Métodos de Seleção Automática de Modelos ARIMA no Software R: Uma Comparação dos Algoritmos do Pacote *Forecast*

Discente: Franklin Piauhy Neto
Orientador: Prof. Dr. Marcelo Ruy

Resumo:
Dentre as aplicações de análise de séries temporais, a previsão de valores futuros é uma das mais utilizadas. Nas previsões, a precisão do método utilizado é um dos fatores críticos em sua adoção. Um dos principais métodos de previsão disponíveis é o modelo ARIMA, cujo ponto negativo é a complexidade para se especificar o modelo mais adequado. Atualmente, com o aumento no poder computacional e a crescente disponibilidade de software estatístico, já é possível encontrar soluções completamente automatizadas, tal como o pacote "*forecast*" escrito na linguagem de programação R. O pacote possui dois algoritmos para a especificação de modelos: busca exaustiva e em etapas (*stepwise*), este último uma opção para se encurtar o tempo de computação. O presente trabalho teve como objetivo testar se há diferença significativa entre as precisões das previsões pontuais geradas por ambos os algoritmos. Para tanto, 2.829 séries anuais, trimestrais e mensais da *M3 Competition* foram selecionadas, previsões com os dois métodos foram executadas e suas precisões foram calculadas. A conclusão foi que para as séries anuais e trimestrais os métodos tiveram em média o mesmo grau de precisão e para as séries mensais o procedimento exaustivo foi em média ligeiramente melhor.
Palavras chave: Séries Temporais Univariadas; Métodos de Previsão; Comparação de Métodos; Erros de Previsão.

Methods for Automatic Selection of ARIMA Models in Software R: A Comparison of the Algorithms of Forecast Package

Abstract:
In time series analysis, forecasting is one of the most important applications. In forecasting, the accuracy of the method used is one of the critical factors in its adoption. One of the main available forecasting methods is ARIMA model, whose downside is the complexity to specify the most adequate model. Currently, with the increase in computational power and in the availability of statistical software, it is already possible to find fully automated solutions, such as the "forecast" package written in the R programming language. The package has two algorithms for specifying models: exhaustive and stepwise search. The latter is an option to shorten computation time. The present work aimed to test whether there is a significant difference between the accuracy of point forecasts generated by both algorithms. For that, 2,829 annual, quarterly and monthly series of the M3 Competition were selected, predictions with the two methods were executed and their accuracies were calculated. The conclusion was that for the annual and quarterly series, the methods had on average the same degree of precision and for the monthly series, the exhaustive procedure was on average slightly better.
Key words: Univariate Time Series; Forecast Methods; Comparison of Methods; Forecast Errors.

1. Introdução

O acelerado desenvolvimento tecnológico das últimas décadas tem feito com que atualmente muitas organizações gerem e coletem diariamente grandes quantidades de dados. Como consequência, é cada vez maior a necessidade do desenvolvimento de ferramentas

computacionais que automatizem a organização, a sumarização e a análise desses dados.

Uma característica particular desse tipo de dados é que eles são ordenados no tempo, formando a chamada série temporal (MORETTIN; TOLOI, 2006). Alguns exemplos são os preços das ações minuto a minuto, a temperatura hora após hora, a chegada de pacientes em uma clínica médica diariamente, a produção semanal de determinado produto, a taxa de desemprego mês a mês em determinada região, o volume anual de importações de um país, dentre outros.

Segundo Morettin e Toloi (2006), os principais objetivos da análise de séries temporais são descrever o comportamento da série, investigar seu mecanismo gerador, procurar periodicidades nos dados e fazer previsões de seus valores futuros, sendo este último um dos mais utilizados, principalmente em se tratando de séries econômicas e financeiras. Para Hyndman e Athanasopoulos (2021), os dois principais métodos estatísticos de previsão de séries temporais são a suavização exponencial e o modelo autorregressivo integrado de médias móveis (ARIMA – *autoregressive integrated moving average*).

De acordo com Morettin e Toloi (2006), a suavização exponencial é uma classe de métodos cujo propósito é separar um padrão de qualquer outro ruído que possa estar contido nos dados e, então, usar esse padrão para prever os valores futuros da série. Os métodos de suavização exponencial decompõem a série temporal em um ou mais componentes e os modelam por meio de relações recursivas. Previsões feitas pelos métodos de suavização exponencial são médias ponderadas de valores passados, com os pesos decaindo exponencialmente, à medida que as observações se tornam mais antigas. Os métodos mais conhecidos e utilizados de suavização exponencial são a suavização exponencial simples, a suavização exponencial de Holt e a suavização exponencial de Holt-Winters.

Segundo Hyndman e Athanasopoulos (2021), ao invés de modelar os componentes da série, os modelos ARIMA exploram outra característica distintiva das mesmas: o fato de dados próximos terem maior relação entre si do que dados separados por grandes intervalos de tempo. Esta característica é resumida matematicamente por uma grandeza denominada autocorrelação serial e o modelo ARIMA utiliza as autocorrelações em diversas defasagens para prever os valores futuros da série.

Diferentemente da suavização exponencial, que possui algumas poucas formulações alternativas, um ponto negativo na adoção prática do ARIMA sempre foi o grau de complexidade teórico necessário para se especificar o modelo mais adequado. Felizmente, com o atual aumento no poder computacional, aliado à crescente disponibilidade de software estatístico dedicado à análise de séries temporais, este panorama está mudando. Já é possível encontrar diversas soluções completamente automatizadas, bastando ao usuário fornecer a série temporal, que o programa sozinho seleciona o modelo ARIMA mais adequado, estima os seus parâmetros e executa as previsões.

