Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

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

Trabalho de Conclusão de Curso

Preenchimento de Valores Faltantes em Séries Temporais Utilizando Árvores de Decisão

Alisson Silva Neimaier

10 de Maio de 2022

Alisson Silva Neimaier

Preenchimento de Valores Faltantes em Séries Temporais Utilizando Árvores de Decisão

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.

Orientadora: Prof[a]. Dr[a]. Taiane Schaedler Prass

Porto Alegre
10 de Maio de 2022

Alisson Silva Neimaier

Preenchimento de Valores Faltantes em Séries Temporais Utilizando Árvores de Decisão

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(a) Orientador(a) e pela Banca Examinadora.

Orientadora:_____
Prof[a]. Dr[a]. Taiane Schaedler Prass, UFRGS
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. Flávio Augusto Ziegelmann, UFRGS
Universidade Federal do Rio Grande do Sul

Prof. Dr. Guilherme Pumi, UFRGS
Universidade Federal do Rio Grande do Sul

Porto Alegre
10 de Maio de 2022

"Existe uma teoria que diz que, se um dia alguém descobrir exatamente para que serve o Universo e por que ele está aqui, ele desaparecerá instantaneamente e será substituído por algo ainda mais estranho e inexplicável. Existe uma segunda teoria que diz que isso já aconteceu.".
- Douglas Adams, O Guia do Mochileiro das Galáxias

Agradecimentos

Agradeço aos meus cachorros, Bandit, Diego, Feijão, Frida e Maria Bethânia pela companhia e amor canino;

Aos meus amigos e familiares, grande parte do que sou hoje, devo a vocês;

À Angelo, Eduardo e Max, por estarem comigo desde sempre e pela amizade que apenas cresce com o tempo;

Aos meus avós, Maria e Nelson, por todo o carinho e cuidado comigo, desde que eu era pequeno até hoje;

Aos meus pais, Elvis e Tatiane, por terem me guiado para que eu seja uma boa pessoa e incentivado minhas ideias malucas;

À minha companheira, Martha, teu apoio e teu amor fazem tudo ser mais fácil e a vida parecer mais bela. Agradeço também por embelezar este trabalho criando a Figura 2.2;

À OBMEP, por ter me mostrado que existiam mais pequenos nerds matemáticos no mundo e pelas oportunidades que continua me apresentando;

À tia dos docinhos e ao café da física por adoçarem e alegrarem os dias no Vale;

Aos colegas e professores, por terem feito do IME a minha casa nesses últimos anos, principalmente ao Gabriel e ao Rafael pela amizade além da estatística;

À minha orientadora do Ensino Médio, Simone, por ter sido minha inspiração para fazer o curso de estatística;

À minha orientadora, Taiane, pela orientação, confiança e por estar sempre disponível para falar sobre estatística, vida e tudo mais;

Aos membros da banca, Flávio e Pumi, agradeço por terem aceitado avaliar esse trabalho;

E a mim mesmo, obrigado Eu.

Resumo

Na literatura existem diversas técnicas para o tratamento de observações faltantes para dados que não são séries temporais. Já no contexto de séries temporais encontram-se alguns trabalhos focados em modelos lineares da família ARIMA. Entretanto, em geral, os artigos não discutem a validade das metodologias propostas para o caso de um grande volume de dados faltantes. Nesse contexto, a identificação da ordem do modelo apropriado para utilização de métodos paramétricos é outro ponto desafiador. Tendo em vista esses fatos, este trabalho aborda uma metodologia para recomposição 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 fazem o papel de variável resposta, enquanto que defasagens correspondentes a tais valores são utilizadas como preditoras, a árvore selecionada pelo algoritmo de treinamento é então utilizada para prever os valores faltantes na resposta. Para investigar a metodologia proposta, foram utilizadas simulações de Monte Carlo, considerando processos da família ARMA e o passeio aleatório, variando o tamanho das séries temporais, os parâmetros dos modelos, a 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 de inputação tradicionais. Os resultados encontrados evidenciam a potencial do método proposto e condizem com o referencial teórico deste estudo.

Palavras-Chave: Séries Temporais, Árvores de Decisão, Valores Faltantes.

Abstract

There are plenty of techniques for the treatment of missing data outside of the time series framework and some in the context of linear time series from the ARIMA family. However, in general, these articles do not discuss the validity of the proposed methodologies in case of a large volume of missing data. In this context, identifying the appropriate model order for the parametric methods is another challenging point. With that in mind, this work proposes a methodology for recomposing time series 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 are treated as the response variable, while the lags corresponding to those values are used as predictors. The tree selected by the training algorithm is then used to predict the missing values in the response. To analyze the proposed methodology, we use Monte Carlo simulations, considering processes from the ARMA family and the random walk processes varying the size of the time series, the model parameters, the proportion of missing values, and the number of predictors. To assess the quality of the recomposition, the decision trees' predictions were compared with those of some traditional inputation methods. The results show the potential of the methodology and are in line with what was built in the theoretical framework of this study.

Keywords: Time Series, Decision Trees, Missing Data.

Sumário

1 Introdução

Uma série temporal é uma sequência de observações ao longo do tempo. Existem diversos métodos para modelagem de séries temporais, porém, quando há a presença de dados faltantes, este pode ser um processo desafiador.

Apesar de alguns padrões como tendência e sazonalidade poderem ser identificados através de análise gráfica, mesmo quando existem dados faltantes, existem outras características que exigem a utilização de técnicas mais complexas, que muitas vezes só podem ser aplicadas na presença de todas as observações. Exemplos de métodos que não podem ser utilizados quando há a presença de valores faltantes são os métodos de análise de flutuação destendenciada (DFA - *detrended fluctuation analysis*) e análise de correlação cruzada destendenciada (DCCA - *detrended cross-correlation analysis*), que são ferramentas úteis na análise de associações dentro e entre séries temporais, inclusive em alguns contextos de não estacionariedade.

Na literatura, existem diversas técnicas para o tratamento de observações faltantes para dados que não são séries temporais (veja, por exmeplo, Peng e Lei, 2021, e referências ali contidas). No contexto de séries temporais encontram-se alguns trabalhos focados em modelos lineares da família ARIMA (autoregressivos integrados de médias móveis) (veja, por exemplo Ljung, 1989; Luceño, 1997; Yodah et al., 2013). Entretanto, em geral, os artigos não discutem a validade das metodologias propostas para o caso de um grande volume de dados faltantes.

Uma das razões pelas quais os Modelos ARMA são atrativos é porquê são um caso particular dos modelos lineares gerais que, pelo teorema da decomposição de Wold (veja Brockwell e Davis, 1991, página 188), podem ser utilizados para descrever qualquer processo estocástico fracamente estacionário. Motivado pela dificuldade associada ao processo de identificação do modelo 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 neste trabalho 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*. Esse ambiente serve como auxiliar na escolha dos parâmetros interessantes a serem estudadas via simulações de Monte Carlo e futuramente será disponibilizado como uma ferramenta para reconstrução de séries temporais reais.

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

2 Definições Iniciais

Para que seja possível uma discussão mais aprofundada sobre os resultados encontrados, é necessária a definição dos principais temas presentes neste estudo. Neste capítulo, 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).

2.1 Séries Temporais

Nesta seção, serão apresentados conceitos importantes relacionados a séries temporais que serão vitais para a construção dos algoritmos e discussões sobre resultados encontrados. Estudos mais detalhados sobre os temas a seguir podem ser encontrados em Brockwell e Davis (1991), Morettin e Toloi (2004) e Shumway e Stoffer (2005).

2.1.1 Processos Estocásticos

Um processo estocástico $\{X_t\}_{t\in T}$ é uma família de variáveis aleatórias definidas em um mesmo espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$, em que $T \neq \emptyset$ é um conjunto de índices arbitrário. Ao modelar uma série temporal, é assumido que cada valor observado x_t é uma realização de uma variável aleatória X_t, desta forma, a série temporal $\{x_t\}_{t\in T_0}$ é interpretada como sendo uma realização, ou parte de uma realização,de um processo estocástico $\{X_t\}_{t\in T}$, sendo T_0 um subconjunto de T. Vale ressaltar que o conjunto T considerado neste trabalho será o dos números inteiros, portanto, sem perda de generalidade, as definições que seguem serão descritas considerando-se esse conjunto.

Em qualquer contexto de modelagem, é necessário ter maneiras de lidar com a dependência entre duas ou mais variáveis aleatórias. Normalmente, são impostas algumas restrições para o tipo de dependência que essas variáveis podem ter, sendo muito comum assumir que essa dependência é estacionária. No que segue, são descritos os diferentes tipos de estacionariedade e como essas propriedades interferem nos estudos realizados.

Definição 2.1.1. Um processo $\{X_t\}_{t\in\mathbb{Z}}$ é dito *fortemente estacionário* se, e somente se suas distribuições finito-dimensionais são invariantes sob translações do tempo, isto é, para qualquer $(t_1, \cdots, t_k) \in \mathbb{Z}^k$,

$$(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h}) \stackrel{d}{=} (X_{t_1}, X_{t_2}, \ldots, X_{t_k}), \quad \forall\, k \geq 1, \quad h \in \mathbb{Z},$$

em que $\stackrel{d}{=}$ significa igualdade em distribuição.

A propriedade de estacionariedade forte é muito desejável, pois com base em apenas uma série temporal é possível tirar conclusões sobre as distribuições marginais e conjuntas de um processo estocástico, porém esta é uma propriedade muito restritiva e difícil de verificar na prática. De forma alternativa, um processo pode ser caracterizado com base nos seus momentos de primeira e segunda ordem e de suas funções de autocovariância e autocorrelação

Definição 2.1.2. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico tal que $\mathrm{Var}(X_t) < \infty$, para todo $t \in \mathbb{Z}$. Então a *função de autocovariância* e a *função de autocorrelação* de $\{X_t\}_{t\in\mathbb{Z}}$ são definidas, respectivamente, como

$$\gamma_X(r,s) = \mathrm{Cov}(X_r, X_s) = \mathbb{E}[(X_r - \mathbb{E}[X_r])(X_s - \mathbb{E}[X_s])],\ r,s \in \mathbb{Z},$$

e

$$\rho_X(r,s) = \frac{\gamma_X(r,s)}{\sqrt{\gamma_X(r,r)\gamma_X(s,s)}},\ r,s \in \mathbb{Z}.$$

As funções de autocovariância e autocorrelação medem o grau de interdependência linear entre as variáveis aleatórias. Como este trabalho pretende realizar previsões utilizando defasagens da própria série temporal como variáveis independentes, é importante entender o comportamento destas funções nos modelos que serão utilizados nas simulações. Finalmente, dada a definição de autocorrelação, é possível enunciar a definição de estacionariedade fraca.

Definição 2.1.3. Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é dito *fracamente estacionário* se, para todo $t \in \mathbb{Z}$,

i) $\mathbb{E}[X_t^2] < \infty$,

ii) $\mathbb{E}[X_t] = \mu,\ \mu \in \mathbb{R}$,

iii) $\gamma_X(r,s) = \gamma_X(r+t, s+t), \forall\ r,s \in \mathbb{Z}$.

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

Observe que, se $\{X_t\}_{t\in\mathbb{Z}}$ é fracamente estacionário, então $\gamma_X(r,s) = \gamma_X(r-s,0)$, para todo $r,s \in \mathbb{Z}$. Portanto, é mais conveniente redefinir a função de autocovariância apenas em termos de uma variável, a distância (*lag*) entre as observações, da seguinte forma

$$\gamma_X(h) := \mathrm{Cov}(X_{t+h}, X_t) = \gamma_X(t+h, h) = \gamma_X(h,0), \quad h \in \mathbb{Z}.$$

Como esta classe de processos é a mais importante para este trabalho, será omitida a palavra "fracamente" da Definição 2.1.3, e será dito que o processo $\{X_t\}_{t\in\mathbb{Z}}$ é estacionário. Um exemplo muito importante de processo estacionário, que é necessário para definir diversos modelos de séries temporais é o processo ruído branco.

Definição 2.1.4. Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é chamado de *ruído branco* com média μ e variância σ^2, denotado por $\{X_t\}_{t\in\mathbb{Z}} \sim \text{RB}(\mu, \sigma^2)$, se, e somente se, $\mathbb{E}[X_t^2] < \infty$, $\mathbb{E}[X_t] = \mu$, para todo $t \in \mathbb{Z}$, e sua função de autocovariância pode ser escrita como

$$\gamma_X(h) = \begin{cases} \sigma^2, & \text{se } h = 0, \\ 0, & \text{se } h \neq 0. \end{cases}$$

Assim como a função de autocorrelação, a função de autocorrelação parcial, definida a seguir, fornece informações sobre a estrutura de dependência do processo. Ela é 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 $X_{t+1}, \cdots, X_{t+h-1}$.

Definição 2.1.5. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário, com função de autocovariância $\gamma_X(\cdot)$. A *função de autocorrelação parcial* de $\{X_t\}_{t\in\mathbb{Z}}$, denotada por $\alpha_X(\cdot)$, é definida como

$$\alpha_X(h) = \phi_{hh}, \quad h \geq 1,$$

onde ϕ_{hh} é o coeficiente da equação

$$\mathcal{P}_{\overline{sp}\{1,X_1,X_2,\cdots,X_h\}}(X_{h+1}) = \phi_{h0} + \sum_{j=1}^{h} \phi_{hj} X_{h+1-j}, \tag{2.1.1}$$

e $\mathcal{P}_{\overline{sp}\{1,X_1,X_2,\cdots,X_h\}}(X_{h+1})$ denota a projeção ortogonal de X_{h+1} no subespaço fechado $\overline{sp}\{1, X_1, X_2, \cdots, X_h\}$ gerado pelas observações anteriores.

A partir das definições iniciais sobre processos estocásticos, é possível descrever as classes de modelos utilizadas nas simulações de Monte Carlo no Capítulo 3. No que segue, além de suas definições, também são discutidas algumas de suas propriedades.

2.1.2 Modelos 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. Os modelos mais simples desta família são os autoregressivos (AR) e os de média móvel (MA), definidos a seguir.

Definição 2.1.6. Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é denominado *autorregressivos de ordem p* ou AR(p) se, e somente se, pode ser escrito como

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{2.1.2}$$

em que $Y_t = X_t - \mu$, $\mu = \mathbb{E}[X_t]$ e $\{\varepsilon_t\}_{t\in\mathbb{Z}} \sim \text{RB}(\mu, \sigma_\varepsilon^2)$.

Definição 2.1.7. Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é denominado *média móvel de ordem q* ou MA(q) se, e somente se, pode ser escrito como

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad t \in \mathbb{Z}, \tag{2.1.3}$$

em que $\mu = \mathbb{E}[X_t]$ e $\{\varepsilon_t\}_{t\in\mathbb{Z}} \sim \text{RB}(\mu, \sigma_\varepsilon^2)$.

Denotando por L o operador defasagem, as equações (2.1.2) e (2.1.3) podem ser reescritas, respectivamente, como

$$\phi(L)(X_t - \mu) = \varepsilon_t \quad \text{e} \quad Y_t = \mu + \theta(L)\varepsilon_t, \quad t \in \mathbb{Z},$$

em que

$$\phi(z) = 1 - \phi_1 z - \cdots = \phi_p z^p \quad \text{e} \quad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \tag{2.1.4}$$

são denominados *polinômios característicos*. Uma característica interessante desses modelos é que, se $\phi(z) \neq 0$ para $z \in \mathbb{C}$, tal que $|z| \leq 1$, então o processo AR(p) estacionário pode ser escrito como um MA(∞) e, se $\theta(z) \neq 0$ para $z \in \mathbb{C}$, tal que $|z| \leq 1$, então o processo MA(q) pode ser escrito como um AR(∞). Essa relação, que é uma consequência do Teorema 2.1.1 apresentado a seguir, permite que tais modelos sejam facilmente identificados com base em suas funções de autocorrelação e autocorrelação parciais, conforme discutido ainda nesta seção. O modelo ARMA por sua vez combina as características dos modelos AR e MA em uma única equação, conforme apresentado abaixo.

Definição 2.1.8. Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é denominado autorregressivo de média móvel de ordem (p, q) ou ARMA(p, q) se pode ser escrito como

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad t \in \mathbb{Z} \tag{2.1.5}$$

em que $Y_t = X_t - \mu$, $\mu = \mathbb{E}[X_t]$ e $\{\varepsilon_t\}_{t\in\mathbb{Z}} \sim \text{RB}(\mu, \sigma_\varepsilon^2)$.

De forma análoga aos modelos AR e MA, é possível reescrever a equação (2.1.5) em termos dos polinômios característicos, isto é,

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

onde $\phi(\cdot)$ e $\theta(\cdot)$ são definidos em 2.1.4. É necessário ressaltar que as equações ARMA não tem uma solução única, de fato, existem infinitas soluções não estacionárias para estes processos (para mais detalhes, veja van der Vaart, 2010). Portanto, alguns autores (como, por exemplo, Brockwell e Davis, 1991) exigem que o processo ARMA seja estacionário por definição para que haja uma única solução para as equações.

O teorema a seguir explicita as condições necessárias e suficientes para a existência e unicidade de uma solução estacionária para o sistema de equações (2.1.5). Observe que, como os modelos AR e MA são casos particulares do modelo ARMA, é suficiente enunciar o teorema para o caso geral.

Teorema 2.1.1. *Se $\phi(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| = 1$, então o sistema de equaçãoes $\phi(L)(X_t - \mu) = \theta(L)\varepsilon_t$ possui uma única solução estacionária dada por*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}, \quad t \in \mathbb{Z}, \tag{2.1.6}$$

em que os coeficientes $\{\psi_j\}_{j\in\mathbb{Z}}$ são determinados através da relação

$$\theta(z)\phi^{-1}(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j.$$

Demonstração. Veja Brockwell e Davis (1991), teorema 3.13, página 88. ∎

No caso particular em que $\phi(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| \leq 1$, isto é, o polinômio não possui raízes nem dentro nem em cima do círculo unitário, conclui-se que $\psi_j = 0$, para todo $j < 0$. Nesse caso, o valor do processo no tempo t depende unicamente do presente e dos valores do processo ruído branco. Essa propriedade é conhecida como *causalidade*. Vários *softwares*, incluindo o *R*, incluem a exigência de que polinômio $\phi(\cdot)$ não possua raízes dentro do círculo unitário em suas funções de simulação e ajuste de modelos ARMA. O teorema que segue justifica o uso dessa restrição pois mostra que todo processo ARMA(p, q) estacionário pode ser reescrito em termos polinômios característicos cujas raízes estão todas fora do círculo unitário.

