidade acaba por zerar coeficientes, agindo como um seletor de variáveis (Tibshirani, 1996). Um terceiro método que se apropria de características tanto do LASSO, quanto da regressão Ridge é o método chamado *Elastic net*. O principal ponto acerca dele é o fato de ele unir as penalidades de LASSO e de Ridge em sua composição, tentando balancear os benefícios que ambos os métodos proporcionam. Há também um método de regularização que é uma variação do LASSO, o AdaLASSO, ou “LASSO Adaptativo", que atribui penalizações distintas a cada coeficiente das covariáveis, incorporando uma propriedade de consistência na seleção de variáveis, que o LASSO não possui em algumas situações (Zou, 2006). Finalmente, ao se tratar de previsões de volatilidade é relevante utilizar métodos baseados em árvores, já que esses mé-

todos são alternativas interessantes por ajustarem bem modelos não lineares. Neste caso em específico, será utilizado o método *Random forest*, uma vez que reduz a variância de estimadores de árvores de regressão (James, 2013).

Através de métodos mais recentes na literatura, como os métodos de *machine learning*, e comparando com um modelo *benchmark* tal qual o modelo HAR, tem-se a possibilidade de explorar diferentes formas de previsões de volatilidade para aplicabilidade prática, especialmente, aos dados financeiros. Essa aplicabilidade prática se traduz em entender quais métodos ou modelos implicam em melhores resultados dependendo do período analisado. Em comparação às previsões obtidas pelo modelo HAR, espera-se que ao incluir mais covariáveis, especificamente 22 covariáveis (que são ou 22 defasagens da volatilidade realizada ou 22 médias defasadas da volatilidade realizada), tendo em vista os métodos de *machine learning*, as previsões de volatilidade realizada apresentem mais qualidade em relação aos erros de previsão. Haja vista a possibilidade de uma das propostas obter melhores resultados do que as outras, em momentos de crise econômica, ou no recente cenário pandêmico, pode-se compreender que dentre outros objetivos tem-se o de averiguar se há diferenças significativas entre as previsões obtidas pelas diversos propostas. Além disso, também é de interesse investigar se o cenário pandêmico afetou as séries de volatilidade e suas previsões, dado que a pandemia de COVID-19 certamente foi um período atípico para o mercado financeiro. Logo, neste estudo, previsões de volatilidade servirão a diversos propósitos.

Neste estudo, compararemos as diversas propostas através do *Model Confidence Set* (MCS), que determina um conjunto de modelos contendo o(s) melhor(es) modelo(s), em um determinado nível de confiança (Peter R. Hansen e Nason, 2011). Portanto, o MCS será usado com o intuito de entender quais dos modelos testados são os melhores para realizar previsões de volatilidade realizada em quatro cenários distintos: momentos pré-pandemia, momentos iniciais da pandemia, momentos durante sua ocorrência, e em todo o período considerado.

Para conduzir o estudo, serão contextualizados conceitos de séries temporais, finanças, e de *machine learning*. Também será apresentada a análise empírica com dados reais da bolsa de valores de São Paulo, e consequentemente, os resultados obtidos. Por fim, serão discutidos os resultados e apresentadas as conclusões do estudo conjuntamente às considerações finais.

2 Metodologia

Dado o tópico deste estudo, é importante mencionar breves conceitos de séries temporais, apresentar conceitos de finanças, e introduzir o modelo HAR e os métodos de *machine learning*, com intuito de construir o arcabouço necessário para previsões de volatilidade realizada.

2.1 Séries temporais

Uma série temporal pode ser definida como qualquer conjunto de observações ordenadas no tempo (Morettin e Toloi, 2008). Ainda, conforme os autores, ao obter uma série temporal $Z(t_1), ..., Z(t_n)$, observada nos instantes $t_1, ..., t_n$, há alguns possíveis interesses

- investigar o mecanismo gerador da série temporal;
- fazer previsões de valores futuros da série;
- descrever apenas o comportamento da série;
- procurar peridiocidades relevantes aos dados.

Dentre os possíveis interesses em séries temporais citados pelos autores, neste estudo, o principal objetivo é realizar previsões de séries temporais através de um modelo clássico e de métodos de *machine learning*. Tal qual eternizado por Morettin e Toloi (2008), as previsões não são um fim, mas um meio de obter informações para posteriores tomadas de decisões. Dessa forma, havendo várias previsões provenientes de diversas propostas, é possível tomar decisões com base nas previsões de melhor qualidade, as quais são determinadas através de comparações entre suas precisões.

2.1.1 Estacionariedade e diferenças

A premissa básica da estacionaridade é que as leis de probabilidade que regem o comportamento de uma série temporal não mudam com o tempo. Em certo sentido, o processo está em equilíbrio (Cryer e Chan, 2008). Ainda, considerando Shumway e Stoffer (2011), pode-se definir que uma série temporal é chamada fracamente estacionária se

- $\mathbb{E}|Z_t| = \mu \quad , \forall t \quad ,$

- $\mathbb{E}|Z_{t^2}| < \infty \quad , \forall t \quad ,$
- $Cov_Z(s,t) = Cov_Z(s-t) \quad , \forall t \text{ e } \forall s \quad .$

Embora a estacionariedade seja importante, muitas séries encontradas na em termos práticos, não são estacionárias, assim sendo, nesses casos, é necessário ajustar os dados originais de forma a transformá-los (Morettin e Toloi, 2008). Uma tranformação comumente utilizada é tomar diferenças sucessivas da série original até ocorrer a obtenção de uma série estacionária. Desse modo, tem-se que a primeira diferença para uma série Z(t), é definida por

$$\Delta Z_t = Z_t - Z_{t-1} \quad , \tag{2.1}$$

a segunda diferença é definida como

$$\Delta^2 Z_T = \Delta[\Delta Z_T] = \Delta[Z_t - Z_{t-t}] \quad , \tag{2.2}$$

escrita de outro modo,

$$\Delta^2 Z_t = Z_t - 2Z_{t-1} + Z_{t-2} \quad . \tag{2.3}$$

Generalizando para a n-ésima diferença de Z(t), tem-se

$$\Delta^n Z_t = \Delta[\Delta^{n-1} Z_t] \quad . \tag{2.4}$$

Habitualmente, tomar uma ou duas diferenças da série acaba sendo suficiente para que a nova série se torne estacionária (Morettin e Toloi, 2008).

2.1.2 Modelo AR

Os processos autorregressivos são regressões sobre si mesmos (Cryer e Chan, 2008). De forma geral, esses modelos baseiam-se no fato de que o valor atual da série pode ser explicado como função de valores passados, considerando o número de passos no passado necessários para prever o valor atual (Shumway e Stoffer, 2011).

Considerando $\tilde{Z}_t = Z_t - \mu$, pode-se escrever $\tilde{Z}_t$, como

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + ... + \phi_p \tilde{Z}_{t-p} + a_t \quad , \tag{2.5}$$

representando um autoregressivo de ordem p. Sendo assim, valor atual da série $\tilde{Z}_t$ é uma combinação linear dos p valores passados mais recentes dela mesma mais o termo a_t que incorpora tudo o que há de novo na série no tempo t que não é explicado pelos valores passados (Cryer e Chan, 2008). Evidenciando o polinômio característico

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - ... - \phi_p x^p \quad , \tag{2.6}$$

e também a equação característica

$$1 - \phi_1 x - \phi_2 x^2 - ... - \phi_p x^p = 0 \quad , \tag{2.7}$$

tem-se que ao assumir que a_t é independente dos valores $\phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + ... + \phi_p \tilde{Z}_{t-p}$, existe uma solução estacionária para a equação 2.7 se, e somente se, as p raízes dessa equação excederem 1 em valor absoluto. Ainda segundo Cryer e Chan (2008), para que as raízes sejam maiores que 1, em valor absoluto, é necessário, mas não suficiente, que ambas

$$\left.\begin{array}{l} \phi_1 + \phi_2 + ... + \phi_p \\ e \qquad |\phi_p| < 1 \end{array}\right\} \quad . \tag{2.8}$$

Logo, é equivalente dizer que o processo será estacionário se a raiz de $\phi(x) = 0$ estiver fora do círculo unitário (Cryer e Chan, 2008) (Morettin e Toloi, 2008).

2.2 Dados de alta frequência e volatilidade

2.2.1 Dados de alta frequência

O uso de dados de alta frequência para previsões de volatilidade é importante por vários motivos, um dos mais relevantes é que dados de menor frequência falham em captar toda a movimentação que ocorre ao longo do dia no mercado de ações (Andersen e Bollerslev, 1998). Em outras palavras, o retorno diário é um valor que resume as mudanças que ocorreram nos preços ao longo do dia, já o conjunto de retornos intradiários consiste em valores efetivos das mudanças que ocorreram nos preços ao longo do dia. Ainda, para estimar volatilidade realizada diária, um valor que seja o resumo do que ocorreu no dia não serve, pois, o que interessa são as mudanças efetivas nos preços, e por isso, se usam retornos intradiários para realizar essa estimação. Maiores explicações acerca de volatilidade realizada diária podem ser vistas na subseção 2.2.3. Outros motivos que corroboram para a relevância do uso de dados de alta frequência são: um princípio estatístico básico de quanto maior é o tamanho de uma amostra, mais informações ela contém; quanto mais se pode observar sobre volatilidade, melhor é o entendimento de seu comportamento ao longo do tempo (Marmitt, 2012).

2.2.2 Volatilidade

A volatilidade é uma medida de variabilidade dos preços dos ativos ao longo do tempo. Além disso, usualmente, se faz referência a ela como sendo o desvio padrão dos retornos dos ativos (Marmitt, 2012). Os motivos pelos quais ocorrem variações nos preços dos ativos são diversos. Em períodos de crises, as repercussões ao mercado financeiro podem ser acentuadas, refletindo em variações grandes nos preços dos ativos (Andersen et al., 2003b). Ainda, variáveis macroeconômicas como inflação, desemprego, consumo e produção também podem acelerar variações nos preços dos ativos (SCHWERT, 1989). Sendo assim, mudanças nos valores de volatilidade ocorrem por diversos fatores. Logo, para realizar previsões de volatilidade, deve-se levar em consideração as dinâmicas do mercado financeiro.

2.2.3 Variância realizada diária e volatilidade realizada diária

Em síntese, a variância realizada diária é a soma dos quadrados dos retornos intradiários ao longo de um dia e sua raiz quadrada é a volatilidade realizada diária. Para chegar nas definições formais de variância realizada e volatilidade realizada, é necessário apresentar algumas definições, por isso, de acordo com (Andersen et al., 2003a) define-se que retornos esperados são iguais a zero para qualquer horizonte de tempo, padroniza-se o intervalo de tempo para M observações intradiárias, e também deve-se assumir que W_t (movimento browniano padrão) e σ_t (volatilidade instantânea) são independentes, além de condicionar a expectativa matemática na trajetória de volatilidade $\{\sigma_{t+\tau}\}_{\tau=0}^{h}$, desse modo, a variância do retorno para um período de tempo h pode ser descrita como

$$\sigma_{t,h}^2 = \int_0^h \sigma_{t+\tau}^2 d\tau \quad , \tag{2.9}$$

essa variância pode ser chamada de variância integrada (VI), logo, a volatilidade para um período de tempo h é igual a integral das volatilidades intradiárias passadas. A VI, no entanto, não é observada e, por ser objeto de interesse, precisa ser estimada. O retorno intradiário no período m e no dia t é obtido da seguinte maneira

$$r_{t,m} = p_{t,m} - p_{t,m-1} \quad , \tag{2.10}$$

para $m = 1, ..., M$ e $t = 1, ..., n$.

Ainda, conforme (Andersen et al., 2003a), tem-se a definição de variância realizada diária como

$$RV_t^2 = \sum_{m=1}^{M} r_{t,m}^2 \quad . \tag{2.11}$$

Sob certas condições envolvendo uma falta de autocorrelação de retornos, os autores demonstraram que a variância realizada da equação anterior é um estimador consistente da variância integrada (VI), portanto, $RV_t \xrightarrow{p} VI_t$.