Dentro desta categoria de soluções automatizadas, destaca-se o pacote *forecast* de Hyndman e Khandakar (2008), escrito na linguagem de programação R (R CORE TEAM, 2021). Pacotes são coleções de funções, dados e códigos compilados que ampliam as capacidades originais do software R. Normalmente são programados por estatísticos computacionais e não raro são o estado da arte em determinada área da Estatística. Como os pacotes e o próprio software são de código aberto, eles estão disponíveis ao escrutínio da comunidade científica com relação a sua correção, eficácia, etc.

O pacote *forecast* possui uma função específica para o ajuste automático de modelos ARIMA denominada "*auto.arima*". Por padrão, a função utiliza um método de força bruta para a seleção de modelos. Dentro de certos limites, uma busca exaustiva é feita, considerando-se todas as combinações possíveis de termos autoregressivos e de médias móveis e o melhor modelo é

aquele que minimiza certa função perda. Para dados sazonais, a quantidade de modelos alternativos a ser considerada pode facilmente chegar à casa das centenas ou até dos milhares, dependendo da quantidade de termos incluídos. Quando o objetivo é se analisar uma grande quantidade de séries simultaneamente, tal método não é prático. Dessa forma, Hyndman e Khandakar (2008) criaram um algoritmo de busca em etapas (*stepwise*), que diminui sobremaneira a quantidade de combinações a serem testadas.

Porém, nem no artigo original ou nas documentações posteriores do pacote são apresentadas comparações do desempenho do procedimento *stepwise* versus o de busca exaustiva. Assim, o presente trabalho tem como objetivo testar se há diferença significativa entre as precisões das previsões pontuais geradas por ambos os métodos, exaustivo e *stepwise*, implementados no pacote *forecast* quando aplicados a séries temporais anuais, trimestrais e mensais. Especificamente, para cada tipo de série (anual, trimestral e mensal), estaremos testando as seguintes hipóteses estatísticas:

H_0: Os dois métodos (exaustivo e *stepwise*) possuem em média a mesma precisão
H_1: Os dois métodos não possuem em média a mesma precisão

Para se atingir tais objetivos, este artigo está dividido em cinco seções, incluindo esta introdução. A seguir é apresentada a revisão da literatura, com os conceitos relacionados aos modelos ARIMA e avaliação da precisão das previsões. Posteriormente são abordados os aspectos metodológicos, os resultados encontrados e as considerações finais do estudo.

2. Revisão da Literatura

Nesta seção são abordados os métodos quantitativos de previsão utilizados neste trabalho e como se medir a precisão das previsões executadas pelos mesmos. Segundo Hyndman e Athanasopoulos (2021), métodos quantitativos são ideais quando duas condições são satisfeitas: informações numéricas a respeito do passado estão disponíveis; é razoável supor que alguns aspectos dos padrões passados continuarão no futuro.

2.1. Modelo ARIMA

Os modelos ARIMA são divididos em sazonais e não sazonais, de acordo com a presença ou ausência de sazonalidade na série, respectivamente. Como o ARIMA sazonal é uma extensão do não sazonal, este último será inicialmente abordado.

De acordo com Morettin e Toloi (2006), o modelo ARIMA não sazonal explora o fato de dados próximos terem maior relação entre si do que dados separados por grandes intervalos de tempo. Esta característica é resumida matematicamente por uma grandeza que varia entre ±1 denominada autocorrelação serial, que seria o grau de associação linear entre valores defasados da série. A autocorrelação na defasagem 1 seria a associação entre y_t e y_{t-1}, na defasagem 2 entre y_t e y_{t-2} e assim sucessivamente. A seguir são explicados os componentes do modelo ARIMA não sazonal: os termos autorregressivos (AR), os termos de médias móveis (MA) e o nível de integração da série (I).

Em um modelo autorregressivo, a variável de interesse é escrita como uma combinação linear de seus valores passados. Assim, um modelo autorregressivo de ordem p, representado por AR(p), é descrito pela equação (1):

$$y_t = c + \phi_1 \cdot y_{t-1} + \phi_2 \cdot y_{t-2} + \cdots + \phi_p \cdot y_{t-p} + \varepsilon_t \quad (1)$$

Onde c é o intercepto e ε_t uma sequência de variáveis aleatórias independentes igualmente distribuídas denominada erro ou resíduo (HYNDMAN; ATHANASOPOULOS, 2021). A

equação (1) é semelhante a uma regressão múltipla, exceto que as variáveis previsoras são valores defasados de y_t.

Ao invés de utilizar os valores passados da variável y_t, outro tipo de modelo utiliza uma média ponderada de distúrbios aleatórios passados e presentes, como mostra a equação (2):

$$y_t = \mu + \varepsilon_t + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \cdots + \theta_q \cdot \varepsilon_{t-q} \quad (2)$$

Onde μ é a média do processo gerador da série temporal. O modelo descrito pela equação (2) é denominado de modelo de médias móveis de ordem q e é representado por MA(q). Os valores de y_t podem ser interpretados como uma média móvel ponderada do erro presente e dos seus últimos q valores.