Teorema 2.1.2. *Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo ARMA(p, q) estacionário, com média μ, que satisfaz as equações*

$$\phi(L)(X_t - \mu) = \theta(L)\varepsilon_t, \quad \{\varepsilon_t\}_{t\in\mathbb{Z}} \sim RB(\mu, \sigma_\varepsilon^2),$$

em que $\phi(z) \neq 0$ e $\theta(z) \neq 0$, para $z \in \mathbb{C}$ tal que $|z| = 1$. Então existem polinômios $\tilde{\phi}(z)$ e $\tilde{\theta}(z)$, de graus p e q, respectivamente, satisfazendo

$$\tilde{\phi}(z) \neq 0 \quad e \quad \tilde{\theta}(z) \neq 0, \quad \text{para todo } z \in \mathbb{C} : |z| \leq 1.$$

e um processo ruído branco $\{\varepsilon_t^*\}_{t\in\mathbb{Z}}$ tais que o processo $\{X_t\}_{t\in\mathbb{Z}}$ satisfaz

$$\tilde{\phi}(L)(X_t - \mu) = \tilde{\theta}(L)\varepsilon_t^*, \quad \{\varepsilon_t\}_{t\in\mathbb{Z}} \sim RB(\mu, \sigma_\varepsilon^2).$$

Demonstração. Veja Brockwell e Davis (1991), proposição 3.5.1, página 105. ∎

Portanto, sem perda de generalidade, serão considerados nas simulações realizadas neste trabalho, apenas processos ARMA satisfazendo a condição de causalidade. Nesse contexto, segue de imediato de (2.1.6) que

$$\mathbb{E}[X_t] = \mu, \quad \text{Var}[X_t] = \sigma_\varepsilon^2 \sum_{k=0}^{\infty} \psi_k^2 \quad \text{e} \quad \rho_X(h) = \frac{\sum_{k=0}^{\infty} \psi_k \psi_{k+|h|}}{\sum_{k=0}^{\infty} \psi_k^2}. \tag{2.1.7}$$

No que segue são apresentadas as expressões exatas para os casos em que $p, q \in \{0, 1\}$, que são o foco das simulações do Capítulo 3.

Suponha que $p = q = 1$. Por simplicidade de notação defina $\phi := \phi_1$ e $\theta := \theta_1$. Assuma que $\phi \neq 0$ e observe que

$$\phi(z) \neq 0, \ z \in \mathbb{C} : |z| = 1 \implies \phi^{-1}(z) = \sum_{k=0}^{\infty} \phi^k z^k.$$

Então, $\psi(z) := \theta(z)\phi^{-1}(z)$, implica

$$\psi(z) = (1 - \theta z) \sum_{k=0}^{\infty} \phi^k z^k = 1 + \sum_{k=1}^{\infty} \phi^k z^k - \theta \sum_{k=1}^{\infty} \phi^{k-1} z^k = 1 + \sum_{k=1}^{\infty} (\phi^k - \theta\phi^{k-1}) z^k.$$

Segue que

$$\psi(z) = \sum_{k=0}^{\infty} \psi_k z^k \quad \iff \quad \psi_0 = 1 \quad \text{e} \quad \psi_k = (\phi^k - \theta\phi^{k-1}), \quad k \geq 1. \tag{2.1.8}$$

Observe que, de (2.1.8) conclui-se que

$$\psi_k^2 = I(k=0) + \Big(1 - \frac{2\theta}{\phi} + \frac{\theta^2}{\phi^2}\Big)\phi^{2k} I(k>0), \quad k \geq 0,$$

e, para $h \neq 0$,

$$\psi_k \psi_{k+|h|} = (\phi^{|h|} - \theta\phi^{|h|-1})I(k=0) + \Big(1 - \frac{2\theta}{\phi} + \frac{\theta^2}{\phi^2}\Big)\phi^{2k+|h|} I(k>0), \quad k \geq 0.$$

Logo,

$$\begin{aligned}\sum_{k=0}^{\infty} \psi_k^2 &= 1 + \left(1 - \frac{2\theta}{\phi} + \frac{\theta^2}{\phi^2}\right)\sum_{k=1}^{\infty}\phi^{2k} = 1 + \left(1 - \frac{2\theta}{\phi} + \frac{\theta^2}{\phi^2}\right)\frac{\phi^2}{(1-\phi^2)} \\ &= 1 + \frac{\phi^2 - 2\theta\phi + \theta^2}{(1-\phi^2)} = 1 + \frac{(\phi-\theta)^2}{(1-\phi^2)}\end{aligned}$$

e, para $h \neq 0$,

$$\begin{aligned}\sum_{k=0}^{\infty} \psi_k \psi_{k+|h|} &= (\phi^{|h|} - \theta\phi^{|h|-1}) + \left(1 - \frac{2\theta}{\phi} + \frac{\theta^2}{\phi^2}\right)\sum_{k=1}^{\infty}\phi^{2k+|h|} \\ &= (\phi^{|h|} - \theta\phi^{|h|-1}) + \left(1 - \frac{2\theta}{\phi} + \frac{\theta^2}{\phi^2}\right)\frac{\phi^{|h|+2}}{(1-\phi^2)} \\ &= (\phi^{|h|} - \theta\phi^{|h|-1}) + \frac{\phi^{|h|+2} - 2\theta\phi^{|h|+1} + \theta^2\phi^{|h|}}{(1-\phi^2)} \\ &= \frac{(1-\phi^2)(\phi^{|h|} - \theta\phi^{|h|-1}) + \phi^{|h|}(\phi-\theta)^2}{(1-\phi^2)} = \frac{(\phi-\theta)(1-\phi\theta)\phi^{|h|-1}}{(1-\phi^2)}.\end{aligned}$$

Substituindo os resultados obtidos acima em (2.1.7) conclui-se que, se $\{X_t\}_{t\in\mathbb{Z}}$ é um processo ARMA(1,1), então $\phi\theta \neq 0$ e

$$\mathrm{Var}[X_t] = \sigma_\varepsilon^2\left(1 + \frac{(\phi-\theta)^2}{(1-\phi^2)}\right) \quad \text{e} \quad \rho_X(h) = I(h=0) + \frac{(\phi-\theta)(1-\phi\theta)\phi^{|h|-1}}{(1-\phi^2)+(\phi-\theta)^2} I(h\neq 0),$$

e, se $\{X_t\}_{t\in\mathbb{Z}}$ é um processo AR(1), então $\phi \neq 0$, $\theta = 0$ e

$$\mathrm{Var}[X_t] = \frac{\sigma_\varepsilon^2}{(1-\phi^2)} \quad \text{e} \quad \rho_X(h) = I(h=0) + \phi^{|h|} I(h \neq 0).$$

Por simplicidade, as propriedades do processo MA(1) foram obtidas separadamente. Observe que, no caso em que $p = 0$ e $q = 1$, temos $\phi = 0$ e

$$\psi(z) = \theta(z) = 1 - \theta z \quad \Longrightarrow \quad \psi_k = I(k=0) + \theta I(k=1), \quad k \geq 0.$$

Portanto,

$$\sum_{k=0}^{\infty} \psi_k^2 = 1 + \theta^2 \quad \text{e} \quad \sum_{k=0}^{\infty} \psi_k \psi_{k+|h|} = (1+\theta^2)I(h=0) + \theta I(|h|=1), \quad h \in \mathbb{Z}.$$

Logo,

$$\mathrm{Var}[X_t] = \sigma_\varepsilon^2(1+\theta^2) \quad \text{e} \quad \rho(h) = I(h=0) + \theta I(|h|=1), \quad h \in \mathbb{Z}.$$

Observa-se ainda que, de (2.1.1) e (2.1.2) é possível concluir que, para um processo AR(1), com $|\phi| < 1$, a função de autocorrelação parcial de $\{X_t\}_{t\in\mathbb{Z}}$ é dada por

$$\alpha_X(h) = \phi I(h = 1), \quad h \geq 1.$$

Já para os modelos MA(1), com $|\theta| \neq 1$, e ARMA(1,1) com $|\phi| \neq 1$ e $|\theta| \neq 1$, a função de autocorrelação parcial é tal que $\alpha_X(h) \neq 0$, para todo $h \geq 1$. Isso deve-se ao fato de que tais processos podem ser reescritos como um AR(∞). A prova formal desses resultados envolve solucionar as equações de Yule-Walker (para mais detalhes, veja Brockwell e Davis, 1991).

2.1.3 Passeio Aleatório

Até o momento foram apresentadas definições e propriedades de processos estocásticos estacionários da família ARMA(p, q). Porém, é interessante estudar também o que acontece quando não é satisfeita a condição de estacionariedade. Um dos processos estocásticos não estacionários mais conhecido é o passeio aleatório cuja definição é apresentada a seguir.

Definição 2.1.9. Um processo estocástico $\{X_t\}_{t\in\mathbb{N}}$ é chamado de *passeio aleatório* se pode ser escrito como

$$X_0 = 0, \quad X_t = X_{t-1} + \varepsilon_t,\ t > 0, \quad \{\varepsilon_t\}_{t\geq 1} \sim \text{RB}(0, \sigma_\varepsilon^2). \tag{2.1.9}$$

Observe que (2.1.9) implica que o processo $\{X_t\}_{t\in\mathbb{N}}$ pode ser reescrito como

$$X_t = \sum_{i=1}^{t} \varepsilon_i, \quad X_0 = 0,\ \{\varepsilon_t\}_{t\geq 1} \sim \text{RB}(0, \sigma_\varepsilon^2),$$

de onde conclui-se que $\mathbb{E}(X_t) = 0$. Dessa forma, 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 e $\mathbb{E}(X_t) = \mu$, para todo $t \in \mathbb{N}$.

Se $\{X_t - \mu\}_{t\in\mathbb{N}}$ é um passeio aleatório, é imediato que

$$\mathbb{E}[X_t] = \mu, \quad \text{Var}[X_t] = \sigma_\varepsilon^2 t \quad \text{e} \quad \rho_X(t, s) = \frac{\min(t, s)}{\sqrt{ts}}.$$

Apesar desse processo não ser estacionário, embora seja um abuso de notação, devido à representação (2.1.9) e por simplicidade de notação, no Capítulo 3 o passeio aleatório será dito ser um processo AR(1) com $\phi = 1$.

2.2 Valores Faltantes

Valores faltantes (ou dados faltantes) ocorrem quando não há um valor associado a uma ou mais observações de uma variável, este é um problema comum que pode acarretar em diversas consequências para a análise estatística. Segundo Molenberghs et al. (2020), a presença de valores faltantes leva a perda de informação proporcionalmente a quantidade de dados faltantes e essa perda de informação pode gerar modelos imprecisos e viesados sobre o parâmetro de interesse.

2.2.1 Mecanismo gerador de dados faltantes

Para obter resultados válidos a partir de um banco de dados incompleto, é necessário pensar sobre a proveniência dos dados faltantes. Normalmente o pesquisador não tem controle sobre os dados ausentes, portanto, é preciso assumir algumas hipóteses sobre o mecanismo gerador e a validade das análises depende da razoabilidade dessas hipóteses. De acordo com Molenberghs et al. (2020) é possível classificar os valores faltantes em 3 categorias, conforme o mecanismo gerador:

- **Completamente aleatórios** (MCAR - *missing completely at random*): quando a probabilidade de uma observação da variável ser um valor faltante é independente dos demais valores de interesse (tanto os observados quanto os valores faltantes). Alguns exemplos são erros humanos de preenchimento de informações e falhas de equipamento.
- **Aleatórios** (MAR - *missing at random*): quando a probabilidade de uma observação da variável ser um valor faltante depende apenas dos valores de interesse observados. Por exemplo, alguém do sexo masculino tem menor chance de preencher um questionário sobre depressão, porém isso depende apenas do sexo e não do nível de depressão do indivíduo.
- **Não aleatórios** (MNAR - *missing not at random*): quando a probabilidade de uma observação da variável ser um valor faltante depende tanto dos valores observados quanto dos valores faltantes que estamos interessados. Os valores faltantes não aleatórios também são chamados de "não-ignoráveis", pois a própria informação dos valores faltantes deve ser modelada. Um exemplo disso, seria se os valores faltantes dependessem do dia da semana em que são coletadas as informações.

Também segundo Molenberghs et al. (2020), não existe um consenso sobre a definição do mecanismo completamente aleatório incluir a condição de independência também às variáveis explicativas. Para evitar os problemas que surgem quando não é feita esta exigência, na definição dada anteriormente foi utilizada a definição de MCAR oriunda de Little (1995). Neste trabalho, os valores faltantes serão assumidos como completamente aleatórios, esse cenário pode não ser factível em alguns problemas reais, conforme Greiner et al. (1997), porém, segundo Hastie et al. (2009) a maior parte dos métodos de imputação fazem esta suposição para serem válidos.

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:

- **Ignorar ou descartar dados:** Existem dois métodos. 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. Este método não é recomendado no contexto de séries temporais, porém, se interessar ao leitor, mais informações podem ser encontradas em Batista e Monard (2003);

- **Estimação:** Procedimentos de estimação por máxima verossimilhança são utilizados para estimar de forma paramétrica um modelo para os dados completos. Segundo Dempster et al. (1977), algumas variações do algoritmo EM (*Expectation-Maximization*) conseguem lidar com a estimação de parâmetros com dados incompletos;

- **Imputação:** Técnicas que buscam preencher os valores faltantes com valores estimados utilizando de relações conhecidas com os valores observados. O método proposto neste trabalho se enquadra nesta categoria, nele será utilizada a média condicional, o que permite maior flexibilidade.

Alguns métodos de imputação em séries temporais, que serão usados para comparação de resultados e estão disponíveis no pacote `imputeTS` (Moritz e Bartz-Beielstein, 2017), estão listados abaixo:

- Medidas de tendência central: média, mediana e moda;

- Médias móveis: com peso simples, linear e exponencial;

- LOCF/NOCB (*last observation carried forward/next observation carried backward*): última observação levada adiante próxima observação trazida para trás;

- Interpolação: interpolação linear, por splines e de Stineman;

- Suavização de Kalman: modelo estrutural estimado via máxima verossimilhança e usando uma representação do espaço de estados do modelo ARIMA, em que a ordem do modelo é escolhida automaticamente;

- Aleatório: seleciona aleatoriamente uma observação da amostra para reconstrução do valor faltante.

2.3 Árvores de Decisão

O aprendizado supervisionado é aplicado em um contexto em que são pré definidas as variáveis explicativas (ou dependentes) e as variáveis respostas (independentes) e deseja-se construir um modelo que seja capaz de descrever da melhor forma possível as respostas a partir das variáveis explicativas. Um modelo de aprendizado supervisionado de fácil interpretação é o de árvores de decisão.

Árvore de decisão é um método de classificação e regressão 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 dados em vários subespaços, de forma que os resultados em cada subespaço final sejam tão homogêneos quanto possível. Nesta seção, são apresentados os principais conceitos relacionados a 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?

Uma árvore é uma estrutura de dados com algumas especificidades, para que seja possível discutir sobre elas, estão listados abaixo alguns dos conceitos principais sobre a estrutura e processos de árvores:

- **Nó raiz:** Representa a totalidade da população ou da amostra;
- **Particionamento:** É o processo de dividir um nó em dois ou mais sub-nós;
- **Nó pai/Nó filho:** Um nó que é dividido em sub-nós é chamado de nó pai, os sub-nós são chamados de nós filhos;
- **Ramos:** Alternativas às tomadas de decisão que particionam um nó pai em dois nós filhos;
- **Nó Terminal/Folha:** Nós com nenhuma partição a partir dele. Indica à decisão a ser tomada (previsão);
- **Sub-árvore/Galho:** Uma sub-seção de uma árvore de decisão;
- **Variáveis auxiliares/Divisões auxiliares (*surrogate*):** quando existem dados faltantes em uma das variáveis independentes, uma variável auxiliar é utilizada para fazer o particionamento da árvore no lugar daquela com dado ausente. Tal variável é denominada *surrogate* e a técnica em questão recebe o nome de *surrogate split*;
- **Poda:** Quando reduzimos o tamanho da árvore de decisão removendo nós.

A Figura 2.1 apresenta uma estrutura simplificada de uma árvore de decisão *binária*, isto é, uma árvore de decisão em que cada nó pai é dividido no máximo dois nós filhos. Nesta figura é possível observar o nó raíz, os ramos, os nós e as folhas.

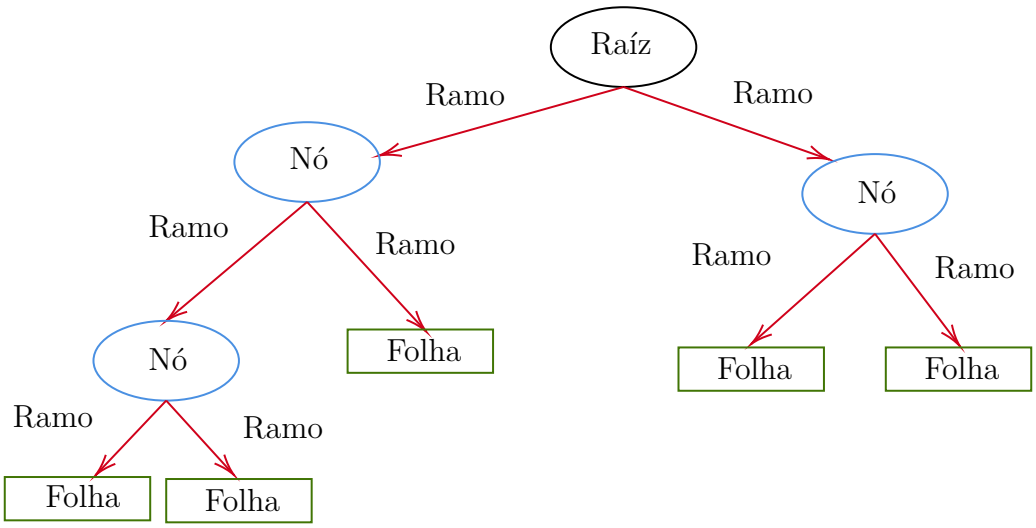

Figura 2.1: Estrutura de uma árvore de decisão.

2.3.2 Como fazer uma árvore crescer

Os métodos de aprendizado supervisionado buscam aprender como prever valores de uma variável resposta $Y \in \mathcal{Y}$ ou, de forma mais geral, de uma função da variável resposta $g(Y)$ a partir de variáveis explicativas $\boldsymbol{X} \in \boldsymbol{\mathcal{X}} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p = \otimes_{j=1}^{p} \mathcal{X}_j$, também denominadas variáveis independentes. No contexto tradicional de árvores de decisão, o par $(\boldsymbol{X}, Y)$ é considerado como sendo um vetor aleatório com distribuição conjunta $\mathbb{P}$ e os dados observados, denotados por $\{\boldsymbol{X}_i, Y_i\}_{i=1}^{n}$, são vistos como uma amostra aleatória de $\mathbb{P}$.

