Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

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

Trabalho de Conclusão de Curso

Previsão da Evolução da Covid-19 utilizando métodos de *Machine Learning*

Giulia Bagatini Carlotto

23 de novembro de 2021

Giulia Bagatini Carlotto

Previsão da Evolução da Covid-19 utilizando 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: Prof. Dr. Flávio A. Ziegelmann

Porto Alegre
Novembro de 2021

Giulia Bagatini Carlotto

Previsão da Evolução da Covid-19 utilizando 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 pelo Orientador e pela Banca Examinadora.

Orientador:_____
Prof. Dr. Flávio A. Ziegelmann, UFGRS
Doutor pela University of Kent at Canterbury, UK

Banca Examinadora:

Prof. MSc. Klaus Böesch, IFSul
Mestre pela Universidade Federal do Rio Grande do Sul – Porto Alegre, RS

Porto Alegre
Novembro de 2021

"Que todos os nossos esforços estejam sempre focados no desafio à impossibilidade. Todas as grandes conquistas humanas vieram daquilo que parecia impossível."
(Charles Chaplin)

Agradecimentos

Começo agradecendo a Deus por, ao longo deste processo complicado e desgastante, ter me feito enxergar o caminho nos momentos em que pensei em desistir.

Aos meus pais, Josiana e Vanderlei, por todo apoio e incentivo que serviram de alicerce para as minhas realizações, pela educação que me proporcionaram e por não medirem esforços para que eu chegasse até esta etapa da minha vida. A vocês eu devo a vida e todas as oportunidades que nela tive e espero um dia poder lhes retribuir.

Ao meu namorado Matheus, pela compreensão, paciência e companheirismo ao longo desses anos e, principalmente, durante o período do projeto. Obrigada por ser meu melhor amigo e sempre me incentivar e acreditar que eu seria capaz de superar mais essa importante jornada.

Aos amigos que tive a oportunidade de conhecer e conviver intensamente durante os últimos anos da faculdade: Bruna, Franciele, Eduardo, Gabriela, Juliana, Lincon, Rafaela, Renan e Tainá. O companheirismo e a troca de experiências foram essenciais nesse caminho árduo. Quero deixar um agradecimento muito especial à Franciele e à Juliana, porque sem o suporte delas, este último semestre teria sido quase impossível. Aos demais colegas da estatística, foi um prazer encontrar vocês nessa caminhada, que possamos nos esbarrar em muitas outras oportunidades da vida.

Aos meus amigos da vida inteira, Maria Eduarda e Pedro. Obrigada por estarem presentes em todos os momentos da minha vida, por sempre me ouvirem e aconselharem e pela amizade incondicional.

Aos professores do Departamento de Estatística da UFRGS, por todo conhecimento compartilhado e pela incansável dedicação para com os alunos. Em especial ao meu orientador, professor Flávio, por ter feito parte da minha jornada na graduação desde 2018, sendo meu orientador na bolsa de Iniciação Científica e neste trabalho, muito obrigada pelo incentivo e confiança e, principalmente, por compreender as adversidades que surgiram nesse momento atípico que estamos vivendo. Agradeço, também, ao professor Klaus por ter aceitado fazer parte da banca deste trabalho e pelas considerações feitas.

Por fim, um agradecimento especial à UFRGS pela elevada qualidade do ensino oferecido, pela oportunidade de aprendizado e crescimento tanto profissional quanto pessoal.

Resumo

Com o surgimento da recente pandemia global da Covid-19, vários modelos de previsão de séries temporais vêm sendo utilizados pelos governos como instrumentos na tomada de decisões, auxiliando na maneira de rastrear e gerenciar a evolução da epidemia. Devido à ampla disponibilidade de dados referentes à crise, torna-se imprescindível que métodos modernos sejam propostos, explorando ao máximo a informação existente. Nesse contexto, propõem-se a utilização de métodos de *Machine Learning* para prever o número de casos e fatalidades do novo vírus em metrópoles brasileiras. A pesquisa baseia-se em dados disponibilizados por órgãos governamentais e integra variáveis medidas diariamente em múltiplos domínios, como dados da evolução da epidemia e covariáveis ambientais. Cabe salientar que nosso objetivo será o de explorar estas técnicas dentro deste problema aplicado, sem a pretensão, entretanto, de que efetivamente atuemos em tempo real para execução das previsões. Ou seja, o trabalho possuirá um caráter *ex post* e não um *ex ante*. Além disso, estaremos realizando um exercício numérico que inclui simulação de Monte Carlo, para auxiliar nas implementações e análises dos dados empíricos, de modo a comparar as performances desses estimadores em um cenário conhecido. A comparação dos métodos é feita através de medidas de erro de previsão. Como resultados obtidos, observou-se que as previsões da quantidade de infectados pela doença ficaram muito próximas aos dados reais coletados para todos os métodos e ligeiramente melhores que as previsões de número de vítimas fatais, que, por sua vez, tiveram variações um pouco maiores em relação aos dados reais. Apesar disso, os resultados verificam a eficácia desses métodos em termos de simulação e previsão da tendência de surto de Covid-19.

Palavras-Chave: Covid-19, Pandemia, Séries Temporais, Aprendizado de Máquina, Previsão.

Abstract

With the emergence of the recent Covid-19 global pandemic, several time series prediction models have been used by governments as decision-making tools, helping to track and manage the evolution of the disease. Due to the wide availability of data related to the crisis, it is essential that modern methods are proposed, exploring the important information available. In this context, it is suggested the use of Machine Learning methods to predict the number of cases and deaths of the new virus in Brazilian metropolises. The research is based on data provided by government agencies and it uses weekly updated variables of multiple domains, such as data on the evolution of the epidemic and environmental covariates. It should be noted that our objective is to explore these technics within this applied problem, without the pretension, however, that we actually act in real time to execute the forecasts. In other words, this research will have an ex post character, not an ex ante. In addition, we will be performing a numerical exercise that includes Monte Carlo simulation, to assist in the implementation and analysis of empirical data, in order to compare the performance of these estimators in a known scenario. Comparison of methods is done through predictive error measures. As a result, it was observed that the forecasts for the number of people infected by the disease were very close to the actual data collected for all methods and slightly better than the forecasts for the number of fatal victims, which, in turn, had slightly greater variations in relation to the real data. Despite this, the results verify the effectiveness of these methods in terms of simulation and prediction of the Covid-19 outbreak trend.

Keywords: Covid-19, Pandemic, Time Series, Machine Learning, Forecast.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A recente pandemia global, consequência da Covid-19, uma doença causada pelo novo Coronavírus SARS-CoV-2, que manifesta um quadro clínico que varia de infecções assintomáticas a quadros respiratórios graves, teve início em Wuhan, China, em dezembro de 2019. Rapidamente, a Covid-19 ocasionou danos extremos em quase todo o mundo (Zhu et al., 2020). O Brasil, por sua vez, teve o primeiro caso da doença no final do mês de fevereiro de 2020 em São Paulo. Alguns dias depois, foi registrada a primeira transmissão interna no país. Assim, foi possível observar inúmeras medidas de proteção sendo expandidas por todo território nacional, como já haviam sendo efetivadas em alguns outros países.

Desde então, tornou-se indispensável trabalhar com os dados coletados sobre o vírus. Estudos como o publicado por um influente grupo do *Imperial College London* (Walker et al., 2020) começaram a ser realizados para prever diferentes cenários da pandemia causada pelo novo Coronavírus. Muitos modelos de previsão de séries temporais vêm sendo utilizados por diversos pesquisadores – modelos como o autorregressivo integrado de médias móveis (ARIMA) empregado por Tandon et al. (2020), por exemplo – para auxiliar governos e órgãos legislativos na maneira de rastrear e gerenciar a evolução da doença.

A grande evolução no número de casos e mortes dessa pandemia pode ser influenciada por diferentes fatores, como por exemplo: temperatura, qualidade e umidade do ar, variáveis que já foram estudadas em ligação com o Covid (Ward et al., 2020; Sehra et al., 2020). Além disso, a literatura que busca fazer previsões de casos e óbitos de Covid-19 utiliza variáveis mais inovadoras, como as buscas no Google (Fantazzini, 2020). Em outras palavras, esses componentes podem ser encarados como covariáveis. O conjunto de dados a ser aplicado nesse estudo, em especial, combina informações do próprio passado da série temporal com essas covariáveis, que podem ser úteis para fazer inferência.

Diante desse cenário, dadas as grandes quantidades de informações e variáveis (em relação ao número de observações) que estão disponíveis, é imprescindível que métodos modernos sejam propostos e utilizados de tal forma que explorem ao máximo essa informação. Em séries temporais, particularmente, há métodos (e modelos) capazes de lidar com grandes números de variáveis, visto que incluem técnicas de selecionar as relevantes, reduzindo a dimensão do banco de dados. Os métodos de aprendizado de máquina (do inglês *Machine Learning*) (Friedman et al., 2001), especialmente, vêm destacando-se e estabelecendo-se como relevantes candidatos na esfera de previsão.

Embora a abordagem de aprendizado de máquina tenha sido definida como ins-

trumento padrão para a modelagem de desastres naturais (Pyayt et al., 2011; Asim et al., 2017) e previsão do tempo (Holmstrom et al., 2016; Zaytar e El Amrani, 2016), sua implementação em questões epidemiológicas ainda é um tópico muito atual. Por esse motivo, surge a motivação de aplicar tais abordagens para análise da recente crise de saúde. O objetivo principal deste trabalho consiste em explorar os resultados do uso combinado de técnicas de séries temporais e aprendizado de máquina para prever a evolução do número de infectados e óbitos por Covid-19 nas três cidades com o maior número de casos e mortes por Covid-19 no Brasil: São Paulo, Rio de Janeiro e Brasília.

Antes do estudo aplicado, disponibilizaremos resultados de um exercício numérico, em que avaliamos a performance dos métodos propostos em termos de simulações de Monte Carlo. Isso é realizado com o intuito de facilitar as implementações dos dados empíricos, mas principalmente, para comparar as performances desses estimadores em um cenário conhecido (simulação), ou seja, em que possamos medir com precisão o melhor estimador em termos de medida de erro, dado que observamos uma resposta sem erros causados por fatores externos, por exemplo. Ainda, na simulação, utilizaremos como conjunto de variáveis independentes as defasagens das covariáveis e, também, da própria variável resposta, que é um procedimento muito similar ao que é implementado para os dados reais, apesar da estrutura de dados (especificamente da variável dependente) ser desconhecida, mas potencialmente diferente dos dados simulados.

Para tanto, estaremos propondo a aplicação de métodos estatísticos envolvendo *Machine Learning* que desempenhem bem no campo de previsão de séries temporais. Em particular, sugerimos a aplicação dos seguintes métodos: LASSO (Tibshirani, 1996), adaLASSO (Zou, 2006) e *Random Forest* (Breiman, 2001). Almejamos também comparar os resultados destes métodos àqueles oriundos de modelos *benchmark*, mais especificamente, ao modelo auto-regressivo integrado de médias móveis (ARIMA), sistematizado por Box et al. (1970).

Além desta introdução, o trabalho está dividido da seguinte forma: no capítulo 2 é apresentada a metodologia de pesquisa, que descreve tecnicamente todas as abordagens estudadas. No capítulo 3 tem-se a Análise Numérica, composta de simulações de Monte Carlo e análise empírica e que apresenta o passo a passo realizado para obtenção dos resultados, os quais serão apresentados na sequência. Por fim, no capítulo 4 são apresentadas as considerações finais do estudo realizado, bem como sugestões para estudos futuros.

2 Metodologia

Este capítulo tem como motivação introduzir a base teórica para os conceitos e métodos empregados e discutidos neste estudo, estruturando-se em quatro tópicos principais. Inicialmente, é indispensável mostrar características centrais do vasto conjunto de ferramentas para a compreensão de dados que iremos usufruir, isto é, de *Machine Learning*. Na sequência, vamos instituir os fundamentos teóricos baseados em séries temporais e, ainda, na associação do aprendizado de máquina com séries temporais, seguidos da exemplificação das técnicas a serem abordadas no decorrer do trabalho. Finalmente, será apresentada uma visão geral da aplicação das técnicas de *Machine Learning* na modelagem da pandemia de Coronavírus, além de trazermos outras definições essenciais para o discernimento do estudo.

2.1 Introdução a *Machine Learning*

Em 1959, Arthur Samuel definiu o aprendizado de máquina como o "campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados" e foi, provavelmente, o pioneiro nessa área de pesquisa (Samuel, 1959).

As técnicas de *Machine Learning* fazem uso de computação para programar algoritmos capazes de aprender com os dados de modo a serem treinados. Assim, os algoritmos aprimoram com a experiência e com o tempo, refinando um modelo que pode ser usado para prever o valor da variável de interesse da maneira mais precisa possível com base no aprendizado anterior. Em síntese, pode-se interpretar o aprendizado de máquina como uma aplicação computacional que gera previsões com base em propriedades conhecidas e aprendidas com dados de treinamento, encontrando padrões nos dados (Hastie et al., 2009).

