UFSM

CIÊNCIA e NATURA

ISSN 2179-460X
OPEN ACCESS

Ci. e Nat., Santa Maria, v. 46, e84257, 2024• https://doi.org/10.5902/2179460X84257
Submissão: 27/06/23 • Aprovação: 20/05/24 • Publicação: 29/11/24

Statistics

Preenchimento de valores faltantes em séries temporais utilizando árvores de decisão

Missing values imputation in time series using decision trees

Alisson Silva Neimaier [I], **Taiane Schaedler Prass** [I]

[I]Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

RESUMO

O preenchimento de valores faltantes em séries temporais é um problema que tem recebido pouca atenção. Os estudos encontrados na literatura geralmente se concentram em modelos lineares da família ARIMA e não discutem a validade das metodologias propostas para casos com um grande volume de dados faltantes, nos quais métodos paramétricos tornam-se desafiadores devido ao problema adicional de identificar a ordem do modelo. Para abordar essas questões, este estudo propõe uma metodologia de reconstrução de séries temporais utilizando árvores de decisão, um método de aprendizado de máquina que não assume um modelo paramétrico para os dados. Nessa abordagem, os valores conhecidos da série temporal atuam como a variável resposta, enquanto as defasagens correspondentes são usadas como preditores. A árvore selecionada pelo algoritmo de treinamento é então usada para prever os valores faltantes na resposta. Simulações de Monte Carlo são utilizadas para investigar a metodologia proposta, considerando processos da família ARMA e o passeio aleatório, variando o tamanho da série temporal, parâmetros dos modelos, proporção de valores faltantes e os preditores. Para avaliar a qualidade das reconstruções, as previsões das árvores de decisão foram comparadas com as de alguns métodos tradicionais de imputação. Os resultados demonstram o potencial do método proposto e são consistentes com o arcabouço teórico deste estudo. Para promover a metodologia proposta, foi desenvolvido um aplicativo em Shiny que está disponível publicamente.

Palavras-chave: ARMA; Passeio aleatório; Árvores de decisão; Dados faltantes; Imputação

ABSTRACT

Filling in missing values in time series is a problem that has received little attention. The studies found in the literature generally focus on linear models from the ARIMA family and do not discuss the validity of proposed methodologies for cases with a large volume of missing data, in which parametric methods become challenging due to the additional problem of identifying the order of the model. To address these

issues, this study proposes a methodology for time series reconstruction using decision trees, a machine learning method that does not assume a parametric model for the data. In this approach, the known values of the time series act as the response variable, while corresponding lags are used as predictors. The tree selected by the training algorithm is then used to predict the missing values in the response. Monte Carlo simulations are used to investigate the proposed methodology, considering processes from the ARMA family and the random walk, while varying the size of the time series, model parameters, proportion of missing values, and the predictors. To evaluate the quality of the reconstructions, the predictions of the decision trees are compared with those of some traditional imputation methods. The results demonstrate the potential of the proposed method and are consistent with the theoretical framework of this study. To promote the proposed methodology, a shiny application has been developed and made publicly available.

Keywords: ARMA; Random walk; Decison trees; Missing data; Imputation

1 INTRODUÇÃO

A literatura de modelagem de séries temporais é vasta, mas a presença de dados faltantes pode tornar o processo de seleção de modelos desafiador. Alguns padrões como tendência e sazonalidade podem ser identificados através de análise gráfica, mesmo quando existem dados faltantes. Porém, existem outras características (por exemplo, a estrutura de dependência) que exigem a utilização de técnicas mais complexas que muitas vezes só podem ser aplicadas em dados completos.

Modelos autoregressivos de médias móveis (ARMA) estacionários são casos particulares dos modelos lineares gerais que, pelo teorema da decomposição de Wold (veja Brockwell and Davis, 1991, página 188), podem ser utilizados para descrever qualquer processo estocástico fracamente estacionário. Essa é uma das características que torna tais modelos tão atrativos e faz com que os modelos lineares da família ARIMA (autoregressivos integrados de médias móveis) sejam amplamente estudados e aplicados na literatura. Estudos envolvendo dados faltantes e modelos dessa classe podem ser encontrados, por exemplo, em Ljung (1989); Luceño (1997); Yodah et al. (2013). Entretanto, nos estudos envolvendo processos ARIMA, geralmente os autores não discutem a validade das metodologias propostas para o caso de um grande volume de dados faltantes.

Motivado pela dificuldade associada ao processo de identificação do modelo ARMA quando a quantidade de dados faltantes é muito grande, este trabalho pretende abordar uma metodologia para recomposição de séries temporais que não assuma um modelo paramétrico. Métodos de aprendizado de máquina (*machine*

learning) aparecem como uma alternativa neste contexto. Dentre as abordagens já utilizadas na literatura envolvendo métodos de aprendizado de máquina podemos citar Dergachev et al. (2001). No artigo em questão os autores apresentam um método para recuperar os dados faltantes em séries temporais que se baseia em modelar os dados por meio de variedades (*manifolds*) de pequenas dimensões em combinação com redes neurais. Por meio de um aplicação a dados reais os autores mostram que é possível recuperar de forma satisfatória lacunas onde dados foram propositalmente deixados de fora, em cenários que o total de dados faltantes chega a 50%.

Neste trabalho é proposta uma metodologia de preenchimento de dados faltantes em séries temporais que utiliza árvores de decisão (Breiman et al., 1984). Árvores de decisão é um método não paramétrico, flexível quanto às variáveis explicativas e capaz de lidar facilmente com valores faltantes. Tendo em vista essas características, essa abordagem se mostra bastante promissora dentro do escopo deste trabalho. Na metodologia proposta, o preenchimento de valores faltantes utilizando árvores de decisão é feito tomando como variável dependente os valores observados da série temporal e como variáveis explicativas as observações anteriores e/ou posteriores a estas. O modelo de regressão ajustado a estes dados é então utilizado para obter as previsões para os dados faltantes.

Devido à dificuldade em se obter resultados teóricos referentes à metodologia proposta, consideramos simulações de Monte Carlo para analisar o desempenho do método. Neste trabalho são consideradas séries temporais simuladas a partir de modelos ARMA e de um passeio aleatório, para que seja possível estudar o desempenho do método proposto em contexto de estacionariedade e não estacionariedade em uma família de modelos de séries temporais amplamente estudada. Além de variar o tamanho das séries temporais, os parâmetros dos modelos e a proporção de valores faltantes, também foi explorado o desempenho da metodologia em termos da quantidade de preditores utilizados. Para a implementação do método proposto utiliza-se o software R, versão 4.1.3 (R Core Team, 2022). Além de ser um software livre e flexível para o desenvolvimento de algoritmos de estatística, ele dispõe de bibliotecas com métodos tradicionais de reconstrução de séries temporais, que são utilizadas para fins de comparação da qualidade do preenchimento dos valores faltantes. Além disso, para motivar a utilização do método

proposto, criou-se uma interface interativa para a análise gráfica e recomposição das séries temporais utilizando a ferramenta `Shiny`.

Este artigo é organizado como segue, na Seção 2 são descritos os principais conceitos teóricos envolvendo séries temporais, valores faltantes e árvores de decisão, a serem utilizados no decorrer do trabalho, na Seção 3 são descritas as simulações de Monte Carlo realizadas e apresentados os resultados encontrados, a Seção 4 é dedicado à apresentação do aplicativo *Shiny*. Finalmente, as conclusões são descritas na Seção 5.

2 REFERENCIAL TEÓRICO

No que segue são apresentados conceitos envolvendo o objeto de estudo (séries temporais), o problema relacionado ao objeto de estudo (valores faltantes) e a ferramenta utilizada para resolver tal problema (árvores de decisão). Os conceitos envolvendo séries temporais são vitais para a construção dos algoritmos e discussões sobre resultados encontrados. Estudos mais detalhados podem ser encontrados em Van der Vaart (2010); Brockwell and Davis (1991); Morettin and Toloi (2004); Shumway and Stoffer (2005).

2.1 Séries Temporais

Uma série temporal é um conjunto de observações ordenadas no tempo. Para fins de modelagem, assume-se que uma série temporal é uma realização, ou parte de uma realização, de um processo estocástico $\{X_t\}_{t\in T}$, onde $T \neq \emptyset$ é um conjunto de índices. Sendo assim, quando falamos em propriedades da série temporal, na verdade estamos nos referindo às propriedades do processo estocástico que deu origem à série temporal.

As propriedades de estacionariedade forte e ergodicidade estão entre as mais desejáveis. Isso deve-se ao fato que, se $\{X_t\}_{t\in T}$ é fortemente estacionário e ergódico, então, como consequência, com base em apenas uma série temporal é possível tirar conclusões sobre as distribuições marginais e conjuntas do processo. Porém estas são propriedade muito restritivas e difíceis de verificar na prática. Sendo assim, com muita frequência, o foco dos estudos são processos fracamente estacionários. Nesse contexto, as quantidades de interesse passam a ser os momentos de primeira e segunda ordem

e as funções de autocovariância e autocorrelação. Na prática, é possível investigar se a hipótese de estacionariedade fraca é plausível observando o gráfico da série temporal pois, se essa propriedade for válida, os dados devem flutuar em torno de uma média constante, com variabilidade estável ao longo do tempo.

As funções de autocovariância e autocorrelação medem o grau de interdependência linear entre as variáveis aleatórias. Em particular, a função de autocorrelação parcial é uma ferramenta importante quando deseja-se investigar a correlação entre as variáveis X_t e X_{t+h} após a remoção das dependências lineares das variáveis aleatórias intermediárias $X_{t+1}, \cdots, X_{t+h-1}$. Como este trabalho pretende realizar previsões utilizando defasagens da própria série temporal como variáveis explicativas, é importante entender o comportamento destas funções para modelos que serão utilizados nas simulações, isto é, processos autoregressivos de médias móveis (ARMA) e o passeio aleatório.

2.1.1Processos ARMA

Segundo Van der Vaart (2010) os processos ARMA são uma versão de regressão linear para séries temporais, em que as variáveis explicativas são os valores anteriores dessa série temporal e o erro adicionado é um processo de médias móveis. Formalmente, dizemos que $\{X_t\}_{t\in\mathbb{Z}}$ é um processo ARMA(p,q) se existem $\mu \in \mathbb{R}$, polinômios $\phi(\cdot)$ e $\theta(\cdot)$ de graus p e q, respectivamente, e um processo ruído branco $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ com média zero e variância σ_ε^2, tais que

$$\phi(L)(X_t - \mu) = \theta(L)\varepsilon_t, \quad t \in \mathbb{Z}. \tag{1}$$

Essa igualdade deve ser entendido como “pontualmente quase certamente” no espaço de probabilidade subjacente. Ressalta-se que a solução de (1) não é única. De fato, sem inicialização ou a imposição de condições adicionais, existem infinitas soluções não estacionárias para esses sistema de equações (para mais detalhes, veja Van der Vaart, 2010). Como forma de garantir unicidade, alguns autores exigem que, por definição, o processo ARMA seja estacionário (veja, por exemplo, Brockwell and Davis, 1991). Sem perda de generalidade os polinômios $\phi(\cdot)$ e $\theta(\cdot)$ (denominados polinômios

característicos) são usualmente escritos como

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \quad \text{e} \quad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

Ainda, por simplicidade de notação, quando $p = 1$ e $q = 1$ define-se $\phi := \phi_1$ e $\theta := \theta_1$, respectivamente. Os modelos mais simples da família ARMA são os autoregressivos (AR) e os de média móvel (MA), obtidos quando $q = 0$ e $p = 0$, respectivamente.

Uma condição necessária e suficiente para a existência e unicidade de uma solução estacionária para o sistema de equações (1) é que $\phi(\cdot)$ não possua raízes no círculo unitário complexo. Tal solução é dada por (veja Brockwell and Davis, 1991, teorema 3.13, página 88)

$$X_t = \mu + \sum_{k=-\infty}^{\infty} \psi_k \varepsilon_{t-k}, \quad t \in \mathbb{Z}, \tag{2}$$

onde os coeficientes $\{\psi_k\}_{k\in\mathbb{Z}}$ são unicamente determinados através da relação $\theta(z)\phi^{-1}(z) = \sum_{k=-\infty}^{\infty} \psi_k z^k$. No caso particular em que $\phi(\cdot)$ também não possui raízes dentro do círculo unitário complexo, conclui-se que $\psi_k = 0$, para todo $k < 0$. Nesse caso, o valor do processo no tempo t depende unicamente do presente e dos valores passados do processo ruído branco. Essa propriedade é conhecida como *causalidade*. Uma característica interessante é que um processo AR(p) estacionário e causal pode ser escrito como um MA(∞) e um processo MA(q), tal que $\theta(\cdot)$ não possui raízes no círculo unitário complexo ou dentro dele, pode ser escrito como um AR(∞). Essa relação permite que tais modelos sejam facilmente identificados com base em suas funções de autocorrelação e autocorrelação parciais.

Vários *softwares*, incluindo o R (R Core Team, 2022), permitem apenas a simulação e ajuste de modelos ARMA causais. O uso dessa restrição é justificado pelo fato que todo processo ARMA(p, q) estacionário pode ser reescrito em termos de um novo processo ruído branco e de polinômios característicos, com os mesmos graus p e q dos polinômios originais, porém tais que as raízes estão todas fora do círculo unitário (veja Brockwell and Davis, 1991, proposição 3.5.1, página 105). Portanto, sem perda de generalidade, serão considerados nas simulações realizadas neste trabalho, apenas processos ARMA

estacionários e causais. Nesse contexto, segue de imediato de (2) que

$$\mathbb{E}(X_t) = \mu, \quad \mathrm{Var}(X_t) = \sigma_\varepsilon^2 \sum_{k=0}^{\infty} \psi_k^2 \quad \text{e} \quad \rho(h) = \left(\sum_{k=0}^{\infty} \psi_k \psi_{k+|h|}\right)\left(\sum_{k=0}^{\infty} \psi_k^2\right)^{-1}, \tag{3}$$

já os valores da função de autocorrelação parcial $\alpha(\cdot)$ podem ser obtidos, de forma recursiva, a partir $\rho(\cdot)$, aplicando-se o algoritmo de Durbin–Levinson. Mais especificamente, $\alpha(h) = \phi_{h,h}$, onde $\phi_{1,1} = \rho(1)$,

$$\phi_{h+1,h+1} = \frac{\rho(h+1) - \sum_{j=1}^{h} \phi_{h,j}\rho(h+1-j)}{1 - \sum_{j=1}^{h} \phi_{h,j}\rho(j)} \quad \text{e} \quad \phi_{h+1,j} = \phi_{h,j} - \phi_{h+1,h+1}\phi_{h,h+1-j}, \tag{4}$$

para $1 \leq j \leq h$.