A combinação do modelo autorregressivo AR(p) com o de médias móveis MA(q) resulta no modelo denominado ARMA(p,q). Para muitas séries encontradas na prática, para a obtenção de um modelo parcimonioso, a inclusão de termos autorregressivos e de médias móveis é uma solução adequada. Assim, os modelos AR(p) e MA(q) são casos especiais do modelo ARMA(p,q) tomando-se $q = 0$ e $p = 0$, respectivamente.

Os 3 modelos apresentados anteriormente são aplicáveis a séries que possuem média constante (estacionárias na média). Séries com tendência estocástica podem ser transformadas em séries estacionárias na média tomando-se diferenças entre seus valores sucessivos. Se após aplicarmos a 1ª diferença, a série estabilizar sua média, ela é dita integrada de ordem 1. Caso isso não ocorra, aplica-se a 2ª diferença e assim sucessivamente. Uma série integrada de ordem d é aquela que necessitou de d diferenças para tornar-se estacionária na média. Se uma série integrada de ordem d puder ser modelada por um processo ARMA(p,q), teremos um modelo ARIMA(p,d,q) não sazonal, cuja equação é dada por (3):

$$\Delta^d y_t = c + \phi_1 \cdot \Delta^d y_{t-1} + \cdots + \phi_p \cdot \Delta^d y_{t-p} + \varepsilon_t + \theta_1 \cdot \varepsilon_{t-1} + \cdots + \theta_q \cdot \varepsilon_{t-q} \quad (3)$$

Na equação (3), y é a variável de interesse, c é o intercepto, Δ^d é a quantidade de diferenciações (se não forem necessárias, são excluídas da equação), ε é uma sequência de variáveis aleatórias independentes igualmente distribuídas e os termos ϕ e θ são parâmetros a serem estimados a partir dos dados. O intercepto c é dado pela equação (4):

$$c = \left(1 - \phi_1 - \cdots - \phi_p\right) \cdot \mu \quad (4)$$

O ARIMA sazonal modela adicionalmente as autocorrelações presentes nas defasagens múltiplas do período sazonal. Assim, por exemplo, se em uma série mensal os meses de um ano tiverem associação com o mesmo mês do ano anterior, haverá uma autocorrelação na defasagem 12. Se a associação se estender para dois anos no passado, haverá uma autocorrelação na defasagem 24 e assim sucessivamente. Um modelo ARIMA sazonal é formado adicionando-se à equação (3) termos autorregressivos, de médias móveis e diferenciações sazonais, sendo representados por: ARIMA(p,d,q)(P,D,Q)$_m$, onde m é o período sazonal. Assim, o modelo sazonal tem a mais que o não sazonal: D diferenças sazonais, para eliminar a sazonalidade estocástica e estabilizar a média; P termos autorregressivos sazonais (y_{t-m}, y_{t-2m}, …., $y_{t-P.m}$); e Q termos de médias móveis sazonais (ε_{t-m}, ε_{t-2m}, …, $\varepsilon_{t-Q.m}$).

De acordo com Hyndman e Athanasopoulos (2021), o procedimento geral para ajustar modelos ARIMA sazonais aos dados é o seguinte: (a) transforme os dados para estabilizar a variância, caso necessário; (b) se a série for não estacionária na média, tome diferenças (d e D) até que isso ocorra; (c) determine p, q, P e Q; (d) ajuste o melhor modelo do passo anterior e teste se

os resíduos são ruído branco (sequência de variáveis aleatórias independentes e igualmente distribuídas); (e) implemente o modelo para executar as previsões. No caso de modelos não sazonais, basta excluir os termos sazonais das etapas anteriores.

Segundo Hyndman e Athanasopoulos (2021), o passo (b) é feito por meio dos chamados testes de raízes unitárias, disponíveis em software estatístico. Os passos (a), (d) e (e) também estão disponíveis em programas voltados à análise de séries temporais.

Tradicionalmente o passo (c) é feito por meio da análise do padrão de autocorrelação dos dados. Isto se deve ao fato de que cada modelo ARIMA exibe um comportamento particular teórico para cada combinação de p, q, P e Q. Na prática, entretanto, tal método não é simples. Primeiramente por ser subjetivo e dependente da experiência do analista. Além disso, o padrão das autocorrelações amostrais pode divergir bastante do padrão teórico devido ao erro amostral. Para se contornar tal limitação, foram propostos métodos automáticos de seleção de modelos, como será visto a seguir.

2.2. Métodos Automáticos de Modelagem ARIMA

Para se evitar a subjetividade e, principalmente, o alto grau de complexidade teórica necessário para se especificar um modelo ARIMA (passo (c)), Morettin e Toloi (2006) explicam que na prática é utilizado um método baseado em uma função penalizadora. A ideia básica é determinar a combinação de p, q, P e Q que minimiza uma função que indica a falta de ajuste do modelo. Utilizando-se uma busca exaustiva, ajustam-se todos os modelos possíveis com p, q, P e Q variando entre, por exemplo, 0 e 5 e verifica-se qual combinação minimiza a função penalizadora. As funções mais utilizadas são o critério de informação de Akaike corrigido (AICc) e o critério de informação bayesiano (BIC).

