O Uso do Google Trends para Melhorar a Previsão de Séries Temporais: O Caso da Taxa de Desocupação no Brasil

Discente: Paloma Fernandes Santos
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. Este artigo tem como objetivo verificar se uma regressão dinâmica com o uso do Google Trends melhora o desempenho das previsões para a série mensal taxa de desocupação das pessoas de 14 anos ou mais de idade no Brasil, relativamente os métodos de suavização exponencial e ARIMA puros, medido pelo erro absoluto médio e pela raiz quadrada do erro quadrático médio. A conclusão foi que o uso do índice do Google Trends com a palavra-chave escolhida não mostrou ganhos em sua adoção no quesito melhoria da precisão das previsões para a série em questão.
Palavras chave: Séries Temporais; Previsões; ARIMA; Suavização Exponencial; Regressão Dinâmica; Google Trends.

Using Google Trends to Improve Time Series Forecasting: the Case of Unemployment Rate in Brazil

Abstract:
Among the applications of Time Series Analysis, forecasting future values is one of the most used. In forecasts, the accuracy of the method used is one of the critical factors in its adoption. This article aims to verify whether a dynamic regression using Google Trends improves the performance of forecasts for the monthly unemployment rate series of people aged 14 years and over in Brazil, relative to exponential smoothing and ARIMA methods, measured by the mean absolute error and the root mean squared error. The conclusion was that the use of Google Trends index with the chosen keyword showed no gains in its adoption in terms of improving the accuracy of forecasts for the series in question..
Key words: Time Series; Forecasts; ARIMA; Exponential Smoothing; Dynamic Regression; Google Trends.

1. Introdução

Previsões são antecipações de um ou mais eventos futuros, podendo ser categorizadas como de curto, médio ou longo prazo, de acordo com o número de períodos à frente utilizados. Consistem em uma etapa muito importante no processo de tomada de decisão em diversas áreas, tais como em gerenciamento de operações, marketing, finanças, economia, controle industrial, demografia, dentre outros.

De acordo com Singer (1997) p.39: “por causa das coisas que não sabemos que não sabemos, o futuro é em grande parte imprevisível. Mas alguns desenvolvimentos podem ser antecipados, ou pelo menos imaginados com base no conhecimento existente”.

Segundo o Dicionário Aurélio, a palavra previsão é um substantivo feminino que significa ato ou feito de prever, antevisão, estudo ou exame feito com antecedência (FERREIRA, 1975).

Logo, é a estimação de uma ou mais variáveis em um momento futuro. Por meio da previsão é possível modificar comportamentos para um melhor desempenho e adequação ao que está por vir, diminuindo os riscos na tomada de decisão (HARRISON; STEVENS, 1976).

Em muitas situações, previsões envolvem o uso de séries temporais, que são sequencias numéricas ordenadas no tempo. Esses dados são denominados contínuos quando são registrados initerruptamente ao longo do tempo e discretos quando são coletados em intervalos específicos, normalmente regulares (MORETTIN; TOLOI, 2006).

Para Hyndman e Athanasopoulos (2018), uma serie temporal é composta de um ou mais elementos básicos: tendência, sazonalidade, variações cíclicas e variações irregulares. A tendência é um movimento de aumento ou decréscimo persistente e de longo prazo nos dados. Sazonalidade ocorre quando a série é afetada por fatores tais como a época do ano ou dia da semana e é um padrão que se repete com uma periodicidade conhecida e fixa. Um ciclo é um padrão que se repete com alguma regularidade, mas com uma periodicidade não fixa, normalmente longa e às vezes até desconhecida. Tradicionalmente o componente cíclico é agrupado à tendência e ambos são analisados conjuntamente. As variações irregulares são o elemento que confere à série seu comportamento estocástico e são modeladas como uma sequência de variáveis aleatórias com alguma distribuição de probabilidade.

De acordo com Morettin e Toloi (2006), outra característica distintiva das séries temporais é 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 (dados separados por 1 unidade de tempo) seria a associação entre y_t e y_{t-1}, na defasagem 2 entre y_t e y_{t-2} e assim sucessivamente.

Segundo Hyndman e Athanasopoulos (2018), os dois principais métodos estatísticos de previsão de séries temporais exploram tais características distintivas das mesmas. São eles: a suavização exponencial e o modelo autorregressivo integrado de médias móveis (ARIMA – *autoregressive integrated moving average*). A suavização exponencial é uma classe de métodos de previsão cujo propósito é distinguir um padrão de comportamento de qualquer outro ruído que possa estar contido nas observações e, então, usar esse padrão para prever valores futuros da série. Já o ARIMA, ao invés de modelar os componentes da série, utiliza a autocorrelação serial nas diversas defasagens para prever os valores futuros da série baseado em seus valores passados e presente.