A volatilidade realizada diária é a raiz quadrada da variância realizada diária (Junior e Pereira, 2013). Em outras palavras, a volatilidade realizada é o desvio padrão realizado. Em vista disso, se escreve

$$\sqrt{RV_t^2} = \sqrt{\sum_{m=1}^{M} r_{t,m}^2} \quad , \tag{2.12}$$

mais resumidamente,

$$RV_t = \sqrt{\sum_{m=1}^{M} r_{t,m}^2} \quad . \tag{2.13}$$

Em várias ocasiões, em econometria, é comum o uso de RV_t para denotar variância realizada. Neste estudo, seguiu-se o mesmo de (Corsi, 2008), em que o termo RV_t é usado como volatilidade realizada (raiz quadrada da variância realizada), posto isto, a notação foi ajustada para a situação[1].

2.2.4 Modelo HAR

Em séries temporais, comumente, o objetivo é analisar a dependência temporal da série em um única frequência de variação temporal, isto é, ou diária, ou mensal ou ainda anual. Todavia, o modelo *Heterogeneous autoregressive model* (HAR) propõe adicionar num mesmo modelo volatilidades realizadas em diferentes frequências de variações temporais. Sendo assim, ele é um modelo AR heterogêneo (Alvarenga, 2015) (Corsi, 2008).

Definindo a volatilidade parcial como a volatilidade gerada por um determinado componente do mercado, o modelo proposto pode ser descrito como uma cascata aditiva de volatilidades parciais. Considerando um modelo hierárquico com apenas três componentes de volatilidade correspondentes para horizontes de tempo de um dia (d), uma semana (w) e um mês (m), denota-se respectivamente: $\tilde{\sigma}_t^{(d)}, \tilde{\sigma}_t^{(w)}, \tilde{\sigma}_t^{(m)}$ (Corsi, 2008).

[1]Foram usadas notações tal qual em (Junior e Pereira, 2013), exceto para a variância realizada diária, cuja notação foi modificada.

Tendo em vista o processo de retorno de alta frequência como determinado pelo componente de volatilidade de frequência mais alta, neste caso, o componente diário, com, $\tilde{\sigma}_t^{(d)} = \sigma_t^{(d)}$, então, o retorno do processo é descrito por

$$r_t = \sigma_t^{(d)} \epsilon_t \quad , \tag{2.14}$$

com $\epsilon_t \sim NID(0, 1)$.

Assim, o modelo HAR pode ser escrito como

$$RV_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + w_{t+1d} \quad , \tag{2.15}$$

com, $w_{t+1d} = \tilde{w}_{t+1d}^{(d)} + w_{t+1d}^{(d)}$.

Por conseguinte, abrindo o modelo descrito por Corsi (2008), pode-se ver a estrutura AR com certas condições nos coeficientes.

Irregularidade temporal, sazonalidade intradiária e microestrutura de mercado

Existem três conceitos que se relacionam e explicam um pouco do que ocorre no pregão diariamente, são eles: irregularidade temporal, sazonalidade intradiária e microestrutura de mercado. A irregularidade temporal remete ao fato de negociações ocorrerem livremente ao longo do dia, de forma que exista assincronia entre elas. Há alguns pontos problemáticos nessa situação: transações acontecem ao longo do dia, para o mesmo ativo, com preços diferentes; ativos diferentes têm frequências de transações diferentes, sofrendo de formas diferentes com a irregulatidade temporal (Marmitt, 2012). Já a sazonalidade intradiária representa um padrão que ocorre diariamente no pregão envolvendo a volatilidade. Ela começa alta na abertura, vai decaindo ao longo do dia e volta a crescer no fechamento do pregão (WOOD et al., 1985). Além disso, os dados de alta frequência também trazem problemas estruturais, por conterem mais informações, acabam mais sujeitos a ruídos (Taylor, 2011). Esses ruídos são conhecidos como microestruturas de mercado, ou seja, são fricções que ocorrem a cada momento no tempo sendo incorporadas aos preços dos ativos. Pode-se afirmar que as microestruturas de mercado surgem tanto por causa da irregularidade temporal entre transações quanto pela sazonalidade intradiária (Marmitt, 2012).

Para a realização da análise empírica, deve-se levar em consideração que os conceitos apresentados estão presentes nos retornos intradiários de ativos, e consequentemente, as medidas realizadas construídas a partir dos retornos também consideram esses conceitos.

2.3 Métodos de *Machine Learning*

O interesse em usar métodos de *machine learning* para realizar previsões para volatilidade reside em uma explicação evidenciada por Konzen e Ziegelmann (2016), de que quando o número de variáveis ou parâmetros é grande em comparação com o tamanho da amostra que se quer analisar, os métodos tradicionais de previsões são desafiados considerando interpretabilidade, estimativa e eficácia do modelo. Dessa forma, encontrar outros meios de prever volatilidade pode ser extremamente útil. Neste estudo, consideram-se 22 covariáveis e os métodos usados são ou métodos de regularização ou árvores de regressão.

2.3.1 Métodos de regularização

Conforme Goodfellow et al. (2017), métodos de regularização são técnicas usadas para calibrar métodos de *machine learning* objetivando obter bons desempenhos. Além disso, o uso desses métodos serve para atingir melhores previsões através da remoção de ruídos e deixando o modelo mais simples, generalizável e com melhor performance. No presente estudo, praticamente todos os métodos de *machine learning* usados para realizar previsões de volatilidade são de regularização, com exceção da *Random Forest*, que é um método de árvore.

Regressão RIDGE

A regressão Ridge é um método frequentemente utilizado considerando dados que sofram multicolinearidade, pois quando ocorre a multicolinearidade, os estimadores de mínimos quadrados têm variâncias grandes, o que pode resultar em valores preditos distantes dos valores reais.

Em síntese, Ridge é um método em que ocorre inserção de restrição nos coeficientes ao introduzir um fator de penalidade ao método tradicionalmente usado para obtenção de estimativas de coeficientes. A penalidade da regressão Ridge é efetivada nos quadrados dos coeficientes. Sendo assim, pode-se escrever

$$\hat{\beta}^{Ridge} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin} \sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{j=1}^{k} \beta_j x_{jt} \right)^2 \quad , \tag{2.16}$$

estando sujeito à

$$\sum_{j=1}^{k} (\beta_j)^2 \leq s \quad , \tag{2.17}$$

onde o parâmetro $s \geq 0$ controla a penalidade. Pode-se reescrever a expressão, de forma que

$$\hat{\beta}^{Ridge} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin} \left\{ \sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{j=1}^{k} \beta_j x_{jt} \right)^2 + \lambda \sum_{j=1}^{k} (\beta_j)^2 \right\} \quad , \tag{2.18}$$

tal que, quanto maior o valor de λ, com $\lambda > 0$, maior serão as penalidades dos coeficientes. Se $\lambda = 0$, não ocorre penalização e as estimativas se reduzem às calculadas pelo MQO.

Método LASSO

Tibshirani (1996) propôs o método LASSO (*Least absolute shrinkage and selection operator*) com o intuito de minimizar a soma residual dos quadrados, desde que a soma do valor absoluto dos coeficientes fosse menor que uma constante. Dada a natureza dessa restrição, ela tende a produzir alguns coeficientes que são exatamente zero, ou seja, é viável obter estimativas e selecionar variáveis simultaneamente. A ideia presente no método LASSO é simples e pode ser aplicada a uma variedade de modelos estatísticos. (Tibshirani, 1996).

Em resumo, o método LASSO adiciona uma penalização ao MQO. Assim, busca-se a solução de

$$\hat{\beta}^{LASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin} \sum_{t=1}^{T}\left(y_t - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ji}\right)^2 \quad , \tag{2.19}$$

sujeita à restrição

$$\sum_{j=1}^{k}|\beta_j| \leq s \quad , \tag{2.20}$$

onde o parâmetro $s \geq 0$ controla a penalidade. Pode-se reescrever a expressão, de forma que

$$\hat{\beta}^{LASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin}\left\{\sum_{t=1}^{T}\left(y_t - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ji}\right)^2 + \lambda\sum_{j=1}^{k}|\beta_j|\right\} \quad , \tag{2.21}$$

em que $\lambda > 0$. Novamente, se $\lambda = 0$, não ocorre penalização e as estimativas se reduzem às calculadas por MQO.

Método *Elastic Net*

O método *Elastic Net* (EL) tem um propósito: combinar as penalizações dos métodos LASSO e Ridge, de maneira a unir as características dos dois métodos. Este método foi proposto como método de regularização e seleção de variáveis, além disso, ele encoraja um efeito de agrupamento, isto é, preditores fortemente correlacionados tendem a estar dentro ou fora do modelo juntos (Zou e Hastie, 2005).

Esse método usa um procedimento de estimação de duas etapas, isto é, primeiro para cada λ fixo ele encontra os coeficientes de regressão Ridge, em seguida, faz um encolhimento considerando o método LASSO que faz uma quantidade dupla de encolhimento. Contudo, eventualmente, pode haver aumento de viés. Tem-se a fórmula

$$\hat{\beta}^{EL} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin}\left\{\sum_{t=1}^{T}\left(y_t - \beta_0 - \sum_{j=1}^{k}\beta_j x_{jt}\right)^2 + \lambda\left[\frac{(1-\alpha)}{2}\sum_{j=1}^{k}(\beta_j)^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right]\right\} \quad , \tag{2.22}$$

em que a nova penalidade é $\frac{\lambda\cdot(1-\alpha)}{2}$ para o termo associado à penalidade da regressão Ridge, sendo que a divisão por 2 é apenas uma conveniência matemática para otimização. Além disso, $\lambda \cdot \alpha$ é a nova penalidade para o termo associado à penalidade do método LASSO.

Em termos práticos, considera-se que α controla a "mistura" entre as duas penalidades e λ controla o valor da penalização. Consequentemente, α aceita valores entre 0 e 1, de forma que

- se $\alpha = 0$ e $\lambda = 0$, não ocorre penalização e as estimativas se reduzem às calculadas pelo método de mínimos quadrados ordinários;
- se $\alpha = 0$ e $\lambda \neq 0$, ocorre apenas penalização associada à regressão Ridge;
- se $\alpha = 1$ e $\lambda \neq 0$, ocorre apenas penalização associada ao método LASSO;
- se $0 < \alpha < 1$ e $\lambda \neq 0$, ocorre penalização associada ao método *Elastic Net*.

Método AdaLASSO

Existem cenários em que o LASSO não é consistente para seleção de variáveis tal qual nas ocasiões em que há variáveis muito correlacionadas. Nesta conjuntura, foi proposta outra versão do LASSO, denominada AdaLASSO (*Adaptive LASSO*), onde pesos adaptativos são usados com intuito de penalizar os coeficientes de forma não homogênea. Assim sendo, ao atribuir penalizações distintas a cada coeficiente das covariáveis, há incorporação de uma propriedade de oráculo, chamada consistência na seleção de variáveis (Zou, 2006).

Ainda considerando Zou (2006), tem-se que sob algumas suposições um estimador de mínimos quadrados penalizado tem propriedades de oráculo se

- $\lim_{n\to\infty} P(A_n^* = A) = 1$, ou seja, é consistente em seleção de variáveis;
- $\sqrt{n}(\hat{\beta}_A^{*(n)} - \beta_A^*) \to_d N(0, \sigma^2 \times C_{11}^{-1})$, tal que, há estimativas de coeficientes diferentes de zero que seguem assintoticamente a mesma distribuição que os estimadores do método de mínimos quadrados ordinários, quando a equação é estimada através desse método, apenas com as variáveis relevantes;

onde $A = \{j : \beta_j^* \neq 0\}$ é o conjunto verdadeiro de coeficientes não-nulos, e $A_n^* = \{j : \hat{\beta}^{*(n)} \neq 0\}$ representa o conjunto de coeficientes não-nulos estimados. Portanto,

$$\hat{\beta}^{AdaLASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin} \sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{j=1}^{k} \beta_j x_{jt} \right)^2 \quad , \tag{2.23}$$

sujeito à

$$\sum_{j=1}^{k} w_j |\beta_j| \quad , \tag{2.24}$$

ou,

$$\hat{\beta}^{AdaLASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin} \left\{ \sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} w_j |\beta_j| \right\} \quad , \tag{2.25}$$

onde, $w = |\hat{\beta}^*|^\tau$ e $\tau > 0$, e $\hat{\beta}^*$ é um estimador n-consistente de β. Note que

- se $\tau = 0$, o que ocorre é que $w = 1$, logo, o método se reduz ao método LASSO;
- frequentemente usa-se $\tau = 1$, então, $w = |\hat{\beta}^*|$, embora quaisquer valores para $\tau > 1$ sejam aceitáveis.

