Algoritmos

MDPI

Artigo

Seleção Variável em Previsão de Séries Tempoesas Usando Florestas Aleatórias

Cristo Tyralis * **e Geórgia Papacharalampous**

Departamento de Recursos Hídricos e Engenharia Ambiental, Escola de Engenharia Civil, Universidade Técnica Nacional de Atenas, Iroon Polytechniou 5, 157 80 Zografou, Grécia; papacharalampous.georgia@gmail.com ou gpapacharalampous@itia.n tua.gr

***** Correspondência: montchrister@gmail.com ou hristos@itia.ntua.gr; Tel.: +30-21-0514-8526

Recebido: 29 de julho de 2017; Aceito: 1 de outubro de 2017; Publicado: 4 de outubro de 2017

Resumo: A previsão da série temporal usando algoritmos de aprendizado de máquina ganhou popularidade recentemente. Random forest é um algoritmo de aprendizado de máquina implementado na previsão de séries temporências; no entanto, a maioria de suas propriedades de previsão permaneceram inexploradas. Aqui, nós nos aproximamos de avaliar o desempenho de florestas aleatórias em uma etapa de previsão usando dois grandes conjuntos de dados de séries de curto prazo com o objetivo de sugerir um conjunto ideal de variáveis preditoras. Além disso, comparamos seu desempenho com métodos de benchmarking. O conjunto de dados first é composto por 16.000 séries temporais simuladas de uma variedade de modelos ARFIMA (Autoregressive Fractionally Integrated Moving Average). O segundo conjunto de dados consiste em 135 séries médias anuais de tempo de temperatura. O maior desempenho preditivo de RF é observado quando se utiliza um baixo número de variáveis preditoras recentes. Esse resultado pode ser útil em aplicações futuras relevantes, com a perspectiva de alcançar maior precisão preditiva.

Palavras-chave: ARFIMA; ARMA; aprendizado de máquina; um passo à frente da previsão; florestas aleatórias; previsão de séries tempor previstas; seleção variável

1. Introdução

1.1. Previsão da Série Temporal e Florestas Aleatórias

O uso de algoritmos de aprendizagem de máquina na modelagem preditiva (ver foare a definição [1]) e no caso específico de previsão de séries temporizadas aumentou consideravelmente recentemente [2]. Aplicações relevantes podem ser encontradas nos artigos de revisão [3-6].

A previsão da série temporal pode ser classificada em duas categorias de acordo com o horizonte de previsão, ou seja, previsão de uma etapa e várias etapas [2,7]. A previsão de uma etapa (examinada aqui) é útil em aplicações únicas, por exemplo, na avaliação de métodos de previsão utilizando simulações [8], previsão de engenharia [9,10], previsão ambiental [11,12]e previsão financeira [13-17]. Além disso, as estratégias de previsão recursiva e de várias etapas do DiRec dependem da previsão de um passo. Em particular, as previsões de um passo à frente são usadas na estratégia recursiva, enquanto em cada etapa o valor previsto da etapa anterior é usado como entrada adicional, sem alterar o modelo de previsão. A estratégia DiRec combina a estratégia recursiva, mas o modelo de previsão muda a cada etapa [7].

A floresta aleatória (RF) é um algoritmo de locação de máquinas introduzido em [18], que pode ser usado para classificação e regressão. É popular porque pode ser aplicado a uma ampla gama de problemas de previsão, tem alguns parâmetros para sintonizar, é simples de usar, foi aplicado com sucesso a muitos problemas práticos e pode lidar com pequenos tamanhos de amostra, espaços de características de alta dimensão e estruturas de dados complexas [19,20]. Uma revisão e uma simples apresentação do algoritmo RF podem ser encontradas em [20-22]. A regressão usando RF pode ser implementada para fins de previsão de séries temporâneos. Aplicações representativas podem ser encontradas em muitos campos científicos,

incluindo as de engenharia [23,24], ciências ambientais e geofísicas [25-27], estudos financeiros [28,29], e medicina [30], com

Algoritmos **2017**, *10*, 114; doi:10.3390/a10040114www.mdpi.com/journal/algoritmos

desempenho variado. Além disso, pequenos conjuntos de dados são usados nesses aplicativos; portanto, os resultados não podem ser generalizados.

O desempenho do algoritmo RF depende do ajuste de seus parâmetros e da seleção variável (também conhecida como selection de recursos). Procedimentos para estimar o conjunto ideal de parâmetros são propostos em diversos estudos. Esses procedimentos consideram o desempenho do algoritmo dependendo das variáveis selecionadas [22,31], o número de árvores [32,33], o número de possíveis direções para divisão em cada nó de cada árvore [20,34] (p. 199), e o número de exemplos em cada célula, abaixo do qual a célula não é dividida [20,35]]. Há também alguns estudos que examinam as propriedades de RF em um contexto teórico [19]. O desempenho dos algoritmos de aprendizagem de máquina é geralmente avaliado por meio de estudos de caso único. Às vezes, grandes conjuntos de dados compostos por estudos do mundo real são usados para benchmarking e comparação de vários algoritmos [7,36-40]. Além disso, [41] propõe o uso de simulações como uma alternativa considerável. As respectivas aplicações podem ser encontradas em [42,43].

1.2. Um quadro para avaliar o desempenho de florestas aleatórias na previsão da série temporal

Em problemas habituais de regressão, é dada uma amostra composta por observações da variável dependente e das respectivas observações das variáveis preditoras. O modelo de regressão é treinado usando esta amostra. A previsão é então realizada quando novas observações das variáveis preditoras são obtidas. Se decidirmos usar menos variáveis preditores do que o incluído na amostra, então o tamanho do conjunto de treinamento não muda. Além disso, é importante que a inclusão de variáveis preditoras sem importância não impacte seriamente o desempenho preditivo da RF, como indicado em [34] (p. 489). Por outro lado, o uso de RF para previsão de séries temporizou não é idêntico ao simples caso de regressão. Neste caso, o papel das variáveis preditoras é tomado por valores anteriores da série temporal (variáveis defasadas). Portanto, aumentar o number das variáveis preditoras, ou seja, as variáveis defasadas selecionadas, inevitavelmente resulta na redução da duração do conjunto de treinamento. O uso de menos variáveis preditoras, em vez disso, pode reduzir as informações obtidas pelo conhecimento disponível do temporâneo. A computação analítica do desempenho não é atingível, devido à complexidade do algoritmo RF [20]. No entanto, uma resposta empírica poderia ser dada no contexto de um grande experimento. Além disso, neste experimento de simulação o rf could ser comparado aos métodos de benchmarking, cujo desempenho é conhecido a priori, teoricamente. Aqui realizamos um grande experimento de simulação, no qual usamos 16.000 séries temporais simuladas de uma variedade de modelos ARFIMA (Autoregressive Fractionly Integrated Moving Average) (44,45). Estudos semelhantes utilizando simulações da família de modelos ARFIMA podem ser encontrados em [46,47], embora em menor grau e implementando diferentes métodos. Os estudos de simulação são efetivamente complementados quando integrados a estudos de caso; consequentemente, também aplicamos os métodos a um conjunto de dados de 135 temperaturas instrumentais médias anuais do conjunto de dados em [48]. O comprimento de cada série temporal simulada e observada é de 101.

Comparamos o desempenho de cinco algoritmos na previsão do 101º valor de cada série de tempo, quando encaixados aos seus primeiros 100 valores. Os cinco algoritmos implementados no estudo são dois métodos ingênuos, um modelo ARFIMA, o método e o algoritmo RF. Os métodos ingênuos geralmente funcionam bem em séries temporentas previstas [37] e os processos ARFIMA são métodos tradicionais que são frequentemente usados para a previsão de séries temporáticas [45]. O método theta foi introduzido recentemente e também é um dos métodos de previsão mais bem sucedidos [49]. Utilizamos vários versions do algoritmo RF em relação ao número de variáveis preditoras, enquanto a otimização do conjunto de parâmetros é realizada utilizando métodos propostos em [50,51]. As métricas utilizadas para a comparação são o erro absoluto médio (MAE), o erro médio esquartejado (MSE), o erro percentual absoluto médio (MAPE) e o coeficiente de regressão linear. A análise realizada aqui diz respeito a séries temporâneas simuladas a partir de processos estocásticos estacionários, enquanto modelos estacionários também são usados para modelar a série de tempo de temperatura.

1.3. Objetivo do Estudo

O objetivo principal do nosso estudo é investigar como o desempenho da RF está relacionado à seleção variável em uma etapa de previsão de séries de curto prazo. O quadro proposto na Seção 1.2 pode ajudar a fornecer uma resposta empâmica para o problema da seleção variável. Em conclusão, mostra-se que um baixo número de variáveis defasadas recentes tem melhor desempenho, destacando a importância do comprimento do conjunto de treinamento. O algoritmo RF é provado ser competente no tempo de previsão de series, embora não tenha exibido o melhor desempenho na maioria dos casos examinados. Isso não implica que o RF seja de pouco valor prático. Em vez disso, como foi mostrado em [12], não há um método de previsão universalmente melhor, therefore, em aplicações práticas devemos usar múltiplos métodos para obter uma boa previsão. Nesses casos, enfatizamos que devem ser utilizadas metodologias adequadas para a otimização do desempenho de cada método. Nossa contribuição diz respeito à optimização do desempenho de previsão da RF.

2. Métodos e Dados

Na Seção 2, apresentamos os métodos e dados utilizados no presente estudo. Em particular, apresentamos as definições dos modelos utilizados para simulação e teste, bem como os métodos de forsting. Além disso, apresentamos os dados de temperatura, também utilizados para o teste dos métodos. Os códigos e dados para reproduzir o estudo estão disponíveis como material suplementar (ver apêndice A). Os cálculos foram realizados utilizando-se a linguagem de programação R [52].

2.1. Métodos

2.1.1. Definição de Modelos ARMA e ARFIMA

Aqui, fornecemos as definições dos processos estocásticos autoregressivos da média móvel (ARMA) e ARFIMA, enquanto o leitor também é referido [45] (pp. 6-65, 489-494) para informações further. Uma série temporal em tempo discreto é definida como uma sequência de observações *x1*, *x2*, . . , de um certo fenômeno, enquanto o tempo t é declarado como um subscrito para cada valor $_{xt}$. Uma série temporal pode ser modelada por um processo estocástico. Esta última é uma sequência de variáveis aleatórias $_{x1}$, $\underline{x2}$, ... As variáveis aleatórias são sublinhadas de acordo com a notação usada em [53].

Consideremos um processo estocástico estacionário de variáveis aleatórias normalmente distrificadas. A média μ do processo estocástico é definida por:

$$\mu := \mathrm{E}[\underline{xt}], \tag{1}$$

A função de desvio padrão σ do processo estocástico é definida por:

$$\sigma == (\mathrm{Var}[_{xt}])^{0,5} \tag{2}$$

A função de covariância entre $\underline{xt}$ e $\underline{xt}_{+k}$, γ_k do processo estocástico é definida por:

$$\gamma_k := \mathrm{E}[(_{xt} - \mu)(_{xt+k} - \mu)] \tag{3}$$

A função de correlação entre $\underline{xt}$ e $\underline{xt}_{+k}$, ρk do processo estocástico é definida por:

$$_{rk} := \gamma k/\sigma 2 \tag{4}$$

Um processo estocástico estacionário normal $\{\underline{at}\}$ é chamado de processo de ruído branco, se for uma sequência de variáveis aleatórias não corrigidas. Consideremos daqui em diante que o ruído branco é uma variável com média zero, a menos que mencionada o contrário, e o desvio padrão σ_a.

Definimos o processo estocástico $\{yt\}$ por:

$$_{yt} := \underline{xt} - \mu \tag{5}$$

Consideremos o operador B, que é definido por:

$$\underline{Bjxt} = \underline{xt}-\text{j} \tag{6}$$

Em seguida, o operador $\varphi_p(B)$ é definido por:

$$\varphi_p(B) =(1 -\varphi_{1B} - \ldots\ldots -\varphi\, pBp) \tag{7}$$

O processo estocástico $\{_{xt}\}$ é um modelo AR(p) autoregressive, se:

$$\varphi_p(B)yt = \underline{at}, \tag{8}$$

que podem ser escritos na seguinte forma:

$$yt = \varphi\, 1yt-1 + \ldots + \varphi\, pyt-p + \underline{at} \tag{9}$$

Consideremos também o operador $_{\varphi q}(B)$, que é definido por:

$$_{\varphi q}(B) := (1 + \phi_{1B} + \ldots \tag{10}$$

O processo estocástico $\{\underline{xt}\}$ é um modelo MA(q) médio móvel, se:

$$yt = \varphi q(B)\, \underline{at}, \tag{11}$$

que podem ser escritos na seguinte forma:

$$yt = \underline{at} + \phi 1at-1 + \ldots + \varphi qat-q \tag{12}$$

O processo estocástico $\{\underline{xt}\}$ é um modelo arma(p, q) de média móvel autoregressive, se

$$\varphi_p(B)yt = \varphi q(B)\underline{at}, \tag{13}$$

que podem ser escritos na seguinte forma:

$$yt = \varphi\, 1yt-1 + \ldots + \varphi\, pyt-p + \underline{at} + \underline{\varphi 1at}-1 + \ldots + \underline{\varphi qat}-q \tag{14}$$

Vamos ∈ (−0,5, 0,5). O processo estocástico $\{\underline{xt}\}$ é um ARFIMA (p, d, q), se

$$\varphi_p(B)(1 - B)\underline{dxt} = _{\varphi q}(B)\underline{at} \tag{15}$$

2.1.2. Simulação de modelos ARMA e ARFIMA

Simulamos os processos ARMA e ARFIMA usando o arima.sim incorporado na função R [52] e o algoritmo fracdiff.sim do fracdiff do pacote R fracdiff [54], respectivamente.

