Previsão de divulgações do CAGED a partir de dados do Google Trends[*]

Forecasting CAGED Releases Using Google Trends Data

Daniel Arruda e Luiz Carlos Day Gama[**]

Resumo: Objetiva-se investigar a correlação entre dados do Google Trends e o saldo de empregos criados em cada mês (CAGED), entre janeiro de 2010 e junho de 2019, além de sugerir um modelo de previsão que permita antecipar o indicador econômico. O modelo proposto é comparado a outro univariado através da aderência das previsões fora da amostra de cada um aos dados reais. Os resultados demonstram que há correlação entre o CAGED e as séries de buscas no Google para alguns termos e que esses dados melhoram o desempenho preditivo de um modelo básico para o indicador do mercado de trabalho.

Palavras-chave: CAGED. Emprego. Google Trends. Big data. Previsão.

Abstract: This research study was conducted to investigate the correlation between Google Trends data and the balance of monthly jobs created (CAGED), from January 2010 to June 2019. A forecasting model that allows the anticipation of the economic indicator was suggested. The proposed model is compared to a univariate one, through the adherence of forecasts out of the sample for each of the real data of the period. Results demonstrate that there is a correlation between CAGED and Google search series for some terms. Besides, we found that these data improve the predictive performance of a basic model for the labor market indicator.

Keywords: CAGED. Employment. Google Trends. Big data. Forecasting.

JEL: C22. E27.

[*] Submissão: 24/08/2022 | Aprovação: 27/12/2022 | DOI: 10.5380/re.v44i84.87339

[**] Respectivamente: (1) Economista na LCA Consultores, Brasil | ORCID: 0000-0003-4447-2360 | E-mail: danielferreira1714@gmail.com | (2) Professor Titular I no Ibmec-BH, Brasil | ORCID: 0000-0001-7026-2709 | E-mail: luiz.gama@professores.ibmec.edu.br

1. Introdução

A informação obtida em tempo real se faz cada vez mais indispensável no mundo atual. No âmbito econômico, os agentes hoje baseiam suas análises majoritariamente em divulgações de dados por órgãos oficiais, como o IBGE e o Banco Central, que geralmente demoram algumas semanas para serem consolidadas e disponibilizadas.

O Cadastro Geral de Empregados e Desempregados (CAGED), por exemplo, é um indicador econômico divulgado pelo Ministério da Economia que reflete o número de admitidos e desligados no regime celetista, entre outras informações relevantes, em todo o território nacional. Esse dado, portanto, está intimamente relacionado a diversas outras variáveis, sendo então de suma importância para compreender o cenário econômico nacional.

Entretanto, assim como outras estatísticas oficiais, os dados do CAGED são divulgados com atraso em relação ao mês do exercício correspondente (aproximadamente de um mês). Dessa forma, para todos aqueles envolvidos na tarefa de antecipar a evolução da conjuntura econômica (como bancos, empresas privadas, investidores, órgãos do governo e pesquisadores), faz sentido buscar dados alternativos disponibilizados em tempo real, cuja correlação contemporânea seja historicamente elevada com o dado oficial do CAGED (*nowcast*).

Um candidato que surge na literatura econômica recente é o *Google Trends*. Essa ferramenta, criada pela empresa de tecnologia Google, sintetiza todas as pesquisas feitas no seu buscador por um determinado termo e em determinado período, e disponibiliza uma série temporal de índices que refletem a variação do interesse do público por aquele termo (a informação pode ainda ser segmentada para a região geográfica de interesse do usuário).

Diversos artigos, principalmente internacionais, têm apontado o alto poder preditivo das séries temporais extraídas de pesquisas no google na antecipação da evolução de variáveis econômicas, tanto desagregadas, como vendas de automóveis (Choi; Varian, 2012), quanto agregadas, como inflação (Guzman, 2011). Vale notar que boa parte dos artigos nessa área busca prever variáveis ligadas à evolução do emprego formal, porém, também é possível encontrar outros artigos cujo foco esteja, por exemplo, na antecipação de aumento de volatilidade no mercado financeiro (Perlin *et al*., 2016).

O buscador do Google é amplamente utilizado ao redor do mundo, atingindo mais de 1,2 trilhão de pesquisas no ano (ou 40.000 por segundo, em média[1]). Dessa forma, é natural esperar que indivíduos na procura de ocupação laboral usem essa ferramenta para obter informações sobre vagas de emprego e realizar suas candidaturas. Uma vez que o número de pessoas contratadas em determinado mês seja proporcional ao número daqueles que procuraram emprego, então é possível que os índices de busca do *Google Trends* por termos como "emprego" sejam correlacionados a um saldo maior de criação de vagas de trabalho, antecipando, em certa medida, o dado do CAGED.

O objetivo desse trabalho é responder se os dados obtidos de frequência de buscas na internet, como no *Google Trends*, permitem a antecipação e previsão da evolução de uma variável econômica ligada ao emprego (divulgações do CAGED), no contexto brasileiro recente (2010 a 2019).

Para isso, são comparados dois modelos de previsão para dados do CAGED. O primeiro é um modelo puramente autorregressivo, enquanto o segundo é um ARIMAX, ou seja, um modelo que incorpora como variáveis independentes as próprias defasagens relevantes da variável dependente, além de variáveis exógenas em termos contemporâneos, que são os dados do *Google Trends*. O principal critério de comparação é a proximidade dos dados projetados por cada modelo em relação aos dados reais observados em determinado período (projeção fora da amostra). Obviamente, o modelo que apresenta o menor erro é considerado superior. Caso o modelo que incorpora os dados do Google tenha menor erro de previsão, isso será uma evidência adicional da relevância das pesquisas na internet para se conhecer, em tempo real, a evolução das variáveis econômicas em determinada região.

Uma conclusão obtida é que a utilidade do índice de pesquisas como regressor em um modelo depende fortemente da escolha pelos termos corretos, isto é, a infinidade de termos que podem ser pesquisados pelo indivíduo exige que o pesquisador tenha um método que possibilite encontrar os dados mais valiosos, assim como evitar correlações espúrias. Por exemplo, enquanto o índice de buscas por "emprego" possui forte correlação com a série do CAGED e consiste em um

[1] Disponível em <https://www.internetlivestats.com/google-search-statistics/>. Acesso em: 19 mar 2020.

regressor estatisticamente significante no modelo ajustado, o mesmo não ocorre para o termo "trabalho".

Na segunda seção é feito um estudo de pesquisas nacionais e internacionais que buscaram entender a viabilidade de se usar as estatísticas disponibilizadas pelo Google para antever variáveis econômicas, especialmente aquelas ligadas ao mercado de trabalho. Na sequência, é feita uma análise descritiva dos dados usados nesse trabalho, seguido de uma descrição metodológica da análise realizada. Os resultados obtidos são então detalhados, analisados e expostos na forma de gráficos e tabelas, enquanto a última seção explicita as conclusões do presente trabalho.