2.3.2 Critério na escolha de parâmetros

Métodos mais frequentes para escolha de parâmetros em *machine learning*, podem não ser ideais para serem usados neste estudo em específico, devido à natureza dos dados que serão utilizados, ou seja, por causa das correlações temporais existentes entre as observações. Nesta análise, será usado o *Bayesian information criterion* (BIC), que segundo os autores Zhang et al. (2010), seleciona parâmetros de regularização (neste caso, o λ) de tal forma que o modelo verdadeiro seja consistentemente identificado (Zhang et al., 2010).

Além de selecionar λ para todos os métodos de regularização através do BIC, também serão selecionados valores para α, para o método *Elastic Net*. Ainda, para o método AdaLASSO, o valor de τ foi fixado em 1, seguindo (Konzen e Ziegelmann, 2016).

2.3.3 Métodos de árvores

São métodos usados para regressão ou classificação, sendo assim, é plausível estratificar ou segmentar o espaço do preditor em várias regiões simples. Para fazer uma previsão para alguma observação, habitualmente usa-se a média ou a moda das observações de treinamento na região a que ela pertence. Métodos baseados em árvore, normalmente, não são competitivos com outras abordagens de aprendizado supervisionado em termos de acurácia de previsão (James, 2013).

Neste estudo, será usado apenas o método de regressão *Random Forest*. Para entender o que é uma Random Forest, floresta aleatória na tradução literal, é necessário entender o que é uma árvore de decisão.

Árvores de decisão (*Regression trees*)

Neste estudo, as árvores de decisão são chamadas árvores de regressão, por serem provenientes de regressão. Uma árvore de regressão é uma estrutura de dados com algumas especificidades. Cada árvore possui um nó raiz, de onde parte toda a população ou amostra sobre a qual se quer realizar previsões, então, há o particionamento, ou seja, a divisão da árvore, sendo que cada particionamento origina os ramos da árvore. Por fim, com a árvore crescida, há o nó Terminal, cuja partição é inexistente. Logo, a partir desse nó final tem-se as previsões (James, 2013). Exemplifica-se a partir de uma figura retirada de James (2013).

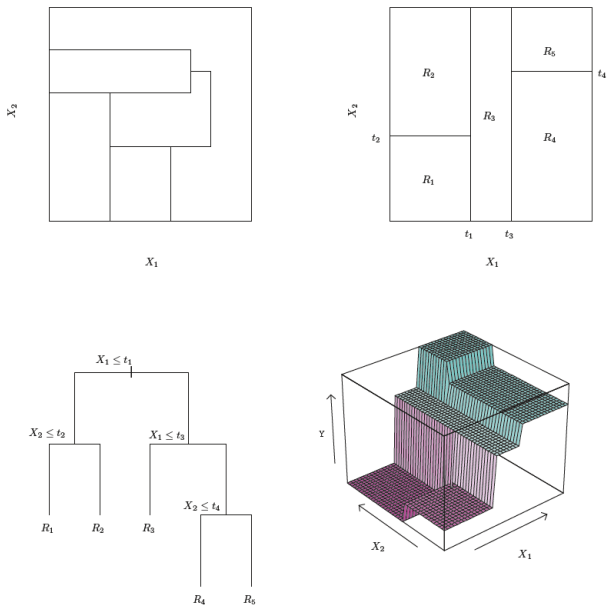


Figura 2.1: *Árvore de decisão*
Fonte: (James, 2013).

- **Canto superior esquerdo**: Uma partição do espaço de recursos bidimensionais que poderia não resultar da divisão binária recursiva.
- **Canto superior direito**: A saída de uma divisão binária recursiva em um exemplo bidimensional.
- **Canto inferior esquerdo**: Uma árvore correspondente à partição no painel superior direito.
- **Canto inferior direito**: Um gráfico em perspectiva da superfície de previsão correspondente a essa árvore.

Fonte: (James, 2013)

Random Forest

Um conjunto de árvores de regressão é melhor em performance do que uma única árvore, e esse conjunto de árvores pode ser considerado uma *Random forest*. Afinal, se houver muitos modelos relativamente não correlacionados (árvores de regressão) operando em conjunto, haverá melhor performance em relação a qualquer um dos modelos individuais (Breiman, 1993).

Como as árvores de regressão que crescem profundamente tendem a aprender padrões irregulares, elas superajustam seus conjuntos de treinamento, de forma que tenham viés baixo, mas variância muito alta. Então, *Random forests* podem calcular a média de várias árvores de decisão profundas, treinadas em diferentes partes do mesmo conjunto de treinamento, de tal forma que há redução da variância de estimadores (Hastie et al., 2003). Logo, random Forest é um método mais eficiente do que árvores de regressão individuais.

2.4 Comparações de previsões

2.4.1 *Model Confidence Set* (MCS)

O alicerce deste estudo é a comparação de previsões, para tanto, uma forma de fazer essa comparação é através do *Model Confidence Set* (MCS) desenvolvido por Peter R. Hansen e Nason (2011). O objetivo do MCS é determinar um conjunto de modelos constituído de forma a conter o(s) melhor(es) modelo(s) com um determinado nível de confiança (Peter R. Hansen e Nason, 2011).

Mais detalhadamente, o procedimento consiste em uma sequência de testes de equivalência (σ_M) que permitem construir um conjunto de modelos superiores, onde a hipótese nula é de que todos os modelos dentro deste conjunto se igualam em relação à eficácia de predição, e são superiores aos que estão fora do conjunto, com um determinado nível de confiança. Havendo rejeição da hipótese nula, o modelo com a pior performance é removido. Sendo assim, até que o teste de equivalência não seja rejeitado, o processo se repete. Finalmente, o conjunto final de modelos é chamado *Model Confidence Set* (Peter R. Hansen e Nason, 2011).

Para realizar tais procedimentos, foram calculadas as perdas associadas às previsões de volatilidade de cada modelo, portanto, foi necessário usar funções de perda,

descritas em (Hansen e Lunde, 2005). Apresentam-se as funções utilizadas.

$$MSE_1 \equiv n^{-1} \sum_{t=1}^{n} (\sigma - h_t)^2 \quad ; \tag{2.26}$$

$$MAE_1 \equiv n^{-1} \sum_{t=1}^{n} |\sigma - h_t| \quad . \tag{2.27}$$

Além dessas funções, também escolheu-se a estatística para o teste em cada etapa da iteração, de forma que a estatística de teste discutida em Peter R. Hansen e Nason (2011) e utilizada neste estudo, é

$$T_{max,M} = \underset{i \in M}{max}(t_{i.}) \quad . \tag{2.28}$$

onde,

$$t_{i.} = \frac{\overline{d}_{i.}}{\sqrt{\widehat{var}(\overline{d}_{i.})}} \quad , \tag{2.29}$$

para $i, j \in M$, onde $\overline{d}_{i.}$ é a perda amostral do i-ésimo modelo em relação à média entre modelos em M. Para mais detalhes, veja (Peter R. Hansen e Nason, 2011).

2.4.2 Teste Diebold-Mariano

Apresentado por Diebold e Mariano (1995), o teste Diebold-Mariano é amplamente utilizado na literatura de séries temporais, visto que ele determina se as duas previsões são significativamente diferentes. Para tal, o teste considera diferenciais de perda $\{d_t\}_{t=1}^{T}$, de forma que para uma função de perda ao quadrado tem-se

$$d_t = e_t^2 - \breve{e}_t^2 \quad , \tag{2.30}$$

em que, e_t^2 é o erro de previsão e $\breve{e}_t^2$ é o erro de referência. Supondo que o diferencial de perda é uma série estacionária, havendo uma amostra com erros, tem-se que $\bar{d}$ é a média dessa amostra, sendo que ela converge assintoticamente para distribuição normal, de forma que

$$\sqrt{T}\bar{d} \xrightarrow{d} N(\mu, 2\pi f_d(0)) \quad . \tag{2.31}$$

Logo, para o teste de Diebold-Mariano, a estatística para testar se há igualdade na precisão de previsão é calculada por

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \quad , \tag{2.32}$$

onde $2\pi \hat{f}_d(0)$ é uma estimativa consistente (Diebold e Mariano, 1995).

3 Análise empírica

Foram usados dados da bolsa de valores de São Paulo (BOVESPA), onde montou-se um portfolio com pesos iguais (*naive*) de ativos da bolsa, em cima do qual foram calculadas as volatilidades realizadas baseadas em retornos de 5min dos preços deste portfólio. A partir de um vetor contendo valores para volatilidades realizadas diárias, foram calculadas previsões pelos métodos de *machine learning* e pelo modelo HAR. Por fim, as previsões foram comparadas através do *Model Confidence Set* (MCS). Todos os cálculos, análises e gráficos foram realizados no software R Core Team (2022) sob a interface do Posit team (2022).

3.1 Dados

O banco de dados contém dados de alta frequência, usados por (Ricco, 2021): são retornos financeiros intradiários usados para construir o portfólio e suas medidas de variabilidade, de dez ativos de diferentes segmentos constando entre os ativos mais negociados no mercado financeiro da Bovespa, compreendendo o período de 02 de julho de 2018 até 31 de dezembro de 2020. Os ativos presentes nos dados, são apresentados na Tabela (3.1).

Tabela 3.1: Informações dos dados iniciais

Ativo	Símbolo	Setor da economia
Ambev S.A.	ABEV3	Indústria alimentícia
B3	B3SA3	Financeiro
Bradesco S.A.	BBDC4	Financeiro
Intermedica S.A.	GNDI3	Saúde
Itaú S.A.	ITUB4	Financeiro
JBS S.A.	JBSS3	Indústria alimentícia
Magazine Luiza S.A.	MGLU3	Consumo e varejo
Petrobras S.A.	PETR4	Óleo e Gás
Suzano S.A.	SUZB3	Indústria de papel e celulose
Vale S.A.	VALE3	Indústria de mineração

3.2 Vetor de volatilidades realizadas diárias

O banco de dados mencionado será chamado de banco de dados inicial, afinal, os dados desse primeiro banco foram usados apenas para construção de outros bancos de dados para a realização do estudo. Os preços intradiários dos ativos financeiros correspondem a dez colunas do banco de dados inicial. Deste banco inicial, construiu-se o portfolio com pesos iguais e obtiveram-se seus preços intradiários, sendo estes armazenados em um vetor. A seguir, usou-se uma função do pacote *highfrequency* (Boudt et al., 2022) para

- calcular os retornos intradiários do portfólio;
- calcular variância realizada diária com base no vetor de retornos intradiários criados.

Por fim, obteve-se o vetor de variâncias realizadas diárias, e calculou-se a raiz quadrada de cada um de seus valores, criando o vetor de volatilidades realizadas diárias que é o objeto relevante nesse estudo.

3.3 Separação em bancos de dados

A separação do banco de dados foi importante para alcançar os objetivos do estudo, relembrando, eles são: obter previsões de volatilidade realizada e posteriores comparações entre as previsões obtidas por diferentes propostas; e entender as possíveis implicações que a pandemia de COVID-19 teve para essas previsões.

- **Banco 1**: Chama-se "Banco 1", aquele contruído com covariáveis (volatilidades realizadas) que são defasagens de valores presentes da volatilidade realizada até a ordem 22 considerando o período de 02/07/2018 até 31/12/2020, totalizando 597 observações;
- **Banco 2**: Chama-se "Banco 2", aquele contruído com covariáveis que são médias de defasagens da volatilidade realizada presente, considerando a média aritmética simples em relação ao número de dias passados até a ordem 22 considerando o período de 02/07/2018 até 31/12/2020, totalizando 597 observações.