2.1.3. Previsão utilizando modelos ARMA e ARFIMA

Aqui apresentamos como podemos usar os modelos ARMA e ARFIMA em uma previsão de séries de tempo um passo à frente. As aplicações específicas dos métodos de previsão são apresentadas na Seção 2.1.8. Consideramos uma série temporal de n observações. Consideremos também um modelo instalado nas observações n e posteriormente utilizado para prever xn_{+1}. Que $_{xn}$ e ψ representem a última observação e a previsão de xn_{+1}, respectivamente. Os métodos que utilizam os modelos ARMA e ARFIMA podem ser usados como benchmarks nos experimentos de simulação. No FACT, espera-se que esses métodos sejam melhores que os demais, quando aplicados à série temporal sintética, uma vez que estes últimos são simulados utilizando modelos ARMA ou ARFIMA (ver Seção 2.1.8). Examinamos dois casos.

No primeiro caso assumimos que as séries de tempo são modeladas por um ARMA(p, q), em que p e q são predefinidos, enquanto os parâmetros devem ser estimados. Encaixamos os modelos ARMA aos dados usando a arima incorporada na função R. Durante o processo de implementação, os valores de p, q são definidos iguais aos valores respeitosos utilizados no experimento de simulação correspondente. A função

aplica o método de máxima probabilidade para estimar os valores dos parâmetros AR e MA φ_1, ..., φ_p, $\varphi 1$, ..., φq dos modelos. Utilizamos o modelo ARMA instalado no modo de previsão, implementando a previsão construída em função R [52].

No segundo caso, assumimos que a série temporal pode ser modelada por um modelo ARFIMA(p, d, q) com parâmetros desconhecidos. Encaixamos um modelo ARFIMA aos dados usando a função arfima da previsão do pacote R [55,56]. Os valores de p, d, q são estimados utilizando-se o Critério de Informação Akaike com uma correção para tamanhos de amostra finito (AICc), com d variando entre −0,5 e 0,5. A função aplica o método de máxima probabilidade para estimar os valores dos parâmetros AR e MA φ_1, ..., φ_p, $\varphi 1$, ..., φq dos modelos. Os procedimentos de seleção do pedido e os procedimentos de estimativa dos parâmetros são explicados, por exemplo, em [57] (Capítulo 8.6). Utilizamos o modelo ARFIMA instalado no modo de previsão, implementando a função de previsão do pacote R de previsão.

2.1.4. Previsão usando métodos ingênuos

Utilizamos dois métodos de previsão ingênuos, a última ingênua observação e a média ingênua, que estão entre os benchmarks mais utilizados [57] (Capítulo 2.3). As previsões produzidas são dadas por (16) e (17) respectivamente:

$$\psi = xn \tag{16}$$

$$\psi = (x_1 + . + x_n)/n \tag{17}$$

Os métodos ingênuos aqui apresentados são benchmarks, pois são os métodos de previsão mais simples em termos de complexidade teórica e custo computacional.

2.1.5. Previsão Utilizando o Método Theta

Implementamos uma versão simples do method de previsão de teta, que foi introduzido em [49], através da função tetaf da previsão do pacote R. Theta usa uma determinada série temporal para criar duas ou mais séries temporândias auxiliares com diferentes curvaturas locais (modificadas) em relação ao original, ou seja, as "Linhas Theta". Estes últimos são extrapolados separadamente e posteriormente combinados com pesos iguais para produzir a previsão. A modificação das curvaturas locais é implementada utilizando-se o "Theta-coeficiente" φ, que é aplicado às segundas diferenças da série de tempo.

A versão aqui adotada usa duas linhas Theta, para $\varphi = 0$ e $\varphi = 2$. Esta versão específica teve um bom desempenho no M3-Competition [37], enquanto mais tarde [58] provou sua equivalência à simples suavização exponencial (SES) com deriva. Para os métodos de previsão de suavização exponential, e particularmente para sua implementação em uma estrutura espacial estatal, o leitor é referido [57] (Capítulo 7) e [59], respectivamente.

2.1.6. Florestas Aleatórias

RF é um termo vago [20]. Aqui usamos o algoritmo RF original introduzido em [18]. Em particular, apresentamos a versão do algoritmo usado para regressão. A versão específica do algoritmo foi implementada no pacote randomForest R [60]. O restante da seção é amplamente reproduzido a partir de [20] adaptações with, enquanto notação idêntica é usada em [20,60]. A presente Seção não se destina a fornecer uma descrição detalhada da RF. O leitor interessado pode encontrar mais detalhes em [20], enquanto os cálculos são reprodutíveis (Apêndice A).

Assumimos que o $\underline{\boldsymbol{u}}$ é um vetor aleatório com elementos k. O objetivo é prever v estimando a função de regressão:

$$m(\boldsymbol{u}) = E[\underline{v} | \ \underline{\boldsymbol{u}} \ = \boldsymbol{u}] \tag{18}$$

dada a amostra de montagem:

$$_{Ss} = ((_{u1}, v1), . .$$

que são realizações independentes da variável aleatória ($\underline{\boldsymbol{u}}$, $\underline{v}$). Portanto, o objetivo é construir uma estimativa *de ms* da função m.

Uma floresta aleatória é um preditor construído pelo cultivo *de* árvores de regressão aleatórias M. Para a árvore j-th da família, o valor previsto em $\boldsymbol{u}$ é denotado por $ms(\boldsymbol{u}; \ \underline{\varphi j}, _{Ss})$, onde $_{\phi 1}$, ..., $\underline{\varphi M}$ são variáveis aleatórias independentes, distribuídas como $\underline{\varphi}$ e independentes de Ss. A variável aleatória $\underline{\varphi}$ é usada para

resamar o conjunto de montagem antes do cultivo de árvores individuais e para selecionar as direções sucessivas para a divisão.

A previsão é então dada pela média dos valores previstos de todas as árvores.

Antes de construir cada árvore, *bs* observations são escolhidos aleatoriamente a partir dos elementos de ***u***. Essas observações são usadas para cultivar a árvore. Em cada célula da árvore, uma divisão é realizada pela maximização do critério CART (definido em [20]) selecionando variáveis mtry aleatoriamente entre as *k* original, escolhendo o melhor ponto variável/dividido entre o mtry e dividindo o nó em dois nódulos filhas. O crescimento da árvore é interrompido quando cada célula contém menos de pontos de nodesize.

Os *parâmetros utilizados* em RF são *bs* ∈ {1, . . , *s*}, mtry ∈ {1, ... ∈ ∈ . . . Na maioria dos estudos, concorda-se que o aumento do número de árvores não diminui o desempenho preditivo; no entanto, resulta em um aumento do custo computacional. Oshiro et al. [32] sugere um intervalo entre 64 e 128 árvores em uma floresta baseada em experimentos. Kuhn e Johnson [34] (p. 200) sugerem o uso de pelo menos 1000 árvores. Probst e Boulesteix [33] sugerem que o maior ganho de desempenho é obtido ao treinar 100 árvores. No presente estudo, utilizamos *M* = 500 árvores.

O número de pontos de dados amostrados em cada árvore e o número de exemplos em cada célula abaixo do qual a célula não é dividida são definidos iguais aos seus valores padrão no pacote randomForest R, ou seja, igual ao tamanho da amostra e nodesize = 5, respectivamente. Estes são relatados como bons valores [20,35].

A mtry do parâmetro é estimada durante a fase de validação através da implementação da função trainControl do caret do pacote R [50,51]. A mtry ideal é encontrada usando resampato de bootstrap, ou seja, amostrando aleatoriamente o teste de validação com reseleção (o tamanho da amostra bootstrap deve ser igual ao tamanho do teste de validação), encaixando o modelo na amostra bootstrap e medindo seu desempenho no set restante, ou seja, os dados não selecionados pela amostragem de botas. Vinte e cinco iterações são usadas no procedimento de ressupimento. A mtry ideal é encontrada calculando o desempenho médio dos modelos montados, enquanto a busca pelo valor ideal do parâmetro é realizada em uma grade. O erro quadrado médio raiz (RMSE) é usado para medir o desempenho. O mtry ideal minimiza o RMSE. Detalhes sobre o histórico teórico do método podem ser encontrados em [34] (pp. 72-73). A função de regressão de Equation (18) é definida pelo modelo, que é montado no teste de validação usando o valor ideal de mtry.

2.1.7. Previsão da Série Temporal usando florestas aleatórias

O uso de RF para previsão de séries temporáticas de uma etapa é simples e semelhante à maneira como a RF pode be usada para regressão. Que *g* seja a função obtida a partir do treinamento de RF, que será utilizado para previsão xn_{+1}, dado *x1*, . . Se usarmos variáveis *k* defasadas , então o xn_{+1} previsto é dado pela seguinte equação para $t = n + 1$:

$$_{xt} = g(xt_{-1}, \ldots, xt_{-k}),\ t = k + 1, \ldots$$

A função *g* não está em forma fechada, mas pode ser obtida treinando o algoritmo RF usando um conjunto de treinamento de tamanho $n - k$. Em cada amostra do conjunto de fitação a variável dependente é *xt*, para $t =$ k + 1, . . Quando o número de variáveis preditoras *k* aumenta, o tamanho do conjunto de treinamento $n - k$ diminui (um exemplo é apresentado na Figura 1). O conjunto de treinamento, que inclui $n - k$ amostras, é criado utilizando a função CasesSeries do pacote Rminer R [61,62]. Por fim, a montagem é realizada utilizando-se a função de trem do pacote Caret R e o valor previsto de xn_{+1} é obtido utilizando a função predito do pacote Caret R.

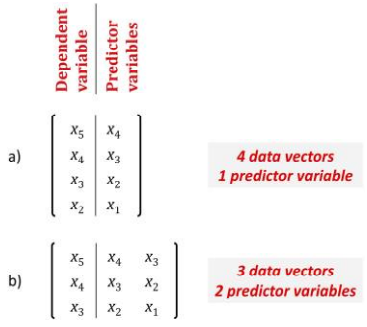

Figura 1. Esboço explicando como a amostra de treinamento muda com o número de variáveis preditoras para séries temporais com $n = 5$ e (**a**) $k = 1$; b K = 2.

Kuhn e Johnson [34] (p. 463) definem a importância variável como a quantificação geral da relação entre as variáveis preditoras e o valor previsto. Aqui usamos a primeira medida de importância variável em RF conforme definido no pacote RandomForest R [60]. A partir do procedimento de treinamento utilizando variáveis *k* preditor, mantemos as variáveis com importância variável positiva e definimos um novo conjunto de variáveis preditoras, o que inclui essas variáveis. Então repetimos o procedimento de previsão. Neste caso, o tamanho da amostra depende do índice mínimo do subconjunto das variáveis preditoras quando previsto *em* xn_{+1}, por exemplo, se o índice mínimo for $n + 1 - k$, então o conjunto de treinamento é de tamanho $n - k$. No entanto, o uso de menos variáveis preditoras diminui o custo computacional.

2.1.8. Resumo dos Métodos

Na Tabela 1, resumimos os métodos apresentados nas Seções 2.1.3-2.1.7. Na Tabela 2, apresentamos a categoria de dados aos quais aplicamos cada caso específico dos métodos baseados no modelo ARMA e ARFIMA na Seção 3. Na Tabela 3, apresentamos os 12 diferentes procedimentos de seleção variável utilizados pelo algoritmo RF.

Mesa 1. Resumo dos métodos apresentados nas Seções 2.1.3-2.1.7 e sua abreviação como usado na Seção 3.

Método	Secção	Breve Explicação
arfima	2.1.3	Usa modelos ARMA ou ARFIMA equipados
ingênuo1	2.1.4	Previsão igual ao último valor observado, Equação (16)
ingênuo2	2.1.4	Previsão igual à média do conjunto montado, Equação (17)
theta	2.1.5	Usa o método theta
Rf	2.1.7	Usa florestas aleatórias, Equação (20)

Mesa 2. Métodos apresentados na Seção 2.1.3 para previsão utilizando modelos ARMA e ARFIMA e suas aplicações específicas na Seção 3.

Método	Aplicação na Seção 3
1º caso na Seção 2.1.3	Simulações da família de modelos ARMA
2º caso na Seção 2.1.3	Simulações da família dos modelos ARFIMA, com $d6 = 0$
2º caso na Seção 2.1.3	Dados de temperatura

Mesa 3. Procedimentos de seleção variável utilizados pelo algoritmo RF.