2.1.2 Passeio Aleatório

Um processo estocástico $\{X_t\}_{t\in\mathbb{N}}$ é chamado de passeio aleatório se pode ser escrito como

$$X_0 = 0, \quad X_t = X_{t-1} + \varepsilon_t, \quad t > 0, \tag{5}$$

onde $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é um processo ruído branco com média zero e variância σ_ε^2. É fácil ver que $\{X_t\}_{t\in\mathbb{N}}$ pode ser reescrito como $X_t = \sum_{i=1}^{t} \varepsilon_i$, para $t > 1$, de onde conclui-se que $\mathbb{E}(X_t) = 0$. Além disso, se $\{Y_t\}_{t\in\mathbb{N}}$ é um passeio aleatório e $X_t = \mu + Y_t$, então $\{X_t - \mu\}_{t\in\mathbb{N}}$ é um passeio aleatório com média zero,

$$\mathbb{E}(X_t) = \mu, \quad \mathrm{Var}(X_t) = \sigma_\varepsilon^2 t \quad \text{e} \quad \mathrm{Cov}(X_t, X_s) = \frac{\min(t,s)}{\sqrt{ts}}, \quad ts > 0.$$

Devido à representação (5), diremos que passeio aleatório é um AR(1) com $\phi = 1$. Ressaltamos, entretanto, que para os autores que definem o processo ARMA como sendo a solução estacionária de (1), essa nomenclatura é um abuso de notação.

2.2 Valores Faltantes

Conforme discutido em Molenberghs et al. (2020), a presença de valores faltantes (ou dados faltantes) é uma questão comum que pode ter várias consequências para a análise estatística. Isso deve-se ao fato que, na presença de

dados faltantes ocorre necessariamente uma perda de informação e uma redução na precisão com que parâmetros de interesse podem ser estimados. Além disso, é natural que essa redução na precisão está diretamente relacionada à quantidade de dados faltantes e é influenciada, até certo ponto, pelo método de análise. Além disso, em certas circunstâncias, dados faltantes podem introduzir viés e, portanto, levar a inferências enganosas sobre os parâmetros de interesse. A possibilidade de viés é um motivo de grande preocupação e acaba sendo o que torna a análise de dados incompletos mais desafiadora. Em particular, quando há dados faltantes, a validade de qualquer método de análise exigirá certas suposições sobre as razões pelas quais os valores faltantes ocorrem.

No que segue, discutimos os mecanismos de dados faltantes e qual deles se aplica ao estudo realizado neste trabalho. Os conceitos aqui apresentados são discutidos com mais detalhes em Molenberghs et al. (2020). Também apresentamos uma breve descrição dos métodos clássicos utilizados no processamento de dados faltantes.

2.2.1 Mecanismos de dados faltantes

Seja $\boldsymbol{Y} = (Y_1, \cdots, Y_n)'$ o vetor com n valores da variável resposta e X a matriz $n \times p$ de covariáveis associada a $\boldsymbol{Y}$. Seja $\boldsymbol{R} = (R_1, \cdots, R_n)'$ o vetor cuja coordenada R_i assume o valor 0 ou 1, se Y_i foi observado ou um valor faltante, respectivamente, para $1 \leq i \leq n$. Observe que, dado R_i, o vetor $\boldsymbol{Y}$ pode ser particionado em dois subvetores $\boldsymbol{Y}^o$ e $\boldsymbol{Y}^m$ que correspondem às componentes observadas e faltantes de $\boldsymbol{Y}$, respectivamente. Essas duas componentes são comumente denominados "dados observados" e "dados faltantes", respectivamente. Por sua vez, $\boldsymbol{Y}$ é um vetor hipotético, que seria obtido caso não existissem dados faltantes. O mecanismo de dados faltantes descreve a probabilidade de que uma resposta seja observada ou esteja faltando. Mais especificamente, ele estabelece um modelo probabilístico para a distribuição dos indicadores de resposta $\boldsymbol{R}$ condicional a $\boldsymbol{Y}^o$, $\boldsymbol{Y}^m$ e X. Conforme Rubin (1976), ao considerar como $\boldsymbol{R}$ está relacionado $\boldsymbol{Y}$ e X, identificamos três categorias de mecanismos de dados faltantes: (i) faltando de forma completamente aleatória (MCAR - *missing completely at random*), (ii) faltando de forma aleatória (MAR - *missing at random*) e (iii) faltando de forma não aleatória (NMAR - *missing not at random*).

Dizemos que os dados estão faltando de forma completamente aleatória quando $\boldsymbol{R}$ é independente de $\boldsymbol{Y}^o$ e $\boldsymbol{Y}^m$, ou seja, quando a probabilidade de observações da resposta estarem faltando não tem relação com os valores específicos, que deveriam ter sido obtidos, ou com o conjunto de respostas observadas. Na presença de covariáveis, não existe um consenso universal na literatura em termos da definição de MCAR também pressupor independência dos valores faltantes em X ou não. Sendo assim, seguindo a ideia de Little (1995) reservamos o termo MCAR para o caso em $P(\boldsymbol{R} \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m, X) = P(\boldsymbol{R})$ e, quando $P(\boldsymbol{R} \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m, X) = P(\boldsymbol{R} \mid X)$, diremos que o mecanismo de dados faltantes é "covariável-dependente". Os dados são considerados faltando de forma aleatória quando a probabilidade de que as respostas estejam faltando depende do conjunto de respostas observadas $\boldsymbol{Y}^o$, mas não tem relação com os valores específicos $\boldsymbol{Y}^m$ que, em princípio, deveriam ter sido obtidos. Em particular, os dados são MAR quando $\boldsymbol{R}$ é condicionalmente independente de $\boldsymbol{Y}^m$, dado $\boldsymbol{Y}^m$, isto é, $P(\boldsymbol{R} \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m, X) = P(\boldsymbol{R} \mid \boldsymbol{Y}^o, X)$. Finalmente, dizemos que os dados estão faltando de forma não aleatória quando a probabilidade de que as respostas estejam faltando está relacionada aos valores específicos $\boldsymbol{Y}^m$ que deveriam ter sido obtidos, além dos valores que foram efetivamente obtidos $\boldsymbol{Y}^o$. Portanto, os dados são NMAR se $P(\boldsymbol{R} \mid \boldsymbol{Y}, X) = P(\boldsymbol{R} \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m, X)$ depender de pelo menos uma componente de $\boldsymbol{Y}^m$.

Neste trabalho, assume-se que os valores faltantes são gerados pelo mecanismo MCAR. Conforme Greiner et al. (1997), esse cenário pode não ser factível em alguns problemas reais, porém, segundo Hastie et al. (2009) a maior parte dos métodos de imputação fazem esta suposição para sua validade.

2.2.2 Métodos clássicos para processamento de dados faltantes

Conforme Pratama et al. (2016), de modo geral, os métodos de tratamento de valores faltantes podem ser divididos em 3 principais categorias: (i) ignorar ou descartar dados, (ii) estimação e (iii) imputação. Existem dois métodos para ignorar ou descartar dados. O primeiro é conhecido como análise de casos completos (*complete case analysis*), que consiste em remover quaisquer observações com dados faltantes e o segundo é o descarte de variáveis (*case deletion*), que exclui as variáveis dependendo da quantidade de observações faltantes. Mais informações sobre essa abordagem

pode ser encontrada em Batista and Monard (2003). Ressaltamos, entretanto, que ela não é recomendada no contexto de séries temporais. Na categoria de estimação destacam-se os procedimentos de estimação por máxima verossimilhança, que são utilizados para estimar de forma paramétrica um modelo para os dados completos. Além disso, segundo Dempster et al. (1977), algumas variações do algoritmo EM (*Expectation-Maximization*) conseguem lidar com a estimação de parâmetros com dados incompletos. Por fim, técnicas de imputação buscam preencher os valores faltantes com valores estimados fazendo uso de relações conhecidas com os valores observados. O método proposto neste trabalho, que utiliza a média condicional, se enquadra nessa categoria.

Na literatura encontramos diversos métodos de inputação em séries temporais univariadas (veja, por exemplo, Moritz and Bartz-Beielstein, 2017, e referências ali contidas). Dentre eles destacam-se os métodos implementados no pacote `imputeTS` (Moritz and Bartz-Beielstein, 2017), que serão utilizados como forma de comparação neste trabalho: (i) o uso de medidas de tendência central como a média, mediana ou moda, (ii) o método de médias móveis com peso simples, linear ou exponencial, (iii) as técnicas última observação levada adiante (LOCF - *last observation carried forward*) e próxima observação trazida para trás (NOCB - *next observation carried backward*), (iv) intepolação linear, por splines ou de Stineman, (v) o método de suavização de Kalman, onde considera-se um modelo estrutural estimado via máxima verossimilhança e usando uma representação do espaço de estados do modelo ARIMA (cuja ordem é escolhida automaticamente) e (vi) o método aleatório que seleciona aleatoriamente uma observação da amostra para reconstrução do valor faltante.

2.3 Árvores de Decisão

Apredizagem estatística (*statistical learning*) refere-se a um vasto conjunto de ferramentas para modelagem e compreensão de dados complexos (James et al., 2013). Essas ferramentas podem ser classificadas como supervisionadas ou não supervisionadas. O aprendizado supervisionado envolve a construção de um modelo estatístico para prever/estimar uma saída (isto é, uma variável resposta ou dependente) com base em uma ou mais entradas (isto é, variáveis explicativas ou independentes). Com o aprendizado não supervisionado, há entradas, mas não há

saída de supervisão, entretanto, podemos aprender relacionamentos e estruturas a partir dos dados. Neste trabalho abordamos um problema que se enquadra no contexto de aprendizado supervisionado e a ferramenta que utilizaremos é o modelo de árvores de decisão. Nesta seção, são apresentados os principais conceitos relacionados à teoria de árvores de decisão. Para uma leitura mais detalhada sobre a aplicação do método, veja James et al. (2013) e Hastie et al. (2009) e para uma visão mais teórica sobre o assunto, veja Murphy (2012).

2.3.1 O que é uma árvore de decisão

Uma árvore de decisão é um algoritmo de aprendizado supervisionado não paramétrico que simula a sequência lógica de tomada de decisões de um ser humano, criando um fluxograma de perguntas e respostas em que a resposta final é a decisão a ser tomada. O algoritmo para modelos de árvore de decisão particiona repetidamente os espaço das entradas em vários subespaços, de forma que os resultados em cada subespaço final sejam tão homogêneos quanto possível. Ao finalizar o processo obtemos uma estrutura de árvore hierárquica, que consiste de nós internos e externos conectados por ramos. Um nó pode ainda ser classificado como nó pai ou nó filho, dependendo de sua origem: um nó que é dividido em subnós é chamado de nó pai, os subnós são chamados de nós filhos.

São denominados nós internos o nó raíz (ou nó inicial), que recebe a totalidade dos dados, e os nós intermediários, que são criados a partir de testes lógicos. Por sua vez, os nós externos são os nós terminais (ou folhas), que são nós com nenhuma partição a partir deles. Tais nós indicam à decisão final a ser tomada (previsão), quando o algoritmo atinge aquele ponto. Em cada nó interno um teste lógico é aplicado de forma que esse nó é particionado em dois ou mais subnós. Os ramos que conectam os nós aos subnós indicam as possíveis decisões a serem tomadas em cada teste.

2.3.2 Como crescer uma árvore

Seja Y uma variável resposta que toma valores em um espaço amostral $\mathcal{Y}$ e seja $\boldsymbol{X} = (X_1, \cdots, X_p)'$ um vetor de variáveis explicativas, tomando valores em um espaço amostral $\boldsymbol{\mathcal{X}}$. Os métodos de aprendizado supervisionado buscam aprender como

prever valores Y ou, de forma mais geral, de uma função $g(Y)$ a partir de $\boldsymbol{X}$. No contexto tradicional de árvores de decisão, o par $(\boldsymbol{X}, Y)$ é considerado como sendo um vetor aleatório com distribuição conjunta P e os dados observados, denotados por $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$, são vistos como uma amostra aleatória de P.

Existem diferentes algoritmos para construção das árvores de decisão, que variam em método e nos tipos de variáveis que suportam. Dentre eles destacam-se: (i) ID3 (*Iterative Dichotomiser 3* - Quinlan, 1986), para o qual a variável resposta deve ser binária e as variáveis explicativas categóricas , (ii) CHAID (*Chi-Squared Automatic Interaction Detector* - Kass, 1980), para o qual a variável resposta e as variáveis explicativas devem ser categóricas e (iii) CART (*Classification and regression Tree* - Breiman et al., 1984), que não impõe restrições sobre as variáveis. Neste trabalho, será utilizada a versão do algoritmo CART (também conhecido como C&RT) implementada no pacote `rpart` (Therneau and Atkinson, 2019) do R, em que o algoritmo recebe o nome de RPART (*Recursive Partitioning And Regression Trees*). O algoritmo CART cria recursivamente uma partição do espaço de entradas (*input*) $\boldsymbol{\mathcal{X}}$ e realiza as previsões no espaço de saídas (*output*) $\mathcal{Y}$. Para o problema abordado neste trabalho, $\boldsymbol{\mathcal{X}} = \mathbb{R}^p$ e $\mathcal{Y} = \mathbb{R}$. Uma descrição completa desse algoritmo pode ser encontrada no manual fornecido pelos autores do pacote `rpart`[1].