Além disso, já foi mencionado que um dos desafios abordados, para quem realiza previsões de volatilidade, são cenários econômicos atípicos, como momentos de crises econômicas e políticas e também em casos de pandemias, principalmente, considerando o recente cenário pandêmico de COVID-19. Então, ao cruzar os bancos alternativos 1 e 2 (distintos em suas covariáveis) com o cenário pandêmico, tiveram-se quatro bancos de dados, particionados através dos dois bancos anteriores em datas substancialmente relevantes para a pandemia de COVID-19.

- **Banco 1 - pré-pandemia**: Chama-se "Banco 1", por ter as características do Banco 1; e "pré-pandemia", pois ele foi considerado de 02/07/2018 até 31/12/2019, totalizando 348 observações;
- **Banco 1 - em pandemia**: Chama-se "Banco 1", por ter as características do Banco 1; e "em pandemia", pois ele foi considerado de 01/01/2020 até 31/12/2020, totalizando 249 observações;

- **Banco 2 - pré-pandemia**: Chama-se “Banco 2”, por ter as características do Banco 2; e “pré-pandemia”, pois ele foi considerado de 02/07/2018 até 31/12/2019, totalizando 348 observações;

- **Banco 2 - em pandemia**: Chama-se “Banco 2”, por ter as características do Banco 2; e “em pandemia”, pois ele foi considerado de 01/01/2020 até 31/12/2020, totalizando 249 observações.

O principal atrativo em criar esses quatro bancos de dados é analisar possíveis diferenças entre previsões realizadas pré-pandemia e durante a pandemia, além de compreender se mudanças na composição dos bancos de dados, criados a partir de um mesmo vetor, podem impactar em qual das propostas produz melhores previsões. Lembrando que o modelo HAR não leva em consideração as 22 covariáveis que diferenciam os bancos, as previsões realizadas seguem a equação 2.15.

3.4 Procedimentos, previsões e comparações

Com os bancos de dados montados, foi simples realizar as previsões, para tal, os procedimentos aplicados foram similares. Basicamente, para cada proposta, foi criado um *loop* cuja finalidade era criar uma janela móvel de tamanho constante, de modo que os dados usados para gerar cada modelo mudassem dentro da janela utilizada. Dentro do *loop* foram armazenadas previsões em um vetor, logo, ao final do processo, haviam seis vetores com previsões de modelos gerados por métodos de *machine learning* e pelo modelo HAR.

3.4.1 Treino e teste

Todos os bancos foram divididos considerando as mesmas proporções, isto é, $\frac{2}{3}$ para treinar o modelo e $\frac{1}{3}$ para realizar as previsões.

- **Bancos (1 e 2)**: Com 597 observações totais, 398 observações (em janela móvel) para treinar o modelo e 199 observações para realizar as previsões;

- **Bancos (1 e 2) - pré-pandemia**: Com 348 observações totais, 232 observações (em janela móvel) para treinar o modelo e 116 observações para realizar as previsões;

- **Bancos (1 e 2) - em pandemia**: Com 249 observações totais, 166 observações (em janela móvel) para treinar o modelo e 83 observações para realizar as previsões.

3.4.2 Ínicio da Pandemia de COVID-19

Com a existência do interesse em investigar se o cenário pandêmico afetou as séries de volatilidade e suas previsões no início da pandemia, os bancos 1 e 2 foram separados de formas distintas, sem considerar as proporções $\frac{2}{3}$ para treinar o modelo e $\frac{1}{3}$ para realizar novas previsões. Logo, essa investigação específica buscou analisar o ínicio da pandemia, pois é o período que se imaginava haver um pico de volatilidade maior. Isto é, escolheu-se um período em que a pandemia ainda não havia iniciado no

Brasil e também os primeiros meses de sua ocorrência, consequentemente, momentos de maior incerteza refletida na volatilidade.

- Com 471 observações totais, sendo o período considerado de 02/07/2018 até 30/06/2020;
- Com 310 observações (em janela móvel) para treinar o modelo, sendo o período considerado de 02/07/2018 até 31/10/2019;
- Com 161 observações para realizar as previsões, sendo o período considerado de 01/11/2019 até 30/06/2020.

3.4.3 Funções

De acordo com as particularidades de se trabalhar com séries temporais, tornou-se necessário utilizar o BIC para encontrar melhores valores para os parâmetros (vide seção 2.3.2). Contudo, um pacote muito utilizado no software R, chamado *glmnet* (Friedman et al., 2010) (Simon et al., 2011), encontra o melhor valor do parâmetro λ de outra forma, por conseguinte, foram exploradas outras funções baseadas no pacote *glmnet*, mas que utilizassem o BIC. Dessa forma, as funções utilizadas para geração de modelos[1] e previsões[2] foram retiradas de repositório *online* em que o autor das funções havia aplicado-as em outro estudo em situação semelhante.

3.4.4 Ajuste de parâmetros

Cada método ou modelo tem suas particularidades para ajustes de parâmetros. Para criação de modelos através do métodos de regularização, foram usados valores para os parâmetros, tais quais

- $\alpha = 0$ e seleção de melhor λ através do BIC, para a regressão Ridge;
- $\alpha = 1$ e seleção de melhor λ através do BIC, para o método LASSO;
- $\alpha = 1$, $\tau = 1$ e seleção de melhor λ através do BIC, para o método AdaLASSO;
- seleção de melhor α, de valores entre 0.01 e 0.99 e seleção de melhor λ através do BIC, para o método Elastic net.

Os parâmetros para λ foram encontrados pela função já citada na subseção 3.4.3. Já os valores para α foram usados de acordo com o pacote *glmnet* (Friedman et al., 2010) (Simon et al., 2011), com exceção do método *Elastic net*, cujo melhor valor para α foi encontrado através de um *loop* colocado dentro do *loop* inicial, para criação de 100 modelos com α variando entre 0.01 e 0.99, sendo escolhido o α do modelo com o menor BIC, através da função citada na subseção 3.4.3. Sendo assim, para cada volta do *loop* inicial, foi repetido o procedimento para encontrar o melhor α. Em muitas voltas do *loop*, para o método *Elastic net*, o que acabou acontecendo foi que o valor de α ficou muito próximo de 1, o que fez com que os vetores de previsões de *Elastic net* e de LASSO, ficassem muito similares, o que será possível notar através de gráficos no capítulo 4.

[1] https://github.com/gabrielrvsc/HDeconometrics/blob/master/R/ic.glmnet.R
[2] https://github.com/gabrielrvsc/HDeconometrics/blob/master/R/predict.ic.glmnet.R

Além disso, para o modelo HAR, não foram traçados demais parâmetros, continuou-se a realizar previsões um passo à frente, porém, com as características de aplicação do modelo HAR, todos os procedimentos realizados através do pacote *HARModel* (Sjoerup, 2019).

Finalmente, para a *Random forest*, foi usada uma função do pacote *randomForest* (Liaw e Wiener, 2002), colocou-se uma "semente" dentro do *loop* que variava de acordo com o número de iterações usadas, para que fosse possível reproduzir resultados fielmente em outras reaplicações. Além disso, foram usadas 1000 árvores de regressão para a construção da *Random forest*, na tradução literal, "Floresta aleatória". Esse número foi baseado na premissa de que quanto mais árvores melhor, contudo, a partir de um número de árvores, essa premissa não ocorre mais (Breiman, 1993). Em outras palavras, não adianta adicionar infinitas árvores à regressão, pois sua performance não irá melhorar. Neste sentido, o valor de 1000 árvores foi escolhido.

3.4.5 *Model Confidence Set* (MCS)

O *Model Confidence Set* foi apresentado na subseção 2.4.1, no entanto, para a análise empírica foram ajustados alguns pontos. O pacote utilizado para realizar os procedimentos foi o *MCS* (Bernardi, 2017). Foram usadas duas funções de perda apresentadas em Hansen e Lunde (2005) para calcular as perdas associadas às previsões de volatilidade. Além disso, para todos os procedimentos MCS, foram estabelecidos valores fixos para nível de confiança e para o número de amostras *bootstrap* usadas para construir cada teste. Sendo o nível de significância fixado em 50%, e o número de amostras *bootstrap* fixado em 25000. Ainda, para cada procedimento MCS foi usada a estatística de teste (T_{max}) em cada etapa da iteração. Para cada banco de dados, foram calculados dois *Model Confidence Sets*, combinando as duas funções de perda e a estatística de teste. Isso se deve, ao interesse em testar se um único conjunto de modelos iria se sobressair sobre os outros, o que aconteceu em todas as situações para quatro dos seis bancos de dados criados.

3.4.6 Teste Diebold-Mariano

Para realizar o teste Diebold-Mariano, mencionado na subseção 2.4.2, foi usada uma função do pacote *forecast* (Hyndman et al., 2023) (Hyndman e Khandakar, 2008), de forma que fosse usado o *default* de cada parâmetro. Em outras palavras, foi especificada a hipótese alternativa de que as previsões têm diferentes níveis de precisão, além disso, o poder do teste foi igual a 2 e foi usado 1 para o horizonte de previsão.

3.4.7 Análises gráficas

Todas as análises gráficas foram geradas através dos pacotes *ggplot2* (Wickham, 2016) e *plotly* (Sievert, 2020), tal que cores foram escolhidas arbitrariamente para mostrar previsões de diferentes propostas, de forma a criar uma identidade visual.

Além disso, os gráficos mostram o modelo HAR nas diferenciações dos bancos apenas para realizar comparações, pois as 22 covariáveis que diferenciam os bancos 1 e 2 não foram usadas nos cálculos do modelo HAR para realizar previsões.

4 Resultados

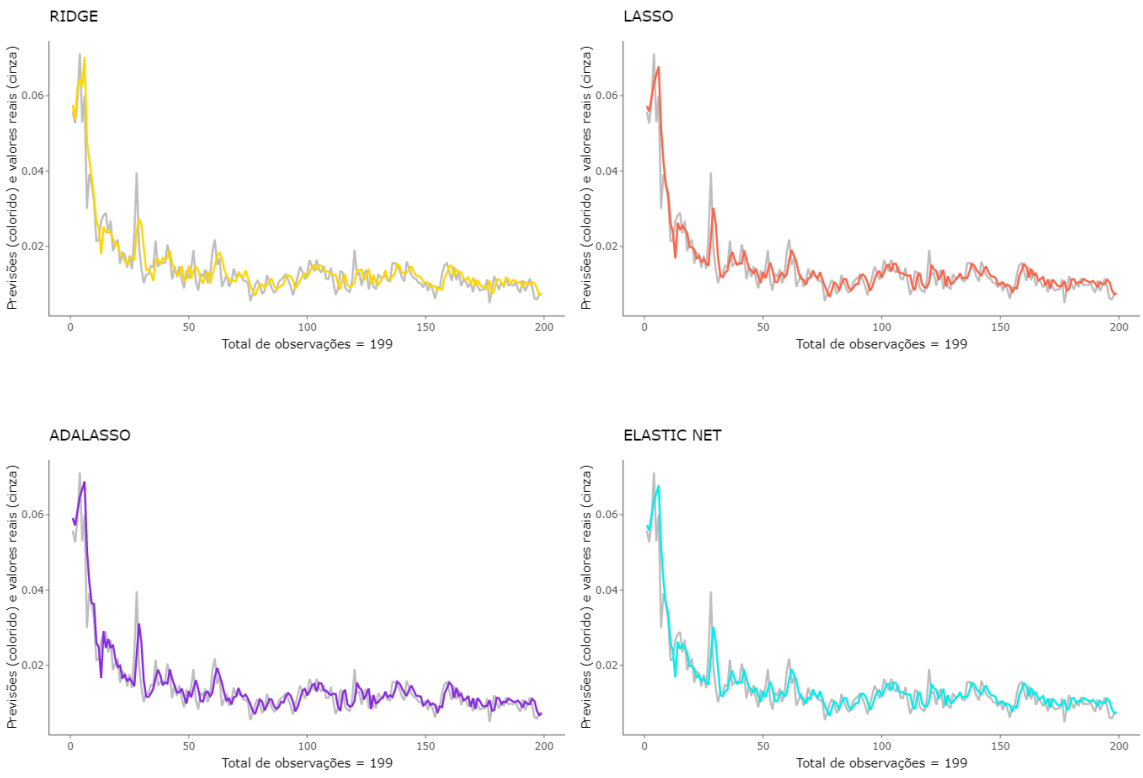
Com os procedimentos concluídos e considerando as previsões armazenadas em diferentes vetores, pode-se analisar os resultados obtidos nos diferentes bancos de dados criados.