Contrariamente à análise de regressão, que utiliza uma ou mais variáveis independentes para prever a variável dependente y_t, ambos os métodos apresentados anteriormente não utilizam informações externas, isto é, os valores futuros de y_t são uma projeção de seus próprios valores passados e presente. Porém, em muitas situações, informações externas à série estão disponíveis e podem melhorar a precisão das previsões. A união da regressão linear, simples ou múltipla, com o ARIMA gera o chamado modelo de regressão dinâmica. Em linhas gerais, é um modelo de regressão convencional, porém com uma adaptação fundamental – o termo de erro é um processo ARIMA ao invés de uma sequência de variáveis aleatórias independentes e igualmente distribuídas, como na regressão comum (HYNDMAN; ATHANASOPOULOS, 2018).

Atualmente, com a internet sendo um dos principais meios de acesso à informação e o Google o seu principal mecanismo de buscas, em tese abre-se a possibilidade de se utilizar os índices de busca do Google como uma informação externa para se melhorar as previsões de determinadas séries temporais (SHIKIDA et al., 2012). Segundo os autores, a ideia básica é que se um indivíduo busca na internet informações sobre modelos de automóvel, há uma chance de ele estar interessado em comprar o bem em um futuro próximo. Ou se ele procura informações

sobre seguro-desemprego, é possível que o indivíduo tenha sido demitido recentemente.

No caso do desemprego e outras estatísticas oficiais (p. ex., o PIB), seu anúncio é sempre feito com certa defasagem, isto é, o valor da estatística de determinado mês só é conhecido alguns meses depois. Dessa forma, caso o volume de buscas no Google seja relacionado com a variável de interesse, ele poderia ser utilizado para se prever, com custo mínimo, o valor da estatística de certo período enquanto o anúncio oficial não é divulgado. Isso é denominado na literatura de *nowcasting*, ou seja, a previsão do presente ou de um futuro muito próximo (CHOI; VARIAN, 2012).

Segundo Shikida et al (2012), o Google Trends é uma ferramenta do Google que fornece um índice adimensional, entre 0 e 100, que informa o volume relativo de pesquisas para uma dada palavra-chave num certo período de tempo, em uma região específica. Por meio dele, é possível saber a evolução das pesquisas que os usuários de internet fizeram no Google sobre determinado assunto ou palavra-chave.

O presente trabalho tem como objetivo geral verificar se o uso do Google Trends melhora a precisão das previsões da série mensal "taxa de desocupação das pessoas de 14 anos ou mais de idade", apurada pelo IBGE por meio da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) e com início em março de 2012.

O objetivo específico é verificar se a regressão dinâmica utilizando como variável regressora o índice do Google Trends tem um desempenho melhor que os métodos de suavização exponencial e ARIMA, medido pelo erro absoluto médio e pela raiz quadrada do erro quadrático médio. A ideia básica é que se a regressão dinâmica não gerar erros de previsão menores que os métodos tradicionais que utilizam apenas os valores da própria série temporal, não há justificativa prática para sua adoção.

Muito embora haja trabalhos com essa abordagem, eles têm algumas limitações. Choi e Varian (2012) utilizaram o Google Trends para prever as solicitações de seguro desemprego nos Estados Unidos. Porém, foram utilizados modelos ARIMA simplificados (especificamente um modelo autorregressivo de ordem 1), o que pode impactar os erros de previsão. Além disso, devido às muitas particularidades do Brasil (econômica, social, demográfica, etc.), seria interessante estudar o desempenho dos métodos em nosso contexto em particular.

Shikida et al (2012) estudaram o uso do Google Trends na previsão da taxa de desocupação brasileira usando modelos ARIMA sem a simplificação anterior. Entretanto, na especificação do modelo de regressão dinâmica, não se permitiu defasagem entre a variável regressora e a variável resposta. Isto é, o modelo apenas associava a taxa de desocupação de um mês com o índice do Google daquele mesmo mês. Muitas vezes é possível melhorar o desempenho de um modelo de regressão dinâmica utilizando-se regressores defasados no tempo. Ou seja, permite-se que y_t seja associado a x_t, x_{t-1}, x_{t-2}, etc.

Além disso, ambos os trabalhos apenas compararam a regressão dinâmica com seu respectivo ARIMA puro (sem o índice do Google). No presente artigo serão utilizados como *benchmarks* tanto o ARIMA quanto a suavização exponencial, pois esse é o padrão nos estudos de séries temporais (MAKRIDAKIS; HIBON, 2000; MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018).

Este artigo está dividido em cinco seções, além desta introdução. A seguir é apresentada a revisão da literatura, com os conceitos relacionados aos Modelos ARIMA e Regressão Dinâmica, Suavização Exponencial 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 (2018), métodos quantitativos são ideais quando informações numéricas a respeito do passado estão disponíveis e é razoável supor que alguns aspectos dos padrões passados continuarão no futuro.