2. Revisão da Literatura

A ferramenta de busca do Google é amplamente utilizada por pessoas de diversos países, gerando diariamente enormes bases de dados relativas às pesquisas dos usuários. Segundo o blog[2] oficial da empresa, em 2015 foi anunciada uma expansão da ferramenta *Google Trends*, possibilitando acesso direto e em tempo real às principais tendências de buscas ao redor do mundo, que já ultrapassavam 100 bilhões de pesquisas ao mês.

Com base nesse cenário, foi natural o surgimento de alguns trabalhos acadêmicos, tanto no Brasil quanto no exterior, que avaliassem a utilidade desses dados de alta frequência para auxiliar na previsão de variáveis econômicas, sob a hipótese de que a informação disponível refletiria a percepção dos diversos agentes econômicos (consumidores e produtores) e orientasse suas ações.

4.1 Literatura internacional

Choi e Varian (2009, 2012) são pioneiros no uso da ferramenta *Google Trends* para antecipação de dados econômicos. O objetivo dos autores foi demonstrar que enquanto estatísticas oficiais de governos e outros órgãos de pesquisa apresentam certo atraso de divulgação, dados de alta frequência podem ser usados para estimar antecipadamente os indicadores econômicos presentes (*nowcast*). Os autores encontram resultados que sugerem que modelos que utilizam dados do *Google Trends* apresentam desempenho superior a modelos

[2] Disponível em <https://googleblog.blogspot.com/2015/06/a-new-window-into-our-world-with-real.html>. Acesso em: 01 mar 2020.

autorregressivos, considerando séries de dados como vendas de veículos de 2004 a 2011 nos Estados Unidos.

Askitas e Zimmermann (2009) testam o uso de dados do *Google Trends* com o objetivo de prever a taxa de desemprego mensal em um período de turbulência para a economia alemã, de janeiro de 2004 a abril de 2009, quando informações em tempo real seriam ainda mais valiosas. Os autores usam a especificação ECM (*error correction model*) e sugerem que há forte correlação entre séries de pesquisa por palavras-chave do Google e a taxa de desemprego na Alemanha.

Visando testar o poder preditivo de pesquisas na internet, D'Amuri e Marcucci (2009) realizam comparações entre 520 modelos ARMA distintos, pelo método de previsões fora da amostra, e sugerem um índice de dados extraídos do google de pesquisas pelo termo "*jobs*", como o melhor *leading indicator* para se projetar a taxa de desemprego dos Estados Unidos, entre 2004 e 2009, inclusive em relação a modelos que usam dados relativos a pedidos de seguro-desemprego.

Guzman (2011) busca responder se dados de alta frequência são capazes de revelar as expectativas dos agentes econômicos em tempo real nos Estados Unidos. O autor utiliza dados do *Google Trends*, em comparação a indicadores de baixa frequência (36 pesquisas periódicas) e à taxa de inflação implícita nos juros, e avalia que aqueles possuem validade na antecipação da inflação em 12 meses utilizando metodologias de projeção fora da amostra e testes de causalidade Granger.

Especialmente para o caso de uma economia emergente, Chadwick e Sengul (2012) visam testar a validade de se usar dados do *Google Trends* para antecipar a taxa de desemprego não agrícola na Turquia, de 2005 a 2012, país onde apenas 45% da população tinha acesso à internet (segundo dados do *World Bank*, esse número já ultrapassava 70% [3]em 2018). Os dados do Google são referentes a pesquisas por termos como "*job announcements*" e relacionados, além de nomes de sites de carreira populares. Os modelos lineares propostos apresentam desempenho superior aos puramente autorregressivos e são de grande utilidade,

[3] Disponível em <https://data.worldbank.org/indicator/IT.NET.USER.ZS>. Acesso em: 11 mar 2020.

uma vez que os dados oficiais do país são divulgados com atraso médio de três meses.

Já em artigo teórico, Edelman (2012) busca expor os fatores relevantes relacionados ao uso de bases de dados disponíveis na internet. Segundo o autor, a internet não apenas aumenta a quantidade de dados disponíveis para o trabalho do economista, como também reduz o custo de coletar informação e permite, em consequência, mais trabalhos empíricos realizados em menor tempo.

Einav e Levin (2014) buscam analisar como o avanço do chamado *Big Data* - definido como dados disponíveis em tempo real e em larga escala, compostos por novos tipos de variáveis e menos estruturados - pode impactar a pesquisa econômica e a análise de políticas públicas. Segundo os autores, as grandes bases de dados podem melhorar a eficiência das operações do governo, além de permitir novos campos de pesquisa econômica ainda não explorados.

Já Tuhkuri (2016) se vale de mais de 35 milhões de consultas obtidas do google por 13 termos com alto volume de buscas, como "*unemployment benefits*", todos ligados a benefícios de seguro-desemprego, e demonstra que esses dados melhoram o poder preditivo de modelos que projetam a taxa de desemprego dos Estados Unidos no curto prazo, para o período de 2004 a 2014. A metodologia empregada se baseia em testes de causalidade de Granger e estudo da função de correlação cruzada.

Por fim, Koop e Onorante (2018) afirmam que séries do *Google Trends* podem ser usadas não somente para melhorar a previsão de variáveis desagregadas, mas também de variáveis macroeconômicas convencionais, como desemprego e inflação. Os autores utilizam indicadores econômicos mensais para os Estados Unidos, a partir de 2004, e os dados do google são obtidos partindo da busca pela própria variável em questão, acrescentando os termos relacionados sugeridos pelo buscador.

4.2 Literatura nacional

Perlin *et al.* (2016) buscam identificar o impacto de séries específicas do *Google Trends* em oscilações do mercado financeiro, para quatro países de língua inglesa (Estados Unidos, Reino Unido, Austrália e Canadá), no período de 2005 a 2014, sob a hipótese de que a ferramenta refletiria a percepção dos agentes

econômicos em geral (e entre eles *traders* e investidores). A metodologia do estudo se baseia na projeção de modelos VAR e em testes de causalidade de Granger. Os autores identificam termos específicos de pesquisa (como "*stocks*") correlacionados tanto com o aumento da volatilidade no mercado de capitais, quanto com a queda dos retornos dos ativos, sugerindo que investidores utilizam determinados termos em suas pesquisas majoritariamente antes das decisões de vendas.

Guimarães Filho (2017) busca avaliar o impacto do uso de dados do *Google Trends*, como pesquisas pelo termo "emprego", na antecipação das estatísticas do CAGED divulgadas pela Secretaria do Trabalho. O estudo se baseia na comparação do modelo que incorpora *Google Trends* com outro modelo autorregressivo acrescido de *dummies* sazonais. O período de análise é de janeiro de 2004 a julho de 2016. Como principais resultados o autor conclui que modelos que utilizam *Google Trends* têm desempenho melhor que outros puramente autorregressivos e a inclusão de outras variáveis explicativas macroeconômicas não é capaz de melhorar a performance preditiva.