Porém, quando o objetivo é se analisar uma grande quantidade de séries simultaneamente, tal método não é prático. Hyndman e Khandakar (2008) criaram um algoritmo de busca em etapas (*stepwise*), que diminui sobremaneira a quantidade de combinações a serem testadas. O procedimento básico é o seguinte:

1. Inicia-se o processo ajustando-se os quatro modelos abaixo e seleciona-se aquele cujo AICc ou BIC seja mínimo (denominado de modelo corrente). Se $d + D \leq 1$, os modelos são ajustados com intercepto ($c \neq 0$), caso contrário, não ($c = 0$):
    - ARIMA(2, d, 2) se $m = 1$, ou ARIMA(2, d, 2)(1, D, 1)$_m$ se $m > 1$;
    - ARIMA(0, d, 0) se $m = 1$, ou ARIMA(0, d, 0)(0, D, 0)$_m$ se $m > 1$;
    - ARIMA(1, d, 0) se $m = 1$, ou ARIMA(1, d, 0)(1, D, 0)$_m$ se $m > 1$;
    - ARIMA(0, d, 1) se $m = 1$, ou ARIMA(0, d, 1)(0, D, 1)$_m$ se $m > 1$.
2. São consideradas até treze variações no modelo corrente. Caso o novo modelo tenha menor AICc ou BIC ele se torna o modelo corrente e o processo é repetido, sendo finalizado quando não for possível achar um novo modelo com menor AICc ou BIC:
    - Um dos parâmetros p, q, P ou Q varia ±1 a partir do modelo corrente;
    - Ambos os parâmetros p e q variam ±1 a partir do modelo corrente;
    - Ambos os parâmetros P e Q variam ±1 a partir do modelo corrente;
    - O intercepto c é incluído, se no modelo corrente $c = 0$, ou excluído caso contrário.

Além disso, o algoritmo tem uma série de restrições, de forma a evitar problemas de convergência e para garantir o retorno de um modelo final válido.

2.3. Avaliação da Precisão das Previsões

Para Hyndman e Athanasopoulos (2021), é importante avaliar a precisão das previsões utilizando previsões genuínas. Dessa forma, os resíduos do modelo não são uma indicação confiável dos erros de previsão. A precisão somente pode ser determinada considerando-se o desempenho do modelo em dados novos e que não foram utilizados em seu ajuste.

Uma solução é separar os dados disponíveis em duas partes: os dados de treino e os dados de teste. O primeiro é usado para estimar os parâmetros do modelo, enquanto o segundo é utilizado para se avaliar a precisão das previsões. Como os dados de teste não foram usados no ajuste do modelo, eles são um indicador da qualidade do modelo de previsão em dados novos.

O erro de previsão é a diferença entre o valor observado e o previsto. Segundo Hyndman e Athanasopoulos (2021), a precisão das previsões pode ser avaliada de três maneiras: por meio dos erros dependentes de escala, dos erros porcentuais ou dos erros escalonados.

Os erros dependentes de escala estão na mesma unidade de medidas da série original. Muito embora sejam mais simples e intuitivos, a desvantagem é que não se pode comparar erros advindos de séries que estejam em unidades diferentes. Os erros mais comuns deste tipo são o erro absoluto médio (EAM) e a raiz quadrada do erro quadrático médio (REQM).

Os erros percentuais não possuem a desvantagem anterior, pois são livres de unidade. Entretanto, são indefinidos quando $y_t = 0$ e só são aplicáveis em séries medidas em uma escala de razão. Os principais são o erro percentual absoluto médio e o erro percentual absoluto médio simétrico.

Uma alternativa aos erros percentuais, por não possuírem as desvantagens anteriores, são os erros escalonados. Dentre os erros escalonados, Hyndman e Athanasopoulos (2021) propõem o uso do erro escalonado absoluto médio (MASE – *mean absolute scaled error*), definido como a média aritmética dos valores absolutos dos q_j dados pela equação (5).

$$q_j = \frac{e_j}{\frac{\sum_{t=m+1}^{T}|y_t - y_{t-m}|}{T-m}} \tag{5}$$

Onde $m = 1$ para série anuais, $m = 4$ para séries trimestrais e $m = 12$ para séries mensais. De acordo com Pellegrini (2012), o MASE é robusto a *outliers*, tem uma escala significativa, é amplamente aplicável e não é afetado por problemas de indeterminação, exceto se a série tiver variância zero (todas as observações forem iguais). Neste artigo, este erro será utilizado para se comparar a precisão dos dois métodos de previsão.

3. Metodologia

Inicialmente, foi montado o banco de dados com as séries temporais a serem analisadas. As séries foram retiradas da M3 *Competion* (MAKRIDAKIS; HIBON, 2000). Esta foi uma competição onde *experts* foram convidados a fornecer previsões para 3003 séries temporais reais utilizando métodos variados. As séries da M3 *Competition* foram escolhidas por representarem uma variedade de aplicações de séries temporais (havia séries financeiras, micro e macro econômicas, demográficas e industriais). Para as análises, retivemos apenas as séries com periodicidade anual, trimestral e mensal, como indica a tabela 1.

De posse das 2.829 séries, cada uma foi dividida em duas partes: dados de treino e de teste. Seguindo o mesmo procedimento utilizado na competição, os dados de teste foram: os últimos 6 valores para as séries anuais, os últimos 8 valores para as séries trimestrais e os últimos 18 para as séries mensais. Por consequência, esses foram os horizontes de previsão utilizados.

Tabela 1: Classificação das séries temporais retiradas da M3 *Competition*

Período	Tipo de Série						Total
	Micro	Indústria	Macro	Financeira	Demográfica	Outras	
Anual	146	102	83	58	245	11	645
Trimestral	204	83	336	76	57	0	756
Mensal	474	334	312	145	111	52	1428
Total	824	519	731	279	413	63	**2.829**

Fonte: Adaptado de Makridakis e Hibon (2000) p. 454.

