

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

SAMUEL GUIMARÃES FILHO

**GOOGLE TRENDS PARA PREVISÃO DE VARIÁVEIS MACRO: USO NO BRASIL
ATRAVÉS DO ALGORITMO AUTOMETRICS**

SÃO PAULO

2017

SAMUEL GUIMARÃES FILHO

**GOOGLE TRENDS PARA PREVISÃO DE VARIÁVEIS MACRO:
USO NO BRASIL ATRAVÉS DO ALGORITMO AUTOMETRICS**

Dissertação apresentada à Escola de
Economia de São Paulo da Fundação
Getúlio Vargas, como requisito para
obtenção de título de Mestre em Economia

Campo de Conhecimento: Economia

Orientador: Prof. Dr. Emerson Fernandes
Marçal

SÃO PAULO

2017

Guimarães Filho, Samuel.

Google Trends para previsão de variáveis macro : uso no Brasil através do algoritmo autometrics / Samuel Guimarães Filho. - 2017.

37 f.

Orientador: Emerson Fernandes Marçal

Dissertação (mestrado) - Escola de Economia de São Paulo.

1. Google (Firma). 2. Ferramentas de busca na Web. 3. Mercado de trabalho - Estatística. I. Marçal, Emerson Fernandes. II. Dissertação (mestrado) - Escola de Economia de São Paulo. III. Título.

CDU 331.6

SAMUEL GUIMARÃES FILHO

**GOOGLE TRENDS PARA PREVISÃO DE VARIÁVEIS MACRO:
USO NO BRASIL ATRAVÉS DO ALGORITMO AUTOMETRICS**

Dissertação apresentada à Escola de Economia de
São Paulo da Fundação Getúlio Vargas,
como requisito para obtenção do título de Mestre
em Economia

Campo de conhecimento: Economia

Data da aprovação:

10 / 02 / 2017

Banca examinadora:

Prof. Dr. Emerson Fernandes Marçal
(Orientador)
FGV - EESP

Prof. Dr. Rogerio Mori
FGV – EESP

Prof. Dr. Diogo de Prince Mendonça
UNIFESP

AGRADECIMENTOS

Neste espaço, gostaria de poder agradecer a todos que estiveram comigo ao longo deste processo e que, sem eles, a conclusão deste trabalho não seria possível.

Primeiramente, gostaria de agradecer aos meus pais pelo apoio e incentivo incondicional aos estudos e a buscar meus sonhos e objetivos mesmo que distante de casa.

Agradeço também a minha esposa pelo amor, apoio e, sobretudo, pela compreensão ao longo destes anos de estudos a mais.

Agradeço também aos meus colegas de classe. Felizmente, ao longo do curso, grande amizades foram feitas e, indubitavelmente, estas serão levadas por toda a vida.

Por fim, agradeço aos professores de Escola de Economia da FGV pelos ensinamentos ao longo do curso e, sobretudo, ao meu orientador, Emerson Fernandes Marçal, pela paciência, ajuda e condução do trabalho ao longo deste processo.

RESUMO

Este trabalho tem como objetivo testar se o uso do Google Trends como variável exógena melhora a previsão do dado mensal do CAGED em relação a modelos que usam apenas as próprias defasagens. Para a seleção do modelo foi utilizado o algoritmo *Autometrics* e para comparação de modelos o utilizado o *Model Confidence Set*. Além disto, o modelo que utiliza o *Google Trends* foi comparado com previsões dos analistas de Mercado. Os resultados encontrados apontam que o modelo que utiliza o *Google Trends* como variável exógena é superior ao modelo que utiliza apenas a própria defasagem. No entanto, este modelo, não foi capaz de superar os analistas de mercado.

Palavras-chave: Mercado de trabalho, CAGED, *Google Trends*, seleção automática de modelos, *Autometrics*, *Model Confidence Set*

ABSTRACT

This work aims to test if the use of Google Trends as an exogenous variable improves the prediction of the monthly data for Brazilian Formal Job Creation (CAGED) compared to a model that uses only the lags themselves. For the selection of the model was used the algorithm *Autometrics* and for model comparison the Model Confidence Set. In addition, the model that uses Google Trends data will be compared with some market analyst's forecasts. The results show that the model the uses the Google data as an exogenous variable is superior to the model that only uses the lag itself. However, this model was not able to overcome the market analysts.

Keywords: Labour Market, Formal Job Creation, Google Trends, Automated Model Selection, *Autometrics*, *Model Confidence Set*

LISTA DE FIGURAS

Figura 1: CAGED - Saldo acumulado 12 meses	12
Figura 2: Iniciando no DGP até modelos específicos	16
Figura 3: Índice para palavra “emprego” no Brasil	20
Figura 4: Google Correlate para a série da primeira diferença do CAGED.....	21
Figura 5: Seleção de Palavras do Google	22
Figura 6: Tela da Bloomberg com informações sobre as projeções dos Analistas	27
Figura 7: Selecionando o Modelo utilizando as variáveis do Google Trends	29
Figura 8: Seleção entre o Modelo Google e o modelo Naive.....	29
Figura 9: Selecionando o modelo com variáveis do Google Trends e Macroeconômicas.....	30
Figura 10: Seleção entre Modelo Google e Modelo Google_Macro.....	30
Figura 11: Seleção entre o Modelo Google e Previsões dos Analistas de Mercado	30

LISTA DE TABELAS

Tabela 1: Teste Raiz Unitária para Variáveis do Google Trends	23
Tabela 2: Resultado Teste Raiz Unitária para série do CAGED.....	24
Tabela 3: Variáveis Macro Econômicas.....	25
Tabela 4: Resultado Teste Raiz Unitária para Variáveis Macroeconômicas.....	26
Tabela 5: Resultado da Primeira Etapa.....	31
Tabela 6: Resultado de Segunda Etapa.....	31
Tabela 7: Coeficientes das Variáveis Macro regredidas contra o Erro do Modelo Google	32
Tabela 8: Resultado da Terceira Etapa	33

SUMÁRIO

1. INTRODUÇÃO	11
2. REVISÃO DA LITERATURA.....	14
2.1. Google Trends.....	14
2.2. Now-casting e Indicadores Antecedentes e Coincidentes	15
2.3. Seleção Automática de Modelos	16
2.3.1. Autometrics	17
2.4. Model Confidence Set	18
3. METODOLOGIA E DESCRIÇÃO DA BASE DE DADOS	20
3.1. O que é o Google Trends?	20
3.2. Selecionado as Palavras	21
3.3. Tratamento do Índice extraído do Google Trends	22
3.4. CAGED e Tratamento do CAGED.....	24
3.5. Variáveis Econômicas e Tratamento das Variáveis.....	25
3.6. Modelo Naive	26
3.7. Benchmark de Mercado: Projeção dos Analistas	27
3.8. Tamanho da Base da Dados e Período Analisado	27
3.9. Variáveis Explicativas Elegíveis e Modelo Google.....	28
4. RESULTADOS	29
5. CONCLUSÃO	34
6. LIMITAÇÕES E POSSÍVEIS EXTENSÕES.....	35
7. REFERÊNCIAS.....	36

1. INTRODUÇÃO

Para qualquer economista, obter informações sobre o andamento da economia de qualquer país é muito importante. Além disto, seja em nível de governo, seja para o empresário, ter uma melhor noção de como anda a atividade “em tempo real” torna a tomada de decisão, seja da gestão de estoques, do gerenciamento de expectativas, ou mesmo esforços de arrecadação, mais precisa.

No entanto, os dados econômicos são divulgados com um certo atraso pelos órgãos estatísticos uma vez que precisam, sempre, ajustar os dados brutos que coletam. No caso brasileiro, por exemplo, o dado do PIB de um determinado trimestre é divulgado com um atraso de aproximadamente dois meses em relação ao término do trimestre em questão.

