

**INSPER**

**MESTRADO PROFISSIONAL EM ECONOMIA**

**NATACHA PEREZ ALBARELLA**

**USO DE DADOS DE BUSCA NA INTERNET NA ESTIMAÇÃO DE INDICADORES  
ECONÔMICOS**

**São Paulo**

**2017**

**NATACHA PEREZ ALABRELLA**

**USO DE DADOS DE BUSCA NA INTERNET NA ESTIMAÇÃO DE INDICADORES  
ECONÔMICOS**

*Dissertação* apresentada ao Insper, como parte dos requisitos para a obtenção do título de mestre em Economia.

Área de Concentração: Macroeconomia

Orientador: Prof. Dr. Rinaldo Artes

**São Paulo**

**2017**

Albarella, Natacha Perez

Uso de dados de busca na internet na estimação de indicadores econômicos /Natacha Perez Albarella; orientador: Rinaldo Artes – São Paulo: Insper, 2018. 41f. Dissertação (Mestrado – Programa de Mestrado Profissional em Economia. Área de concentração: Macroeconomia) – Insper Instituto de Ensino e Pesquisa.

## FOLHA DE APROVAÇÃO

Natacha Perez Albarella

Uso de dados de busca na internet na estimação de indicadores econômicos

Dissertação apresentada ao Programa de Mestrado Profissional em Economia do Insper Instituto de Ensino e Pesquisa, como requisito para obtenção do título de Mestre em Economia.

Área de concentração: Macroeconomia

### Banca Examinadora

Prof. Dr. Rinaldo Artes

Orientador

Instituição: Insper      Assinatura: \_\_\_\_\_

Prof. Dr. Gustavo Monteiro de Athayde

Instituição: Insper      Assinatura: \_\_\_\_\_

Dr. Luis Fernando Pereira Azevedo

Instituição: EESP/FGV      Assinatura: \_\_\_\_\_

## **AGRADECIMENTOS**

Ao meu orientador, por sua paciência, dedicação e profissionalismo.

Aos meus pais, por ensinarem a priorizar e dar valor aos estudos, pela dedicação e amor.

Aos meus amigos, que apesar de minha ausência, sempre me apoiaram e estiveram ao meu lado.

A Isabel, por me ajudar a acreditar em mim.

Ao meu marido, pelo amor, apoio e paciência.

## RESUMO

ALBARELLA, Natacha Perez. **Uso de dados de busca na internet na estimação de indicadores econômicos**, 2017. Dissertação de Mestrado – Insper Instituto de Ensino e Pesquisa, São Paulo, 2017.

Este trabalho visa a analisar se o uso de dados de busca na internet como variável em modelos consagrados para projeção de variáveis macroeconômicas é capaz de melhorar seu poder preditivo. Partindo da lei de Okun, Curva de Phillips e Curva IS, a taxa de desemprego, de inflação e o crescimento do PIB no Brasil serão estimados e projetados. As séries de busca na internet obtidas no Google Trends são sumarizadas através da análise de componentes principais e a primeira componente principal de cada tema é incluída nos modelos originais. A amostra está em frequência trimestral compreende o período de 2004 a 2017. No caso da taxa de desemprego, o modelo aumentado apresentou aderência maior aos dados do que o modelo original, porém seu poder preditivo quase não se alterou. Com relação aos modelos de inflação o desempenho preditivo melhorou. Já os modelos de crescimento do produto pioraram em aderência e poder de projeção.

Palavras-chave: big data, Inflação, Desemprego, PIB, Google, internet

## ABSTRACT

ALBARELLA, Natacha Perez. **Using Internet search data for estimating economic indicators**, 2017. Masters dissertation – Insper Instituto de Ensino e Pesquisa, São Paulo, 2017.

This paper analyzes whether the use of Internet search data as a variable in consecrated forecasting macroeconomic models is able to improve its predictive power. Based on the Okun law, the Phillips curve and the IS curve, the unemployment rate, inflation rate, and GDP growth in Brazil are estimated and projected. The web search series obtained through Google Trends has been summarized through the analysis of principal components and the first component of each theme is included in the original models. The sample is in quarterly frequency and comprised the period from 2004 to 2017. In the case of the unemployment rate, the increased model presented greater adherence to the data than the original model, but its predictive power almost did not change. Regarding inflation models, the predictive performance improved. The models of GDP growth have worsened in adhesion and forecasting power.

Keywords: big data, Inflation, Unemployment, GDP, Google, Internet.

JEL Code: C22, C53, C82, E17

## SUMÁRIO EXECUTIVO

Com o avanço da internet nos últimos vinte anos, houve aumento expressivo no volume de dados armazenados. Esse volume de informação – sem precedentes na história – pode nos dizer muito sobre quem somos: medos, desejos, comportamentos e decisões que tomamos. De temas profundos à corriqueiros, é possível aprender mais sobre a psique humana hoje do que seria possível imaginar vinte anos atrás.

Dados de compras online, vídeos visualizados, mensagens de celular, dados de geolocalização, históricos médicos, postagens em redes sociais; em resumo, toda interação com a internet deixa seu registro. Há muita informação disponível, entretanto, esses dados são capazes de melhorar modelos macroeconômicos?

Investigou-se aqui o uso potencial de dados de busca na internet para prever variáveis macroeconômicas importantes, especificamente taxa de inflação, taxa de desemprego e crescimento do PIB no Brasil.

Com auxílio de uma pesquisa de opinião na internet, os termos de busca foram traçados e, através da técnica econométrica de análise de componentes principais, novas variáveis foram criadas.

Assim como sugeria a literatura, as buscas na internet podem oferecer alguma contribuição na percepção dos agentes que não pertencem ao mercado financeiro com relação à inflação, ou seja, pode atuar como uma nova fonte de informação quanto as expectativas de inflação do público e auxiliar nas projeções.

Os modelos usados na estimação e projeção de crescimento do produto e de desemprego não apresentaram resultados muito significativos, mas isso pode se dever a dificuldade de se obter os termos de busca mais apropriados para sumarizar a sensibilidade da população em relação a um tema específico. Uma combinação diferente de termos de busca poderia gerar resultados totalmente distintos dos apresentados aqui.

O uso de dados de busca na internet pode ser vantajoso uma vez que o Google Trends é atualizado diariamente. Essa nova fonte de dados é flexível e pode ser convertida facilmente em séries de frequência mais baixa para uso em modelos tradicionais. No entanto, tentou-se demonstrar aqui que as séries de busca de dado podem melhorar o desempenho de alguns modelos, mas não substituir as estatísticas oficiais.

Existem ainda inúmeras aplicações potenciais em economia e finanças para tal fonte, como *nowcasting* e projeções outras variáveis macroeconômicas e de retorno no mercado de commodities e mercado de ações.

## Sumário

1. INTRODUÇÃO .....	11
2. REVISÃO BIBLIOGRÁFICA.....	13
3. METODOLOGIA .....	16
4. RESULTADOS.....	25
5. CONCLUSÃO .....	34
6. REFERÊNCIA BIBLIOGRÁFICA .....	36
APÊNDICE .....	38

## 1. INTRODUÇÃO

Com o avanço da internet nos últimos vinte anos, houve aumento expressivo no volume de dados armazenados. De fato, em 2014, o volume de dados disponíveis era de 1 septilhão ( $1 \times 10^{24}$ ). Estimava-se, então, que até 2020 esse número terá se multiplicado por 6<sup>1</sup>. O número de internautas no mundo era de 3,2 bilhões em 2015, segundo a União Internacional das Telecomunicações, órgão vinculado à Organização das Nações Unidas. Segundo o IBGE, em 2015, 58% da população brasileira tinha acesso à internet – representando 102 milhões de internautas.

Esse volume de informação – sem precedentes na história – pode nos dizer muito sobre quem somos: medos, desejos, comportamentos e decisões que tomamos. De temas profundos à corriqueiros, é possível aprender mais sobre a psique humana hoje do que seria possível imaginar vinte anos atrás.

A internet é em si uma fonte de dados que tem o potencial de alterar a forma como economistas interagem com os dados, ao reduzir o custo de armazenamento, atualização e distribuição, aumentando a informação disponível com custos reduzidos (EDELMAN, 2012).

Quando se realiza uma busca na internet, por exemplo através do Google, o termo pesquisado é armazenado e usado para direcionar propagandas relacionadas ao termo da busca. Dados de compras online, vídeos visualizados, mensagens de celular, dados de geolocalização, históricos médicos, postagens em redes sociais; em resumo, toda interação com a internet deixa seu registro.

Há muita informação disponível, entretanto, esses dados são capazes de melhorar modelos macroeconômicos? Aqui, investiga-se o uso potencial de dados de busca na internet para prever variáveis macroeconômicas importantes, especificamente taxa de inflação, taxa de desemprego e crescimento do PIB no Brasil.

A hipótese embutida na análise é de que as pessoas revelam informações úteis sobre suas necessidades, desejos, interesses e preocupações através de seu comportamento na internet; os termos buscados em sites de busca refletem essas informações.

Com o auxílio de modelos tradicionais de estimação das variáveis macroeconômicas de interesse, pretende-se estimar os modelos originais e versões aumentadas, que incluem variáveis que sintetizam os termos de busca na internet, e espera-se que estes modelos tenham performance superior na projeção fora da amostra do que os modelos tradicionais.