4.1 Banco 1

Os resultados obtidos pelos modelos construídos considerando o banco 1 com defasagens dos dias anteriores, serão apresentados por meio de gráficos e do *Model confidence set* (MCS). Além disso, será apresentada uma tabela com os valores dos erros de cada proposta em relação ao modelo *benchmark*.

4.1.1 Banco 1 - Completo

Análise gráfica



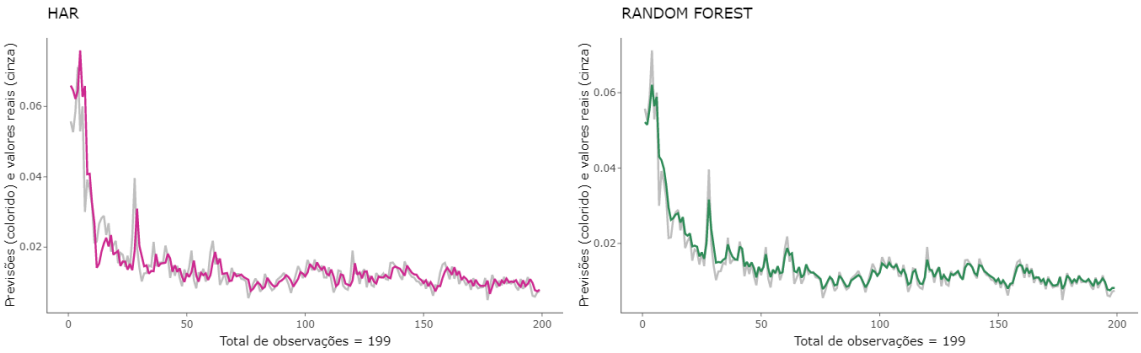
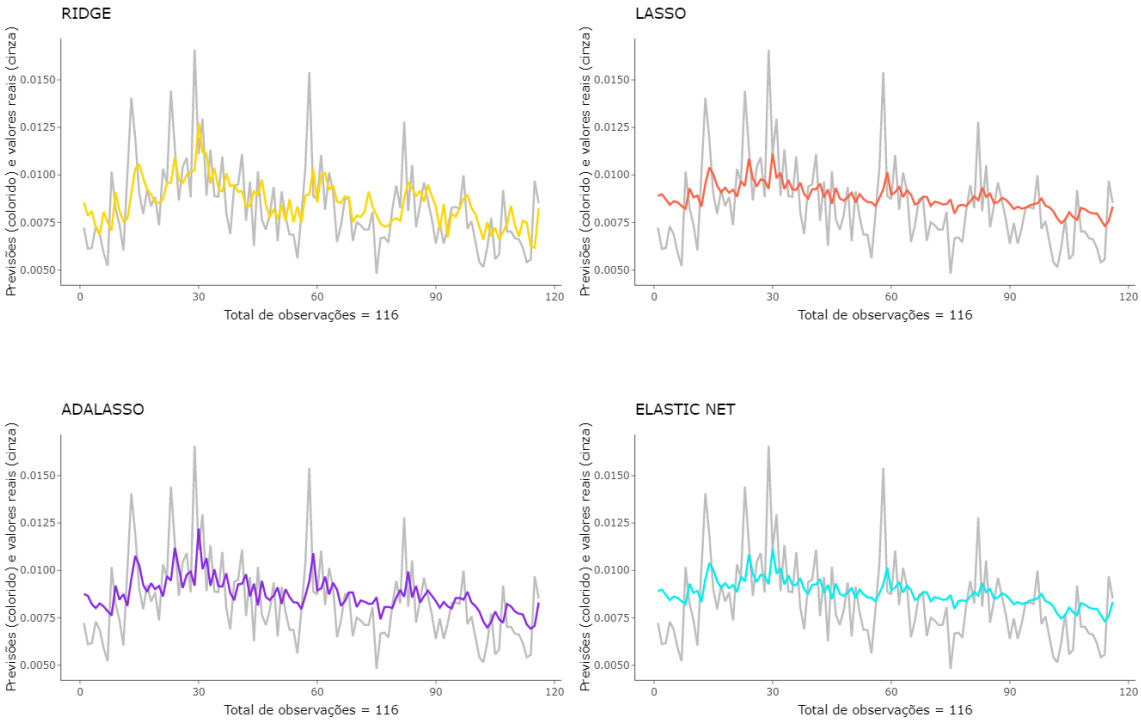


Figura 4.1: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 199 observações.

É possível perceber que as previsões de todos os conjuntos de modelos analisados parecem estar ajustadas aos valores reais, embora nenhum dos conjuntos de modelos consiga captar picos de volatilidade nos valores reais corretamente.

4.1.2 Banco 1 - Pré-pandemia

Análise gráfica



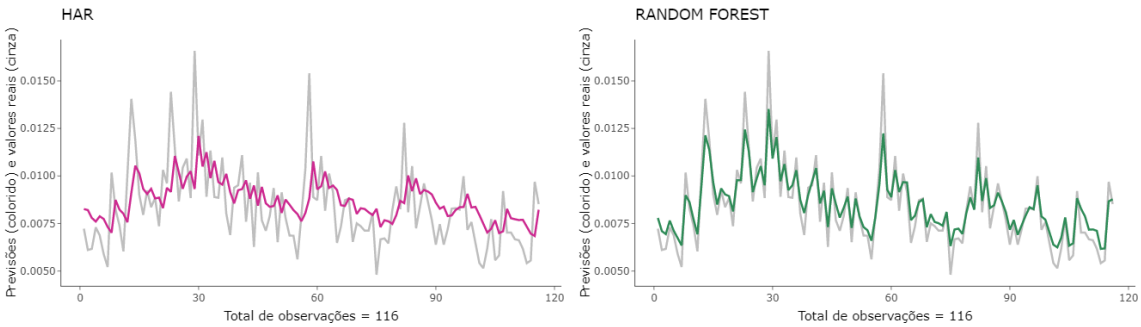
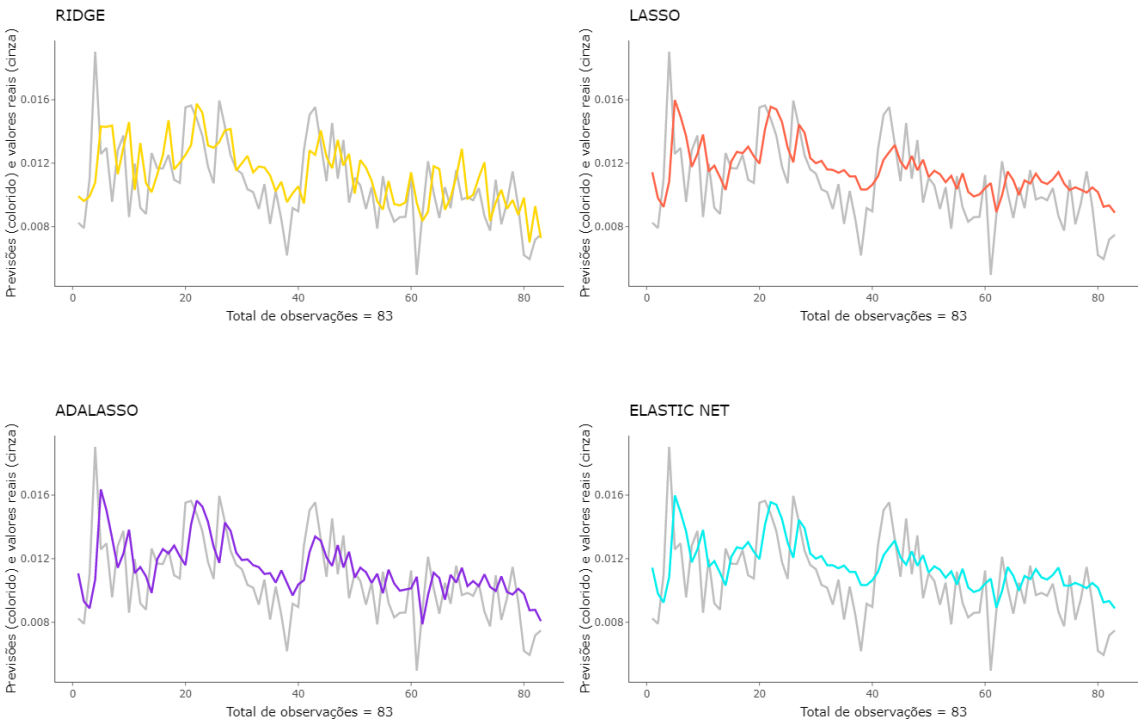


Figura 4.2: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 116 observações.

Os gráficos mostram visualmente que algumas previsões parecem melhores do que outras, por exemplo, as previsões realizadas através de modelos gerados pelo método *Random forest* parecem se ajustar muito bem aos dados.

4.1.3 Banco 1 - Em pandemia

Análise gráfica



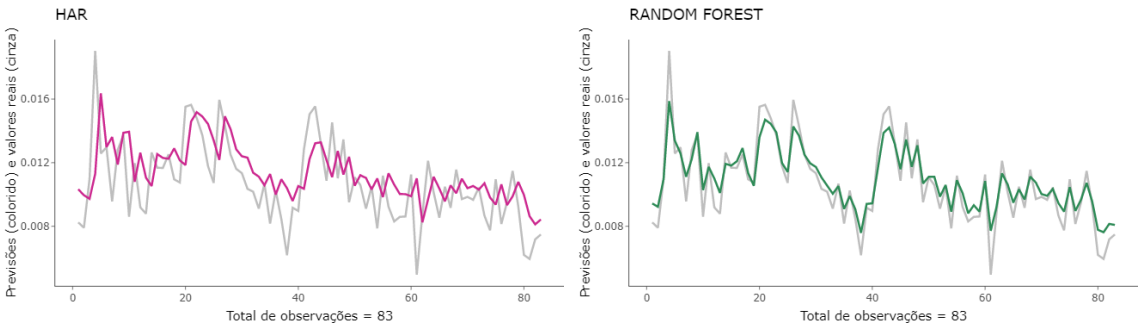
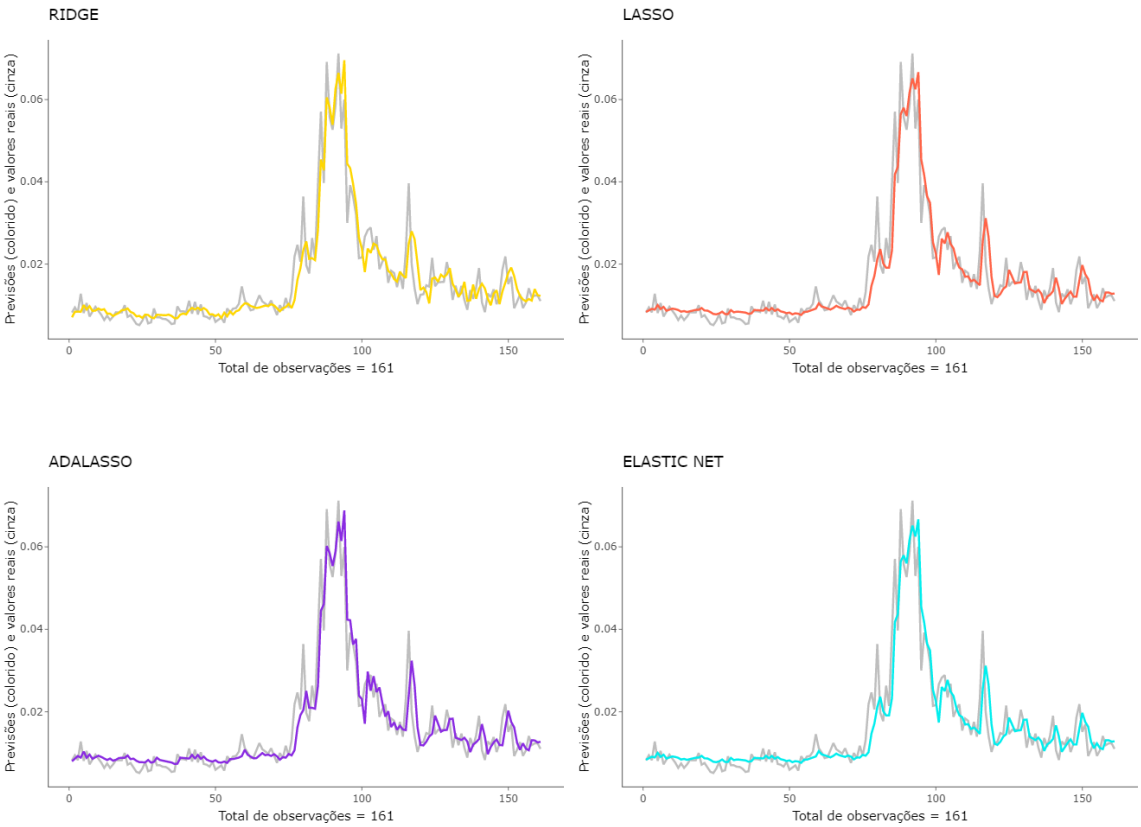


Figura 4.3: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 83 observações.