Ao mesmo tempo, milhares de pesquisas são realizadas diariamente na rede (www). Pessoas pesquisam sobre os mais variados assuntos, tais como viagens, curiosidades, vagas de emprego, doenças, etc. A pergunta que se coloca neste trabalho, então, é se é possível utilizar dados de pesquisas realizadas pelos usuários na internet, em especial, pesquisas do Google (*Google Trends*) para aperfeiçoar a previsão sobre dados econômicos.

Em outras palavras, o objetivo central do presente trabalho é, portanto, testar se o uso do *Google Trends* como variável exógena melhora a previsão de dados econômicos no Brasil em relação a modelos que usam apenas as próprias defasagens.

Já em 2009, Choi e Varian se fazem a mesma pergunta e, além de ter como objetivo familiarizar os leitores do artigo com esta nova ferramenta, chegam a evidências positivas sobre o uso do *Google Trends* e, ao longo do tempo, desenvolvem novos estudos que serão melhor detalhados nas próximas seções,

Google Trends é um índice normalizado que permite analisar o comportamento sobre o volume de pesquisas de determinada palavra ao longo do tempo. Além disto, autor não tem conhecimento até o presente de outro trabalho que tenha o Brasil como foco nesta análise.

Neste trabalho, seu uso será na tentativa de melhor prever o mercado de trabalho, em especial, o CAGED – cadastro geral de empregados e desempregados –, divulgado pelo ministério do trabalho, que mede, para determinado mês, o saldo entre admissões e demissões de trabalhos formais em todo território nacional.

O motivo principal pelo qual este trabalho se concentrará no mercado de trabalho se deve ao fato de haver uma série de trabalhos no exterior que também testam o uso do *Google Trends* para melhorar a previsão de dados relativos a emprego.

Além disto, a análise do mercado de trabalho é fundamental para qualquer economia de mercado. Além de ser uma medida associada de bem estar dos indivíduos, a taxa de desemprego tem importante correlação com a formação dos salários e inflação.

Ainda, para o Brasil, dado a relevância no consumo das famílias no produto, identificar, portanto, possíveis pontos de virada na tendência do mercado de trabalho se torna crucial. Por fim, este tema se torna ainda mais relevante no atual contexto da profunda crise econômica na qual o país se encontra.

Desde setembro de 2014, houve a perda de mais de 2 milhões e meio de empregos com carteira assinada. No acumulado de 12 meses, para o mês de junho de 2016 verifica-se o pior resultado em mais de quinze anos, superando inclusive o período da grande crise financeira global de 2008.

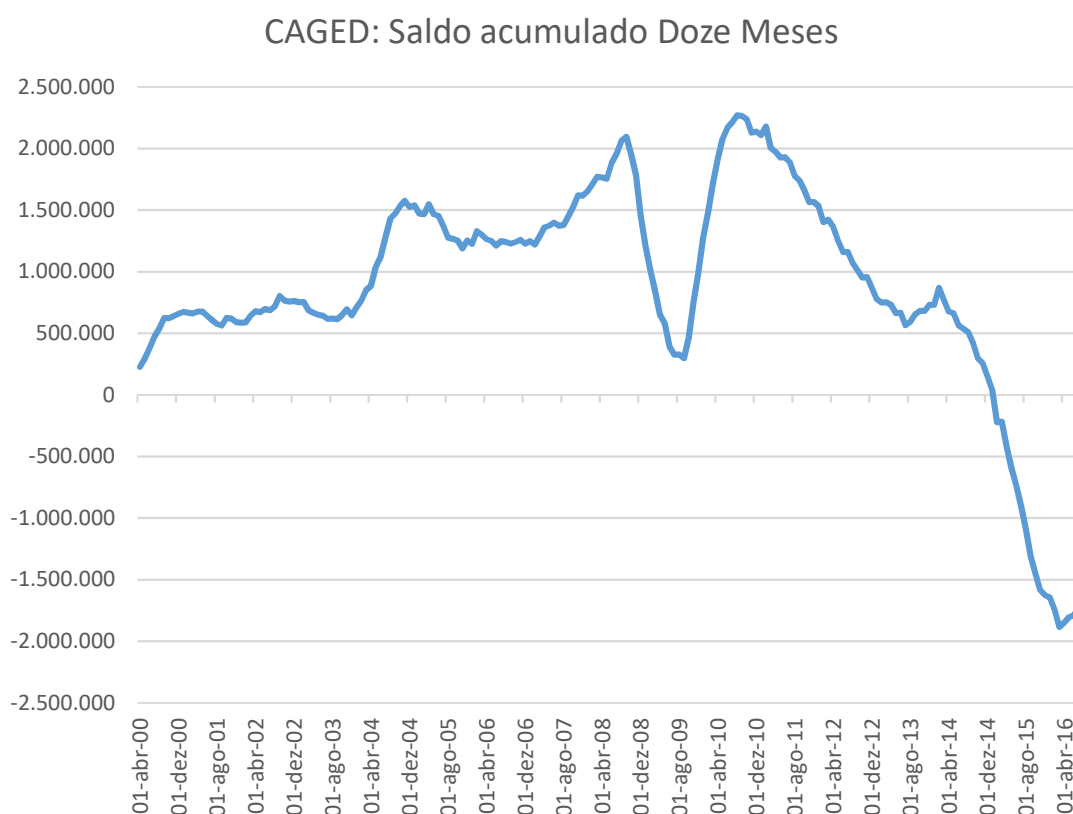


Figura 1: CAGED - Saldo acumulado 12 meses

Para responder esta pergunta, alguns procedimentos foram realizados. Pode-se dividir o procedimento realizado em três etapas. Inicialmente, através da pesquisa de palavras relacionadas a emprego para a território brasileiro no Google, foi estimado o melhor modelo para prever o CAGED. Este modelo foi comparado, então, a um modelo *naive*, que utilizava apenas a própria defasagem do CAGED e as *dummies* sazonais.

Na segunda etapa, um novo modelo foi estimado. Desta vez, além de incorporar as informações disponíveis no *Google Trends*, há a adição de variáveis macroeconômicas. Em seguida, este novo modelo foi comparado ao modelo estimado na primeira etapa, o qual utilizava apenas as informações do *Google Trends*.

Por fim, na terceira etapa, o modelo “vencedor” da segunda etapa foi comparado a um *benchmark* de mercado. Neste caso, será utilizada a projeção dos analistas de mercado disponível na *Bloomberg* no dia da divulgação do dado do CAGED.

Para a seleção do melhor modelo foi utilizado o algoritmo *Autometrics* (Doornik, 2009). Já para a comparação de modelos, foi utilizado o critério de seleção *Model Confidence Set* (Hansen, Lunde e Nason, 2010).

Algumas conclusões que serão posteriormente detalhadas: (i) o modelo que utiliza o *Google Trends* é superior ao modelo que utiliza as próprias defasagens do CAGED, (ii) A adição das variáveis macro não foi capaz de melhorar o modelo que utilizava apenas o *Google Trends*, (iii) os modelos citados acima não foram capazes de superar as previsões dos analistas de mercado disponíveis na *Bloomberg*.

O presente trabalho será dividido nas seguintes seções: (i) Revisão da Literatura, (ii) Metodologia e Descrição da Base de Dados, (iii) Resultados, (iv) Conclusão, (v) Limitações e Possíveis Extensões; (vi) Referências.