---

<sup>1</sup> <https://oglobo.globo.com/sociedade/tecnologia/estudo-da-emc-preve-que-volume-de-dados-virtuais-armazenados-sera-seis-vezes-maior-em-2020-12147682>

Na primeira seção apresenta-se a teoria que embasa o uso de dados de busca na internet em modelos econômicos. Na segunda seção é discutida a forma de obtenção dos dados e o tratamento recebido para utilização em estimações econômicas. Na terceira parte é apresentada a análise empírica e os resultados obtidos. A quarta seção traz a conclusão.

## 2. REVISÃO BIBLIOGRÁFICA

Segundo (WARD; BARKER, 2013), todo o trabalho de analisar e armazenar um grande volume de informação é conhecido como *big data*. Einav e Levin (2014) afirmam que o uso de *Big Data* é mais comum para fins de estimação de modelos preditivos. Empresas como Amazon e Netflix, por exemplo, possuem modelos de previsão sobre quais livros ou filmes um indivíduo pode se interessar (e adquirir). Um outro exemplo são os produtos da Apple, que contam com o recurso de auto completar frases com base em padrões de uso.

Igboayaka (2015) afirma que redes sociais (como Facebook e Twitter) são usadas para expressar opiniões, um estado de espírito ou perspectivas que podem ser mineradas para se obter os sentimentos dos usuários em relação à situação econômica. De acordo com dados do CONECTA, comunidade online de pesquisa, a principal finalidade de acesso na internet no Brasil entre jovens de 15 a 32 anos é a conexão com redes sociais. O segundo uso mais comum da internet é a busca de informações.

O website “google.com” é o site de busca mais popular do mundo, com uma participação de mercado de 78% (para celulares e tablets, esta proporção é de 95%) em junho de 2017<sup>2</sup>. O Google Trends é a ferramenta do Google que sumariza as estatísticas de busca.

Artigos recentes têm investigado a utilização da ferramenta para melhorar estimativas de variáveis macroeconômicas. Uma das principais motivações para o uso de dados de busca de internet é sua alta frequência. Indicadores de atividade econômica podem ser divulgados com defasagem de semanas e até meses de sua ocorrência.

Utilizando dados obtidos pelo Google Trends, Choi e Varian (2012) conseguem estimar o comportamento de atividade real como vendas no varejo, vendas de automóveis, vendas de residências e viagens. Em um trabalho anterior<sup>3</sup>, os mesmos autores usam dados da mesma ferramenta para prever o estado corrente do mercado de trabalho norte americano, estimando variáveis como pedidos de auxílio desemprego e a taxa de desocupação da economia.

Askitas e Zimmermann (2009) encontraram uma correlação forte entre um índice da atividade criado através do Google Trends e a taxa de desemprego na Alemanha. Koop e Onorante (2013) também utilizam dados do Google Trends para

---

2

<sup>3</sup> (CHOI e VARIAN, 2009)

capturar mudanças estruturais no comportamento de variáveis macroeconômicas convencionais, como emprego, inflação e dados de produção.

Com foco em estimar expectativas de inflação, Guzman (2011) afirma que as buscas na internet podem ser interpretadas como uma medida de expectativas reveladas, uma vez que as pessoas pesquisam por tópicos sobre os quais desejam aprender mais, ou ainda assuntos com os quais estejam mais preocupadas. Um bom exemplo disso é um indivíduo que não está se sentindo bem recorrer à ferramenta de busca na internet e procurar por seus sintomas a fim de descobrir se pode ter uma doença específica. Da mesma forma, uma pessoa preocupada com suas finanças e que acredita que está vivendo um momento de alta generalizada nos preços, pode usar a internet para buscas sobre a inflação.

Stock e Watson (2001) afirmavam que o trabalho de todo macroeconomista se resume a quatro tarefas: descrever e analisar dados econômicos, quantificar o que se sabe sobre a estrutura macroeconômica, aconselhar a autoridade monetária e fazer projeções. Neste último, o teste final para o um modelo econométrico é seu desempenho fora da amostra<sup>4</sup>.

Uma projeção fora da amostra simulada, divide a amostra usando na estimação do modelo apenas parte dela. Gera-se projeção um passo à frente e se reestima o modelo a cada passo, simulando como seriam as projeções em tempo real, mas sendo conduzido de forma retrospectiva, usando dados reais.

De fato, Guzman (2011) mostra que o índice gerado a partir do Google Trends tem um desempenho fora da amostra melhor do que os indicadores de expectativas tradicionais para o mercado norte-americano.

A justificativa para essa melhora é o fato do indicador de buscas na internet ter frequência de divulgação mais alta (pode chegar a ser *real time* com o uso de *feeds*). Conclui-se ainda que mídias sociais podem ajudar na busca por ancoragem de expectativas de inflação por parte do Federal Reserve por facilitar a “profecia autorrealizável”.

O Brasil, por ter uma população adulta que viveu um período de hiperinflação recentemente, é um bom candidato para criação de um índice de expectativas baseado em buscas na internet, uma vez que tal população deve apresentar uma preocupação maior com o assunto. Dessa forma, assim como em Guzman (2011), um aumento nas buscas de internet por termos relacionados a “inflação” pode ser interpretado como um aumento na preocupação com o tema.

Apesar de o Banco Central divulgar em alta frequência (semanal) o relatório Focus com as expectativas para diversos indicadores macroeconômicos, esse relatório leva em conta apenas a percepção dos agentes do mercado financeiro. Ao

---

<sup>4</sup> “the ultimate test of a forecasting model is its out-of-sample performance” (STOCK e WATSON, 2001).

passo que as buscas na internet forneceriam informações sobre os assuntos mais presentes na mente do público em geral.

No Brasil, Azevedo (2017) busca, através de notícias e do Google Trends captar a variação de sentimentos de agentes e averiguar se há relação entre estas medidas e variações dos ciclos econômicos e na precificação de ativos.

Seguindo o trabalho de Koop e Onorante (2013), pretende-se encontrar evidências de que dados de busca na internet ajudam a melhorar estimativas e projeções de curto prazo de algumas variáveis macroeconômicas no Brasil, quais sejam, inflação, taxa de desemprego e PIB.

### 3. METODOLOGIA

Assume-se aqui, como em Azevedo (2017), que os sites de busca podem fornecer informação precisa sobre mudanças no estado da economia, em particular das pessoas. Aqui usaremos estatísticas de busca da empresa Google, o Google Trends é a ferramenta que sumariza as estatísticas de busca. A ferramenta apresenta um índice do volume de pesquisa de determinado termo no site de busca da empresa, e permite que seus resultados sejam discriminados por localização. As séries de busca estão disponíveis com dados desde 2004.

A Tabela 1 apresenta as variáveis que serão estimadas, e projetadas, na frequência trimestral, para o horizonte de um trimestre à frente. Todos as variáveis são obtidas no banco de dados do IBGE.

**Tabela 1: Variáveis dependentes**

	<b>Variável</b>	<b>Fonte</b>
Inflação	Índice de Preços ao Consumidor Amplo (IPCA), todos os itens	IBGE
Taxa de desemprego	Taxa de desocupação, nacional	IBGE
PIB	Produto Interno Bruto	IBGE

Para projeção das variáveis dependentes estimaremos um modelo convencional, apenas com variáveis explicativas oficiais. Um segundo modelo, aumentado, será estimado com uma variável explicativa a mais, aquela obtida através da ferramenta Google Trends.

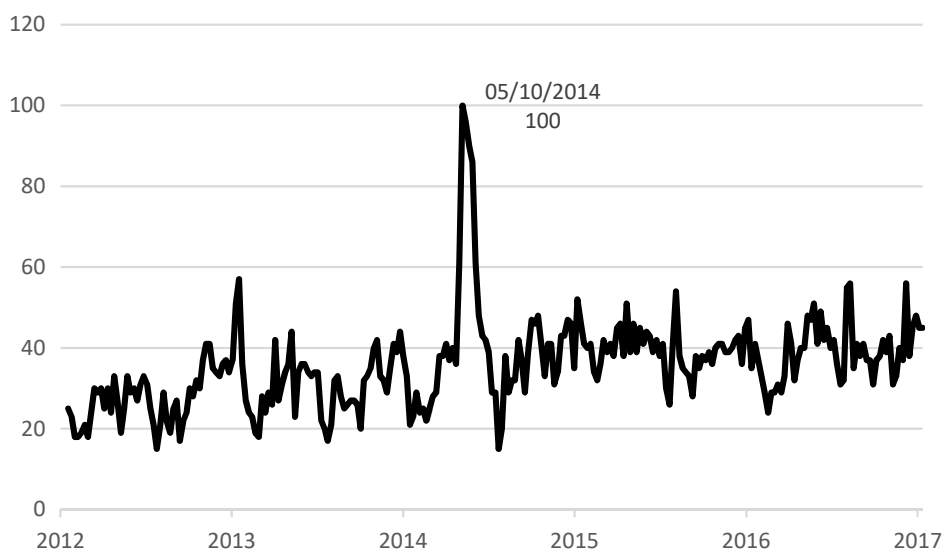
#### 3.1. O Google Trends

Ao se realizar uma busca sobre um termo no site Google Trends<sup>5</sup>, o resultado apresentado é um gráfico. A série apresentada no gráfico é normalizada entre 0 e 100, os números representam o interesse relativo pelo tópico pesquisado, comparado a todos outros assuntos para uma dada região e data. Assim, o momento em que houve maior volume de buscas sobre o termo pesquisado assume o valor 100 e os demais pontos da série são obtidos através da divisão do volume daquele momento pelo de maior procura e depois multiplicado por 100. A frequência dos dados apresentados pode chegar até minuto a minuto. É possível fazer o download da série obtida.