2.1. Modelos ARIMA e Regressão Dinâmica

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.

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 denominada autocorrelação serial, que seria o grau de associação linear entre valores defasados da série. 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).

De acordo com Hyndman e Athanasopoulos (2018), em um modelo autorregressivo de ordem p, representado por AR(p), a variável de interesse (y_t) é escrita como uma combinação linear de seus valores passados (y_{t-1}, y_{t-2}, etc.). Outra opção para se modelar a autocorrelação serial dos dados é utilizar uma média ponderada dos erros aleatórios presente e passados na equação (ε_t, ε_{t-1}, etc.). Este modelo, é denominado de médias móveis de ordem q e é representado por MA(q). A combinação do modelo autorregressivo com o de médias móveis resulta no modelo ARMA(p,q). Para muitas séries encontradas na prática, a combinação de termos autorregressivos e de médias móveis geram modelos com menor número de parâmetros, relativamente aos AR(p) e MA(q) puros.

Os modelos anteriores são para séries estacionárias na média (i.e., com média constante). 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:

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

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 1 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 (2018), 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.

Hyndman e Athanasopoulos (2018) explicam que um modelo de regressão dinâmica consiste de um modelo de regressão linear simples ou múltipla onde a suposição de que os erros são independentes foi relaxada, ou seja, é permitido que haja autocorrelação entre os mesmos. Dessa forma, a regressão dinâmica é uma regressão linear com erros que seguem um processo ARIMA.

Assim, os objetivos e interpretação de uma regressão dinâmica são idênticos ao da regressão comum, ou seja, estimar ou prever a média populacional de uma variável métrica y, tendo por base valores conhecidos de um ou mais regressores x. Cada variável regressora é ponderada de forma que seu respectivo peso denota a associação daquela variável com y, controlando ou mantendo fixo o efeito das demais (HAIR et al, 2009).

Uma extensão interessante da regressão dinâmica relativamente à comum é a possibilidade da utilização de regressores defasados no tempo, tais como x_{t-1}, x_{t-2}, etc., permitindo que valores passados da variável regressora sejam associados a valores contemporâneos da variável a ser prevista (y_t).

2.2. Suavização Exponencial

De acordo com Hyndman e Athanasopoulos (2018), os métodos de suavização exponencial decompõem e modelam cada componente da série por meio de relações recursivas. Os métodos mais conhecidos de suavização exponencial são a suavização exponencial simples, a suavização exponencial de Holt e a suavização exponencial de Holt-Winters.

A suavização exponencial simples é utilizada em séries localmente constantes, isto é, sem tendência e sem sazonalidade, e consiste de uma única equação, onde o próximo valor da série é uma média ponderada do valor atual e de todos os valores passados, com os pesos decaindo em progressão geométrica. Ou seja, quanto mais antigo o dado, menor a sua contribuição para as previsões futuras. Os pesos utilizados podem ser determinados de uma maneira *ad hoc* ou estimados dos dados históricos.

A suavização exponencial de Holt aplica-se a séries que possuem tendência apenas. Neste caso, tem-se duas equações recursivas, uma para o nível local da série e outra para a sua taxa de aumento/decréscimo por unidade de tempo. Por fim, a suavização exponencial de Holt-Winters é utilizada em séries que apresentam sazonalidade e possui 3 relações recursivas, uma para cada componente. Há dois tipos de suavização de Holt-Winters: a com sazonalidade aditiva, usada em séries com amplitude constante e a com sazonalidade multiplicativa, para séries com amplitude variável.

Segundo Hyndman e Athanasopoulos (2018), os métodos de suavização exponencial originais não têm base probabilística, sendo apenas relações recursivas. Assim, eles não geram intervalos de predição, somente estimativas pontuais. Atualmente, a suavização exponencial foi expandida e ganhou uma base probabilística denominada modelo de espaço de estados. Hyndman et al. (2008) sintetizaram e expandiram o trabalho de diversos autores e criaram uma taxionomia para os métodos de suavização exponencial: combinando o erro com a tendência e a sazonalidade existem 30 modelos, onde os métodos de suavização simples, de Holt e de Holt-Winters são

casos especiais. Tal classificação é denominada pelos autores de ETS (*error, trend and seasonality*). O quadro 1 mostra a classificação proposta por Hyndman et al. (2008). Para cada um dos 15 métodos do quadro 1 existem duas variantes, uma com o erro aditivo e outra com o erro multiplicativo, gerando 30 métodos de suavização. A suavização exponencial simples é o modelo (N,N), a suavização de Holt, o modelo (A,N) e a de Holt-Winters, os modelos (A,A) e (A,M).