Método	Variáveis Explicativas
rf05, rf10, rf15, rf20, rf25, rf30, rf35, rf40, rf45, rf50	Usa as últimas 5, ...
rf20imp, rf50imp	Utiliza as variáveis mais importantes das últimas 20 e 50 variáveis, respectivamente

2.1.9. Métricas

Aqui definimos as métricas utilizadas nas comparações. O erro (E) é definido por:

$$E := \psi - xn_{+1} \tag{21}$$

O erro absoluto (AE) é definido por:

$$AE := |\psi\, xn_{+1}| \tag{22}$$

O erro quadrado (SE) é definido por:

$$SE := (\psi - xn_{+1})^2 \tag{23}$$

O erro percentual (PE) é definido por:

$$PE := (\psi - xn_{+1})/xn_{+1} \tag{24}$$

O erro percentual absoluto (APE) é definido por:

$$APE := |\psi\, xn_{+1}/xn_{+1}| \tag{25}$$

As métricas definidas em Equações (21)–(25) podem ser usadas para a avaliação do desempenho em experimentos únicos. Para avaliar o desempenho de previsão usando um conjunto de simulações múltiplas, as métricas devem ser resumidas. Que *N* seja o número dos experimentos de previsão conducionados. Que o número de série de cada experimento *seja* declarado como um subscrito a cada valor métrico $_{Ei}$, AEi, SEI, PEi e APEi. Em seguida, a média dos erros (MoE) é definida por:

$$\text{MoE} := (E_1 + \ldots + E_N)/N \quad (26)$$

A mediana dos erros (MdoE) é definida por:

$$\text{MdoE:= mediana (E1, . . } \quad (27)$$

A média dos erros absolutos (MoAE) é definida por:

$$\text{Moae := (AE1 + . . } \quad (28)$$

A mediana dos erros absolutos (MdoAE) é definida por:

$$\text{MdoAE := mediana (AE1, . . } \quad (29)$$

A média dos erros quadrados (MoSE) é definida por:

$$\text{MoSE := (SE1 + . . } \quad (30)$$

A mediana dos erros quadrados (MdoSE) é definida por:

$$\text{MdoSE := mídia(SE1, . . } \quad (31)$$

A média dos erros percentuais (MoPE) é definida por:

$$\text{MoPE := (PE1 + . . PEN)/}N \quad (32)$$

A mediana dos erros percentuais (MdoPE) é definida por:

$$\text{MdoPE := mediana(PE1, ... , PEN) A} \quad (33)$$

média dos erros percentuais absolutos (MoAPE) é definida por:

$$\text{MoAPE := (APE1 + . . + APEN)/}N \quad (34)$$

A mediana dos erros percentuais absolutos (MdoAPE) é definida por:

$$\text{MdoAPE := mediana (APE1, . . } \quad (35)$$

Os meios e as medianas podem diferir substancialmente quando os outliers são observados; portanto, ambos são úteis na avaliação do desempenho de previsão. Outra métrica útil do desempenho de previsão, ao avaliar múltiplos experimentos de forma simultaneousamente, é a inclinação da regressão. Que *eu* também seja declarado como um subscrito a cada previsão ψ_i e seu valor real correspondente $xn_{+1,i}$. Estima-se que o coeficiente de regressão a (ou inclinação da regressão) meça a dependência de $\psi_1, \ldots, \psi_N$ em $xn_{+1,1}, \ldots, xn_{+1,N}$, quando esta dependência é expressa pelo seguinte modelo de regressão linear:

$$\psi_i = a\, xn_{+1,i} + b \quad (36)$$

Na Tabela 4, apresentamos a faixa dos valores das métricas e seus valores quando a previsão é perfeita.

Mesa 4. Métricas of previsão de desempenho, seu intervalo e respectivos valores quando a previsão é perfeita.

Métrica	Equação	Gama	Valores métricos para previsão perfeita
erro	(21)	[-∞, ∞]	0
erro absoluto	(22)	[0, ∞]	0
erro quadrado	(23)	[0, ∞]	0
erro percentual	(24)	[-∞, ∞]	0

erro percentual absoluto	(25)	[0, ∞]	0
coeficiente de regressão linear	(36)	[-∞, ∞]	1

Uma discussão sobre a adequação das métricas para medir o desempenho dos métodos de previsão pode ser encontrada em [57] (Capítulo 2.5) e [63], enquanto [64,65] discutem extensivamente métricas ligadas ao coeficiente de regressão. Aqui preferimos use as métricas da Tabela 4, devido à sua simplicidade.

2.2. Dados

Aqui apresentamos os dados utilizados no presente estudo. Usamos dois grandes conjuntos de dados. A primeira foi obtida a partir de simulações dos modelos apresentados na Seção 2.1.2. O segundo conjunto de dados é um conjunto de temperaturas anuais observadas.

2.2.1. Série tempo simulada

Foram utilizados 16 modelos da família de modelos ARFIMA apresentados na Seção 2.1.1. Cada modelo corresponde a um experimento de simulação apresentado na Tabela 5. Cada experimento de simulação inclui 1000 séries temporânuladas de tamanho 101 dos modelos da Tabela 5.

Mesa 5. Experimentos de simulação, seus respectivos modelos e seus parâmetros definidos. Consulte a Seção 2.1.1 para obter as definições dos parâmetros.

Experimentar	Modelo	Parâmetros
1	ARMA(1, 0)	$f_1 = 0,6$
2	ARMA(1, 0)	$f_1 = -0,6$
3	ARMA(2, 0)	$f_1 = 0,6$, $f_2 = 0,2$
4	ARMA(2, 0)	$f_1 = -0,6$, $\phi_2 = 0,2$
5	ARMA(0, 1)	$i_1 = 0,6$
6	ARMA(0, 1)	$i_1 = -0,6$
7	ARMA(0, 2)	$i_1 = 0,6$, $i_2 = 0,2$
8	ARMA(0, 2)	$i_1 = -0,6$, $\varphi 2 = -0,2$
9	ARMAS(1, 1)	$f_1 = 0,6$, $i_1 = 0,6$
10	ARMAS(1, 1)	$f_1 = -0,6$, $\varphi 1 = -0,6$
11	ARMAS(2, 2)	$f_1 = 0,6$, $f2 = 0,2$, $i1 = 0,6$, $i2 = 0,2$
12	ARFIMA(0, 0,40, 0)	
13	ARFIMA(1, 0,40, 0)	$f_1 = 0,6$
14	ARFIMA(0, 0,40, 1)	$i_1 = 0,6$
15	ARFIMA(1, 0,40, 1)	$f_1 = 0,6$, $i_1 = 0,6$
16	ARFIMA(2, 0,40, 2)	$f_1 = 0,6$, $f2 = 0,2$, $i1 = 0,6$, $i2 = 0,2$

2.2.2. Conjunto de dados de temperatura

O conjunto de dados do mundo real inclui 135 séries médias anuais de tempo de temperatura instrumental extraídas do banco de dados de [48]. O banco de dados inclui temperaturas médias mensais observadas. Extraímos estações, que incluíram temperaturas mensais médias para o period 1916-2016, ou seja, 101 anos de observações. Descartamos estações com dados faltantes. Nós retratamos as 135 estações restantes na Figura 2. As estações cobrem uma parte considerável da superfície da Terra, portanto, podem representar comportamentos diversos observados em séries temporânicas geofísicas. As temperaturas médias anuais foram obtidas pela média por ano dos valores mensais.

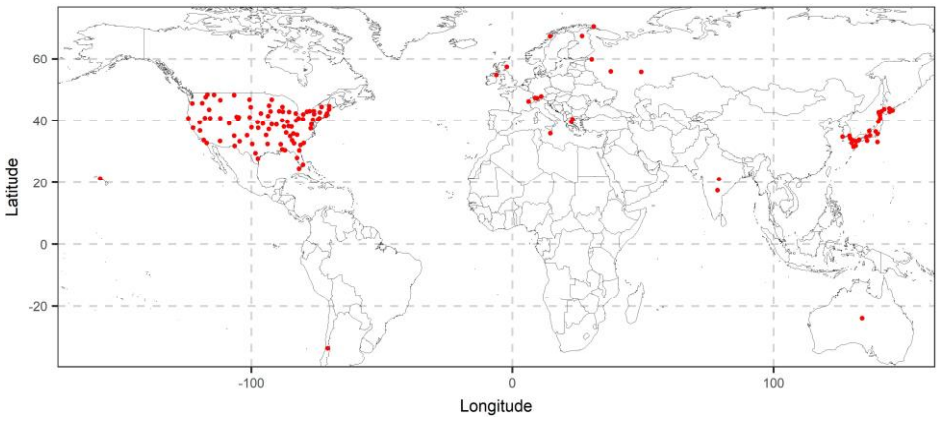

Figura 2. Mapa de locais para as 135 estações utilizadas na análise.

3. Resultados

Aqui apresentamos os resultados da aplicação dos métodos aos dados apresentados na Seção 2. Usamos dois conjuntos de dados, incluindo séries temporidas de tamanho 101. Os dois conjuntos de dados incluem séries temporâneas simuladas e temperaturas observadas. A montagem é realizada nos primeiros 100 valores da série temporal, enquanto os methods são comparados na previsão do 101º valor. A aplicação dos métodos a conjuntos de dados específicos foi apresentada na Seção 2.1.8.

3.1. Simulações

Apresentamos os resultados desde a aplicação dos métodos até a série temporal simulada. Cada experimento de simulação dura 30h, enquanto o tempo de computação para todos os métodos que excluem o RF é de alguns minutos. Em particular, apresentamos as estratégias de caixa dos erros absolutes e dos erros. Além disso, apresentamos as medianas dos erros absolutos, medianas dos erros quadrados e os coeficientes de regressão. Os métodos rf foram otimizados minimizando a RMSE, portanto, as figuras mais representativas de seu desempenho preditivo são as que apresentam os erros quadrados. Dado que as informações apresentadas pelos boxplots dos erros quadrados (disponíveis nas informações complementares) são mínimas, apresentamos principalmente aqui resultados relacionados aos errors absolutos das previsões. As duas abordagens produzem resultados equivalentes, como mostrado nas seções atuais. Além disso, o objetivo principal do presente estudo é a otimização do desempenho preditivo da RF, portanto, os resultados dos métodos dessarmados na Seção 2, devem ser avaliados utilizando-se diversos métodos. Resultados detalhados de todos os experimentos podem ser encontrados na análise.html arquivo no apêndice A.

Na Figura 3, apresentamos um exemplo típico em que testamos os métodos quando aplicados ao Série de 1000 time simulada do modelo ARMA(1, 0) com $\varphi_1 = 0{,}6$. Neste exemplo típico, observamos que todos os métodos têm desempenho semelhante em relação aos erros absolutos e quadrados. O método rf30 é o melhor entre os métodos rf, enquanto arfima tem o desempenho de best e ingênua2 tem o pior desempenho. O método arfima é aproximadamente 5% melhor do que os melhores métodos rf. Os métodos ingênuos1 e têm desempenho semelhante aos melhores métodos rf. O método rf mais simples, ou seja, rf1, tem um bom desempenho. Na verdade, seu desempenho é comparável aos melhores métodos rf, ou seja, rf20imp, rf25 e rf30. Quanto ao uso de variáveis importantes, introduzidas com os métodos rf20imp e rf50imp, seu desempenho é semelhante ao dos métodos rf20 e rf50, respectivamente.

Ao olhar para os coeficientes de regressão, o padrão de desempenho dos métodos é claro. Os métodos rf são melhores quando se usam menos variáveis preditoras. O método ingênuo1 tem um desempenho melhor do que todos os métodos, seguido pelos methods theta e arfima.

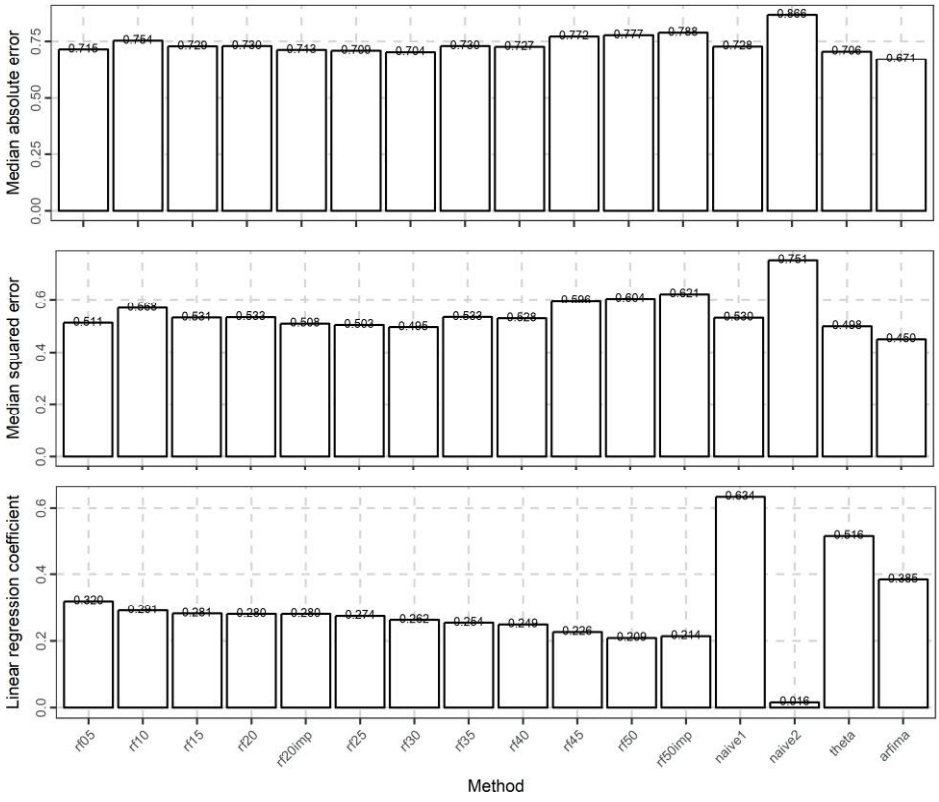

Figura 3. Parcelas das medianas dos erros absolutos, medianas dos erros quadrados e coeficientes de regressão ao prever o 101º valor de 1000 séries temporais simuladas de um modelo ARMA(1,0) com $\varphi_1 = 0{,}6$.