A maioria dos problemas em *Machine Learning* compreende uma das duas categorias: não supervisionado ou supervisionado. O aprendizado não supervisionado caracteriza-se por não possuir variável dependente (de saída). O supervisionado, por sua vez, ocorre quando as variáveis de entrada e saída são dadas, e o intuito é explicar a saída em termos das variáveis independentes (James et al., 2013). Como o objetivo deste estudo é a previsão dos números de casos e mortes pela Covid-19 em grandes cidades, isto é, possui uma saída predefinida, iremos trabalhar apenas com aprendizado supervisionado.

Durante o aprendizado, os parâmetros do modelo preditor são ajustados automaticamente aos dados, na tentativa de minimizar o erro de predição, aumentando a capacidade de generalização. Existem, ainda, os hiperparâmetros, que devem ser ajustados buscando evitar o *overfitting*, ou seja, quando o modelo gerado se torna

muito especializado no conjunto de treinamento, obtendo baixo desempenho quando confrontado com novos dados. O *underfitting* também deve ser contido, que ocorre quando o algoritmo não se ajusta adequadamente ao padrão, obtendo um baixo desempenho frente ao conjunto de treinamento (James et al., 2013).

2.1.1 Aprendizado Supervisionado

Conforme James et al. (2013), o objetivo principal no aprendizado supervisionado é construir um modelo capaz de aprender a relação da resposta (variável dependente) com os preditores (variáveis independentes), para, posteriormente, ser utilizado na predição da resposta para observações futuras.

Podemos dizer que é uma maneira de aprendizagem por experiência. Em outras palavras, os algoritmos tentam adquirir conhecimento analisando um conjunto de dados previamente investigado, denominado dados de treinamento, que contém informações capazes de fazer os algoritmos aprenderem a conexão entre os valores de variáveis independentes e dependentes. Além disso, o algoritmo tenta desenvolver um modo de prever a saída desejada de uma entrada baseando-se nesses dados conhecidos. Em seguida, outros dados podem ser alimentados no algoritmo para realizar predições acerca de novos dados, com valores desconhecidos para as variáveis dependentes, chamados dados de teste (Shalev-Shwartz e Ben-David, 2014).

2.2 Séries temporais

De acordo com Box et al. (1970), uma série temporal é um conjunto de observações obtidas sequencialmente ao longo do tempo. Uma característica importante deste tipo de dados é que as observações vizinhas são dependentes e precisamos sempre levar em conta essa ordem temporal. Alguns conceitos são essenciais para a compreensão da teoria ligada a séries temporais e serão descritos a seguir.

Ao analisarmos uma ou mais séries temporais, a representação gráfica dos dados é fundamental, visto que pode revelar padrões de comportamento relevantes. Sendo assim, o ideal é que o gráfico temporal anteceda qualquer análise.

Conforme citam Morettin e Toloi (2006), as séries temporais são compostas por quatro elementos básicos, denominados componentes, e podem ser classificados em:

- tendência: o movimento dos dados a longo prazo, que pode ser causado por qualquer aspecto que afete a variável de interesse (a longo prazo);
- variações cíclicas: variações com grau de regularidade, mas com período superior a um ano;
- variações sazonais: existente quando os fenômenos ocorridos durante o tempo se repetem a cada período idêntico, podendo ser determinística, quando seu padrão sazonal é regular e estável no tempo, ou estocástica, quando a componente sazonal da série varia com o tempo;
- variações irregulares: são variações aleatórias, que não apresentam regularidade.

Ainda segundo Morettin e Toloi (2006), uma das suposições mais frequentes que se faz acerca de uma série temporal é a de que ela é (fracamente) estacionária, isto é, se desenvolve no tempo aleatoriamente ao redor de uma média constante, retratando alguma forma de equilíbrio estável. No entanto, na prática, grande parte das séries encontradas apresentam alguma característica de não-estacionariedade, como tendência e/ou sazonalidade.

Para fazer a verificação de estacionariedade podemos utilizar testes de raiz unitária. Um dos mais empregados na literatura é o Teste de Dickey-Fuller Aumentado (Said e Dickey, 1984), em que temos como hipóteses:

H_0: tem raiz unitária, ou seja, não é estacionária
H_1: não tem raiz unitária, ou seja, é estacionária

Desse modo, se o p-valor do teste for significativo a um $\alpha = 0.05$, temos evidências para rejeitar a hipótese de que a série é não estacionária.

Tendo grande parte dos procedimentos de análise estatística de séries temporais supondo que as séries sejam estacionárias, se estas não forem, será necessário transformar os dados originais. Um tipo especial de filtro, muito útil para remover uma componente de tendência, consiste em tomar diferenças sucessivas da série original, até se obter uma série estacionária. Para isso, vamos considerar a série temporal $Y(t_1), ..., Y(t_n)$, observada nos instantes $t_1, ..., t_n$. A primeira diferença de $Y(t)$ é definida por

$$\Delta Y(t) = Y(t) - Y(t-1), \tag{2.1}$$

a segunda diferença é

$$\Delta^2 Y(t) = Y(t) - 2Y(t-1) + Y(t-2). \tag{2.2}$$

De modo geral, a n-ésima diferença de $Y(t)$ é

$$\Delta^n Y(t) = \Delta[\Delta^{n-1} Y(t)]. \tag{2.3}$$

Sob circunstâncias normais, será suficiente tomar uma ou duas diferenças para que a série se torne estacionária.

Uma forma bastante simples de eliminar o efeito sazonal, por sua vez, é tomar médias sazonais e, para isso, pode-se recorrer ao filtro de médias móveis (MM). A técnica de médias móveis consiste em calcular a média aritmética das r observações mais recentes, isto é,

$$M_t = \frac{Y_t + Y_{t-1} + ... + Y_{t-r+1}}{r}. \tag{2.4}$$

Assim, M_t é uma estimativa que não pondera as observações mais antigas, o que é coerente, dado que o parâmetro varia suavemente com o tempo. Ou seja, a cada período, a observação mais antiga é substituída pela mais recente, calculando-se uma média nova. No entanto, uma desvantagem da técnica é que ela deve ser utilizada somente para prever séries estacionárias, senão, a precisão das previsões obtidas será muito pequena, visto que os pesos atribuídos às r observações são todos iguais e nenhum peso é dado às observações anteriores a esse período (Morettin e Toloi, 2006).

Em síntese, nota-se que o estudo de séries temporais ocorre para diversas finalidades, como para a análise de várias características ligadas à série, a investigação

do mecanismo gerador da mesma e à previsão do futuro com base no conhecimento do passado. Sabe-se que grande parte dos conjuntos de dados da realidade científica são de natureza temporal. Logo, verificamos que a previsão de séries temporais ocupa um papel extremamente relevante na ciência, engenharia e negócios (Palit e Popovic, 2006). Neste trabalho, o interesse está em predizer valores futuros com base em valores passados e, para isso, vamos introduzir algum referencial teórico de aprendizado de máquina ligado diretamente a previsão de séries temporais.

2.2.1 *Machine Learning* na previsão de séries temporais

Sabe-se que há muito tempo, o campo da previsão de séries temporais tem sido influenciado pelos mais diversos modelos desenvolvidos na literatura. Por exemplo, em meados dos anos setenta, foram introduzidos os modelos mais frequentemente aplicados: modelos ARIMA, propostos por Box et al. (1970). Porém, com a crescente disponibilidade de grandes quantidades de variáveis e informações tornou-se fundamental que técnicas modernas e robustas fossem propostas e utilizadas de maneira a explorarem ao máximo essa informação. Em vista disso, especialmente na última década, os métodos de *Machine Learning* vem ganhando cada vez mais destaque, estabelecendo-se como fortes candidatos no domínio da previsão, principalmente quando comparados aos modelos clássicos (Ahmed et al., 2010; Lippi et al., 2013).

Os estudos na área de aprendizado de máquina tiveram início com o desenvolvimento do modelo de Rede Neural Artificial (RNA), apresentado por Fitch (1944). Por exemplo, Werbos e John (1974) mostram que as RNA superam os métodos estatísticos clássicos, como as abordagens de Box et al. (1970). Posteriormente, os conceitos iniciais foram estendidos a outras técnicas, que passaram a ser utilizadas nas mais diversas áreas. Um exemplo é mostrado em Zakaria e Shabri (2012), onde usam do algoritmo de *Support Vector Machine*, SVM (Cortes e Vapnik, 1995), para previsão do fluxo de água em locais não medidos.

Nos dias atuais, nos deparamos com uma ampla gama de métodos de *Machine Learning* e, por isso, selecionar a ferramenta mais adequada para prever uma série pode ser difícil. Nas próximas seções, os métodos específicos a serem empregados neste trabalho serão explicados.

2.3 Algoritmos

As seções seguintes fornecem as definições e a compreensão básica dos algoritmos referentes aos métodos selecionados para serem implementados no escopo do trabalho. Neste estudo, o foco central são os algoritmos de *Machine Learning* supervisionados utilizados para modelagem preditiva de respostas quantitativas.

2.3.1 Regularização

Um dos focos principais nesta pesquisa é o estudo de métodos de regularização (também conhecidos como de encolhimento, do inglês *Shrinkage*). Utilizamos os métodos LASSO e adaLASSO com a finalidade de reduzir a dimensionalidade do espaço paramétrico e atingir melhores previsões. Os métodos de regularização que

estaremos utilizando neste trabalho destacam-se pela interpretabilidade, enquanto outras abordagens que serão empregadas, pela flexibilidade.

Segundo James et al. (2013), a abordagem dos métodos de regularização baseia-se em um modelo linear, que pode ser descrito em sua forma vetorial como

$$y_i = \beta_0 + X_i^T \beta + \epsilon_i, \tag{2.5}$$

onde $y_i \in \mathbb{R}$ é a variável resposta, $X_i = (x_{1i}, ..., x_{ki})^T \in \mathbb{R}^k$ é o conjunto de preditores, $\beta = (\beta_1, ..., \beta_k)^T$ é o conjunto de parâmetros e β_0 é uma constante.

O método de mínimos quadrados ordinários (MQO) é baseado na minimização da soma dos quadrados dos resíduos (SQR):

$$\hat{\beta} = \underset{\beta_0,\beta_1,...,\beta_k}{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ji}\beta_j \right)^2. \tag{2.6}$$

Os métodos de regularização voltam-se para o fato de que os coeficientes são estimados usufruindo de um procedimento de ajuste alternativo. Como as estimativas de MQO podem prejudicar a precisão das previsões, por possuírem frequentemente grande variância (apesar do baixo viés), encolher o conjunto de coeficientes, permitindo um pouco de viés para reduzir a variância das previsões, tende a auxiliar nesta precisão (Friedman et al., 2001).

Os métodos de encolhimento ajustam um modelo envolvendo todos os k preditores. No entanto, é usada uma técnica que restringe as estimativas dos coeficientes, ou ainda, que reduz estimativas dos coeficientes (para preditores redundantes) em direção a zero. Dependendo do tipo de encolhimento realizado, alguns dos coeficientes podem ser estimados em exatamente zero (no método LASSO, por exemplo). Então, eles também podem executar seleção de variáveis. Portanto, tais métodos constituem uma alternativa satisfatória ao estimar parâmetros em grandes dimensões.

Least Absolute Shrinkage and Selection Operator (LASSO)

O *Least Absolute Shrinkage and Selection Operator* (LASSO) foi proposto por Tibshirani (1996) e é uma das técnicas mais conhecidas de regularização. O objetivo do método é estimar um modelo capaz de gerar previsões com pequena variância e ainda, que possa determinar o conjunto de preditores que melhor expliquem a variável resposta, reduzindo a zero os parâmetros correspondentes às variáveis irrelevantes. Efron et al. (2004) demonstraram que, sob algumas condições, o LASSO pode manipular mais variáveis do que observações.

As estimativas LASSO são obtidas através da minimização dos quadrados dos resíduos impondo penalização na soma dos valores absolutos do conjunto de coeficientes, com isso, é definido como:

$$\hat{\beta}^{LASSO} = \underset{\beta_0,\beta_1,...,\beta_k}{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ji}\beta_j \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\}, \tag{2.7}$$

onde $\lambda \geq 0$ controla a severidade da penalização e é determinado por técnicas baseadas em dados, como validação cruzada (do inglês *cross-validation*) ou o uso de critérios de informação. Quando $\lambda = 0$, as estimativas LASSO serão iguais às de

Mínimos Quadrados Ordinários (MQO), que é provavelmente o método mais popular para estimação de parâmetros.