A cada uma das 2.829 séries de treino foram ajustados dois tipos de modelos, utilizando a função "*auto.arima*" do pacote *forecast*: ARIMA com especificação por meio de busca exaustiva e por meio de busca *stepwise*. Especificamente, a função executa automaticamente os passos de (b) a (e) descritos na seção 2.1, exceto a análise residual, que é efetuada por outras funções do pacote. O passo (a) – transformação estabilizadora da variância – não foi executada, pois há evidências que a mesma não afeta a precisão das previsões pontuais do modelo ARIMA (SILVA; RUY, 2019). A análise residual também foi excluída, primeiro por ser impraticável se analisar individualmente cada uma das séries e, segundo, pelo fato de que a qualidade do modelo será aferida pela precisão de suas previsões e não por meio de seus resíduos.

Uma vez ajustados ambos os modelos aos dados de treino, o pacote faz automaticamente as previsões. Por fim, dadas as previsões e os dados de teste, o pacote também calcula de forma automática os erros de previsão, neste caso, o MASE. Todos os cálculos foram efetuados no software R versão 4.1.1. e no pacote *forecast* versão 8.15. O Apêndice 1 mostra os comandos utilizados e necessários para se reproduzir as análises aqui apresentadas.

4. Resultados

A tabela 2 exibe as estatísticas descritivas dos erros MASE de cada método de previsão aplicados às séries, bem como seus tempos médios de computação em segundos. O tempo médio foi calculado dividindo-se o tempo total de computação pela quantidade de séries. A figura 1 mostra os diagramas de caixa dos logaritmos naturais dos erros MASE dos métodos. Como a distribuição dos erros é assimétrica à direita e sua amplitude é muito diferente em função da periodicidade das séries, utilizou-se a transformação logarítmica para se obter simetria e diminuir a amplitude, facilitando assim a visualização conjunta dos erros.

Pela inspeção da tabela 2, pode-se verificar que para dada periodicidade, o procedimento *stepwise* tem um tempo de computação médio menor. Além disso, a economia em tempo computacional (diferença entre os tempos dos métodos) aumenta com o aumento da periodicidade das séries.

Na figura 1 e na tabela 2 também se observa que quanto maior a periodicidade, menores os erros médios e medianos e menores as variabilidades em torno desses valores. Era esperado que as séries anuais apresentassem maiores erros, primeiramente por terem menos termos que as demais, impactando negativamente a precisão das estimativas dos modelos. Em segundo lugar, de acordo com Makridakis, Spiliotis e Assimakopoulos (2018), séries sazonais são mais fáceis de serem previstas, pois o efeito sazonal domina a flutuação da série, ao passo que em séries não sazonais, o principal fator são as variações irregulares (aleatórias). As séries mensais obtiveram erros menores que as trimestrais. Uma possível explicação seria o tamanho das mesmas. Em média, as séries mensais analisadas continham 117 termos, as trimestrais continham 49 termos e as anuais, 28. Na prática, as séries mensais são normalmente maiores,

seguidas pelas trimestrais, por conta da frequência maior de registros no tempo.

Tabela 2: Comparação entre os métodos de previsão

Período	Método	MASE						Tempo Médio (segundos)
		Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo	
Anual	Exaustivo	0,041	0,997	1,898	2,999	3,524	37,012	0,41
	Stepwise	0,041	0,994	1,894	2,971	3,496	37,012	0,04
Trimestral	Exaustivo	0,054	0,503	0,835	1,189	1,556	10,554	2,81
	Stepwise	0,033	0,500	0,822	1,185	1,531	10,554	0,09
Mensal	Exaustivo	0,061	0,497	0,710	0,854	1,024	5,407	20,34
	Stepwise	0,047	0,498	0,718	0,875	1,043	7,998	1,16

Fonte: Dados da pesquisa.

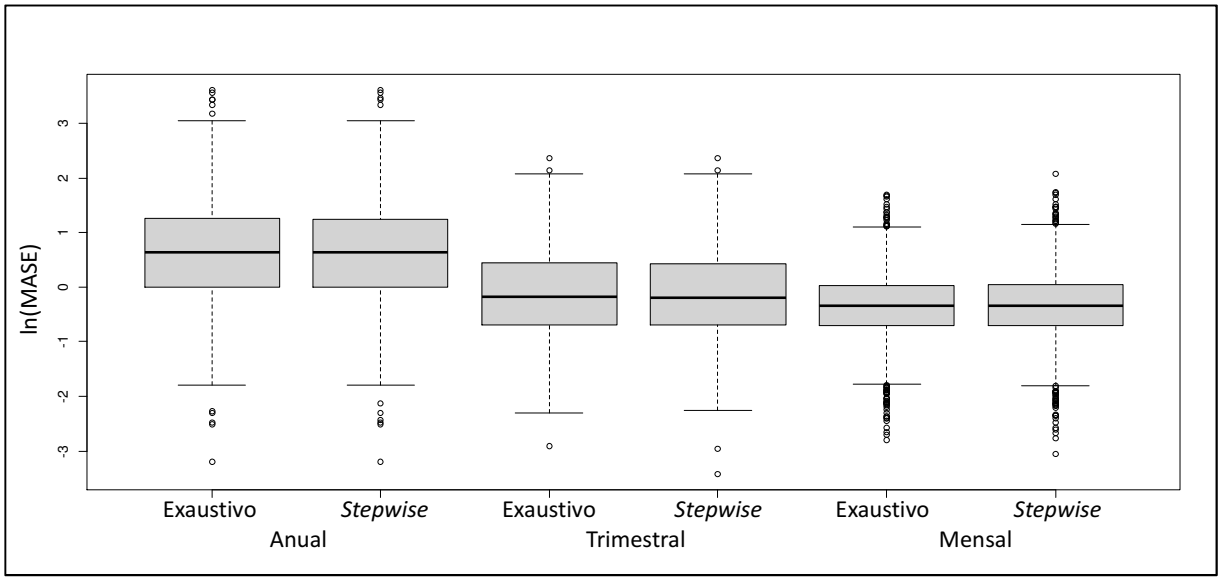

Figura 1 – Logaritmo natural dos erros MASE em função do método e do período das séries. *Fonte*: Dados da pesquisa.