Já Perez Albarella (2017) busca avaliar se dados de buscas na internet (extraídos do *Google Trends*), selecionados por pesquisas com internautas, são úteis para melhorar a previsão de modelos macroeconômicos consagrados, como a lei de Okun, a curva de Phillips e a curva IS. As variáveis estimadas são inflação, taxa de desemprego e PIB, todas divulgadas pelo IBGE entre 2004 a 2017. A metodologia se baseia na comparação de previsões fora da amostra dos modelos tradicionais e dos modelos que incorporam os dados de buscas no Google. A autora conclui que *Google Trends* são estatísticas úteis na previsão de variáveis econômicas, embora não substituam as variáveis explicativas dos modelos tradicionais.

Com base na literatura revisada, esse trabalho busca fornecer evidência adicional sobre a capacidade dos dados extraídos de frequência de buscas no Google de antecipar indicador econômico ligado ao emprego, para o contexto brasileiro recente.

3. Base de Dados

Para a obtenção, manipulação dos dados, estatísticas descritivas e para a estimação dos modelos econométricos foi utilizado o *software* estatístico *R*.

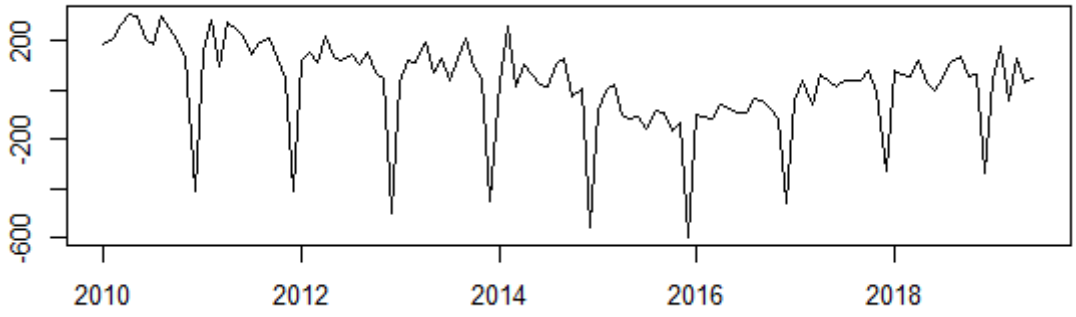
4.1 Caged

O Cadastro Geral de Empregados e Desempregados (CAGED) é um dispositivo legal, incorporado ao Ministério da Economia, que possui a função de registrar admissões e desligamentos sob o regime da CLT, informados por milhares de empresas em todo o território nacional. Dessa forma, o indicador está relacionado à evolução de diversas variáveis econômicas, como Produto Interno Bruto (PIB), taxa de desemprego e índices de inflação.

Embora os dados do CAGED sejam de fundamental importância para todos aqueles envolvidos na análise da conjuntura econômica brasileira, a divulgação ocorre, historicamente, no final do mês seguinte àquele do exercício analisado, representando um atraso aproximado de 30 dias.

Neste trabalho, utiliza-se o saldo (admitidos menos desligados) de empregos gerados em frequência mensal divulgado atualmente pelo Ministério da Economia em seu *website*[4], abrangendo o período de 1992 até 2019.

Figura 1 **– Série do CAGED**



Fonte – Ministério da Economia.

Observando o gráfico acima nota-se que a série de dados do CAGED possui caráter fortemente sazonal, apresentando de forma recorrente uma retração nos

[4] Disponível em < http://pdet.mte.gov.br/caged>. Acesso em: 07 mar 2020.

meses de dezembro. Esse fato pode ser explicado pela demissão dos funcionários contratados temporariamente para as vendas de fim de ano. Ademais, vê-se uma leve tendência de aumento do saldo após a recessão econômica, sugerindo que uma leve retomada do emprego estivesse em curso no período.

4.2 Google Trends

O *Google Trends* é uma ferramenta ligada ao buscador da empresa norte-americana Google que disponibiliza, em tempo real, estatísticas que sintetizam a evolução das pesquisas feitas por termos específicos, em bases semanais ou mensais, desde 2004.

Segundo Choi e Varian (2012), o Google disponibiliza um índice de série temporal do volume de pesquisas por determinado termo e em determinada área geográfica. Esse índice é calculado com base na razão entre as buscas pelo termo e o número total de pesquisas na mesma região no período sob análise. Posteriormente, os resultados obtidos são normalizados para o intervalo entre 0 e 100, sendo cada valor proporcional ao número de pesquisas. Há ainda certa flexibilidade, no sentido de que o índice considera todas as ocorrências de pesquisas que incluem o termo buscado e não somente quando há correspondência exata. Dessa forma, o índice equivalente ao termo "emprego" contará também as buscas por "emprego para economista", por exemplo.

Os termos buscados foram "emprego", "empresas" e "trabalho", sendo a motivação detalhada na seção 4.1. O período de análise é definido de janeiro de 2010 até junho de 2019, por dois motivos específicos. Primeiramente, buscou-se um período em que o uso da ferramenta digital se fazia mais comum na procura por colocação profissional no Brasil e, portanto, são eliminadas as observações mais antigas. Em segundo lugar, a série de buscas pelo termo "emprego" apresentou *outliers* significativos para os meses de agosto e setembro de 2019 (acima do dobro da média histórica), possivelmente em função da aprovação da medida provisória da Liberdade Econômica na Câmara, que alterava pontos referentes à Consolidação das Leis do Trabalho (CLT). Dada a dificuldade de antever eventos políticos, esse período é retirado da base de dados, de forma a não enviesar o resultado das previsões.

A abrangência dos dados é nacional. Assim como o CAGED, existe a possibilidade de desagregação por municípios, de forma a analisar ao longo de um determinado período em quais cidades o termo foi mais ou menos pesquisado. Entretanto, não há a possibilidade de ser feita uma análise intra-urbana, ou seja, como a procura pelo termo variou dentro de um município ao longo de um período, caracterizando uma limitação em comparação aos dados do CAGED.

Outra possível limitação dos dados do *Google Trends* e que deve ser considerada está relacionada à validade de conteúdo dos índices de pesquisa, ou seja, se a pesquisa na internet sobre um determinado índice é qualitativamente consistente com a questão que se pretende medir. Um exemplo apontado por Mellon (2014) é a palavra "jobs", que mensura a proeminência das questões trabalhistas, não tendo relação com Steve Jobs, antigo presidente da *Apple*. Obviamente é uma desvantagem em comparação aos dados tradicionais, como CAGED. Com isso, em futuras replicações, o contexto precisa ser considerado para mudanças nos índices de pesquisa, comparando-os com as medidas existentes e deve-se tomar cuidado com o significado dos termos de pesquisa, para evitar inferências incorretas.

Por se tratar de uma abordagem recente na literatura sobre mercado de trabalho a utilização de dados do *Google Trends*, no anexo do trabalho é incluído o código para obtenção dos dados.