Nas Figuras 4-6, apresentamos os boxplots dos erros absolutos para os casos dos modelos ARMA e ARFIMA. Parece que, na maioria dos casos, os métodos rf com menos variáveis preditoras são melhores em comparação com os métodos rf com muitas variáveis preditoras de acordo com o median dos erros absolutos. A gama de erros absolutos parece ficar mais ampla ao usar mais variáveis preditoras, bem como o número e a magnitude dos outliers. As diferenças entre todos os casos do rf são pequenas. O método arfima é o melhor, como esperado, enquanto os métodos ingênuos e têm desempenho preditivo variado.

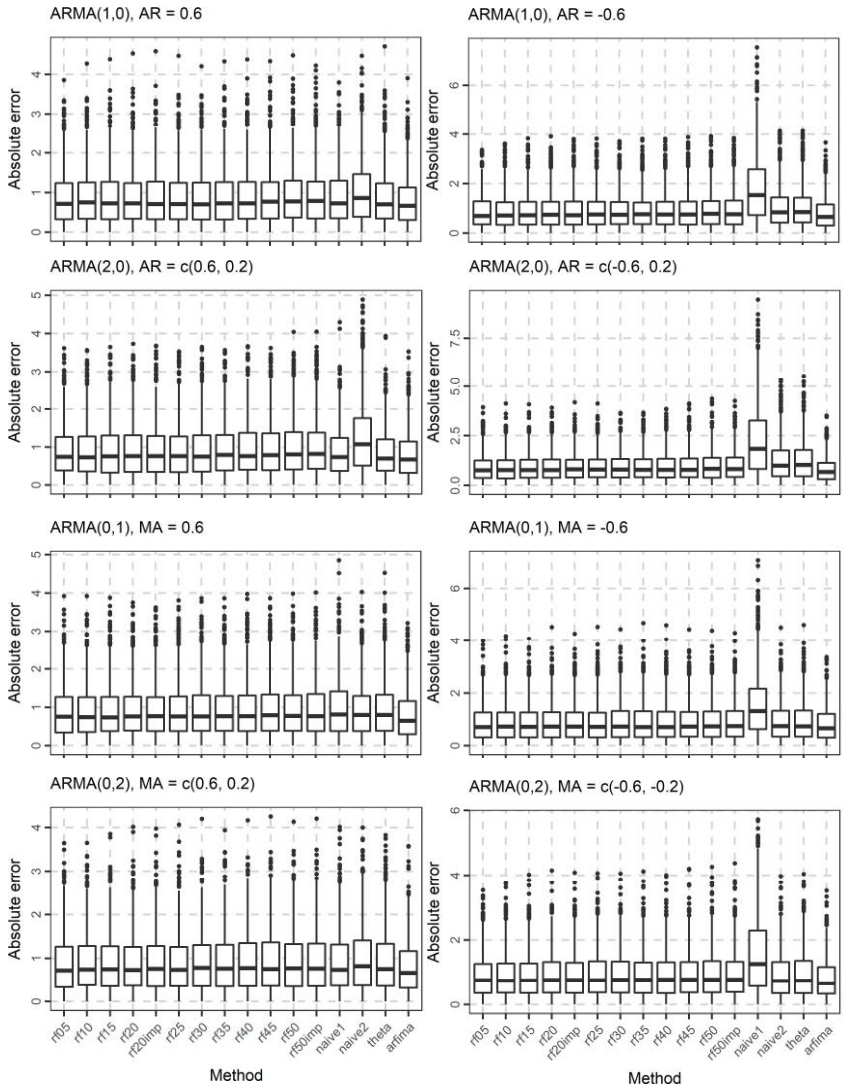

Figura 4. Boxplots dos erros absolutos ao prever o 101º valor de 1000 séries temporais simuladas de cada modelo ARMA(p, 0) e ARMA(0, q) utilizado no presente estudo.

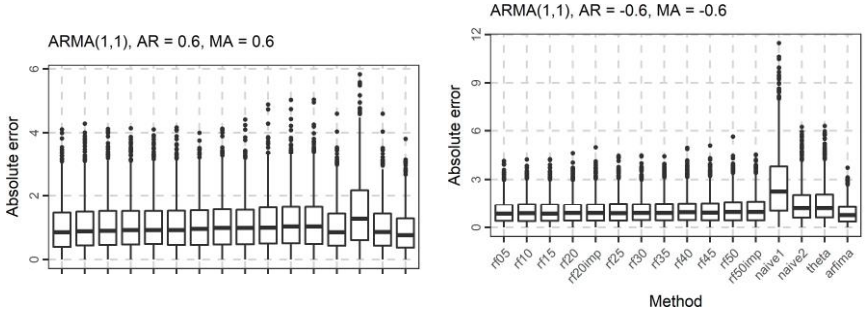

Figura 5. *Cont.*

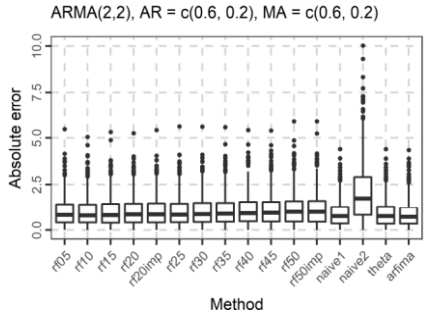

Figura 5. Boxplots dos erros absolutos ao prever o 101º value de 1000 séries temporais simuladas de cada modelo ARMA(*p*, *q*) utilizado no presente estudo.

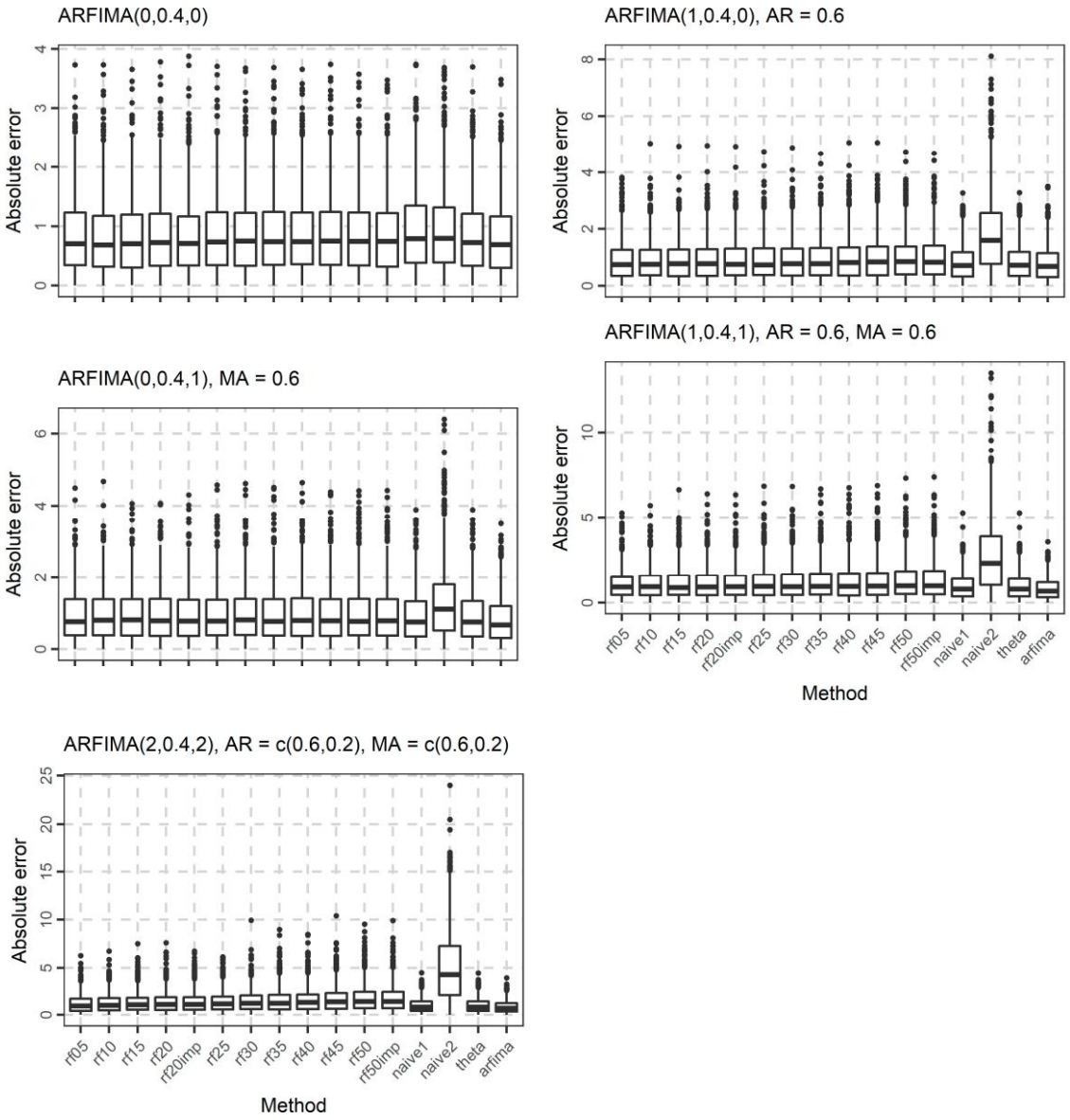

Figura 6. Boxplots dos erros absolutos ao prever o 101º valor de 1000 séries temporais simuladas de cada modelo ARFIMA utilizado no presente estudo.

A destribução dos erros de previsões também é de interesse. Nas Figuras 7-9, apresentamos os boxplots dos erros para os casos ARMA e ARFIMA. Todos os métodos são imparciales, no sentido de que o erro de previsão mediana é aproximadamente igual a 0, enquanto a maioria dos boxplots são simétricos em torno de 0. A dispersão dos erros apresentados nas Figuras 7-9 já foi quantificada pelos boxplots dos erros absolutos nas Figuras 4-6.

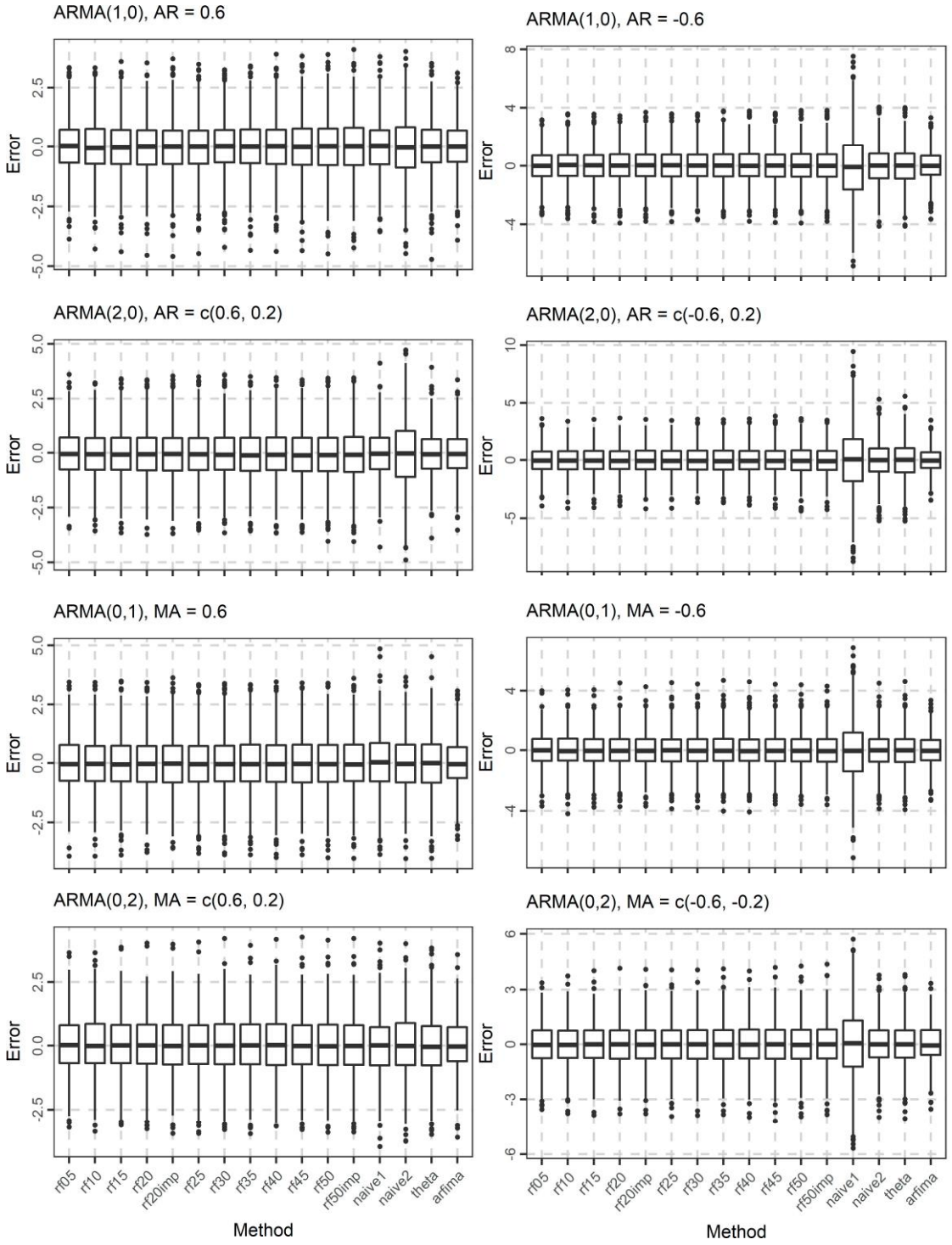

Figura 7. Boxplots dos erros ao prever o 101º valor de 1000 séries temporais simuladas de cada modelo ARMA(*p*, 0) e ARMA(0, *q*) utilizados no presente estudo.