2. REVISÃO DA LITERATURA

2.1. Google Trends

Os trabalhos que serão citados abaixo tentam responder se o uso do *Google Trends* como variável exógena melhora a previsão do modelo em relação ao modelo auto regressivo e as evidências são que o *Google Trends* melhora a previsão. O ponto que se faz aqui é que este trabalho não está interessado em prever o futuro, mas sim, se com a ajuda do *Google Trends* é possível fazer uma melhor previsão sobre presente (Choi & Varian, 2012).

Ainda Choi & Varian (2012) mostra que o uso do *Google Trends* pode ajudar a fazer previsões de curto prazo para o número de pedidos de seguro desemprego e venda de veículos nos Estados Unidos. Mais especificamente, quando comparados a modelos AR1, há uma melhora de 5.95% no erro absoluto médio para o caso de pedidos de seguro desemprego e de 10,6% para a venda de veículos.

D'Amuri, Marcucci (2012) sugerem o uso do *Google Trends* como melhor *leading indicator* para a taxa de desemprego nos Estados Unidos. Fazendo previsões fora da amostra, eles mostram que o melhor modelo que utiliza o *Google Trends* apresenta uma performance superior ao que utiliza *leading indicators* tradicionais, reduzindo o erro quadrático médio em 29%.

Tuhkuri (2014, 2015 e 2016) utiliza *Google search volumes* para prever taxa de desemprego nos 28 países da união Europeia e conclui que, de fato, o uso do Google ajuda na previsão de desemprego, no entanto, este uso é limitado a previsões de curto prazo.

Os trabalhos citados acima utilizam os dados para países desenvolvidos. Será que é possível estender este estudo para países emergentes?

Há estudos para países emergentes e os resultados seguem sendo promissores. Chadwick, Sengul (2012) mostram que para a Turquia, quando comparado a modelos que usam apenas as defasagens da própria taxa de desemprego, os modelos que contêm o *Google Trends* melhoram a previsão da taxa de desemprego mensal.

No Brasil, o presente autor desconhece estudos relacionados ao uso do *Google Trends* para previsões de tendência do mercado de trabalho. Há no entanto, um trabalho sobre a aplicação do *Google Trends* para outros fins. Caldeira, Perlin, Pontuschka, Santos (2015) encontram resultados positivos se pesquisas de palavras relacionadas a finanças podem ter impacto em

incerteza de mercado, log-retorno e volume negociado no mercado acionário para 4 países de língua inglesa.

2.2. Now-casting e Indicadores Antecedentes e Coincidentes

De acordo com Banbura, Giannone, Modugno e Reichlin (2013), *now-casting* é chamado da definição do presente, do futuro muito próximo e do passado muito recente. É a contração da palavra *now*, que significa agora, com *forecasting*, que significa previsão.

A importância de se fazer uma previsão sobre alguma variável macroeconômica é que, muitas vezes, como dito na introdução, há um atraso na divulgação dos dados. Para o CAGED, objeto deste trabalho, sua divulgação ocorre com um atraso de quase um mês, normalmente na última semana do mês seguinte ao mês de referência do dado.

Ainda de acordo com Banbura, Giannone, Modugno e Reichlin (2013), princípio básico do *now-casting* é explorar informações que são disponibilizadas antes e com maior frequência que a variável de interesse para se obter uma estimativa mais cedo sobre o dado antes de sua divulgação oficial.

Envolve, portanto, obter a projeção de uma variável de interesse baseado em um conjunto de variáveis disponíveis, Ω_v , onde v está associado a tempo de algum dado em particular. Banbura, Giannone, Modugno e Reichlin (2013) chama a atenção que este v não deve ser confundido com o t . Normalmente, dado que não há uma perfeita sincronização de divulgação dos dados utilizados para fazer a previsão e a variável de interesse, v e t podem até mesmo possuir frequências diferentes.

No caso do presente trabalho, com a exceção dos dados do *Google Trends* que possuem frequência semanal, todas as variáveis que serão utilizadas e analisadas possuem frequência mensal. Para contornar este problema, mesmo para a variável do *Google Trends* semanal, como será mostrado mais adiante, será utilizado o dado semanal como se ele fosse mensal. Desta forma, para cada palavra pesquisada, para um dado mês, haverá quatro variáveis representando as quatro semanas.

2.3. Seleção Automática de Modelos

De acordo com Doornik, Hendry (2013), o processo gerador de dados (DGP – *Data Generation Process*) é o mecanismo econômico que opera no mundo real. Muitas vezes, estas relações são muito complexas, de difícil mensuração.

Uma vez que o DGP é impossível de modelar, há a necessidade, então, de reduzir a análise para um tamanho que seja manejável. O que se busca, então, é encontrar modelos que, de uma forma simplificada, baseada em uma série de variáveis observáveis, consiga explicar melhor o DGP. São os chamados processo local gerador de dados LDGP (*Local Data Generation Process*).

Realizar este procedimento, ou seja, esta modelagem partindo do LDGP até um modelo específico é chamado de geral-para-específico (GET – *General to Specific*). O ponto inicial do processo de redução de modelo é o modelo geral irrestrito (GUM – *General Unrestricted Model*). Neste modelo, há a inclusão de todas as variáveis que, em teoria, podem explicar o fenômeno estudado. Novamente, Doornik, Hendry (2013), descreve em uma figura o processo que se permite ir do desconhecido DGP até o modelo específico.

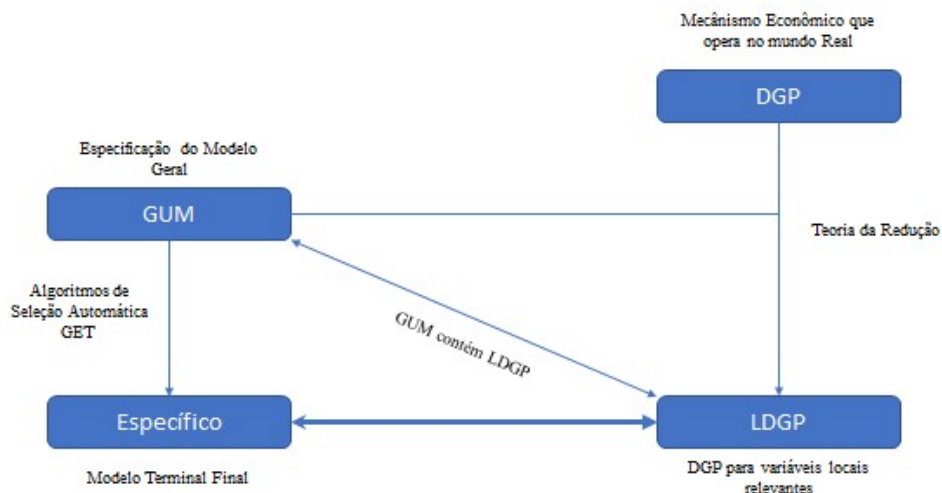


Figura 2: Iniciando no DGP até modelos específicos

Há um *trade-off* no tamanho do GUM. Um GUM maior aumenta a chance de reter variáveis indesejadas. Enquanto isto, um GUM menor pode significar que variáveis chaves são omitidas.

Quando o GUM inicial é grande, realizar os GET's pode consumir muito tempo. A implementação computacional, portanto, do GET's é uma solução conveniente. Neste caso o GET's é chamado de PcGET's.

2.3.1. Autometrics

Algoritmo para seleção automática de modelos (Doornik, 2009). Está disponível no software Oxmetrics 7.0.

Como dito acima, quando há muitos regressores, o trabalho para se realizar os GET's pode demandar bastante tempo, uma vez que o número de possíveis caminhos cresce rapidamente, já que há 2^k possibilidades. Por exemplo, se há 15 regressores, está se falando de 32.768 possibilidades! O *Autometrics* irá adotar vários caminhos de redução, não sendo necessário, portanto, percorrer os 2^k caminhos possíveis.