De acordo com Konzen e Ziegelmann (2016), o valor do parâmetro de ajuste λ na Eq. (2.7) é tradicionalmente escolhido por meio de validação cruzada em uma estrutura *cross-section*. No entanto, em uma configuração de série temporal, escolher o parâmetro λ usando o Critério de Informação Bayesiano (BIC) é mais adequado. Além disso, selecionar o modelo através desse critério de informação é mais rápido do que usar validação cruzada e tem algumas vantagens teóricas em alguns casos. Por exemplo, Zou et al. (2007) mostram que é possível estimar de forma consistente os graus de liberdade do LASSO usando o BIC.

Também, vale ressaltar que, no contexto de séries temporais, em especial, a penalização LASSO é aplicada em alguns estudos, como em Li e Chen (2014) e Garcia et al. (2017).

***Adaptive* LASSO (adaLASSO)**

O *Adaptive* LASSO foi proposto por Zou (2006), após verificar que pode haver situações nas quais o LASSO não é consistente na seleção de variáveis. A abordagem do adaLASSO emprega diferentes pesos para diferentes coeficientes e seu estimador é definido como:

$$\hat{\beta}^{adaLASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ji}\beta_j \right)^2 + \lambda \sum_{j=1}^{k} \omega_j |\beta_j| \right\}, \tag{2.8}$$

onde $\omega_j = |\hat{\beta}_j^*|^{-\tau}$, que, por sua vez, é o coeficiente de uma estimativa de primeiro passo e $\tau > 0$, visto que se $\tau = 0$ temos o método LASSO tradicional. O valor mais comumente usado para o parâmetro τ é 1.

Os pesos individuais ω_j auxiliam na seleção das variáveis relevantes, assim, penalizam-se mais os coeficientes das variáveis que aparentam ser pouco importantes. Isso porque uma variável relevante x_j tende a possuir um coeficiente $\hat{\beta}_j^*$ grande, o que faz diminuir o peso ω_j atribuído ao coeficiente daquela variável e, pelo outro lado, se a variável x_j for irrelevante, o coeficiente $\hat{\beta}_j^*$ tende a ser pequeno e implicará em um peso ω_j grande.

Zou (2006) sugere utilizar $\hat{\beta}^{MQO}$ como $\hat{\beta}^*$, a menos que a colinearidade seja uma preocupação. Nesse caso, recomenda $\hat{\beta}^{Ridge}$ para o melhor ajuste da regressão *Ridge* - que é um método que encolhe o conjunto de coeficientes ao impor uma penalização na soma dos quadrados dos mesmos -, por ser mais estável que por MQO.

2.3.2 Métodos *Ensemble*

Os métodos chamados de *Ensembles* combinam decisões para melhorar o desempenho do sistema como um todo. Estudos empíricos e teóricos mostraram que ambos problemas de classificação e regressão apresentam, frequentemente, melhor desempenho preditivo em comparação com um único modelo (Bauer e Kohavi, 1999; Kleinberg, 2000), evidentemente isso ocorre às custas da interpretabilidade dos modelos, por isso, são mais flexíveis. No contexto de séries temporais, o uso de determinados algoritmos desta classe cresce cada vez mais (Hillebrand e Medeiros, 2010; Dantas et al., 2017). Abaixo, estudaremos o método de *Random Forests*.

Para introduzir seu conceito, é importante exemplificarmos uma árvore de decisão, algoritmos que podem ser aplicados tanto em problemas de regressão (resultados contínuos) quanto de classificação. Uma árvore de decisão consiste na estratificação ou segmentação do espaço do preditor em várias regiões simples. A fim de prever uma dada observação, usualmente é utilizada a média ou a moda das observações de treinamento na região a que pertence (James et al., 2013). Um exemplo com cinco regiões da abordagem de árvores de decisão é mostrado na figura 2.1.

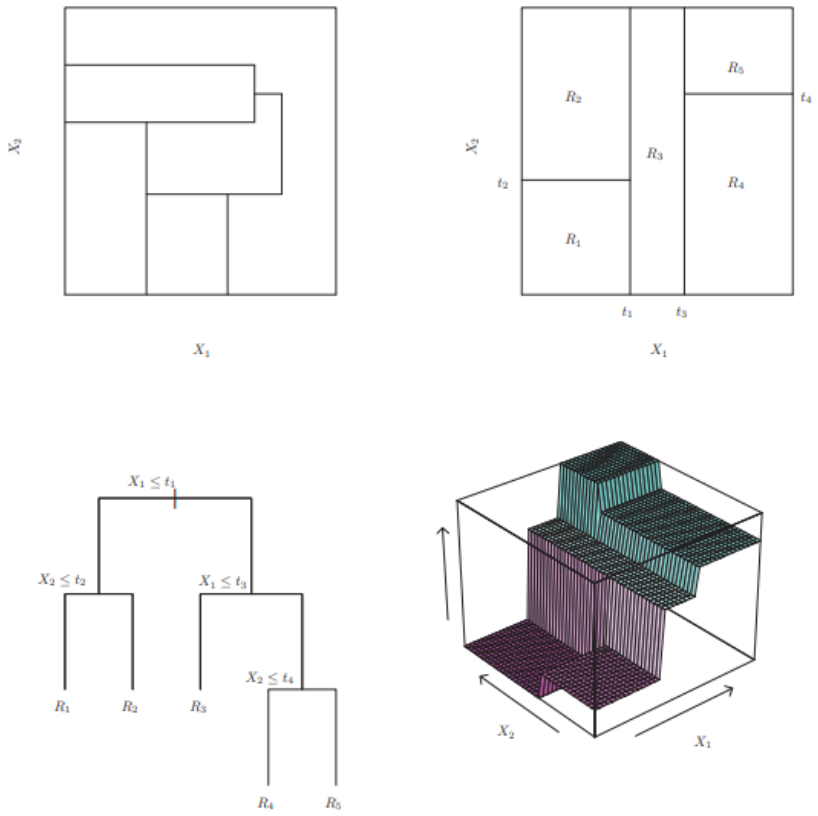


Figura 2.1: Representações gráficas de uma árvore de decisão
Superior esquerdo: uma partição de espaço de recurso bidimensional que não pode resultar da divisão binária recursiva. Superior direito: a saída da divisão binária recursiva em um exemplo bidimensional. Inferior esquerdo: uma árvore que corresponde à partição no canto superior direito. Inferior direito: um gráfico em perspectiva da superfície de previsão correspondente a essa árvore. **Fonte:** James et al. (2013).

Em outras palavras, uma árvore de decisão normalmente inicia com um único nó, que é dividido em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades até a finalização da árvore. Apesar da árvore de decisão ter tendência de ajustar-se bem aos dados, não é muito eficaz para fazer previsões. Para contornar essa situação é que será usada a abordagem *Ensemble*.

***Random Forests* (RF)**

O método de *Random Forests*, introduzido por Breiman (2001), pode ser interpretado como uma coleção de árvores de decisão combinadas, em que cada uma delas é treinada em uma fração distinta do conjunto de dados, fazendo com que

sejam diferentes entre si. Então, uma vez que as árvores tenham sido treinadas, o método combina a saída das árvores de decisão individuais para gerar a saída final. O algoritmo de *Random Forests* é capaz de entender quais variáveis são mais importantes durante o treinamento.

Na etapa de construção das árvores de decisão, quando os nós são divididos no algoritmo de *Random Forests*, cada árvore faz a divisão usufruindo um subconjunto de m preditores, que são variáveis escolhidas aleatoriamente do conjunto completo de p preditores, o que acaba fornecendo mais informações básicas para a previsão de saída (Liaw et al., 2002). Usualmente, o número de preditores considerados em cada divisão para problemas de regressão é $p/3$, ou seja, aproximadamente igual à um terço do número total de preditores. Nesta etapa é utilizado o *bootstrap* (Efron, 1979), que é um método de reamostragem onde as amostras selecionadas podem ser repetidas na seleção.

Os principais parâmetros ajustáveis deste método são o número de árvores de decisão a serem treinadas, que aumenta a complexidade e a capacidade preditiva do algoritmo; e o método de divisão, que é a forma como a divisão do nó é aplicada. Além disso, nesse método, o peso de cada árvore é igual (a $1/B$, em que B é o número de árvores).

Em séries temporais, a técnica de *Random Forests* é implementada, por exemplo, em Garcia et al. (2017) e, também, em Kane et al. (2014).

2.3.3 Modelos *Benchmark*

Uma vez que este trabalho apresenta uma breve análise comparativa de métodos de aprendizado de máquina com alternativas como os modelos ARIMA para prever o surto de Covid-19, forneceremos uma introdução básica a essa técnica.

Modelo ARIMA

O modelo autorregressivo integrado de médias móveis (ARIMA), introduzido por Box et al. (1970), é uma família de equações que descrevem uma série temporal dependente de seus valores anteriores, isto é, de suas próprias defasagens ou/e termos de erro do modelo (normalmente definido com um ruído branco ou os choques do modelo), de maneira que a equação possa ser usada para a estimação de valores futuros.

Em outras palavras, podemos especificá-lo pela ordem (número de defasagens) do modelo (p), grau de diferenciação (número de vezes em que os dados tiveram valores passados subtraídos (d) e ordem do modelo de média móvel (q). Assim, é possível descrevê-lo como ARIMA (p, d, q) e a equação do mesmo define-se por

$$\begin{aligned} Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \\ \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_q \epsilon_{t-q} + \epsilon_t, \end{aligned} \tag{2.9}$$

em que Y_t é a série diferenciada, β_1 é o coeficiente do primeiro termo AR, p é a ordem do termo AR, ϕ_1 é o coeficiente do primeiro termo MA, q é a ordem do termo MA e ϵ_t é o erro.

Em trabalhos da recente pandemia, observamos, por exemplo, Dehesh et al. (2020) empregando o modelo ARIMA para estimar a propagação do vírus em países como China, Itália, Coréia do Sul, Irã e Tailândia. Além disso, Tandon et al. (2020)

averiguam a aplicação do mesmo para a previsão do número de casos confirmados na Índia e Ribeiro et al. (2020) para o número acumulado de infectados no Brasil.

2.3.4 *Model Confidence Set*: comparação das previsões

De acordo com (Hansen et al., 2011), o *Model Confidence Set* (MCS) é um procedimento que permite comparar vários modelos (de uma coleção M^0) simultaneamente, considerando suas habilidades preditivas e, após, identificar um subconjunto M^*, que seja superior em termos de previsão, dado um nível de confiança.

A técnica baseia-se em um teste de equivalência (δ_M) e em uma regra de eliminação (e_M). O teste de equivalência é aplicado a cada etapa e, no primeiro passo, assume-se como hipótese nula que todos os modelos em M_0 são iguais em termos de capacidade de predição. Caso rejeitemos essa hipótese, a regra de eliminação remove o modelo com pior performance. O processo é repetido até que no subconjunto de modelos restante, o teste de equivalência não seja rejeitado. É importante citar que o procedimento MCS é implementado através de um procedimento de *bootstrap*, em que a estatística de teste é construída a partir da estatística t.

O resultado do MCS irá conter apenas alguns modelos (ou o melhor) quando dados informativos estiverem disponíveis; caso contrário, iremos obter um resultado com muitos (ou todos) modelos.

2.4 Modelagem da pandemia

As próximas seções são a respeito especificamente do uso de métodos de *Machine Learning* para modelar a pandemia e algumas possibilidades e desafios que os mesmos trazem.

2.4.1 O uso de *Machine Learning*

A utilização de algoritmos de *Machine Learning* retrata um tema muito atual para a área da saúde (Chiavegatto Filho, 2015). O grande progresso em pesquisas deste domínio, principalmente quando nos referimos à modelagem preditiva, ocorre associado ao aumento da demanda por métodos mais qualificados, acessíveis, ágeis e efetivos para pacientes, médicos e organizações.

No último ano, particularmente, atentamos ao amplo destaque que os métodos de *Machine Learning* vêm ganhando nos mais diversos meios de comunicação. Este relevante papel deve-se ao fato de que pesquisadores das mais diversas áreas passaram a usufruir dessa tecnologia na luta contra a pandemia de Coronavírus por todo o mundo (Medeiros et al., 2020; Vaishya et al., 2020). Com esse tipo de métodos na previsão de tendências da pandemia, é possível avaliar a eficácia de intervenções e direcionar medidas adequadas para reduzir o contágio.

Os resultados preditivos do uso de técnicas de aprendizado de máquina aplicadas em séries temporais, assim como a coleção de variáveis explicativas, diferem dependendo da variável usada como resposta, visto que neste trabalho serão feitas previsões de números de novos casos e mortes diários de Coronavírus em cidades brasileiras. Desse modo, é necessário introduzir o conjunto de características (variáveis independentes) utilizadas na previsão: temperatura, qualidade e umidade do

ar e buscas no Google de palavras relacionadas à Covid-19, que podem ser encaradas como covariáveis. O conjunto de dados a ser aplicado combina informações do próprio passado da série temporal com essas covariáveis.