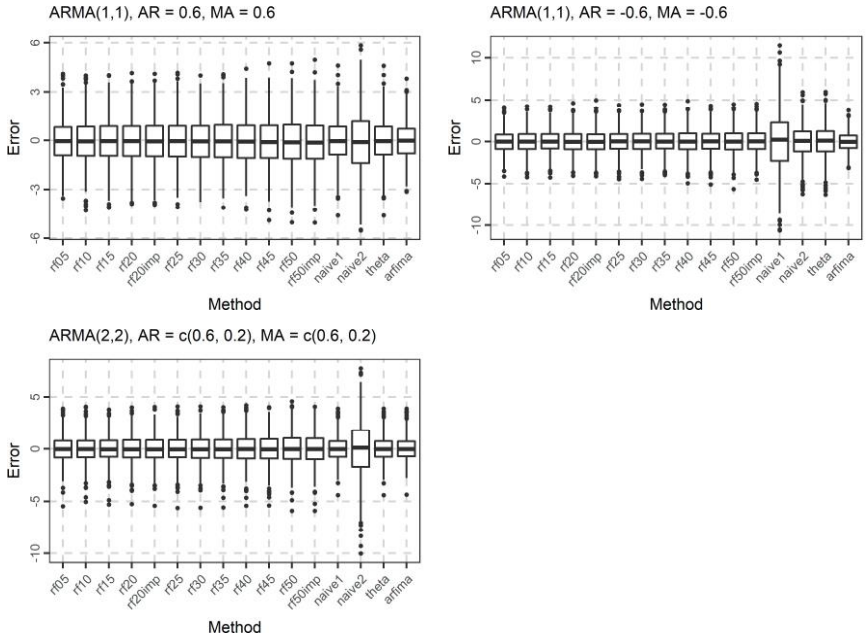

Figura 8. Boxplots dos erros ao prever o 101º valor de 1000 séries temporais simuladas de cada modelo ARMA(*p*, *q*) utilizado no presente estudo.

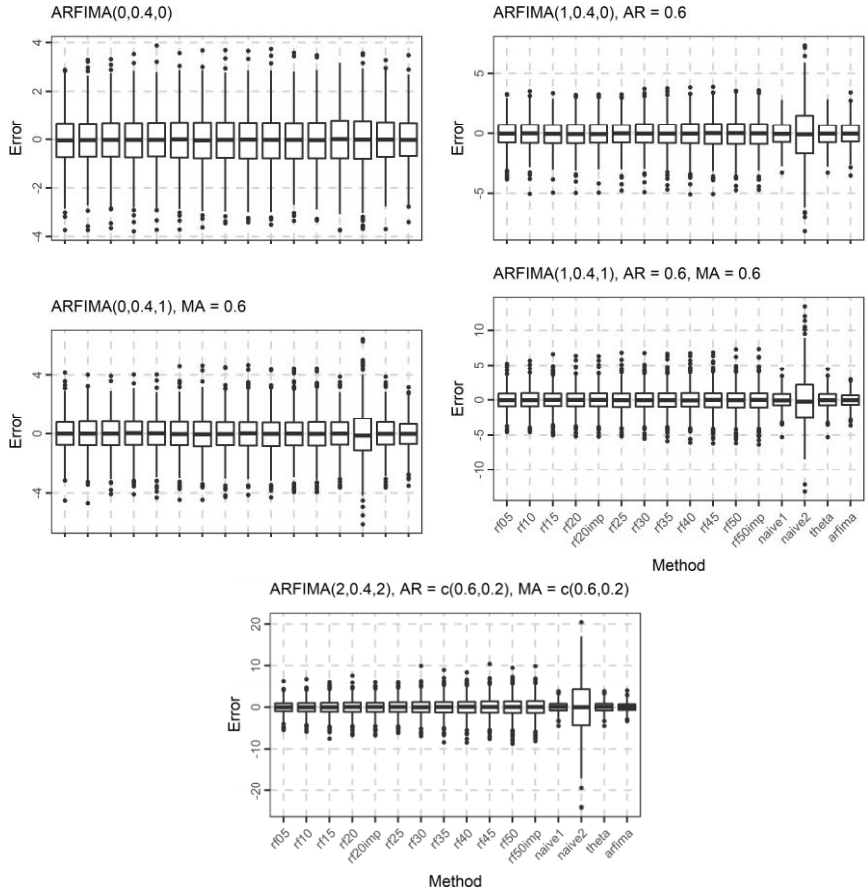

Figura 9. Boxplots dos erros ao prever o 101º valor de 1000 séries tempo simuladas de cada Model ARFIMA utilizado no presente estudo.

Dada a grande quantidade de experimentos e resultados, bem como as pequenas diferenças entre o desempenho dos diversos métodos rf, resumimos os resultados utilizando os rankings dos métodos em cada experimento, por exemplo, na Figura 10, apresentamos os rankings de todos os métodos nos experimentos de simulação em relação à média e mediana dos erros absolutos. Devido às pequenas diferenças entre todos os métodos rf, um mapa de calor apresentando os valores computados para a média e mediana dos erros absolutos (e não suas classificações) seria quase monocromático e, portanto, não apropriado para o propósito do presente estudo.

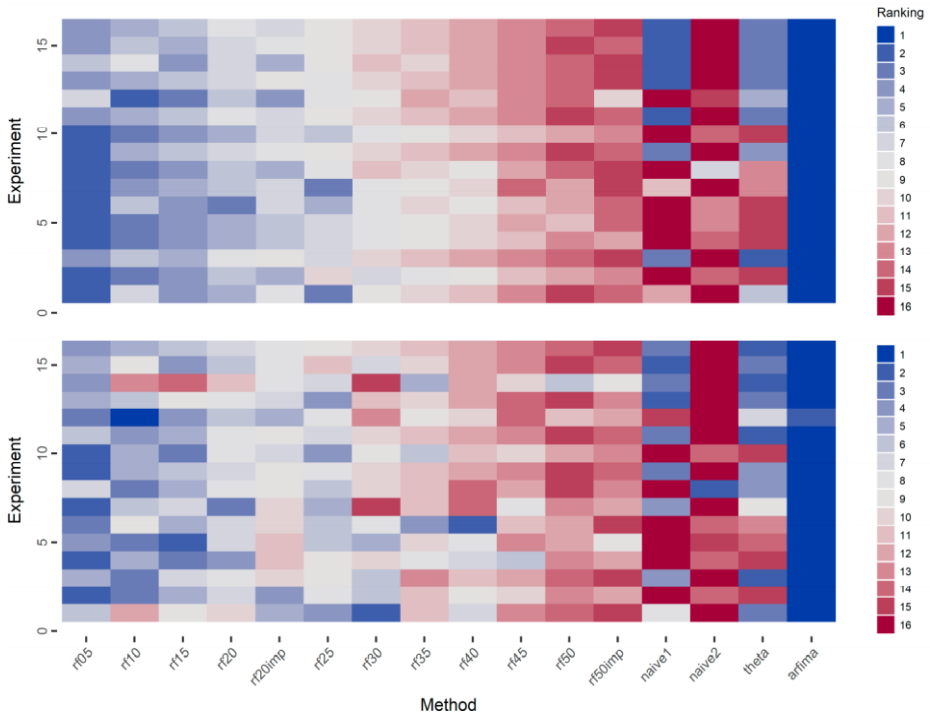

Figura 10. Classificação de métodos dentro de cada experimento de simulação com base na média (**superior**) e mediana (**inferior**) dos erros absolutos. Melhores methods são apresentados com menor valor de classificação e cores azuis.

Quanto aos meios dos erros absolutos, o método arfima apresenta o melhor desempenho em todos os casos. Na maioria dos casos dos experimentos de simulação ARMA (experimentos numerados de um a 11), observamos que os métodos rf são melhores do que os outros métodos. Observa-se um padrão no qual o desempenho dos métodos rf diminui quando o número de variáveis preditoras aumenta. O método rf20imp é geralmente melhor do que o método rf20, enquanto o método rf50imp é pior do que o método rf50. Os métodos ingênuos e têm desempenho semelhante. No caso das simulações ARFIMA, os resultados são semelhantes quanto ao desempenho dos métodos arfima e à intercomatuação entre os métodos rf. No entanto, neste caso, os métodos ingênuos1 e têm um desempenho melhor do que os métodos rf.

Ao olhar para as medianas dos erros absolutos, o método arfima tem o melhor desempenho. Ao contrário dos casos dos meios dos erros absolutos, o pattern do desempenho decrescente dos métodos rf, quando o aumento do número de variáveis preditoras é menos uniforme. No entanto, o padrão geral sugere que o desempenho da RF diminui com o aumento do número de variáveis preditoras. É importante ressaltar a competência dos métodos rf na previsão de séries temporais de pequenas séries temporais. Na verdade, a RF apresenta melhor desempenho em comparação com todos os métodos, exceto o benchmark arfima. Isso definitivamente vale a pena considerar.

Além disso, na Figura 11, apresentamos o ranking dos métodos em relação à média e mediana dos erros quadrados. O padrão é semelhante ao da Figura 10. De fato, o aumento do número de variáveis preditoras diminui o desempenho do algoritmo RF.

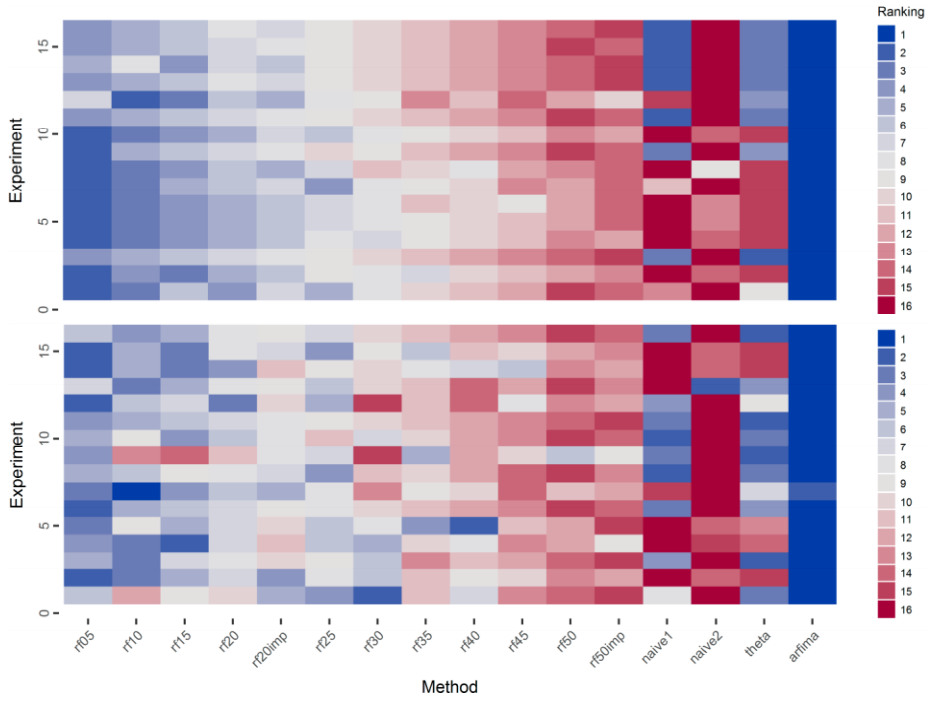

Figura 11. Classificação de métodos em cada experimento de simulação com base na média (**superior**) e mediana (**inferior**) dos erros quadrados. Melhores métodos são apresentados com menor valor de classificação e cores azuis.