No fim, poderá haver mais de um modelo terminal. Para estes, casos, como será resumido em seguida, o *Autometrics* usa o critério de *Schwarz* como critério de desempate. O modelo final será aquele com o menor SC (Critério de *Schwarz*)

A principal decisão de ajuste para o algoritmo é o nível de significância α na qual a seleção ocorre. Para o presente trabalho, sempre será utilizado um nível de significância de 0.1% ($\alpha = 0,001$). Além disso, como parâmetro do *Autometrics*, foi selecionado o método de *dummies* de Saturação (IIS – *Impulse-Indicator Saturation*). No final do processo, é objetivo do algoritmo que se produza um modelo congruente, em outras palavras, sem má-especificação.

Em resumo, de acordo com Castle, Hendry, Kitov (2013), o *Autometrics* consiste nas seguintes etapas:

- (i) GUM: modelo geral irrestrito (GUM, em inglês) é o ponto de partida para a pesquisa. Ele deve basear-se em considerações teóricas gerais para aninhar o LDGP
- (ii) Pré-Busca: antes da busca propriamente dita, uma pré-busca para redução de defasagens é implementada para remover defasagens insignificantes, acelerando

os procedimentos de busca. A pré-busca só é realizada se o número de variáveis não excede o número de observações ($N < T$).

- (iii) Caminhos de Busca: O Autometrics usa uma árvore para explorar os caminhos. Começando com o GUM, o algoritmo remove a variável menos significativa determinada pela menor razão t. Cada remoção se constitui em um ramo da árvore. Para cada redução, existe uma única sub-árvore que é, então, submetida a um *backtest* contra o GUM inicial usando um teste F. Se este teste falha, nenhum sub-nó deste ramo é considerado. Este procedimento é efetuado até não haver nenhuma variável a mais que possa ser removida, ao nível de significância previamente estabelecido α . Quando isto acontece, o modelo pode ser considerado como o terminal.
- (iv) Teste Diagnóstico: cada modelo terminal é submetido a uma série de testes dado um dado nível de significância. Os testes são: Normalidade, Heterocedasticidade, teste de Chow, auto correlação residual e heterocedasticidade autoregressiva condicional.
- (v) Desempate: Como resultado da busca em árvores, múltiplos modelos terminais válidos pode ser encontrados. Como critério de desempate é utilizado o Critério de Schwarz.

2.4. Model Confidence Set

Para a seleção de modelos é utilizado o *Model Confidence Set* (MCS). Muitas vezes temos uma gama de modelos e é pertinente se perguntar qual deles é “melhor”. Hansen, Lunde, Nason (2010), afirmam que é possível reduzir sua amostra de modelos para uma amostra menor que contém os melhores modelos dado um nível de significância.

Ainda segundo Hansen, Lunde e Nason (2010) o objetivo do MCS é determinar o conjunto de modelos, M^* , que consistem nos melhores modelos dado um conjunto de modelos iniciais, M^0 , onde o termo “melhor” é definido em termos de um critério que é selecionado por quem realiza o teste.

O procedimento do MCS é baseado em um teste de equivalência, δ_M ; e uma regra de eliminação, e_M . Este teste de equivalência é aplicado ao conjunto $M = M^0$, caso δ_M seja

rejeitado, há evidências que os objetos em M não são bons e, neste caso, e_M é utilizado para eliminar um objeto que possua uma performance inferior. Este procedimento é repetido até que δ_M seja aceito. Nesta hora, MCS é definido então como o conjunto dos modelos sobreviventes.

Utilizando-se o mesmo nível de significância, α , em todos os testes, este procedimento garante que $\lim_{n \rightarrow \infty} P(M^* \subset \hat{M}_{1-\alpha}^*) \geq 1 - \alpha$, e no caso onde haja um objeto mais forte temos que $\lim_{n \rightarrow \infty} P(M^* = \hat{M}_{1-\alpha}^*) = 1$.

O procedimento MCS também retorna um p-valor para cada um dos objetos. Para um dado objeto, $i \in M^0$, o p-valor do MCS, \hat{p}_i , é o limite para o qual $i \in \hat{M}_{1-\alpha}^*$, se e somente se $\hat{p}_i \geq \alpha$. Consequentemente, um objeto com um pequeno p-valor para MCS é improvável que seja uma das melhores alternativas de M^0 .

3. METODOLOGIA E DESCRIÇÃO DA BASE DE DADOS

3.1.O que é o *Google Trends*?

É um índice (<https://support.google.com/trends>) que, ao longo do tempo, mensura o volume de pesquisas realizadas pelos usuários do Google, para uma determinada área geográfica.

Este índice é obtido dividindo-se o número de pesquisas da palavra desejada pelo número total de pesquisas realizadas. Por fim, o índice é normalizado. O maior número para o índice é 100 e o menor é 0.

Ainda, para o cálculo do índice, é escolhida uma amostra sem viés de pesquisas do Google. Como forma de ajuste, o Google desconsidera pesquisas realizadas por um mesmo I.P..

Os dados estão disponíveis desde 1 de janeiro de 2004 e é possível selecionar o índice representativo apenas do território brasileiro.

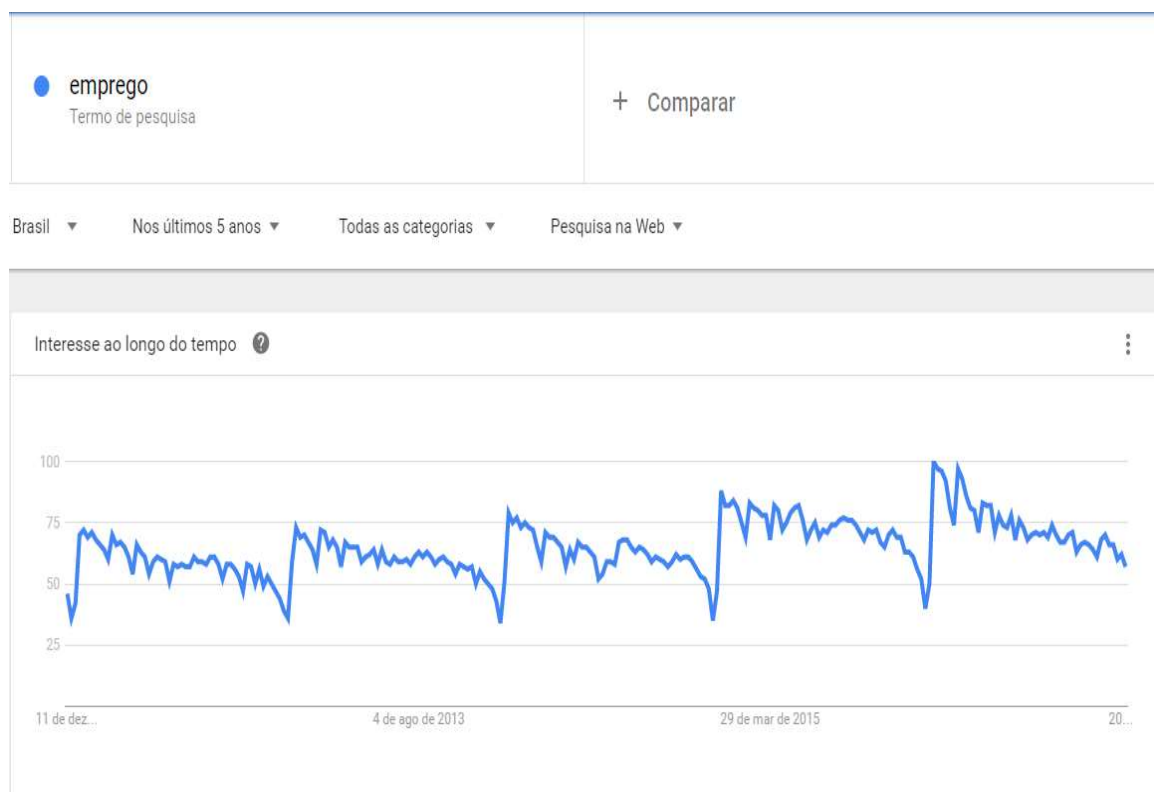


Figura 3: Índice para palavra “emprego” no Brasil

3.2. Selecionado as Palavras

Para selecionar as palavras, foi utilizado outra ferramenta do Google: *Google Correlate*. Em <https://www.google.com/trends/correlate/faq>, é possível obter informações sobre este outro instrumento. Através do uso *Google Correlate* é possível achar buscas no Google que possuam padrão similar ao longo do tempo com alguma série em particular. Em outras palavras, é possível identificar séries que possuam correlação elevada com a série de interesse, independente se esta correlação seja espúria ou não.