Existem diferentes algoritmos para construção das árvores de decisão, que variam em método e nos tipos de variáveis que suporta. Abaixo, estão listados alguns dos mais conhecidos:

- *Iterative Dichotomiser 3* (ID3): Variável resposta binária, variáveis explicativas categóricas (Quinlan, 1986);
- *Chi-Squared Automatic Interaction Detector* (CHAID): Variável resposta e variáveis explicativas categóricas (Kass, 1980);
- *Classification and regression Tree* (CART): Sem restrições sobre as variáveis (Breiman et al., 1984).

Neste trabalho, será utilizada a versão do algoritmo CART (ou C&RT) implementada no pacote `rpart` do R (Therneau e Atkinson, 2019), 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 (*imput*) $\boldsymbol{\mathcal{X}}$ e realiza as previsões no espaço de saídas (*output*) $\mathcal{Y}$. Para o problema abordado neste trabalho, $\mathcal{X} = \mathbb{R}^p$ e $\mathcal{Y} = \mathbb{R}$. Uma descrição completa desse algoritmo pode ser encontrada no manual fornecido pelos autores do pacote `rpart`. De maneira informal e resumida, podemos dizer que o algoritmo 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, para cada nó $A = \otimes_{j=1}^{p}[\ell_j, r_j] \subset \mathbb{R}^p$, 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 (veja Josse et al., 2019, para mais detalhes)

$$(j^*, z^*) = \underset{(j,z)\in S}{\arg\min}\Bigg\{\mathbb{E}\Bigg[\Big(Y - \mathbb{E}\big[Y|X_j \leq z, \boldsymbol{X} \in A\big]\Big)^2 I(X_j \leq z, \boldsymbol{X} \in A) + \Big(Y - \mathbb{E}\big[Y|X_j > z, \boldsymbol{X} \in A\big]\Big)^2 I(X_j > z, \boldsymbol{X} \in A)\Bigg]\Bigg\}. \tag{2.3.1}$$

Para qualquer nó A, a otimização acima é equivalente a solucionar o seguinte problema

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

em que $\mathcal{P}_c$ 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 desse processo 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

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

Na Figura 2.2 é possível visualizar o exemplo de uma árvore de decisão com 5 nós terminais que geram as regiões $\{R_1, \cdots, R_5\}$. Para o nó raíz, o par que minimiza a equação 2.3.1 é $(1, t_1)$, ou seja, o ponto t_1 da variável X_1.

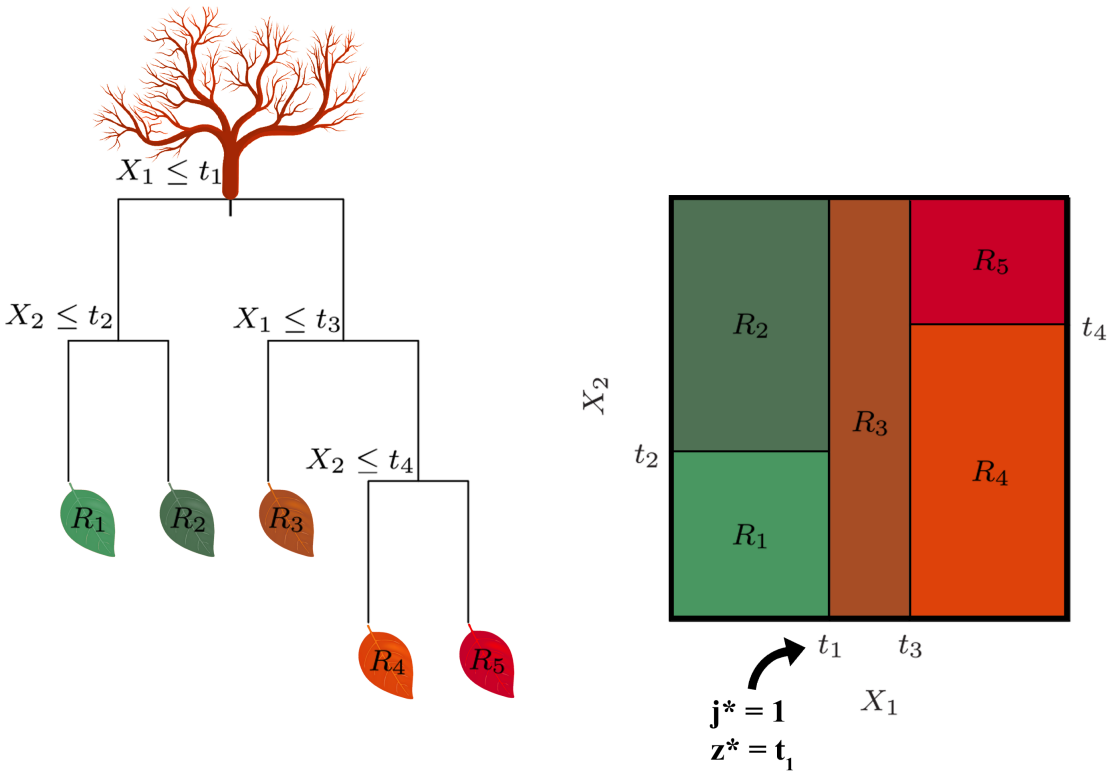

Figura 2.2: Resultado de uma partição recursiva binária em um exemplo bivariado (à direita) e a árvore de decisão correspondente à esta partição (à esquerda).

A melhor solução para o problema acima seria a que minimizasse o erro calculado a partir de $f(\boldsymbol{X})$. 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, faz-se a seguinte pergunta: como decidir o tamanho da árvore?

2.3.3 Como podar uma árvore

A estratégia padrão para decidir o tamanho de uma árvore é criar uma árvore grande $\mathcal{T}_0$ e podá-la até encontrar a sub-árvore que tenha o menor erro em um grupo de teste. O erro é estimado utilizando o método de validação-cruzada (*cross-validation*). Como fazer a validação-cruzada para todas as sub-árvores seria uma tarefa muito pesada, utilizamos um método chamado *cost complexity pruning* (também conhecido

como *weakest link pruning*) para pegarmos apenas um pequeno conjunto das sub-árvores. 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 sub-árvore $\mathcal{T}_\alpha \subset \mathcal{T}_0$ que minimiza

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

em que $|\mathcal{T}|$ é o número de nós terminais da árvore $\mathcal{T}$, $\mathscr{X}_m = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\} \cap R_m$, 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 sub-árvore $\mathcal{T}$ e o quão bem ela se ajusta ao grupo de treino. Quando $\alpha = 0$, a sub-árvore $\mathcal{T}_\alpha$ 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 sub-árvores 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 sub-árvores 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 sub-árvore correspondente a α. O processo para construção e poda de árvores pode ser resumido conforme o algoritmo abaixo (James et al., 2013)

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;

2. Aplique *cost complexity pruning* para obter a sequência de melhores sub-árvores como uma função de α.

3. Use validação cruzada K-*fold* para escolher α. Isto é, divida as observações do grupo de teste em K subconjuntos. Para cada $k = 1, \cdots, K$:

   - Repita os passos 1 e 2 em todos os folds, menos o k-ésimo.
   - Avalie o erro quadrático médio de previsão utilizando o k-ésimo fold, como uma função de α.

   Para cada α, calcule a média dos resultados e selecione o α que minimiza o erro médio.

4. Retorne a sub-árvore do passo 2 que corresponde ao valor escolhido de α.

2.3.4 Rpart

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 e Atkinson, 2019), o algoritmo implementado segue

o que foi descrito por Breiman et al. (1984). A função `rpart` aceita como entrada diversos argumentos para personalizar o ajuste do modelo de árvores de decisão. Os argumentos utilizados neste trabalho e a descrição dos mesmos são explicitados abaixo:

- **`formula:`** a fórmula do modelo, como na função `lm` e outras funções de ajuste do modelo *R*. Nela são indicadas a variável resposta Y e as variáveis independentes $\boldsymbol{X}$. As variáveis em questão podem estar definidas no ambiente global ou em um objeto da classe `data.frame`. No segundo caso, o objeto deve ser indicado através do argumento `data` e os nomes utilizados na fórmula deve coincidir com os nomes das colunas correspondentes no objeto;

- **`data:`** um objeto da classe `data.frame` que corresponde ao banco de dados utilizado na `formula`;

- **`na.action:`** indica a maneira como a função lida com valores faltantes. O padrão é `na.rpart` e faz com que sejam retiradas da modelagem apenas as linhas do banco de dados cuja variável resposta seja um valor faltante ou que todas as covariáveis possuam valores faltantes;

- **`method:`** tipo de regra de divisão a ser usada. Se o método estiver ausente, a rotina tentará fazer uma estimativa inteligente com base na variável resposta. Se Y é um objeto da classe `Surv` (proveniente do pacote `survival`) então `method = "exp"` é assumido, se Y tem 2 colunas então `method = "poisson"` é assumido, se Y é um fator então `method = "class"` é assumido, caso contrário `method = "anova"` é assumido;

- **`control:`** lista de opções que controlam o algoritmo `rpart`. Os parâmetros disponíveis para modificação, que podem ser definidos através desse lista são os argumentos da função `rpart.control`.

A função **rpart.control:** permite acessar e redefinir controles adicionais para a modelagem das árvores de decisão, são eles:

- **`minsplit` e `minbucket:`** definem, respectivamente, o número mínimo de observações necessárias em um nó para que seja considerada mais uma divisão e o número mínimo de observações em qualquer nó terminal. Se apenas um desses argumentos for especificado a função automaticamente utiliza a relação `minsplit = minbucket*3`. O padrão da função é `minsplit = 20` e `minbucket = 7`;

- **`cp:`** parâmetro de complexidade utilizado para realizar uma pré-poda. Divisões que não melhoram o modelo por um fator de pelo menos `cp` são ignorados. O principal objetivo é reduzir o tempo para computar todas as divisões, eliminando as que claramente não valem a pena. O padrão é 0.01;

- **`maxcompete:`** é um parâmetro cujo propósito é auxiliar na visualização e interpretação dos resultados. O *output* final retém a informação de quais foram as *maxcompete* melhores partições depois da escolhida. O padrão é 4. Esse argumento não tem efeito no tempo computacional e afeta pouco o uso da memória;

- **`maxsurrogate:`** número de partições substitutas mantidas no *output* final. O padrão é 5;

- **`usesurrogate:`** define como os *surrogates* devem ser utilizados no processo de divisão. As possíveis opções são

  - 0: apenas exibição. Se uma observação possui um valor ausente para a regra de divisão primária ela não é considerada nos passos seguintes.
  - 1: usar substitutos, em ordem, para dividir os casos cuja variável primária está faltando. Se todos os substitutos estiverem faltando, a observação não será dividida.
  - 2: se todos os substitutos estão faltando, o algoritmo utiliza a regra do voto majoritário. Esse é o padrão e corresponde à recomendação de Breiman et al. (1984).

  Observação: Nos casos em que uma observação não é considerada nas divisões seguintes, a previsão associada a tal observação é obtida com base no nível máximo que a observação atingiu na árvore. Por exemplo, se `usesurrogate = 0` e a variável X_j que deve ser utilizada para fazer a primeira partição da árvore possuir um valor falante para a i-ésima observação, então $\hat{Y}_i$ será definido como sendo a média global da variável resposta (pois nesta etapa as observações encontram-se todas no nó raiz) e a observação $(Y_i, \boldsymbol{X}_i)$ não será carregada para o próximo nível da árvore, isto é, ela não aprecerá nos nós filhos criados a partir do nó raiz.

- **`surrogatestyle:`** controla como é feita a seleção do melhor substituto. Quando definido como 0 (padrão), o desempenho da potencial variável substituta é analisado levando-se em conta o total de classificações corretas; quando definido como 1, utiliza-se o percentual correto, calculado sobre os valores não faltantes da substituta. A primeira opção penaliza mais severamente as covariáveis com um grande número de valores ausentes. A descrição fornecida no manual do `rpart` não deixa claro como esse critério se aplica nos casos em que as variáveis envolvidas são contínuas;

- **`xval:`** número de validações cruzadas. O padrão é 10;

- **`maxdepth:`** profundidade máxima do nó final da árvore. O nó raiz conta como profundidade 0. Valores maiores que 30 produzirão resultados sem sentido em máquinas de 32 bits. O padrão é 30.

3 Estudos de Simulação

Neste capítulo apresentamos os resultados de uma simulação de Monte Carlo realizada com o objetivo de investigar a qualidade da reconstrução de séries temporais, utilizando árvores de decisão. São consideradas séries simuladas provenientes de modelos ARMA(p, q) e também de um passeio aleatório[1], com diferentes proporções de valores faltantes (ρ). Além de variar a proporção de dados faltantes, também são considerados diferentes tamanhos de amostras e diversos cenários para as covariáveis. Para fins de comparação, são também empregados os métodos de imputação listados na Seção 2.2.2.

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 cenário considerado, as amostras $\{X_t\}_{t=1}^{n}$ com valores faltantes, foram geradas seguindo-se os passos descritos a seguir.

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

Passo 2 Para os modelos ARMA, a série temporal com média zero é gerada utilizando a função `arima.sim`, que primeiramente gera a parte de médias móveis do modelo e depois calcula a parte autoregressiva recursivamente, conforme o algortimo descrito no final desta lista. 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

$$X_t = \mu + \sum_{i=1}^{t} \varepsilon_i, \quad t = 1, \cdots, m.$$

Observe que, neste caso, não é necessário o *burn-in*, ou seja, $b = 0$.

Passo 3 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

[1] Por simplicidade de notação, no que segue nos referimos ao passeio aleatório como sendo um modelo AR(1) com $\phi = 1$.

transformadas em valores faltantes. Desta forma, a série temporal com a inclusão dos valores faltantes é escrita como

$$X_t^{\text{miss}} = \begin{cases} \texttt{NA}, & \text{se } t \in T_1, \\ X_t, & \text{se } t \in T_1^C. \end{cases}$$

Passo 4 Repetir o processo $r = 1000$ vezes, de tal forma que ao final, serão 1000 replicações de cada modelo distinto.

Algoritmo da função `arima.sim`

1. Dada a sequência $\{\varepsilon_t\}_{t=-b+1}^{m}$, define-se $\xi_t = \varepsilon_t$, se $q = 0$, e

$$\xi_t = \left(\varepsilon_t + \sum_{k=1}^{q} \theta_k \varepsilon_{t-k}\right) I(t > -b + q), \text{ se } q > 0, \text{ para } t \in \{-b+1, \cdots, m\}.$$

2. Dada a sequência $\{\xi_t\}_{t=-b+1}^{m}$, define-se $y_t = \xi_t$, se $p = 0$, e

$$Y_t = \sum_{k=1}^{p} \phi_k \xi_{t-k} I(t - k \geq -b + 1), \text{ se } p > 0, \text{ para } t \in \{-b+1, \cdots, m\}.$$

3. A sequência $\{Y_t\}_{t=1}^{m}$ corresponde a uma amostra de tamanho m de um processo ARMA(p, q) com média zero.

3.2 Cenários testados

Em todos os cenários analisados considerou-se

- proporção de valores faltantes $\rho \in \{0.1, 0.2, 0.5, 0.8\}$;
- 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)};
- tamanho da amostra $n \in \{100, 500, 1000\}$.

Em relação aos parâmetros dos modelos,

- 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};
- para o modelo ARMA$(1, 1)$, $\phi = 0.7$ e $\theta = 0.4$, para que fosse condizente com a série temporal gerada no artigo de Prass e Pumi (2021).

3.3 Estimativas

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 Subseção 2.2.2 e a partir do método proposto, seguindo o algoritmo abaixo

1. Definir X_t^{miss} como sendo a variável resposta.
2. Construir a matriz de covariáveis, utilizando os h_1 passos anteriores e h_2 passos posteriores de X_t^{miss}.
3. 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`.
4. 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`.
5. 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`.

3.4 Métricas

Sejam $\{X_t\}_{t=1}^n$ a série temporal original, isto é, sem valores faltantes e $\hat{X}_t$ o valor estimado para preencher os valores faltantes X_t^{miss}, para $t \in T_1$. Neste trabalho, as métricas utilizadas para medir a qualidade da reconstituição dos valores faltantes são o erro quadrático médio (EQM) e o erro absoluto percentual médio (EAPM), respectivamente definidas por

$$\text{EQM} = \frac{1}{\lfloor n\rho \rfloor} \sum_{t \in T_1} (X_t - \hat{X}_t)^2 \quad \text{e} \quad \text{EAPM} = \frac{1}{\lfloor n\rho \rfloor} \sum_{t \in T_1} \left| \frac{X_t - \hat{X}_t}{X_t} \right|.$$

Nas próximas seções, são apresentados os resultados referentes à comparações dos métodos de preenchimento de valores faltantes, com base nas medidas definidas acima. Na Seção 3.5 são discutidas as diferenças entre os preenchimentos das árvores, dependendo do número de covariáveis utilizadas no modelo. A Seção 3.6 é dedicada à comparação entre as árvores obtidas com um $\boldsymbol{h}$ específico, determinado com base nos estudos realizados na Seção 3.5, e os outros métodos.

3.5 Comparação das árvores

Primeiramente, analisou-se o desempenho das árvores de decisão em termos de proporção de dados faltantes e das defasagens utilizadas como covariáveis. Esta análise teve como objetivo verificar a relação do número de covariáveis utilizadas nas árvores de decisão e a qualidade de previsão em cada um dos cenários descritos na Seção

3.2 para decidir quais os valores de $\boldsymbol{h}$ seriam utilizados na comparação das previsões das árvores de decisão com os outros métodos de imputação.

Nas Figuras 3.1 - 3.3 são apresentados os box-plots referentes à 1000 replicações do logaritmo do erro quadrático médio de previsão para os cenários em que as árvores de decisão foram obtidas considerando-se amostras de tamanho $n \in \{100, 500, 1000\}$, respectivamente, do modelo AR(1), para $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$. Nas Figuras 3.7 - 3.9 são apresentados os resultados referentes aos modelos MA(1), com $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$, e a Figura 3.13 correspondem ao modelo ARMA(1,1) com $\phi = 0.7$ e $\theta = 0.4$. De forma análoga, os resultados para o EAPM são apresentados nas Figuras 3.4 - 3.6 para o modelo AR(1), nas Figuras 3.10 - 3.12 para o modelo MA(1) e na Figura 3.14 para o modelo ARMA$(1, 1)$. Em cada figura considera-se ainda $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ e $\boldsymbol{h} \in \{$(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)$\}$.

De forma geral observa-se que o EQM 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 EQM aumenta com $|\phi|$ e $|\theta|$, sendo o aumento mais marcante no caso dos modelos AR, onde $|\phi| = 1$ corresponde à um processo não estacionário. Os valores do EAPM apresentam um comportamento semelhante ao EQM. No que segue, são descritos os comportamentos das métricas para cada modelo separadamente, em termos de $\boldsymbol{h}$.