<sup>5</sup> <https://trends.google.com/trends/?hl=en>

Por exemplo, quando se deseja saber qual é o histórico de busca por inflação, no site do Google, no Brasil, nos últimos 5 anos<sup>6</sup>, obtemos o resultado presente no Gráfico 1. O momento de maior interesse no período pelo tema “inflação”, e consequentemente momento de maior volume de buscas no site, ocorreu na semana de 05/10/2014. Nesse momento, o índice apresentado assumiu o valor 100. Nos demais pontos da série, o volume de buscas em cada semana é dividido pelo volume ocorrido em 05/10/2015 e multiplicado por 100, gerando a série que observamos.

**Gráfico 1 – Resultado de buscas pelo termo “inflação” no site do Google, no Brasil, nos últimos 5 anos – dados semanais**



Ainda na mesma página que apresenta os resultados de buscas pelo tema escolhido, é possível obter uma lista de “consultas relacionadas”. Usuários que pesquisaram o tema escolhido, no exemplo “inflação”, também pesquisaram por temas presentes nessa lista. No exemplo anterior, as duas principais consultas relacionadas são “o que é inflação?” e “inflação no Brasil”. A ferramenta disponibiliza as 25 buscas mais relacionadas com o tema de interesse.

Uma outra forma de extrair os dados de busca no site do Google é através de agregações automáticas de categorias. Assim, ao invés de buscar especificamente por uma palavra, é possível buscar por entretenimento, esportes, negócios, etc. e descobrir quais são os tópicos mais buscados relacionados à agregação escolhida. Por exemplo, a busca pelo tópico “desemprego” traz como tópico correlato “Reforma da Previdência”.

<sup>6</sup> <https://trends.google.com/trends/explore?geo=BR&q=infla%C3%A7%C3%A3o>

Aqui pretende-se extrair do Google Trends dados referentes a buscas pelas variáveis apresentadas na Tabela 1 desde o início do banco de dados, em janeiro de 2004, até junho de 2017.

### 3.2. Seleção de termos

Existem muitos termos de busca que podem ser usados para a finalidade deste trabalho. Da mesma forma, há tratamentos diferenciados na literatura para obtenção dos mesmos. A fim de encontrar a maior variedade de termos relacionados aos temas de interesses, quais sejam, nível de atividade econômica, inflação e desemprego, foi realizada uma enquete com internautas sobre os temas, na qual os respondentes deveriam indicar quais termos de busca digitariam no Google caso houvesse interesse pelos temas mencionados. Durante o processo de criação das perguntas, houve preocupação em não usar o termo de interesse para que os entrevistados não fossem influenciados pelas palavras em si. O campo de resposta era livre e os respondentes foram instruídos a se imaginar em cada uma das situações e simplesmente responder da mesma forma que escreveriam no site de busca. As perguntas estão presentes na Tabela 2.

**Tabela 2 – Questionário da pesquisa de termos usados em buscas na internet**

Tema de interesse	Pergunta
Inflação	Se você sentisse que os preços em geral estão muitos altos ou baixos e fosse pesquisar por algum termo no Google, qual seria?
Desemprego	Se você estivesse procurando por um emprego novo, quais termos buscaria no Google?
PIB	Se você sentisse que o país está indo bem (ou mal), com relação à economia, quais termos buscaria no Google?

A pesquisa foi lançada nas redes sociais no dia 02 de agosto de 2017, com o auxílio da plataforma gratuita Survey Monkey<sup>7</sup>. A pesquisa teve um caráter puramente exploratório, com a finalidade de obter ideias e expandir o horizonte de termos que pessoas possam buscar no dia-a-dia. Não houve, portanto, a intenção de encontrar uma amostra representativa da população brasileira, como por região do país, sexo, idade ou grau de instrução. Houve participação de 48 internautas, sendo que alguns deles responderam às questões da mesma forma. Por exemplo, na pergunta

<sup>7</sup> <https://pt.surveymonkey.com/>

relacionada à desemprego mais de um internauta afirmou que buscaria pelo termo "vaga de emprego".

A partir das repostas obtidas, assim como em Koop e Onorante (2013), foram usados também os termos mais populares relacionados à busca – as "consultas relacionadas", conforme reportado pelo Google Trends. As respostas e termos relacionados estão presentes nas Tabela 12, Tabela 13 e Tabela 14<sup>8</sup>.

Com relação ao desemprego, foram excluídas respostas muito específicas e pessoais que não seriam relacionadas ao nível agregado de desemprego, por exemplo, uma das repostas foi "vaga de gerente". Neste caso, a resposta foi substituída apenas pela palavra "vaga". Foram omitidas buscas relacionadas no caso em que estas apresentaram resultados iguais a anteriores ou ainda iguais a uma resposta do questionário. No total, foram obtidas 58 séries de buscas para o desemprego.

Para as respostas relacionadas a inflação, mais uma vez, alguns termos de busca foram excluídos por serem muito específicos e não serem listados entre as buscas mais recorrentes dentro do Google Trends, por exemplo a busca por "oscilações de preços". Algumas palavras apresentaram tópicos relacionados que não faziam sentido para o propósito aqui apresentado, por exemplo, a resposta "carestia" trouxe como temas relacionados os termos "locusta" e "inseto" que não se relacionam ao tema "inflação" e, por isso, foram excluídos. Foram obtidas assim, 45 séries relacionadas à inflação.

Com relação ao crescimento do PIB, foram obtidas 43 séries.

De posse dos grupos de séries de termos selecionados, será aplicada análise de componentes principais para cada grupo. A primeira componente de cada grupo será, então, analisada e utilizada na estimação dos modelos econométricos.

### **3.3. Modelos**

Nesta seção apresentam-se os modelos que serão estimados para economia brasileira tendo por objetivo final projetar as variáveis de interesse.

#### **3.3.1. Taxa de desemprego**

A existência de uma relação negativa entre atividade econômica e a mudança no emprego foi proposta por Okun (1962). O aumento do produto determinaria a

---

<sup>8</sup> VIDE APÊNDICE.

evolução nos níveis de emprego, uma vez que as empresas precisariam aumentar o número de trabalhadores para elevar a produção (BLANCHARD, 2011).

A Equação (1), que representa a lei de Okun, pode ser representada da seguinte forma:

$$u_t - u_{t-1} = -\theta(g_{yt} - g_y) + v_t \quad (1)$$

onde,  $u_t$  é a taxa de desemprego no período  $t$ ,  $\theta$  é um parâmetro positivo,  $g_{yt}$  é a taxa de crescimento no período  $t$ ;  $g_y$  é o produto potencial e  $v_t$  é o resíduo. Se  $g_{yt} > g_y$ , ou seja, a economia apresenta crescimento acima do potencial, o desemprego tende a cair.

A Tabela 3 apresenta um resumo das variáveis que serão usadas na tentativa de estimar a equação de Okun para economia brasileira. Os dados, em frequência trimestral, abrangem o período do primeiro trimestre de 2012 ao segundo trimestre de 2017.

**Tabela 3 - Sumário de variáveis, fonte de dados e sinais esperados na estimação da Lei de Okun**

<b>Painel A: variável dependente</b>	
Variável	Definição/fonte
Taxa de desemprego encadeada	Taxa de desocupação da PNAD trimestral a partir do 1º trimestre de 2012, calculada pelo IBGE. A série foi dessazonalizada e usada em primeira diferença.
<b>Painel B: variáveis explicativas</b>	
Variáveis	Definição/fonte
Hiato do produto	Hiato calculado como a diferença logarítmica entre produto trimestral e o produto potencial. A série trimestral dessazonalizada do produto é calculada pelo IBGE. O produto potencial foi estimado por filtro HP. Espera-se sinal negativo.
PC1_ur	Primeira componente principal das séries de busca na internet relacionadas ao desemprego, usada em primeira diferença. Espera-se sinal positivo.

### 3.3.2. Inflação

A curva de Phillips está entre os instrumentos mais utilizados entre banqueiros centrais e acadêmicos para projeção da taxa de inflação. A relação estabelecida pelo economista neozelandês William Phillips retrata o comportamento da inflação em função do desemprego, ou outras medidas do nível de atividade econômica.

O Banco Central do Brasil divulgou no Relatório de Inflação do segundo trimestre de 2017 uma revisão se seu modelo agregado de pequeno porte. No boxe do referido relatório é apresentada curva de Phillips utilizada e reproduzida aqui na equação (2):

$$\pi_t^L = \sum_{i>0} a_{1i} E_t \pi_{t+1} + \sum_{j>0} a_{2j} \pi_{t-j} + \sum_{k>0} a_{3k} \pi_{t-k}^* + \sum_{l>0} a_{4l} h_{t-l} + \sum_{n \geq 0} \sum_p a_{5n}^p Z_{t-n}^{\pi,p} + \varepsilon_t \quad (2)$$

onde  $\pi_t^L$  é a inflação de preços livres;  $E_t \pi_{t+1}$  é a expectativa em t de inflação medida pelo IPCA i trimestres à frente;  $\pi_t$  é a inflação medida pelo IPCA;  $\pi_t^*$  é uma medida de inflação externa;  $h_t$  é uma medida de hiato do produto;  $Z_t^{\pi,p}$  é a variável de controle  $p$  e  $\varepsilon_t$  é um termo de erro. Os parâmetros estimados foram sujeitos à restrição de verticalidade da curva de Phillips de longo prazo, ou seja,  $\sum_{i>0} a_{1i} + \sum_{j>0} a_{2j} + \sum_{k>0} a_{3k} = 1$ .