De maneira informal e resumida, podemos dizer que o algoritmo CART determina automaticamente quais variáveis serão utilizadas para criar as partições, a posição das partições e também qual topologia (*shape*) a árvore deve ter. De maneira formal, dado um nó $A = [\ell_1, r_1] \times \cdots \times [\ell_p, r_p] \subset \mathbb{R}^p$, o CART encontra a melhor partição (j^*, z^*) no conjunto de possíveis partições $S = \{(j, z), j \in [1, p] \cap \mathbb{N}, z \in [\ell_j, r_j]\}$, em que j indica o índice da variável para a qual é feita a partição e z a posição em que ocorre a partição. Mais especificamente, a melhor partição (j^*, z^*), em um dado nó A, é qualquer uma das soluções para o problema de otimização dado por (para mais detalhes, veja Josse et al., 2019)

$$(j^*, z^*) = \underset{(j,z)\in S}{\arg\min}\Big\{\mathbb{E}\Big(\big[Y - \mathbb{E}\big(Y|E_{j,z} \cap E\big)\big]^2 I\big(E_{j,z} \cap E\big) + \big[Y - \mathbb{E}\big(Y|E_{j,z}^c \cap E\big)\big]^2 I\big(E_{j,z}^c \cap E\big)\Big)\Big\}, \quad (6)$$

onde $E := [\boldsymbol{X} \in A]$ e $E_{j,z} := [X_j \leq z]$. Para um nó A qualquer (fixo), a otimização acima

[1]Acessível em https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

é equivalente a solucionar o seguinte problema

$$h^* = \underset{h \in \mathcal{P}_c}{\arg\min} \Big\{ \mathbb{E}\Big(\big[Y - h(\boldsymbol{X})\big]^2 \mathbb{I}(\boldsymbol{X} \in A) \Big) \Big\},$$

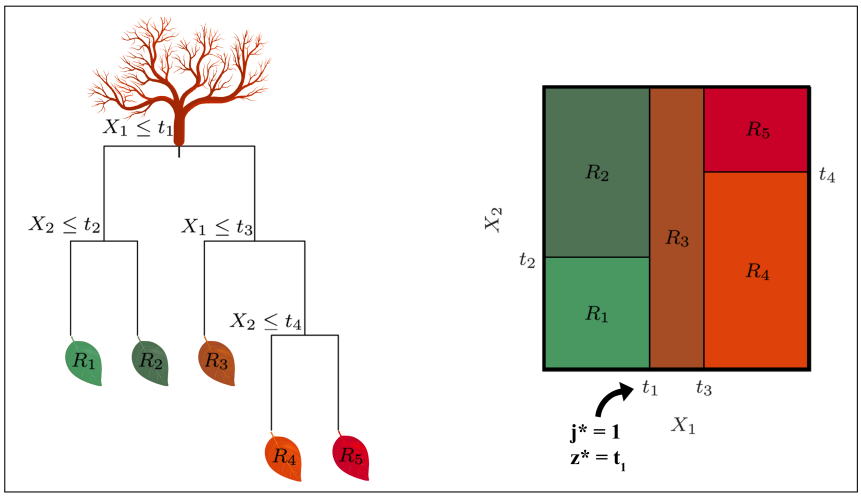
onde $\mathcal{P}_c$ é o conjunto das funções constantes por partes em $A \cap [X_j \leq z]$ e $A \cap [X_j > z]$, para $(j, z) \in S$.

O resultado do processo de partição descrito acima é uma árvore $\mathcal{T}$, com M nós terminais (folhas) e uma partição $\{R_1, \cdots, R_M\}$ de $\mathbb{R}^p$. A função de predição correspondente à árvore $\mathcal{T}$ é dada por

$$f(\boldsymbol{X}) = \sum_{m=1}^{M} c_m \mathbb{I}(\boldsymbol{X} \in R_m), \quad c_m = \mathbb{E}(Y | \boldsymbol{X} \in R_m). \tag{7}$$

A título de ilustração, na Figura 1 é possível visualizar uma árvore de decisão com 5 nós terminais que dão origem às regiões $\{R_1, \cdots, R_5\}$. Nesse exemplo, para o nó raíz, o par que minimiza a equação 6 é $(j^*, z^*) = (1, t_1)$, ou seja, o valor t_1 no intervalo onde a variável X_1 assume valores.

Figura 1 – Exemplo de uma árvore de decisão



Legenda: Resultado de uma partição recursiva binária em um contexto bivariado (à direita) e a árvore de decisão correspondente à esta partição (à esquerda)

2.3.3 Como podar uma árvore

É intuitivo que a melhor solução para o problema de construção de uma árvore seria aquela que minimizasse o erro calculado a partir da função de predição $f(\cdot)$, dada em (7). Porém, em uma árvore grande (com muitos nós) podem ocorrer problemas

de sobreajuste e em uma árvore pequena informações importantes podem não ser capturadas. Sendo assim, uma questão relevante é como decidir o tamanho da árvore de decisão? A estratégia padrão para decidir o tamanho de uma árvore é criar uma árvore grande $\mathcal{T}_0$ e podá-la até encontrar a subárvore que tenha o menor erro em um grupo de teste. O erro é usualmente estimado através do método de validação-cruzada (*cross-validation*). Como fazer a validação-cruzada para todas as subárvores seria uma tarefa muito pesada, utiliza-se um método chamado *cost complexity pruning* (também conhecido como *weakest link pruning*) para analisarmos apenas um pequeno conjunto das subárvores. A ideia geral do algoritmo é explicada a seguir.

Considere uma sequência de árvores indexadas por um parâmetro de ajuste α não negativo. Para cada valor de α, existe uma subárvore $\mathcal{T}_\alpha \subset \mathcal{T}_0$ que minimiza

$$C_\alpha(\mathcal{T}) = \sum_{m=1}^{|\mathcal{T}|} \sum_{i \in I_m} (Y_i - \hat{Y}_m)^2 + \alpha|\mathcal{T}|, \quad \hat{Y}_m = \frac{1}{|\mathscr{X}_m|} \sum_{i \in I_m} Y_i,$$

onde $|\mathcal{T}|$ é o número de nós terminais da árvore $\mathcal{T}$, I_m é o conjunto de índices definido por $I_m = \{i : \boldsymbol{X_i} \in \mathscr{X}_m\}$, $\mathscr{X}_m$ é a região determinada por $\mathscr{X}_m = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\} \cap R_m$, R_m é o subconjunto do espaço de preditores correspondente a m-ésima folha e $|\mathscr{X}_m|$ é o número de observações na m-ésima folha. O parâmetro de ajuste α controla o *trade-off* entre a complexidade da subárvore $\mathcal{T}$ e o quão bem ela se ajusta ao grupo de treino. Quando $\alpha = 0$, a subárvore $\mathcal{T}_\alpha$ vai ser simplesmente $\mathcal{T}_0$. Entretanto, quanto maior o valor de α, maior é o preço a se pagar por ter uma árvore com muitos nós terminais e, portanto, $C_\alpha(\mathcal{T})$ tende a ser minimizado por subárvores menores.

A medida que α cresce os ramos da árvore são podados de uma forma encaixada (*nested*) e previsível (James et al., 2013; Hastie et al., 2009): os nós internos vão sendo aglutinados, dois a dois, até que reste um único nó. Obtém-se assim uma sequência de subárvores em função de α, que contém $\mathcal{T}_\alpha$. A escolha do α é então feita utilizando-se um conjunto de validação ou usando validação cruzada. Uma vez que α é determinado, retorna-se ao conjunto de dados completo para obter a subárvore correspondente a α. O processo para construção e poda de árvores pode ser resumido conforme o algoritmo abaixo (James et al., 2013)

Passo 1: Construa uma árvore grande $\mathcal{T}_0$, utilizando o grupo de treino e o método de divisão binária recursiva, parando apenas quando cada nó terminal tiver um

número de observações menor ou igual a um mínimo pré-determinado;

Passo 2: Aplique *cost complexity pruning* para obter a sequência de melhores subárvores como uma função de α.

Passo 3: Use validação cruzada K-*fold* para escolher α. Para isso, divida as observações do grupo de teste em K subconjuntos e, para cada $k = 1, \cdots, K$: (a) repita os passos 1 e 2 em todos os folds, menos o k-ésimo e (b) avalie o erro quadrático médio de previsão utilizando o k-ésimo fold, como uma função de α. Então, para cada α, calcule a média dos resultados e selecione o α que minimiza o erro médio.

Passo 4: Retorne a subárvore do passo 2 que corresponde ao valor escolhido de α.

Para a implementação do método de preenchimento de dados faltantes proposto neste trabalho, a construção e poda das árvores é realizada utilizando-se o algoritmo implementado na função `rpart` (do pacote homônimo) do *R*. Segundo a documentação do próprio pacote (Therneau and Atkinson, 2019), o algoritmo implementado segue o que foi descrito por Breiman et al. (1984).

3 ESTUDOS DE SIMULAÇÃO

Para investigar a qualidade da reconstrução de séries temporais utilizando árvores de decisão conduzimos um estudo de simulação de Monte Carlo. Nesse estudo são consideradas séries simuladas provenientes de modelos ARMA estacionários e também de um passeio aleatório que, por simplicidade de notação, será denominado um modelo AR(1) com $\phi = 1$. Além de variar a proporção de dados faltantes (ρ), também são considerados diferentes tamanhos de amostras (n) e diversos cenários para as covariáveis. Para fins de comparação, são também empregados os métodos de imputação listados na Seção 2.2.

3.1 Processo gerador de dados

Nesta simulação foram considerados apenas processos ARMA(p, q) com $p, q \in \{0, 1\}$ e $\mu = 100$. Para cada modelo considerado foram realizadas $r = 1000$ replicações e, para cada replicação as amostras $\{X_t\}_{t=1}^{n}$ com valores faltantes, foram geradas seguindo-se os passos descritos a seguir.

PGD1: Define-se $\varepsilon \sim \mathcal{N}(0,1)$ e obtem-se uma amostra i.i.d. $\{\varepsilon_t\}_{t=-b+1}^{m}$ de ε, onde $m = 1000$ é o tamanho amostral e $b = 100$ é o tamanho da amostra de *burn-in*.

PGD2: Para os modelos ARMA, a série temporal com média zero é gerada utilizando a função `arima.sim` do pacote `stats` (R Core Team, 2022) do `R`. Somando-se a média μ, eliminado-se as b primeiras observações e tomando as primeiras m observações da amostra restante, obtém-se a amostra desejada $\{X_t\}_{t=1}^{m}$, sem dados faltantes. O passeio aleatório foi gerado usando a representação dada por $X_t = \mu + \sum_{i=1}^{t} \varepsilon_i$. Neste caso, não é necessário o *burn-in*, ou seja, $b = 0$.

PGD3: Para cada valor de n considerado, um conjunto T_1, com $\lfloor n\rho \rfloor$ elementos é selecionado a partir de uma amostra aleatória simples sem reposição do conjunto $T \in \{1, 2, \cdots, n\}$, as observações da série temporal $\{X_t\}_{t\in T_1}$ são transformadas em valores faltantes. Desta forma, a série temporal com a inclusão dos valores faltantes é dada por $X_t^{\text{miss}} = \texttt{NA} \times \mathbb{I}(t \in T_1) + X_t \times \mathbb{I}(t \in T_1^C)$.

Para cada cenário obtido variando-se o modelo e a configuração das covariáveis, considerou-se como proporção de valores faltantes $\rho \in \{0.1, 0.2, 0.5, 0.8\}$, em combinação com os tamanho da amostra $n \in \{100, 500, 1000\}$. Em termos de valores de parâmetros, para os modelos AR(1) e MA(1) considerou-se $\phi, \theta \in$ {-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1}, já para o modelo ARMA$(1,1)$ fixou-se $\phi = 0.7$ e $\theta = 0.4$, de forma a reproduzir o cenário do artigo Prass and Pumi (2021).

A função de autocorrelação $\rho(h)$, para $h \in \mathbb{Z}$, para o modelo AR(1), MA(1) e ARMA(1,1) é dada, respectivamente, por

$$\rho(h) = \phi^{|h|}, \quad \rho(h) = \mathbb{I}(h = 0) + \frac{\theta}{1+\theta^2}\mathbb{I}(|h| = 1)$$

e

$$\rho(h) = \mathbb{I}(h = 0) + \frac{(\phi+\theta)(1+\phi\theta)\phi^{|h|-1}}{(1-\phi^2)+(\phi+\theta)^2}\mathbb{I}(h \neq 0).$$

A função de autocorrelação parcial $\alpha(k)$, para $k > 1$, para o modelo AR e MA, respectivamente, é dada por

$$\alpha(k) = \phi\mathbb{I}(k = 1) \quad \text{e} \quad \alpha(k) = \frac{(-\theta)^k(1-\theta^2)}{1-\theta^{2(k+1)}}.$$

Ressaltamos que, no caso do modelo ARMA(1,1), não é possível apresentar uma fórmula fechada para $\alpha(k)$, porém, tais valores podem ser facilmente obtidos utilizando-se a relação (4). Com base nesses resultados, as covariáveis nas árvores de decisão foram tomadas como sendo as defasagens $(X_{t-h1}^{\text{miss}}, \cdots, X_{t-1}^{\text{miss}}, X_{t+1}^{\text{miss}}, \cdots X_{t+h2}^{\text{miss}})$ da resposta X_t^{miss}, com $\boldsymbol{h} = (h_1, h_2) \in \{(1,0),(1,1),(2,0),(2,2),(5,0),(5,5)\}$. Observa-se que, em todos os cenários considerados, para $h_1, h_2 > 5$, a dependência entre X_t e X_{t+h} é pequena (ou nula), tanto em termos de autocorrelação, quanto de autocorrelação parcial.