Modelo AR(1):

- $\rho = 0.8$: quando a proporção de valores faltantes é muito alta, não há uma distinção muito clara de quais valores de $\boldsymbol{h}$ resultaram nas melhores previsões, independente do tamanho da amostra. O valor $\boldsymbol{h} = (5, 5)$ produziu resultados ligeiramente melhores em alguns cenários;

- $|\phi| = 0.1$: neste caso não há diferença aparente entre as previsões, independentemente dos valores de n, ρ e $\boldsymbol{h}$;

- $|\phi| = 0.5$: neste caso não é possível notar diferença na qualidade das previsões quando $n = 100$. Para os tamanhos de amostras maiores, é visível que os valores das métricas são menores quando são utilizadas informações tanto do passado quanto do futuro da amostra, ou seja, quando $\boldsymbol{h} \in \{(1, 1), (2, 2), (5, 5)\}$;

- $|\phi| = 0.9$: neste caso é visível que os valores das métricas 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;

- $\phi = 1$: no contexto de não estacionariedade é possível notar que os valores das métricas diminuem conforme aumentamos o número de covariáveis no modelo e os melhores resultados foram obtidos quando $\boldsymbol{h} = (5, 5)$.

Modelo MA(1):

- $\rho = 0.8$: não há uma distinção clara de quais valores de $\boldsymbol{h}$ resultaram nas melhores previsões, independente do tamanho da amostra, apenas que o caso em que $\boldsymbol{h} = (5, 5)$ foi melhor em alguns casos;

- $|\phi| = 0.1$: não há diferença aparente entre as previsões, independentemente dos valores de n, ρ e $\boldsymbol{h}$;

- $|\phi| \geq 0.5$: neste caso, não é possível notar diferença na qualidade das previsões quando $n = 100$, para os tamanhos de amostras maiores, é visível que os valores das métricas são menores quando são utilizadas informações tanto do passado quanto do futuro da amostra, ou seja, quando $\boldsymbol{h} \in \{(1,1),(2,2),(5,5)\}$;

Modelo ARMA(1,1):

- Para todos os casos, é notável que os valores das métricas são menores quando são utilizadas informações tanto do passado quanto do futuro da amostra, ou seja, quando $\boldsymbol{h} \in \{(1,1),(2,2),(5,5)\}$.

Em todos os cenários testados houve um desempenho melhor dos métodos que 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, na próxima seção, será considerado apenas o caso em que $\boldsymbol{h} = (5,5)$, ou seja, que foram utilizadas como covariáveis na construção das árvores os 5 passos anteriores e posteriores a observação faltante.

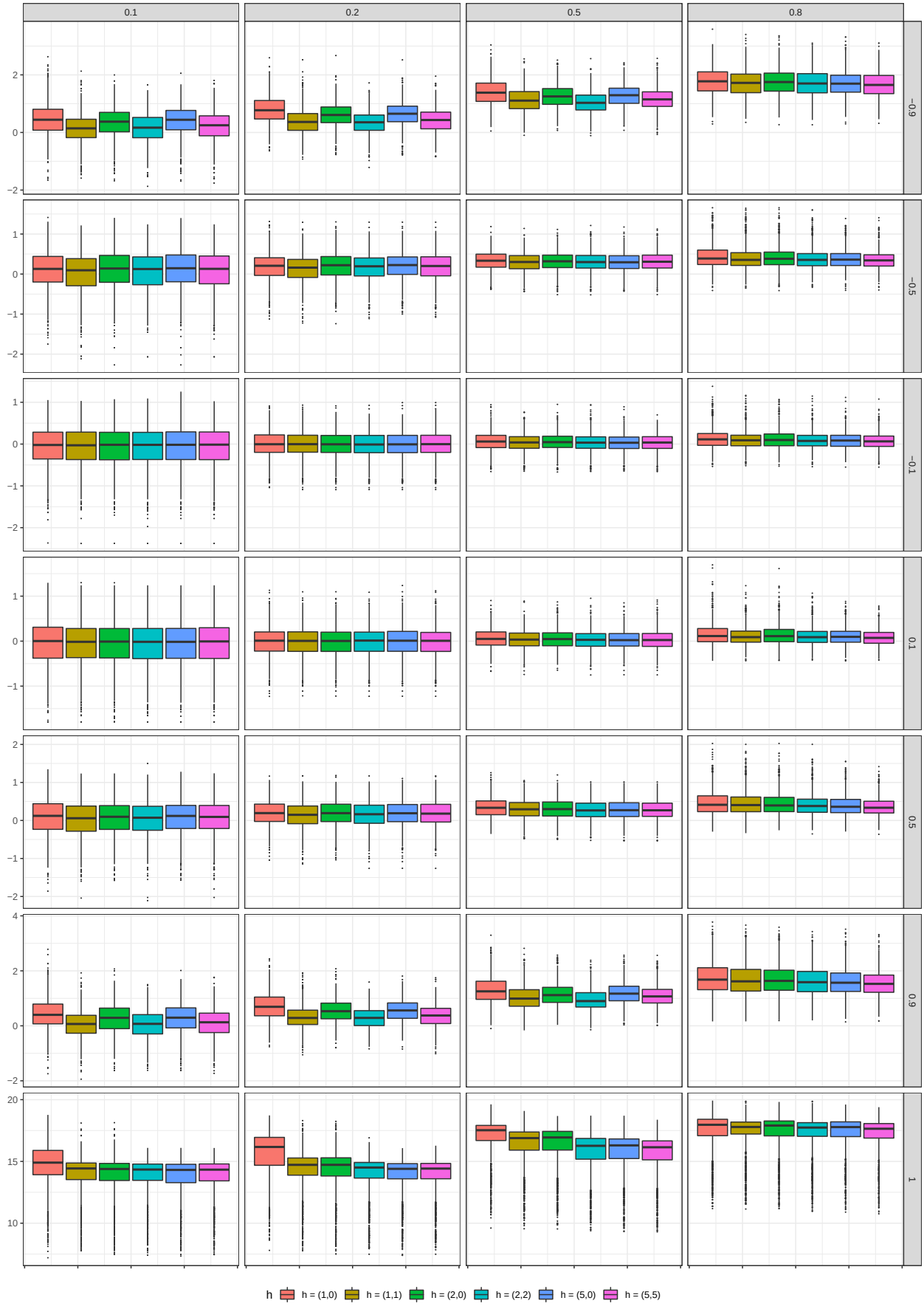

Figura 3.1: Box-plots do logaritmo do EQM para o modelo AR(1) com $n = 100$, $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas), utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

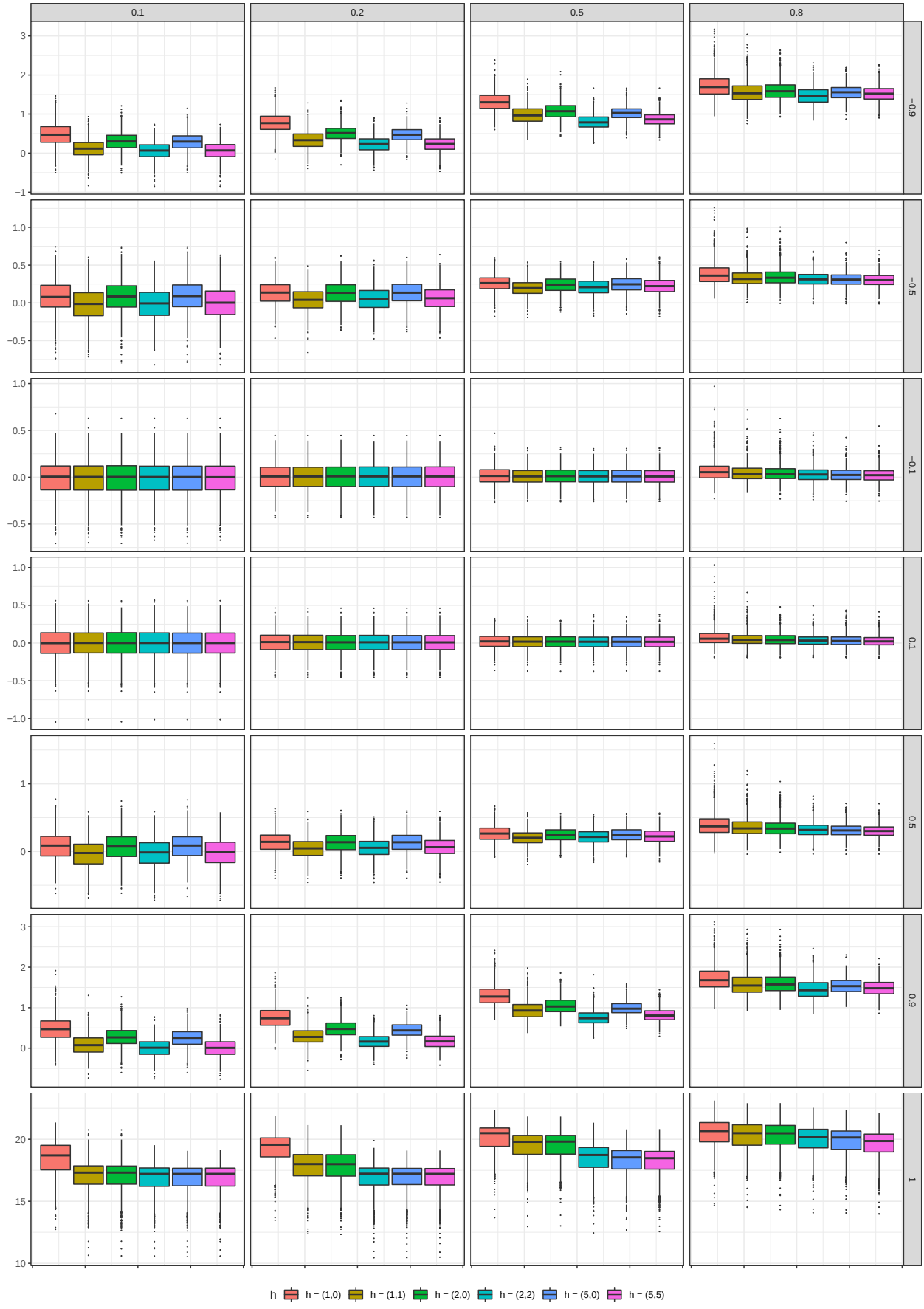

Figura 3.2: Box-plots do logaritmo do EQM para o modelo AR(1) com $n = 500$, $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

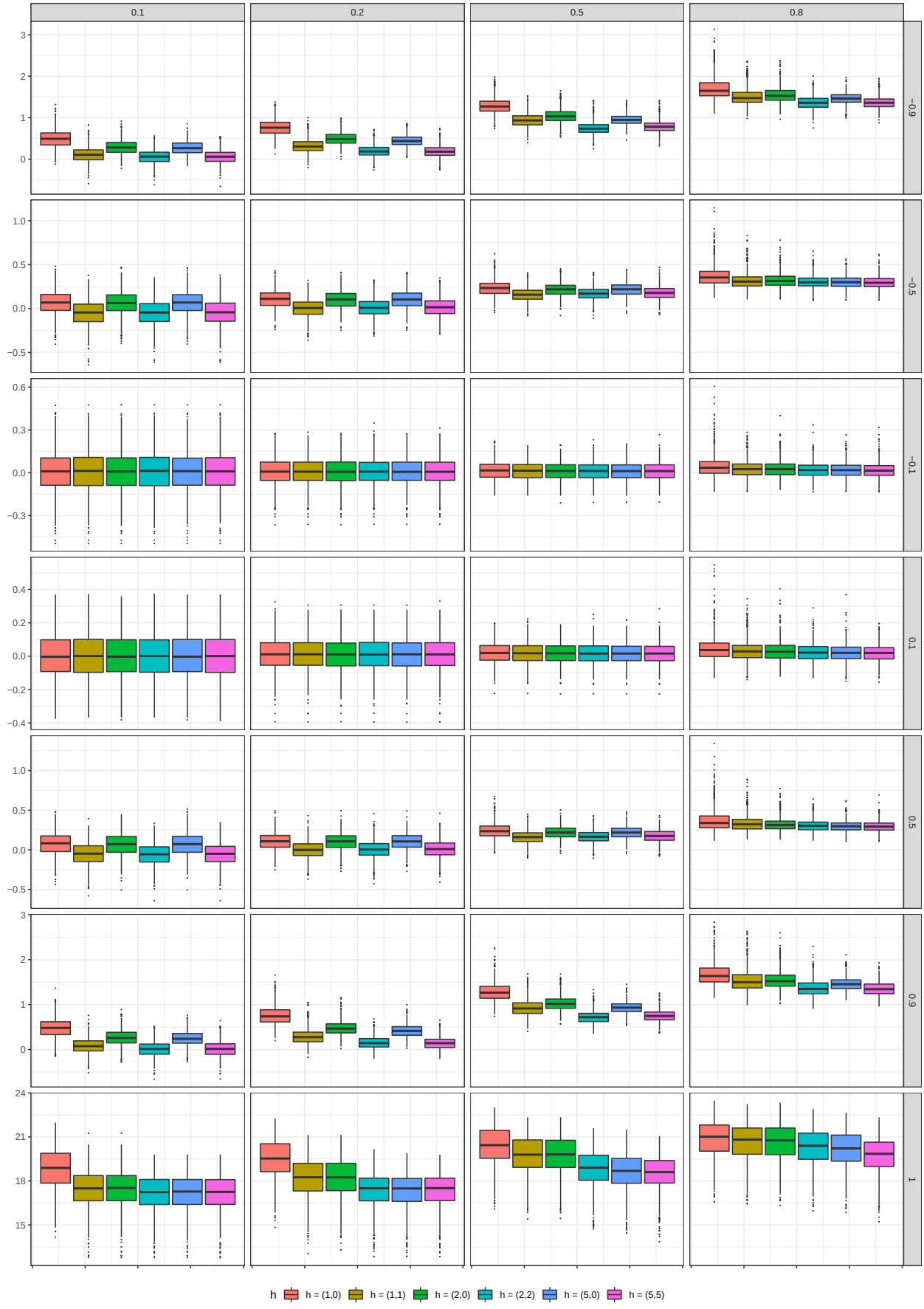

Figura 3.3: Box-plots do logaritmo do EQM para o modelo AR(1) com $n = 1000$, $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

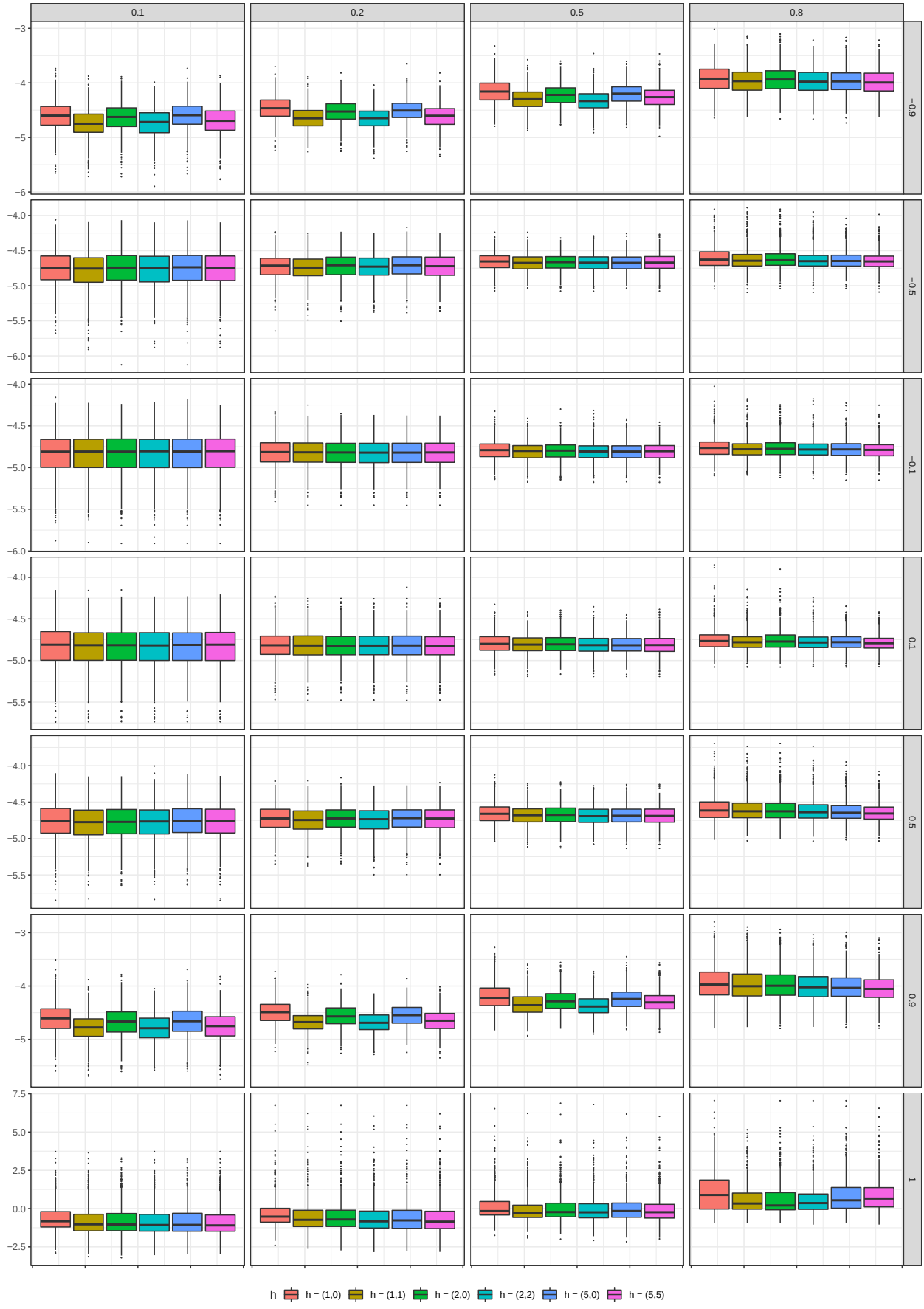

Figura 3.4: Box-plots do logaritmo do EAPM para o modelo AR(1) com $n = 100$, $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