Quadro 1: Classificação dos Métodos de Suavização Exponencial

Tendência	Sazonalidade		
	N	A	M
N (nenhuma)	N,N	N,A	N,M
A (aditiva)	A,N	A,A	A,M
Ad (aditiva e amortecida)	Ad,N	Ad,A	Ad,M
M (multiplicativa)	M,N	M,A	M,M
Md (multiplicativa e amortecida)	Md,N	Md,A	Md,M

Fonte: Hyndman et al. (2008), p. 12.

2.3. Avaliação da Precisão das Previsões

Para Hyndman e Athanasopoulos (2018), 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. O erro de previsão é diferente do resíduo do modelo, pois eles são aplicados em dados diferentes (teste e treino, respectivamente). Segundo Hyndman e Athanasopoulos (2018), 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) e suas fórmulas encontram-se nas equações 2 e 3, respectivamente.

$$EAM = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{2}$$

$$REQM = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \tag{3}$$

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 (2018) propõem o uso do erro escalonado absoluto médio (MASE – *mean absolute scaled error*). No presente artigo, como será analisada uma única série temporal, serão utilizados os erros dependentes de escala EAM e REQM, devido sua facilidade de interpretação.

3. Metodologia

Inicialmente foram obtidas as séries temporais a serem analisadas. A série mensal "taxa de desocupação das pessoas de 14 anos ou mais de idade" foi retirada do banco de dados do IPEA – Instituto de Pesquisa Econômica Aplicada (http://www.ipeadata.gov.br). No momento do acesso, ela cobria o período que ia de março de 2012 até janeiro de 2021.

Para a obtenção dos índices do Google Trends, acessou-se a página do Trends na internet (https://trends.google.com.br) e inseriu-se como termo de busca a palavra-chave "seguro-desemprego" para o período de 01/01/2012 a 31/03/2021. O período de cobertura dessa série foi maior devido ao fato de haver a possibilidade de se utilizar regressores defasados na regressão dinâmica.

De posse das 2 séries, cada uma foi dividida em duas partes: os dados de treino e os de teste. Seguindo o mesmo procedimento da *M3* e da *M4 Competition* para dados mensais, os dados de teste foram os últimos 12 valores das séries, ou seja, o horizonte de tempo para as previsões foi de 12 meses. A *M4 Competition* (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018) e sua antecessora, a *M3 Competiton* (MAKRIDAKIS; HIBON, 2000), foram competições onde *experts* foram convidados a fornecer previsões para séries reais usando métodos variados e são referências no campo de previsões de séries temporais.

A estimação dos modelos, as previsões e os seus respectivos erros foram executados no software R (R CORE TEAM, 2021) versão 4.1.0 utilizando o pacote *forecast* versão 8.14 desenvolvido por Hyndman e Khandakar (2008). Pacotes são coleções de funções, dados e códigos compilados e 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 à verificação da comunidade científica com relação a sua correção, eficácia, etc.

O ajuste de modelos ARIMA é feito automaticamente pelo pacote *forecast*. Ele determina se diferenciações são necessárias utilizando o chamado de teste de raízes unitárias e seleciona as ordens p, q, P e Q minimizando uma função perda, por padrão o critério de informação de Akaike corrigido (AICc). A seguir, os parâmetros do modelo são estimados pelo método de máxima verossimilhança. Por fim, o pacote faz automaticamente as previsões para o horizonte de tempo solicitado e calcula os erros de previsão. O pacote também é capaz de estimar modelos de regressão dinâmica, escolhendo de forma totalmente automática o processo ARIMA que melhor se ajusta aos resíduos do modelo de regressão fornecido.

A suavização exponencial também foi executada no pacote *forecast.* É digno de nota que algumas combinações do quadro 1 são instáveis, assim o pacote não implementa todos os 30 modelos listados. Para séries com tendência e sazonalidade, como a deste trabalho, o pacote permite as seguintes combinações: AAA, MAA, AAdA, MAdA, MAM, MAdM, MMM, MMdM. No presente artigo, estes 8 modelos de suavização exponencial formam, juntamente com o ARIMA puro, os *benchmarks* para se avaliar a precisão das previsões da regressão dinâmica.

Porém, após a primeira rodada de análise dos dados notou-se que os dados de teste cobriam o período de fevereiro de 2020 a janeiro de 2021, coincidindo com a pandemia do novo coronavírus, fazendo com que todos os métodos de previsão falhassem, gerando erros de

previsão substanciais. Isso se deve ao fato de que uma das premissas dos métodos quantitativos não foi atendida, a de que é razoável supor que alguns aspectos dos padrões passados continuarão no futuro. O impacto da pandemia na economia e, por consequência, no mercado de trabalho foi tão pronunciado que tal suposição não se sustenta. Isso introduz muito ruído na análise, não permitindo uma comparação justa e adequada entre os métodos de previsão.