Importante notar que apenas a parcela referente aos preços livres (aproximadamente 75% do índice) – preços de mercado – é estimada através da curva de Phillips. Os demais componentes do IPCA (administrados) tem seus preços reajustados por contrato, não respondendo a movimentos do ciclo de negócios.

A Tabela 4 apresenta um resumo das variáveis usadas na tentativa de replicar a curva de Phillips especificada pelo Banco Central. Os dados, em frequência trimestral, contemplam o período de 2004 a 2017.

**Tabela 4 - Sumário de variáveis, fonte de dados e sinais esperados na estimação da curva de Phillips**

<b>Painel A: variável dependente</b>	
Variável	Definição/fonte
Inflação de preços livres	Varição acumulada no trimestre dos preços livres no IPCA. Serie é calculada e disponibilizada pelo BC.
<b>Painel B: variáveis explicativas</b>	
Variáveis	Definição/fonte
Expectativa de inflação	Inflação esperada pelos agentes do mercado financeiro nos trimestres subsequentes, de acordo com o Relatório Focus do Banco Central. Espera-se sinal positivo.
Inflação passada	Inflação medida pelo IPCA, acumulada nos trimestres anteriores. O IPCA é calculado e disponibilizado pelo IBGE. Espera-se sinal positivo.

Inflação externa	Índice de variação de preços commodities, medidos em reais. Estimado e disponibilizado pelo Banco Central. Calculou-se a variação média no trimestre. Espera-se sinal positivo.
Hiato do produto	Hiato calculado como a diferença logarítmica entre produto trimestral e o produto potencial. A série trimestral dessazonalizada do produto é calculada pelo IBGE. O produto potencial foi estimado por filtro HP. Espera-se sinal positivo.
Taxa de câmbio nominal	Taxa de cambio reais por dólar. Calculou-se a variação nominal média no trimestre a partir de dados diários disponibilizados pelo Banco Central. Espera-se sinal positivo.
Oceanic Niño Index (ONI)	Série provida pelo <i>Climate Prediction Center</i> , vinculado ao <i>National Oceanic and Atmospheric Administration</i> (NOAA) – EUA. A variável captura choques de oferta advindos de fatores climáticos, valores positivos indicam presença de El Niño (impactando positivamente nos preços de alimentos). Espera-se sinal positivo.
PC1_cpi	Primeira componente principal das séries de busca na internet relacionadas à inflação, usada em primeira diferença. Espera-se sinal positivo.

### 3.3.3. PIB

Mais uma vez fazendo uso de modelos consagrados na macroeconomia, para estimação e posterior projeção do crescimento do PIB será utilizada a curva IS. A curva IS (Investimento-Poupança, do inglês Investment-Savings) relaciona o nível de produto a suas defasagens, e à taxa de juros real.

Também no Relatório de inflação do segundo trimestre de 2017, o Banco Central apresentou a curva IS utilizada em seu modelo de pequeno porte, reproduzida na equação (3):

$$h_t = \beta_0 + \sum_{i>0} \beta_{1i} h_{t-1} + \sum_{j>0} \beta_{2j} r_{t-j} + \sum_{k>0} \beta_{3k} \Delta sup_{t-k} + \sum_{l>0} \beta_{4l} (h_{t-l}^* + \Delta \bar{y}_{t-l}^* - \Delta \bar{y}_{t-l}) + \sum_{n \geq 0} \sum_p \beta_{5n}^p Z_{t-n}^{h,p} + u_t \quad (3)$$

onde  $h_t$  é o hiato do produto,  $r_t$  é a taxa de juros real,  $\Delta sup_t$  é a variação do superávit primário estrutural,  $h_t^*$  é o hiato do produto mundial relevante para a economia nacional,  $\Delta \bar{y}_t^*$  é o crescimento do produto potencial mundial,  $\Delta \bar{y}_t$  é o crescimento do produto potencial doméstico,  $Z_t^{h,p}$  é a variável de controle e  $u_t$  é um termo de erro.

A Tabela 7 apresenta um resumo das variáveis usadas na tentativa de replicar a curva IS especificada pelo Banco Central. Os dados, em frequência trimestral, contemplam o período de 2004 a 2017.

Tabela 5 - Sumário de variáveis, fonte de dados e sinais esperados na estimação da curva IS

**Painel A: variável dependente**

Variável	Definição/fonte
Hiato do produto	Hiato calculado como a diferença logarítmica entre produto trimestral e o produto potencial. A série trimestral dessazonalizada do produto é calculada pelo IBGE. O produto potencial foi estimado por filtro HP. Espera-se sinal positivo.

**Painel B: variáveis explicativas**

Variáveis	Definição/fonte
Taxa de juros real	Taxa de juros nominal <i>swap</i> pré-DI de 360 dias deflacionada pela expectativa de inflação relativa ao período de vigência do contrato. Calculado pela BM&F, corresponde a uma troca de posições de um fluxo de caixa a taxa fixa "Pré" por um fluxo de caixa a taxa flutuante "DI". A expectativa de inflação 12 meses à frente, dos agentes de mercado, é divulgada pelo Banco Central no Relatório Focus. Foi calculada a taxa real média no trimestre. Espera-se sinal negativo
Superávit Primário estrutural	Resultado o governo central, excluídas receitas e despesas não recorrentes, como proporção do produto potencial. A variação do primário estrutural também é conhecida como impulso fiscal. Espera-se sinal positivo.
Hiato do produto mundial	Hiato calculado como a diferença logarítmica entre produto trimestral e o produto potencial mundial dos principais parceiros comerciais do Brasil. O produto, medido em dólares, de Argentina, França, Alemanha, China e EUA foi ponderado pela participação na corrente comércio do Brasil. Espera-se sinal positivo.
Produto potencial mundial	O produto potencial foi estimado por filtro HP e foi obtido crescimento interanual. Espera-se sinal positivo.
Confiança do consumidor	Índice de confiança do consumidor da FGV. Foi obtida a variação trimestral e a variável foi usada como controle. Espera-se sinal positivo.
CDS	Credit Default Swap de 5 anos, série obtida na Bloomberg. Por medir o prêmio de risco do país, espera-se sinal negativo.
Taxa de câmbio nominal	Taxa de cambio reais por dólar. Calculou-se a variação nominal média no trimestre a partir de dados diários disponibilizados pelo Banco Central.
PC1_gdp	Primeira componente principal das séries de busca na internet relacionadas ao crescimento do produto. Espera-se sinal positivo.

### 3.4. Análise das projeções

Dentre os objetivos aqui apresentados está a avaliação dos modelos para sua finalidade preditiva. Para tanto, foram utilizadas 3 medidas diferentes de análise de projeção: raiz do erro quadrático médio (RMSE), erro absoluto médio (MAE) e a média do erro percentual absoluto (MAPE).

As estatísticas de erro de projeção são calculadas conforme abaixo:

$$RMSE = \sqrt{\sum_{t=T+1}^{T+h} (\hat{y}_t - y_t)^2 / h}$$

$$MAE = \sum_{t=T+1}^{T+h} |\hat{y}_t - y_t| / h$$

$$MAPE = 100 * \sum_{t=T+1}^{T+h} \left| \frac{\hat{y}_t - y_t}{y_t} \right| / h$$

onde, a amostra projetada é  $j = T+1, T+2, \dots, T+h$ ,  $\hat{y}_t$  é o valor projetado e  $y_t$  é o valor observado.

Em todos os casos, quanto menor o valor da estatística, menor o erro e melhor a habilidade de projeção do modelo de acordo com aquele critério.

## 4. RESULTADOS

Assim como em (CHOI e VARIAN, 2012), pretende-se estimar modelos tradicionais para as variáveis de interesse e comparar seu desempenho preditivo fora da amostra com modelos aumentados, nos quais incluem-se as primeiras componentes principais obtidas dos índices de busca do Google Trends. Como medidas de acurácia para previsões fora da amostra serão usadas a raiz do erro quadrado médio (RMSE), o erro médio absoluto e o erro percentual absoluto médio (MAPE).

### 4.1. Análise de componentes principais

A análise de componentes principais modela a estrutura de variância de um grupo de variáveis observáveis usando combinações lineares de tais variáveis. As combinações lineares, ou componentes, podem ser usadas em análises subsequentes, os coeficientes das combinações, ou *loadings*, podem ser usados para interpretar tais componentes.

As componentes principais são obtidas através da decomposição de autovalores da matriz de covariância. A primeira componente principal é a combinação linear, de norma igual a 1, com máxima variância. As componentes subsequentes maximizam a variância entre as combinações lineares de norma 1 que sejam ortogonais às componentes anteriores.

Assim, de acordo com Azevedo (2017), por meio da análise de componentes principais é possível reduzir um grande número de variáveis a poucos índices, permitindo que a variabilidade de tais séries seja usada de maneira parcimoniosa e garantindo mais graus de liberdade para o modelo.

As séries de termos buscados foram dessazonalizadas e centralizadas. Para cada conjunto de séries foi extraída a primeira componente principal e usada como variável explicativa de um modelo aumentado para projeção.

#### 4.1.1. Taxa de desemprego

A análise de componentes principais das 58 séries obtidas através do Google Trends, dessazonalizadas e padronizadas pode ser observada na Tabela 15<sup>9</sup>.

---

<sup>9</sup> Ver Apêndice.

A primeira componente (PC1\_ur) tem autovalor 33,27 e corresponde a 57% da variância total. As duas primeiras componentes correspondem a mais de 75% da variação total. A PC1\_ur possui *loadings* que variam entre -0,16 e +0,17. Como algumas das variáveis tem correlação negativa com a taxa de desemprego, podem aparecer *loadings* com sinal negativo. É possível interpretar tais séries como positivamente correlacionadas ao emprego. Os maiores *loadings* correspondem à busca pelos termos “vagas”, “tirar carteira de trabalho”, “entrevista de emprego” e “seguro desemprego”.