É possível ainda, inserir no próprio site séries customizadas pelo usuário e escolher para qual área geográfica se deseja fazer a pesquisa. No presente trabalho, duas séries serão inseridas: o próprio CAGED e a primeira diferença do CAGED.

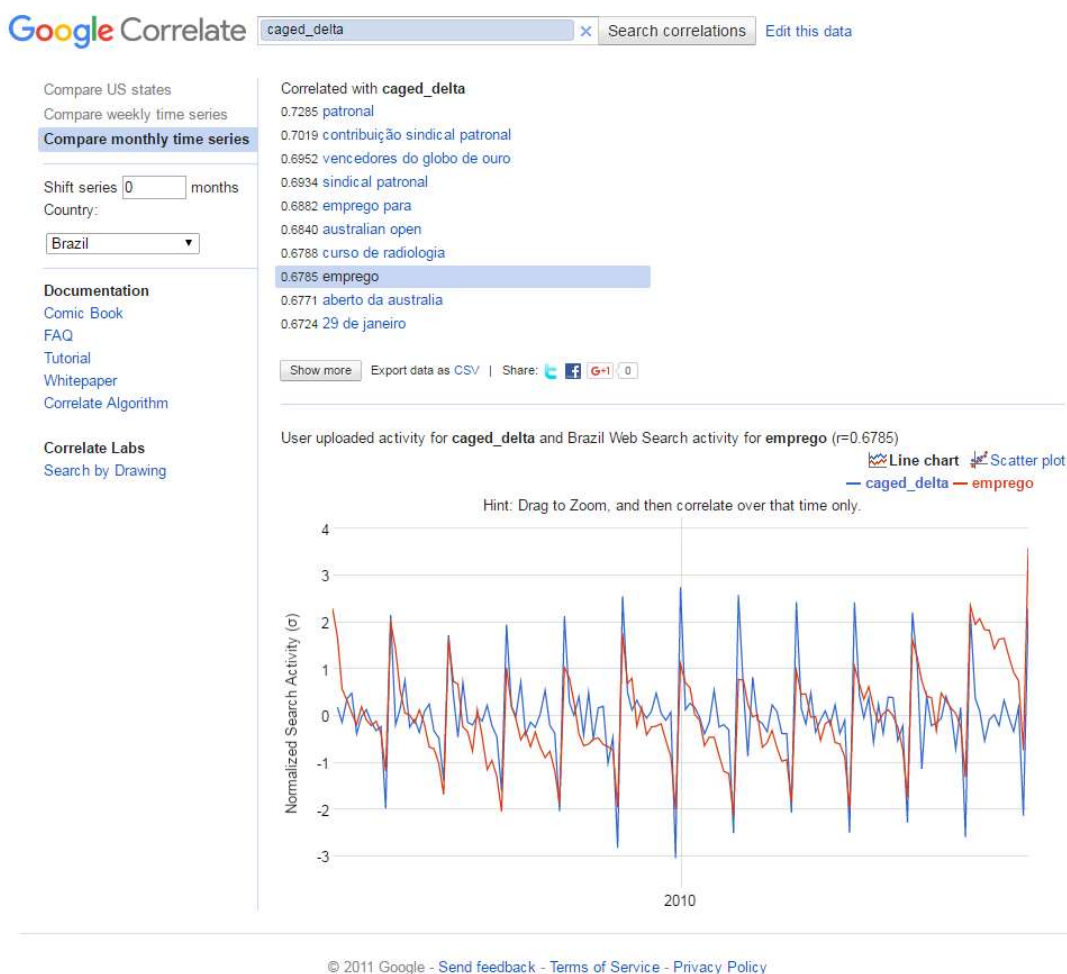


Figura 4: Google Correlate para a série da primeira diferença do CAGED

Das palavras que apareceram com maior correlação com as séries inseridas, foram escolhidas aquelas com algum sentido econômico. Por exemplo, como visto na Figura acima, não faz sentido econômico, no contexto CAGED, a pesquisa “vencedores do globo de outro”, apesar de possuir correlação elevada com a série inserida.

É de se esperar, no entanto, que quando alguma pessoa busque emprego na internet, ela digite alguma palavra relacionada ao tema. Pode-se citar exemplos de palavras: emprego, currículo, vaga de emprego.

Foram escolhidas, portanto, as sete palavras que fizessem algum sentido econômico e possuíssem maior correlação com as séries inseridas. São elas: emprego, patronal, rh, empresas em, trabalho em, treinamentos, emprego para.



Figura 5: Seleção de Palavras do Google

3.3.Tratamento do Índice extraído do *Google Trends*

A frequência do índice do *Google Trends* é semanal. No entanto, a periodicidade do CAGED é mensal. Como conciliar estas informações? Neste trabalho, para cada palavra, cada mês foi dividido em quatro semanas.

Há uma hipótese importante para a transformação dos dados: todos os dias, sejam eles dia de semana ou final de semana, possuem o mesmo peso. Por exemplo, se na semana o número índice foi 67, todas os dias desta semana possuem o mesmo índice. Desta forma, foi possível construir quatro números índices fazendo uma média aritmética, para cada palavra, por mês, da seguinte maneira:

1ª semana: do dia 1 ao dia 7

2ª semana: do dia 8 ao dia 14

3ª semana: do dia 15 ao dia 21

4ª semana: do dia 22 até o final do mês

Além disso, foi utilizado a primeira diferença do número índice, já que 25 das 28 variáveis, de acordo com o teste *Augmented Dickey-Fuller*, possuem séries não estacionárias. Os resultados podem ser vistos abaixo na tabela. As variáveis que aparecem na figura em amarelo são as estacionárias. Uma vez que, o p-valor é muito baixo, portanto posso rejeitar a hipótese H_0 de que a série seja não estacionária.