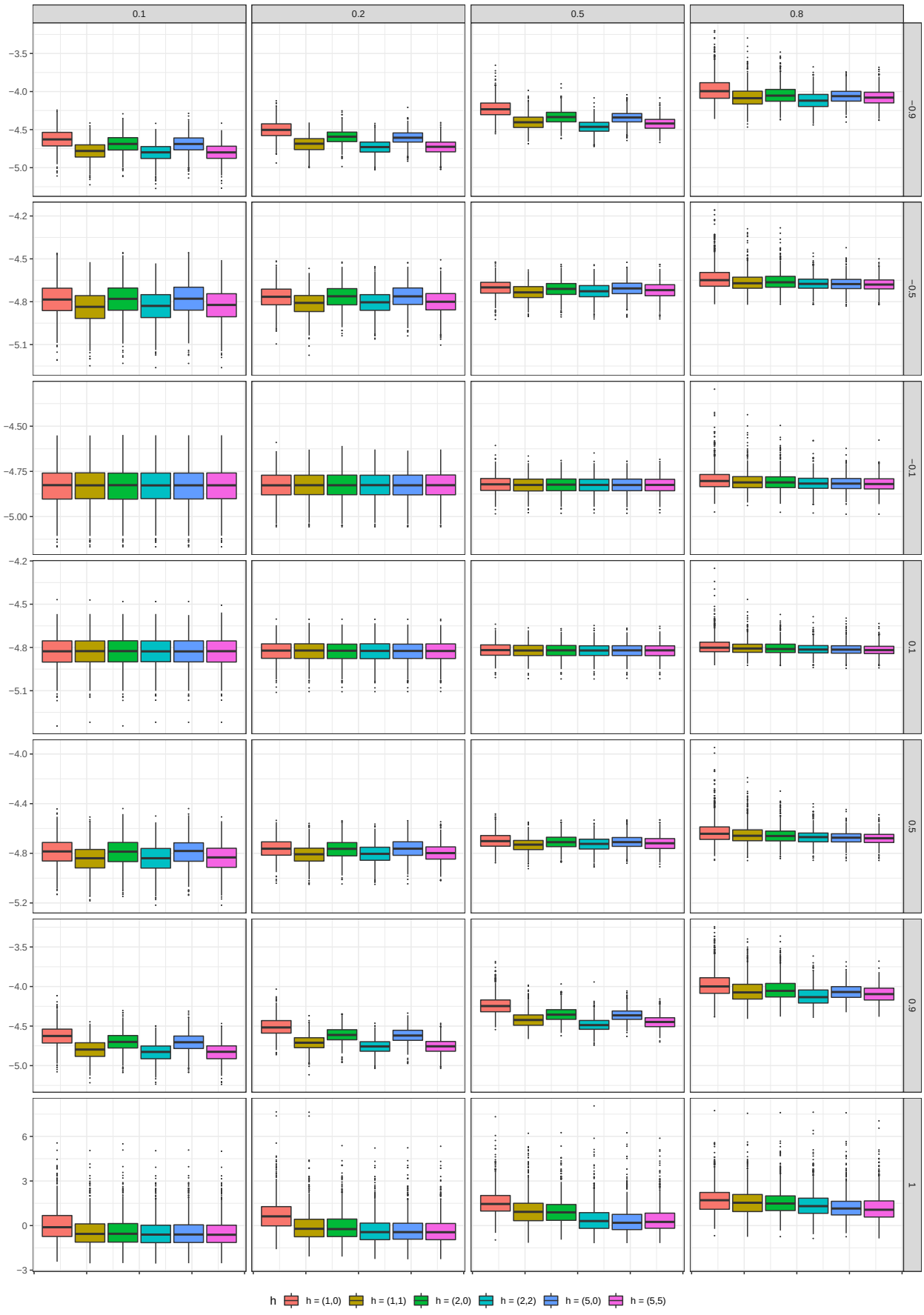

Figura 3.5: Box-plots do logaritmo do EAPM para o modelo AR(1) com $n = 500$, $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

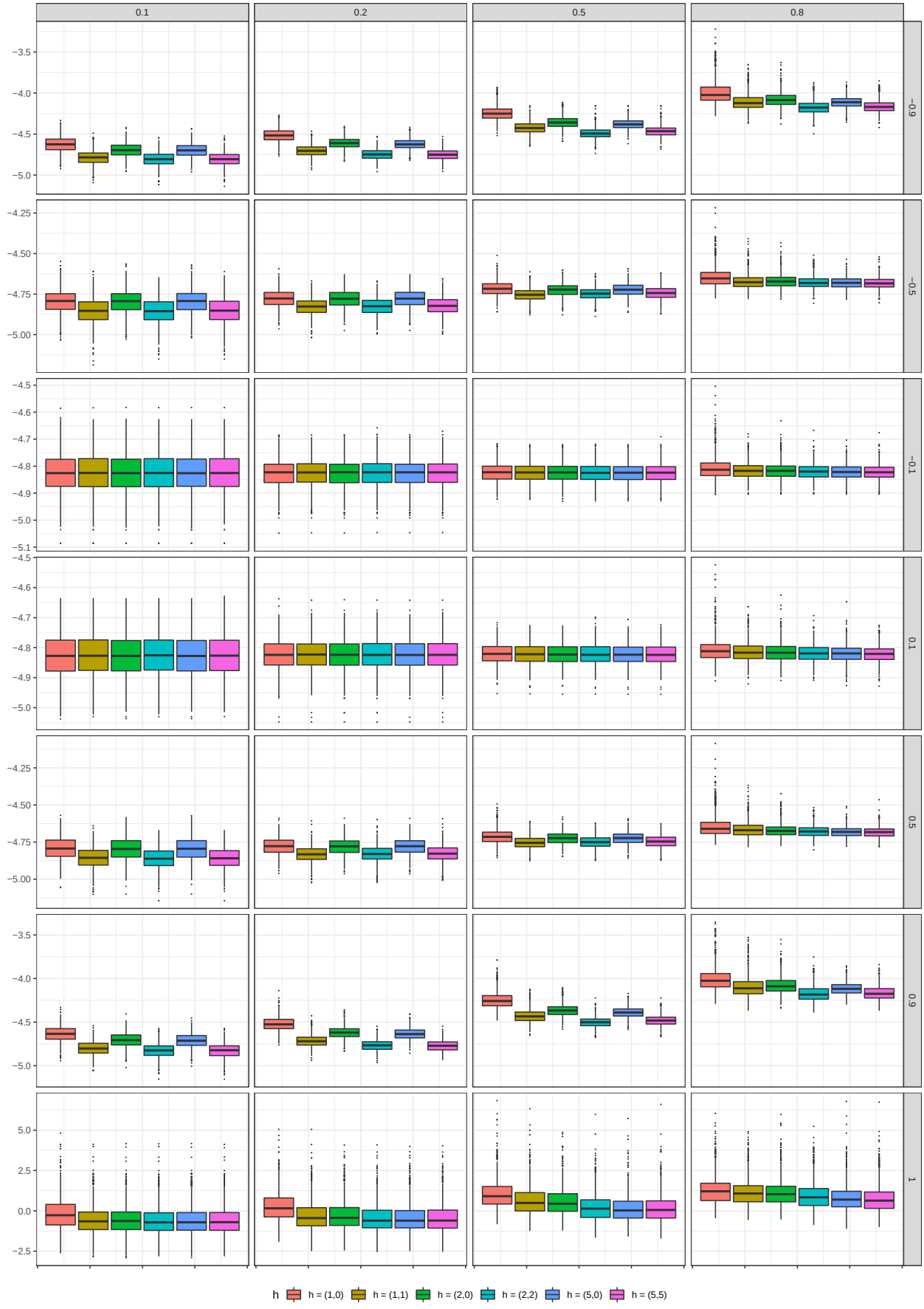

Figura 3.6: Box-plots do logaritmo do EAPM do modelo AR(1) com $n = 1000$, $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

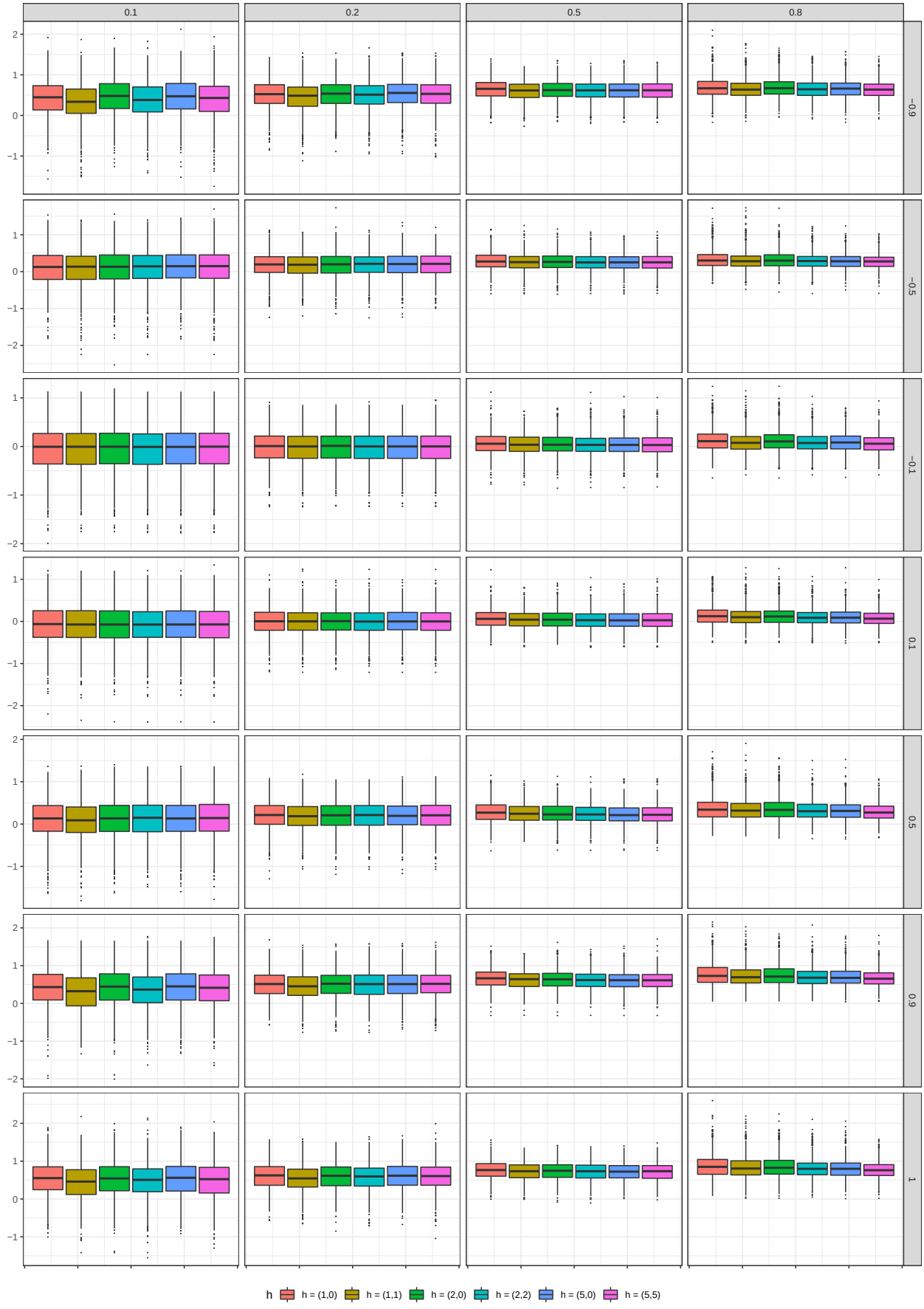

Figura 3.7: Box-plots do logaritmo do EQM para o modelo MA(1) com $n = 100$, $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

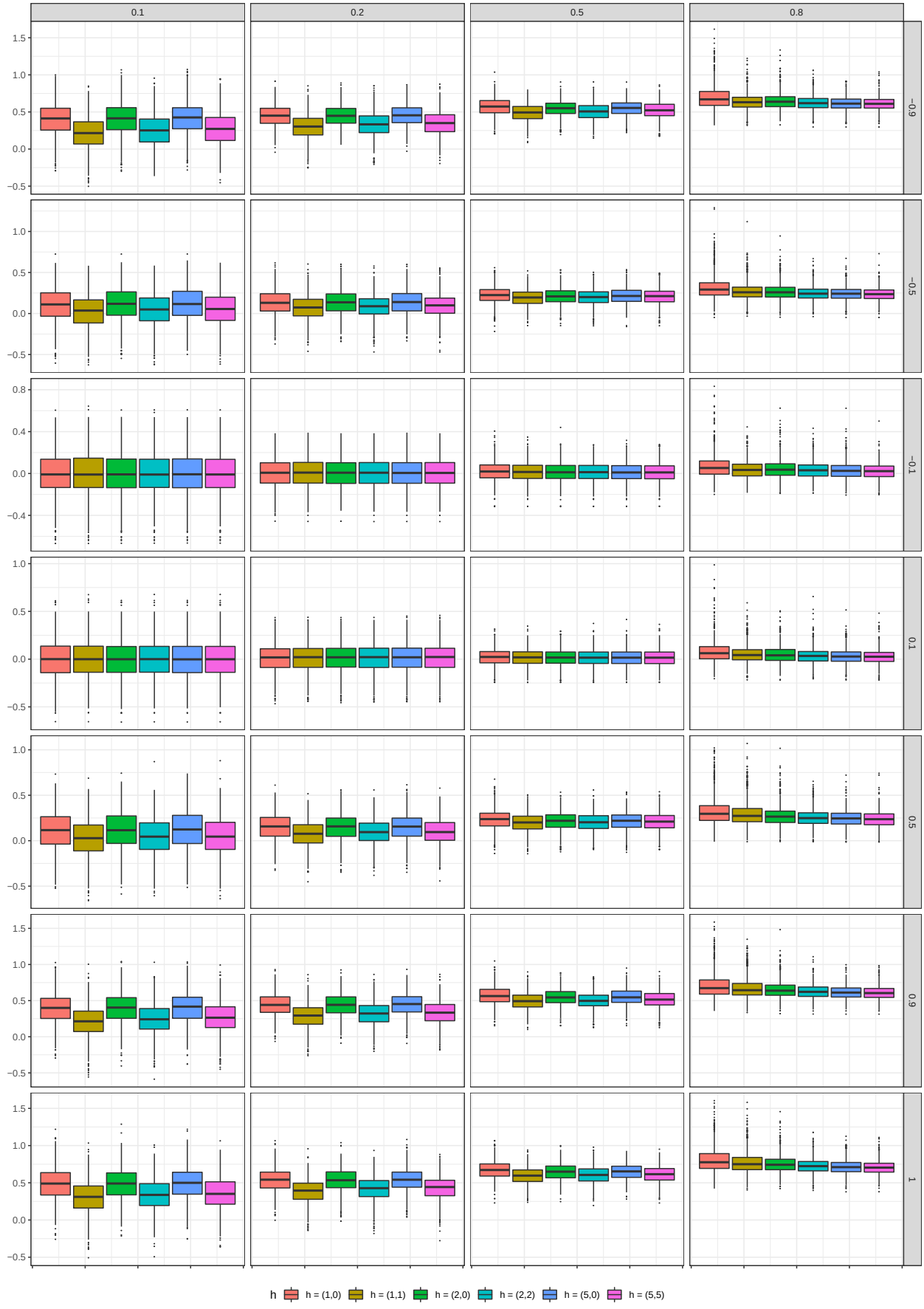

Figura 3.8: Box-plots do logaritmo do EQM para o modelo MA(1) com $n = 500$, $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

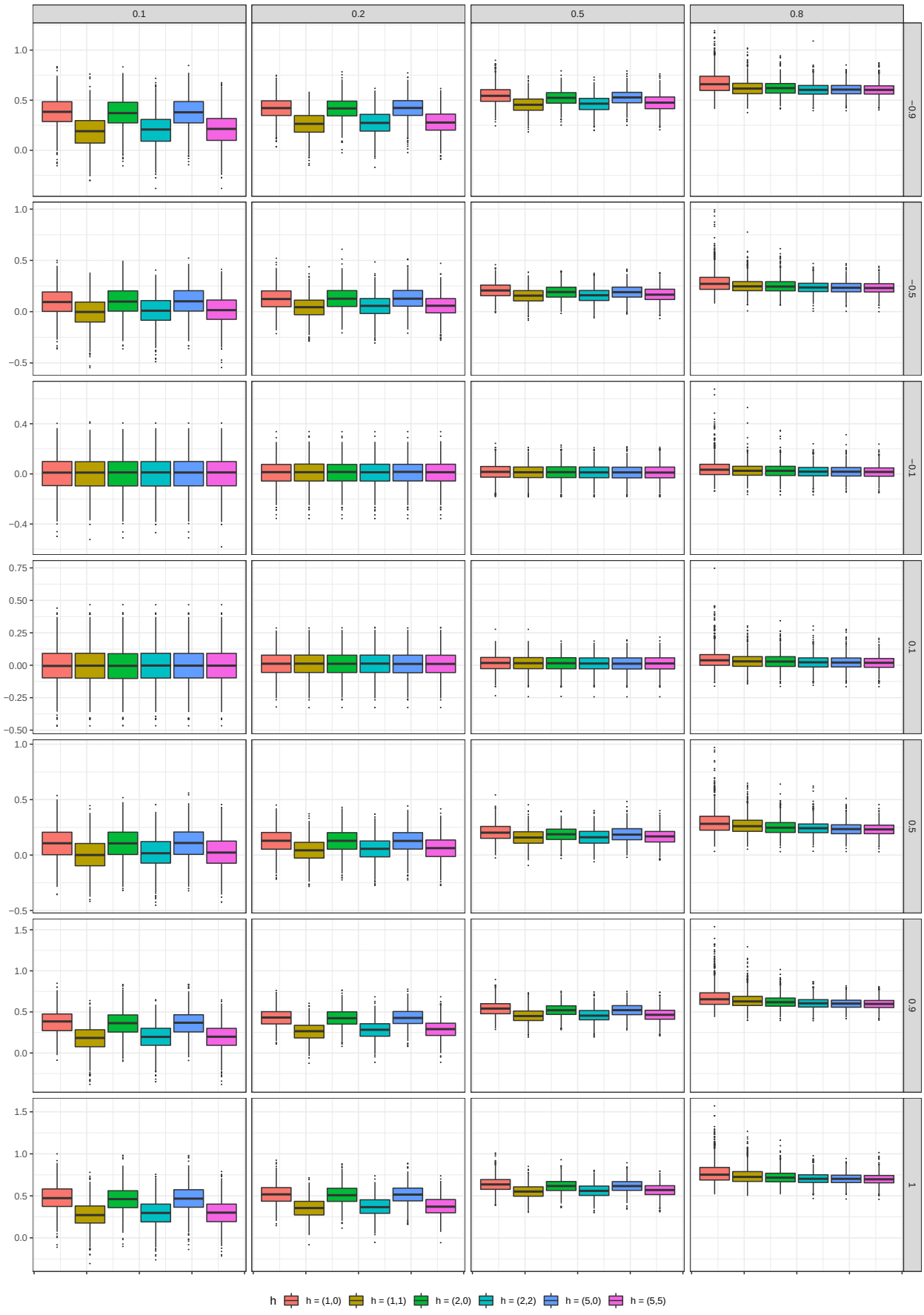

Figura 3.9: Box-plots do logaritmo do EQM para o modelo MA(1) com $n = 1000$, $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