Na Figura 12, apresentamos o ranking dos métodos em relação ao coeficiente de regressão. Em geral, os métodos arfima são os de melhor desempenho, enquanto o ingênuo2 é o pior método. Observamos um padrão em relação à série temporal simulada arma em que os métodos ingênuos1 e são os mais asaltados nos experimentos simulados 1, 3, 5, 7, 9 e 11. Nesses experimentos, os valores dos parâmetros dos modelos ARMA são positivos. Em contraste, os experimentos de simulação 2, 4, 6, 8 e 10 correspondem aos modelos ARMA com pelo menos um parameter negativo.

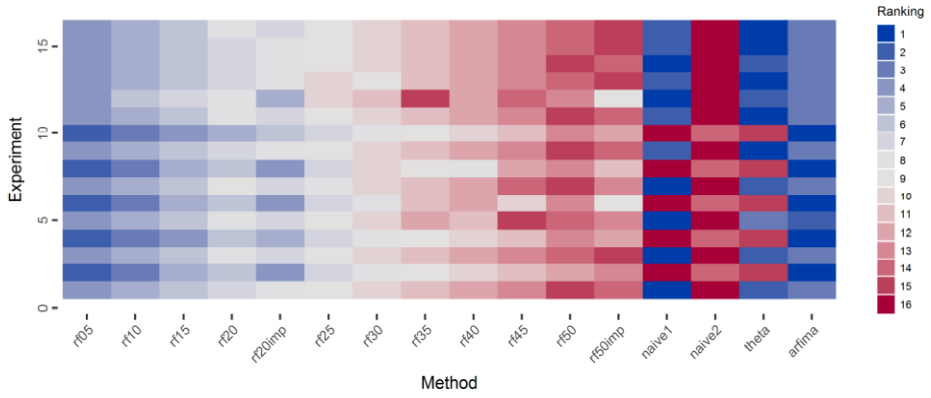

Figura 12. Classificação de métodos em cada experimento de simulação com base nos coeficientes de regressão. Melhores métodos são apresentados com menor valor de classificação e cores azuis.

Em relação aos métodos rf, eles são quase uniformemente melhores quando usam menos variáveis preditoras. Além disso, os métodos utilizando variáveis importantes são melhores, em comparação com os respectivos métodos, nos quais se utiliza o número total de variáveis preditoras.

3.2. Análise de temperatura

Quanto à aplicação dos métodos ao conjunto de dados de temperatura, privamos os resultados completos no arquivo temperature_analysis.html no apêndice A. Aqui apresentamos alguns resultados necessários para a discussão seguinte. Mais detalhadamente, na Figura 13, apresentamos os boxplots das previsões e sua comparação com os valores do teste. Em termos dessa comparação específica, o ingênuo1 apresenta o melhor desempenho, uma vez que a mediana e o alcance de suas previsões se assemelham muito aos valores de teste correspondentes. O método ingênuo2 é o pior. Os outros métodos são semelhantes.

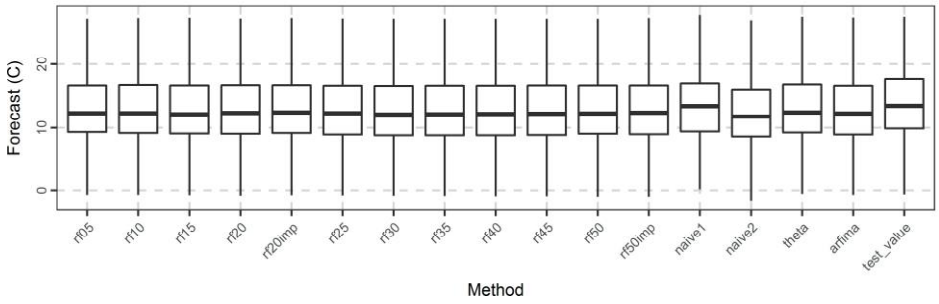

Figura 13. Boxplot de previsões de temperatura para cada método.

Para destacar possíveis diferenças entre os métodos rf, apresentamos ainda algumas comparações com base nos valores métricos. Na Figura 14, apresentamos os boxplots do erro, erro absoluto, erro quadrado e valores de erro percentual absoluto. Os erros são principalmente negativos. Além disso, as respectivas distribuições dos métodos rf estão mais próximas entre si do que das respectivas distribuições de ingênuos1, ingênuos2 e, enquanto a distribuição dos errors do método arfima é bastante mais próxima das dos métodos rf. O mesmo se aplica às distribuições das outras métricas. Os outliers muito produzidos em termos de erro percentual absoluto por quase todos os métodos podem ser explicados pela presença de valores de temperatura of, que são próximos de zero.

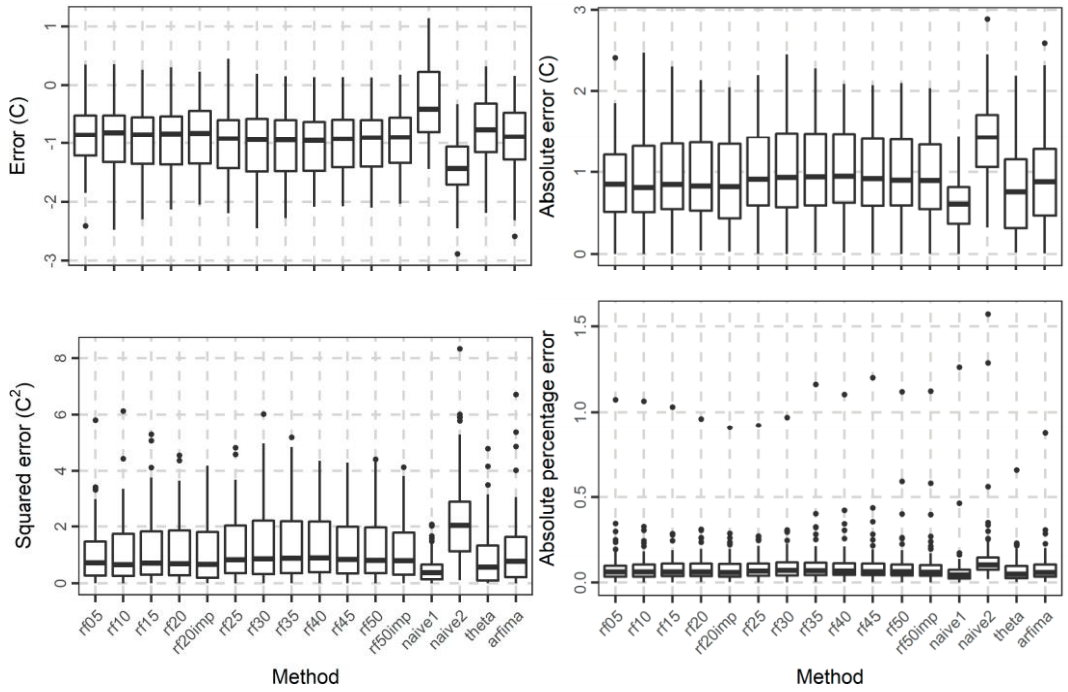

Figura 14. Boxplots do erro, erro absoluto, erro quadrado e valores de erro percentual absoluto das previsões de temperatura para todos os métodos.

Além disso, a Figura 15 concentra-se na performance média dos métodos em relação a todas as métricas computadas (apresentadas na Seção 2.1.9). Uma comparação com esse desempenho pode facilitar a classificação dos métodos. No que diz respeito a este ranking em particular, os métodos ingênuos1 e theta são os melhores, enquanto o método ingênuo2 é claramente o pior. O método arfima exibe pior desempenho do que ingênuo1 e, mas melhor do que o método ingênuo2.

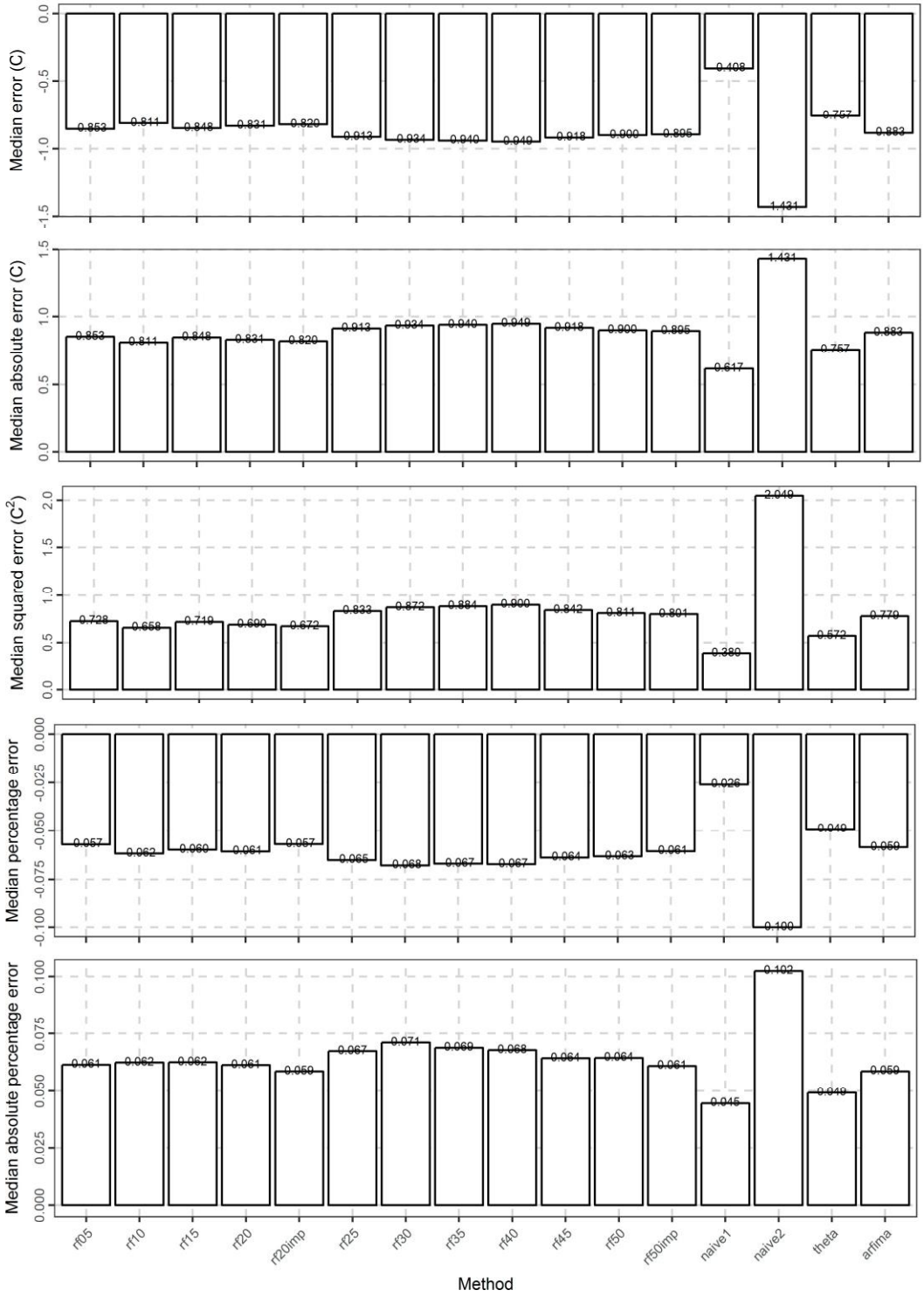

Figura 15. Barplots das medianas de vários tipos de métricas que medem o erro das previsões de temperatura para todos os métodos. Os tipos de erros são retratados nos eixos verticais.

Em relação aos métodos rf, o melhor desempenho é exibido pelo método rf10 em termos da mediana de erros, mediana de erro absoluto e mediana de erros quadrados. Pelo contrário, em relação à mediana dos erros percentuais e à mediana dos erros percentuais absolutos, o melhor método rf é o rf20imp. Em geral, os métodos rf10, rf15, rf20, rf20, rf20imp estão próximos em termos dos valores medianos das métricas. O mesmo se aplica ao desempenho de todos os métodos rf restantes, que, no entanto, são piores. O método arfima tem um desempenho semelhante com the
piores métodos rf.

Na Figura 16, apresentamos os coeficientes de regressão entre os valores previstos e os de teste. Aqui as diferenças entre os métodos são insignificantes. O melhor método é o, seguido pelos métodos rf05 e arfima. Além disso, o coeficiente de regressão aumenta com o aumento do número de variáveis preditoras.

Além disso, usar as variáveis importantes resulta em melhores resultados, por exemplo, rf20imp é melhor que rf20.