Tabela 1: Teste Raiz Unitária para Variáveis do *Google Trends*

Palavra	Semana	Nome Variável	P-Valor
Emprego	Semana 1	VAR1	0.994627
	Semana 2	VAR2	0.917747
	Semana 3	VAR3	0.875136
	Semana 4	VAR4	0.710661
Patronal	Semana 1	VAR5	0.419165
	Semana 2	VAR6	0.004283
	Semana 3	VAR7	0.620818
	Semana 4	VAR8	0.707918
RH	Semana 1	VAR9	0.905051
	Semana 2	VAR10	0.93299
	Semana 3	VAR11	0.820185
	Semana 4	VAR12	0.870124
Empresas em	Semana 1	VAR13	0.565071
	Semana 2	VAR14	0.61877
	Semana 3	VAR15	0.617498
	Semana 4	VAR16	0.772278
Trabalho em	Semana 1	VAR17	0.485287
	Semana 2	VAR18	0.502476
	Semana 3	VAR19	0.568789
	Semana 4	VAR20	0.819226
Treinamentos	Semana 1	VAR21	0.661445
	Semana 2	VAR22	1.43E-19
	Semana 3	VAR23	0.283785
	Semana 4	VAR24	9.79E-14
Emprego para	Semana 1	VAR25	0.896874
	Semana 2	VAR26	0.50863
	Semana 3	VAR27	0.743356
	Semana 4	VAR28	0.862439

3.4.CAGED e Tratamento do CAGED

O CAGED indica a diferença entre admissões e desligamentos para um determinado mês. É disponibilizado pelo ministério do trabalho em seu site (<http://pdet.mte.gov.br/caged?view=default>). “o CAGED ... foi criado como instrumento de acompanhamento e de fiscalização do processo de admissão e de dispensa de trabalhadores regidos pela CLT, com o objetivo de assistir os desempregados e de apoiar medida contra o desemprego...”.

As declarações de estabelecimentos com movimentação são prestadas até o dia 7 do mês subsequente, exceto para o trabalhador em percepção do Seguro-Desemprego, cuja declaração deve ser enviada no dia do início das atividades.

A periodicidade do dado é mensal e sua abrangência é todo o território nacional. Há cerca de 900 mil estabelecimentos declarantes por mês. Além do saldo, há ainda a informação sobre os salários do fluxo de admitidos e desligados, e não da totalidade do estoque de trabalhadores

Ainda é possível obter a quebra do saldo por setor da economia. São eles: Extrativa Mineral, Indústria de Transformação, Serviços Industriais de Utilidade Pública, Comércio, Serviços, Administração Pública, Agropecuária e Outros.

No presente trabalho, será utilizado apenas o saldo total, uma vez que o objetivo do trabalho é estimar o CAGED consolidado. Por fim, para as simulações será realizada uma transformação do número. Ao invés de se utilizar o próprio número do CAGED como divulgado pelo ministério do trabalho, será utilizada a sua primeira diferença, delta CAGED, pelo fato do CAGED ser uma série não estacionária de acordo com o teste *Augmented Dickey-Fuller*.

Tabela 2: Resultado Teste Raiz Unitária para série do CAGED

Variável	P-Valor
CAGED	0.336619

3.5. Variáveis Econômicas e Tratamento das Variáveis

No presente trabalho, foram utilizadas variáveis macroeconômicas, cujas séries também sejam de periodicidade mensal e que fossem disponíveis desde janeiro de 2004, em linha com o *Google Trends* que, como dito acima, começou a ser divulgado neste período.

Além disso, para as séries que possuíam vários componentes, foram selecionados aqueles relativos a emprego. No final, segue, na tabela abaixo, as variáveis macroeconômicas que foram utilizadas neste trabalho:

Tabela 3: Variáveis Macro Econômicas

Nome da Série	Fonte	Periodicidade	Estacionaridade	Dados mais recente quando da Divulgação do CAGED referente ao mês t
PIM: Produção Industrial Mensal	IBGE	Mensal	Não Estacionária	t_{-1}
Indicadores Industriais - Indústria de Transformação: Componente de Emprego	CNI	Mensal	Não Estacionária	t_{-1}
Indicadores Industriais - Indústria de Transformação: Componente de Horas Trabalhadas	CNI	Mensal	Não Estacionária	t_{-1}
Levantamento de Conjuntura: Componente de Total de Pessoal Ocupado	FIESP	Mensal	Não Estacionária	t_{-1}
Sondagem da Indústria - FGV: Índice de Confiança	FGV	Mensal	Não Estacionária	t
Sondagem da Indústria - FGV: Nível de Utilização da Capacidade Instalada	FGV	Mensal	Estacionária	t
Índice de Confiança do Consumidor	FCESP	Mensal	Não Estacionária	t
BACEN - Índice de Atividade Econômica	BACEN	Mensal	Não Estacionária	t_{-1}
Índice de Volume de Vendas no Varejo - Brasil	IBGE	Mensal	Não Estacionária	t_{-1}

Na tabela abaixo, encontra-se o resultado do teste *Augmented Dickey-Fuller* para verificar a estacionaridade ou não da série. Para as séries não estacionárias, foi considerado a sua variação e não o nível.

Tabela 4: Resultado Teste Raiz Unitária para Variáveis Macroeconômicas

Variável	Nome Variável	P-Valor	Estacionaridade
i	PIM	0.673385	não estacionária
ii	CNI: Componente de Emprego	0.994227	não estacionária
iii	CNI: Componente de Total de Horas Trabalhadas	0.771259	não estacionária
iv	FIESP: Total de Pessoal Ocupado	0.992557	não estacionária
v	FGV: Índice de Confiança	0.4129	não estacionária
vi	FGV: NUCI	0.052736	estacionária
vii	FCESP: Confiança do Consumidor	0.793475	não estacionária
viii	IBC-Br	0.959144	não estacionária
ix	PMC	0.998303	não estacionária

3.6. Modelo *Naive*

Como dito anteriormente, com o auxílio do *Autometrics*, o modelo selecionado utilizando as variáveis do Google como variáveis exógenas, inicialmente, foi comparado a um modelo *naive* que considerou apenas a própria defasagem do CAGED e as *dummies* sazonais.

A pergunta que se colocou anteriormente, no entanto, foi qual modelo *naive* escolher. Neste trabalho, quatro modelos *naives* foram testados: (i) AR1 sem as *dummies* sazonais, (ii) AR1 com as *dummies* sazonais, (iii) AR12 sem as *dummies* sazonais e (iv) AR12 com as *dummies* sazonais.

Dentre os quatro modelos, o modelo AR1 com as *dummies* sazonais, de acordo com o MCS, apresentou o melhor desempenho. Neste modelo, o CAGED em um mês t é previsto utilizando-se do CAGED no mês $t-1$.

$$y_t = \beta_0 + \sum_{i=1}^{10} \beta_i D_{it} + \beta_{11} y_{t-1} + a_t$$

3.7. Benchmark de Mercado: Projeção dos Analistas

Neste trabalho, como *benchmark* da projeção de mercado, será utilizada a mediana das projeções dos analistas disponíveis na *Bloomberg* na hora da divulgação do dado do CAGED.

Como é possível ver na figura abaixo, está disponível, para cada mês, além da mediana (*Median Estimate*), a média (*Average Estimate*), a maior estimativa (*High Estimate*), menor estimativa (*Low Estimate*) e o número de analistas (*Number of Estimates*) que disponibilizaram sua projeção. É possível também saber individualmente a projeção de cada analista.



Figura 6: Tela da Bloomberg com informações sobre as projeções dos Analistas

Fonte: *Bloomberg*

3.8. Tamanho da Base de Dados e Período Analisado

Como dito acima, no presente trabalho, dado que o *Google Trends* é disponível apenas a partir de 2004, foi utilizado para as contas os dados mensais do CAGED de janeiro de 2004 até julho de 2016. Para as demais variáveis macro, também foi utilizado a mesma janela temporal.

Foram realizadas previsões mensais fora da amostra para o período compreendido entre janeiro de 2012 e julho de 2016, totalizando 55 previsões.

3.9. Variáveis Explicativas Elegíveis e Modelo Google

Na primeira etapa, onde se buscará o melhor modelo para previsão do CAGED utilizando apenas as informações do Google, para cada iteração, haverá as seguintes variáveis explicativas elegíveis: (i) três primeiras defasagens do delta CAGED, (ii) as *dummies* sazonais e, (iii) para cada uma das sete palavras selecionadas acima, 4 variáveis, representando as quatro semanas do mês, bem como, suas respectivas três primeiras defasagens. Ou seja, para cada iteração, o GUM inicial possui mais de 80 variáveis.