3.2 Processo de imputação dos dados

Para cada uma das replicações dos cenários testados, a série temporal $\{X_t^{\text{miss}}\}_{t=1}^n$ foi reconstituída utilizando-se dos métodos listados na Seção 2.2 (disponíveis no pacote `imputeTS`) e a partir do método proposto, seguindo o algoritmo abaixo

PID1: Definir X_t^{miss} como sendo a variável resposta, para $h_1 < t \leq n - h_2$.

PID2: Construir a matriz de covariáveis, utilizando os h_1 passos anteriores e h_2 passos posteriores de X_t^{miss}.

PID3: Utilizando apenas as respostas e respectivas covariáveis correspondentes aos índices $t \notin T_1$, treinar o modelo de árvore de decisão utilizando a função `rpart` com os seguintes valores para os argumentos: `minsplit = 6`, `cp = 0.01`, `maxcompete = 4`, `maxsurrogate = 5`, `usesurrogate = 0`, `xval = 10`, `maxdepth = 30`.

PID4: Podar a árvore utilizando a função `prune`, com o parâmetro `cp` recebendo o menor valor de erro calculado a partir do método de validação cruzada feito pelo `rpart`.

PID5: Utilizando a árvore de decisão obtida no passo anterior, prever os valores de X_t^{miss}, para $t \in T_1$ com a função `predict`, obtendo-se assim a sequência de valores preditos $\{\hat{X}_t\}_{t \in T_1}$.

Para avaliar a qualidade da reconstituição dos valores faltantes utilizamos como métrica o erro absoluto percentual médio (MAPE - *mean absolute percentage error*), dado por $\text{MAPE} = \lfloor n\rho \rfloor^{-1} \sum_{t \in T_1} |X_t - \hat{X}_t||X_t|^{-1} \times 100\%$. Para fins de validação do método proposto, primeiramente avaliamos a qualidade de preenchimento das árvores para diferentes valores de h_1 e h_2, a fim de definir a melhor configuração de covariáveis. Em

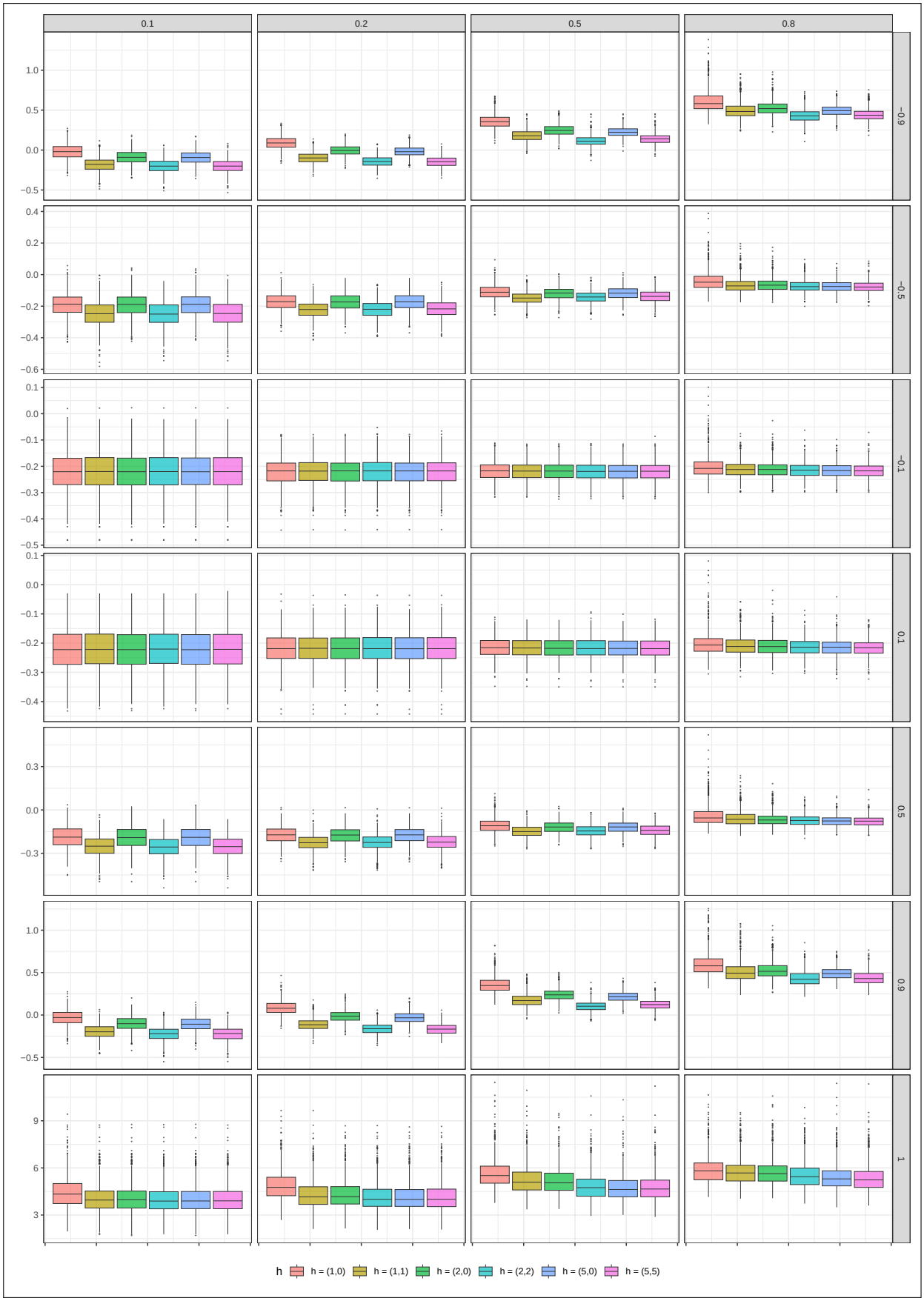
seguida, comparamos os resultados obtidos com árvores construídas com um valor específico de $\boldsymbol{h}$ e os outros métodos de imputação.

3.3 Resultados da imputação via árvores de decisão

Nesta seção descrevemos os resultados relativos às árvores de decisão. Com base nos resultados obtidos deseja-se verificar se existe relação entre o número de covariáveis utilizadas nas árvores de decisão e a qualidade de previsão em cada um dos cenários considerados. Nas Figuras 2 e 3 são apresentados os box-plots referentes à 1000 replicações do logaritmo do MAPE para os cenários em que as árvores de decisão foram obtidas considerando-se amostras de tamanho $n = 1000$ dos modelos AR(1) e MA(1), respectivamente. Os gráficos para os casos $n \in \{100, 500\}$ encontram-se no material suplementar. A Figura 4 corresponde ao modelo ARMA$(1, 1)$, para $n \in \{100, 500, 1000\}$. Em cada figura considera-se ainda a proporção de valores faltantes $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ e as defasagens $\boldsymbol{h} \in \{(1, 0), (1, 1), (2, 0), (2, 2), (5, 0), (5, 5)\}$. De forma geral observa-se que o MAPE aumenta com ρ e diminui com n. Esse comportamento é esperado dado que, quanto maior o tamanho da amostra de treinamento, melhor as árvores capturam a estrutura dos dados. Observa-se ainda que o MAPE aumenta com $|\phi|$ e $|\theta|$, sendo o aumento mais marcante no caso dos modelos AR, onde $|\phi| = 1$ corresponde à um processo não estacionário. No que segue, são descritos os resultados particulares para cada modelo, em termos de $\boldsymbol{h}$.

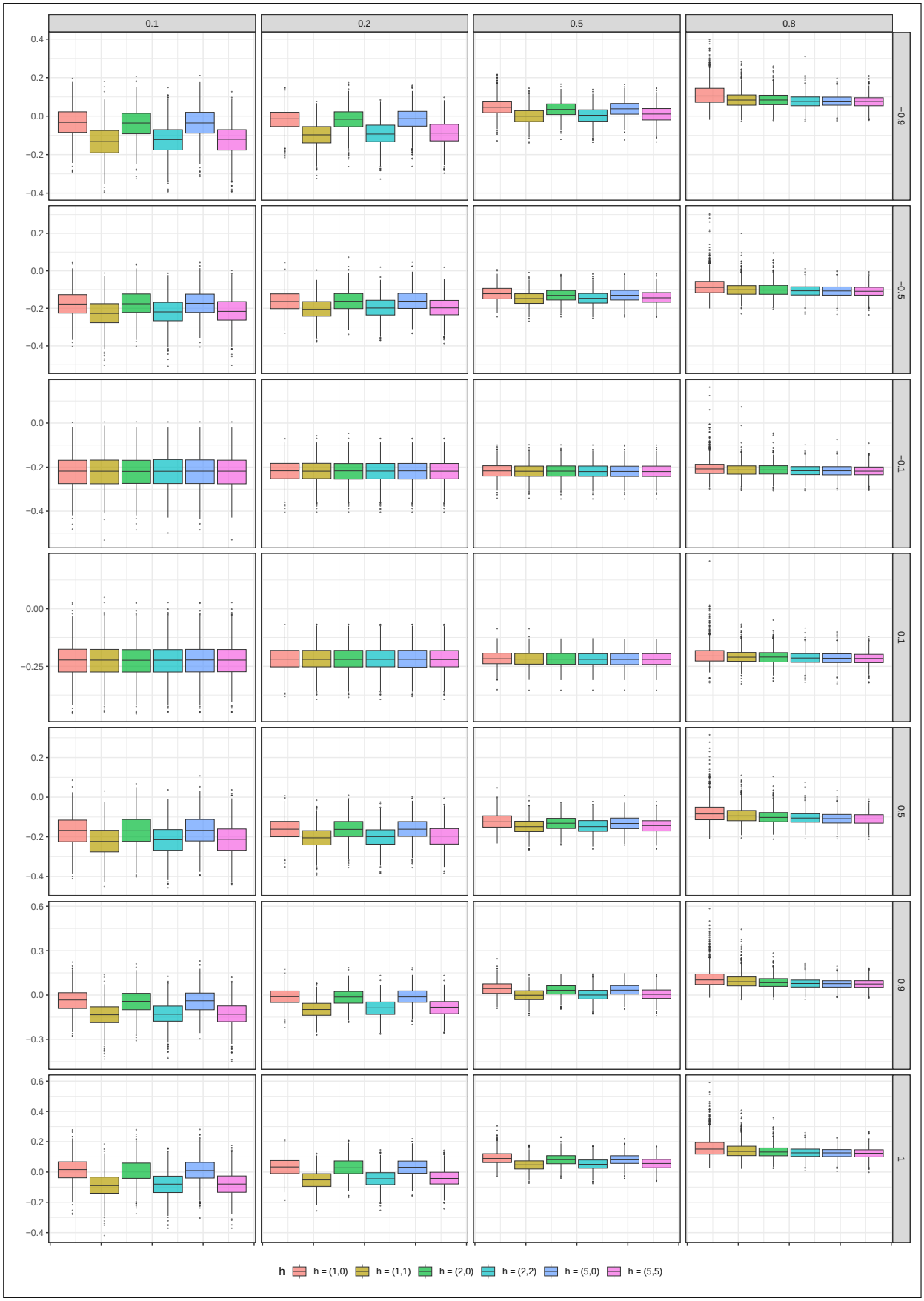
Para o modelo AR(1), quando a proporção de valores faltantes é muito alta ($\rho = 0.8$), embora o valor $\boldsymbol{h} = (5, 5)$ tenha produzido resultados ligeiramente melhores em alguns cenários, não parece haver um valor de $\boldsymbol{h}$ que produza melhores previsões, independente do tamanho da amostra. De forma semelhante, quando $|\phi| = 0.1$ não há diferença aparente entre as previsões, independentemente dos valores de n, ρ e $\boldsymbol{h}$. Para $|\phi| \in \{0.5, 0.9\}$, os valores do MAPE são menores quando são utilizadas informações tanto do passado quanto do futuro da amostra, ou seja, quando $\boldsymbol{h} \in \{(1, 1), (2, 2), (5, 5)\}$, para qualquer valor de n, exceto para o caso em que $n = 100$ e $\phi = 0.5$. No contexto de não estacionariedade ($\phi = 1$) os valores das métricas diminuem conforme aumentamos o número de covariáveis no modelo e os melhores resultados foram obtidos quando $\boldsymbol{h} = (5, 5)$.