A variável PC1\_ur não é estacionária, segundo teste de Dickey-Fuller Aumentado (apresentou p-valor de 25,2% para hipótese nula de raiz unitária). Para que a série pudesse ser utilizada em uma equação estimada por mínimos quadrados, foi obtida a primeira diferença, estacionarizando-a.

#### **4.1.2. Inflação**

Através do Google Trends foram obtidas 45 séries de termos de busca relacionados à inflação. Os resultados da análise de componentes principais das séries após ajuste sazonal e normalização está presente na Tabela 16<sup>10</sup>.

A primeira componente principal (PC1\_cpi) apresenta autovalor 24,3 e responde por 54% da variância do conjunto de 45 séries. Os maiores pesos foram atribuídos à busca pelos termos “tabela fipec”, “quanto custa” “melhor preço” e “Buscapé<sup>11</sup>”, com *loadings* variando entre -0,19 e +0,19. Assim, como em Guzman (2011), interpreta-se que o aumento de busca por termos relacionados à inflação revelam uma preocupação maior com o tema e a PC1\_cpi será considerada uma medida de expectativa de inflação revelada.

Como a variável PC1\_cpi apresentou evidência de raiz unitária (p-valor de 14,5% no teste de Dickey-Fuller Aumentado), foi utilizada em primeira diferença.

#### **4.1.3. PIB**

Com relação às series de termos de busca relacionadas ao crescimento do PIB, a análise das componentes principais está presente na Tabela 17<sup>12</sup>.

A primeira componente principal (PC1\_gdp) apresenta autovalor 29,4 e corresponde à 68% da variância das 43 séries. Esta é a primeira componente com

---

<sup>10</sup> Ver Apêndice.

<sup>11</sup> O Buscapé é uma ferramenta online que compara preços, lojas e produtos.

<sup>12</sup> Ver Apêndice.

maior proporção de variância entre as 3 apresentadas. Além disso, chama atenção o fato de apenas o termo “recessão” ter apresentado *loading* negativo.

## 4.2. Estimações

Visando avaliar se os resultados de pesquisa na internet melhoram a capacidade de projeção de um modelo econômico, inicialmente a estimação foi realizada com uma amostra mais curta e a variável de interesse foi projetada para o período subsequente.

### 4.2.1. Taxa de desemprego

A taxa de desemprego foi estimada através da Lei de Okun para o período do segundo trimestre de 2013 ao primeiro trimestre de 2017. Também foi realizada estimação do modelo em janela móvel e a variável de interesse foi projetada um passo à frente para avaliação dos modelos. Os resultados da estimação estão presentes na Tabela 6.

**Tabela 6 - Estimação da taxa de desemprego (2013T2 - 2017T1)**

Variáveis	I	II	III
C	0,4*** (3,9)	0,6 (0,9)	0,4*** (4,76)
Hiato do produto	-14,8*** (-3,9)		-15,6*** (-5,6)
PC1_ur (-1)		0,09 (1,3)	0,08** (2,0)
AR(1)	0,6*** (3,9)	0,90*** (8,6)	0,6*** (3,5)
<b>Estatísticas comparativas</b>			
R <sup>2</sup> ajustado	87,9%	81,7%	89,2%
AIC	-0,98	-0,46	-1,04
D-W	2,14	1,38	2,11

Os símbolos \*, \*\* e \*\*\* representam os níveis de significância de 10%, 5% e 1% respectivamente. Estatística-t entre parênteses.

Erros-padrão robustos de White.

No modelo original (I), que leva em conta apenas o desvio em relação ao produto potencial, o sinal do hiato obtido está dentro do esperado, é negativo e estatisticamente significativo, assim um crescimento acima do potencial leva à redução do desemprego. Seu poder explicativo é de 87,9%, segundo o R-quadrado ajustado.

Quando se tenta estimar a taxa de desemprego apenas usando apenas a primeira diferença da primeira componente das séries do Google Trends, o sinal é o esperado, positivo, indicando que de fato o aumento do interesse no tema desemprego revelado através da plataforma de busca antecipa movimentos na taxa de desemprego. O modelo (II) tem poder explicativo de 81,7% e segundo D-W apresenta viés altista.

No modelo aumentado (III) com a inclusão da variável que sintetiza as buscas na internet o desvio em relação ao produto potencial segue se comportando dentro do esperado. A nova variável apresentou o sinal esperado, ou seja, o crescimento de buscas na internet no trimestre anterior está positivamente correlacionado com o aumento do desemprego no trimestre subsequente. A variável é estatisticamente significativa a 5%.

O poder explicativo do modelo aumentado é maior, 89.2% segundo R-quadrado ajustado. Comparando ainda os modelos I e III, também segundo o critério de informação de Akaike, que penaliza a inclusão de mais variáveis no modelo<sup>13</sup>, há evidências modestas de melhora.

A análise dos resíduos observados no Gráfico 2<sup>14</sup> confirma que os modelos são estáveis e apresentam reversão à média, com resíduos sem presença de raiz unitária. Os resíduos foram submetidos a testes de Dickey-Fuller aumentado que rejeitaram a hipótese de a presença de raiz unitária (p-valores entre 0,02% e 0,46%).

Os três modelos foram posteriormente estimados em uma janela móvel de 8 trimestres projetando-se o trimestre subsequente para criação da série de projeção fora da amostra. A Tabela 7 sumariza os resultados de avaliação de poder de projeção dos modelos acima.

---

<sup>13</sup> O critério de informação de Akaike (AIC) é calculado como:

$$AIC = -2\left(\frac{L}{T}\right) + k\log(T)/T$$

onde  $L$  é o log da máxima verossimilhança do modelo e  $k$  é o número de parâmetros estimados, usando  $T$  observações.

Assim, quanto menor o valor da estatística, melhor o modelo.

<sup>14</sup> Ver Apêndice.

**Tabela 7 - Estatísticas de avaliação de projeção da taxa de desemprego (2T14-1T17)**

Modelos	Projeção (média 1T17 -2T17)	RMSE	MAE	MAPE
I	13,3	0,16	0,14	1,70%
II	13,3	0,13	0,11	1,30%
III	13,5	0,16	0,13	1,53%

Apesar de alguma melhora no modelo com a inclusão da variável de busca na internet, a projeção não se alterou significativamente. As estatísticas de avaliação de projeção dos modelos II e III, apesar de menores, diferem pouco do modelo original. Para ilustrar a diferença nas projeções provenientes de cada modelo, calculou-se a projeção média para o ano de 2017. As projeções provenientes dos três modelos sugeriram taxa de desemprego entre 13,3% e 13,5%, enquanto a taxa de desemprego no mesmo período foi de 13,0%.

Assim, o erro dos modelos não é grande, cerca de 40 pontos-base, porém a adição da nova variável não melhorou significativamente as projeções. De fato, no caso do modelo III, a projeção média do primeiro semestre ficou mais distante do valor observado, reforçando a percepção de que as variáveis de busca não substituem as estatísticas oficiais.

#### 4.2.2. Inflação

A taxa de inflação dos preços livres foi estimada através da curva de Phillips para o período do quarto trimestre de 2004 ao primeiro trimestre de 2017, posteriormente projetou-se a taxa de desemprego no período subsequente (2T2017). Os resultados da estimação estão presentes na Tabela 8

**Tabela 8 - Estimação da taxa de inflação trimestral de preços livres do IPCA (2004T4 - 2017T1)**

Variáveis	I	II	III
Expectativa de Inflação (t+1)	0,93*** (10,4)	0,90*** (10,6)	
IPCA (t-1)	0,14 (1,63)	0,12 (1,35)	0,76*** (10,15)
Inflação de preços de commodities (t-1)	0,05*** (4,25)	0,06*** (4,44)	0,04** (1,97)

Hiato (t-2)	0,04 (0,82)	0,06 (1,35)	0,08 (1,31)
Varição taxa de câmbio nominal (t-2)	0,01 (1,34)	0,01 (1,49)	-0,002 (-0,15)
ONI (t-1)	0,03 (0,26)	0,06 (0,54)	0,05 (0,31)
PC1_CPI (t-2)		0,003* (1,64)	0,004** (2,08)

#### Estatísticas comparativas

R <sup>2</sup> ajustado	40,2%	41,5%	-24,7%
AIC	-7,34	-7,35	-6,61
D-W	1,73	1,72	1,88

Os símbolos \*,\*\* e \*\*\* representam os níveis de significância de 10%, 5% e 1% respectivamente.  
Estatística-t entre parêntese. Erros-padrão robustos de White.

No modelo original (I), os sinais dos coeficientes atenderam a expectativa e são positivos. A soma dos coeficientes de expectativa de inflação, inflação passada e de preços de commodities é estatisticamente igual a 1, segundo teste de Wald de restrição de coeficientes (com p-valor de 4,18%). No entanto, o hiato do produto e as variáveis de controle não foram estatisticamente significantes.

O poder explicativo do modelo I é baixo, cerca de 40% de acordo com o R<sup>2</sup> ajustado e a estatística de Durbin Watson indica viés de baixa na estimação.