Assim, eliminou-se do conjunto de dados o período de abril de 2020 em diante, pois os primeiros casos reportados de covid-19 no Brasil foram no final de fevereiro e o impacto maior na economia começou a se manifestar a partir de abril. Portanto, foi executada uma nova rodada de análises com a seguinte subdivisão: dados de treino indo de março de 2012 até março de 2019 e dados de teste indo de abril de 2019 até março de 2020. O resultado das análises é reportado a seguir.

4. Resultados

Inicialmente, verificou-se se a série y_t (taxa de desocupação) era estacionária na variância ou se seria necessária uma transformação de Box-Cox para se estabilizar a mesma (HYNDMAN; ATHANASOPOULOS, 2018). O software R tem uma função que determina o expoente ótimo da transformação e seu intervalo de confiança 95%. Como tal intervalo conteve o número 1, nenhuma transformação fez-se necessária. A seguir, por meio do método automático de seleção de modelos ARIMA do pacote *forecast*, ajustou-se um ARIMA puro (sem os índices do Trends) aos dados de treino. Utilizando o critério de minimização do AICc, o algoritmo determinou como ótimo o modelo sazonal ARIMA$(2,1,0)(1,1,0)_{12}$.

Na sequência, foram testados diversos modelos de regressão dinâmica. Devido a limitações de espaço e para maior clareza na exposição, serão reportados apenas os que tiveram o melhor desempenho nos dados de teste (menores EAM e REQM). O primeiro foi um modelo com 3 regressores, isto é, o índice do Trends contemporâneo (x_t), o índice com uma defasagem temporal (x_{t-1}) e o índice com duas defasagens (x_{t-2}). O segundo foi um modelo de regressão apenas com x_{t-2}. Em ambos os casos, o pacote *forecast* determinou automaticamente como melhor ARIMA para os erros da regressão dinâmica o modelo ARIMA$(0,1,2)(2,0,0)_{12}$.

Uma vez que os modelos ARIMA com e sem regressores foram diferentes entre si, optou-se por testar as seis combinações entre ambos: os dois ARIMA puros e as duas regressões dinâmicas com ambas as estruturas de erro cada. Assim, cada um dos seis modelos foi ajustado aos dados de treino, as previsões para os 12 meses à frente foram determinadas e os erros de previsão calculados, como indica a Tabela 1. Os comandos executados no software R encontram-se no Apêndice 1, juntamente com as séries temporais utilizadas.

Feito isso, partiu-se para os cálculos com os 8 modelos de suavização exponencial sazonais descritos na seção 2.2. O procedimento foi análogo. Os erros de previsão de cada modelo ETS estão na Tabela 1 e os comandos executados no R encontram-se no Apêndice 1.

Pelos dados das Tabela 1 é possível verificar que dentre os modelos ARIMA, o melhor foi o ARIMA$(0,1,2)(2,0,0)_{12}$ e único regressor (x_{t-2}) com erros REQM de 0,1846 pontos percentuais (p.p.) e EAM de 0,1413 p.p. Isto é, a cada mês o método errou em média, para cima ou para baixo, este valor da taxa de desocupação em p.p. Como a taxa de desocupação neste período foi em torno de 12%, esses erros são baixos, em torno de 1,2% a 1,5% em termos relativos. Porém, analisando-se o desempenho do modelo ARIMA$(0,1,2)(2,0,0)_{12}$ puro, percebe-se que a diferença deste para a regressão dinâmica é muito pequena. Ou seja, a inclusão do regressor x_{t-2} teve um impacto incremental nos erros de previsão. O REQM diminuiu 0,0041 p.p em termos absolutos ou 2,16% em termos relativos. O EAM diminui 0,0019 p.p. em termos absolutos ou 1,34% em termos relativos.

Tabela 1: Erros de Previsão dos Modelos Considerados

Modelo	Especificação	REQM	EAM	Regressores
ARIMA	ARIMA(0,1,2)(2,0,0)$_{12}$	0,1891	0,1445	x_t, x_{t-1}, x_{t-2}
	ARIMA(0,1,2)(2,0,0)$_{12}$	0,1846	0,1413	x_{t-2}
	ARIMA(2,1,0)(1,1,0)$_{12}$	0,9399	0,8510	x_t, x_{t-1}, x_{t-2}
	ARIMA(2,1,0)(1,1,0)$_{12}$	0,9466	0,8552	x_{t-2}
	ARIMA(2,1,0)(1,1,0)$_{12}$	0,9315	0,8465	—
	ARIMA(0,1,2)(2,0,0)$_{12}$	0,1887	0,1432	—
ETS	AAA	0,2540	0,2075	—
	AAdA	0,3147	0,2570	—
	MAA	0,4921	0,4522	—
	MAdA	0,3434	0,3040	—
	MAM	0,2625	0,1964	—
	MAdM	0,1753	0,1249	—
	MMM	0,2200	0,1642	—
	MMdM	0,1530	0,1143	—

Fonte: Elaboração Própria.