Figura 2 – Resultados para o modelo AR(1)



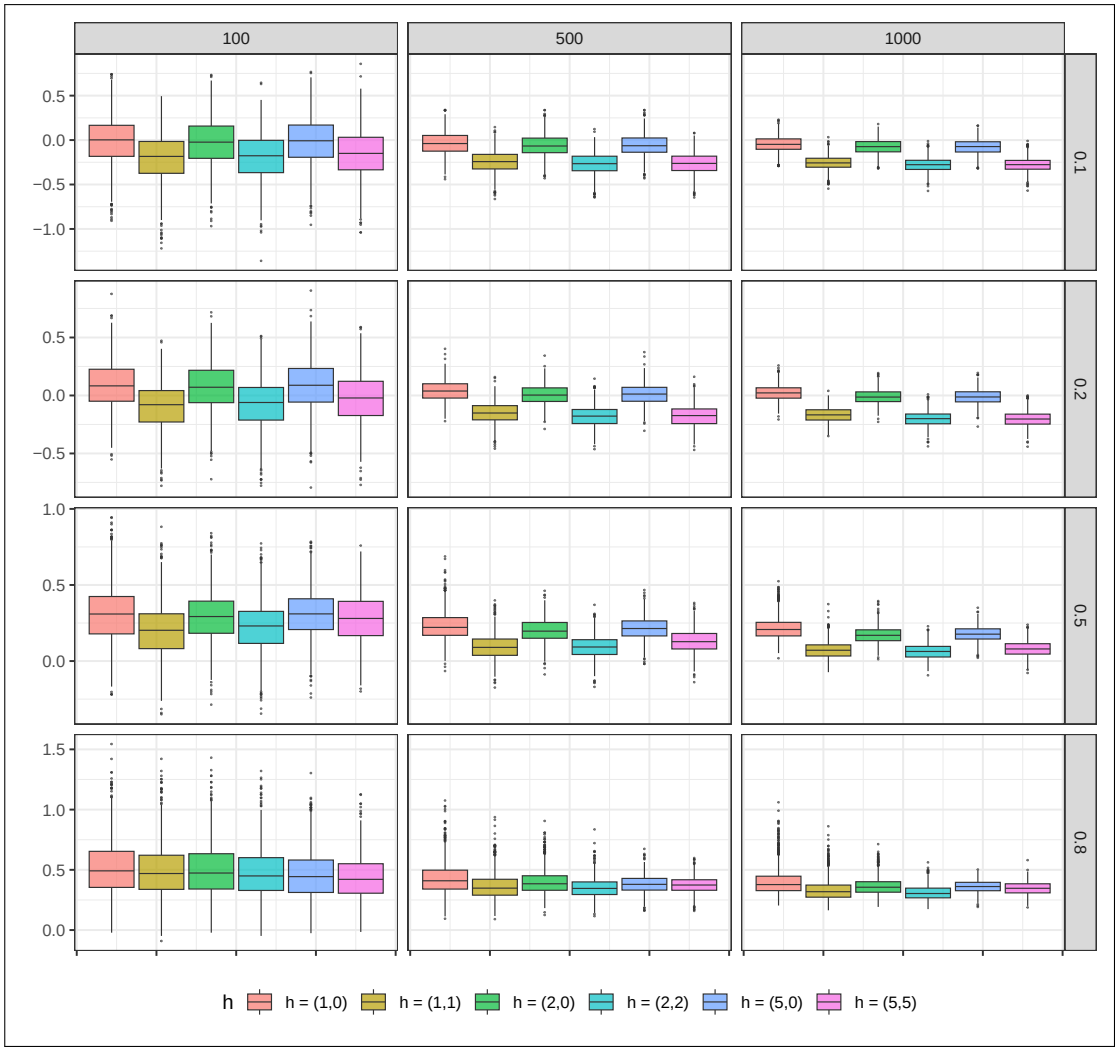
Legenda: Box-plots do logaritmo do MAPE do modelo AR(1) com $n = 1000$, $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel)

Figura 3 – Resultados para o modelo MA(1)



Legenda: Box-plots do logaritmo do MAPE do modelo MA(1) com $n = 1000$, $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel)

Figura 4 – Resultados para o modelo ARMA$(1,1)$



Legenda: Box-plots do logaritmo do MAPE do modelo temporais ARMA$(1,1)$ com $n \in \{100, 500, 1000\}$, $\phi = 0.7$, $\theta = 0.4$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$

Os resultados obtidos para o modelo MA(1) são muito parecidos com os do modelo AR(1). Quando $\rho = 0.8$, não existe diferença aparente entre os valores do MAPE para os diferentes valores de $\boldsymbol{h}$, independente do tamanho da amostra. De forma semelhante, quando $|\phi| = 0.1$ não há diferença aparente entre as previsões, independentemente dos valores de n, ρ e $\boldsymbol{h}$. Já para $|\phi| \geq 0.5$ não é possível notar diferença na qualidade das previsões quando $n = 100$ enquanto que para os tamanhos de amostras maiores, os valores do MAPE são menores quando $\boldsymbol{h} \in \{(1,1), (2,2).(5,5)\}$. Finalmente, para o modelo ARMA$(1,1)$, independente de n e ρ, os valores do MAPE são menores quando são utilizadas informações tanto do passado quanto do futuro da amostra, ou seja, quando $\boldsymbol{h} \in \{(1,1), (2,2), (5,5)\}$.

3.4 Resultados da comparações entre métodos

Em todos os cenários testados as árvores de decisão apresentaram melhor desempenho quando utilizavam informações tanto do passado quanto do futuro e (principalmente) no caso de não estacionariedade as previsões foram melhores quanto maior o número de variáveis explicativas. Portanto, para comparação com os demais métodos, considerou-se apenas árvores com $\boldsymbol{h} = (5, 5)$, ou seja, árvores em que foram utilizadas como covariáveis os 5 passos anteriores e posteriores à observação faltante. Para melhor visualização das tabelas e gráficos, os nomes dos métodos reportados foram reduzidos à siglas, conforme apresentado na Tabela 1.

Tabela 1 – Siglas utilizadas

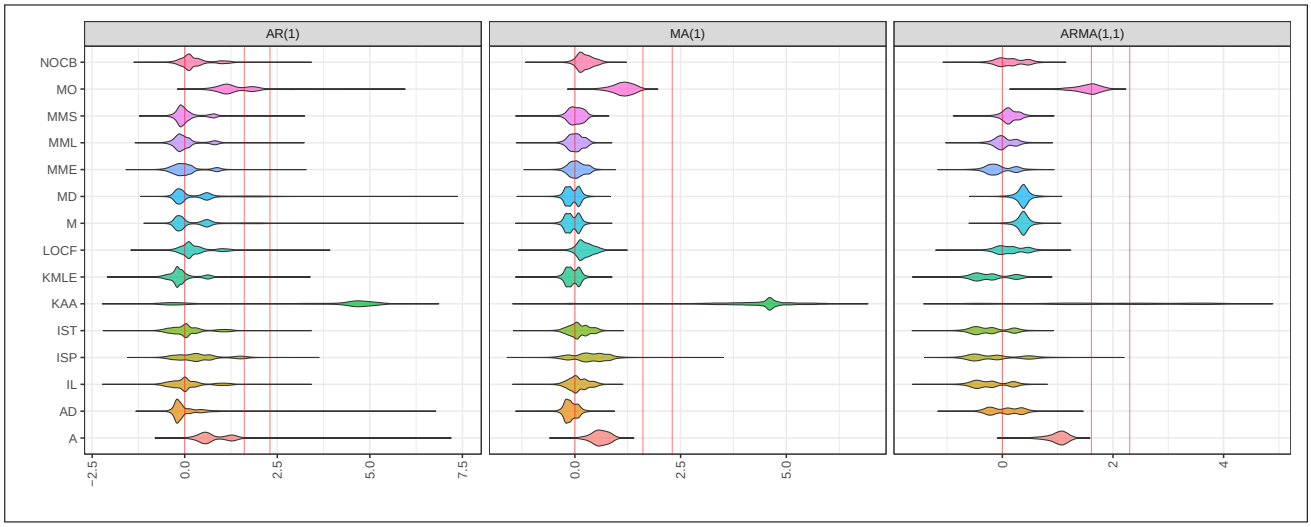
Sigla	Método	Sigla	Método
AD	árvores de decisão	LOCF	última observação levada adiante
M	média	NOCB	próxima observação trazida para trás
MD	mediana	A	método aleatório
MO	moda	KMLE	suavização de Kalman - modelo estrutural estimado via máxima verossimilhança
MMS	médias móveis - peso simples	KAA	suavização de Kalman - utiliza de uma representação do espaço de estados do modelo ARIMA com ajuste do modelo via função `auto.arima`
MML	médias móveis - peso linear		
MME	médias móveis - peso exponencial		
IL	interpolação linear		
ISP	interpolação por splines		
IST	interpolação de Stineman		

Legenda: Métodos de preenchimento de valores faltantes e as siglas correspondentes

Na Figura 5 são apresentados gráficos de violino para os valores do logaritmo do MAPE, para cada modelo e método considerados na simulação. Optou-se por essa escala para melhor visualização e comparação dos resultados. Os valores em questão são globais, desconsiderando-se os coeficientes do modelo, proporção de dados faltantes e tamanho amostral. Para fins de referência, as linhas vermelhas verticais correspondem aos casos MAPE $\in \{1\%, 5\%, 10\%\}$. Para auxiliar na interpretação dos resultados, na Tabela 2 são apresentadas estatísticas descritivas (Est.) para os valores de MAPE: mínimo (Min), média (M), máximo (Max), e os quantis 50%, 95% e 99%. Assim como nos gráficos, os valores são globais, desconsiderando-se os coeficientes do modelo, proporção de dados faltantes e tamanho amostral. O único caso em que o número de replicações é menor do que o planejado é para o método de suavização de Kalman com modelo estrutural estimado via máxima verossimilhança (KMLE) onde

ocorreram falhas no algoritmo resultado em NA. No total, ocorreram falhas em 22 cenários para o modelo AR e 26 para o modelo MA. O total de falhas por cenário variou de 1 replicação à 8 replicações sendo que em apenas 4 cenários ocorreram falhas em 5 ou mais replicações. Portanto, para o KMLE as estatísticas descritivas foram calculadas sobre as replicações que não falharam.

Figura 5 – Valores globais do MAPE



Legenda: Gráficos de violino para os valores do logaritmo do MAPE. Nos Valores globais, desconsiderando-se os coeficientes do modelo, proporção de dados faltantes e tamanho amostral. As linhas vermelhas verticais correspondem aos casos MAPE $\in \{1\%, 5\%, 10\%\}$

Pela Figura 5 e pela Tabela 2 observa-se que, de forma geral, os métodos apresentam o melhor desempenho para o modelo MA e o pior desempenho para o modelo AR. Além diso, os resultados para o modelo ARMA estão mais próximos dos resultados do modelo MA do que do AR. Dentre todos os métodos, o que apresentou o pior desempenho foi o método de suavização de Kalman que utiliza a função `auto.arima` (KAA), seguido dos métodos moda (MO) e aleatório (A). Em particular, para o KAA, o MAPE foi maior do que 10% em mais do que 50% das replicações realizadas para os modelos AR e MA. A média (M) e a mediana (MD) apresentaram comportamento misto no caso do modelo AR. Nesse caso, aproximadamente 50% das replicações resultaram em MAPE menor que 1%, porém, uma pequena proporção das replicações resultou em MAPE muito maior que 10%. Os métodos última observação levada adiante (LOCF) e próxima observação trazida para trás (NOCB) apresentam resultados bons e muito semelhantes entre si, apresentando MAPE maior que 3.87% e 3.85%, respectivamente, apenas no caso AR e para 1% das replicações. Interpolação

por spline (ISP), em geral, apresenta valores de MAPE ligeiramente maiores do que os outros dois métodos de interpolação (IL e IST), que por sua vez, são muito semelhantes entre si. Os métodos de médias móveis (MMS, MML e MME) são muito semelhantes entre si e, junto com o método suavização de Kalman que utiliza máxima verossimilhança (KMLE), estão entre os métodos com os menores valores de MAPE. Árvores de decisão (AD), que são o foco deste trabalho, apresentam um desempenho ligeiramente melhor do que KMLE, no contexto de modelos MA e semelhante aos métodos de médias móveis, no caso dos modelos AR e ARMA.

Tabela 2 – Estatísticas descritivas (Est.) para os valores de MAPE

	Est	A	AD	IL	ISP	IST	KAA	KMLE	LOCF	M	MD	MME	MML	MMS	MO	NOCB
AR	Min	0.45	0.27	0.11	0.21	0.11	0.11	0.12	0.23	0.33	0.31	0.20	0.26	0.29	0.82	0.25
	1%	1.05	0.59	0.46	0.52	0.46	0.47	0.46	0.64	0.62	0.63	0.51	0.55	0.60	1.50	0.64
	50%	1.99	0.89	1.00	1.36	1.03	94.37	0.85	1.17	0.97	0.98	0.96	0.94	0.96	3.53	1.17
	M	3.48	1.20	1.24	1.82	1.27	78.57	0.99	1.43	2.06	2.04	1.15	1.12	1.13	5.52	1.43
	95%	11.45	2.59	3.08	4.93	3.16	183.30	1.97	3.11	7.27	7.17	2.50	2.35	2.25	17.51	3.12
	99%	21.03	5.57	3.84	6.41	3.93	227.74	2.39	3.87	14.88	14.34	2.99	2.81	2.69	29.53	3.85
	Max	1333.39	880.07	30.54	37.54	30.47	960.64	29.69	50.44	1850.35	1582.18	26.63	25.13	25.50	384.47	30.51
MA	Min	0.55	0.25	0.23	0.20	0.23	0.23	0.25	0.26	0.25	0.26	0.30	0.25	0.25	0.84	0.31
	1%	1.03	0.61	0.61	0.57	0.60	0.79	0.61	0.77	0.61	0.62	0.63	0.64	0.65	1.48	0.77
	50%	1.83	0.89	1.05	1.43	1.07	100.00	0.93	1.24	0.93	0.93	1.04	1.02	1.02	3.18	1.24
	M	1.86	0.92	1.11	1.55	1.13	111.94	0.96	1.29	0.96	0.96	1.07	1.04	1.04	3.22	1.29
	95%	2.53	1.19	1.66	2.59	1.70	305.70	1.22	1.81	1.22	1.23	1.47	1.38	1.35	4.60	1.81
	99%	2.80	1.39	1.89	3.64	1.93	634.44	1.41	2.08	1.39	1.40	1.66	1.55	1.52	5.24	2.08
	Max	4.05	2.55	3.11	33.60	3.13	1025.79	2.41	3.45	2.40	2.31	2.62	2.41	2.23	7.14	3.38
ARMA	Min	0.91	0.31	0.20	0.24	0.20	0.24	0.20	0.30	0.55	0.55	0.31	0.36	0.41	1.14	0.34
	1%	1.55	0.59	0.41	0.38	0.40	0.49	0.41	0.64	0.94	0.94	0.53	0.62	0.70	2.08	0.62
	50%	2.81	1.06	0.76	0.79	0.76	7.34	0.77	1.12	1.46	1.46	0.89	1.00	1.14	4.81	1.13
	M	2.76	1.09	0.84	0.99	0.85	17.27	0.86	1.20	1.46	1.46	0.97	1.05	1.17	4.74	1.20
	95%	3.51	1.62	1.35	1.95	1.39	61.84	1.43	1.78	1.77	1.79	1.44	1.44	1.53	6.52	1.78
	99%	3.82	1.92	1.66	2.88	1.76	100.00	1.68	2.23	2.06	2.09	1.79	1.73	1.81	7.30	2.19
	Max	4.89	4.31	2.26	9.01	2.52	133.02	2.46	3.45	2.88	2.94	2.56	2.48	2.55	9.37	3.16

Caption: Estatísticas descritivas (Est.) para os valores de MAPE: mínimo (Min), média (M), máximo (Max), e os quantis 50%, 95% e 99%. Valores globais, desconsiderando-se os coeficientes do modelo, proporção de dados faltantes e tamanho amostral

Após analisar o desempenho global dos métodos de preenchimento, faremos uma descrição detalhada do comportamento dos mesmos, que permite identificar o contexto em que os métodos tendem a produzir os melhores/piores resultados. As Tabelas 3 e 4 apresentam a posição dos métodos de preenchimento de valores faltantes em termos de MAPE, considerando amostras tamanho $n = 1000$ de séries temporais AR(1) e MA(1), respectivamente. Nessas tabelas considera-se $\phi, \theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1.0\}$, $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ e $n = 1000$. Os resultados para $n \in \{100, 500\}$ encontram-se disponíveis no material suplementar. Os valores entre parêntesis, correspondem à média do MAPE para 1000 replicações.