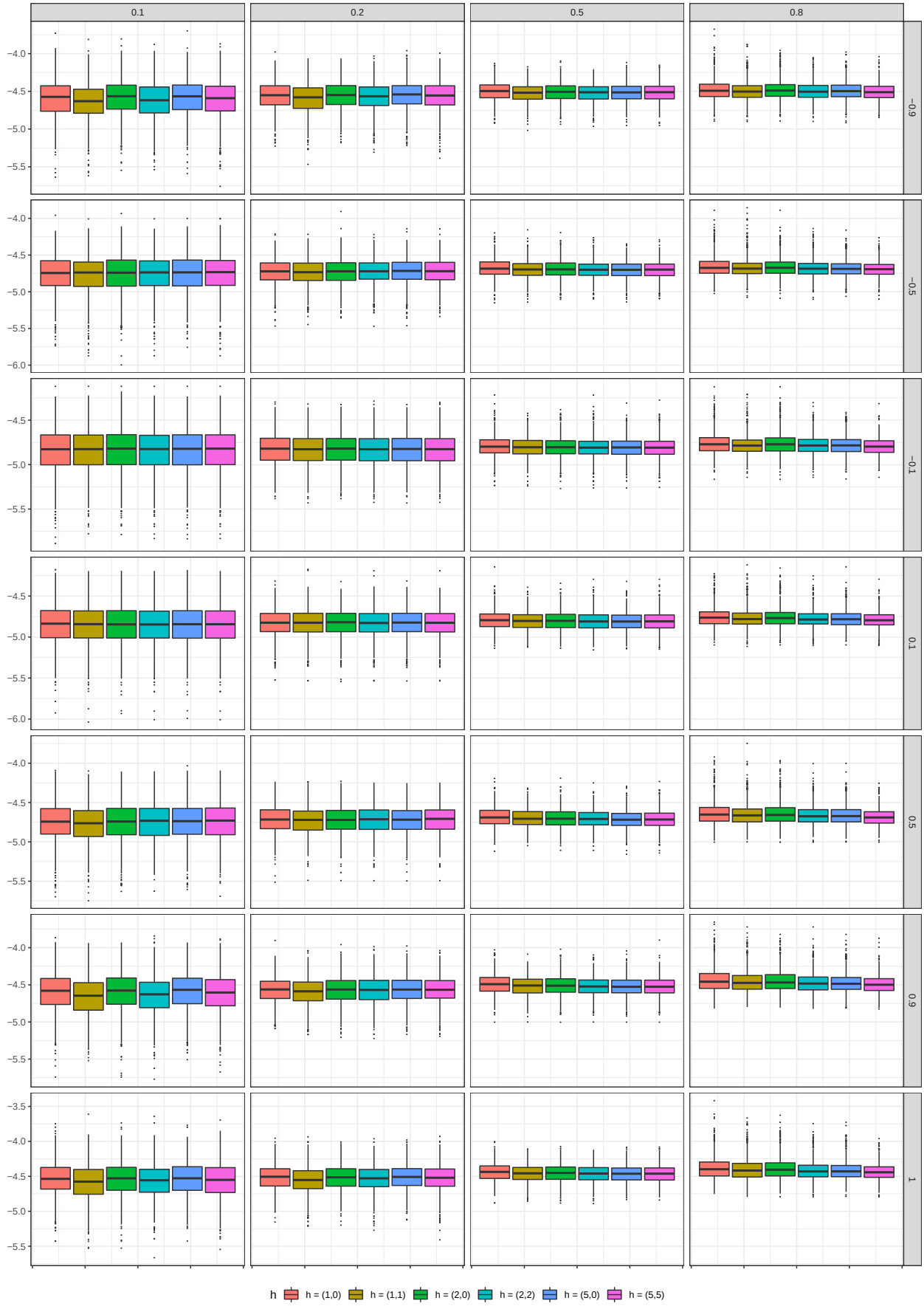

Figura 3.10: Box-plots do logaritmo do EAPM do modelo MA(1) com $n = 100$, $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

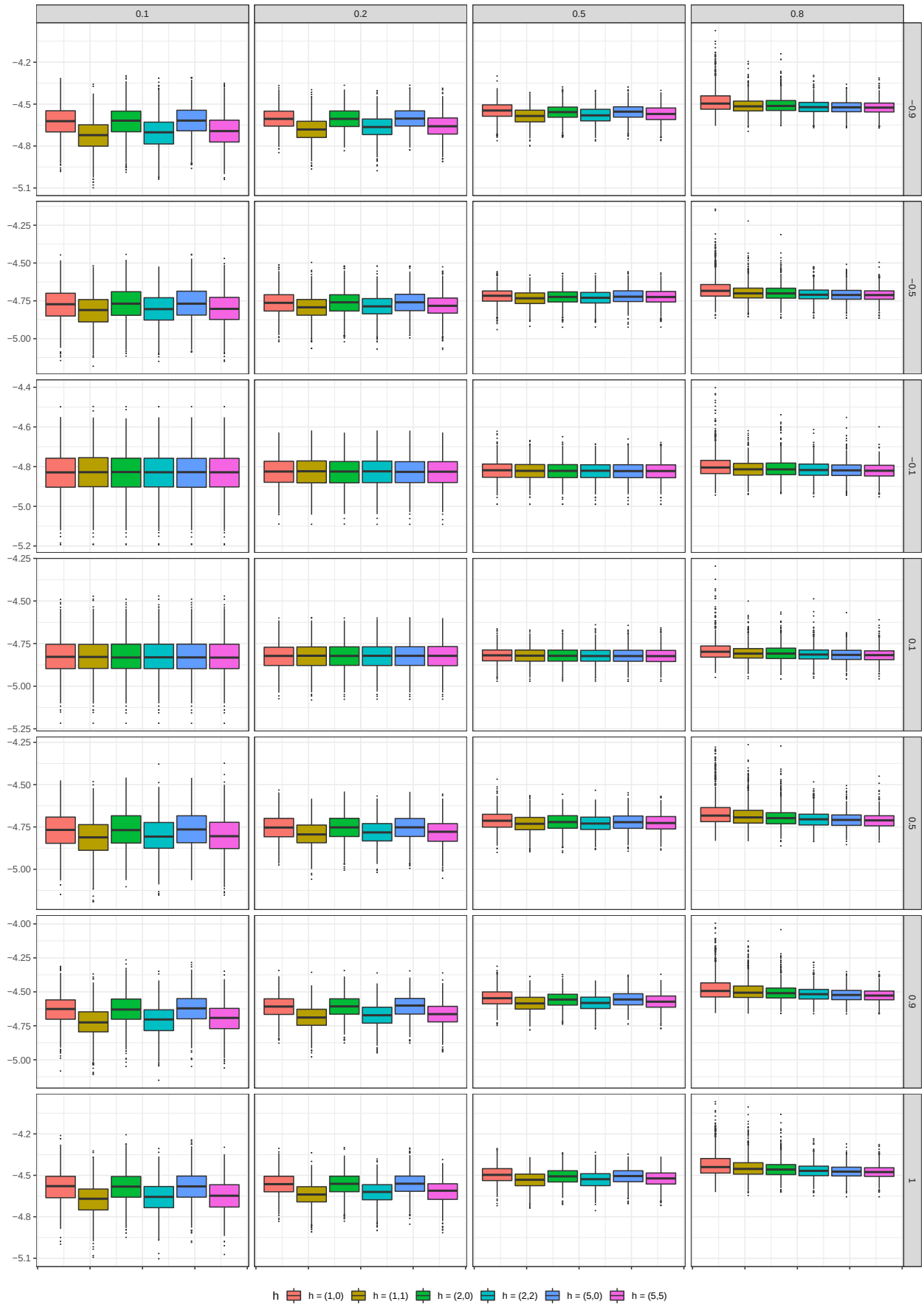

Figura 3.11: Box-plots do logaritmo do EAPM do modelo MA(1) com $n = 500$, $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

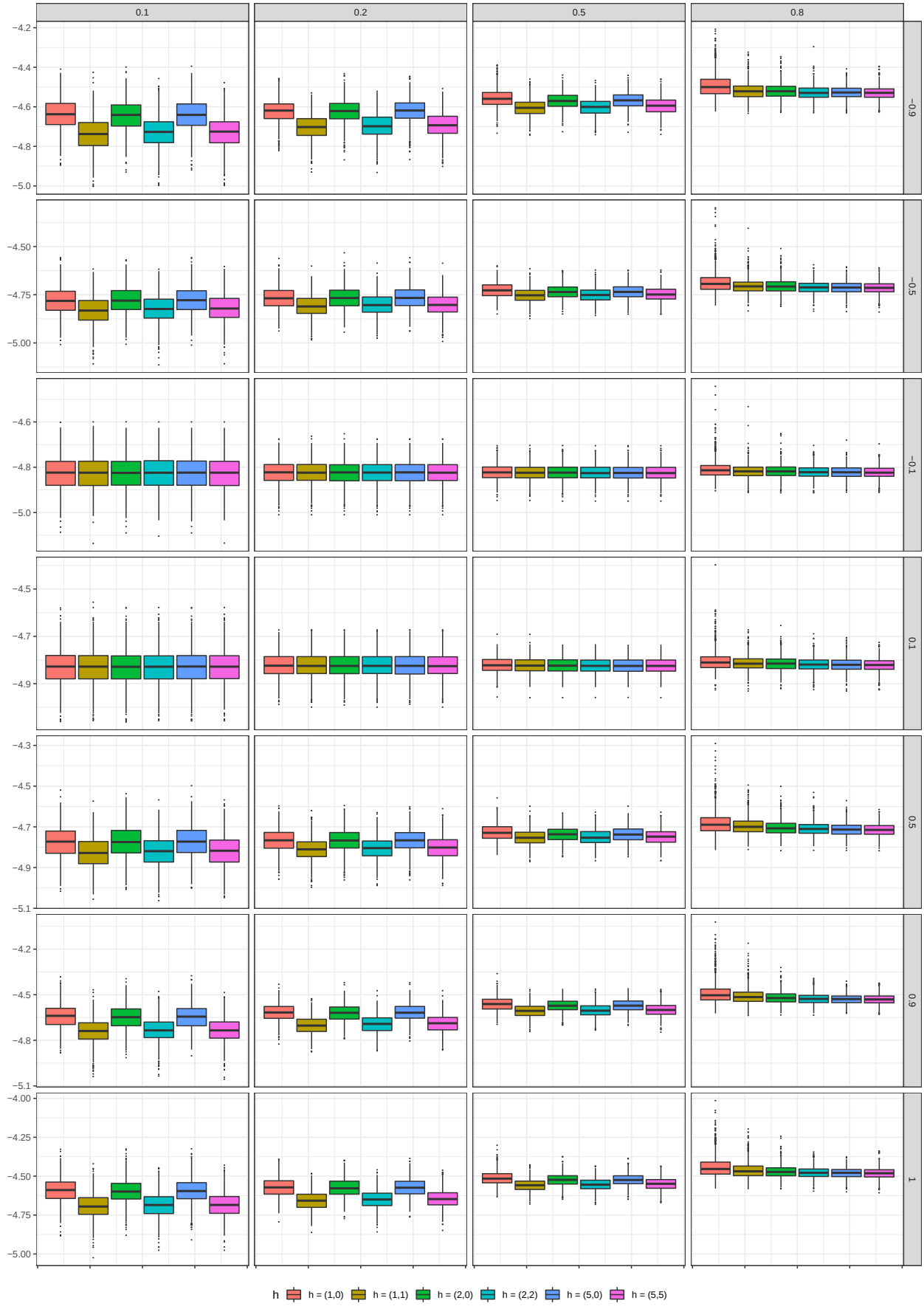

Figura 3.12: Box-plots do logaritmo do EAPM do modelo MA(1) com $n = 1000$, $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas) utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$ (cada painel).

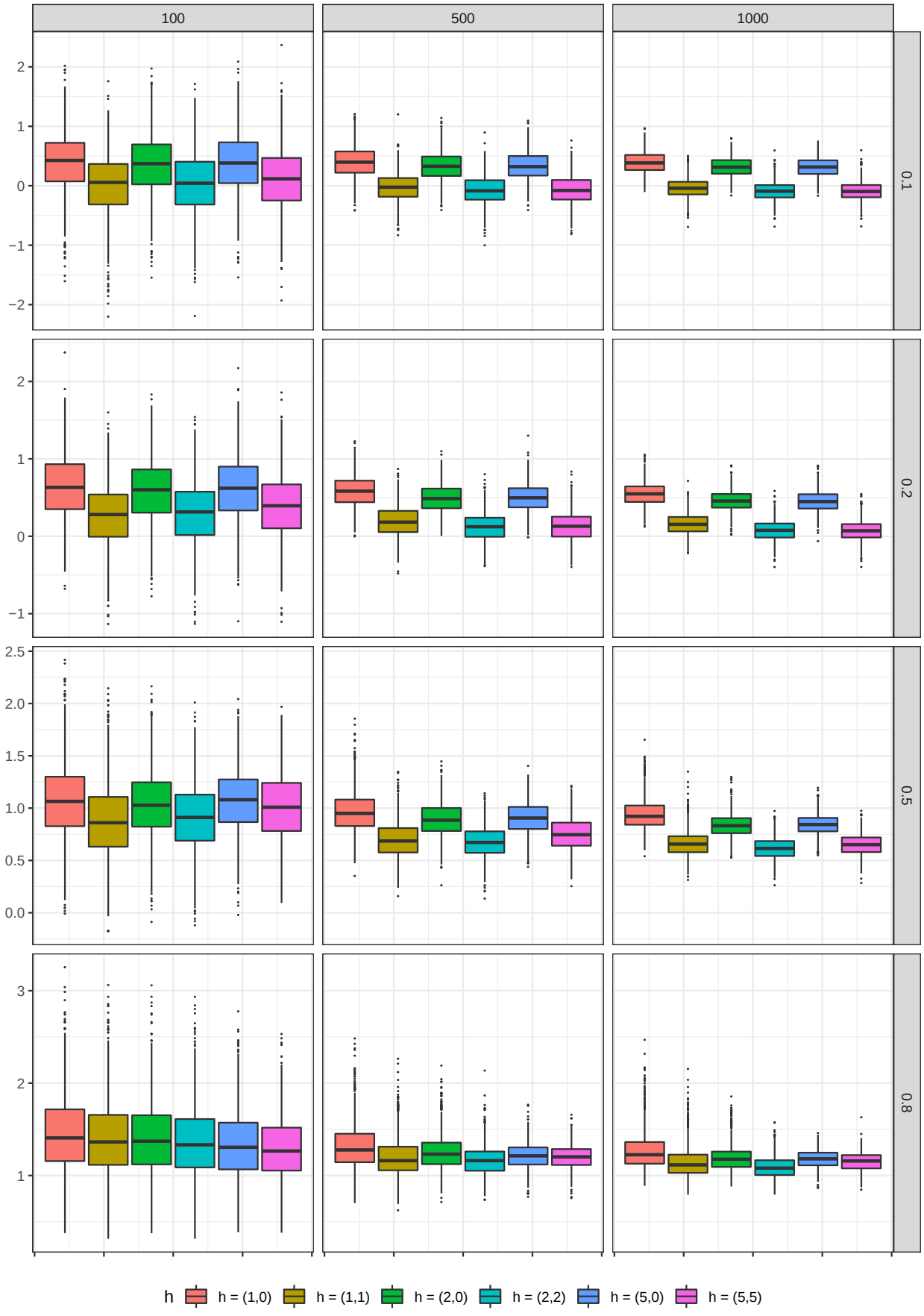

Figura 3.13: Box-plots do logaritmo do EQM do modelo ARMA$(1,1)$ com $n \in \{100, 500, 1000\}$, $\phi = 0.7$, $\theta = 0.4$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$.

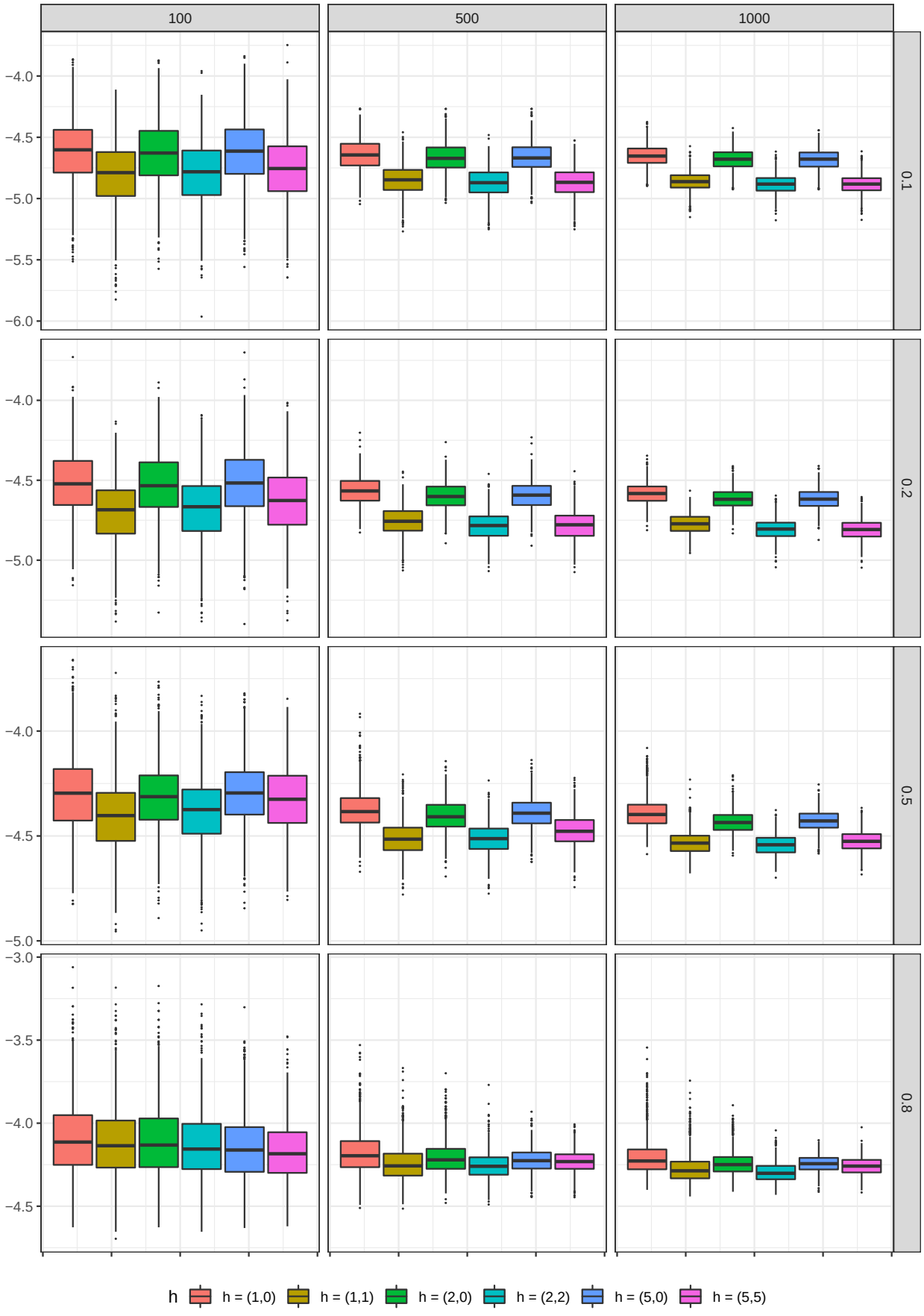

Figura 3.14: Box-plots do logaritmo do EAPM do modelo temporais ARMA$(1,1)$ com $n \in \{100, 500, 1000\}$, $\phi = 0.7$, $\theta = 0.4$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ utilizando como método de imputação árvores de decisão com $\boldsymbol{h} \in \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$.