Para se testar a hipótese de que os dois métodos têm em média a mesma precisão *versus* que as precisões médias são diferentes, deve-se analisar as diferenças entre os MASE obtidos para cada série. Isto se deve ao fato de se ter duas amostras dependentes (emparelhadas). Como cada série foi analisada por dois métodos diferentes, os dados foram coletados aos pares (MONTGOMERY; RUNGER, 2012).

Segundo Montgomery e Runger (2012), o teste paramétrico adequado é o teste *t* emparelhado. Entretanto, este teste tem como suposição que as diferenças seguem uma distribuição normal, o que não ocorre com os dados da presente pesquisa. Alternativas seriam utilizar um teste não paramétrico, que demanda simetria para ser válido para comparação de médias, ou a técnica denominada *bootstrap*.

De acordo com Silva Filho (2010), o *bootstrap* é particularmente atrativo porque exige menos suposições e geralmente fornece respostas mais precisas do que os métodos tradicionais. É uma técnica de reamostragem que permite quantificar a incerteza, gerando o cálculo de erros-padrão, intervalos de confiança e teste de significância. Neste trabalho, foi utilizado o pacote "*wBoot*" de Weiss (2016) para efetuar o teste de comparação de médias de amostras emparelhadas por

meio do método de *bootstrap* não paramétrico com correção de vício acelerado (BCa *bootstrap*). A seguir, serão mostrados os resultados da aplicação do *bootstrap* nos erros MASE das séries anuais, trimestrais e mensais.

Quando executado nas 645 séries anuais, o procedimento exaustivo gerou um MASE menor em 79 delas, o procedimento *stewpise* gerou um erro menor em 81 séries e houve empate nas outras 485, isto é, ambos os métodos chegaram ao mesmo modelo final e, portanto, às mesmas previsões e erros. Como séries anuais não possuem componente sazonal, a quantidade de modelos diferentes não é grande, logo era esperado que ambos os métodos concordassem na maioria das vezes. Aplicando-se o teste de comparação de médias de amostras emparelhadas via BCa *bootstrap* aos dados, obtém-se um valor-p de 0,261, superior ao maior nível de significância tradicionalmente utilizado de 10%. Portanto, falha-se em rejeitar a hipótese nula H_0, ou seja, não há evidências que os dois métodos não possuam em média a mesma precisão.

Nas 756 séries trimestrais, o procedimento exaustivo obteve melhor desempenho (menor erro) em 239 delas, foi pior em 246 e houve empate nas 271 séries restantes. O valor-p do teste de hipótese foi de 0,707, também evidenciando que não se pode rejeitar a hipótese de não diferença média entre os métodos.

Por fim, para as 1428 séries mensais, o procedimento exaustivo foi melhor em 613, foi pior em 530 e houve empate em 285 ocasiões. O valor-p do teste de hipótese foi inferior a 0,001, ou seja, o valor-p é inferior ao menor nível de significância usual de 1%. Portanto, rejeita-se H_0 e conclui-se que há evidências que os dois métodos não possuem em média a mesma precisão.

Dado um resultado significativo, para se determinar qual método gera menores erros médios, basta comparar os valores médios da tabela 2, uma vez que a média das diferenças é igual à diferença das médias. Da tabela verifica-se que em média o procedimento *stepwise* gera erros de previsão maiores em 0,021 unidades (0,875–0,854) em termos absolutos.

Porém, Montgomery e Runger (2012) chamam a atenção na diferença entre significância estatística e significância prática. Segundo os autores, é necessário ser cuidadoso ao interpretar os resultados dos testes para amostras grandes, uma vez que qualquer desvio do valor usado em H_0 será detectado, mesmo quando a diferença for de pouca significância prática.

Particularmente, é difícil determinar se uma diferença média absoluta de 0,021 unidades do erro MASE é grande ou não. É mais informativo se determinar em termos percentuais o quanto em média o erro do procedimento *stepwise* é maior. Para isso, dividiu-se cada uma das diferenças dos erros entre os métodos pelo MASE do respectivo procedimento exaustivo e calculou-se a média aritmética, obtendo-se o valor de 0,047. Ou seja, em média, o procedimento *stepwise* gera erros 4,7% maiores em termos relativos.

Uma medida de significância prática comumente utilizada é o denominado *d* de Cohen ou tamanho do efeito. Lakens (2013) classifica essa medida como um dos principais resultados a serem reportados em estudos empíricos, pois ela permite o reporte do efeito observado em uma métrica padronizada, que pode ser entendida a despeito da escala usada no estudo.