4. Metodologia

4.1 Análise dos dados

Um desafio que surge na utilização de dados do *Google Trends* é a escolha dos termos que serão utilizados, dentre uma infinidade de possíveis candidatos. Para tanto, faz-se necessário um método objetivo, visando mitigar a existência de possíveis vieses do autor. Koop e Onorante (2013), por exemplo, iniciam a busca pela série da própria variável econômica que se deseja projetar, incluindo posteriormente os termos relacionados, sugeridos pelo próprio Google. Perez Albarella (2017), por sua vez, conduz uma enquete com 48 internautas sobre quais termos buscariam se desejassem obter mais informações sobre determinada variável econômica.

Utilizando o método de Koop e Onorante (2013), inicia-se com a série de buscas pelo termo "emprego", posteriormente incorporando também os termos "empresas" e "trabalho", sugeridos pelo Google como relacionados ao primeiro. Os termos obtidos possuem aderência aos encontrados por Guimarães Filho (2017), que utiliza a ferramenta *Google Correlate* para encontrar as melhores séries para projeção de dados do CAGED.

Define-se uma série temporal como uma coleção de observações feitas sequencialmente ao longo do tempo. Em geral, uma característica comum a esse tipo de dado é a presença de autocorrelação. Em outras palavras, existe associação linear entre um de seus componentes (geralmente o presente, Xt) e seu passado ou futuro, representados respectivamente por Xt−h ou Xt+h, em que h ∈ N é um inteiro positivo, denominado lag ou defasagem.

De forma geral, na presença de autocorrelação serial os estimadores de mínimo quadrados ordinários (MQO) continuam sendo não viesados e consistentes, porém deixam de ser eficientes. Em resumo, o modelo MQO tradicional não mais pode ser aplicado (Gujarati; Porter, 2011). Portanto, os modelos a serem aplicados devem dar conta do problema citado.

Para que as inferências estatísticas em séries temporais tenham validade é necessário que os resíduos da série estimada sejam estacionários, ou seja, a média e variância sejam constantes ao longo do tempo e que a covariância entre as observações dependa apenas da distância entre elas e não do tempo (BUENO, 2011).

Utiliza-se o teste *Augmented Dickey-Fuller (ADF)* para testar a estacionariedade das séries obtidas. Segundo Gujarati e Porter (2011), o teste ADF é uma expansão do teste *Dickey-Fuller* (DF), flexibilizando a hipótese de termo de erro não correlacionado. O procedimento ocorre pela verificação da existência de raiz unitária na série temporal analisada, isto é, se segue algum tipo de passeio aleatório[5], o que dificultaria a modelagem adequada. Dessa forma, um p-valor inferior a 5% implica a rejeição da hipótese nula (não estacionariedade), de forma que a série será estacionária. Um p-valor alto, por outro lado, mostra que não podemos rejeitar a hipótese nula e, portanto, não há como afirmar que a série não

[5] Um passeio aleatório gaussiano pode ser definido como $X_t = X_{t-1} + \varepsilon_t$, em que ε_t é o termo de erro aleatório, que se assume seguir distribuições normais independentes com média zero e variância constante.

seja um passeio aleatório. Havendo séries não estacionárias, será tomada a primeira diferença dos dados, o que possibilitará a modelagem correta.[6]

Uma vez obtidas as quatro séries temporais relevantes, busca-se compreender previamente se há alguma relação entre elas. Em teoria, pode-se esperar que os índices de busca no Google tenham correlação contemporânea (ou com poucas defasagens) positiva com os dados do CAGED. Isso ocorreria porque um aumento no número de indivíduos em busca de ocupação laboral provavelmente levaria tanto a mais pesquisas no buscador (considerando que boa parte das buscas por emprego ocorre no ambiente *online)* quanto a uma divulgação de um saldo de empregos criados mais robusta, tendo em vista que uma parcela dos indivíduos teria sucesso na busca por empregos.

Para uma análise inicial dessa hipótese, serão traçados os gráficos da evolução das variáveis para todo o período, permitindo uma comparação visual das séries. Além disso, também serão gerados gráficos para a função de autocorrelação da variável do CAGED (que provavelmente exibirá um valor alto na décima segunda defasagem, dado o fator sazonal da série, além de outras defasagens relevantes), e também uma função de correlação cruzada (FCC), entre dados do CAGED e do *Google Trends*, o que permitirá ter uma primeira impressão sobre a relação entre as variáveis (aqui espera-se que haja correlação contemporânea entre todas as variáveis), além de servir de base para a idealização de um modelo.

4.2 Modelo básico

Inicialmente, é definido um modelo básico puramente univariado para os dados do CAGED. Esse modelo, portanto, não utiliza os dados obtidos via *Google Trends*, e será estimado apenas para fins de comparação com o segundo modelo. São utilizados os dados do período de janeiro de 2010 até dezembro de 2018, de forma que os dados mais recentes são, posteriormente, utilizados para avaliação das projeções fora da amostra.

De acordo com Morettin (2011), uma série temporal pode ser modelada por meio de processo autorregressivo integrado e de médias móveis (ARIMA). Segundo o autor, um modelo ARMA (p,q) é dado pela equação em diferenças

$$X_t = \theta_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \qquad (1)$$

[6] Para maiores detalhes ver Bueno (2011).

em que $\varepsilon_t \sim RB(0, \sigma_\varepsilon^2)$, em que RB significa que os erros são do tipo ruído branco, ou seja, estacionários.

A partir disso, o autor define um modelo ARIMA como um processo onde a sua diferença é modelada conforme um ARMA (p,q), e acrescenta que os casos particulares que apresentam sazonalidade são tratados tomando a diferenciação sazonal da série, obtendo por fim o modelo ARIMA sazonal multiplicativo (SARIMA) de ordem (p,d,q) (P, D, Q).

Nessa etapa, será estimado o melhor modelo SARIMA que se adeque aos dados do CAGED, utilizando funções do software estatístico R, e o ajustamento do modelo será verificado tanto de forma visual, quanto pela função de autocorrelação e distribuição dos resíduos.

4.3 Modelo proposto

A partir das informações fornecidas pela função de autocorrelação cruzada, serão estimados três modelos ARIMAX para a variável dependente (CAGED), utilizando tanto as próprias defasagens mais significativas (o que certamente incluirá uma defasagem sazonal), quanto as três séries do *Google Trends* como variáveis exógenas, incluídas uma a uma, de forma contemporânea, visto que os dados estão disponíveis em tempo real e, portanto, antes da divulgação do dado do CAGED daquele mês. A estrutura do modelo é exposta abaixo

$$X_t = \theta_0 + \phi_p X_{t-p} + \varepsilon_t + \theta_1 GT_t^{emprego} + \theta_2 GT_t^{empresas} + \theta_3 GT_t^{trabalho} \quad (2)$$

em que X_t representa a primeira diferença dos dados do CAGED, X_{t-p} representa todas as suas defasagens significantes, conforme a função de correlação cruzada, GT_t são as séries extraídas do *Google Trends* em termos contemporâneos, e ε_t é o termo de erro da regressão.