A variável obtida com a primeira componente das séries do Google Trends relacionadas à inflação foi introduzida como uma nova medida de expectativa de inflação e apenas a segunda defasagem apresentou o sinal esperado. De fato, o coeficiente da expectativa de inflação medida pelo Banco Central diminui, ou seja, a nova variável “rouba” poder explicativo. O sinal obtido foi positivo, conforme esperado. A restrição de longo prazo da curva de Phillips continuou sendo respeitada e o teste de Wald confirmou que a soma dos coeficientes das três primeiras variáveis e da nova variável introduzida é estatisticamente igual a 1 (com p-valor de 4,95%). Ainda assim, as variáveis de controle permanecem não significativas estatisticamente.

As estatísticas comparativas sugerem leve melhora no modelo após a introdução da nova variável, com aumento do R<sup>2</sup> ajustado e redução do critério de informação de Akaike.

Extrapolando o argumento de que a nova variável pode capturar as expectativas de inflação, excluiu-se no modelo III a expectativa de inflação dos agentes de mercado. A variação da taxa de câmbio nominal passou a ter sinal negativo, ou seja, a depreciação do real ante o dólar reduziria a inflação de preços livres, o que não ocorre. Ademais, apesar de a variável nova aumentar seu coeficiente, a inflação passada aumenta muito mais sua participação, levando a impressão errada

de que a inércia inflacionária é muito maior. Pode-se afirmar, portanto, que o modelo piora muito; o R-quadrado ajustado chega a ser negativo o que sinaliza problemas na estimação. Os problemas encontrados nessa estimação reforçam o ponto de que as variáveis criadas com auxílio de buscas na internet, auxiliam na projeção mas não podem substituir as estatísticas oficiais.

A análise dos resíduos observados no Gráfico 3 confirma que os dois primeiros modelos são estáveis e apresentam reversão à média, com resíduos sem presença de raiz unitária. Os resíduos foram submetidos a testes de Dickey-Fuller aumentado que rejeitaram a hipótese de presença de raiz unitária (os dois modelos de interesse registraram p-valor 0%).

Os dois primeiros modelos foram re-estimados em uma janela móvel de 20 trimestres projetando-se o trimestre subsequente para criação da série de projeção fora da amostra. A Tabela 9 sumariza os resultados de avaliação de poder de projeção dos modelos acima.

Apesar de melhora modesta no modelo com a inclusão da variável de busca na internet, a projeção melhorou significativamente, em média, no período analisado. Para ilustrar, projetou-se a taxa de inflação de preços livres no segundo trimestre de 2017, obteve-se 0,63% de inflação acumulado no trimestre através do 1º modelo original (I), enquanto o modelo aumentado (II) projetou inflação acumulada no trimestre de 0,41%. A inflação de preços livres no referido período foi de 0,25%.

**Tabela 9 - Estatísticas de avaliação de projeção da inflação de preços livres do IPCA (4T09-1T17)**

Modelos	Projeção (2T17)	RMSE	MAE	MAPE
I	0,61%	0,5%	5,9%	28,3%
II	0,41%	0,4%	5,7%	25,6%

Considerando como dada inflação de preços administrados no segundo trimestre de 2017 em 0,12%/t e recalculando o IPCA cheio usando os pesos de preços livres e administrados no 2T17, chegamos à projeção final de IPCA pelo modelo original de 0,49% t/t e do modelo aumentado de 0,37% t/t, o último distando apenas 0,12 pontos percentuais do IPCA acumulado no trimestre.

### **4.2.3. Crescimento do PIB**

O hiato do produto foi estimado através da curva IS para o período do terceiro trimestre de 2006 ao primeiro trimestre de 2017, posteriormente projetou-se o hiato no período subsequente (2T2017). Os resultados da estimação estão presentes na Tabela 10.

Tabela 10 - Estimação do hiato do produto (2006T3 - 2017T1)

Variáveis	I	II	III
c	0,02** (2,62)	0,02*** (2,74)	0,06*** (6,00)
Hiato (t-1)	0,61*** (4,10)	0,61*** (3,94)	
Taxa de Juros Real (t-2)	-0,28** (-2,34)	-0,34** (-2,52)	-0,75*** (-6,02)
Variação do superávit primário estrutural (t-2)	1,03* (1,81)	0,95* (1,65)	2,50** (3,18)
Crescimento PIB Mundial* (t-4)	0,04 (1,34)	0,03 (1,25)	0,06 (1,36)
Variação confiança do consumidor (t-2)	0,08** (2,44)	0,08** (2,20)	0,08** (2,18)
D(CDS (t-4))	-0,64** (-2,26)	-0,55** (-2,02)	-0,88*** (-2,70)
Variação taxa de câmbio nominal (t-1)	-0,07*** (-4,20)	-0,06*** (-3,15)	-0,07* (-1,87)
PC1_GDP (t-1)			0,002** (2,56)
PC1_GDP (t-4)		0,001 (1,11)	
Estatísticas comparativas			
R <sup>2</sup> ajustado	78,0%	77,9%	56,2%
AIC	-6,40	-6,38	-5,71
D-W	2,03	2,00	1,03

Os símbolos \*, \*\* e \*\*\* representam os níveis de significância de 10%, 5% e 1% respectivamente

Estatística-t entre parênteses. Erros-padrão robustos de White.

(\*) Soma do hiato do produto mundial com a diferença de crescimento do PIB potencial mundial e nacional

No modelo original (I) todos os coeficientes apresentaram sinal dentro do esperado, no entanto a variável que representa o crescimento do PIB mundial não foi estatisticamente significativa. Seu poder explicativo é de 78%, segundo o R-quadrado ajustado.

O modelo aumentado (II), com a inclusão da variável que sintetiza as buscas na internet, não se altera muito em relação ao modelo original. A quarta defasagem da variável foi a única que apresentou o sinal correto, porém não é estatisticamente significativa.

O modelo III substitui a defasagem do hiato pela defasagem da componente principal, uma vez que a variável é positivamente correlacionada com o crescimento do produto, como observado na análise dos *loadings* da componente principal. O poder explicativo do modelo III é menor, 56,2% segundo R-quadrado ajustado, e possui viés altista de acordo com estatística DW.

Comparando ainda os modelos segundo o critério de informação de Akaike, que penaliza a inclusão de mais variáveis no modelo, há evidências de piora.

A análise dos resíduos observados no Gráfico 4<sup>15</sup> confirma que os modelos são estáveis e apresentam reversão à média, com resíduos sem presença de raiz unitária. Os resíduos foram submetidos a testes de Dickey-Fuller aumentado que rejeitaram a hipótese de a presença de raiz unitária (com p-valores entre 0% e 0,04%).

As três versões do modelo foram re-estimados em uma janela móvel de 20 trimestres projetando-se o trimestre subsequente para criação da série de projeção fora da amostra. A Tabela 11 sumariza os resultados de avaliação de poder de projeção dos modelos acima.

**Tabela 11 - Estatísticas de avaliação de projeção do PIB (3T11-1T17)**

Modelos	Projeção (2T17)	RMSE	MAE	MAPE
I	0,6%	0,7%	0,6%	1,31
II	0,4%	0,6%	0,6%	1,38
III	0,0%	1,1%	1,0%	1,99

Durante o período analisado, segundo duas estatísticas de avaliação, a inclusão na nova variável praticamente não altera o poder preditivo do modelo, a terceira, MAPE, ainda sinaliza para piora nas projeções.

Como exemplo, temos o segundo trimestre de 2017, no qual a economia brasileira cresceu 0,5% a/a. Os modelos estimados se aproximaram do crescimento real, com estimativas entre 0% a/a e 0,6% a/a. Entretanto, os modelos que incluíam a primeira componente principal das séries do Google Trends geraram projeções mais distantes do que o modelo original.

<sup>15</sup> Ver Apêndice.

## 5. CONCLUSÃO

O avanço tecnológico, a maior inserção da internet e a rápida disseminação de dados que deram início ao chamado *big data*, permitiram aos economistas explorar um novo conjunto de informações.

As buscas na internet e divulgações em mídias sociais são canais de comunicação que refletem o sentimento coletivo. As pessoas podem conhecer a opinião umas das outras e divulgar sua própria opinião em sites como Twitter, Facebook e blogs. Pessoas revelam seus pensamentos e preocupações em mecanismos de busca, como Google, Yahoo e Bing. Essas atividades combinadas criam um processo de revelação de expectativas em tempo real.

Esta dissertação utilizou dados de buscas na internet para construir novas variáveis que auxiliassem na estimação de modelos amplamente consagrados entre os macroeconomistas e melhorar seu poder preditivo.

Com auxílio de uma pesquisa de opinião na internet, os termos de busca foram traçados e através da análise de componentes principais as novas variáveis foram criadas.

No caso da taxa de desemprego, o modelo aumentado apresentou aderência maior aos dados do que o modelo original, assim como visto em Askitas e Zimmermann (2009). Entretanto, seu poder preditivo quase não se alterou.

Com relação aos modelos de inflação, houve pouca alteração nas estatísticas comparativas dos modelos, mas a performance preditiva melhorou. Assim como sugeria a literatura de Guzman (2011), as buscas na internet podem oferecer alguma contribuição na percepção dos agentes que não pertencem ao mercado financeiro com relação ao tema.

Já os modelos de crescimento do produto pioraram em aderência e poder de projeção. No entanto, a análise de componentes principais mostrou que os termos relacionados ao crescimento econômico formam um vetor de pesos positivos com exceção do termo de busca “recessão”.

O uso de dados de busca na internet pode ser vantajoso uma vez que o Google Trends é atualizado diariamente. Essa nova fonte de dados é flexível e pode ser convertida facilmente em séries de frequência mais baixa para uso em modelos tradicionais. No entanto, provou-se aqui que as séries de busca de dado podem melhorar o desempenho de alguns modelos, mas não substituir as estatísticas oficiais.