Tabela 3 – Posição dos métodos em termos de MAPE para o modelo AR(1)

ϕ	ρ	Posição em termos de MAPE														
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º
-0.9	0.1	AD	M	MD	KMLE	MMS	MML	MME	IL	LOCF	NOCB	IST	A	ISP	MO	KAA
		(0.82)	(1.83)	(1.83)	(1.86)	(2.12)	(2.25)	(2.43)	(3.33)	(3.34)	(3.34)	(3.40)	(3.96)	(5.11)	(7.10)	(193.88)
	0.2	AD	M	MD	KMLE	MMS	MML	MME	LOCF	NOCB	IL	IST	A	ISP	MO	KAA
		(0.87)	(1.84)	(1.84)	(1.86)	(2.13)	(2.26)	(2.43)	(3.16)	(3.16)	(3.17)	(3.25)	(3.92)	(4.75)	(7.03)	(175.48)
	0.5	AD	M	MD	KMLE	MMS	MML	MME	IL	NOCB	LOCF	IST	A	ISP	MO	KAA
		(1.15)	(1.84)	(1.84)	(1.87)	(2.17)	(2.27)	(2.39)	(2.73)	(2.76)	(2.76)	(2.82)	(3.82)	(4.17)	(6.76)	(141.57)
	0.8	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	A	ISP	MO	KAA
		(1.56)	(1.84)	(1.84)	(1.88)	(2.20)	(2.26)	(2.37)	(2.40)	(2.47)	(2.56)	(2.56)	(3.57)	(4.30)	(6.18)	(111.80)
-0.5	0.1	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.78)	(0.92)	(0.92)	(0.93)	(1.00)	(1.06)	(1.14)	(1.45)	(1.48)	(1.54)	(1.55)	(2.04)	(2.04)	(3.70)	(172.12)
	0.2	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	ISP	A	MO	KAA
		(0.80)	(0.92)	(0.92)	(0.94)	(1.01)	(1.07)	(1.15)	(1.42)	(1.46)	(1.51)	(1.51)	(2.00)	(2.03)	(3.66)	(158.76)
	0.5	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.87)	(0.92)	(0.92)	(0.94)	(1.05)	(1.10)	(1.17)	(1.32)	(1.36)	(1.41)	(1.41)	(1.95)	(1.96)	(3.50)	(131.01)
	0.8	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	A	ISP	MO	KAA
		(0.92)	(0.93)	(0.93)	(0.94)	(1.12)	(1.16)	(1.23)	(1.24)	(1.28)	(1.34)	(1.34)	(1.81)	(2.13)	(3.15)	(107.97)
-0.1	0.1	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	ISP	A	MO	KAA
		(0.80)	(0.81)	(0.81)	(0.81)	(0.86)	(0.87)	(0.91)	(1.05)	(1.06)	(1.18)	(1.19)	(1.35)	(1.79)	(3.21)	(112.86)
	0.2	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.86)	(0.88)	(0.92)	(1.05)	(1.06)	(1.18)	(1.18)	(1.37)	(1.76)	(3.17)	(109.56)
	0.5	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.89)	(0.91)	(0.96)	(1.04)	(1.07)	(1.16)	(1.16)	(1.47)	(1.71)	(3.04)	(102.04)
	0.8	M	MD	AD	KMLE	MMS	MML	IL	MME	IST	NOCB	LOCF	A	ISP	MO	KAA
		(0.80)	(0.81)	(0.81)	(0.82)	(0.97)	(0.99)	(1.04)	(1.04)	(1.07)	(1.14)	(1.15)	(1.58)	(1.73)	(2.76)	(97.36)
0.1	0.1	M	MD	AD	KMLE	MML	MMS	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.81)	(0.81)	(0.81)	(0.81)	(0.84)	(0.84)	(0.85)	(0.93)	(0.94)	(1.08)	(1.09)	(1.15)	(1.78)	(3.23)	(78.90)
	0.2	M	MD	AD	KMLE	MML	MMS	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.84)	(0.84)	(0.86)	(0.93)	(0.95)	(1.09)	(1.09)	(1.18)	(1.77)	(3.19)	(82.12)
	0.5	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.88)	(0.88)	(0.91)	(0.96)	(0.98)	(1.10)	(1.11)	(1.31)	(1.70)	(3.05)	(87.25)
	0.8	M	AD	MD	KMLE	MMS	MML	IL	MME	IST	LOCF	NOCB	A	ISP	MO	KAA
		(0.81)	(0.81)	(0.81)	(0.82)	(0.95)	(0.96)	(1.00)	(1.02)	(1.03)	(1.12)	(1.12)	(1.57)	(1.62)	(2.75)	(92.56)
0.5	0.1	IL	IST	MME	KMLE	AD	MML	MMS	ISP	M	MD	NOCB	LOCF	A	MO	KAA
		(0.74)	(0.75)	(0.76)	(0.77)	(0.78)	(0.80)	(0.85)	(0.87)	(0.92)	(0.92)	(0.94)	(0.94)	(2.04)	(3.69)	(25.86)
	0.2	IL	IST	MME	KMLE	AD	MML	MMS	ISP	M	MD	LOCF	NOCB	A	MO	KAA
		(0.77)	(0.78)	(0.78)	(0.80)	(0.81)	(0.81)	(0.86)	(0.91)	(0.92)	(0.92)	(0.97)	(0.97)	(2.02)	(3.64)	(29.80)
	0.5	MME	IL	MML	IST	AD	KMLE	MMS	M	MD	NOCB	LOCF	ISP	A	MO	KAA
		(0.85)	(0.85)	(0.86)	(0.87)	(0.87)	(0.88)	(0.90)	(0.92)	(0.92)	(1.05)	(1.06)	(1.09)	(1.95)	(3.48)	(44.37)
	0.8	M	MD	AD	KMLE	MML	IL	MMS	IST	MME	LOCF	NOCB	ISP	A	MO	KAA
		(0.92)	(0.92)	(0.93)	(0.93)	(0.99)	(1.00)	(1.01)	(1.03)	(1.03)	(1.18)	(1.19)	(1.50)	(1.81)	(3.15)	(68.46)
0.9	0.1	IL	KMLE	IST	MME	ISP	MML	AD	LOCF	NOCB	MMS	MD	M	KAA	A	MO
		(0.61)	(0.62)	(0.62)	(0.69)	(0.70)	(0.77)	(0.81)	(0.85)	(0.85)	(0.88)	(1.82)	(1.82)	(2.99)	(3.78)	(6.75)
	0.2	IL	KMLE	IST	MME	ISP	MML	AD	MMS	LOCF	NOCB	MD	M	A	KAA	MO
		(0.64)	(0.64)	(0.64)	(0.70)	(0.73)	(0.78)	(0.85)	(0.88)	(0.89)	(0.89)	(1.81)	(1.81)	(3.75)	(5.44)	(6.68)
	0.5	IL	KMLE	IST	MME	MML	ISP	MMS	LOCF	NOCB	AD	M	MD	KAA	A	MO
		(0.75)	(0.75)	(0.76)	(0.80)	(0.85)	(0.90)	(0.94)	(1.07)	(1.07)	(1.14)	(1.81)	(1.81)	(3.59)	(3.68)	(6.48)
	0.8	IL	KMLE	IST	MML	MME	MMS	ISP	NOCB	LOCF	AD	M	MD	A	MO	KAA
		(1.05)	(1.06)	(1.08)	(1.12)	(1.12)	(1.19)	(1.40)	(1.48)	(1.48)	(1.55)	(1.82)	(1.82)	(3.47)	(5.98)	(15.17)
1.0	0.1	IL	KMLE	KAA	IST	MME	ISP	MML	LOCF	NOCB	MMS	AD	MD	M	A	MO
		(0.62)	(0.62)	(0.62)	(0.63)	(0.70)	(0.70)	(0.80)	(0.89)	(0.89)	(0.91)	(1.99)	(10.85)	(11.13)	(16.81)	(23.77)
	0.2	KMLE	IL	IST	MME	ISP	KAA	MML	MMS	NOCB	LOCF	AD	MD	M	A	MO
		(0.67)	(0.67)	(0.67)	(0.74)	(0.77)	(0.77)	(0.82)	(0.93)	(0.94)	(0.97)	(2.10)	(11.98)	(12.45)	(17.37)	(23.90)
	0.5	IL	KMLE	IST	MME	MML	ISP	MMS	KAA	NOCB	LOCF	AD	MD	M	A	MO
		(0.77)	(0.77)	(0.78)	(0.83)	(0.89)	(0.92)	(0.99)	(1.07)	(1.16)	(1.17)	(3.54)	(11.37)	(11.76)	(17.50)	(23.53)
	0.8	IL	KMLE	IST	MME	MML	MMS	ISP	LOCF	NOCB	KAA	AD	MD	M	A	MO
		(1.11)	(1.11)	(1.14)	(1.23)	(1.23)	(1.34)	(1.44)	(1.76)	(1.78)	(3.00)	(6.59)	(11.25)	(11.58)	(16.85)	(22.99)

Legenda: Posição dos métodos de preenchimento de valores faltantes em termos de MAPE, considerando amostras tamanho $n = 1000$ de séries temporais AR(1) com $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1.0\}$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. Os valores entre parêntesis, correspondem à média do MAPE para as 1000 replicações

Tabela 4 – Posição dos métodos em termos de MAPE para o modelo MA(1)