Muitos aspectos específicos da propagação da doença, como número de indivíduos recuperados, efeitos sazonais e o tempo entre a fase aguda da doença e a morte, não foram considerados. Isso porque, para a maioria das cidades, não existem dados completos relativos a essas variáveis.

A observação do comportamento da Covid-19 em muitos países indica incerteza e complexidade (Zhong et al., 2020). Nesse contexto, temos modelos tradicionais que devem ser adaptados às características específicas de cada localidade avaliada, como a vulnerabilidade à infecção em virtude de alterações nas intervenções de saúde pública, para que possam proporcionar resultados mais confiáveis (Cooper et al., 2020). Isso estabelece um limite na capacidade de generalização e robustez de tais modelos padrões.

Portanto, o crescente interesse em técnicas com maior capacidade de generalização e predição, fez com que o aprendizado de máquina começasse a se destacar na previsão da recente pandemia, obtendo resultados consideráveis quando comparados aos métodos tradicionais (Ardabili et al., 2020). Embora a abordagem de *Machine Learning* tenha sido empregada na modelagem de surtos ou pandemias anteriores, como na do Zika vírus (Carlson et al., 2018) e da influenza H1N1 (Yin et al., 2018), há menos estudos na literatura dedicados a Covid-19 e, consequentemente, a contribuição do trabalho consiste em explorar estes métodos para previsão da evolução da pandemia.

3 Análise Numérica

Nas próximas seções será apresentada a análise numérica usada para o desenvolvimento desse trabalho. A estruturação de toda parte aplicada, que contempla análises e cálculos, será realizada através do *software* livre R (R Core Team, 2019) sob a interface do RStudio (2020). A versão do R utilizada foi a 4.0.3.

Assim, com a finalidade de avaliar o desempenho de previsão dos métodos de aprendizado de máquina no contexto de séries temporais, realizamos, inicialmente, um exercício numérico com simulações de Monte Carlo, que será essencial para entender o funcionamento de cada método em um cenário conhecido, onde possamos medir com maior precisão qual desempenha melhor, dado que observaremos uma resposta sem erros causados por fatores externos, por exemplo. O estudo contempla, ainda, uma análise de dados empíricos relacionados a Covid-19 e, apesar da estrutura (e características) de dados ser diferente do que será observado na simulação, o procedimento de implementação é similar, por utilizarmos como conjunto de variáveis independentes as defasagens das covariáveis e, também, da própria variável resposta.

O processo de estimação das equações dos métodos de regularização contou com o auxílio da função `ic.glmnet` do pacote `HDeconometrics` para otimização dos parâmetros β_j e λ, que foram selecionados de acordo com o BIC. Para o adaLASSO, empregamos a regressão Ridge no primeiro passo para computar os pesos ω_j da variável x_j. No caso de *Random Forest*, o processo de estimação foi feito através da função `randomForest`, com número de árvores $B = 500$ e o número mínimo de observações em cada folha igual a 15, ou seja, parâmetros fixos.

As previsões de y_t deste capítulo foram comparadas através das medidas de Raiz Quadrada do Erro Quadrático Médio (rEQM) e Erro Absoluto Médio (EAM), de modo a avaliar o quanto o valor predito ($\hat{y}_t$) para a resposta de uma observação se aproxima de seu valor observado (y_t). As definições dos erros são representadas por:

$$rEQM = \sqrt{\frac{1}{n_r\ T_0}\sum_{i=1}^{n_r}\sum_{t=1}^{T_0}(\hat{y}_{it} - y_{it})^2}, \tag{3.1}$$

$$EAM = \frac{1}{n_r\ T_0}\sum_{i=1}^{n_r}\sum_{t=1}^{T_0}|\hat{y}_{it} - y_{it}|. \tag{3.2}$$

Em ambos algoritmos apresentados neste capítulo, tanto na seção 3.1 quanto na 3.2, as últimas 30 observações das séries foram retiradas (não foram usadas para a estimação) para avaliar as previsões fora da amostra 1 passo à frente. Ou seja, nesse caso, a equação de previsão (só 1 passo a frente) dado o conjunto de informação até

o instante anterior $(t-1)$ varia de modelo pra modelo. Especialmente para a análise empírica, realizamos também a previsão 7 passos à frente, sendo que nesse caso, a equação de estimação usa informações até o instante $t-7$.

Assim, para as previsões foi utilizado um esquema em expansão, ou seja, conforme foram surgindo novas observações, mantivemos as primeiras, então o tamanho de amostra aumentou a cada nova previsão.

3.1 Simulação

Assim como implementado em Konzen e Ziegelmann (2016), nesta seção estamos empregando o método de Monte Carlo com 500 replicações, em que simulamos 10 séries temporais independentes. Todas seguem um processo AR(1) da forma $x_{i,t} = 0.6x_{i,t-1} + u_{i,t}$, onde $u_{i,t} \sim N(0,1)$, $i = 1, \ldots, 10$. Assim, consideramos o seguinte processo gerador de dados:

$$\begin{aligned} y_t = 0.8y_{t-1} &+ 0.6x_{1,t-1} + 0.3x_{1,t-2} - 0.5x_{2,t-1} - 0.2x_{2,t-2} + 0.4x_{3,t-2} \\ &+ 0.4x_{4,t-1} - 0.3x_{5,t-1} + 0.2x_{6,t-1} + \epsilon_t, \quad t = 1, 2, \ldots, T, \end{aligned}$$

onde $\epsilon_t \sim N(0,1)$.

Os métodos apresentados na seção 2.3 foram implementados, usufruindo-se 10 defasagens de y e 10 de x_j, $j = 1, \ldots, 10$, totalizando 110 preditores candidatos. Ainda, comparamos o desempenho dos métodos em três diferentes tamanhos da série, $T = \{50, 500, 1000\}$.

Na tabela 3.1 estão os resultados dos erros das previsões 1 passo à frente, em que apresentamos os valores médios de rEQM e EAM entre as 500 replicações de Monte Carlo e, em negrito, estão os menores valores desses erros para cada tamanho de amostra.

Para o menor tamanho de amostra ($T = 50$), notamos uma superioridade do LASSO, isto é, o menor erro de previsão para uma série temporal. Analisando o caso intermediário, com $T = 500$, percebemos que o adaLASSO produziu uma previsão ligeiramente mais precisa. Por fim, para o maior tamanho de amostra ($T = 1000$), os métodos LASSO e adaLASSO possuem erros de previsão muito semelhantes.

Tabela 3.1: Erros de previsão 1 passo à frente

T	**50**	**500**	**1000**
Médias dos rEQMs			
LASSO	**2,227**	1,033	1,014
adaLASSO	2,336	**1,006**	**0,998**
Random Forest	3,814	2,288	2,047
Médias dos EAMs			
LASSO	**1,779**	0,828	0,815
adaLasso	1,854	**0,808**	**0,802**
Random Forest	3,140	1,849	1,654

Nota: Os valores em negrito indicam o método com os menores valores de rEQM e EAM para cada tamanho de amostra.

Observamos ainda que, em todos os casos, LASSO e adaLASSO superam o método *Random Forest* neste contexto trabalhado. Esse resultado é coerente, uma vez

que o modelo utilizado é de fato um modelo linear. De modo geral, métodos baseados em um modelo linear têm um melhor desempenho quando o verdadeiro modelo é linear. E espera-se que métodos baseados em modelos não lineares tenham um melhor desempenho para quando o verdadeiro modelo é não linear, se os parâmetros forem escolhidos através de um procedimento adequado de validação.

Os boxplots dos erros de previsão são exibidos na Figura 3.1. Os gráficos dos métodos LASSO e adaLASSO, para ambas as medidas de erro, possuem a forma parecida, com pequena variação, enquanto o *Random Forest* dispõem de uma variabilidade maior. Os gráficos corroboram todos os resultados discutidos na tabela 3.1.

3.2 Análise Empírica

Nesta seção, o objetivo é comparar os métodos já definidos na aplicação de dados epidemiológicos, mais especificamente, da recente pandemia.

3.2.1 Banco de Dados

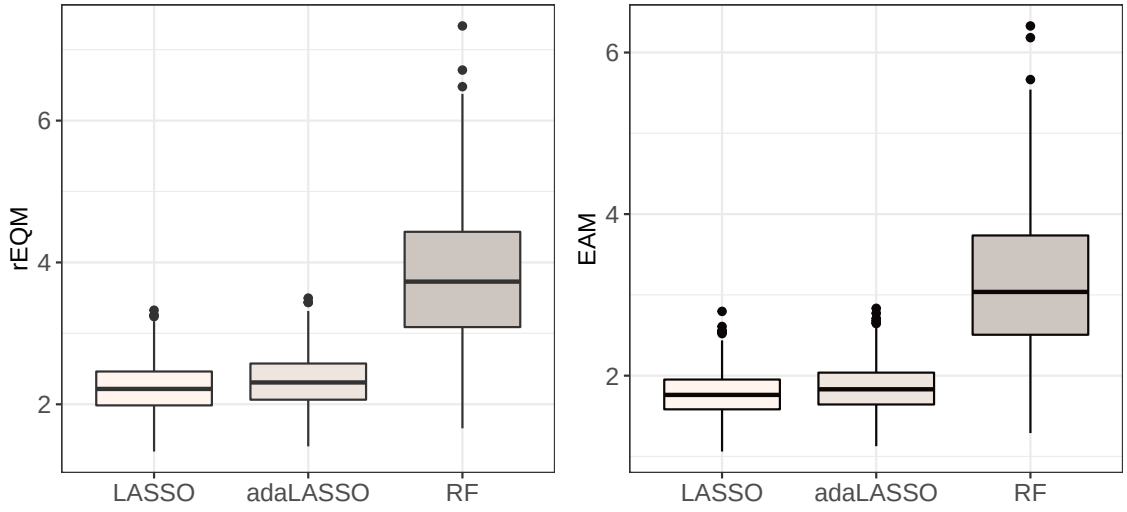
Uma etapa fundamental é a preparação do banco de dados para o uso dos métodos. Os dados da aplicação foram obtidos por meio de um *site* conceituado, referência no monitoramento da Covid-19 no Brasil: o Brasil.io, que compila boletins epidemiológicos das 27 Secretarias Estaduais de Saúde e disponibiliza uma base de dados com a série histórica de casos e óbitos confirmados diariamente por município, que serão justamente as variáveis a serem incorporadas neste projeto referentes à evolução da pandemia.

Além disso, diferentes fatores externos influenciam os grandes números da pandemia. Em especial, nesse estudo, usaremos temperatura, qualidade e umidade do ar e buscas no Google de palavras relacionadas à pandemia, componentes que podem ser encarados como covariáveis. O conjunto de dados a ser aplicado nesse estudo combina informações do próprio passado da série temporal com essas covariáveis, que podem ser úteis para fazer inferência.

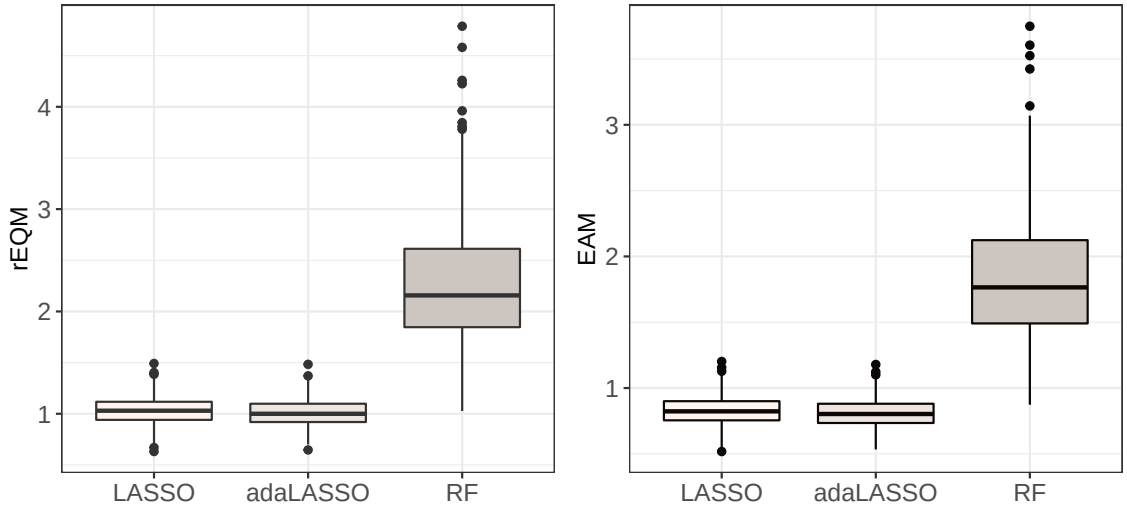
As variáveis ambientais foram coletadas através do *site* do Instituto Nacional de Meteorologia, órgão do Ministério da Agricultura, Pecuária e Abastecimento, que fornece diversas informações meteorológicas à sociedade brasileira.