Shikida et al (2012), ao analisarem esta série, reportaram uma redução de 8% no EAM com a introdução do índice do Trends. Porém, os EAM do estudo deles foram em torno de 0,80 p.p., muito superiores ao do nosso melhor modelo. Pode-se notar na Tabela 1 que os modelos ARIMA(2,1,0)(1,1,0)$_{12}$ com e sem regressores tiveram erros dessa magnitude (em torno de 0,85 p.p.). Ou seja, é possível que a melhora de 8% que Shikida et al (2012) encontraram foi devido ao fato que o modelo ARIMA que eles especificaram pode não ter sido ótimo. Tivessem eles especificado outro modelo ARIMA puro, talvez a redução teria sido menor.

Observando-se os dados da Tabela 1, percebe-se que no geral os modelos de suavização exponencial foram melhores para essa série. Em particular, os modelos MMdM e MAdM foram muito superiores aos modelos ARIMA(0,1,2)(2,0,0)$_{12}$ com e sem regressores. O melhor modelo ETS (MMdM) diminuiu o REQM em 0,0316 p.p. ou 17,12% e o EAM em 0,027 p.p. ou 19,11%, relativamente ao ARIMA(0,1,2)(2,0,0)$_{12}$ com único regressor.

Camilo e Ruy (2019) já haviam estudado o desempenho relativo desses dois métodos (ARIMA puro e ETS) para séries não sazonais e, muito embora em média seus desempenhos fossem idênticos, para algumas séries havia uma disparidade tão grande quanto essa observada. Aparentemente o mesmo ocorre para séries sazonais. Dessa forma, seria interessante tentar verificar quais são as características distintivas das séries que levam a resultados tão diferentes e se há algum tipo de padrão. Logicamente, uma maneira prática de se resguardar dos efeitos dessas variações desconhecidas no desempenho do métodos é utilizar mais de um método de previsão e tirar a média (ou mediana) de seus resultados, conforme se havia concluído na *M4 Competition* (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018).

Tomadas em conjunto, as comparações da regressão dinâmica com o modelo ARIMA puro e com o ETS permitem concluir que o uso do índice do Google Trends com a palavra-chave escolhida (“seguro desemprego”) não mostrou ganhos em sua adoção no quesito melhoria da precisão das previsões para a série “taxa de desocupação das pessoas de 14 anos ou mais de idade”.

5. Considerações Finais

Este artigo teve como objetivo verificar se uma regressão dinâmica com o uso do Google Trends melhoraria o desempenho das previsões para a série mensal "taxa de desocupação das pessoas de 14 anos ou mais de idade", relativamente os métodos de suavização exponencial e ARIMA puros, medido pelo erro absoluto médio e pela raiz quadrada do erro quadrático médio. Como o método não gerou erros de previsão menores que os métodos tradicionais que utilizam apenas os valores da própria série temporal, sua adoção rotineira não se justificativa. Os resultados da presente pesquisa indicam que o desempenho relativo dos métodos de previsão depende mais dos fatores característicos da própria série, tais como heteroscedasticidade, não linearidade e grau de aleatoriedade (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018).

Porém, este resultado requer ser analisado do ponto de vista de seu grau de generalização e de suas limitações. A primeira diz respeito à palavra-chave utilizada. O termo de busca "seguro desemprego" tem relação apenas com os desempregados que fazem jus ao benefício. O Brasil tem alto grau de informalidade no mercado de trabalho, o que torna esse termo um indicador não tão eficiente, diferentemente de outros países com estrutura mais formal de mão de obra. Talvez o uso de mais de um *leading indicator* pudesse mitigar essa baixa confiabilidade. Assim, trabalhos futuros poderiam verificar se o uso de um conjunto de indicadores melhoraria a qualidade das previsões.

Além das características do mercado de trabalho, o Brasil também tem um alto nível de exclusão digital, o que diminui a associação entre o volume de buscas e a série estudada. A última PNAD Contínua de divulgação anual (IBGE, 2021) indica que dos 72.929 mil domicílios brasileiros, em 17,3% deles não havia internet. Adicionalmente, das 183.296 mil pessoas com 10 anos ou mais, 26,3% delas ou possuem celular sem acesso à internet ou não possuem o aparelho e 21,7% delas não haviam feito uso da internet em pelo menos um momento nos 90 dias que antecederam a data da entrevista.