Todas as previsões parecem muito similares entre si, com exceção das previsões realizadas pelos modelos do método *Random forest*, pois estas parecem se ajustar muito bem aos dados. Embora, seja pertinente ressaltar que nenhum conjunto de modelos conseguiu efetuar previsões que captassem valores extremos para volatilidade.

4.1.4 Banco 1 - Início da pandemia

Análise gráfica



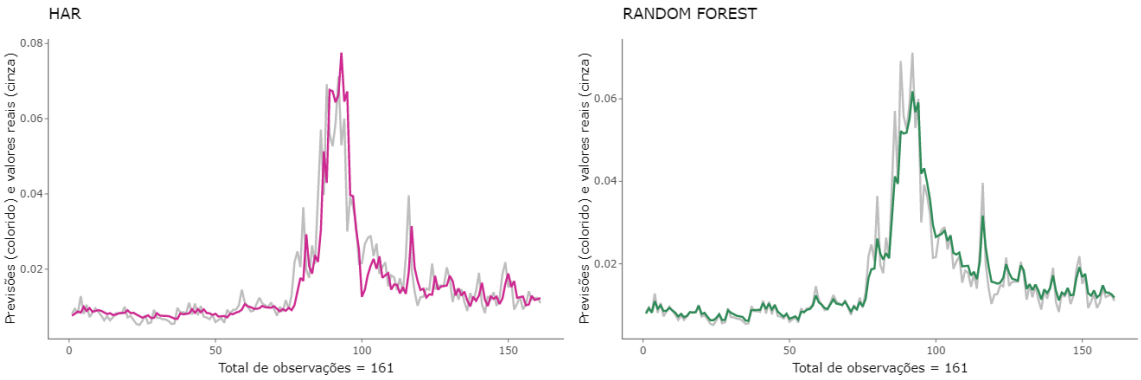


Figura 4.4: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 161 observações.

Através dos gráficos, nota-se que picos de volatilidade existiram no início da pandemia, como já se imaginava, afinal, momentos de incerteza afetam mercados financeiros. Por mais que as previsões desses picos não sejam as melhores possíveis, ao que tudo indica, as propostas conseguiram realizar previsões razoáveis para um período conturbado.

Model confidence set

Para as duas funções de perda (MSE_1, MAE_1) combinadas com a estatística de teste (T_{max}), o *Model confidence set* mostrou que *Random forest* obteve as melhores previsões para todos os particionamentos do banco 1. De modo que esse método foi relevante para todos os períodos analisados, ou seja, para dados pré-pandemia, durante a pandemia e também de início da pandemia, além de dados compreendendo todo o período do banco 1.

Erros

Considerando as tabelas (4.1 e 4.2), os erros (MSE e MAE) de cada proposta serão apresentados em relação ao modelo *benchmark* (HAR). Os valores maiores que 1 (em negrito) mostram que o modelo HAR foi melhor em relação à proposta comparativa para o erro analisado. Já os valores menores que 1 mostram que a proposta foi melhor que o modelo *benchmark* para o erro em questão, e quanto menor é esse valor, melhor é a proposta.

Na tabela 4.1, nota-se que há muitos valores maiores que 1, principalmente para as propostas LASSO, AdaLASSO e *Elastic Net*, ainda que esses valores sejam muito próximos de 1, indicando que não há grandes diferenças entre as propostas e o modelo *benchmark*. Apenas a proposta *Random forest* se destaca, de tal forma que para o banco com dados durante a pandemia para o erro MSE a relação entre os erros de *Random forest* e do modelo HAR chega a ser menor que 0.2, isto é, a perda de *Random forest* teve redução de mais de 80% do erro em relação ao HAR, evidenciando sua superioridade.

Para o ínicio da pandemia e considerando todo o período, a tabela 4.2 mostra que todas as propostas são melhores que o modelo *benchmark*, por mais que algumas tenham reduções de 9% ou 12% do erro e outras tenham redução de mais de 80%

Tabela 4.1: Erros (MAE e MSE) para o banco 1

Propostas	Pré-pandemia		Em pandemia	
	MAE	**MSE**	**MAE**	**MSE**
Ridge	0.9009032	0.8567754	0.9937581	0.9939560
LASSO	**1.0138424**	**1.0108056**	**1.0469962**	**1.0560291**
AdaLASSO	**1.0027552**	0.9783263	**1.0217731**	**1.0309573**
Elastic Net	**1.0138910**	**1.0109478**	**1.0437452**	**1.0498606**
Random Forest	0.4167323	0.1761831	0.3946012	0.1688946

Tabela 4.2: Erros (MAE e MSE) para o banco 1

Propostas	Início da pandemia		Todo período	
	MAE	**MSE**	**MAE**	**MSE**
Ridge	0.7300153	0.3694290	0.8714640	0.5495609
LASSO	0.7815565	0.4531022	0.8814723	0.6022350
AdaLASSO	0.7692803	0.4262201	0.9148127	0.6326439
Elastic Net	0.7822315	0.4539610	0.8843297	0.6049605
Random Forest	0.4694729	0.2245583	0.4446502	0.1840271

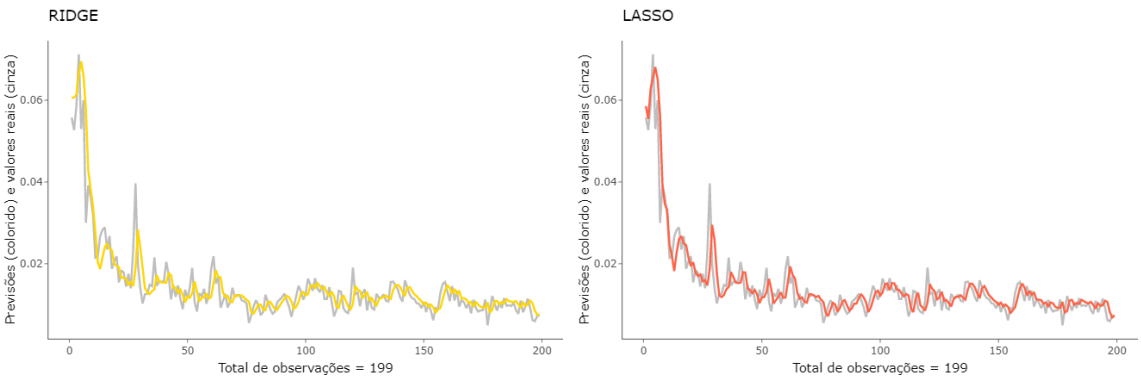
do erro em relação ao modelo *benchmark*. Novamente, é notável a superioridade de *Random forest*, devido aos seus baixos valores.

4.2 Banco 2

Para os resultados obtidos pelos modelos construídos considerando o banco 2 com covariáveis de médias aritiméticas simples dos dias anteriores, analisam-se os gráficos e o *Model confidence set*. Ademais, será apresentada uma tabela com os valores dos erros de cada proposta em relação ao modelo *benchmark*.

4.2.1 Banco 2 - Completo

Análise gráfica



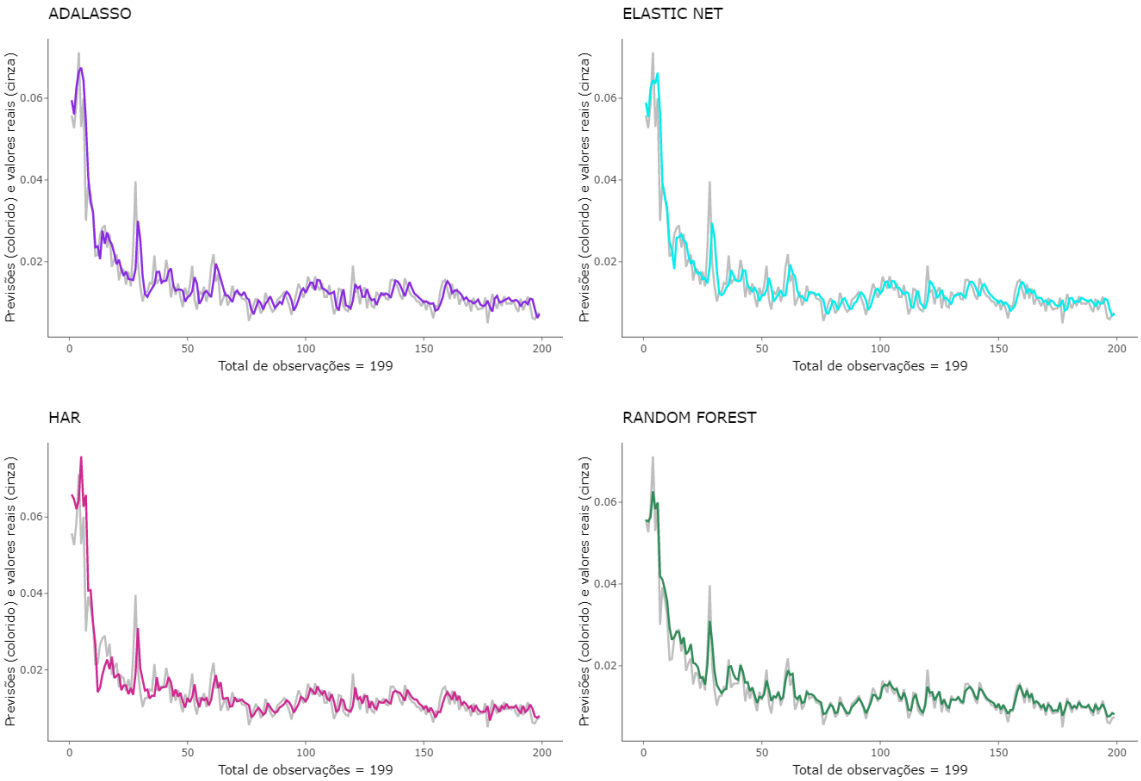
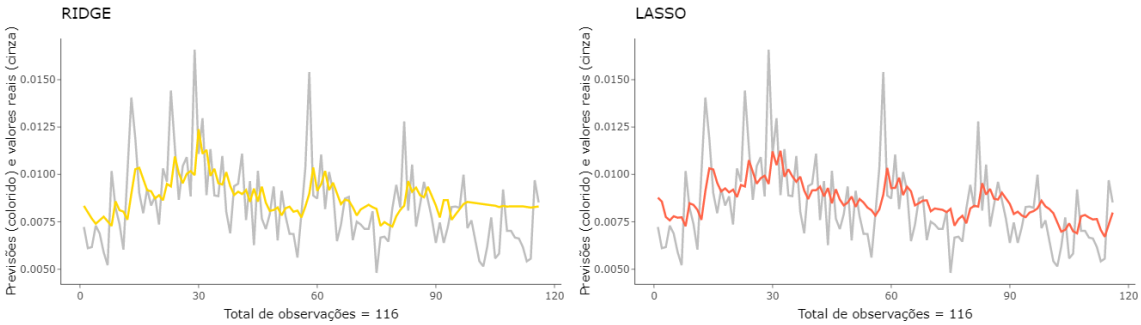


Figura 4.5: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 199 observações.