Ainda, conforme resultados mostrados em (Oliveira, 2021), que comprovam que as covariáveis de buscas no Google possuem uma boa capacidade de predição dos casos de Covid-19 no estado do RS, utilizamo-as em nossas análises. Dessa forma, através do pacote `gtrendsR`, captamos as buscas feitas no Google por palavras relacionadas a automedicação para a Covid-19 e ao interesse em locais e eventos que potencialmente são mais propensos a transmissão do vírus - como bares, carnaval, etc - nas cidades de interesse.

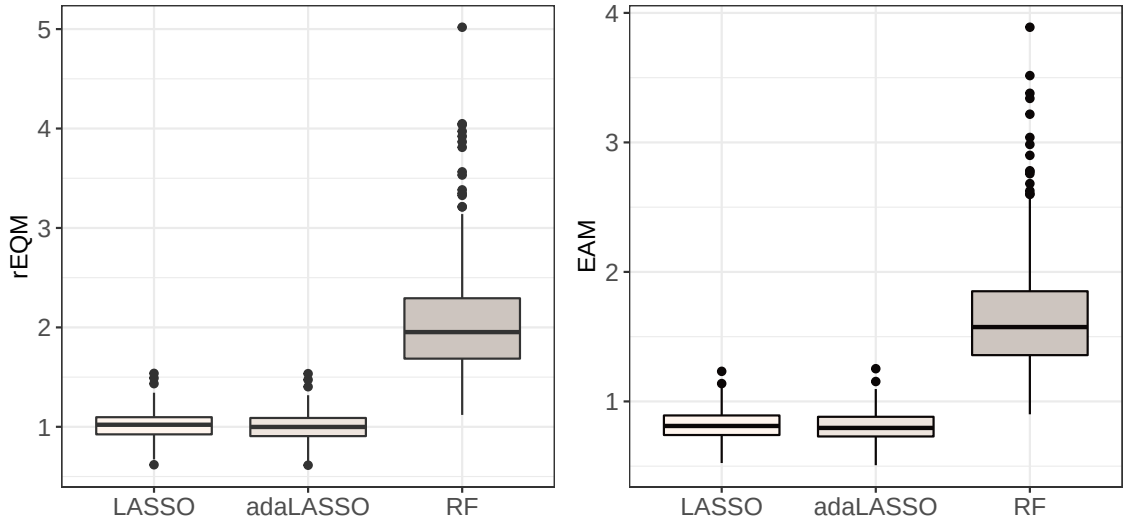
Cabe destacar que o foco do trabalho é nas três cidades com maior número de mortes e casos de Covid-19 no Brasil, segundo dados do Brasil.io: São Paulo, Rio de Janeiro e Brasília. Os dados coletados de todas variáveis utilizadas neste estudo são diários e respectivos ao período compreendido entre o primeiro caso confirmado de cada cidade (25/02/20 em São Paulo, 06/03/20 no Rio de Janeiro e 07/03/20 em Brasília) e 31 de agosto de 2021.



(a) T = 50



(b) T = 500



(c) T = 1000

Figura 3.1: Boxplots dos erros de previsão por método estudado

3.2.2 Critérios para comparações e implementação no R

Na modelagem dos dados empíricos, foram estimados todos os métodos e modelos apresentados na seção 2.3 bem como seus erros de previsão, que são apresentados na próxima subseção. Além dos erros das equações 3.1 e 3.2 analisados, avaliamos o Erro Relativo Médio (ERM) para a aplicação, visto que assim poderemos compreender se os métodos e modelos desempenharam melhor para a previsão de mortes ou de casos da Covid-19. A definição do ERM é dada por:

$$ERM = \frac{1}{n_r\ T_0} \sum_{i=1}^{n_r} \sum_{t=1}^{T_0} \left(\frac{|\hat{y}_{it} - y_{it}|}{y_{it}} \right). \tag{3.3}$$

Aplicamos o filtro de médias móveis em cada uma das séries trabalhadas, ou seja, ao invés de contabilizar apenas os casos registrados nas últimas 24 horas, estaremos somando os dados mais recentes com os dos seis dias anteriores e dividindo o resultado por sete, de modo a retirar qualquer oscilação que possa vira atrapalhar na análise.

Conforme apresentado na seção 2.2, para entender se as séries temporais estudadas estavam no âmbito estacionário ou não, utilizamos o `adf.test` do pacote `tseries` do R. Desse modo, se aceitarmos a hipótese nula de que cada respectiva série é não estacionária, optaremos por tirar a diferença da mesma e assim, finalmente, aplicar nosso métodos. Vale lembrar que a previsão está sendo realizada para as médias móveis de mortes e casos de Covid-19 e cada uma das variáveis especificadas na subseção 3.2.1 será utilizada como covariável. Especificamente na análise empírica, o número de defasagens usado para cada variável é 14.

No caso do modelo *benchmark* ARIMA, as covariáveis utilizadas são apenas as defasagens da própria variável resposta. Utilizamos o algoritmo de previsão automática do ARIMA, `auto.arima`, para selecionar um modelo adequado através da escolha da ordem p, d, q, efetuada através de testes de raiz unitária e o critério de AIC (*Akaike Information Criterion*), sendo escolhido a ordem que adotar o menor valor para esse critério.

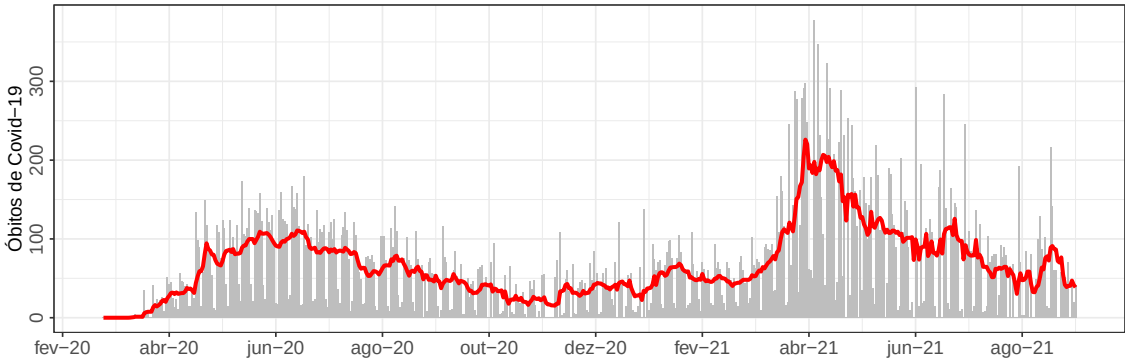
Como dito anteriormente, separamos uma parte da amostra para estimação e analisamos o desempenho das previsões 1 passo a frente fora dessa amostra. Para os dados empíricos estaremos implementando, também, a previsão 7 passos à frente para cada uma das variáveis resposta (MM de mortes e de casos), bem como para cada cidade e método. Realizamos as previsões 1 passo à frente nessa etapa mais como um exercício de previsão, dado que acreditamos que tomando 7 passos à frente tenhamos resultados mais úteis para esse tipo de problema real.

Por fim, na análise empírica, para testar se as previsões produzidas pelo conjunto de métodos são significativamente diferentes do *benchmark* e também quais são os melhores métodos, consideramos a abordagem de *Model Confidence Set* (MCS), conforme exemplificado na subseção 2.3.4. Para realizar esse teste, utilizamos a função `MCSprocedure`.

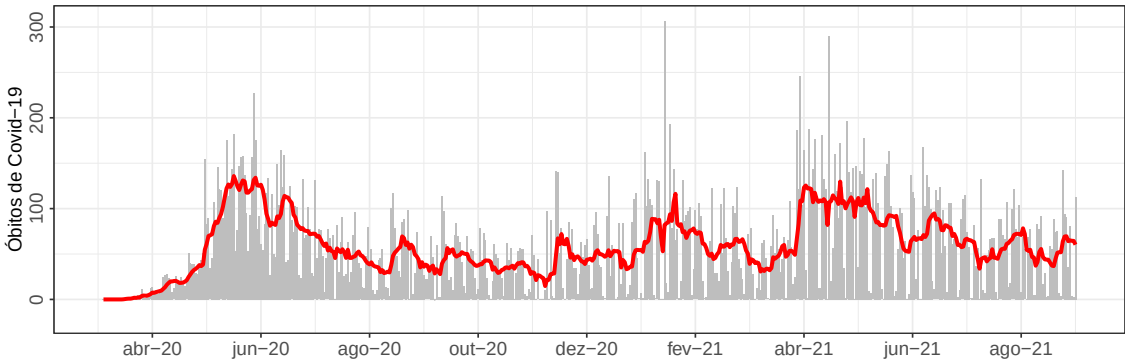
3.2.3 Resultados

Os dados usados na aplicação e as características inerentes a eles já foram elucidadas ao longo desse trabalho. As figuras 3.2 e 3.3 podem ser observadas para entendermos melhor o comportamento das séries temporais relacionadas à evolução de óbitos e casos de Covid-19 em cada cidade estudada. A linha vermelha dos

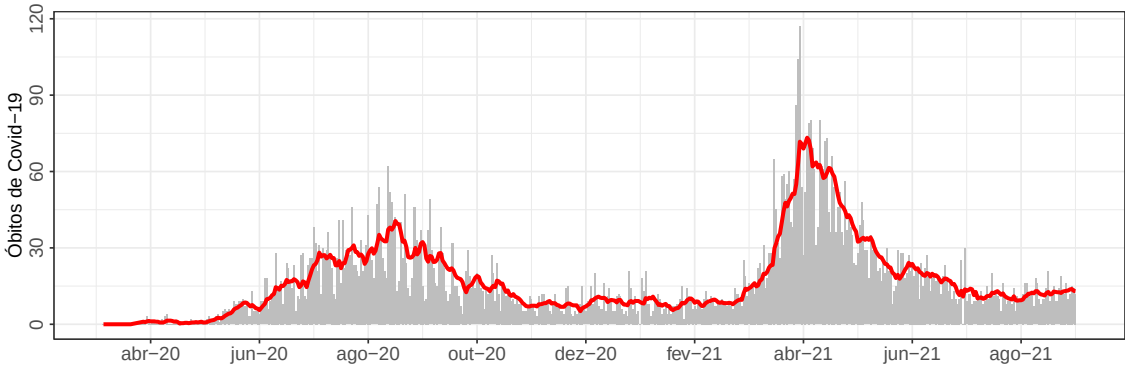
gráficos representa as médias móveis e, em cinza, temos o número de casos/mortes brutos diários em São Paulo, Rio de Janeiro e Brasília, respectivamente.



(a) São Paulo



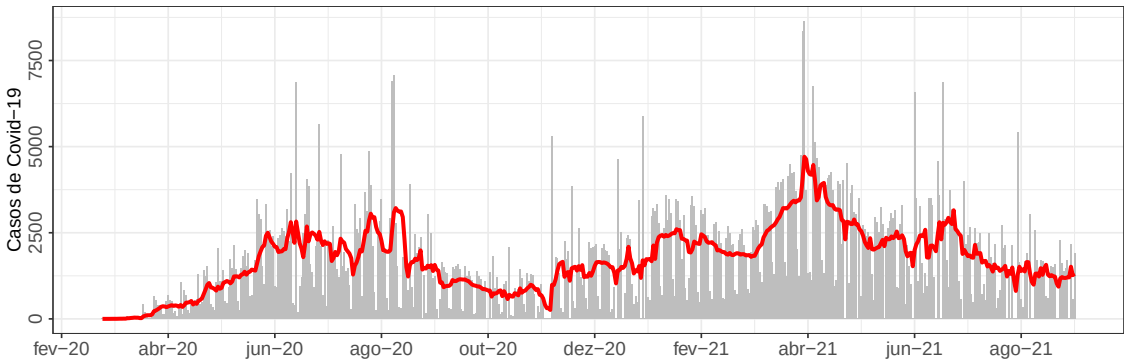
(b) Rio de Janeiro



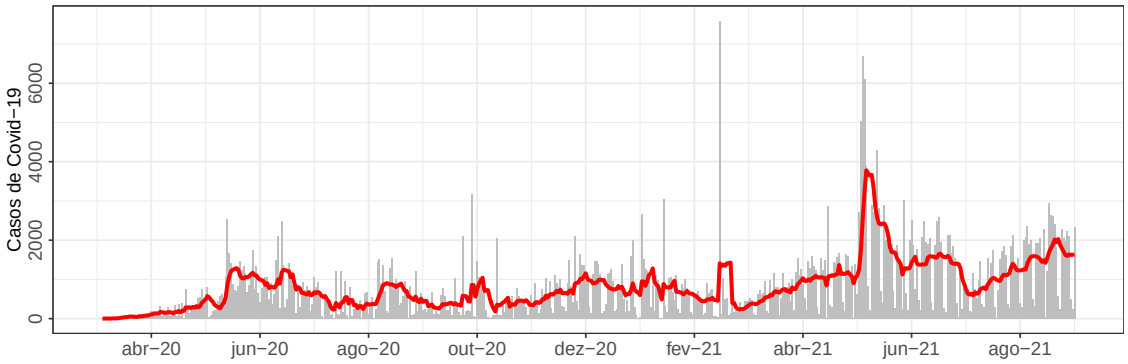
(c) Brasília

Figura 3.2: Gráficos das Médias Móveis de Mortes por cidade brasileira
A linha vermelha representa as médias móveis e em cinza temos as mortes brutas (diárias) observadas.