Além disso, vale destacar a dificuldade de se obter os termos de busca mais apropriados para sumarizar a sensibilidade da população em relação a um tema

específico. Uma combinação diferente de termos de busca poderia gerar resultados totalmente distintos dos apresentados aqui.

Existem ainda inúmeras aplicações potenciais em economia e finanças para tal fonte, como *nowcasting* e projeções outras variáveis macroeconômicas e de retorno no mercado de commodities e mercado de ações.

## 6. REFERÊNCIA BIBLIOGRÁFICA

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

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

AZEVEDO, L. F. P. **Impactos econômicos e financeiros de notícias**. Fundação Getúlio Vargas. [S.l.]. 2017.

BLANCHARD, Olivier. **Macroeconomia**. 5ª edição. 2011.

CARLIN, W.; SOSKICE, D. **Macroeconomics: Institutions, instability, and the financial system**. [S.l.]: Oxford University Press, USA, 2014.

CHOI, H.; VARIAN, H. Predicting initial claims for unemployment benefits. **Google Inc**, p. 1-5, 2009.

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

DE MAURO, A.; GRECO, M.; GRIMALDI, M. **What is big data? A consensual definition and a review of key research topics**. AIP conference proceedings. [S.l.]: [s.n.]. 2015. p. 97-104.

EDELMAN, B. Using internet data for economic research. **The Journal of Economic Perspectives**, v. 26, 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.

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

IGBOAYAKA, J.-V. C. E. **Using Social Media Networks for Measuring Consumer Confidence: Problems, Issues and Prospects**. Université d'Ottawa/University of Ottawa. [S.l.]. 2015.

KOOP, G.; ONORANTE, L. **Macroeconomic nowcasting using Google probabilities**. **University of Strathclyde**, 2013.

LOHR, S. The age of big data. **New York Times**, v. 11, 2012.

MCAFEE, A.; BRYNJOLFSSON, E.; OTHERS. Big data: the management revolution. **Harvard business review**, v. 90, p. 60-68, 2012.

OKUN, Arthur M. The gap between actual and potential output. In: **Proceedings of the American Statistical Association**. 1962.

RELATÓRIO, DE INFLAÇÃO. Banco Central do Brasil, jun. 2017 (publicação trimestral). Disponível em <http://www.bcb.gov.br/htms/reinf/port/2017/09/ri201709P.pdf>

SCOTT, S. L.; VARIAN, H. R. Bayesian variable selection for nowcasting economic time series. In: \_\_\_\_\_ **Economic analysis of the digital economy**. [S.l.]: University of Chicago Press, 2015. p. 119-135.

STOCK, J. H.; WATSON, M. W. Vector autoregressions. **The Journal of Economic Perspectives**, v. 15, p. 101-115, 2001.

SUHOY, T. **Query indices and a 2008 downturn: Israeli data**. Bank of Israel. [S.l.]. 2009.

TAYLOR, J. B. **Discretion versus policy rules in practice**. Carnegie-Rochester conference series on public policy. [S.l.]: [s.n.]. 1993. p. 195-214.

VARIAN, H. R. Big data: New tricks for econometrics. **The Journal of Economic Perspectives**, v. 28, p. 3-27, 2014.

WARD, J. S.; BARKER, A. Undefined by data: a survey of big data definitions. **arXiv preprint arXiv:1309.5821**, 2013.

## APÊNDICE

**Tabela 12 - Respostas à pergunta relacionada ao desemprego e termos relacionados do Google Trends**

Respostas	Termos relacionados
Carreira	plano de carreira, carreira profissional, cargos e salários, planejamento de carreira
carteira de trabalho	tirar carteira de trabalho
desemprego	seguro desemprego, desemprego brasil,
emprego	entrevista, entrevista emprego, entrevista de emprego perguntas, entrevista de trabalho,
head hunter	michael page, recrutador, hays, robert half, page personnel
linkedin	linkedin Brasil, linkedin vagas, login linkedin, catho, vagas catho, empregos catho, login catho, catho curriculum, infojobs, infojobs vagas, infojobs empregos, login infojobs, indeed, indeed vagas, indeed empregos
modelo de curriculo	curriculum, curriculum modelo, curriculum vitae, como fazer curriculo, fazer curriculo
oportunidade	oportunidade de emprego, oportunidade de trabalho
oportunidade de emprego	oportunidade de trabalho
recrutadores	emprego, linkedin
site de vagas de emprego	site de empregos, site de empregos grátis
Vaga	vagas emprego, agencia de emprego
vagas abertas	vagas de emprego abertas
vagas de emprego	site vagas de emprego
vagas de trabalho	vagas de emprego
vagas em aberto	concurso em aberto, concursos aberto

**Tabela 13 - Respostas à pergunta relacionada à inflação e termos relacionados do Google Trends**

Respostas	Termos relacionados
Carestia	
comparação de preços	site de comparação de preços, comparar preços, buscape, bondfaro, submarino, extra, walmart, site de compras, compras coletivas*
comprar [nome do produto]	onde comprar, quero comprar,
cupom de desconto	cupom [loja específica]
custo de vida	custo de vida Brasil
diferença de preço	
IGP	IGP-DI, igpm, IGP FGV, Índice IGP, reajuste aluguel, indice de reajuste
Inflação	inflação brasil, o que é inflação, taxa inflação, índice inflação, juros, IPCA
IPCA	acumulado ipca, ipca indice, inpc, taxa ipca, selic, o que é ipca
preço alto	
preço baixo	menor preço
preço	melhor preço, bom preço
preço médio	Fipe, tabela fipe
promoções	sine vagas de emprego, vagas de emprego gratis
quanto custa	vagas de emprego

**Tabela 14 – Respostas à pergunta relacionada ao crescimento do PIB e termos relacionados do Google Trends**

Respostas	Termos relacionados
balança comercial	balança comercial brasileira, superavit, superavit comercial, deficit, balança de pagamentos
confiança do consumidor	índice de confiança do consumidor
crescimento	crescimento do brasil, crescimento e desenvolvimento, crescimento economico, taxa de crescimento, crescimento pib, fatores de crescimento, crescimento pib brasil, crescimento economico brasil, crescimento e desenvolvimento economico, desenvolvimento econômico
crescimento do PIB	crescimento do pib brasileiro
desemprego	seguro desemprego, desemprego brasil,
economia	brasil economia, economia do brasil, economia brasileira, notícias economia,
ibovespa	ações
idh	idh brasil
índice de confiança	
investimento	fundo investimento, o que é investimento, banco investimento, qual melhor investimento, investimento direto, poupança, qual o melhor investimento, tesouro direto, investimento cdb, franquias, simulador investimento
juros	taxa de juros, financiamento, taxa de juros de financiamento, qual a taxa de juros, taxa de juros brasil
PIB	PIB Brasil, o que é PIB, PIB per capita
recessão	brasil recessão, o que é recessão, recessão econômica, recessão significado,
salário	salário mínimo, salário mínimo valor, 13 <sup>o</sup> salário, salário mínimo atual
Selic	juros Selic, o que é Selic, Selic hoje
situação econômica	
taxa de crescimento	

**Tabela 15 - Análise das componentes principais das séries relacionadas ao desemprego**

<b>Painel A: Autovalores das componentes principais</b>			
Componente Principal	Autovalores	Proporção	Proporção acumulada
1	33,27	0,57	0,57
2	11,15	0,19	0,77
3	3,11	0,05	0,82
4	2,68	0,05	0,87
5	1,63	0,03	0,89
6	0,96	0,02	0,91
7	0,82	0,01	0,92
8	0,73	0,01	0,94
9	0,53	0,01	0,95
10	0,45	0,01	0,95
⋮	⋮	⋮	⋮
58	0,00	0,00	1,00

<b>Painel B: Loadings da primeira componente principal</b>	
Variável	loadings
carreira	-0.03
plano de carreira	-0.14
carreira profissional	-0.09
carteira de trabalho	0.15
tirar carteira de trabalho	0.17
cargos e salários	-0.17
linkedin	0.16
linkedin vagas	0.15
linkedin brasil	0.13
login linkedin	0.13
catho	-0.03
vagas catho	0.12
empregos catho	-0.04
login catho	0.13
modelo de curriculo	-0.14
como fazer curriculo	0.02
fazer um curriculo	0.05
curriculum vitae	-0.16
infojobs	0.15
infojobs empregos	0.15
infojobs vagas	0.16
infojobs login	0.16
indeed	0.15
indeed vagas	0.15
indeed emprego	0.15

vagas	0.17
vagas emprego	0.17
oportunidade	-0.13
oportunidade de emprego	-0.12
oportunidade de trabalho	0.03
vagas de emprego	0.17
agencia de emprego	-0.16
vagas abertas	0.15
vagas de emprego abertas	0.12
vagas de emprego	0.17
sine vagas de emprego	0.17
vagas de emprego gratis	0.11
site de vagas de emprego	0.12
site de empregos	-0.15
site de empregos gratis	0.06
vagas de trabalho	0.16
vagas em aberto	0.08
concurso em aberto	0.14
concursos aberto	0.08
desemprego	0.12
seguro desemprego	0.15
desemprego brasil	-0.12
emprego	0.10
entrevista	0.14
entrevista de emprego	0.16
perguntas entrevista de emprego	0.14
entrevista trabalho	-0.06
head hunter	-0.16
Recrutador	-0.15
michael page	-0.11
hays	0.05
robert half	0.11
page personnel	0.09