No caso de amostras emparelhadas, o tamanho do efeito é calculado dividindo-se a média das diferenças pelo desvio-padrão das diferenças. Quando aplicado aos dados da presente pesquisa, obtém-se aproximadamente 0,10. Ou seja, a diferença das médias (efeito) é 10% de seu desvio-padrão. Segundo Lakens (2013), é usual utilizar as seguintes diretrizes para interpretar os resultados: efeito pequeno ($d = 0,2$), efeito médio ($d = 0,5$) e efeito grande ($d = 0,8$). Portanto, o efeito observado no presente estudo pode ser considerado muito pequeno. Porém, o próprio autor alerta que esses valores são arbitrários e que não devem ser interpretados rigidamente, pois efeitos pequenos podem ter consequências drásticas, como por exemplo, uma intervenção que evita mortes, não importa o quão pequena seja, sempre será válida.

5. Considerações Finais

O presente artigo teve como objetivo testar se havia diferença significativa entre as precisões das previsões pontuais geradas pelos métodos ARIMA automáticos de busca exaustiva e *stepwise* implementados no pacote *forecast* quando aplicados a séries temporais anuais, trimestrais e mensais. Para tanto 2.829 séries com essas periodicidades foram analisadas por ambos os métodos e os erros MASE de cada foram comparados utilizando-se *bootstrap* não paramétrico com correção de vício acelerado. Para as séries anuais e trimestrais, o procedimento *stepwise* se mostrou uma alternativa viável à busca exaustiva, pois além de gerar em média o mesmo nível erro, consumiu bem menos esforço computacional.

Porém, para as séries mensais, onde o ganho de tempo é maior, o procedimento *stepwise* foi ligeiramente inferior, gerando erros em média 4,7% maiores ou um efeito padronizado de 0,10. Muito embora essa diferença possa ser considerada muito pequena, ela existe e deve ser levada em consideração na escolha entre os métodos. Ou seja, se ela será relevante ou não vai depender do custo do erro, da quantidade de séries a serem analisadas e do tempo disponível para computação. Por exemplo, se o custo do erro for alto, o procedimento exaustivo pode ser mais adequado. Por outro lado, se uma decisão baseada nas previsões for urgente, o *stepwise* pode ser melhor. A título de exemplo, para as 1428 séries mensais analisadas neste trabalho, o tempo de processamento do procedimento *stepwise* foi de 28 minutos e o do exaustivo excedeu 8 horas. Pode ser que em certas circunstâncias não haja todo esse tempo disponível para a tomada de decisão.

Adicionalmente, mesmo os dois métodos tendo em média a mesma precisão para as séries anuais e trimestrais e o exaustivo sendo ligeiramemte melhor em séries mensais, houve casos em que um dos métodos teve um desempenho muito superior e se sobressaiu. Mais interessante ainda foram as situações onde o método *stepwise* foi melhor, significando que a busca exaustiva gerou sobreajuste aos dados (sobreajuste é quando o modelo tem bom desempenho nos dados de treino, mas generaliza mal nos dados de teste). Assim, trabalhos futuros poderiam tentar verifcar quais seriam as características distintivas dessas séries que levaram a tal resultado tão díspar e se há algum tipo de padrão. A diferença entre a precisão dos métodos pode depender de fatores característicos das séries tais como heteroscedasticidade, não linearidade e grau de aleatoriedade (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018).

Outro resultado deste estudo que merece ser aprofundado em trabalhos futuros é relativo ao fato de séries sazonais serem mais fáceis de serem previstas. Wang, Smith e Hyndman (2006) criaram uma medida padronizada entre 0 e 1 que indica a força do componente sazonal. Seria interessante verificar se esta medida, juntamente com outros atributos sazonais, ajudam a explicar o grau de previsibilidade de uma série sazonal.

Uma limitação deste trabalho e que enseja trabalhos futuros é relativa ao horizonte das previsões e o tipo de erro utilizado (MASE). Neste artigo, as comparações entre as precisões dos métodos foram baseadas no MASE de previsões 6, 8 e 18 períodos a frente. Seria interessante verificar se os resultados aqui obtidos se mantém variando o tipo de erro e os períodos de tempo das previsões, principalmente para previsões mais curtas.

Uma segunda limitação do trabalho e que também precisa de mais estudos é relativa ao uso do critério AICc para a escolha do melhor modelo. Muito embora Hyndman e Athanasopoulos (2021) o recomendem, seria interessante verificar a sensibilidade dos resultados ao se utilizar o critério BIC. Como essa função penaliza mais que o AICc os modelos mais complexos, seria interessante verificar se a questão do sobreajuste observado com a busca exaustiva em algumas séries seria alterada. Além disso, nesse artigo foram utilizados todos os ajustes padrão na função "*auto.arima*", então os resultados também têm de ser interpretados neste contexto.

Referências

HYNDAMN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. 3. Ed. Melbourne: OTexts, 2021. Disponível em: https://otexts.org/fpp3/. Acesso em: 25 out. 2021.

HYNDMAN R. J.; KHANDAKAR, Y. Automatic Time Series Forecasting: the forecast package for R. **Journal of Statistical Software**, v. 27, n. 3, p. 1-22, 2008.

LAKENS, D. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: a practical primer for t-tests and ANOVAs. **Frontiers in Psychology**, v. 4, november, p. 1-12, 2013.

MAKRIDAKIS, S.; HIBON, M. The M3 Competition: results, conclusions and implications. **International Journal of Forecasting**, v. 16, n. 4, p. 451-476, 2000.