Na segunda etapa, haverá como variáveis explicativas elegíveis, além das citadas acima, as variáveis macroeconômicas e suas respectivas três primeiras defasagens. Nesta etapa, para cada iteração, o tamanho do GUM inicial é superior a da primeira etapa e o número de variáveis a cada iteração passa de 110.

Desta forma, não haverá um modelo Google definido e que valha a todo instante do tempo. Uma vez que, para cada iteração, através dos parâmetros selecionados para o algoritmo do *Autometrics*, variáveis diferentes podem ser selecionadas para compor o modelo final.

4. RESULTADOS

Como dito acima, podemos dividir o presente trabalho e, consequentemente, os resultados em três etapas:

- (i) Através do uso do *Autometrics*, encontrou-se o melhor modelo utilizando as palavras do Google e, em seguida, comparou-se com um modelo *naive*, neste caso, um modelo AR(1) com *dummies* sazonais;



Figura 7: Selecionando o Modelo utilizando as variáveis do *Google Trends*

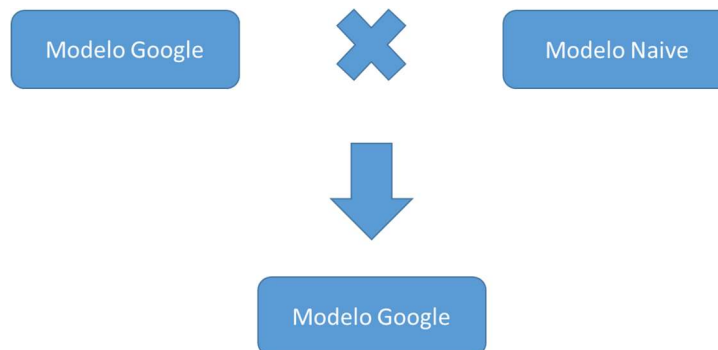


Figura 8: Seleção entre o Modelo Google e o modelo *Naive*

- (ii) Com este melhor modelo em mãos, foi verificado se a inclusão de variáveis macroeconômicas diminuía o erro de previsão do Modelo;



Figura 9: Selecionando o modelo com variáveis do *Google Trends* e Macroeconômicas

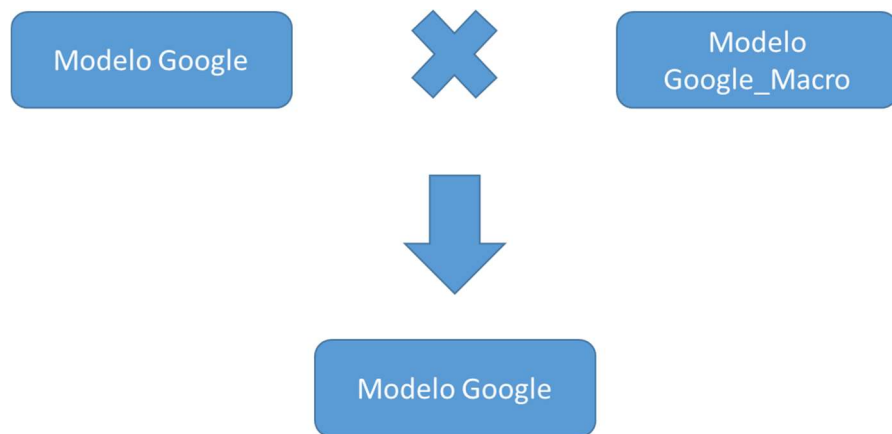


Figura 10: Seleção entre Modelo Google e Modelo Google_Macro

- (iii) Comparou-se o modelo “vencedor” da segunda etapa com um *benchmark* de mercado: projeção dos analistas disponível na *Bloomberg* na hora da divulgação do CAGED;

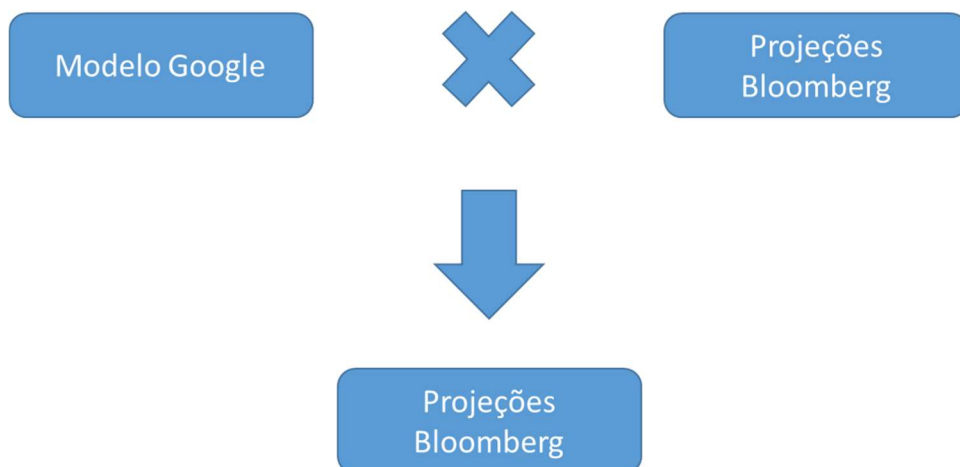


Figura 11: Seleção entre o Modelo Google e Previsões dos Analistas de Mercado

Para a primeira etapa, para cada mês previsto, o *Autometrics* busca a melhor combinação possível dentre as variáveis disponíveis até o mês em questão. Para cada mês também foi calculada a previsão para o CAGED utilizando o modelo *naive*, neste caso, AR(1) com as *dummies* sazonais. Com a previsão em mãos, para a comparação de modelos foi utilizado o critério do MCS. A Tabela 5 mostra o P-value para cada um dos dois modelos. Vale frisar que, de acordo ainda com o MCS, o modelo com p-value 1 é dito melhor. Logo, utilizando-se o MCS como critério de escolha, o modelo que utiliza o Google supera o modelo *naive*.

Tabela 5: Resultado da Primeira Etapa

	mse(*10 ³)	p-value
Modelo Google	3,051.24	1.0000
Modelo AR com dummies sazonais	3,787.09	0.1836

Para a segunda etapa, para cada mês previsto, além do modelo Google selecionado acima, o *Autometrics* buscou a melhor combinação possível dentre as variáveis disponíveis até o mês em questão. A diferença agora é que, dentre as variáveis disponíveis, além das variáveis Google, há as variáveis macroeconômicas apresentadas acima e suas primeiras três defasagens. Novamente, com a previsão em mãos, para a comparação de modelos foi utilizado o critério MCS. Resultado encontra-se na tabela abaixo.

Tabela 6: Resultado de Segunda Etapa

	mse(*10 ³)	p-value
Modelo Google	3,051.24	1.0000
Modelo Google+Variáveis_Economicas	3,366.83	0.5952

Como pode ser visto, a adição de variáveis macro não melhorou o modelo. A princípio, a intuição indicaria que a adição das variáveis macroeconômicas deveria melhorar a capacidade de previsão do modelo. Há alguma maneira de ratificarmos o resultado indicado pelo MCS?

Para responder esta pergunta, verificou-se que, para um grau de confiança de 95%, os p-valores dos coeficientes das variáveis macro regredidas contra o erro do modelo Google calculado na primeira etapa são superiores a 0,05, como pode ser visto na Tabela 7. Ou seja, não são significativos. Para algumas variáveis foi possível regredir o erro apenas contra suas defasagens, uma vez que, como visto na descrição da base de dados, para algumas variáveis macro, para a data t , ainda não há a correspondente divulgação.