As previsões de todos os conjuntos de modelos analisados não parecem muito distintas entre si, isto é, elas parecem estar ajustadas aos valores reais, apesar da ressalva de que nenhum dos conjuntos de modelos consegue captar mudanças abruptas de volatilidade dos valores reais.

4.2.2 Banco 2 - Pré-pandemia

Análise gráfica



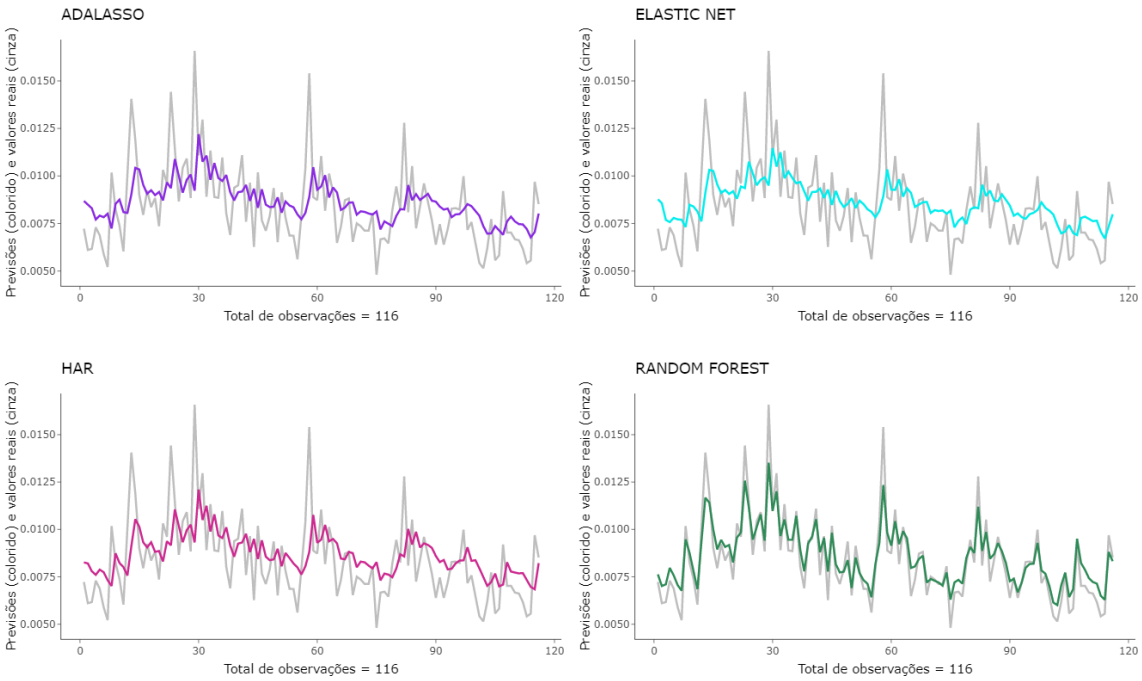
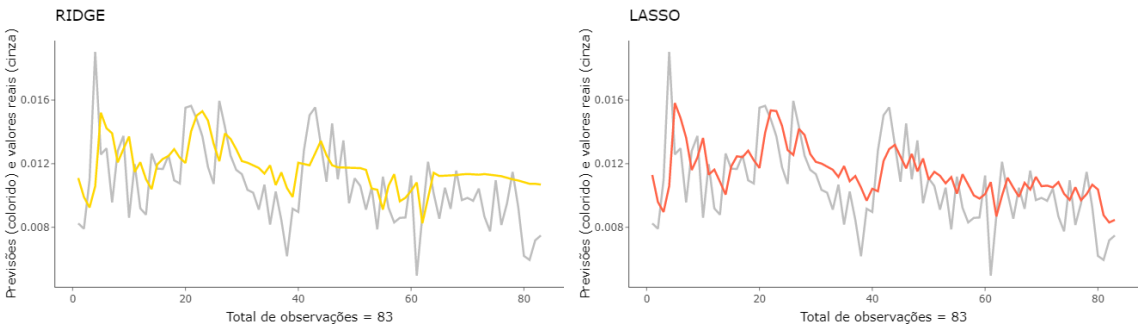


Figura 4.6: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 116 observações.

Por mais que os gráficos para os modelos provenientes dos métodos de regularização aparentem estar melhor ajustados para esse banco, o gráfico que claramente mostra previsões muito boas para os valores reais é relativo ao método *Random forest*.

4.2.3 Banco 2 - Em pandemia

Análise gráfica



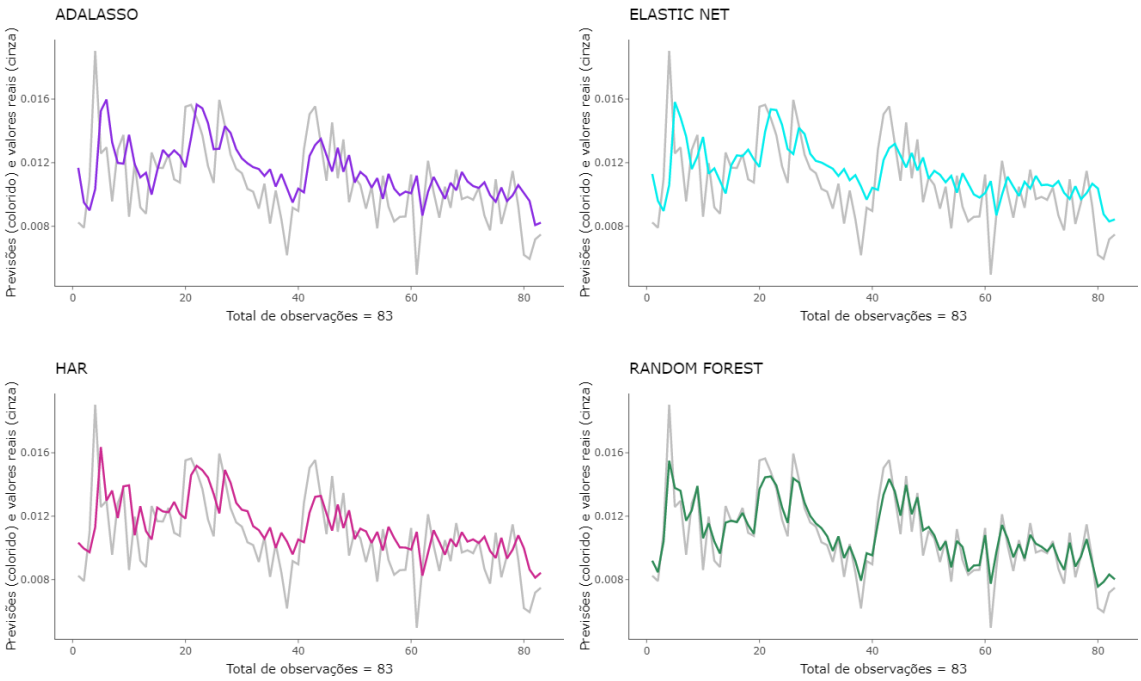
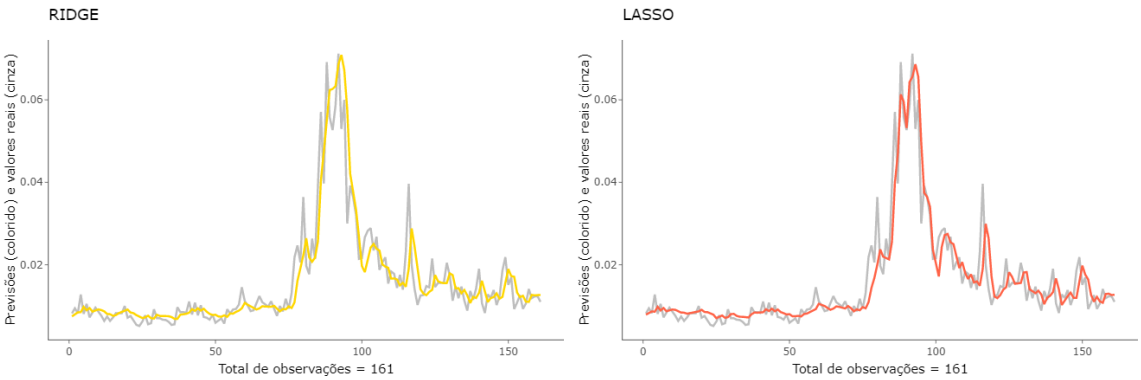


Figura 4.7: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 83 observações.

Pelo gráfico relativo à regressão Ridge, parece que essas previsões são as piores. As demais previsões não parecem muito melhores. Novamente, o melhor gráfico corresponde ao das previsões do conjunto de modelos gerados pelo método *Random forest*.

4.2.4 Banco 2 - Início da pandemia

Análise gráfica



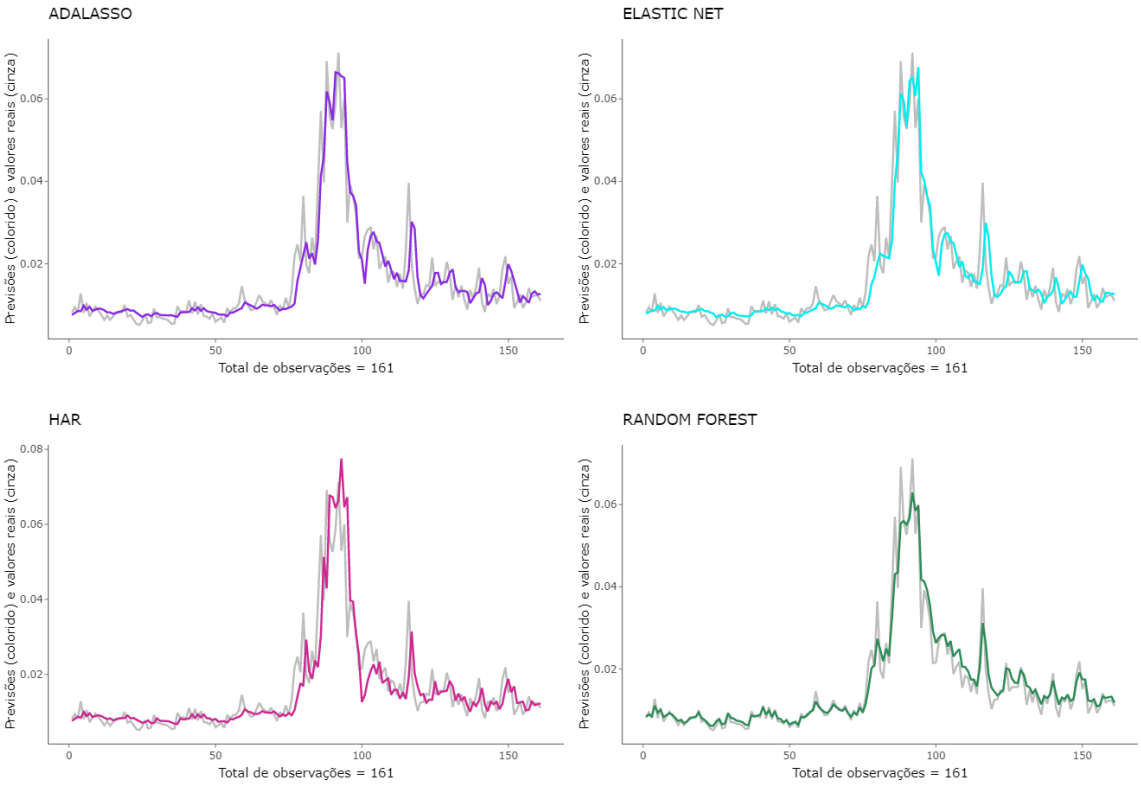


Figura 4.8: Previsões de cada modelo (linhas coloridas) em contraste com os valores reais (linhas cinzas) para as 161 observações.

Constatando a existência de picos de volatilidade no início da pandemia, nota-se que as previsões realizadas parecem razoáveis, considerando os métodos de *machine learning* e modelo HAR.

Model confidence set

O *Model confidence set* mostrou que a *Random forest* produziu as melhores previsões para cada um dos dois procedimentos realizados para todos os particionamentos do banco 2, considerando as duas funções de perda (MSE_1, MAE_1) combinadas com a estatística de teste (T_{max}). Isto é, *Random forest* se destacou para dados pré-pandemia, durante a pandemia, início da pandemia e também para todos os dados do banco 2.