MAKRIDAKIS, S.; SPILIOTIS, E.; ASSIMAKOPOULOS, V. The M4 Competition: results, findings, conclusion and way forward. **International Journal of Forecasting**, v. 34, n. 4, p. 802-808, 2018.

MONTGOMERY, D. C.; RUNGER, G.C. **Estatística Aplicada e Probabilidade para Engenheiros**. 5. Ed. Rio de Janeiro: LTC, 2012.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de Séries Temporais**. 2. Ed. São Paulo: Blucher, 2006.

PELLEGRINI, T. R. **Uma Avaliação de Métodos de Previsão Aplicados à Grandes Quantidades de Séries Temporais Univariadas**. 2012. Dissertação (Mestrado em Estatística) – Centro de Ciências Exatas e Tecnológicas, Universidade Federal de São Carlos, São Carlos, 2012.

R CORE TEAM (2021). **R**: a language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. Disponível em http:www.R-project.org, 2021.

SILVA FILHO, A. S. Inferência em Amostras Pequenas: métodos bootstrap. **Revista de Ciências Exatas e Tecnologia**, v. 5, n. 5, p. 115-126, 2010.

SILVA, R. G. F. L.; RUY, M. Avaliação da Precisão dos Modelos ARIMA com e sem Transformação Estabilizadora da Variância na Previsão de Séries Temporais. *In:* SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, XXVI, 2019, Bauru, SP. **Anais [...].** Bauru: Unesp, 2019. Disponível em: https://simpep.feb.unesp.br/anais_simpep.php?e=14. Acesso em: 25 out. 2021.

WANG, X.; SMITH, K. A.; HYNDMAN, R. J. Characteristic-Based Clustering for Time Series Data. **Data Mining and Knowledge Discovery**, v. 13, n.3, p.335-364, 2006.

WEISS, N. A. **wBoot:** bootstrap methods. R package version 1.0.3. Disponível em https://CRAN.R-project.org/package=wBoot, 2016.

Apêndice 1 – Comandos utilizados no software R

```
### Preparação do R
# instala os pacotes (única vez)
install.packages("forecast", dependencies = TRUE)
install.packages("wBoot", dependencies = TRUE)
install.packages("Mcomp")
# carrega os pacotes (todo início de sessão)
library(forecast)
library(wBoot)
library(Mcomp)
### Montagem da bases de dados: anual, trimestral e mensal
# Anual
yearly = subset(M3, 1)
N = 645
anual = vector(mode = "list", length = N)
for(i in 1: N){
anual[[i]] = ts(c(as.numeric(yearly[[i]]$x), as.numeric(yearly[[i]]$xx)),
start = 1, frequency = 1)}
remove(yearly)
# Trimestral
quart = subset(M3, 4)
N = 756
tri = vector(mode = "list", length = N)
for(i in 1: N){
tri[[i]] = ts(c(as.numeric(quart[[i]]$x), as.numeric(quart[[i]]$xx)), start
= 1, frequency = 4)}
remove(quart)
# Mensal
mont = subset(M3, 12)
N = 1428
mes = vector(mode = "list", length = N)
for(i in 1: N){
mes[[i]] = ts(c(as.numeric(mont[[i]]$x), as.numeric(mont[[i]]$xx)), start =
1, frequency = 12)}
remove(mont)
# Remove variáveis auxiliares e salva o resultado:
remove(i, N)
save("anual", "mes", "tri", file = "dados.RData")
#
#### Análise dos dados
## Função para calcular o MASE dos modelos ARIMA - busca exaustiva e
stepwise
## Horizonte de Previsão: 6 para dados anuais, 8 para dados trimestrais, 18
para dados mensais
# Função para calcular o MASE - argumentos: série (x) e lógico - TRUE é o
procedimento stepwise e FALSE o exaustivo
mase = function(x, stp = TRUE){
n = length(x)
if (frequency(x) == 1) {
K = 6
} else if (frequency(x) == 4) {
K = 8
} else {
K = 18
}
treino = window(x, start = time(x)[1], end = time(x)[n-K])
teste = window(x, start = time(x)[n-K+1], end = time(x)[n])
mod = auto.arima(treino, stepwise = stp, approximation = stp)
prev = forecast(mod, h = K)
erro = accuracy(prev, teste)[2,6]
```

```
return(erro)}
#
anual_step = as.vector(as.numeric(lapply(anual, mase, TRUE)))
anual_exau = as.vector(as.numeric(lapply(anual, mase, FALSE)))
#
tri_step = as.vector(as.numeric(lapply(tri, mase, TRUE)))
tri_exau = as.vector(as.numeric(lapply(tri, mase, FALSE)))
#
mes_step = as.vector(as.numeric(lapply(mes, mase, TRUE)))
mes_exau = as.vector(as.numeric(lapply(mes, mase, FALSE)))
#
save("anual_step", "anual_exau", "tri_step", "tri_exau", "mes_step",
"mes_exau", "anual", "mes", "tri", file = "resultados.RData")
#
### Bootstrap
set.seed(269506); boot.paired.bca(x = anual_step, y = anual_exau,
alternative = c("two.sided"), conf.level = 0.99, R = 99999, null.hyp = 0)
set.seed(269506); boot.paired.bca(x = tri_step, y = tri_exau, alternative =
c("two.sided"), conf.level = 0.99, R = 99999, null.hyp = 0)
set.seed(269506); boot.paired.bca(x = mes_step, y = mes_exau, alternative =
c("two.sided"), conf.level = 0.99, R = 99999, null.hyp = 0)
```