Tabela 7: Coeficientes das Variáveis Macro regredidas contra o Erro do Modelo Google

Variável	Defasagem	p-valor Coeficiente
IBGE - PIM	0	N.D.
	1	0.5406
	2	0.645
	3	0.6456
CNI - Componente de Emprego	0	N.D.
	1	0.239
	2	0.9809
	3	0.7611
CNI - Componente de Horas Trabalhadas	0	N.D.
	1	0.3939
	2	0.4335
	3	0.7608
FIESP - Componente de Total de Pessoal Ocupado	0	N.D.
	1	0.4831
	2	0.4967
	3	0.9334
FGV - Sondagem da Indústria	0	0.8109
	1	0.1272
	2	0.6696
	3	0.9709
FGV - NUCI	0	0.8392
	1	0.7215
	2	0.2868
	3	0.3778
FCESP - Índice de Confiança do Consumidor	0	0.3079
	1	0.5059
	2	0.9601
	3	0.5852
BACEN- IBC-br	0	N.D.
	1	0.5536
	2	0.2064
	3	0.3895
IBGE - Vendas no Varejo	0	N.D.
	1	0.4602
	2	0.0961
	3	0.4372

Para a terceira e última etapa, através do MCS, comparou-se o modelo calculado na primeira etapa com um benchmark de mercado, neste caso a projeção dos analistas disponíveis na *Bloomberg* na hora da divulgação do CAGED. Como pode ser visto na Tabela 8 abaixo, neste caso, o modelo Google calculado na primeira etapa e usado agora não foi capaz de superar o *benchmark* de acordo com o MCS.

Tabela 8: Resultado da Terceira Etapa

	mse(*10 ³)	p-value
Modelo Google	3,051.24	0.2083
Benchmark Bloomberg	2,507.92	1.0000

5. CONCLUSÃO

No presente trabalho testou-se se o uso de dados de pesquisas na rede, *Google Trends*, como variáveis exógenas, para prever o CAGED é superior a modelos que usam apenas as próprias defasagens. Os resultados encontrados são positivos. Para a seleção dos modelos que utilizavam as variáveis do *Google Trends* foi utilizado o algoritmo *Autometrics*.

Em um primeiro momento, foi verificado que o modelo que utiliza o *Google Trends* e o *Google Correlate* tem desempenho superior ao modelo AR(1) com variáveis de ajuste sazonal de acordo com o critério de seleção do *Model Confidence Set*. No entanto, a dúvida que surgiu foi se haveria como melhorar ainda mais este modelo. O que se pensou foi a inclusão de dados macroeconômicos.

Em um segundo momento, portanto, ainda utilizando o MCS, surpreendentemente, a inclusão de variáveis macroeconômicas não aumentou a capacidade preditiva do modelo.

Por fim, a mediana da previsão dos analistas disponíveis na *Bloomberg* se apresenta, no entanto, como a melhor estimativa para o CAGED, superando, então, o modelo encontrado na primeira etapa de acordo com o critério de seleção MCS. Com o objetivo de tentar explicar o resultado desta última etapa, o autor tentou buscar na literatura alguma possível explicação.

Bates and Granger (1969), Clemen (1989) concluem que uma combinação de previsões pode possuir um erro quadrático médio menor que as previsões originais. Mais recentemente, Makridakis and Hibon (2000), analisando um maior número de métodos, também chegam a conclusão vários modelos, quando combinados, possuem melhor precisão, na média, que os modelos individuais.

Esta, portanto, pode ser uma das explicações pela qual a mediana da previsão dos analistas disponíveis na *Bloomberg* é superior, de acordo com o MCS, ao modelo que utiliza o *Google Trends* como variável exógena, mesmo que o modelo que utiliza o *Google Trends* não componha esta mediana.

Outra possível explicação poderá ser encontrada no tópico a seguir.

6. LIMITAÇÕES E POSSÍVEIS EXTENSÕES

Como possível limitação pode-se citar o fato de que no presente trabalho assumiu-se que, independente do dia da semana pesquisado, estes dias apresentariam o mesmo peso. Pode-se verificar na realidade que haja mais procuras nos finais de semana, alterando-se assim os pesos relativos entre os dias.

Além disso, sobre o fato do modelo Google não ser capaz de bater o *benchmark Bloomberg*, pode-se argumentar que há meses nos quais o ministro do trabalho, responsável pela divulgação do dado do CAGED, antecipa alguma informação sobre o número a ser divulgado, como aconteceu no mês de junho de 2016.

No dia 17 de junho de 2016, falando sobre o CAGED referente à maio que seria divulgado na tarde do dia de 24 junho, o ministro antecipou que haveria um aumento do número de desempregados no país em comparação com os meses anteriores, o que de fato ocorreu. (<http://oglobo.globo.com/economia/ministro-do-trabalho-antecipa-que-houve-aumento-no-numero-de-desempregados-no-pais-16470122?versao=amp>).

Desta forma, houve, naquele dia, a possibilidade de algum analista que possuísse alguma projeção mais otimista alterasse sua projeção, tendo em vista que a *Bloomberg* permite que se altere a projeção em seu sistema até um dia útil anterior a divulgação do dado.

7. REFERÊNCIAS

BANBURA, M., D. GIANNONE, M. MODUGNO, AND L. REICHLIN (2013). Now-casting and the real-time data-flow. In G. Elliot and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, pp. 195-237. Amsterdam: Elsevier-North Holland.

CASTLE, JENNIFER L., HENRY, KITOV, OLEG I. Forecasting and Nowcasting Macroeconomic Variables: A methodological Overview. Discussion Paper Series, University of Oxford, n. 674, 2013.

CHADWICK, MG. and SENGUL, G (2012) Nowcasting Unemployment Rate in Turkey: Let's Ask Google Working Papers 1218, Research and Monetary Policy Department, Central Bank of the Republic of Turkey.

CHOI, H. and VARIAN, H. 2012. "Predicting the Present with Google Trends", *Economic Record*, 88: 2–9.

CLEMN, R. T. (1989) Combining forecasts: a review and annotated bibliography, *International Journal of Forecasting*, 5, 559-83.

D'AMURI, F. and MARCUCCI, J., 2012. "The predictive power of Google searches in forecasting unemployment," Temi di discussione (Economic working papers) 891, Bank of Italy, Economic Research and International Relations Area.

DOORNIK, J. A. (2009). Autometrics. In Honour of David Hendry.

DOORNIK, J. A. and HENDRY, D. A.. "Empirical Econometric Modelling – PC GIVE™ 14: Vol I".

HANSEN, P. R., LUNDE, A., & NASON, J. M. (2010). The model confidence set. Available at SSRN 522382.

MAKRIDAKIS, S. and HIBON, M. (2000), "The M3-Competition: Results, Conclusions and Implication," International Journal of Forecasting, 16, 451-476.

PERLIN, MARCELO; PORTELA, A. ; CALDEIRA, J. ; PONTUSCHKA, MARTIN . Can we predict the financial markets based on Google's search queries?. In: 15º Encontro Brasileiro de Finanças, 2015, São Paulo - SP. Anais do 15º Encontro Brasileiro de Finanças, 2015.

TUHKURI, J. (2014). Big Data: Google Searches Predict Unemployment in Finland. ETLA Reports 31.

TUHKURI, J. (2015). Big Data: Google Searches Predict Unemployment in Finland. *NTTS 2015 Conference Proceedings*.

TUHKURI, J. (2016). Forecasting Unemployment with Google Searches. *ETLA Working Papers* 35.

BATES, J.M. and GRANGER, C.W.K. (1969). The Combination of Forecasts. *Operations Research Quarterly* 20, 451-468.