Assim como no caso anterior, o modelo será ajustado utilizando apenas os dados de janeiro de 2010 a dezembro de 2018, e as projeções fora da amostra serão comparadas aos dados divulgados para janeiro a junho de 2019.

Os modelos serão então comparados por meio da significância das variáveis independentes, do coeficiente de determinação e do Critério Bayesiano de Schwarz

(BIC). Uma vez escolhido o melhor modelo, será avaliada graficamente a função de autocorrelação dos resíduos e dos resíduos ao quadrado, visando analisar a capacidade de ajustamento e a ausência de heterocedasticidade.

Na sequência, a hipótese de normalidade dos resíduos será verificada primeiro visualmente, pelo histograma da sua distribuição, e posteriormente por dois testes específicos. Gujarati e Porter (2011) apresentam o teste de Jarque-Bera, que calcula uma estatística com base na assimetria e curtose dos dados, e avalia a probabilidade de serem provenientes de uma distribuição normal. Entretanto, como esse teste é assintótico, será também realizado o teste de Shapiro-Wilk, que funciona bem para amostras pequenas e médias.

4.4 Projeção fora da amostra

Por fim, o modelo básico estimado será comparado ao modelo proposto, com o objetivo de entender a possível relevância de se utilizar os dados do *Google Trends* como variáveis para previsão do CAGED. Para isso, são avaliadas as projeções fora da amostra de cada modelo, isto é, para cada uma das alternativas serão geradas projeções para os seis meses posteriores (equivalente a janeiro de 2019 até junho de 2019) e os valores obtidos são comparados com os dados reais observados para o período. O mesmo método é empregado em diversos artigos semelhantes, como por exemplo em Guzman (2011), onde é usado para avaliar projeções de inflação.

Muitos indicadores são candidatos a serem utilizados para avaliar e comparar as projeções do modelo. Em Perez Albarella (2017), por exemplo, utiliza-se a raiz quadrada do erro quadrático médio (RMSE), o erro absoluto médio (MAE) e a média do erro percentual absoluto (MAPE). Visando simplificar o trabalho, será avaliado apenas o resultado obtido pelo RMSE, conforme a equação abaixo:

$$RMSE = \sqrt{\frac{\sum_1^N (f_t - o_t)^2}{N}} \qquad (3)$$

em que N é o tamanho da amostra, f_t são os valores projetados e o_t são os valores observados.

O melhor modelo, portanto, será aquele cujas projeções apresentarem um menor valor para o RMSE, o que indica melhor aderência aos dados reais e melhor capacidade preditiva.

5. Resultados

Inicialmente, é empregado o teste ADF para todas as séries de dados a serem utilizadas nos modelos. Conforme mostrado na tabela abaixo, todas as séries obtidas pelo *Google Trends* são não estacionárias, a um nível de significância de 5%.

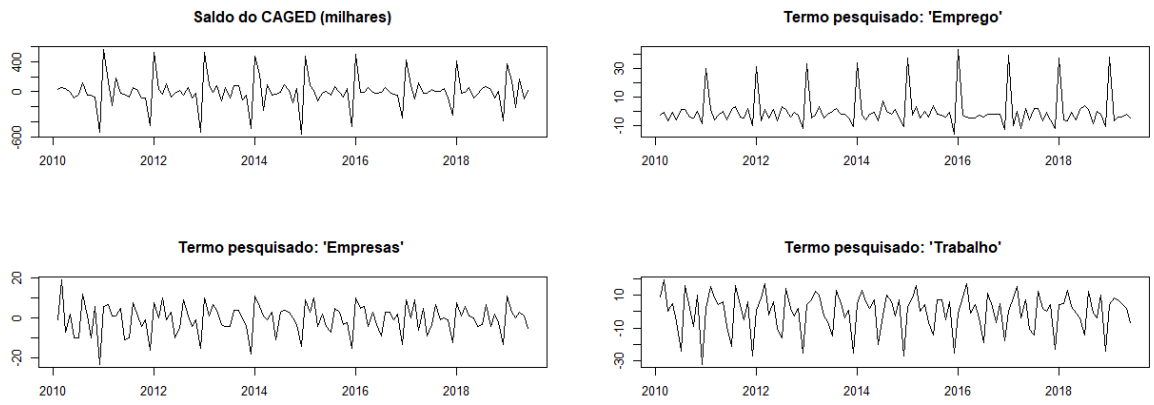
Tabela 1 – Teste ADF

	CAGED	**Emprego**	**Empresas**	**Trabalho**
Estatística	-4,00	0,48	-1,49	-0,83
P-valor	0,01	0,46	0,14	0,35

Fonte: Elaboração do autor.

Para solucionar o problema é tomada a primeira diferença de todas as séries temporais, que passam a apresentar características de estacionariedade, de acordo com o teste ADF. As séries em primeiras diferenças são apresentadas nos gráficos abaixo, e vê-se que preservam a característica sazonal citada anteriormente.

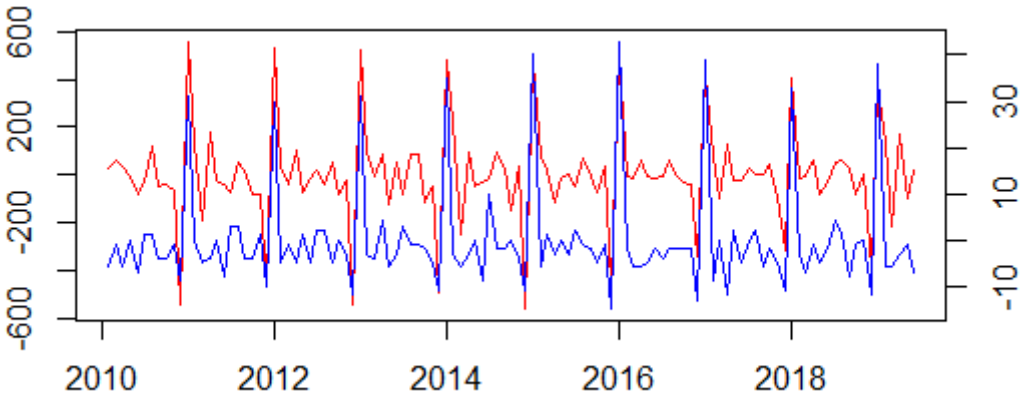
Figura 2 **– Séries do CAGED e do Google Trends**



Fonte: Ministério da Economia e Google Trends.

A busca por um modelo de previsão para a variável do CAGED se inicia pela análise visual das séries obtidas, por meio dos gráficos da Figura 1. A primeira relação que se encontra é que os dados do CAGED e o índice de pesquisas pelo termo "emprego" apresentam, aparentemente, o mesmo efeito sazonal, com fortes quedas em dezembro (explicadas anteriormente), o que fica mais claro plotando as duas séries, já tomadas em primeiras diferenças, no mesmo gráfico.