---

**Tabela 16 - Análise das componentes principais das séries relacionadas à inflação**

<b>Painel A: Autovalores das componentes principais</b>			
Componente Principal	Autovalores	Proporção	Proporção acumulada
1	24,30	0,54	0,54
2	8,87	0,20	0,74
3	2,93	0,07	0,80
4	2,25	0,05	0,85
5	1,43	0,03	0,88
6	1,04	0,02	0,91
7	0,85	0,02	0,93
8	0,63	0,01	0,94
9	0,47	0,01	0,95
10	0,37	0,01	0,96
⋮	⋮	⋮	⋮
45	0,00	0,00	1,00

**Painel B: Loadings da primeira componente principal**

Variável	loadings
carestia	0.00
comparação de preços	-0.17
comprar	0.19
onde comprar	0.19
quero comprar	0.12
site de comparação de preços	-0.08
buscape	0.18
submarino	0.06
bondfaro	0.06
site de compras	0.12
extra	0.19
walmart	0.18
Cupom de desconto	0.14
custo de vida	0.04
custo de vida brasil	-0.07
diferença de preço	0.18
igp	-0.10
reajuste aluguel	0.07
indice de reajuste	-0.02
igpm	-0.19
igp di	-0.16
inflação	-0.18
inflação brasil	-0.16
o que é inflação	-0.15
taxa inflação (Brasil)	-0.16

índice inflação	-0.19
juros	-0.07
IPCA	-0.14
acumulado ipca	-0.01
ipca indice	-0.14
o que é ipca	-0.13
selic	-0.18
inpc	-0.20
taxa ipca	0.05
preço alto	0.18
preço baixo	0.18
menor preço	0.18
preço	0.18
melhor preço	0.19
bom preço	0.17
preço médio	0.17
fipe	0.20
tabela fipe	0.19
promoções	0.11
quanto custa	0.19
carestia	0.00
comparação de preços	-0.17
comprar	0.19
onde comprar	0.19
quero comprar	0.12
site de comparação de preços	-0.08
buscape	0.18
submarino	0.06
bondfaro	0.06
site de compras	0.12
extra	0.19
walmart	0.18
Cupom de desconto	0.14

---

**Tabela 17 - Análise das componentes principais das séries relacionadas ao crescimento do PIB**

<b>Painel A: Autovalores das componentes principais</b>			
Componente Principal	Autovalores	Proporção	Proporção acumulada
1	29.44	0.68	0.68
2	2.61	0.06	0.75
3	2.32	0.05	0.80
4	1.52	0.04	0.83
5	1.35	0.03	0.87
6	1.00	0.02	0.89
7	0.92	0.02	0.91
8	0.61	0.01	0.92
9	0.55	0.01	0.94
10	0.42	0.01	0.95
⋮	⋮	⋮	⋮
43	0,00	0,00	1,00

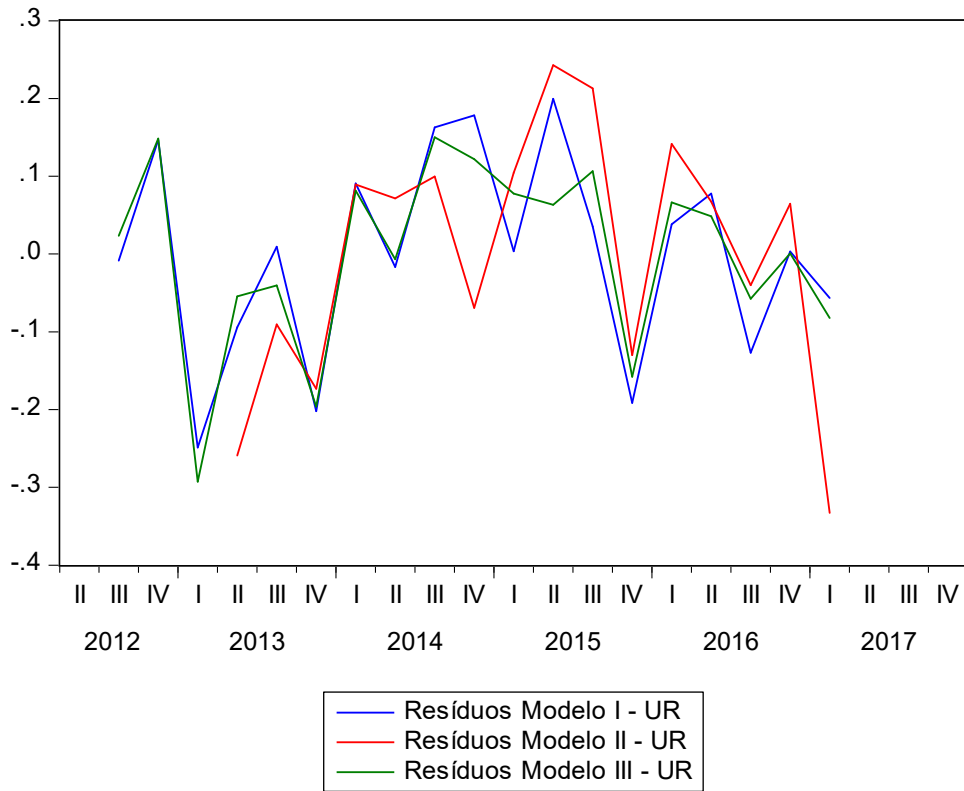
**Painel B: *Loadings* da primeira componente principal**

Variável	<i>loadings</i>
pib	0.18
pib brasil	0.17
o que é PIB	0.16
pib per capita	0.17
balança comercial	0.18
balança comercial brasileira	0.17
superavit	0.15
superavit comercial	0.15
deficit	0.14
deficit comercial	0.13
balança de pagamentos	0.17
confiança do consumidor	0.14
indice de confiança do consumidor	0.02
crescimento	0.18
crescimento do brasil	0.18
crescimento e desenvolvimento	0.17
crescimento economico	0.18
taxa de crescimento	0.17
crescimento pib	0.17
fatores de crescimento	0.06
crescimento pib brasil	0.17
crescimento economico brasil	0.16
crescimento e desenvolvimento economico	0.15
desenvolvimento economico	0.18

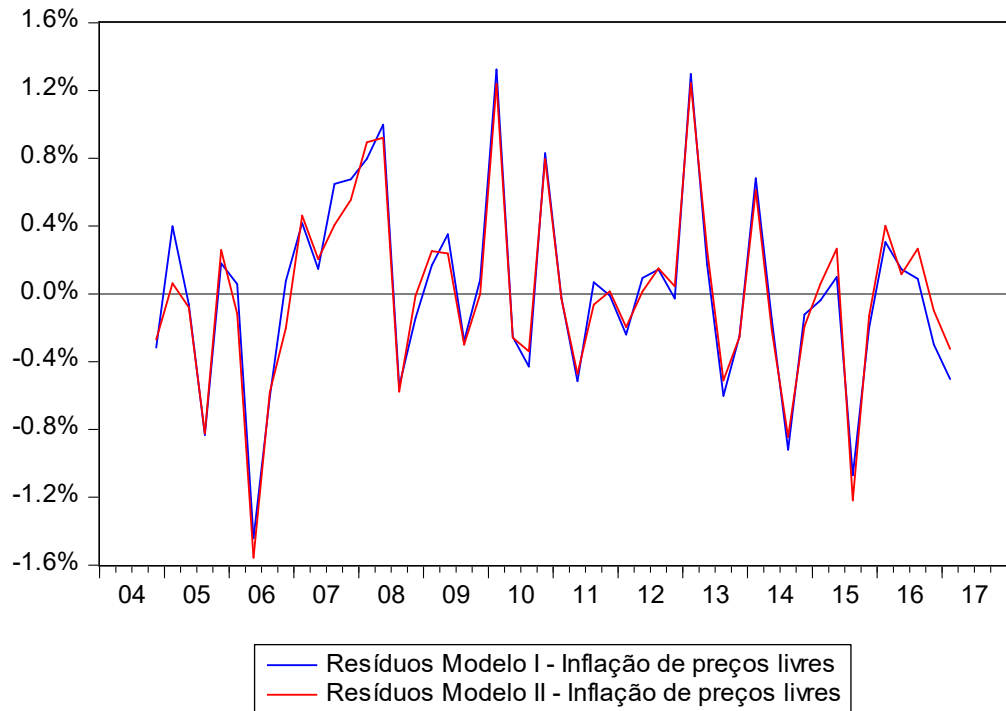
crescimento do PIB	0.15
crescimento do pib brasileiro	0.14
crescimento do pib brasileiro	0.14
economia	0.18
brasil economia	0.18
economia do brasil	0.18
economia brasileira	0.18
noticias economia	0.12
terra economia	0.12
ibovespa	0.05
ações	0.14
idh	0.17
idh brasil	0.14
investimento	0.17
juros	0.08
recessão	-0.02
salário	0.10
selic	0.17
situacao economica	0.17
pib	0.18
pib brasil	0.17
o que é PIB	0.16
pib per capita	0.17
balança comercial	0.18
balança comercial brasileira	0.17
superavit	0.15
superavit comercial	0.15
deficit	0.14
deficit comercial	0.13
balança de pagamentos	0.17
confiança do consumidor	0.14
indice de confiança do consumidor	0.02
crescimento	0.18
crescimento do brasil	0.18

---

**Gráfico 2 - Resíduos das estimações de taxa de desemprego diferenciada**



**Gráfico 3 - Resíduos das estimações de inflação de preços livres do IPCA**



**Gráfico 4 - Resíduos das estimações do hiato do produto**