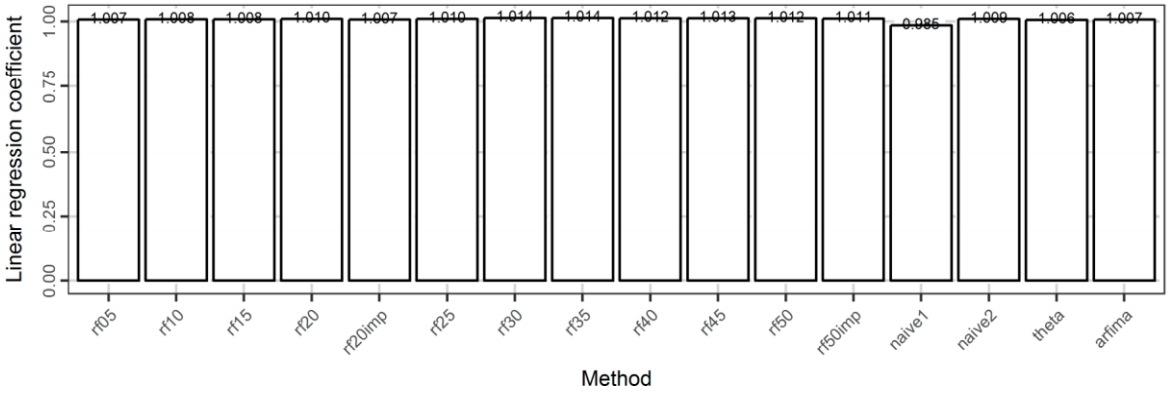

Figura 16. Partes do coeficiente de regressão do modelo de linhalar entre os valores previstos e os valores de teste do conjunto de dados de temperatura.

4. Discussão

A maioria dos estudos com foco na previsão de séries temporais visa propor um método de previsão baseado em seu desempenho (em comparação com outros métodos, geralmente alguns), quando aplicado dentro de um pequeno número de estudos de caso (por exemplo, [8,14,26,28,30]). Reconhecendo este fato específico e visando fornecer uma contribuição tangível na previsão de séries temporâneas usando RF, realizamos um extenso conjunto de experimentos computacionais. Esses experimentos foram projetados para criar um fundo empírico sobre o efeito da seleção variável no desempenho de previsão, ao implementar a RF em uma etapa de previsão de séries de curto prazo. Examinamos dois grandes conjuntos de dados, um conjunto de dados composed de séries temporizais simuladas e um conjunto de dados de temperatura. Este último é utilizado complementar ao primeiro para achados mais conciso.

O resultado fornecido é claro para o procedimento selecionado de ajuste de parâmetros e para ambos os conjuntos de dados examinados e, portanto, gostaríamos de sugerir sua consideração em aplicações futuras. A RF teve um desempenho melhor principalmente ao usar algumas variáveis preditoras recentemente defasadas. Esse resultado poderia ser explicado pelo fato de que o aumento do número de variáveis defasadas inevitavelmente leva à redução da duração do conjunto de treinamento e concomitantemente à redução das informações exploradas da série temporal original durante a fase de montagem do modelo. Ao contrário, o uso de muitas variáveis preditoras em outros tipos de problemas de regressão não afeta seriamente o desempenho preditivo de RF [34] (p. 489).

Em geral, os resultados sugerem que o algoritmo RF tem um bom desempenho em uma etapa antes da previsão de séries de curto prazo. No entanto, outros algoritmos de previsão podem exibir melhor performance em alguns casos. Isso não se aplica a todos os experimentos realizados, nem implica que a RF não deve ser usada em séries de tempo de um passo à frente, prevendo a favor desses outros algoritmos, como mostrado em [24-26,30]. De fato, em [12] o leitor interessado pode encontrar uma comparação em pequena escala em 50 séries temporísticas geofísicas, o que ilustra o fato de que um algoritmo pode ter um desempenho melhor ou pior dependendo da série temporal examinada.

Além disso, focamos no algoritmo rf e tentamos melhorar sua performance, sob a condição de que este algoritmo fosse considerado competente. Portanto, há uma questão de saber se a condição de que o algoritmo de florestas aleatórias é competente é verdadeira, o que será respondido a seguir. As diferenças entre os métodos de previsão de the são em sua maioria pequenas e o desempenho do algoritmo RF é sempre comparável com o desempenho de outros algoritmos (esta é a razão pela qual usamos a classificação dos métodos nos heatmaps e não os valores das métricas). O algoritmo RF foi encontrado competente nos experimentos usando o conjunto de dados simulado, sendo pior do que o algoritmo arfima (uma boa referência para o caso da série temporal simulada) apenas, mas em geral melhor em comparação com os outros métodos (métodos ingênuos1, ingênuos2 e teta). O algoritmo RF foi encontrado competente no experimento usando o conjunto de dados de temperatura, sendo melhor do que os métodos arfima e ingênuo2 e pior do que ingênuo1 e. Verificou-se também que o RF subótimo (ou seja, aqueles que utilizam

muitas variáveis preditoras) foram em sua maioria piores que os outros métodos, enquanto o RF otimizado foi competente, destacando assim o papel da seleção variável na melhoria do desempenho da RF.

Com base neste fato, gostaríamos de propor o uso de uma variedade de algoritmos em cada aplicação de previsão. Além disso, uma vez que acreditamos que todas as escolhas metodológicas (por exemplo, a seleção variável na previsão do aprendizado de máquina) devem ser baseadas na teoria, ou pelo menos em evidências (que podem ser fornecidas por estudos extensivos, como o presente), em vez de serem feitas à vontade, sugerimos que o desempenho de todos os algoritmos a serem usados em uma aplicação específica deveria ter sido previamente otimizado.

Além da principal contribuição do estudo present, alguns outros achados notáveis também foram derivados através das análises. Esses achados são resumidos posteriormente e destacam os méritos do quadro metodológico adotado. Em geral, observamos pequenas diferenças nos vários métodos de RF que, no entanto, podem ser essenciais para algumas aplicações. É importante ressaltar que métodos simples podem ter um desempenho extremamente bom em relação ao resto em esquemas específicos de autocorrelação, como também concluído em outros grandes comparais (por exemplo, [37]) ou relatados por evidências em estudos menores (por exemplo, [12]). Particularmente para o experimento realizado em séries tempoetrais do mundo real, observamos que o erro absoluto mínimo na previsão da temperatura do próximo ano usando RF foi aproximadamente equal para 0,6 ˚C. Este resultado é definitivamente importante para as geociências, sugerindo que a temperatura é um processo difícil de prever.

Além disso, o presente estudo é um dos primeiros a aplicar benchmarking para avaliar o desempenho de previsões de florestas aleatórias na previsão de séries temporâneas e, portanto, uma discussão sobre sua eficácia é significativa. A realização de um desempenho optimal de longo prazo de um algoritmo específico de aprendizagem de máquina requer uma estratégia firme, que poderia resultar através de estudos como o presente. O benchmarking mostrou-se útil para destacar a pequena magnitude das diferenças entre os vários métodos rf, fornecendo um limite superior e inferior no desempenho deste último, permitindo assim uma das principais conclusões, ou seja, que a RF é competente em uma etapa antes da previsão de pequenas séries temporências. Portanto, reconhecemos os benefícios da implementação de benchmarking dentro de comparações em larga escala para criar uma imagem mais fiel da bondade do desempenho caracterizando os algoritmos de previsão que dificilmente podem ser examinados analiticamente, por exemplo, outros algoritmos de aprendizagem de máquina.

Finalmente, a reprodutibilidade foi uma consideração fundamental deste artigo. Os códigos estão disponíveis nas informações complementares, enquanto o leitor é encorajado a usá-los para fins de verificação ou para atividades futuras de pesquisa.

5. Conclusões

Florestas aleatórias estão entre os algoritmos de aprendizado de máquina mais populares. A melhoria de seu desempenho em uma etapa de previsões de séries de curto prazo certamente valeu a pena tentar, uma vez que eles também provaram ser um algoritmo de previsão competente. Essas melhorias de desempenho e competências são sugeridas por um extenso conjunto de experimentos, que usam grandes conjuntos de dados em conjunto
com benchmarking de desempenho. Os resultados desses experimentos indicam claramente que o uso de algumas variáveis recentes como preditores durante o processo de montagem leva a uma maior precisão preditiva para o algoritmo florestal aleatório usado neste estudo.

Claro, há limitações em nosso estudo. Em particular, os resultados foram obtidos utilizando-se séries de tempo curto, enquanto pode haver melhores procedimentos para encontrar os conjuntos ideais de parâmetros. No entanto, este é o maior experimento realizado com o objetivo acima mencionado. Esperamos que este resultado seja útil em aplicações futuras.

Materiais Complementares: O material suplementar está disponível online na www.mdpi.com/1999-4893/10/4/ 114/s1.

Agradecimentos: Agradecemos a três revisores anônimos e ao Editor Acadêmico cujas sugestões e comentários ajudaram a melhorar consideravelmente o manuscrito.

Contribuições do Autor: Os dois autores trabalharam no artigo e na implementação das metodologias utilizando a linguagem de programação R de forma igual e em todos os seus aspectos. No entanto, Hristos Tyralis é o principal escritor deste artigo e Georgia Papacharalampous concebeu a ideia.

Conflitos de Interesse: Os autores não declaram conflito de interesses.

Apêndice Um

Os materiais suplementares encontrados como um arquivo zip em http://dx.doi.org/10.17632/nr3z96jmbm.1 incluem todos os códigos e dados utilizados no manuscrito. Há um arquivo readme com instruções detalhadas sobre how para executar cada código e salvar os resultados. Dois arquivos .html, incluindo os resultados e visualizações aqui apresentados, bem como mais resultados que não poderiam ser incluídos aqui por razões de brevidade, mas melhorar nossa descoberta, também pode ser encontrado no material flexível.

As análises e visualizações foram realizadas em R Programming Language [52] utilizando-se as embalagens R contributivas caret [50,51], previsão [55,56], fracdiff [54], gdata [66], ggplot2 [67], randomForest [60], readr [68], remodele2 [69], e rminer [61,62].

Referências

1. Shmueli, G. Para explicar ou prever? *Stat. Sci.* **2010**, *25*, 289-310. [CrossRef]
2. Bontempi, G.; Taieb, S.B.; Le Borgne, Y.A. Estratégias de aprendizado de máquina para previsão de séries tempores. Nos *Negócios*
   Inteligência (Notas de Palestras no Processamento de Informações Empresariais); Aufaure, M.A., Zimányi, E., Eds.; Springer:
   Berlim/Heidelberg, Alemanha, 2013; Volume 138, pp. 62-77. [CrossRef]
3. De Gooijer, J.G.; Hyndman, R.J. 25 anos de previsão de séries temporâneos. *Previsão int. J.* **2006**, *22*, 443-473. [CrossRef]
4. Fildes, R.; Nikolopoulos, K.; Crone, S.F.; Syntetos, A.A. Previsão e pesquisa operacional: Uma revisão.
   J. Oper. Res. Soc. **2008**, *59*, 1150-1172. [CrossRef]
5. Weron, R. Previsão de preços de eletricidade: Uma revisão do estado da arte com um olhar para o futuro. *Previsão int. J.* **2014**, *30*, 1030-1081. [CrossRef]
6. Hong, T.; Previsão de carga elétrica probabilística: Uma revisão tutorial. *Previsão int. J.* **2016**, *32*, 914-938. [CrossRef]
7. Taieb, S.B.; Bontempi, G.; Atiya, A.F.; Sorjamaa, A. Uma revisão e comparação de estratégias para a previsão de séries temporísticas de várias etapas com base na competição de previsão nn5. *Especialista Syst. Appl.* **2012**, *39*, 7067-7083. [CrossRef]
8. Mei-Ying, Y.; Xiao-Dong, W. Previsão de séries temporentais caóticas usando máquinas vetoriais de suporte menos quadrados. *Chin. Phys.* **2004**, *13*, 454-458. [CrossRef]
9. Faraway, J.; Chatfield, C. Previsão da série temporal com redes neurais: Um estudo comparativo usando os dados da linha aérea. *J.R. Stat. Soc.C Appl. Stat.* **1998**, *47*, 231-250. [CrossRef]
10. Yang, B.S.; Oh, EsM. Tan, A.C.C. Prognóstico de condição da máquina baseado em árvores de regressão e previsão de um passo à frente. *Mech. Syst. Processo de sinal.* **2008**, *22*, 1179-1193. [CrossRef]
11. Zou, H.; Yang, Y. Combinando modelos de séries temporédias para previsão. *Previsão int. J.* **2004**, *20*, 69-84. [CrossRef]
12. Papacharalampous, G.A.; Tyralis, H.; Koutsoyiannis, D. Previsão de processos geofísicos usando algoritmos estocásticos e de aprendizagem de máquina. Em Processo do 10º Congresso Mundial da EWRA sobre Recursos Hídricos e Meio Ambiente "Panta Rhei", Atenas, Grécia, 5 a 9 de julho de 2017.
13. Pérez-Rodríguez, J.V.; Torra, S.; Modelos Andrada-Félix, J. STAR e ANN: Previsão de desempenho no índice espanhol "Ibex-35". *J. A Empir. O Financ.* **2005**, *12*, 490-509. [CrossRef]
14. Khashei, M.; Bijari, M. Uma nova hibridização de redes neurais artificiais e modelos ARIMA para forecasting de séries temporais. *Appl. Soft Comput.* **2011**, *11*, 2664-2675. [CrossRef]
15. Yan, W. Para a previsão automática da série de tempo usando redes neurais. *IEEE Trans. Neural Netw. Lear. Stat.* **2012**, *23*, 1028-1039. [CrossRef]
16. Babu, C.N.; Reddy, B.E. Um modelo ARIMA-ANN baseado em filtro móvel para prever dados da série temporal. *Appl. Soft Comput.* **2014**, *23*, 27-38. [CrossRef]
17. Lin, L.; Wang, F.; Xie, X. ; Zhong, S. Conjunto de máquinas de aprendizagem extrema baseada em florestas aleatórias para previsão de séries temporências multi-regime. *Especialista Syst. Appl.* **2017**, *85*, 164-176. [CrossRef]
18. Breiman, L. Random Forests. *Mach. Aprenda.* **2001**, *45*, 5-32. [CrossRef]
19. Scornet, E.; Biau, G.; Vert, J.P. Consistência de florestas aleatórias. *Stat.* **2015**, *43*, 1716-1741. [CrossRef]
20. Biau, G.; Scornet, E. Uma visita aleatória guiada pela floresta. *Teste* **2016**, *25*, 197-227. [CrossRef]
21. Hastie, T.; Tibshirani, R.; Friedman, J. *Os Elementos da Aprendizagem Estatística*, 2ª ed.; Nova Iorque, NY, EUA, 2009. [CrossRef]