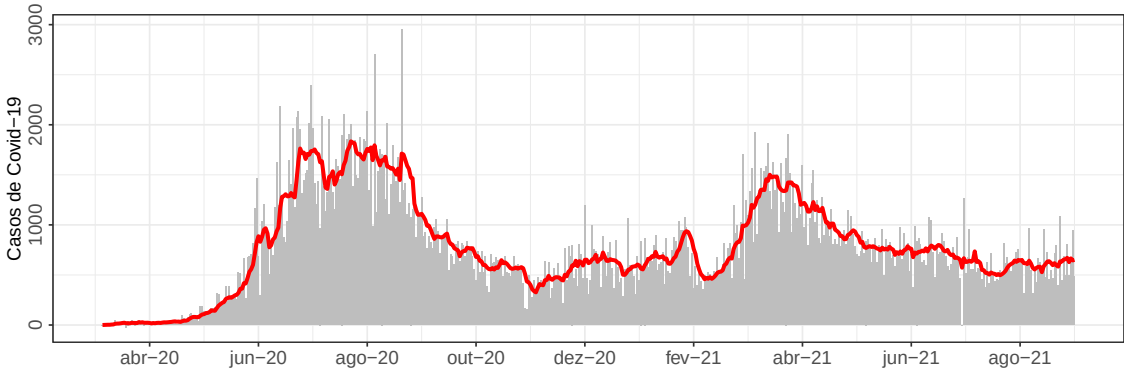
Observando esses gráficos é fácil perceber que apesar dos números serem contabilizados diariamente, ocorrem muitas oscilações entre dias de semana e fins de semana, em virtude da irregularidade no processo de aquisição desses dados. A



(a) São Paulo



(b) Rio de Janeiro



(c) Brasília

Figura 3.3: Gráficos das Médias Móveis de Casos por cidade brasileira
A linha vermelha representa as médias móveis e em cinza temos os casos brutos (diários) observados.

queda de registros nos fins de semana pode ser justificada pela redução de pessoal disponível para preenchê-los ou por fechamento de unidades de saúde ou setores responsáveis por informar os dados às autoridades sanitárias. As planilhas atualizadas no início da semana vão conter, portanto, não apenas com os dados daquele dia, mas todos que aguardavam para serem preenchidos. Com a utilização de médias

móveis, suavizamos esse comportamento que pode atrapalhar a análise, conforme fica perceptível com a linha vermelha.

A fim de avaliar se as variáveis resposta utilizadas seguem um processo estocástico estacionário foi realizado um teste de raiz unitária. Assim, a tabela 3.2 traz os p-valores do teste ADF e conseguimos notar que, exceto a série de médias móveis de casos do Rio de Janeiro, todas as outras aceitam a hipótese nula de não-estacionariedade a um nível de significância de 5%. Desse modo, só não iremos diferenciar a série que, segundo o resultado do teste, traz evidências de estacionariedade. Após diferenciar as séries que apresentaram não estacionariedade através do teste ADF, foi feito o teste novamente na série resultante de modo a garantir sua estacionariedade.

Tabela 3.2: p-valor do Teste de Dickey-Fuller aumentado para as variáveis respostas utilizadas na modelagem preditiva

	São Paulo	**Rio de Janeiro**	**Brasília**
Teste ADF			
MM Mortes	0,694	0,332	0,553
MM Casos	0,688	0,025	0,620

É essencial deixar explícito que também aplicamos o teste ADF em cada covariável utilizada e como tivemos evidências de estacionariedade em todas, adicionamos apenas o resultado do p-valor do teste para as variáveis resposta.

Os gráficos dispostos nas figuras 3.4 e 3.5 correspondem às previsões 1 e 7 passos à frente para o número de óbitos por Covid-19 em cada cidade estudada, respectivamente. A linha vermelha representa os valores preditos para cada método e a preta os valores observados do período mostrado. As previsões de cada horizonte foram executadas para as últimas 30 observações presentes no banco de dados coletado, de 02 a 31 de agosto de 2021. Visualmente e de maneira geral, os gráficos conseguem seguir as trajetórias dos dados observados, isto é, prever "adequadamente"as curvas verdadeiras de mortes. Os resultados mais precisos dessas previsões podem ser vistos na tabela 3.3.

A tabela 3.3 está disposta com os erros de previsão estudados para a MM de mortes. Os menores erros (rEQM e EAM) são indicados em negrito para cada horizonte de previsão e as células em cinza e azul indicam que o método está incluído no MCS de 50% usando a raiz do erro quadrado e o erro absoluto como funções de perda.

Assim, observa-se que os métodos LASSO e *Random Forest* realizaram previsões com performance semelhante para todas as cidades. O ARIMA ganhou destaque entre todos, pois teve os menores valores de rEQM e EAM para a maioria das previsões, com exceção da cidade de São Paulo (previsão 7 passos à frente), em que a previsão mais precisa é apresentada pelo LASSO. Validando essa informação, temos o método ARIMA presente no MCS para todas essas previsões. Essa superioridade do ARIMA pode se justificar pelo fato dos outros modelos serem selecionados por critério de informação (dentro da amostra), o que pode não refletir o melhor modelo para previsão. Nesse caso, sugere-se escolher o melhor modelo já baseado em uma subamostra fora da amostra de treinamento, que fica como ponto adicional para trabalhos futuros.

Na modelagem de MM de mortes, independente da cidade e de quantos passos a

Tabela 3.3: Erros de previsão das Médias Móveis de Mortes por cidade

	1 passo à frente			7 passos à frente		
	rEQM	**EAM**	**ERM**	**rEQM**	**EAM**	**ERM**
São Paulo						
LASSO	3,352	2,607	0,047	**3,296**	**2,552**	0,047
adaLASSO	4,005	3,150	0,058	3,775	2,937	0,055
Random Forest	3,058	2,498	0,044	3,372	2,679	0,046
ARIMA	**1,981**	**1,561**	0,030	4,045	3,366	0,059
Rio de Janeiro						
LASSO	1,665	1,170	0,024	1,636	1,170	0,024
adaLasso	2,210	1,561	0,032	2,237	1,568	0,033
Random Forest	1,832	1,449	0,027	1,658	1,227	0,023
ARIMA	**1,406**	**1,141**	0,024	**1,404**	**1,122**	0,023
Brasília						
LASSO	0,239	0,190	0,016	0,238	0,190	0,016
adaLasso	0,300	0,214	0,017	0,300	0,216	0,018
RandomForest	0,224	0,184	0,015	0,242	0,197	0,016
ARIMA	**0,189**	**0,142**	0,012	**0,231**	**0,175**	0,015

Nota: Os valores em negrito indicam o método com os menores valores de rEQM e EAM para cada horizonte e cidade específicos e as células em cinza e azul indicam que o método está incluído no MCS 50% construído com base na estatística T_{max} usando a raiz do erro quadrado e o erro absoluto.

frente estivermos observando para a previsão, o método que teve pior desempenho, em geral, foi o adaLASSO e ele foi excluído do MCS em todas as previsões geradas. Apesar disso, os erros das previsões tanto 1 quanto 7 passos à frente estão mostrando um bom desempenho de todos os métodos.

As figuras 3.6 e 3.7, por sua vez, correspondem às previsões nos horizontes 1 e 7 para o número de infectados por Covid-19 em cada cidade estudada. Assim como nas figuras anteriores, a linha vermelha representa os valores preditos para cada método e a preta os valores observados do período mostrado. Em geral, podemos perceber que tanto para 1 quanto para 7 passos à frente, as previsões seguem próximas dos valores observados. Para elucidar os resultados dos gráficos, podemos observar a 3.4.

A tabela 3.4 traz os erros de previsão para a MM de casos. Mais uma vez, os menores erros (rEQM e EAM) são indicados em negrito para cada horizonte de previsão e as células em cinza e azul indicam que o método está incluído no MCS de 50% usando a raiz do erro quadrado e o erro absoluto como funções de perda.

Verificamos através da tabela 3.4 que as previsões do método *Random Forest* para a cidade do Rio de Janeiro e de Brasília são superiores as outras em ambos horizontes de previsão, pois possuem erros menores. Vale destacar que o MCS inclui apenas o *Random Forest* para as previsões feitas acerca da média móvel de casos em Brasília. Novamente, o adaLASSO foi excluído de todas as previsões pelo MCS. O método LASSO é o de maior precisão para a previsão 7 passos à frente de São Paulo levando em conta o rEQM, enquanto o *Random Forest* dispõem do menor EAM.

A coluna de ERM das tabelas 3.3 e 3.4 é importante para termos a comparação genuína entre as previsões de diferentes cidades, variáveis resposta ou horizontes de previsão, dado que leva em consideração a escala do problema. Por exemplo, quando observamos os rEQM das MM de casos para São Paulo e Rio de Janeiro na

Tabela 3.4: Erros de previsão das Médias Móveis de Casos por cidade

	1 passo à frente			7 passos à frente		
	rEQM	**EAM**	**ERM**	**rEQM**	**EAM**	**ERM**
São Paulo						
LASSO	69,069	54,718	0,041	**68,817**	**54,481**	0,041
adaLASSO	88,719	71,321	0,054	87,739	70,124	0,053
Random Forest	88,148	62,095	0,046	81,374	52,716	0,039
ARIMA	**66,149**	**49,623**	0,037	82,640	61,415	0,047
Rio de Janeiro						
LASSO	33,974	28,752	0,018	33,371	27,995	0,017
adaLasso	39,381	32,569	0,020	39,547	32,584	0,020
Random Forest	**33,344**	**27,359**	0,017	**26,486**	**21,535**	0,014
ARIMA	34,369	27,791	0,017	36,505	29,360	0,019
Brasília						
LASSO	9,358	6,374	0,010	9,191	6,232	0,010
adaLasso	10,916	7,275	0,011	10,942	7,272	0,012
RandomForest	**7,052**	**4,806**	0,007	**6,823**	**4,914**	0,008
ARIMA	11,057	9,292	0,015	9,784	7,631	0,012

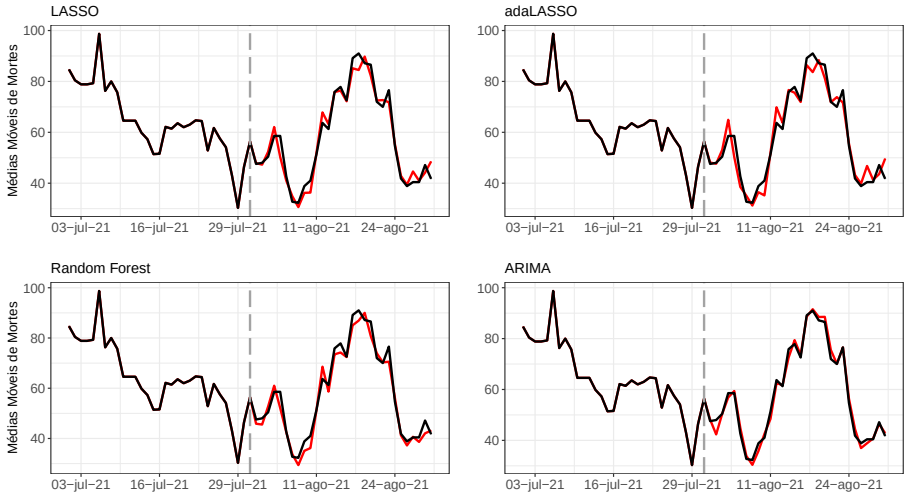
Nota: Os valores em negrito indicam o método com os menores valores de rEQM e EAM para cada horizonte e cidade específicos e as células em cinza e azul indicam que o método está incluído no MCS 50% construído com base na estatística T_{max} usando a raiz do erro quadrado e o erro absoluto.

tabela 3.4, notamos que são muito superiores aos de Brasília. Porém, se voltarmos um pouco aos gráficos da figura 3.3 percebemos a diferença de escala entre eles. Por isso, essa análise deve ser feita através dos resultados da coluna de ERM.