θ	ρ	Posição em termos de MAPE														
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º
-0.9	0.1	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.88)	(1.08)	(1.08)	(1.09)	(1.19)	(1.27)	(1.37)	(1.68)	(1.70)	(1.82)	(1.82)	(2.23)	(2.38)	(4.31)	(505.57)
	0.2	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.92)	(1.08)	(1.08)	(1.09)	(1.20)	(1.28)	(1.37)	(1.66)	(1.69)	(1.79)	(1.79)	(2.25)	(2.36)	(4.25)	(313.72)
	0.5	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	A	ISP	MO	KAA
		(1.01)	(1.07)	(1.07)	(1.09)	(1.25)	(1.31)	(1.41)	(1.58)	(1.62)	(1.69)	(1.69)	(2.28)	(2.30)	(4.09)	(161.58)
	0.8	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	A	ISP	MO	KAA
		(1.08)	(1.08)	(1.08)	(1.10)	(1.33)	(1.37)	(1.47)	(1.47)	(1.52)	(1.58)	(1.59)	(2.11)	(2.55)	(3.67)	(112.25)
-0.5	0.1	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	ISP	A	MO	KAA
		(0.81)	(0.89)	(0.89)	(0.91)	(0.98)	(1.03)	(1.10)	(1.34)	(1.36)	(1.46)	(1.47)	(1.77)	(1.98)	(3.59)	(241.10)
	0.2	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	LOCF	NOCB	ISP	A	MO	KAA
		(0.82)	(0.89)	(0.89)	(0.91)	(0.99)	(1.04)	(1.11)	(1.33)	(1.35)	(1.45)	(1.45)	(1.79)	(1.97)	(3.55)	(207.46)
	0.5	AD	M	MD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.87)	(0.89)	(0.89)	(0.91)	(1.03)	(1.07)	(1.15)	(1.28)	(1.31)	(1.38)	(1.38)	(1.84)	(1.89)	(3.37)	(142.30)
	0.8	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	A	ISP	MO	KAA
		(0.89)	(0.90)	(0.90)	(0.91)	(1.09)	(1.13)	(1.20)	(1.21)	(1.24)	(1.30)	(1.30)	(1.76)	(2.08)	(3.07)	(106.67)
-0.1	0.1	M	AD	MD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.86)	(0.87)	(0.91)	(1.05)	(1.06)	(1.18)	(1.18)	(1.35)	(1.78)	(3.21)	(113.21)
	0.2	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.86)	(0.88)	(0.92)	(1.04)	(1.06)	(1.18)	(1.18)	(1.36)	(1.77)	(3.18)	(109.48)
	0.5	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.90)	(0.92)	(0.96)	(1.04)	(1.07)	(1.16)	(1.16)	(1.46)	(1.70)	(3.04)	(102.97)
	0.8	M	MD	AD	KMLE	MMS	MML	IL	MME	IST	NOCB	LOCF	A	ISP	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.82)	(0.96)	(0.98)	(1.04)	(1.04)	(1.07)	(1.15)	(1.15)	(1.58)	(1.73)	(2.78)	(98.02)
0.1	0.1	M	MD	AD	KMLE	MML	MMS	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.83)	(0.84)	(0.85)	(0.92)	(0.94)	(1.08)	(1.08)	(1.15)	(1.78)	(3.22)	(80.46)
	0.2	M	MD	AD	KMLE	MML	MMS	MME	IL	IST	LOCF	NOCB	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.84)	(0.84)	(0.86)	(0.93)	(0.95)	(1.09)	(1.09)	(1.18)	(1.77)	(3.19)	(83.97)
	0.5	M	MD	AD	KMLE	MMS	MML	MME	IL	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.80)	(0.80)	(0.80)	(0.81)	(0.88)	(0.88)	(0.91)	(0.96)	(0.99)	(1.11)	(1.11)	(1.31)	(1.71)	(3.06)	(88.75)
	0.8	M	AD	MD	KMLE	MMS	MML	IL	MME	IST	NOCB	LOCF	A	ISP	MO	KAA
		(0.80)	(0.81)	(0.81)	(0.82)	(0.95)	(0.96)	(1.00)	(1.02)	(1.03)	(1.12)	(1.12)	(1.58)	(1.62)	(2.76)	(92.65)
0.5	0.1	IL	IST	AD	ISP	MME	MML	KMLE	M	MD	MMS	LOCF	NOCB	A	MO	KAA
		(0.78)	(0.78)	(0.81)	(0.82)	(0.82)	(0.85)	(0.88)	(0.89)	(0.89)	(0.90)	(1.01)	(1.01)	(1.98)	(3.58)	(37.96)
	0.2	IL	IST	AD	MME	MML	KMLE	M	MD	ISP	MMS	LOCF	NOCB	A	MO	KAA
		(0.81)	(0.81)	(0.82)	(0.84)	(0.86)	(0.88)	(0.89)	(0.89)	(0.90)	(0.90)	(1.03)	(1.04)	(1.96)	(3.54)	(42.94)
	0.5	AD	M	MD	KMLE	MML	MME	IL	IST	MMS	NOCB	LOCF	ISP	A	MO	KAA
		(0.87)	(0.89)	(0.89)	(0.89)	(0.91)	(0.91)	(0.91)	(0.93)	(0.94)	(1.12)	(1.12)	(1.16)	(1.90)	(3.39)	(61.25)
	0.8	M	MD	AD	KMLE	MML	MMS	IL	MME	IST	NOCB	LOCF	ISP	A	MO	KAA
		(0.89)	(0.90)	(0.90)	(0.91)	(1.03)	(1.03)	(1.04)	(1.07)	(1.08)	(1.20)	(1.21)	(1.62)	(1.76)	(3.08)	(81.62)
0.9	0.1	ISP	IST	IL	AD	MME	MML	KMLE	MMS	M	MD	LOCF	NOCB	A	MO	KAA
		(0.73)	(0.80)	(0.81)	(0.89)	(0.95)	(1.00)	(1.02)	(1.07)	(1.07)	(1.07)	(1.12)	(1.12)	(2.37)	(4.27)	(18.55)
	0.2	ISP	IST	IL	AD	MME	MML	KMLE	M	MD	MMS	LOCF	NOCB	A	MO	KAA
		(0.85)	(0.86)	(0.87)	(0.92)	(0.96)	(1.01)	(1.05)	(1.07)	(1.07)	(1.08)	(1.16)	(1.17)	(2.35)	(4.23)	(28.18)
	0.5	AD	IL	IST	MME	KMLE	MML	M	MD	MMS	ISP	LOCF	NOCB	A	MO	KAA
		(1.01)	(1.03)	(1.04)	(1.06)	(1.07)	(1.07)	(1.07)	(1.07)	(1.12)	(1.25)	(1.30)	(1.30)	(2.27)	(4.05)	(52.95)
	0.8	M	MD	AD	KMLE	MML	IL	MMS	IST	MME	NOCB	LOCF	ISP	A	MO	KAA
		(1.08)	(1.08)	(1.08)	(1.09)	(1.22)	(1.23)	(1.23)	(1.26)	(1.27)	(1.44)	(1.44)	(1.85)	(2.11)	(3.68)	(76.69)
1	0.1	ISP	IST	IL	AD	MME	MML	KMLE	MMS	M	MD	LOCF	NOCB	A	MO	KAA
		(0.75)	(0.84)	(0.85)	(0.92)	(0.99)	(1.05)	(1.06)	(1.12)	(1.13)	(1.13)	(1.17)	(1.18)	(2.51)	(4.55)	(19.14)
	0.2	ISP	IST	IL	AD	MME	MML	KMLE	M	MD	MMS	LOCF	NOCB	A	MO	KAA
		(0.88)	(0.90)	(0.91)	(0.96)	(1.02)	(1.07)	(1.10)	(1.13)	(1.13)	(1.14)	(1.22)	(1.22)	(2.49)	(4.51)	(28.18)
	0.5	AD	IL	IST	MME	MML	KMLE	M	MD	MMS	ISP	NOCB	LOCF	A	MO	KAA
		(1.06)	(1.08)	(1.09)	(1.11)	(1.12)	(1.13)	(1.13)	(1.13)	(1.18)	(1.31)	(1.36)	(1.36)	(2.40)	(4.28)	(52.36)
	0.8	M	MD	AD	KMLE	MML	IL	MMS	IST	MME	NOCB	LOCF	ISP	A	MO	KAA
		(1.13)	(1.13)	(1.13)	(1.14)	(1.28)	(1.29)	(1.29)	(1.32)	(1.33)	(1.50)	(1.51)	(1.93)	(2.22)	(3.85)	(76.74)

Legenda: Posição dos métodos de preenchimento de valores faltantes em termos de MAPE, considerando amostras tamanho $n = 1000$ de séries temporais MA(1) com $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1.0\}$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. Os valores entre parêntesis, correspondem à média do MAPE para as 1000 replicações

Tabela 5 – Posição dos métodos em termos de MAPE para o modelo ARMA$(1,1)$

ρ	Est.	Posição em termos de MAPE														
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º
								$n = 100$								
0.1	Método	ISP	IST	IL	KMLE	MME	AD	MML	LOCF	NOCB	MMS	M	MD	A	KAA	MO
	MAPE	0.57	0.59	0.60	0.61	0.78	0.87	0.91	0.93	0.93	1.06	1.42	1.42	2.49	3.30	4.12
	MSE	0.54	0.56	0.58	0.58	0.97	1.24	1.32	1.39	1.41	1.76	3.16	3.19	9.27	32.49	20.95
	BIAS	-0.01	-0.01	-0.01	-0.01	0.00	-0.02	0.00	0.00	-0.02	0.00	0.01	0.01	-0.01	-3.07	-4.14
0.2	Método	IST	ISP	IL	KMLE	MME	MML	NOCB	LOCF	AD	MMS	M	MD	A	MO	KAA
	MAPE	0.64	0.64	0.65	0.65	0.81	0.93	0.98	0.99	0.99	1.07	1.43	1.43	2.46	4.08	5.15
	MSE	0.67	0.70	0.69	0.70	1.04	1.37	1.62	1.61	1.62	1.80	3.21	3.25	9.12	20.68	97.01
	BIAS	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-4.10	-4.92
0.5	Método	IL	IST	KMLE	MME	ISP	MML	MMS	NOCB	LOCF	AD	M	MD	A	MO	KAA
	MAPE	0.87	0.87	0.89	0.95	0.97	1.03	1.14	1.24	1.25	1.36	1.44	1.45	2.37	3.86	16.41
	MSE	1.29	1.30	1.34	1.51	1.75	1.74	2.10	2.64	2.68	2.97	3.28	3.33	8.52	18.89	912.99
	BIAS	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	-0.01	0.00	-0.01	0.00	-0.01	-0.02	-3.87	-16.03
0.8	Método	IL	KMLE	IST	MML	MME	MMS	M	MD	AD	NOCB	LOCF	ISP	A	MO	KAA
	MAPE	1.38	1.39	1.42	1.43	1.47	1.49	1.51	1.53	1.56	1.74	1.75	2.08	2.18	3.24	51.44
	MSE	3.15	3.19	3.40	3.38	3.61	3.61	3.56	3.68	3.87	4.96	5.04	8.78	7.32	14.32	4743.41
	BIAS	0.01	0.01	0.01	0.01	0.01	0.02	0.00	-0.01	0.02	0.02	0.00	-0.05	0.00	-3.18	-50.76
								$n = 500$								
0.1	Método	ISP	IST	IL	KMLE	AD	MME	MML	NOCB	LOCF	MMS	MD	M	A	KAA	MO
	MAPE	0.57	0.59	0.60	0.60	0.77	0.79	0.92	0.93	0.93	1.07	1.46	1.46	2.99	4.94	5.25
	MSE	0.53	0.56	0.58	0.58	0.96	0.99	1.35	1.39	1.39	1.80	3.34	3.34	13.01	106.46	31.81
	BIAS	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.02	-4.83	-5.28
0.2	Método	ISP	IST	IL	KMLE	MME	AD	MML	LOCF	NOCB	MMS	M	MD	A	MO	KAA
	MAPE	0.63	0.64	0.65	0.65	0.81	0.83	0.94	0.99	0.99	1.08	1.46	1.46	2.96	5.21	7.84
	MSE	0.68	0.66	0.68	0.68	1.04	1.13	1.38	1.6	1.61	1.82	3.33	3.34	12.80	31.36	263.66
	BIAS	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.00	-0.01	0.00	0.00	-0.01	-0.01	-5.24	-7.72
0.5	Método	IST	IL	KMLE	ISP	MME	MML	MMS	AD	LOCF	NOCB	M	MD	A	MO	KAA
	MAPE	0.84	0.84	0.84	0.92	0.93	1.01	1.13	1.15	1.21	1.22	1.46	1.46	2.87	4.99	14.71
	MSE	1.20	1.19	1.21	1.57	1.44	1.67	2.05	2.15	2.51	2.52	3.36	3.37	12.07	29.14	683.47
	BIAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	-0.01	-5.02	-14.54
0.8	Método	IL	IST	KMLE	MML	MME	MMS	AD	M	MD	LOCF	NOCB	ISP	A	MO	KAA
	MAPE	1.23	1.26	1.29	1.29	1.30	1.36	1.46	1.47	1.47	1.61	1.61	1.66	2.65	4.48	38.39
	MSE	2.58	2.71	2.73	2.8	2.89	3.05	3.36	3.37	3.40	4.35	4.37	5.48	10.47	24.21	2558.49
	BIAS	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.01	-4.50	-38.12
								$n = 1000$								
0.1	Método	ISP	IST	IL	KMLE	AD	MME	MML	LOCF	NOCB	MMS	M	MD	A	KAA	MO
	MAPE	0.58	0.60	0.61	0.61	0.76	0.79	0.92	0.93	0.93	1.07	1.46	1.46	3.17	5.05	5.70
	MSE	0.54	0.57	0.59	0.59	0.93	0.99	1.35	1.40	1.41	1.79	3.36	3.36	14.52	111.62	36.70
	BIAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.03	-4.97	-5.73
0.2	Método	ISP	IST	IL	KMLE	MME	AD	MML	LOCF	NOCB	MMS	M	MD	A	MO	KAA
	MAPE	0.64	0.64	0.65	0.65	0.81	0.82	0.94	0.99	0.99	1.08	1.46	1.46	3.14	5.62	7.91
	MSE	0.69	0.67	0.69	0.69	1.05	1.09	1.39	1.61	1.61	1.83	3.36	3.36	14.28	35.81	266.67
	BIAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.03	-5.65	-7.81
0.5	Método	IST	IL	KMLE	ISP	MME	MML	AD	MMS	LOCF	NOCB	M	MD	A	MO	KAA
	MAPE	0.83	0.84	0.84	0.92	0.93	1.01	1.09	1.13	1.21	1.22	1.46	1.47	3.05	5.43	15.20
	MSE	1.19	1.19	1.20	1.56	1.43	1.67	1.94	2.05	2.50	2.52	3.36	3.37	13.57	33.70	717.58
	BIAS	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	-5.46	-15.05
0.8	Método	IL	IST	MML	MME	KMLE	MMS	AD	M	MD	NOCB	LOCF	ISP	A	MO	KAA
	MAPE	1.22	1.24	1.28	1.29	1.31	1.36	1.41	1.47	1.47	1.60	1.60	1.64	2.84	4.92	36.92
	MSE	2.52	2.65	2.75	2.83	2.77	3.01	3.16	3.37	3.38	4.29	4.31	5.41	11.87	28.37	2273.22
	BIAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	-4.95	-36.74

Legenda: Posição dos métodos de preenchimento de valores faltantes em termos de MAPE, considerando amostras tamanho $n = 1000$ de séries temporais ARMA$(1,1)$ com $(\phi, \theta) = (0.7, 0.4)$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. Os valores do MAPE, correspondem à média para as 1000 replicações. Em cada caso, são apresentados ainda as médias do erro quadrático médio (MSE - *Mean Square Error*) e do vício (BIAS - *Bias*) associados a cada método

Na Tabela 3 observa-se que, para os modelo AR com $-0.9 \leq \phi \leq 0.1$, a média, a mediana e as árvores de decisão sempre aparecem como os três primeiros colocados (não necessariamente nessa ordem), o KMLE aparece na 4ª posição, seguido dos métodos MMS, MML e MME, não necessariamente nessa ordem, exceto para $\phi = 0.1$ e $\rho = 0.8$, caso em que a interpolação linear (IL) aparece em 7º lugar, seguido do MME. Para $\phi = 0.5$ as árvores de decisão (AD) ainda aparecem entre os modelos mais competitivos, mas agora acompanhadas de outros métodos (M, MD, IL, IST, ISP, MME, MML, MMS e KMLE (não necessariamente nessa ordem). Para $\phi = 0.9$ as árvores de decisão (AD) estão na 7ª posição para $\rho \in \{0.1, 0.2\}$ e na 10ª posição para $\rho \in \{0.5, 0.8\}$. Ainda assim, apresentam um bom desempenho com MAPE sempre abaixo de 2%. No caso $\rho = 1$ as árvores de decisão aparecem em 11ª lugar, com MAPE $\in \{1.99\%, 2.10\%, 3.54\%, 6.59\%\}$, respectivamente, para $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. Nesse caso, para os 10 primeiros colocados o MAPE sempre foi menor ou igual a 3% enquanto que o MAPE para os 4 últimos colocados (MD, M, A e MO) sempre foi maior ou igual a 10.85%.