Erros

Dados os erros (MSE e MAE) de cada proposta apresentados nas tabelas (4.3 e 4.4) em relação ao modelo *benchmark* (HAR), tem-se que para os valores maiores que 1 (em negrito), o modelo HAR foi melhor em relação à proposta comparativa para o erro analisado. Todavia, os valores menores que 1 indicam que a proposta foi melhor que o modelo *benchmark* para o erro analisado, e quanto menor é esse valor, melhor é a proposta.

Sendo assim, na tabela 4.3, nota-se que para o banco com dados durante a pandemia a maioria das propostas não foi superior ao modelo HAR, embora muitos dos valores tenham ficados próximos de 1, o que mostra valores muito similares para

Tabela 4.3: Erros (MAE e MSE) para o banco 2

Modelos	Pré-pandemia		Em pandemia	
	MAE	**MSE**	**MAE**	**MSE**
Ridge	**1.0018692**	0.9917201	**1.0987179**	**1.182582**
LASSO	0.9633784	0.9405467	**1.0191054**	**1.032369**
AdaLASSO	0.9872426	0.9731800	**1.0330668**	**1.064278**
Elastic Net	0.9623657	0.9373279	**1.0170247**	**1.028046**
Random Forest	0.4451349	0.2027732	0.4172979	0.190187

Tabela 4.4: Erros (MAE e MSE) para o banco 2

Modelos	Início da pandemia		Todo período	
	MAE	**MSE**	**MAE**	**MSE**
Ridge	0.8287798	0.5996084	0.9235832	0.7497220
LASSO	0.7891665	0.5154433	0.9039817	0.7109504
AdaLASSO	0.7814095	0.4809493	0.9032518	0.6828522
Elastic Net	0.7651059	0.4562706	0.8951990	0.6855024
Random Forest	0.4695232	0.2067033	0.4724605	0.2017870

os erros. Ademais, o método *Random forest* foi visivelmente superior.

Para os dados de ínicio de pandemia e também para os dados de todo período, percebe-se que todas as propostas são superiores ao modelo *benchmark* para os dois erros. Têm propostas com redução de quase 7% do erro, que é exemplificado pela proposta Ridge com o erro MAE para o banco completo. E, há, outras propostas com um pouco mais de 80% de redução em relação ao erro, caso de *Random forest* para o banco completo e o erro MSE. Além disso, o método *Random forest* foi, novamente, superior.

4.3 Comparações entre os bancos 1 e 2

Apresentado na subseção 2.4.2, o teste Diebold-Mariano foi usado com intuito de comparar os quatro distintos cenários entre os bancos 1 e 2, isto é, através das melhores previsões de cada banco (em todos os casos essas previsões são relativas ao *Random forest*). Dessa forma, foram construídos quatro testes (um para cada cenário) e seus resultados são apresentados na tabela 4.5. Para o início da pandemia e todo o período analisado os dois testes rejeitaram a hipótese de não existirem diferenças entre as previsões para banco 1 e 2, já para os momentos pré-pandemia e em pandemia os outros dois testes não rejeitaram essa hipótese para os bancos 1 e 2.

Tabela 4.5: Teste DM

Cenários	Teste Diebold-Mariano	
	Estatística (DM)	**P-valor**
Todo período	-2.3437	0.0201
Início da pandemia	-3.4424	0.0007
Em pandemia	0.00966	0.9923
Pré-pandemia	-0.7042	0.4827

4.4 Discussões

Devem ser discutidos alguns pontos relevantes para todos os bancos de dados. Primeiramente, para dois cenários analisados (*pré-pandemia* e *em pandemia*), não houve diferenças notáveis entre previsões entre os dois bancos de dados criados a partir do vetor de volatilidades realizadas diárias com covariáveis diferentes. Em outras palavras, foi indiferente usar apenas defasagens ou médias de defasagens para obter previsões para os dois bancos. No entanto, para os outros dois cenários (*início da pandemia* e *todo o período*), houve diferenças entre previsões considerando os dois bancos.

Ainda, é necessário incluir nas discussões as considerações acerca dos tamanhos de amostra. O aumento no tamanho da janela móvel usada para realizar as previsões implicou em previsões mais precisas. Em outras palavras, os cenários com maiores janelas móveis, e consequentemente, com maiores vetores de previsões, apresentaram previsões mais precisas. De forma que é interessante analisar em paralelo como a modificação de 22 covariáveis, considerando os bancos 1 e 2, é significativa para os cenários com mais observações e indiferente para os cenários com menos observações.

Há, também, a questão mais importante de toda análise, isto é, o bom desempenho do método *Random forest* em todos os procedimentos, considerando o *Model confidence set* para os dois bancos de dados criados e seus particionamentos.

Uma hipótese para esses resultados é de que não há diferenças significativas de capacidade preditiva entre os métodos de regularização e o modelo HAR como se pensou inicialmente para realização de previsões de volatilidade realizada diária. Pois, esperava-se que ao incluir mais covariáveis, as previsões apresentassem mais qualidade em relação aos erros de previsão. Uma maior qualidade de previsão só ocorreu para o método de árvore utilizado (*Random forest*), e em algumas situações, o que pode não ter relação com incluir mais covariáveis, e sim com as particularidades do método.

Há outra situação evidenciada pelos resultados: variações abruptas de volatilidade não parecem ser previstas de forma precisa por nenhuma das propostas. Esse fato é ilustrado nos vários gráficos em que os picos em cinza, representando variações bruscas de volatilidade, não foram identificados ou previstos. O mesmo não ocorre nos gráficos respectivos aos bancos completos, em que as volatilidades realizadas diárias seguem altas por um certo período e tornam a diminuir, nesses casos, os conjuntos de modelos parecem conseguir realizar previsões razoáveis. Ainda que seja importante evidenciar essa situação, deve-se compreender que esses resultados são razoáveis, uma vez que é muito difícil que alguma das propostas consiga realizar previsões abruptas corretamente, afinal prever picos é difícl, principalmente estando fora da amostra.

5 Considerações finais

A previsão de volatilidade dos ativos financeiros tem vital importância para o mercado financeiro. Portanto, comparar previsões de volatilidade realizada diária obtidas, por diferentes propostas, é relevante, principalmente, para entender quais propostas usar em situações específicas.

Das diferenças preditivas entre as propostas, tem-se que não foram identificadas diferenças preditivas entre o modelo *benchmark* (HAR) e os demais métodos de regularização, dado que nem o modelo HAR, ou tampouco os métodos de regularização estiveram incluídos no *Model confidence set* para o nível de significância fixado em 50%. No entanto, pode-se afirmar que ocorreu a prevalência da *Random forest* em relação ao modelo *benchmark* (HAR) para os dois bancos de dados criados e seus particionametos.

Das considerações acerca da divisão dos dados em relação às covariáveis, tem-se que essa divisão foi relevante para os cenários com mais observações, ou seja, com dados no início da pandemia e durante todo o período, todavia, essa relevância não ocorreu para dados pré-pandemia ou durante a pandemia, que são cenários com menos observações.

Das considerações acerca da divisão dos dados devido à pandemia de COVID-19, concluiu-se que tanto previsões obtidas através de conjuntos de modelos gerados por métodos de regularização quanto previsões obtidas pelo modelo HAR foram igualmente relevantes, em outras palavras, não houve diferenças significativas entre essas previsões segundo o *Model confidence set* considerando os cenários distintos. Ainda assim, é factível concluir que entre o modelo *benchmark* (HAR) e o método de árvore utilizado (*Random forest*), em uma performance geral, o método de árvore obteve melhores resultados para todos os particionamentos realizados considerando o cenário pandêmico.

Do interesse em investigar se o cenário pandêmico afetou as séries de volatilidade e suas previsões, notou-se que houve picos de volatilidade no período de início de pandemia. Além disso, *Random forest* se mostrou o método mais relevante para obtenção de previsões para o período.

Das consideranções gerais, tem-se uma conclusão: é benéfico a quaisquer futuros estudos ampliar o uso de métodos de árvores em aplicações para prever volatilidade realizada diária. É importante lembrar que as previsões não são um fim, mas um meio de obter informações para posteriores tomadas de decisões (Morettin e Toloi, 2008). Portanto, havendo possibilidade de analisar mais dados de volatilidade realizada diária, pode-se explorar a possibilidade de outras propostas se destacarem, principalmente por não ser ideal supor que em todos os estudos neste campo esse

método seja sempre o melhor.

Das conclusões finais, tem-se que no presente estudo, para os dados usados, o método *Random forest* se sobressaiu em todos os casos, logo, é importante considerá-lo para realizar previsões de volatilidade realizada diária.

Referências Bibliográficas

Alvarenga, T. C. (2015). Modelo heterogêneo autorregressivo: uma aplicação a dados de mortalidade e a dados climáticos. Master's thesis, Universidade Federal de Lavras, Lavras.

Andersen, T. G. e Bollerslev, T. (1998). Deutsche mark-dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *The Journal of Finance*, 53(1):219–265.

Andersen, T. G., Bollerslev, T., Diebold, F. X., e Labys, P. (2003a). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Andersen, T. G., Bollerslev, T., Diebold, F. X., e Vega, C. (2003b). Micro effects of macro announcements: Real-time price discovery in foreign exchange. *The American Economic Review*, 93(1):38–62.

Bernardi, L. C. . M. (2017). *MCS: Model Confidence Set Procedure*. R package version 0.1.3.

Boudt, K., Kleen, O., e Sjørup, E. (2022). Analyzing intraday financial data in r: The highfrequency package. *Journal of Statistical Software*, 104(8):1–36.

Breiman, L. (1993). *Classification and regression trees*. Chapman Hall.

Corsi, F. (2008). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Cryer, J. D. e Chan, K.-S. (2008). *Time series analysis*. Springer.

Diebold, F. X. e Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business &amp Economic Statistics*, 13(3):253.

Friedman, J., Hastie, T., e Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Goodfellow, I., Bengio, Y., Courville, A., e Bach, F. (2017). *Deep Learning*. MIT Press.

Hansen, P. R. e Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7):873–889.

Hastie, T., Tibshirani, R., e Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., e Yasmeen, F. (2023). *forecast: Forecasting functions for time series and linear models*. R package version 8.20.

Hyndman, R. J. e Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

James, G. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer New York.

Junior, M. V. W. e Pereira, P. L. V. (2013). Modeling and forecasting of realized volatility: Evidence from brazil. *Brazilian Review of Econometrics*, 31(2):315.

Konzen, E. e Ziegelmann, F. A. (2016). LASSO-type penalties for covariate selection and forecasting in time series. *Journal of Forecasting*, 35(7):592–612.

Liaw, A. e Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Marmitt, J. (2012). Dados de alta frequência : averiguando o impacto de microestrutura de mercado e sazonalidade intradiária na detecção de saltos e estimação da variação quadrática. Master's thesis, Universidade Federal do Rio Grande do Sul. Faculdade de Ciências Econômicas. Programa de Pós-Graduação em Economia, Porto Alegre, BR-RS.

Morettin, P. A. e Toloi, C. M. (2008). *Analise de series temporais*. 2 edition.

Peter R. Hansen, A. L. e Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Posit team (2022). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ricco, R. d. A. (2021). Realized semicovariances : empirical applications to volatility forecasting and portfolio optimization. Master's thesis, Universidade Federal do Rio Grande do Sul. Faculdade de Ciências Econômicas.

SCHWERT, G. W. (1989). Why does stock market volatility change over time? *The Journal of Finance*, 44(5):1115–1153.

Shumway, R. H. e Stoffer, D. S. (2011). *Time Series Analysis and Its Applications*. Springer New York.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Simon, N., Friedman, J., Hastie, T., e Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Sjoerup, E. (2019). *HARModel: Heterogeneous Autoregressive Models*. R package version 1.0.

Taylor, S. J. (2011). *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

WOOD, R. A., McINISH, T. H., e ORD, J. K. (1985). An investigation of transactions data for NYSE stocks. *The Journal of Finance*, 40(3):723–739.

Zhang, Y., Li, R., e Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. e Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.