Figura 3 - CAGED (azul) e termo "emprego" (vermelho)

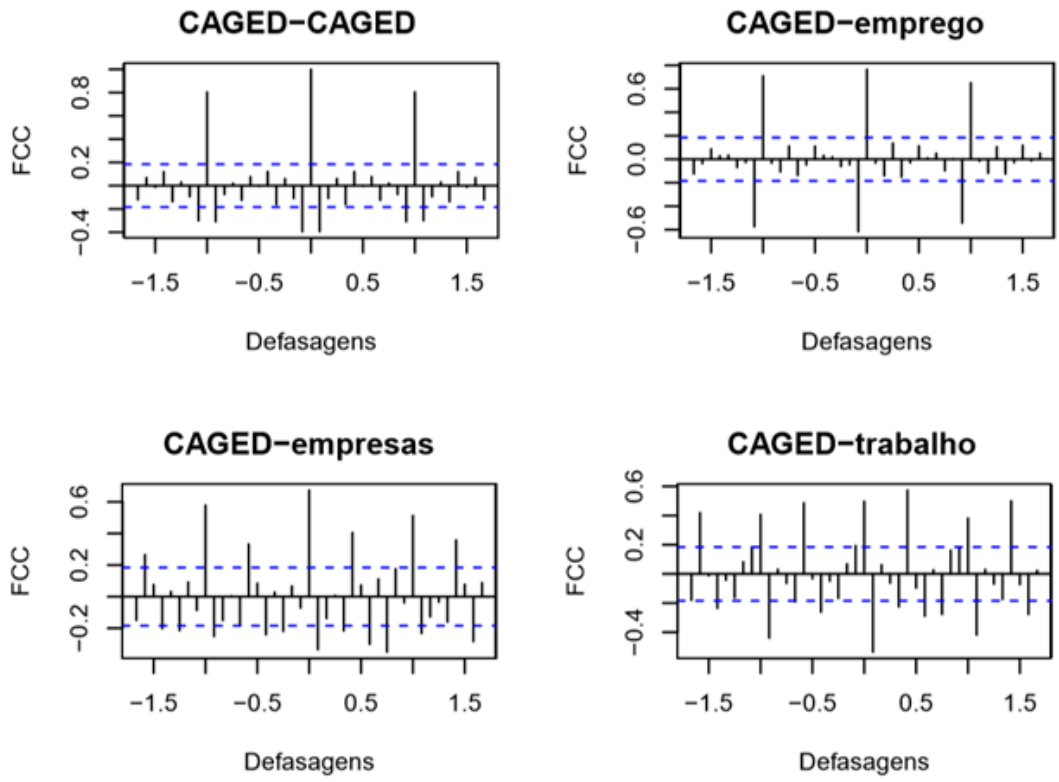


Fonte: Ministério da Economia e Google Trends.

Dessa forma, pode-se chegar a duas conclusões sobre possíveis modelos a serem testados. Primeiramente, no modelo univariado deve ser incluído um termo referente à 12ª defasagem, para incorporar o efeito sazonal apresentado. Em segundo lugar, ao menos o índice para o termo emprego deverá ter correlação contemporânea com a variável dependente e, portanto, representará uma nova informação útil na antecipação da divulgação dos dados do CAGED.

Visando dar mais evidências às conclusões anteriores, além de obter outras relações que podem ser modeladas, são traçados quatro gráficos de correlação cruzada, que avaliam a correlação existente entre os dados do CAGED sem defasagem e os índices do *Google Trends* para diversas defasagens.

Figura 4 – Função de correlação cruzada



Fonte: Ministério da Economia e Google Trends.

As funções de correlação cruzada (FCC) apresentadas possuem o número da defasagem no eixo horizontal, isto é, em zero é analisada a correlação contemporânea (sem nenhuma defasagem) entre duas variáveis, enquanto em um é analisada a décima segunda defasagem (dada a característica de sazonalidade mensal) da variável que está à direita no título do gráfico, em relação aos dados do CAGED sem defasagem. O eixo vertical apresenta o valor da correlação para cada par de séries, e os valores que superam a linha tracejada são considerados estatisticamente significantes. Dessa forma, é possível definir quais defasagens de cada variável independente possuem correlação com os dados da variável dependente. Por exemplo, pode-se afirmar que as defasagens zero, 11 e 12 do índice de buscas pelo termo "emprego" possuem correlação com os dados do CAGED, e então se espera que sejam bons regressores para os modelos a serem discutidos.

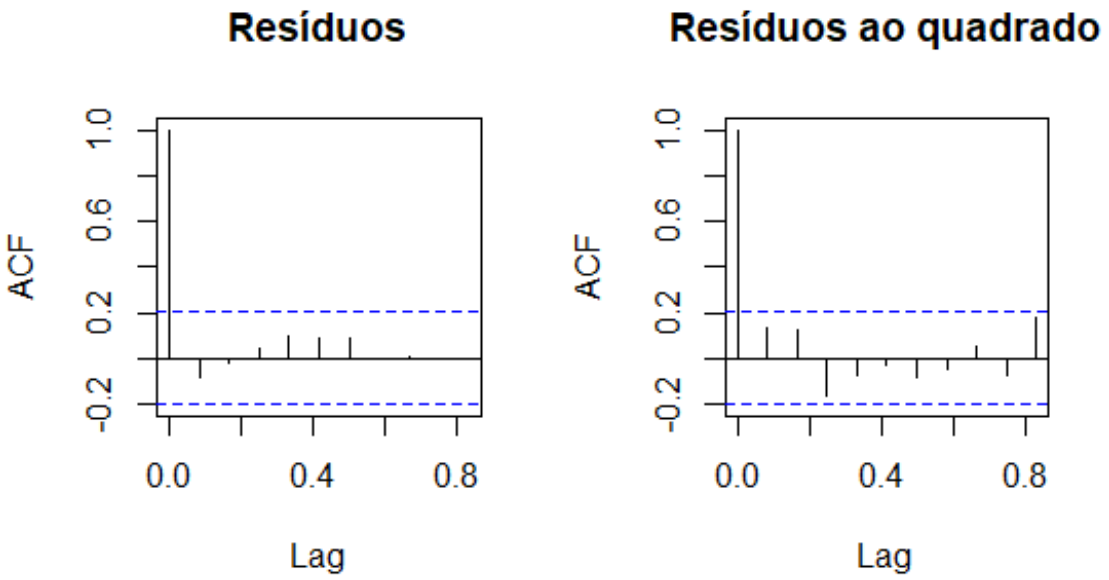
Portanto, os gráficos validam que o modelo univariado do CAGED deve sim contemplar a defasagem sazonal, mas também as 11ª e 13ª defasagens, assim

como a primeira. Também são confirmadas as expectativas de que todos os índices extraídos do *Google Trends* têm, no mínimo, correlação contemporânea com os dados do CAGED, e supostamente serão úteis nos modelos de previsão a serem testados.

Na sequência, é definido o melhor modelo puramente univariado para dados do CAGED, avaliado com base no critério de informação BIC. O modelo obtido é um SARIMA (0,0,1) (0,1,1) [12], possuindo, portanto, uma defasagem de médias móveis e a esperada defasagem sazonal.