3.6 Comparações entre métodos

Nesta seção, o método de preenchimento de valores faltantes utilizando árvores de decisão com número de covariáveis $\boldsymbol{h} = (5, 5)$ foi comparado com os métodos de imputação listados na Seção 2.2.2. Alguns nomes dos métodos foram reduzidos nas tabelas para melhor visualização. As siglas utilizadas são listadas a seguir:

- Int. linear: interpolação linear;
- Int. spline: interpolação por splines;
- Int. Stine: interpolação de Stineman;
- MM. simples: médias móveis com peso simples;
- MM linear: médias móveis com peso linear;
- MM exp.: médias móveis com peso exponencial;
- LOCF/NOCB: última observação levada adiante/Próxima observação trazida para trás;
- Vero.: suavização de Kalman com modelo estrutural estimado via máxima verossimilhança;
- Autoarima.: suavização de Kalman usando uma representação do espaço de estados do modelo ARIMA.

Nas Tabelas 3.1 e 3.2 são apresentados os 3 métodos que obtiveram os melhores resultados no preenchimento de valores faltantes em termos de erro quadrático médio, calculados a partir de 1000 replicações de amostras de tamanho $n \in \{100, 500, 1000\}$ de séries temporais AR(1) com $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ e MA(1) com $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas). De forma similar, nas Tabelas 3.3 e 3.4, são apresentados esses resultados em termos do erro percentual absoluto médio. Já a Tabela 3.5 apresenta os 10 métodos que obtiveram os melhores resultados em termos de ambos os erros calculados a partir de 1000 replicações de amostras de tamanho $n \in \{100, 500, 1000\}$ de séries temporais ARMA$(1, 1)$ com $\phi = 0.7$ e $\theta = 0.4$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas).

De forma geral observa-se que o EQM aumenta com ρ para todos os métodos, porém, diferentemente das árvores de decisão, o aumento do tamanho da amostra não parece afetar a qualidade de previsão dos outros métodos testados. Os valores do EAPM apresentam um comportamento semelhante ao EQM. No que segue, são descritos os comportamentos das métricas para cada modelo separadamente, em termos dos parâmetros ϕ e θ.

Modelo AR(1):

- $\phi = -0.9$: os 3 métodos com o melhor desempenho foram as árvores, média e mediana. O valor das métricas para as árvores foi notavelmente menor do que o calculado para a média e mediana, exceto no caso em que $n = 100$ e $\rho = 0.8$.

- $\phi = -0.5$: os 3 métodos com o melhor desempenho foram as árvores, média e mediana. O valor das métricas para estes 3 métodos foi similar, com as árvores de decisão obtendo os melhores resultados com uma baixa proporção de valores faltantes ($\rho = 0.1$ e 0.2) para todos os tamanhos de amostra e com $\rho = 0.5$, nos casos em que $n \in \{500, 1000\}$.

- $|\phi| = 0.1$: os 3 métodos com o melhor desempenho foram as árvores, média e mediana. Não parece haver real diferença entre a qualidade de preenchimento de valores faltantes neste caso.

- $\phi = 0.5$: nos casos em que a proporção de valores faltantes foi $\rho \in \{0.1, 0.2, 0.5\}$, os métodos com os menores valores para as métricas calculadas foram os de interpolação (linear e de Stineman) e de médias móveis (com peso simples e exponencial), porém no caso com maior proporção de valores faltantes ($\rho = 0.8$), o método de árvores de decisão aparece como uma das melhores alternativas, junto da média e da mediana.

- $\phi \in \{0.9, 1\}$: no caso mais próximo a não estacionariedade e no passeio aleatório, os métodos de preenchimento de valores faltantes com os melhores resultados foram os de interpolação (linear, por splines e de Stineman) e o método de suavização de Kalman, estimado por máxima verossimilhança.

Modelo MA(1):

- $\rho = 0.8$: para qualquer valor de n e θ, a Média sempre apareceu como o melhor método;

- $\theta = -0.9$: os 3 métodos com o melhor desempenho foram as árvores, média e mediana. O valor das métricas para estes 3 métodos foi similar, com as árvores de decisão obtendo os melhores resultados com uma baixa proporção de valores faltantes ($\rho = 0.1$ e 0.2) para todos os tamanhos de amostra e com $\rho = 0.5$, nos casos em que $n \in \{500, 1000\}$.

- $\theta = -0.5$: os 3 métodos com o melhor desempenho foram as árvores, média e mediana. O valor das métricas para estes 3 métodos foi similar, com as árvores de decisão obtendo os melhores resultados com uma baixa proporção de valores faltantes ($\rho = 0.1$) para todos os tamanhos de amostra e com $\rho \in \{0.2, 0.5\}$, nos casos em que $n \in \{500, 1000\}$;

- $|\theta| = 0.1$: os 3 métodos com o melhor desempenho foram as árvores, média e mediana. Não parece haver real diferença entre a qualidade de preenchimento de valores faltantes neste caso;

- $\theta = 0.5$: as árvores tiveram o terceiro melhor desempenho quando $\rho = 0.8$ e também apareceram entre os melhores resultados em todos os cenários com $n \in \{500, 1000\}$;

- $\theta \in \{0.9, 1\}$: as árvores tiveram o terceiro melhor desempenho quando $\rho = 0.8$ e também apareceram entre os melhores resultados quando $\rho = 0.5$ com $n \in \{500, 1000\}$.

Modelo ARMA(1,1):

- Os métodos que apresentaram melhor desempenho neste cenário foram os de interpolação, suavização de Kalman (verossimilhança) e médias móveis, enquanto as árvores de decisão mantiveram-se entre a 5ª e a 10ª melhor opção de método de reconstrução destas séries temporais. Nota-se que o desempenho das árvores de decisão melhora (em termos da posição delas no ordenamento) conforme aumenta o tamanho da amostra, enquanto os métodos que estão no top 4 se mantém estáveis. Observa-se ainda que, a medida que a proporção de dados faltantes aumenta, a diferença no valores das métricas referentes à 1ª e à 10ª diminui (proporcionalmente).

Tabela 3.1: Tabela apresentando os 3 melhores métodos de preenchimento de valores faltantes testados em termos de erro quadrático médio calculado a partir de 1000 replicações para os valores faltantes de amostras de tamanho $n \in \{100, 500, 1000\}$ de séries temporais AR(1) com $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas)

ϕ	Métodos	n = 100				n = 500				n = 1000			
		ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8
-0.9	Top 1	Árvores	Árvores	Árvores	Média	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores
	Top 2	1.417	1.677	3.395	5.447	1.116	1.283	2.423	4.657	1.079	1.212	2.2	3.982
		Média	Média	Média	Mediana	Média	Média	Média	Média	Média	Média	Média	Média
		5.435	5.364	5.368	5.669	5.309	5.331	5.345	5.379	5.29	5.297	5.299	5.314
	Top 3	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana
		5.502	5.438	5.447	5.697	5.312	5.337	5.355	5.404	5.293	5.298	5.304	5.329
-0.5	Top 1	Árvores	Árvores	Média	Média	Árvores	Árvores	Árvores	Média	Árvores	Árvores	Árvores	Média
		1.225	1.297	1.362	1.377	1.01	1.087	1.262	1.347	0.964	1.021	1.197	1.343
	Top 2	Média	Média	Mediana	Mediana	Média	Média	Média	Mediana	Média	Média	Média	Mediana
		1.326	1.35	1.379	1.413	1.329	1.343	1.339	1.353	1.329	1.337	1.335	1.346
	Top 3	Mediana	Mediana	Árvores	Árvores	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana	Árvores
		1.337	1.359	1.413	1.457	1.33	1.345	1.342	1.364	1.329	1.338	1.337	1.351
-0.1	Top 1	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média
		1.016	1.024	1.025	1.058	1	1.006	1.013	1.018	1.016	1.012	1.01	1.015
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana	Mediana
		1.023	1.033	1.034	1.087	1.001	1.008	1.016	1.022	1.017	1.013	1.011	1.017
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Mediana	Árvores	Árvores	Árvores	Árvores
		1.031	1.039	1.05	1.093	1.002	1.01	1.019	1.023	1.017	1.014	1.013	1.018
0.1	Top 1	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média
		1.032	1.024	1.036	1.068	1.019	1.012	1.014	1.023	1.019	1.015	1.012	1.017
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores
		1.037	1.03	1.045	1.088	1.02	1.013	1.016	1.028	1.02	1.016	1.013	1.02
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Mediana
		1.044	1.036	1.061	1.1	1.022	1.015	1.018	1.028	1.022	1.017	1.014	1.02
0.5	Top 1	Int. linear	Int. linear	MM exp.	Média	Int. linear	Int. linear	MM exp.	Média	Int. linear	Int. linear	MM exp.	Média
		0.87	0.926	1.167	1.408	0.884	0.925	1.145	1.343	0.873	0.93	1.142	1.339
	Top 2	Int. Stine	Int. Stine	Int. linear	Mediana	Int. Stine	Int. Stine	Int. linear	Mediana	Int. Stine	MM exp.	Int. linear	Mediana
		0.887	0.946	1.17	1.44	0.901	0.947	1.151	1.349	0.888	0.951	1.147	1.343
	Top 3	MM exp.	MM exp.	Vero.	Árvores	MM exp.	MM exp.	MM linear	Árvores	MM exp.	Int. Stine	MM linear	Árvores
		0.913	0.951	1.181	1.488	0.919	0.947	1.172	1.355	0.91	0.953	1.17	1.349
0.9	Top 1	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear
		0.577	0.614	0.916	2.416	0.583	0.616	0.873	1.849	0.568	0.613	0.872	1.8
	Top 2	Int. Stine	Int. Stine	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.
		0.583	0.623	0.935	2.483	0.585	0.619	0.88	1.866	0.57	0.616	0.877	1.814
	Top 3	Vero.	Vero.	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine
		0.586	0.625	0.952	2.626	0.59	0.628	0.901	1.942	0.575	0.624	0.9	1.889
1	Top 1	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline
		89.206	118.862	420.654	10665.381	101.4	129.582	382.129	5532.162	99.471	125.108	371.971	4963.361
	Top 2	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Vero.	Int. Stine	Int. Stine	Int. Stine	Vero.	Vero.	Int. Stine	Int. Stine
		102.133	132.268	477.478	13183.654	109.572	140.272	445.958	7013.933	105.256	137.413	436.686	6826.628
	Top 3	Vero.	Vero.	Vero.	Vero.	Int. Stine	Vero.	Vero.	Vero.	Int. Stine	Int. Stine	Vero.	Vero.
		104.406	135.343	756.1	19566.309	109.659	142.439	553.666	16918.752	107.203	137.986	527.392	9261.558

Tabela 3.2: Tabela apresentando os 3 melhores métodos de preenchimento de valores faltantes testados em termos de erro quadrático médio calculado a partir de 1000 replicações para os valores faltantes de amostras de tamanho $n \in \{100, 500, 1000\}$ de séries temporais MA(1) com $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas)

		n = 100				n = 500				n = 1000			
θ	Métodos	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8
-0.9	Top 1	Árvores	Árvores	Média	Média	Árvores	Árvores	Árvores	Média	Árvores	Árvores	Árvores	Média
		1.651	1.667	1.756	1.86	1.32	1.342	1.623	1.835	1.173	1.273	1.526	1.825
	Top 2	Média	Média	Mediana	Mediana	Média	Média	Média	Mediana	Média	Média	Média	Mediana
		1.851	1.742	1.782	1.899	1.813	1.717	1.735	1.845	1.722	1.729	1.726	1.829
	Top 3	Mediana	Mediana	Árvores	Árvores	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana	Árvores
		1.864	1.759	1.815	1.922	1.815	1.72	1.737	1.851	1.723	1.73	1.728	1.834
-0.5	Top 1	Árvores	Média	Média	Média	Árvores	Árvores	Árvores	Média	Árvores	Árvores	Árvores	Média
		1.253	1.259	1.269	1.29	1.064	1.119	1.238	1.263	1.03	1.071	1.187	1.258
	Top 2	Média	Mediana	Mediana	Mediana	Média	Média	Média	Árvores	Média	Média	Média	Mediana
		1.28	1.269	1.281	1.322	1.247	1.258	1.257	1.27	1.251	1.256	1.255	1.261
	Top 3	Mediana	Árvores	Árvores	Árvores	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores
		1.287	1.271	1.305	1.339	1.249	1.26	1.261	1.271	1.252	1.257	1.257	1.264
-0.1	Top 1	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média
		1.027	1.034	1.027	1.056	1.014	1.013	1.011	1.016	1.011	1.014	1.012	1.014
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana
		1.035	1.041	1.037	1.081	1.015	1.014	1.014	1.022	1.012	1.014	1.014	1.017
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores
		1.044	1.051	1.051	1.084	1.019	1.016	1.016	1.022	1.012	1.016	1.015	1.017
0.1	Top 1	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média	Média
		1.032	1.021	1.034	1.064	1.006	1.011	1.015	1.02	1.011	1.013	1.013	0.977
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores
		1.038	1.028	1.047	1.087	1.008	1.013	1.017	1.026	1.011	1.014	1.015	0.98
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Mediana
		1.048	1.036	1.059	1.096	1.01	1.016	1.02	1.027	1.013	1.015	1.016	0.98
0.5	Top 1	Int. linear	Int. linear	Média	Média	Int. Stine	Int. linear	Árvores	Média	Int. linear	Int. linear	Árvores	Média
		0.88	1.006	1.191	1.255	0.897	0.998	1.173	1.201	0.908	0.992	1.129	1.226
	Top 2	Int. Stine	Int. Stine	Mediana	Mediana	Int. linear	Int. Stine	Média	Mediana	Int. Stine	Int. Stine	Média	Mediana
		0.882	1.017	1.206	1.288	0.898	1.005	1.193	1.207	0.909	1.001	1.189	1.229
	Top 3	MM exp.	MM exp.	Vero.	Árvores	Árvores	Árvores	Vero.	Árvores	Árvores	Árvores	Mediana	Árvores
		0.989	1.069	1.21	1.309	1.004	1.057	1.196	1.209	0.974	1.012	1.191	1.231
0.9	Top 1	Int. spline	Int. Stine	Int. linear	Média	Int. spline	Int. Stine	Árvores	Média	Int. spline	Int. Stine	Árvores	Média
		0.884	1.224	1.805	1.906	0.879	1.21	1.693	1.824	0.896	1.197	1.6	1.816
	Top 2	Int. Stine	Int. linear	Vero.	Mediana	Int. Stine	Int. linear	Int. linear	Mediana	Int. Stine	Int. linear	Int. linear	Mediana
		1.02	1.239	1.812	1.947	1.009	1.225	1.738	1.832	1.036	1.209	1.738	1.82
	Top 3	Int. linear	Int. spline	Média	Árvores	Int. linear	Int. spline	Int. Stine	Árvores	Int. linear	Int. spline	Int. Stine	Árvores
		1.043	1.291	1.823	1.984	1.031	1.274	1.785	1.839	1.061	1.269	1.784	1.825
1	Top 1	Int. spline	Int. Stine	Int. linear	Média	Int. spline	Int. Stine	Árvores	Média	Int. spline	Int. Stine	Árvores	Média
		0.966	1.348	1.976	2.117	0.975	1.324	1.872	2.013	0.966	1.319	1.77	2.01
	Top 2	Int. Stine	Int. linear	Vero.	Mediana	Int. Stine	Int. linear	Int. linear	Mediana	Int. Stine	Int. linear	Int. linear	Mediana
		1.144	1.36	1.995	2.167	1.139	1.337	1.912	2.023	1.133	1.335	1.913	2.016
	Top 3	Int. linear	Int. spline	MM exp.	Árvores	Int. linear	Int. spline	Int. Stine	Árvores	Int. linear	Int. spline	Int. Stine	Árvores
		1.172	1.41	2.026	2.222	1.163	1.404	1.961	2.03	1.16	1.379	1.963	2.02