Uma limitação adicional é que neste estudo foram utilizados métodos automáticos de seleção de modelos. É possível que um analista experiente utilizando informações a respeito das séries e os métodos tradicionais de identificação chegasse a modelos melhores do que os identificados automaticamente pelo software e que nesses modelos a precisão das previsões diferisse significativamente. Assim, os resultados obtidos devem ser interpretados neste contexto.

Outra limitação do presente artigo, bem como o de estudos correlatos (CHOI; VARIAN, 2012; SHIKIDA et al, 2012) é o uso da regressão dinâmica. Como a variável taxa de desocupação (y_t) afeta o índice de busca (x_t), a regressão não é a maneira mais eficiente de análise dos dados. O ideal seria tratar os dados como uma análise multivariada de séries temporais. Assim, trabalhos futuros poderiam usar o método multivariado denominado "Vector ARMA" na análise desse tipo de problema e verificar se nesta estrutura os resultados são mais promissores.

Referências

CAMILO, G. I.; RUY, M. Avaliação da Precisão dos Métodos ARIMA, Suavização Exponencial e Redes Neurais na Previsão de Séries Temporais Anuais Brasileiras. *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: 01 mar. 2021.

CHOI, H.; VARIAN, H. Predicting the Present with Google Trends. **The Economic Record**, v. 88, Special Issue, p. 2-9, June, 2012.

FERREIRA, A. B. H. **Novo Dicionário da Língua Portuguesa**. Rio de Janeiro: Editora Nova Fronteira, 1975.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. 6. ed. Bookman: Porto Alegre, 2009.

HARRISON, P.; STEVENS, C. Bayesian Forecasting. **Journal of the Royal Statistical Society**. Series B (Methodological), v. 38, n.3, p. 205-247, 1976.

HYNDAMN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. 2. Ed. Melbourne: OTexts, 2018. Disponível em: https://otexts.org/fpp2/. Acesso em: 01 mar. 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.

HYNDMAN, R. J.; KOEHLER, A. B.; ORD, J. K.; SNYDER, R. D. **Forecasting with Exponential Smoothing**: the state space approach. Berlin: Springer-Verlag, 2008.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Pesquisa Nacional por Amostra de Domicílios Contínua**: acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal 2019. Rio de Janeiro: IBGE, 2021. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/17270-pnad-continua.html?edicao=30362&t=sobre. Acesso em: 16 junho 2021.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA (IPEA). Ipeadata: disponível em http://www.ipeadata.gov.br. Acesso em: 01 mar. 2021.

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.

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

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, 2019.

SHIKIDA, C. D.; BYRRO, R. M.; SALVATO, M. A.; ARAUJO JR., A. F. Informações e Política Econômica: um teste para aperfeiçoamento de erros de previsão a partir da utilização do Google Trends. **Revista Gestão & Políticas Públicas**, v. 2, n. 2, p. 197-218, 2012.

SINGER, M. Thoughts of a Nonmillenarian. **Bulletin of the American Academy of Arts and Sciences**, v. 51, n. 2, p. 36-51, 1997.