Tanto a série de autocorrelação dos resíduos, quanto do quadrado dos resíduos, não apresentam valores significativos, evidenciando a qualidade do ajuste e ausência de heterocedasticidade.

Figura 5 – Resíduos do modelo SARIMA



Fonte: elaboração própria.

A normalidade dos resíduos é testada por meio da análise visual do histograma e também pelos testes de Jarque-Bera e Shapiro-Wilk, anteriormente citados. Os resultados dos dois testes, que não rejeitam a hipótese de normalidade, são expostos na Tabela 2.

Tabela 2 **– Testes de normalidade (SARIMA)**

TESTE	Estatística	P-valor
Jarque-Bera	0,3610	0,8348
Shapiro-Wilk	0,9868	0,4704

Fonte: elaboração própria.

Posteriormente, foram testados três novos modelos incorporando as variáveis do Google. Todos os três modelos utilizam as defasagens da variável dependente definidas como relevantes pela função de correlação cruzada, além dos índices do Google em termos contemporâneos incrementados gradativamente em cada modelo.

Tabela 3 **– Resultado das regressões, variáveis em primeira diferença**

	CAGED (1)	CAGED (2)	CAGED (3)
1ª def CAGED	-0,525*** (0,088)	-0,528*** (0,088)	-0,512*** (0,088)
11ª def CAGED	-0,025 (0,040)	-0,019 (0,040)	0,010 (0,045)
12ª def CAGED	0,777*** (0,066)	0,753*** (0,068)	0,749*** (0,068)
13ª def CAGED	0,513*** (0,084)	0,494*** (0,085)	0,474*** (0,086)
Índice “emprego”	2,893** (1,299)	2,329* (1,374)	3,347** (1,552)
Índice “empresas”		1,846 (1,493)	-2,388 (3,403)
Índice “trabalho”			2,618 (1,893)
constante	-0,810	-0,429	-1,065 (6,750)
Observações	94	94	94
R^2	0,898	0,900	0,902
R^2 ajustado	0,892	0,893	0,894
Desv Pad Res	65,682	65,485	65,145
Estatística F	154,681***	129,929***	112,810***
	*p<0,1	**p<0,05	***p<0,01

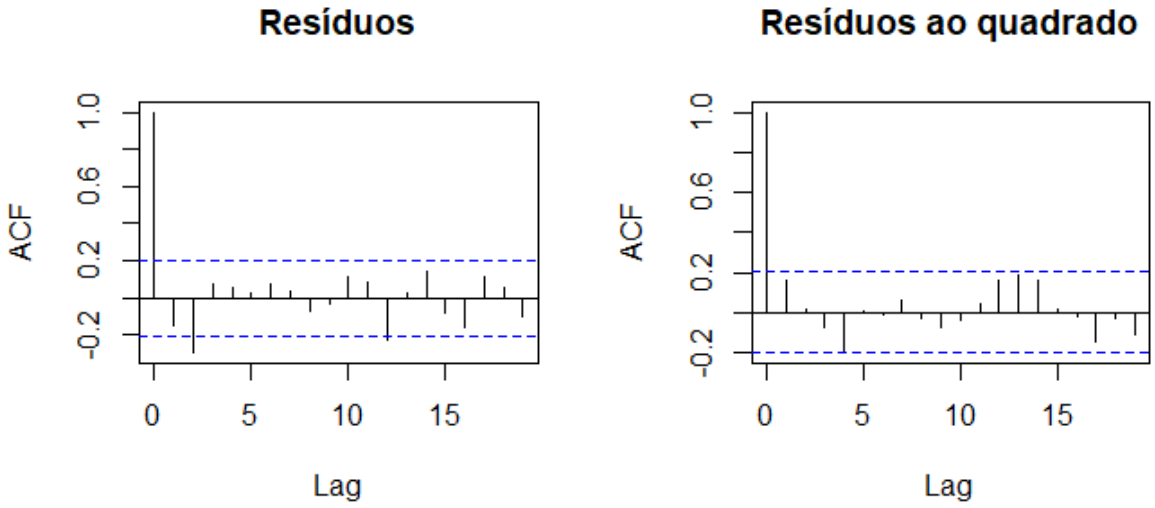
Fonte: elaboração própria.

Observa-se na Tabela 3 que apenas o índice gerado pelo termo “emprego” possui correlação estatisticamente significante com o CAGED, com o esperado sinal positivo, sugerindo o uso do primeiro modelo. Além disso, o critério de informação BIC possui o menor valor para o modelo 1, sugerindo que este se adequa melhor aos dados observados, enquanto os três coeficientes de

determinação ajustados possuem valores aproximadamente iguais. Portanto, o primeiro modelo é escolhido para projetar os dados para os seis meses posteriores.

Conforme feito para o modelo SARIMA, são traçadas as funções de autocorrelação dos resíduos e dos resíduos ao quadrado, além de realizados os mesmos testes de normalidade, que atestam a qualidade do ajuste e normalidade dos resíduos.

Figura 6 – Resíduos do modelo proposto



Fonte: elaboração própria.

Tabela 4 – Testes de normalidade (modelo proposto)

TESTE	Estatística	P-valor
Jarque-Bera	0,3318	0,8472
Shapiro-Wilk	0,9915	0,8198

Fonte: elaboração própria.

Por fim, estimados os dois modelos e utilizando os índices obtidos do *Google Trends* para o período de janeiro a junho de 2019, podemos obter as previsões fora da amostra de cada alternativa e compará-las por meio do método RMSE. Os resultados obtidos são mostrados na tabela abaixo. A primeira linha mostra a previsão feita pelo modelo básico (univariado) para cada mês, enquanto a segunda mostra as projeções obtidas pelo modelo proposto (que incorpora as variáveis do Google). A última linha mostra os valores reais observados no

período, para fins de comparação, enquanto a última coluna mostra o resultado do RMSE para as projeções de cada modelo.

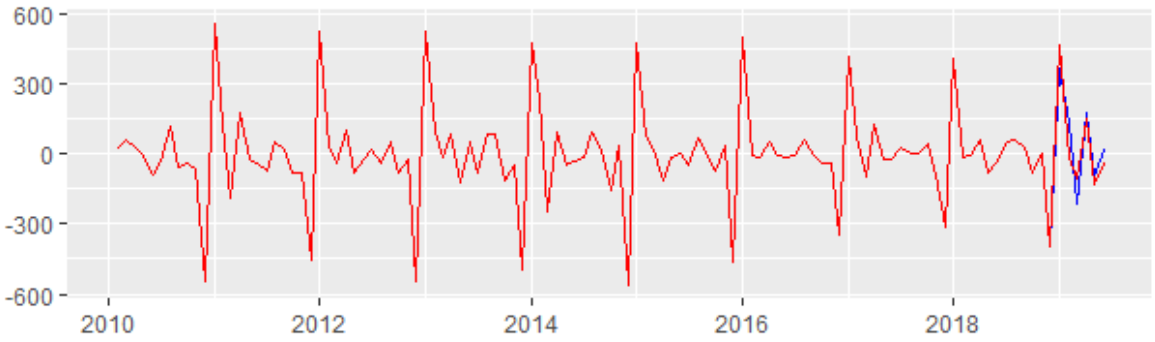
Tabela 5 – Valores projetados e realizado

Modelo	Jan/19	Fev/19	Mar/19	Abr/19	Mai/19	Jun/19	RMSE
SARIMA	437	27	-38	66	-53	-26	103
Proposto	469	-16	-105	150	-126	-40	92
Realizado	369	139	-216	173	-98	16	-

Fonte: Ministério da Economia e dados projetados.