Da Tabela 4 conclui-se que, para o modelo MA, árvores de decisão sempre aparecem entre os quatro melhores modelos, junto com a média, mediana e KMLE (não necessariamente nessa ordem), quando $\phi \leq 0.1$. Quando $\phi \geq 0.5$ aparecem ainda os métodos baseados em interpolação (IL, IST e ISP) e o método de médias móveis com peso exponencial (MME). Esses métodos citados, junto com os outros dois métodos de médias móveis (MMK e MMS) são os únicos que ocupam as 9 primeiras posições.

A Tabela 5 apresenta os resultados referentes ao modelo ARMA$(1, 1)$. Nessa tabela, além da posição dos métodos e do MAPE correspondente, são apresentados ainda o erro quadrático médio (MS) e o vício (B). De forma geral observa-se que, como esperado, o MAPE aumenta com ρ e diminui com n para todos os métodos e modelos. Além disso, os piores desempenhos foram observados para método de suavização de Kalman que utiliza a função `auto.arima` (KAA), o método aleatório (A), a moda (MO) e, em alguns cenários, para interpolação por spline (ISP). A Tabela 5 mostra que, para o modelo ARMA as árvores de decisão aparecem entre a 5ª e a 10ª posição, com valores de MAPE variando de 0.77% a 1.45%. Com excessão dos métodos A, MD e KAA, todos os outros aparecem, pelo menos uma vez, entre os 8 primeiros. Mesmo não ocupando as melhes posições, o desempenho das árvores pode ser considerado satisfatório, tendo

em vista que o menor valor de MAPE obtido foi igual a 0.57%, para o método de interpolação por spline (ISP), enquanto que o maior valor de MAPE entre os quatro primeiros colocados foi 1.43% para o método de médias móveis com peso linear (MML).

4 APLICATIVO SHINY

O *shiny* (Chang et al., 2021) é um pacote do *R* que permite a construção de aplicativos interativos para web. Os aplicativos criados utilizando o `shiny` são compostos por 3 componentes: a interface do usuário (UI), que controla a aparência do aplicativo para o usuário; o *server*, que contém as funcionalidades do aplicativo; e o ShinyApp que cria o aplicativo Shiny. Essa ferramenta pode ser utilizada para o ensino de estatística, visualização e análise de dados e até criação de jogos (alguns exemplos podem ser acessados em https://shiny.rstudio.com/gallery/). Neste trabalho, foi criado um aplicativo com dois objetivos: auxiliar na tomada de decisão para delimitar os cenários a serem testados nos estudos de simulação (Aba 1) e desenvolver uma interface amigável ao usuário para preenchimento de valores faltantes de séries temporais reais (Aba 2).

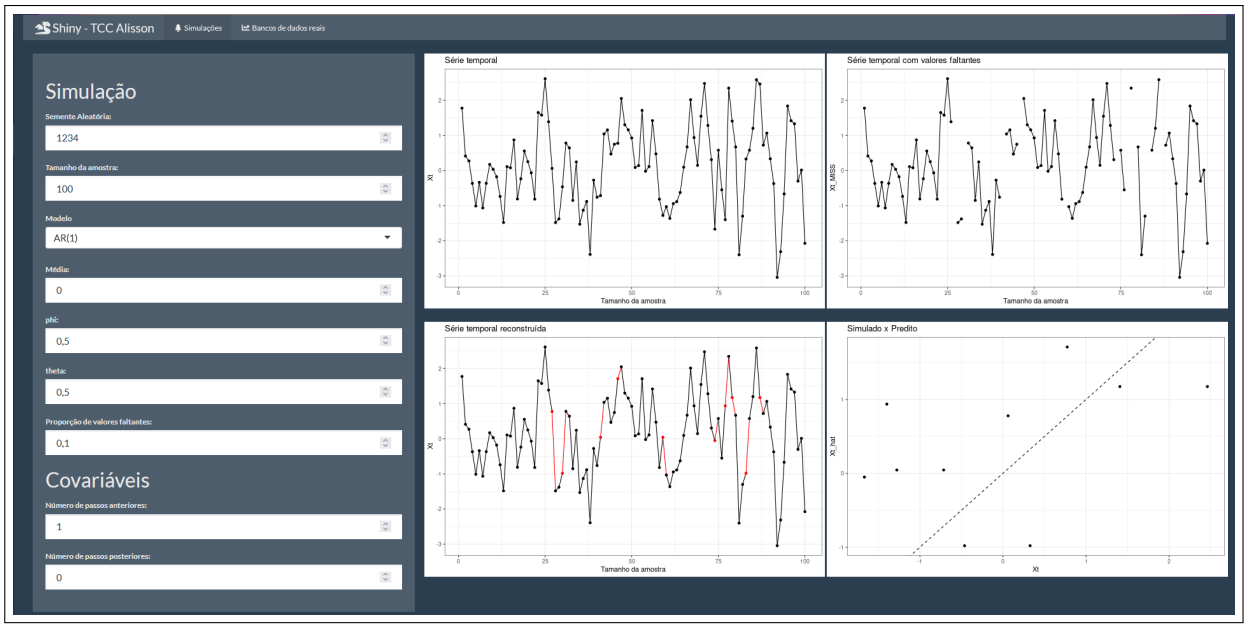
Na Figura 6 é apresentada a Aba 1 do aplicativo. No painel a esquerda é possível alterar os argumentos utilizados para simular as séries temporais com valores faltantes e a quantidade de covariáveis utilizadas nas previsões (os outros argumentos utilizados na modelagem de árvores de decisão são os mesmos da Seção 3.2). No painel principal são apresentados quatro gráficos, a série temporal $\{X_t\}_{t=1}^n$, a série temporal com os valores faltantes $\{X_t^{\text{miss}}\}_{t=1}^n$, a série temporal reconstruída e os valores simulados versus os preditos (da esquerda para direita, cima para baixo).

Na Figura 7 é apresentada a Aba 2 do aplicativo. No painel a esquerda é possível inserir o banco de dados com a série temporal com valores faltantes e alterar os argumentos utilizados pelo algoritmo de árvores de decisão. No painel principal é apresentado o gráfico da série temporal reconstruída e um botão com a opção para baixar o banco de dados com a série temporal preenchida.

Aplicativos Shiny podem ser executados localmente, através do *RStudio* (RStudio Team, 2022), ou em um servidor online, acessando o aplicativo através de um navegador web por qualquer dispositivo com acesso à internet. O aplicativo

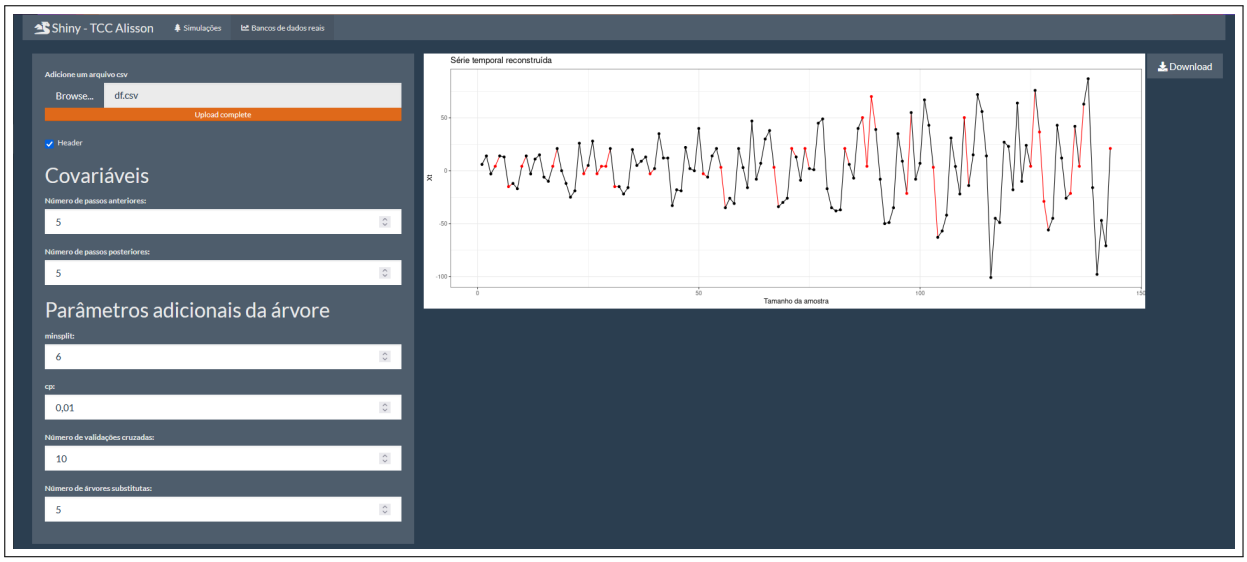
desenvolvido neste trabalho está disponível online, no servidor do *RStudio* para aplicativos Shiny, o *Shinyapps*, no link https://neimaier.shinyapps.io/TCC$_{trees}$/.

Figura 6 – Aba 1 do aplicativo Shiny



Legenda: Aba 1 do aplicativo Shiny, com opções para simular séries temporais AR(1), MA(1) e ARMA$(1,1)$ e reconstruí-las utilizando o método proposto

Figura 7 – Aba 2 do aplicativo Shiny



Legenda: Aba 2 do aplicativo Shiny, em que é possível fazer o *upload* de uma série temporal real que pode ser reconstruída pelo usuário utilizando o método proposto neste trabalho e posteriormente baixada

5 CONCLUSÕES

Neste trabalho foi proposta uma metodologia de preenchimento de dados faltantes em séries temporais baseada em árvores de decisão. Via simulações de Monte Carlo, foi estudado o preenchimento de valores faltantes em séries temporais de modelos AR(1), MA(1), ARMA$(1,1)$ e passeio aleatório, utilizando o método proposto. Também foi comparado o desempenho do método proposto, com métodos clássicos da literatura. Também foi desenvolvida uma ferramenta amigável ao usuário para reconstrução de séries temporais com o método proposto utilizando do pacote `Shiny` do `R`.

Observou-se que a qualidade de reconstrução das séries temporais utilizando árvores de decisão é afetada pela quantidade de covariáveis utilizadas na modelagem. Em particular, os melhores resultados foram obtidos quando foram utilizadas como covariáveis tanto variáveis anteriores quanto posteriores às observações definidas como resposta. Embora os modelos AR, MA e ARMA testados sejam causais e informações sobre o passado deveriam ser suficientes, o resultado obtido não nos surpreende dado que as previsões via árvores de decisão baseiam-se em média condicional.

No contexto de modelos estacionários, as árvores de decisão foram consideradas competitivas. Mesmo no cenário em que ficaram na 10ª posição, os valores de MAPE fica relativamente próximo do MAPE dos primeiros colocados. No caso do passeio aleatório, (modelo AR com $\phi = 1$) as árvores apresentaram um desempenho inferior quando comparadas aos primeiros colocados. Entretanto, mesmo no pior cenário ($\rho = 0.8$), o MAPE ainda está bem abaixo de 10%. Nesse contexto, apenas a média (M), mediana (MD), moda (MO) e o método aleatório (A) foram piores que as árvores de decisão, ficando sempre com MAPE acima de 10%.

REFERÊNCIAS

Batista, G. & Monard, M.-C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17:519–533.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer Science & Business Media, 2 edition.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.7.1.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dergachev, V. A., Gorban, A. N., Rossiev, A. A., Karimova, L. M., Kuandykov, E. B., Makarenko, N. G., & Steier, P. (2001). The filling of gaps in geophysical time series by artificial neural networks. *Radiocarbon*, 43(2A):365–371.

Greiner, R., Grove, A., & Kogan, A. (1997). Knowing what doesn't matter: exploiting the omission of irrelevant data. *Artificial Intelligence*, 97(1-2):345–380.

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York.

Josse, J., Prost, N., Scornet, E., & Varoquaux, G. (2019). On the consistency of supervised learning with missing values. *arXiv:1902.06931*.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 20(2):119–127.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Ljung, G. M. (1989). A note on the estimation of missing values in time series. *Communications in Statistics - Simulation and Computation*, 18(2):459–465.

Luceño, A. (1997). Estimation of missing values in possibly partially nonstationary vector time series. *Biometrika*, 84(2):495–499.

Molenberghs, G., Fitzmaurice, G. M., Kenward, M. G., Tsiatis, A. A., & Verbeke, G. (2020). *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis Group.

Morettin, P. A. & Toloi, C. M. d. C. (2004). *Análise de séries temporais*. Edgard Blucher.

Moritz, S. & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1):207–218.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press.

Prass, T. S. & Pumi, G. (2021). On the behavior of the DFA and DCCA in trend-stationary processes. *Journal of Multivariate Analysis*, 182:104703.

Pratama, I., Permanasari, A., Ardiyanto, I., & Indrayani, R. (2016). A review of missing values handling methods on time-series data. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

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

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Shumway, R. H. & Stoffer, D. S. (2005). *Time Series Analysis and Its Applications (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Therneau, T. & Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

Van der Vaart, A. W. (2010). Time series. Lecture notes for courses “Tijdreeksen”, “Time Series” and “Financial Time Series” held at Vrije Universiteit Amsterdam, 1995-2010.

Yodah, Kihoro, J., Athiany, H., & W, Kibunja, W. (2013). Imputation of incomplete non-stationary seasonal time series data. *Mathematical Theory and Modeling*, 3:142–154.

Contribuições dos autores

1 – Alisson Silva Neimaier
Programa de Pós-Graduação em Estatística - UFRGS
https://orcid.org/0000-0002-7524-0776 • alissonneimaier@hotmail.com
Contribuição: Investigação; Metodologia; Análise Formal; Visualização; Escrita – primeira redação

2 – Taiane Schaedler Prass
Programa de Pós-Graduação em Estatística - UFRGS
https://orcid.org/0000-0003-3136-909X • taiane.prass@ufrgs.br
Contribuição: Conceituação; Metodologia; Supervisão; Escrita – revisão e edição

Como citar este artigo

Neimaier, A. S., & Prass, T. S. (2024). Preenchimento de valores faltantes em séries temporais utilizando árvores de decisão. *Ciência e Natura*, Santa Maria, v.46, e84257. DOI: https://doi.org/10.5902/2179460X84257.