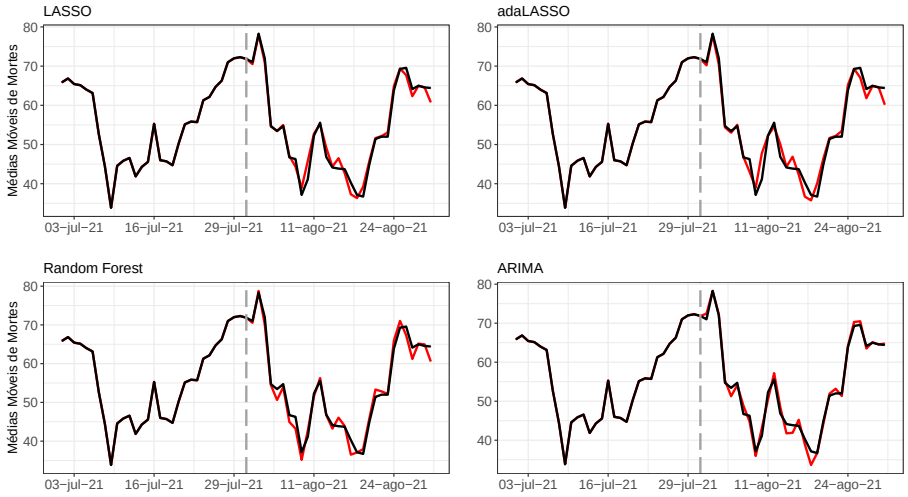
Desse modo, o ERM nos mostra que as previsões de MM de casos acertaram, em média, mais que as de mortes, dado que os resultados desse último são maiores. Podemos, ainda, perceber que as piores previsões de MM de mortes e casos foram relacionadas à cidade de São Paulo, pois quanto maior o ERM, menor a exatidão dos resultados e, em ambas, chegamos a encontrar valores próximos de 0,06.

É importante salientar que os testes que foram feitos com previsões 1 passo à frente servem, especialmente, como forma de avaliar os métodos e modelos. Na prática, previsões 7 passos à frente são mais essenciais, porque podem, de fato, auxiliar no planejamento, desenvolvimento e implementação de avaliações de políticas de resposta à pandemia por governos e agências de saúde pública.

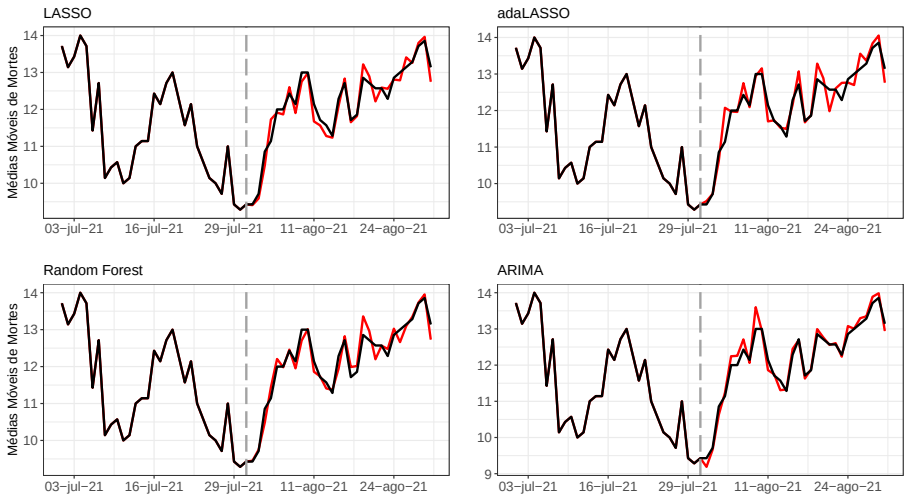
Além disso, também devemos esclarecer que estamos cientes que o número de casos é subnotificado, por isso, a nossa finalidade não era prever adequadamente o número total de infectados. O número relevante e que foi utilizado é correspondente ao número de indivíduos infectados que procuram as unidades de saúde e causam pressão hospitalar.



(a) São Paulo

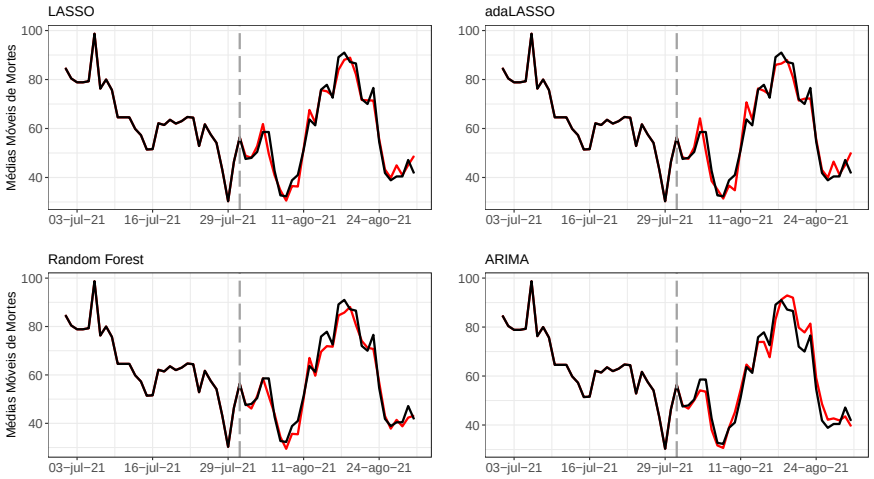


(b) Rio de Janeiro

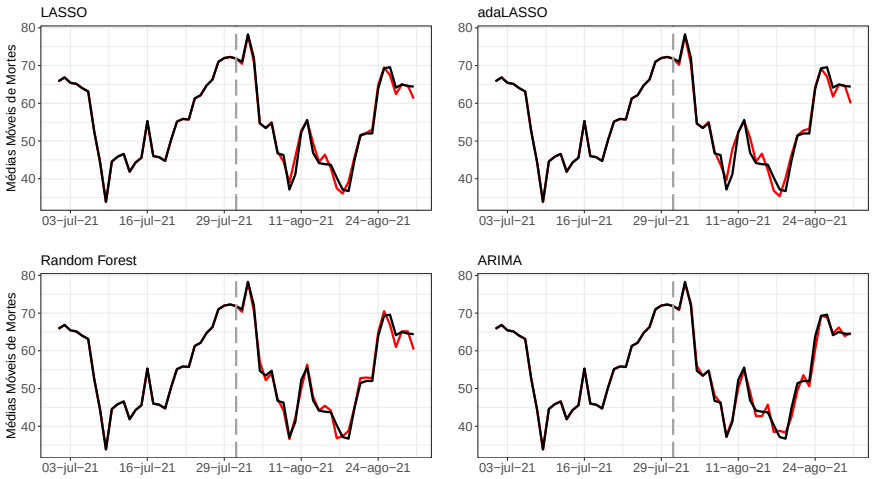


(c) Brasília

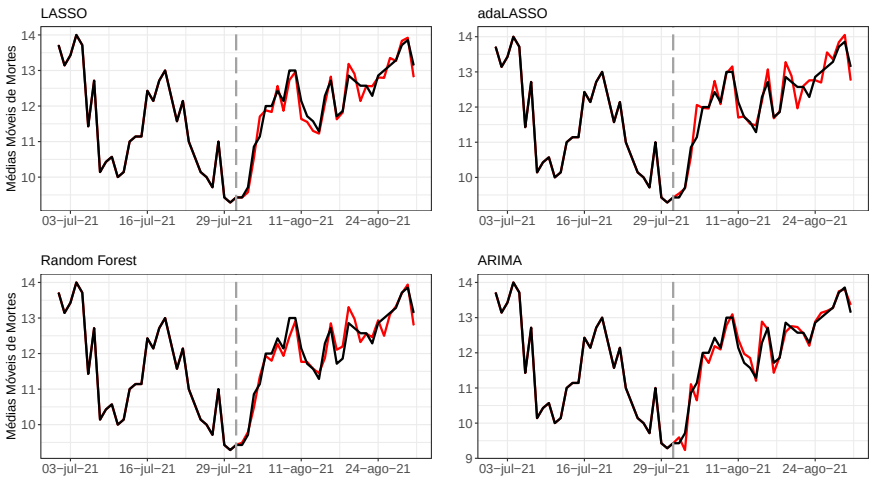
Figura 3.4: Gráficos das previsões 1 passo à frente das médias móveis do número de mortes por cidade
A linha vermelha representa os valores preditos para cada método e a preta os valores observados do período.



(a) São Paulo

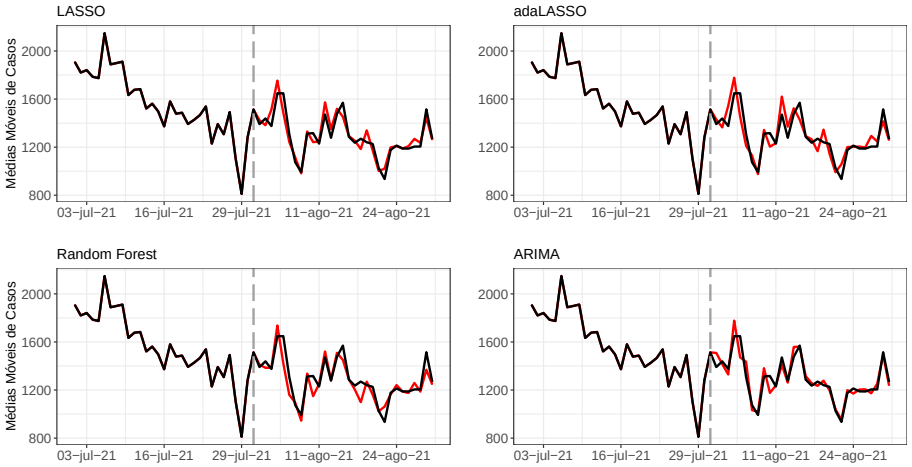


(b) Rio de Janeiro

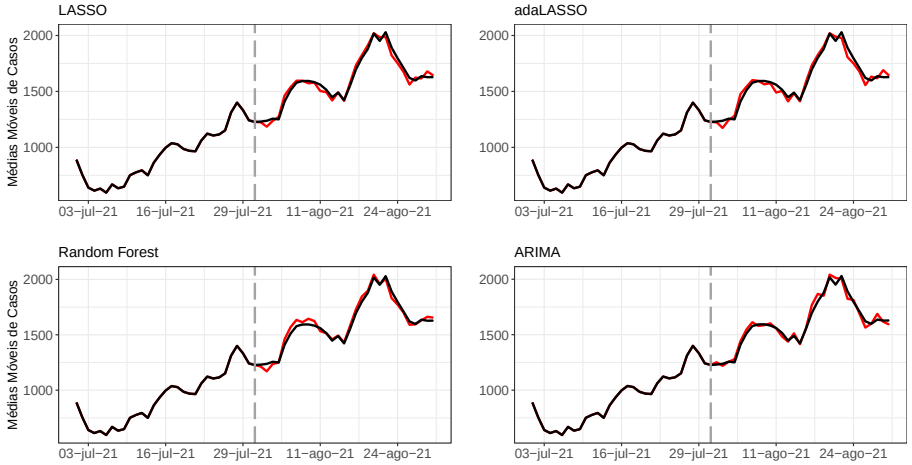


(c) Brasília

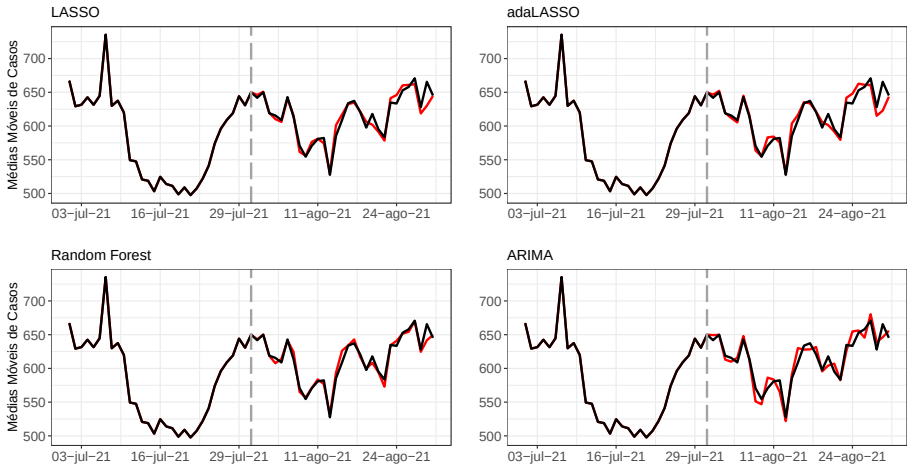
Figura 3.5: Gráficos das previsões 7 passos à frente das médias móveis do número de mortes por cidade
A linha vermelha representa os valores preditos para cada método e a preta os valores observados do período.