Tabela 3.3: Tabela apresentando os 3 melhores métodos de preenchimento de valores faltantes testados em termos de erro absoluto percentual médio calculado a partir de 1000 replicações para os valores faltantes de amostras de tamanho $n \in \{100, 500, 1000\}$ de séries temporais AR(1) com $\phi \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas)

		n = 100				n = 500				n = 1000			
ϕ	Métodos	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8
-0.9	Top 1	Árvores	Árvores	Árvores	Media	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores
		0.009	0.01	0.014	0.019	0.008	0.009	0.012	0.017	0.008	0.009	0.011	0.016
	Top 2	Media	Media	Media	Árvores	Media	Media	Media	Media	Media	Media	Media	Media
		0.019	0.019	0.019	0.019	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018
	Top 3	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana
		0.019	0.019	0.019	0.019	0.018	0.018	0.018	0.019	0.018	0.018	0.018	0.018
-0.5	Top 1	Árvores	Árvores	Media	Media	Árvores	Árvores	Árvores	Media	Árvores	Árvores	Árvores	Media
		0.009	0.009	0.009	0.009	0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009
	Top 2	Media	Media	Mediana	Mediana	Media	Media	Media	Mediana	Media	Media	Media	Mediana
		0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
	Top 3	Mediana	Mediana	Árvores	Árvores	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana	Árvores
		0.009	0.009	0.009	0.01	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
-0.1	Top 1	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana	Mediana
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Mediana	Árvores	Árvores	Árvores	Árvores
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
0.1	Top 1	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Mediana
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
0.5	Top 1	Int. linear	Int. linear	MM exp.	Media	Int. linear	Int. linear	MM exp.	Media	Int. linear	Int. linear	MM exp.	Media
		0.007	0.008	0.009	0.009	0.007	0.008	0.008	0.009	0.007	0.008	0.008	0.009
	Top 2	Int. Stine	Int. Stine	Int. linear	Mediana	Int. Stine	Int. Stine	Int. linear	Mediana	Int. Stine	Int. Stine	Int. linear	Mediana
		0.007	0.008	0.009	0.01	0.008	0.008	0.008	0.009	0.008	0.008	0.008	0.009
	Top 3	MM exp.	MM exp.	Vero.	Árvores	MM exp.	MM exp.	MM linear	Árvores	MM exp.	MM exp.	MM linear	Árvores
		0.008	0.008	0.009	0.01	0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009
0.9	Top 1	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear	Int. linear
		0.006	0.006	0.008	0.012	0.006	0.006	0.008	0.011	0.006	0.006	0.007	0.011
	Top 2	Int. Stine	Int. Stine	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.	Vero.
		0.006	0.006	0.008	0.012	0.006	0.006	0.008	0.011	0.006	0.006	0.008	0.011
	Top 3	Vero.	Vero.	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine	Int. Stine
		0.006	0.006	0.008	0.013	0.006	0.006	0.008	0.011	0.006	0.006	0.008	0.011
1	Top 1	Int. spline	Int. spline	Int. spline	Int. spline	Int. spline	Int. Stine	Int. spline	Int. Stine	Vero.	Int. spline	Int. spline	Int. spline
		0.018	0.022	0.04	0.104	0.003	0.005	0.008	0.018	0.002	0.003	0.004	0.01
	Top 2	Int. Stine	Vero.	Vero.	Vero.	Int. Stine	Vero.	Int. Stine	Int. spline	Int. spline	Vero.	Int. Stine	Int. Stine
		0.018	0.023	0.043	0.113	0.003	0.005	0.009	0.018	0.003	0.003	0.005	0.011
	Top 3	Vero.	Autoarima	Int. Stine	Int. Stine	Vero.	Int. spline	Vero.	Int. linear	Int. Stine	Int. Stine	Vero.	Int. linear
		0.019	0.024	0.044	0.141	0.003	0.005	0.009	0.026	0.003	0.003	0.005	0.015

Tabela 3.4: Tabela apresentando os 3 melhores métodos de preenchimento de valores faltantes testados em termos de erro absoluto percentual médio calculado a partir de 1000 replicações para os valores faltantes de amostras de tamanho $n \in \{100, 500, 1000\}$ de séries temporais MA(1) com $\theta \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9, 1\}$ (linhas) e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas)

θ	Métodos	n = 100				n = 500				n = 1000			
		ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8	ρ =0.1	ρ =0.2	ρ =0.5	ρ =0.8
-0.9	Top 1	Árvores	Árvores	Media	Media	Árvores	Árvores	Árvores	Media	Árvores	Árvores	Árvores	Media
		0.01	0.011	0.011	0.011	0.009	0.009	0.01	0.011	0.009	0.009	0.01	0.011
	Top 2	Media	Media	Mediana	Mediana	Media	Media	Media	Mediana	Media	Media	Media	Mediana
		0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
	Top 3	Mediana	Mediana	Árvores	Árvores	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana	Árvores
		0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
-0.5	Top 1	Árvores	Media	Media	Media	Árvores	Árvores	Árvores	Media	Árvores	Árvores	Árvores	Media
		0.009	0.009	0.009	0.009	0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009
	Top 2	Media	Mediana	Mediana	Mediana	Media	Media	Media	Árvores	Media	Media	Media	Mediana
		0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
	Top 3	Mediana	Árvores	Árvores	Árvores	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores
		0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
-0.1	Top 1	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores	Mediana	Mediana	Mediana
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Mediana	Árvores	Árvores	Árvores
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
0.1	Top 1	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media	Media
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 2	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Mediana	Árvores
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	Top 3	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Árvores	Mediana
		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
0.5	Top 1	Int. linear	Int. linear	Media	Media	Int. Stine	Int. linear	Árvores	Media	Int. linear	Int. linear	Árvores	MM simples
		0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009	0.008	0.008	0.009	-0.265
	Top 2	Int. Stine	Int. Stine	Mediana	Mediana	Int. linear	Int. Stine	Media	Mediana	Int. Stine	Int. Stine	Media	Media
		0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009
	Top 3	Int. spline	MM exp.	Vero.	Árvores	Int. spline	Árvores	Vero.	Árvores	Árvores	Árvores	Mediana	Mediana
		0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009	0.008	0.008	0.009	0.009
0.9	Top 1	Int. spline	Int. spline	Int. linear	Media	Int. spline	Int. spline	Int. linear	Media	Int. spline	Int. spline	Árvores	Media
		0.007	0.009	0.011	0.011	0.007	0.009	0.01	0.011	0.007	0.009	0.01	0.011
	Top 2	Int. Stine	Int. Stine	Int. Stine	Mediana	Int. Stine	Int. Stine	Árvores	Mediana	Int. Stine	Int. Stine	Int. linear	Mediana
		0.008	0.009	0.011	0.011	0.008	0.009	0.01	0.011	0.008	0.009	0.01	0.011
	Top 3	Int. linear	Int. linear	Vero.	Árvores	Int. linear	Int. linear	Int. Stine	Árvores	Int. linear	Int. linear	Int. Stine	Árvores
		0.008	0.009	0.009	0.009	0.008	0.009	0.01	0.011	0.008	0.009	0.01	0.011
1	Top 1	Int. spline	Int. spline	Int. linear	Media	Int. spline	Int. spline	Int. linear	Media	Int. spline	Int. spline	Árvores	Media
		0.008	0.009	0.011	0.012	0.008	0.009	0.011	0.011	0.008	0.009	0.011	0.011
	Top 2	Int. Stine	Int. Stine	Int. Stine	Mediana	Int. Stine	Int. Stine	Árvores	Mediana	Int. Stine	Int. Stine	Int. linear	Mediana
		0.008	0.009	0.011	0.012	0.008	0.009	0.011	0.011	0.008	0.009	0.011	0.011
	Top 3	Int. linear	Int. linear	MM exp.	Árvores	Int. linear	Int. linear	Int. Stine	Árvores	Int. linear	Int. linear	Int. Stine	Árvores
		0.009	0.009	0.011	0.012	0.009	0.009	0.011	0.011	0.009	0.009	0.011	0.011

Tabela 3.5: Tabela apresentando os 10 melhores métodos de preenchimento de valores faltantes testados em termos de erro quadrático médio e erro absoluto percentual médio (linhas) calculado a partir de 1000 replicações para os valores faltantes de amostras de tamanho $n \in \{100, 500, 1000\}$ de séries temporais ARMA$(1,1)$ com $(\phi, \theta) = (0.7, 0.4)$ e $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (colunas)

		$n = 100$				$n = 500$				$n = 1000$			
Métrica	Métodos	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	0.1	0.2	0.5	0.8	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
EQM	Top 1	Int. spline	Int. Stine	Int. linear	Int. linear	Int. spline	Int. Stine	Int. linear	Int. linear	Int. spline	Int. Stine	Int. linear	Int. linear
		0.538	0.672	1.287	3.15	0.53	0.663	1.194	2.58	0.541	0.671	1.189	2.522
EQM	Top 2	Int. Stine	Int. linear	Int. Stine	Vero.	Int. Stine	Int. spline	Int. Stine	Int. Stine	Int. Stine	Int. spline	Int. Stine	Int. Stine
		0.56	0.695	1.299	3.193	0.559	0.681	1.2	2.708	0.57	0.687	1.194	2.647
EQM	Top 3	Int. linear	Int. spline	Vero.	MM linear	Int. linear	Int. linear	Vero.	Vero.	Int. linear	Int. linear	Vero.	MM linear
		0.582	0.697	1.344	3.379	0.581	0.681	1.209	2.725	0.589	0.689	1.198	2.747
EQM	Top 4	Vero.	Vero.	MM exp.	Int. Stine	Vero.	Vero.	MM exp.	MM linear	Vero.	Vero.	MM exp.	Vero.
		0.585	0.7	1.513	3.405	0.581	0.681	1.438	2.795	0.589	0.689	1.43	2.77
EQM	Top 5	MM exp.	MM exp.	MM linear	Media	Árvores	MM exp.	Int. spline	MM exp.	Árvores	MM exp.	Int. spline	MM exp.
		0.974	1.042	1.735	3.558	0.959	1.043	1.568	2.89	0.934	1.05	1.561	2.827
EQM	Top 6	Árvores	MM linear	Int. spline	MM simples	MM exp.	Árvores	MM linear	MM simples	MM exp.	Árvores	MM linear	MM simples
		1.238	1.372	1.748	3.607	0.987	1.132	1.675	3.046	0.99	1.088	1.67	3.006
EQM	Top 7	MM linear	LOCF	MM simples	MM exp.	MM linear	MM linear	MM simples	Árvores	MM linear	MM linear	Árvores	Árvores
		1.324	1.615	2.103	3.613	1.347	1.385	2.053	3.361	1.347	1.39	1.939	3.164
EQM	Top 8	LOCF	NOCB	NOCB	Mediana	NOCB	LOCF	Árvores	Media	LOCF	LOCF	MM simples	Media
		1.39	1.616	2.644	3.683	1.388	1.602	2.153	3.373	1.398	1.606	2.051	3.371
EQM	Top 9	NOCB	Árvores	LOCF	Árvores	LOCF	NOCB	LOCF	Mediana	NOCB	NOCB	LOCF	Mediana
		1.41	1.619	2.676	3.866	1.39	1.612	2.507	3.396	1.406	1.613	2.501	3.382
EQM	Top 10	MM simples	MM simples	Árvores	NOCB	MM simples	MM simples	NOCB	LOCF	MM simples	MM simples	NOCB	NOCB
		1.763	1.8	2.973	4.956	1.796	1.824	2.52	4.353	1.794	1.826	2.524	4.29
EAPM	Top 1	Int. spline	Int. Stine	Int. linear	Int. linear	Int. spline	Int. spline	Int. Stine	Int. linear	Int. spline	Int. spline	Int. Stine	Int. linear
		0.006	0.006	0.009	0.014	0.006	0.006	0.008	0.012	0.006	0.006	0.008	0.012
EAPM	Top 2	Int. Stine	Int. spline	Int. Stine	Vero.	Int. Stine	Int. Stine	Int. linear	Int. Stine	Int. Stine	Int. Stine	Int. linear	Int. Stine
		0.006	0.006	0.009	0.014	0.006	0.006	0.008	0.013	0.006	0.006	0.008	0.012
EAPM	Top 3	Int. linear	Int. linear	Vero.	Int. Stine	Int. linear	Int. linear	Vero.	Vero.	Int. linear	Int. linear	Vero.	MM linear
		0.006	0.006	0.009	0.014	0.006	0.006	0.008	0.013	0.006	0.006	0.008	0.013
EAPM	Top 4	Vero.	Vero.	MM exp.	MM linear	Vero.	Vero.	Int. spline	MM linear	Vero.	Vero.	Int. spline	MM exp.
		0.006	0.007	0.01	0.014	0.006	0.006	0.009	0.013	0.006	0.006	0.009	0.013
EAPM	Top 5	MM exp.	MM exp.	Int. spline	MM exp.	Árvores	MM exp.	MM exp.	MM exp.	Árvores	MM exp.	MM exp.	Vero.
		0.008	0.008	0.01	0.015	0.008	0.008	0.009	0.013	0.008	0.008	0.009	0.013
EAPM	Top 6	Árvores	MM linear	MM linear	MM simples	MM exp.	Árvores	MM linear	MM simples	MM exp.	Árvores	MM linear	MM simples
		0.009	0.009	0.01	0.015	0.008	0.008	0.01	0.014	0.008	0.008	0.01	0.014
EAPM	Top 7	MM linear	NOCB	MM simples	Media	MM linear	MM linear	MM simples	Árvores	MM linear	MM linear	Árvores	Árvores
		0.009	0.01	0.011	0.015	0.009	0.009	0.011	0.015	0.009	0.009	0.011	0.014
EAPM	Top 8	LOCF	LOCF	NOCB	Mediana	NOCB	LOCF	Árvores	Media	LOCF	LOCF	MM simples	Media
		0.009	0.01	0.012	0.015	0.009	0.01	0.011	0.015	0.009	0.01	0.011	0.015
EAPM	Top 9	NOCB	Árvores	LOCF	Árvores	LOCF	NOCB	LOCF	Mediana	NOCB	NOCB	LOCF	Mediana
		0.009	0.01	0.013	0.016	0.009	0.01	0.012	0.015	0.009	0.01	0.012	0.015
EAPM	Top 10	MM simples	MM simples	Árvores	NOCB	MM simples	MM simples	NOCB	LOCF	MM simples	MM simples	NOCB	NOCB
		0.011	0.011	0.014	0.017	0.011	0.011	0.012	0.016	0.011	0.011	0.012	0.016

4 Conclusão

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 de árvores de decisão. Com base nesse estudo foi possível analisar o desempenho deste método em diferentes cenários e verificar se eram compatíveis com o esperado, utilizando-se do referencial teórico de séries temporais, valores faltantes e árvores de decisão. Também foi desenvolvida uma ferramenta amigável ao usuário para reconstrução de séries temporais com o método proposto utilizando do pacote `Shiny` do `R`.

As árvores obtiveram bons resultados em contextos em que a média e mediana foram bons estimadores para os valores faltantes. Isso era esperado, visto que as previsões das árvores de decisão (para variáveis numéricas), nada mais são do que a média condicionada às observações de uma determinada folha, e em séries temporais estacionárias, isto deveria ser próximo da média global.

A qualidade de reconstrução das séries temporais utilizando árvores de decisão é afetada pela quantidade de covariáveis utilizadas na modelagem, melhorando conforme são adicionadas informações, principalmente se são utilizadas no modelos tanto variáveis anteriores quanto posteriores às observações. O método proposto neste trabalho é promissor, principalmente pois os resultados dos preenchimentos parecem melhorar conforme aumenta o tamanho da amostra.

É necessário ressaltar que as árvores de decisão são um método não paramétrico e seria interessante estudar o que aconteceria em modelos que não assumem linearidade dos dados. Portanto, os próximos objetivos desta pesquisa são testar a metodologia criada em séries temporais de modelos não lineares e séries temporais reais, explorar porquê a média não foi uma boa forma de prever os valores faltantes no modelo ARMA(1, 1) e melhorar continuamente a ferramenta criada com o `Shiny`

5 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 5.1 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 3.3). 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).

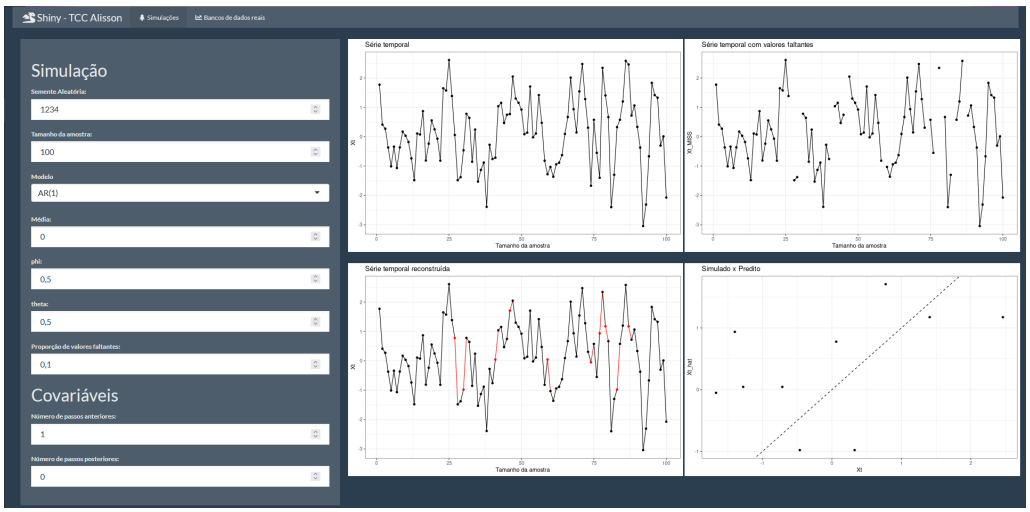

Figura 5.1: Aba 1 do aplicativo Shiny, com opções para simular séries temporais AR(1), MA(1) e ARMA(1, 1) e reconstruí-las utilizando o método proposto.

Na Figura 5.2 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.

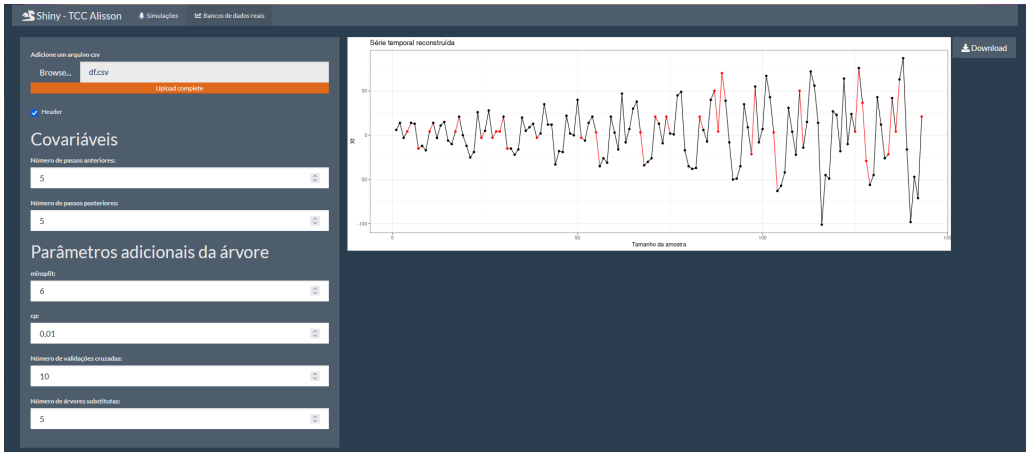

Figura 5.2: Aba 2 do aplicativo Shiny, em que é possível fazer o *upload* de uma série temporal real que pode ser reconstruída pelo usuário utilizando o método proposto neste trabalho e posteriormente baixada.

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/`.

Referências Bibliográficas

Batista, G. e 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., e Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Brockwell, P. J. e 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., e Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.7.1.

Dempster, A. P., Laird, N. M., e 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., e Steier, P. (2001). The filling of gaps in geophysical time series by artificial neural networks. *Radiocarbon*, 43(2A):365–371.

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

James, G., Witten, D., Hastie, T., e 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., 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., Kenward, M., Tsiatis, A., e Verbeke, G. (2020). *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis Group.

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

Moritz, S. e 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.

Peng, L. e Lei, L. (2021). A review of missing data treatment methods.

Prass, T. S. e 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., e 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.

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

Therneau, T. e 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., e W, Kibunja, W. (2013). Imputation of incomplete non-stationary seasonal time series data. *Mathematical Theory and Modeling*, 3:142–154.