22. Verikas, A.; Gelzinis, A.; Bacauskiene, M. Dados de mineração com florestas aleatórias: Um levantamento e resultados de novos testes. *Padrão Recognit.* **2011**, *44*, 330-349. [CrossRef]
23. Herrera, M.; Torgo, L.; Izquierdo, J.; Pérez-García, R. Modelos preditivos para previsão de demanda de água urbana por hora. *J. O Hydrol.* **2010**, *387*, 141-150. [CrossRef]
24. Dudek, G. Previsão de carga de curto prazo usando florestas aleatórias. Em Processo da 7ª Conferência Internacional IEEE Sistemas Inteligentes IS'2014 (Avanços em Sistemas Inteligentes e Computação), Varsóvia, Polônia, 24 a 26 de setembro de 2014; Filev, D., Jabłkowski, J., Kacprzyk, J., Krawczak, M., Popchev, I., Rutkowski, L., Sgurev, V., Sotirova, E., Szynkarczyk, P., Zadrozny, S., Eds.; Springer: Cham, Suíça, 2015; Volume 323, pp. 821-828. [CrossRef]
25. Chen, J.; Li, M.; Wang, W. Estimativa de incerteza estatística usando florestas aleatórias e sua aplicação à previsão de seca. *Matemática. O Probl. Eng.* **2012**, *2012*, 915053. [CrossRef]
26. Naing, W.Y.N.; Previsão de variações mensais de temperatura usando florestas aleatórias. *APRN J. Eng. Appl. Sci.* **2015**, *10*, 10109-10112.
27. Nguyen, T.T.; Huu, Q.N.; Li, M.J. Previsão de níveis de água em séries temporâneas no rio Mekong usando modelos de aprendizado de máquina. Em Proceedings of the 7015 International Conference on Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnã, 8-10 outubro 2015. [CrossRef]
28. Kumar, M.; Thenmozhi, M. Previsão de movimento do índice de ações: Uma comparação de máquinas vetoriais de suporte e floresta aleatória. No *Indian Institute of Capital Markets 9ª Conferência de Mercado de Capitais;* Instituto Indiano de Mercado de Capitais: Vashi, Índia, 2006. [CrossRef]
29. Kumar, M.; Thenmozhi, M. Previsão de retornos do índice de ações usando modelos híbridos florestais ARIMA-SVM, ARIMA-ANN e ARIMA. *Int. J. Bank. Acc. O Financ.* **2014**, *5*, 284-308. [CrossRef]
30. Kane, M.J.; Preço, N.; Uísque, M.; Rabinowitz, P. Comparação dos modelos da série tempo ARIMA e Random Forest para previsão de surtos de influenza aviária H5N1. *BMC Bioinform.* **2014**, 15. [CrossRef] [PubMed]
31. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Seleção variável usando florestas aleatórias. *Padrão Recognit. Letão.* **2010**, *31*, 2225-2236. [CrossRef]
32. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. Quantas árvores em uma floresta aleatória? Em *Machine Learning e Data Mining in Pattern Recognition (Notas de Palestra em Ciência da Computação)*; Perner, P., Ed. Springer: Berlin/Heidelberg, Alemanha, 2012; pp. 154-168. [CrossRef]
33. Probst, P.; Boulesteix Para sintonizar ou não afinar o número de árvores em florestas aleatórias? *arXiv* **2017**, arXiv:1705.05654v1.
34. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Nova York, NY, EUA, 2013. [CrossRef]
35. Díaz-Uriarte, R.; De Andres, S.A. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **2006**, *7*. [CrossRef] [PubMed]
36. Makridakis, S.; Hibon, intervalos de confiança: Uma investigação empírica da série na competição M. *Previsão int. J.* **1987**, *3*, 489-508. [CrossRef]
37. Makridakis, S.; Hibon, M. A Competição M3: Resultados, conclusões e implicações. *Previsão int. J.* **2000**, *16*, 451-476. [CrossRef]
38. Pritzsche, U. Benchmarking de algoritmos clássicos e de aprendizagem de máquina (com ênfase especial em ensacar e impulsionar abordagens) para previsão de séries temporitas. Tese de Mestrado, Ludwig-Maximilians-Universität München, München, Alemanha, 2015.
39. Bagnall, A.; Cawley, G.C. Sobre o uso de configurações padrão de parâmetros na avaliação empírica dos algoritmos de classificação. *arXiv* **2017**, arXiv:1703.06777v1.
40. Salles, R.; Assis, L.; Guedes, G.; Bezerra, E.; Porto, F.; Ogasawara, E. Uma estrutura para benchmarking métodos de aprendizado de máquina usando modelos lineares para previsão de séries temporativas univariadas. Em Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, EUA, 14-19 may 2017; pp. 2338-2345. [CrossRef]
41. Bontempi, G. Estratégias de Aprendizado de Máquina para Previsão de Séries Tempoestas. Escola europeia de Verão de Business Intelligence, Hammamet, Palestra. 2013. Disponível online: https://pdfs.semanticscholar.org/f8ad/a97c142b0a2b1bfe20d8317ef58527ee329a.pdf (acessado em 25 de setembro de 2017).
42. McShane, B.B. Métodos de Aprendizagem de Máquina com Dependência de Séries Tempoesas. Tese de Doutorado, University da Pensilvânia, Filadélfia, PA, EUA, 2010.
43. Bagnall, A.; Bostrom, A.; Grande, J.; Linhas, J. Experimentos simulados de dados para classificação de séries temporísticas parte 1: Comparação de precisão com configurações padrão. *arXiv* **2017**, arXiv:1703.09480v1.
44. Caixa, G.E.P.; Jenkins, G.M. Alguns avanços recentes na previsão e controle. *J.R. Stat. Soc.C Appl. Stat.* **1968**, *17*, 91-109. [CrossRef]

45. Wei, W.W.S. *Time Series Analysis, Univariate e Multivariate Methods*, 2ª ed.; Pearson Addison Wesley: Boston, MA, EUA, 2006; ISBN 0-321-322116-9.
46. Thissen, U.; Van Brakel, R.; De Weijer, A.P.; Melssena, W.J.; Buydens, L.M.C. Usando máquinas vetoriais de suporte para previsão de séries tempores. *O Chemom. Intell. Labrador.* **2003**, *69*, 35-49. [CrossRef]
47. Zhang, G.P. Uma investigação de redes neurais para forecasting linear de séries temporais. *Computação. Res.* **2001**, *28*, 1183-1202. [CrossRef]
48. Lawrimore, J.H.; Menne, M.J.; Gleason, B.E.; Williams, C.N.; Wuertz, D.B.; Vose, R.S.; Rennie, J. Uma visão geral do conjunto mensal de dados de temperatura média da Rede Histórica Global, versão 3. *J. Geofísica. Res.* **2011**, *116*. [CrossRef]
49. Assimakopoulos, V.; Nikolopoulos, K. O modelo teta: uma abordagem de decomposição para a previsão. *Previsão int. J.* **2000**, *16*, 521-530. [CrossRef]
50. Kuhn, M. Construindo modelos preditivos em R usando o pacote de cuidado. *Stat. O Softw*. **2008**, *28*. [CrossRef]
51. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z. Kenkel, B.; A Equipe R Core; et al. *Caret: Treinamento de Classificação e Regressão*, pacote R versão 6.0-76; 2017. Disponível na linha: https://cran.r-project.org/web/packages/caret/index.html (acessado em 7 de setembro de 2017).
52. A Equipe R Core. *R: Uma linguagem e ambiente para computação estatística*; R Foundation para Computação Estatística: Viena, Áustria, 2017.
53. Hemelrijk, J. Sublinhando variáveis aleatórias. *Stat. O Neerl.* **1966**, *20*, 1-7. [CrossRef]
54. Fraley, C.; Leisch, F.; Maechler, M.; Reisen, V.; Lemonte, A. *Fracdiff: ARIMA fracionária aka ARFIMA (p,d,q) Modelos*, R pacote versão 1.4-2; 2012. Disponível online: https://rdrr.io/cran/fracdiff/ (acessado em 2 de dezembro de 2012).
55. Hyndman, R.J.; O'Hara-Wild, M.; Bergmeir, C.; Razbash, S.; Wang, E. *Previsão: Funções de previsão para séries tempores e modelos lineares*, série R versão 8.1; 2017. Disponível online: https://rdrr.io/cran/forecast/ (acessado em 25 de setembro de 2017).
56. Hyndman, R.J.; Khandakar, Y. Previsão automática da série de tempo: O pacote de previsão para R. *Stat. O Softw.* **2008**, *27*. [CrossRef]
57. Hyndman, R.J.; Athanasopoulos, G. *Previsão: Princípios e Prática;* OTexts: Melbourne, Austrália, 2013. Disponível online: http://otexts.org/fpp/ (acessado em 25 de setembro de 2017).
58. Hyndman, R.J.; Billah, B. Desmascarando o método Theta. *Previsão int. J.* **2003**, *19*, 287-290. [CrossRef]
59. Hyndman, R.J.; Koehler, A.B.; Ord, J.K.; Snyder, R.D. *Previsão com Suavização Exponencial: A Abordagem do Espaço Do Estado*; Springer: Berlim/Heidelberg, Alemanha, 2008. [CrossRef]
60. Liaw, A.; Wiener, M. Classificação e Regressão por RandomForest. *R News* **2002**, 2, 18-22.
61. Cortez, P. Mineração de dados com redes neurais e máquinas vetoriais de suporte usando a ferramenta R/rminer . Em *Avanços na Mineração de Dados. Aplicações e Aspectos Teóricos (Notas de Palestra em Inteligência Artificial)*; Perner, P., Ed.; Springer: Berlim/Heidelberg, Alemanha, 2010; Volume 6171, pp. 572-583. [CrossRef]
62. Cortez, P. *Rminer: Métodos de Classificação e Regressão de Mineração de Dados*, R pacote versão 1.4.2; 2016. Disponível online: https://r drr.io/cran/rminer/ (acessado em 2 de setembro de 2016).
63. Hyndman, R.J.; Koehler, A.B. Outro olhar para as medidas de precisão da previsão. *Previsão int. J.* **2006**, *22*, 679-688. [CrossRef]
64. Alexander, D.L.J.; Tropsha, A.; Winkler, D.A. Cuidado com *o R2*: Avaliação simples e inequívoca da precisão da previsão dos modelos QSAR e QSPR. *J. Chem. Modelo.* **2015**, *55*, 1316-1322. [CrossRef] [PubMed]
65. Gramatica, P.; Sangion, A. Um histórico sobre os parâmetros de validação estatística para modelos QSAR: Um esclarecimento sobre métricas e terminologia. *J. Chem. Modelo.* **2016**, *56*, 1127-1131. [CrossRef] [PubMed]
66. Warnes, G.R.; Bolker, B.; Gorjanc, G.; Grothendieck, G.; Korosec, A.; Lumley, T.; MacQueen, D.; Magnusson, A.; Rogers, J. *Gdata: Várias ferramentas de programação R para manipulação de dados*, r pacote versão 2.18.0; 2017. Disponível online: https://cran.r-project.org/web/packages/gdata/index.html (acessado em 6 de junho de 2017).
67. Wickham, H. *Ggplot2: Gráficos Elegantes para Análise de Dados*, 2ª ed.; Springer International Publishing: Cham, Suíça, 2016. [CrossRef]
68. Wickham, H.; Hester, J.; François, R.; Jylänki, J.; Jørgensen, M. *Leia dados de texto retangulares*, versão do pacote R 1.1.1; 2017. Disponível online: https://cran.r-project.org/web/packages/readr/index.html (acessado em 16 de maio de 2017).
69. Wickham, H. Remodelando dados com o package remodele. *Stat. O Softw.* **2007**, *21*. [CrossRef]