Pela tabela, vê-se que o modelo que incorpora o índice de buscas no Google pelo termo "emprego" apresenta um menor RMSE, o que sugere que suas projeções encontram maior aderência com os dados reais divulgados pelo Ministério da Economia, em relação ao modelo puramente univariado. Perez Albarella (2017) também encontra um RMSE ligeiramente mais baixo ao incorporar os dados Google em algumas projeções de variáveis econômicas. As projeções do modelo proposto são mostradas abaixo (em azul), juntamente dos dados reais do CAGED (em vermelho).

Figura 7 – Projeção do modelo proposto



Fonte: Ministério da Economia e dados projetados.

Conforme se vê, a projeção do modelo proposto possui uma aderência substancial ao dado real observado no período, evidenciando a utilidade de se incorporar dados do *Google Trends* em modelos econômicos com finalidade preditiva.

6. Conclusão

Nesse trabalho, buscou-se avaliar a correlação entre as séries de pesquisas do Google e os dados de emprego disponibilizados pelo CAGED, assim como verificar a utilidade de se usar os dados do buscador (de maior frequência) como variável independente em modelos de regressão, buscando antever o indicador econômico e permitir melhores análises prospectivas da conjuntura econômica de um país.

Para isso, foram feitas análises gráficas das séries e de funções de correlação cruzada, além da comparação entre um modelo univariado SARIMA e outro modelo ARIMAX, este último incorporando os dados do Google como variáveis exógenas, para projeções fora da amostra, por meio da mensuração do desvio em relação ao dado real.

Os resultados obtidos, que corroboram os expostos em Guimarães Filho (2017), permitem constatar que há evidências adicionais de que as pesquisas feitas no Google por termos específicos (como "emprego") são úteis para se conhecer antecipadamente e fazer previsões quanto a evolução de alguns indicadores econômicos, como os do mercado de trabalho brasileiro, no período entre 2010 e 2019.

Dessa forma, sugere-se que diversos indicadores de alta frequência, além da maior tempestividade, podem também incorporar informações adicionais que refletem o comportamento dos agentes econômicos em tempo real e, portanto, apesar de não substituírem os modelos de projeção tradicionais, podem constituir novas variáveis que melhorem a capacidade preditiva de alguns modelos econômicos.

Uma possível extensão desse trabalho poderia buscar evidenciar se há causalidade estatística entre as séries do *Google Trends* e dados econômicos diversos. Ademais, pode-se comparar o modelo proposto não somente aos univariados, mas também a modelos multivariados com fundamentação macroeconômica, fornecendo assim mais evidências de que as pesquisas na internet são novas variáveis de importante uso na ciência econômica.

Referências

ASKITAS, N.; ZIMMERMANN, K. F. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, v. 55, n. 2, p. 107-120, 2009.

BRASIL. Ministério do Trabalho. Programa de disseminação das estatísticas do trabalho. Disponível em: <http://pdet.mte.gov.br/caged>.

BUENO, R. L. S. *Econometria de séries temporais*. São Paulo: Cengage Learning, 2011.

CHADWICK, M. G.; SENGUL, G. Nowcasting unemployment rate in Turkey: Let's Ask Google. *Central Bank Review*, v. 15, n. 3, p. 15-40, 2015.

CHOI, H.; VARIAN, H. Predicting the present with Google Trends. *Economic Record*, v. 88, n. s1, p. 2–9, 2012.

D'AMURI, F.; MARCUCCI, J. Google it! Forecasting the US unemployment rate with a Google job search index. *Working Paper 31.2010 Fondazione Eni Enrico Mattei*, 2010.

EDELMAN, B. Using internet data for economic research. *Journal of Economic Perspectives*, v. 26, n. 2, p. 189-206, 2012.

EINAV, L.; LEVIN, J. The data revolution and economic analysis. *Innovation Policy and the Economy*, v. 14, p. 1-24, 2014.

GUIMARÃES FILHO, S. Google Trends para previsão de variáveis macro: uso no Brasil através do algoritmo autometrics. 2017. 37 p. Dissertação (Mestrado em Economia) – Fundação Getúlio Vargas, São Paulo, 2017.

GUJARATI, D. N.; PORTER, D. C. *Econometria básica*. Porto Alegre: McGrawHill/Bookman, 2011.

GUZMAN, G. Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*, v. 36, n. 3, p. 119-167, 2011.

INTERNET LIVE STATS. Google Search Statistics. Disponível em: <https://www.internetlivestats.com/google-search-statistics/>.

KOOP, G.; ONORANTE, L. Macroeconomic nowcasting using Google probabilities. *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A*, v. 40, p. 17-40, 2019.

MORETTIN, P. A. *Econometria financeira: um curso em séries temporais financeiras*. São Paulo: Edgard Blucher, 2011.

PEREZ ALBARELLA, N. Uso de dados de busca na internet na estimação de indicadores econômicos. 2017. 48 p. Dissertação (Mestrado em Economia) – Insper Instituto de Ensino e Pesquisa, São Paulo, 2017.

PERLIN, M.; PORTELA, A.; CALDEIRA, J.; PONTUSCHKA, M. Can we predict the financial markets based on Google's search queries? *Journal of Forecasting*, v. 36, n. 4, p. 454-467, 2016.

THE WORLD BANK. Individuals using the Internet (% of population). Disponível em: <https://data.worldbank.org/indicator/IT.NET.USER.ZS>.

TUHKURI, J. Forecasting unemployment with Google Searches. *ETLA Working Papers, n. 35*, 2016.

Anexo

Obtenção dos dados do Google Trends

```
library(gtrendsR)

library(dplyr)

library(lubridate)

library(tseries)

inicio='2010-01-01'

fim='2019-06-01'

itens<-c("emprego","empresas","trabalho")

for(i in length(itens)){

  g_aux<-gtrends(itens[i], geo = "BR", time = "2010-01-01 2019-06-01")

  g_aux<-g_aux$interest_over_time %>% mutate(date = format(date,"%Y-%m")) 
%>%

    group_by(date) %>% summarise(hits = max(hits))

  aux<-ts(g_aux$hits, start = c(year(inicio),month(inicio)),

          end = c(year(fim),month(fim)), frequency = 12)}
```