(a) São Paulo

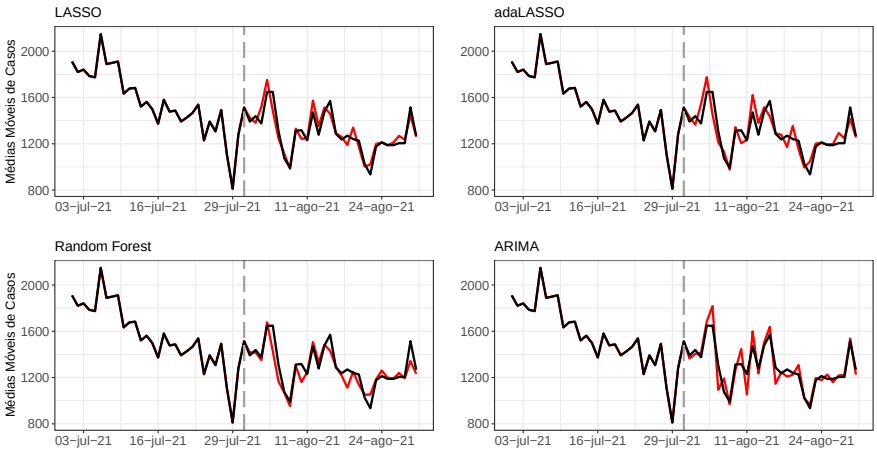


(b) Rio de Janeiro

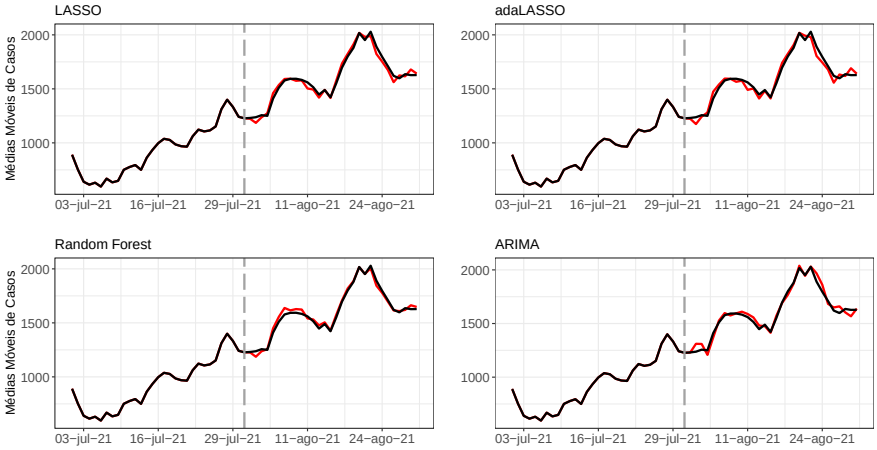


(c) Brasília

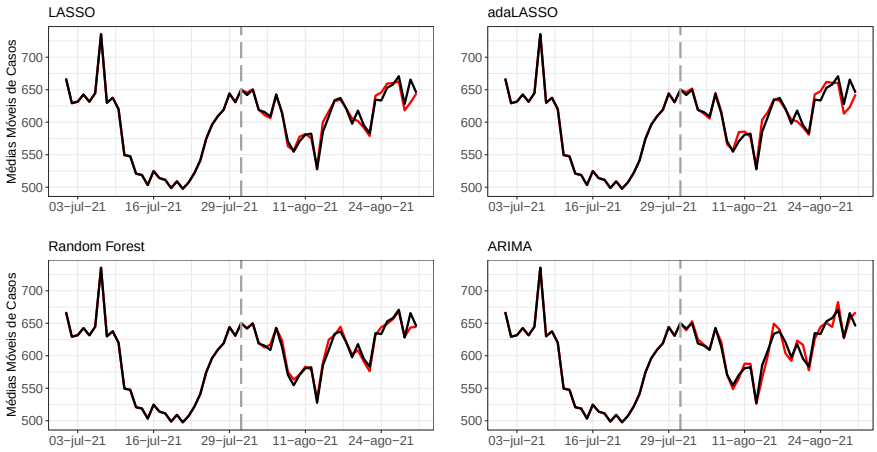
Figura 3.6: Gráficos das previsões 1 passo à frente das médias móveis do número de casos por cidade
A linha vermelha representa os valores preditos para cada método e a preta os valores observados do período.



(a) São Paulo



(b) Rio de Janeiro



(c) Brasília

Figura 3.7: Gráficos das previsões 7 passos à frente das médias móveis do número de casos por cidade
A linha vermelha representa os valores preditos para cada método e a preta os valores observados do período.

4 Considerações Finais

O principal objetivo deste trabalho foi investigar, através dos resultados disponíveis, a competência de métodos mais atuais de *Machine Learning*, LASSO, adaLASSO e Random Forest sobre outros que já sabe-se que desempenham bem na esfera da previsão, como o ARIMA. Para isso, avaliamos os desempenhos por meio de exercícios numéricos, incluindo simulação Monte Carlo e análise de dados empíricos.

No processo de simulação, obtivemos um bom desempenho de previsão especialmente do método adaLASSO, mas também observamos o LASSO com resultados muito semelhantes. Apesar do adaLASSO apresentar resultados mais precisos nesse contexto, não notamos o mesmo comportamento para os dados referentes à evolução da pandemia, visto que para ambas as variáveis resposta analisadas, foi o método com os erros de previsão mais altos. O *Random Forest*, por sua vez, apresentou-se como o método com maior capacidade preditiva na maioria das vezes quando estávamos prevendo as médias móveis de casos de Covid-19. Quando olhamos para os resultados das previsões de médias móveis de mortes, quem se destaca é o ARIMA. Ainda, podemos salientar que olhando especificamente para a previsão 7 passos à frente na análise empírica, indiferentemente da variável resposta, o *Random Forest* mostra superioridade preditiva em seus resultados.

Portanto, não temos como assegurar que um método seja inteiramente o melhor, apenas que existem alguns que funcionam melhor em determinadas situações. Isto é, um método que faz boas previsões para uma determinada estrutura de dados pode não possuir a mesma capacidade para prever outras.

Tendo a pesquisa analisado a influência da tecnologia no decorrer da pandemia e, observando que a utilização de vários métodos foram propostos no âmbito de previsão de séries temporais de diversas estruturas, pudemos mostrar que os métodos propostos são ferramentas eficazes para previsão, principalmente relacionada ao surto de Covid-19 nas cidades brasileiras em questão. Para os dados empíricos, mesmo dispondo de um estudo *ex post facto*, uma vez que tomamos como experimento fatos passados que desenvolveram-se naturalmente, observamos métodos de *Machine Learning* assumindo um papel considerável para futuras pandemias, revelando potenciais benefícios a curto e a longo prazo, que constituem, até mesmo, ações de controle.

Para finalizar, é importante salientar que a análise empírica deste estudo pode ser replicada para outras cidades. Por isso, como sugestão de trabalhos futuros, indica-se a implementação de um código que atue automaticamente para qualquer cidade que possua características próximas às estudadas.

Referências Bibliográficas

Ahmed, N. K., Atiya, A. F., Gayar, N. E., e El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6):594–621.

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., e Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. *Available at SSRN 3580188*.

Asim, K., Martínez-Álvarez, F., Basit, A., e Iqbal, T. (2017). Earthquake magnitude prediction in hindukush region using machine learning techniques. *Natural Hazards*, 85(1):471–486.

Bauer, E. e Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36(1-2):105–139.

Box, G., Jenkins, G., e of STATISTICS., W. U. M. D. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Carlson, C. J., Dougherty, E., Boots, M., Getz, W., e Ryan, S. J. (2018). Consensus and conflict among ecological forecasts of zika virus outbreaks in the united states. *Scientific reports*, 8(1):1–15.

Chiavegatto Filho, A. D. P. (2015). Uso de big data em saúde no brasil: perspectivas para um futuro próximo. *Epidemiologia e Serviços de Saúde*, 24:325–332.

Cooper, I., Mondal, A., e Antonopoulos, C. G. (2020). A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals*, page 110057.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dantas, T. M., Oliveira, F. L. C., e Repolho, H. M. V. (2017). Air transportation demand forecast through bagging holt winters methods. *Journal of Air Transport Management*, 59:116–123.

Dehesh, T., Mardani-Fard, H., e Dehesh, P. (2020). Forecasting of covid-19 confirmed cases in different countries with arima models. *medRxiv*.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Fantazzini, D. (2020). Short-term forecasting of the COVID-19 pandemic using Google Trends data: Evidence from 158 countries. MPRA Paper 102315, University Library of Munich, Germany.

Fitch, F. B. (1944). Mcculloch warren s. and pitts walter. a logical calculus of the ideas immanent in nervous activity. bulletin of mathematical biophysics, vol. 5, pp. 115–133.

Friedman, J., Hastie, T., e Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Garcia, M. G., Medeiros, M. C., e Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, 33(3):679–693.

Hansen, P., Lunde, A., e Nason, J. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hillebrand, E. e Medeiros, M. C. (2010). The benefits of bagging for forecast models of realized volatility. *Econometric Reviews*, 29(5-6):571–593.

Holmstrom, M., Liu, D., e Vo, C. (2016). Machine learning applied to weather forecasting. *Stanford University*, pages 2–4.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kane, M. J., Price, N., Scotch, M., e Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics*, 15(1):276.

Kleinberg, E. M. (2000). A mathematically rigorous foundation for supervised learning. In *International Workshop on Multiple Classifier Systems*, pages 67–76. Springer.

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.

Li, J. e Chen, W. (2014). Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, 30(4):996–1015.

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

Lippi, M., Bertini, M., e Frasconi, P. (2013). Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882.

Medeiros, M., Street, A., Valladão, D., Vasconcelos, G., e Zilberman, E. (2020). Short-term covid-19 forecast for latecomers. *arXiv preprint arXiv:2004.07977*.

Morettin, P. A. e Toloi, C. (2006). Análise de séries temporais. In *Análise de séries temporais*, pages 538–538.

Oliveira, C. (2021). Rio grande do sul sob bandeira preta: uma avaliação do modelo de distanciamento controlado através de uma análise quase experimental baseada em previsões realizadas com o auxílio de buscas no google.

Palit, A. K. e Popovic, D. (2006). *Computational intelligence in time series forecasting: theory and engineering applications*. Springer Science & Business Media.

Pyayt, A. L., Mokhov, I. I., Lang, B., Krzhizhanovskaya, V. V., Meijer, R. J., et al. (2011). Machine learning methods for environmental monitoring and flood protection. *World Academy of Science, Engineering and Technology*, 5(4):118–123.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., e dos Santos Coelho, L. (2020). Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil. *Chaos, Solitons Fractals*, 135:109853.

RStudio (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Said, S. E. e Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.

Sehra, S. T., Salciccioli, J. D., Wiebe, D. J., Fundin, S., e Baker, J. F. (2020). Maximum daily temperature, precipitation, ultra-violet light and rates of transmission of sars-cov-2 in the united states. *Clinical Infectious Diseases*.

Shalev-Shwartz, S. e Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Tandon, H., Ranjan, P., Chakraborty, T., e Suhag, V. (2020). Coronavirus (covid-19): Arima based time-series analysis to forecast near future. *arXiv preprint arXiv:2004.07859*.

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

Vaishya, R., Javaid, M., Khan, I. H., e Haleem, A. (2020). Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*.

Walker, P. G., Whittaker, C., Watson, O., Baguelin, M., Ainslie, K., Bhatia, S., Bhatt, S., Boonyasiri, A., Boyd, O., Cattarino, L., et al. (2020). The global impact of covid-19 and strategies for mitigation and suppression. *Imperial College London*.

Ward, M. P., Xiao, S., e Zhang, Z. (2020). The role of climate during the covid-19 epidemic in new south wales, australia. *Transboundary and Emerging Diseases*.

Werbos, P. e John, P. (1974). Beyond regression : new tools for prediction and analysis in the behavioral sciences /.

Yin, R., Tran, V. H., Zhou, X., Zheng, J., e Kwoh, C. K. (2018). Predicting antigenic variants of h1n1 influenza virus based on epidemics and pandemics using a stacking model. *PloS one*, 13(12):e0207777.

Zakaria, Z. A. e Shabri, A. (2012). Streamflow forecasting at ungaged sites using support vector machines. *Applied Mathematical Sciences*, 6(60):3003–3014.

Zaytar, M. A. e El Amrani, C. (2016). Sequence to sequence weather forecasting with long short-term memory recurrent neural networks. *International Journal of Computer Applications*, 143(11):7–11.

Zhong, L., Mu, L., Li, J., Wang, J., Yin, Z., e Liu, D. (2020). Early prediction of the 2019 novel coronavirus outbreak in the mainland china based on simple mathematical model. *IEEE Access*, 8:51761–51769.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*.

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

Zou, H., Hastie, T., e Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173 – 2192.