Apêndice 1 – Comandos utilizados no software R para o cálculo dos erros de previsões

```
### Utiliza o pacote "forecast"
install.packages("forecast", dependencies = TRUE)
library(forecast)

##### Entrada de Dados
### Dados até janeiro de 2021
y = ts(c(7.9, 7.7, 7.6, 7.5, 7.4, 7.3, 7.1, 6.9, 6.8, 6.9, 7.2, 7.7,
8.0, 7.8, 7.6, 7.4, 7.3, 7.1, 6.9, 6.7, 6.5, 6.2, 6.4, 6.7, 7.2, 7.1,
7.0, 6.8, 6.9, 6.9, 6.8, 6.6, 6.5, 6.5, 6.8, 7.4, 7.9, 8.0, 8.1, 8.3,
8.5, 8.7, 8.9, 8.9, 9.0, 8.9, 9.5, 10.2, 10.9, 11.2, 11.2, 11.3, 11.6,
11.8, 11.8, 11.8, 11.8, 12.0, 12.6, 13.2, 13.7, 13.6, 13.3, 13.0,
12.8, 12.6, 12.4, 12.2, 12.0, 11.8, 12.2, 12.6, 13.1, 12.9, 12.7,
12.4, 12.3, 12.1, 11.9, 11.7, 11.6, 11.6, 12.0, 12.4, 12.7, 12.5,
12.3, 12.0, 11.8, 11.8, 11.8, 11.6, 11.2, 11.0, 11.2, 11.6, 12.2,
12.6, 12.9, 13.3, 13.8, 14.4, 14.6, 14.3, 14.1, 13.9, 14.2), start =
c(2012, 3), frequency = 12)
###
x = ts(c(39, 31, 30, 29, 28, 26, 29, 25, 24, 25, 23, 23, 32, 26, 27,
29, 26, 25, 27, 25, 25, 28, 23, 22, 31, 30, 29, 29, 30, 30, 32, 30,
30, 33, 34, 36, 59, 49, 56, 58, 59, 55, 49, 44, 43, 42, 49, 39, 40,
46, 43, 42, 44, 48, 38, 35, 32, 34, 35, 34, 47, 40, 41, 37, 35, 33,
34, 32, 30, 30, 31, 29, 38, 38, 38, 38, 34, 34, 36, 34, 30, 32, 29,
31, 41, 40, 40, 41, 39, 36, 35, 34, 31, 31, 34, 30, 41, 37, 48, 100,
77, 50, 50, 42, 41, 46, 34, 26, 35, 33, 35), start = c(2012, 1),
frequency = 12)
###
x1 = lag(x, -1)       ### x_{t-1}
####
x2 = lag(x, -2)       ### x_{t-2}
####
dados = ts.intersect(y, x, x1, x2)
### divide os dados e elimina o período a partir de abril/2020
treino = window(dados, start = c(2012, 3), end = c(2019, 3))
teste = window(dados, start = c(2019, 4), end = c(2020, 3))

##### Modelos ARIMA e Regressão Dinâmica - Ajuste e Previsões
### ARIMA(0,1,2)(2,0,0)_12 com x_t, x_{t-1} e x_{t-2}
m1 = Arima(treino[ ,1], c(0,1,2), c(2,0,0), xreg = treino[ ,2:4])
prev_m1 = forecast(m1, xreg = teste[ ,2:4], h = 12)
### ARIMA(0,1,2)(2,0,0)_12 com x_{t-2}
m2 = Arima(treino[ ,1], c(0,1,2), c(2,0,0), xreg = treino[ ,4])
prev_m2 = forecast(m2, xreg = teste[ ,4], h = 12)
### ARIMA(2,1,0)(1,1,0)_12 com x_t, x_{t-1} e x_{t-2}
m3 = Arima(treino[ ,1], c(2,1,0), c(1,1,0), xreg = treino[ ,2:4])
prev_m3 = forecast(m3, xreg = teste[ ,2:4], h = 12)
### ARIMA(2,1,0)(1,1,0)_12 com x_{t-2}
m4 = Arima(treino[ ,1], c(2,1,0), c(1,1,0), xreg = treino[ ,4])
prev_m4 = forecast(m4, xreg = teste[ ,4], h = 12)
### ARIMA(2,1,0)(1,1,0)_12
m0a = Arima(treino[ ,1], c(2,1,0), c(1,1,0))
prev_m0a = forecast(m0a, h = 12)
### ARIMA(0,1,2)(2,0,0)_12
```

```
m0b = Arima(treino[ ,1], c(0,1,2), c(2,0,0))
prev_m0b = forecast(m0b, h = 12)
#
### Erros de Previsão
accuracy(prev_m1, teste[ ,1])
accuracy(prev_m2, teste[ ,1])
accuracy(prev_m3, teste[ ,1])
accuracy(prev_m4, teste[ ,1])
accuracy(prev_m0a, teste[ ,1])
accuracy(prev_m0b, teste[ ,1])

##### Modelos ETS - Ajuste e Previsões
mod1 = ets(treino[ ,1], "AAA", damped = FALSE)        ### AAA
mod2 = ets(treino[ ,1], "AAA", damped = TRUE)         ### AAdA
mod3 = ets(treino[ ,1], "MAA", damped = FALSE)        ### MAA
mod4 = ets(treino[ ,1], "MAA", damped = TRUE)         ### MAdA
mod5 = ets(treino[ ,1], "MAM", damped = FALSE)        ### MAM
mod6 = ets(treino[ ,1], "MAM", damped = TRUE)         ### MAdM
mod7 = ets(treino[ ,1], "MMM", damped = FALSE)        ### MMM
mod8 = ets(treino[ ,1], "MMM", damped = TRUE)         ### MMdM
p1 = forecast(mod1, h = 12)
p2 = forecast(mod2, h = 12)
p3 = forecast(mod3, h = 12)
p4 = forecast(mod4, h = 12)
p5 = forecast(mod5, h = 12)
p6 = forecast(mod6, h = 12)
p7 = forecast(mod7, h = 12)
p8 = forecast(mod8, h = 12)
### Erros de Previsão
accuracy(p1, teste[ ,1])
accuracy(p2, teste[ ,1])
accuracy(p3, teste[ ,1])
accuracy(p4, teste[ ,1])
accuracy(p5, teste[ ,1])
accuracy(p6, teste[ ,1])
accuracy(p7, teste[ ,1])
accuracy(p8, teste[ ,1])
```