Pontifícia Universidade Católica
do Rio de Janeiro

Érico de Mello Penteado

Nowcast de Desemprego usando Google Trends, uma revisão da temática pós-pandemia

Monografia de Final do Curso

Orientador Prof. Gustavo Gonzaga
Co-orientador Dr. Henrique Pires

Rio de Janeiro
Outubro de 2023

Declaro que o presente trabalho é de minha autoria e que não recorri para realizá-lo, a nenhuma forma de ajuda externa, exceto quando autorizado pelo professor tutor.

As opiniões expressas neste trabalho são de responsabilidade única e exclusiva do autor.

Érico de Mello Penteado

Aos meus pais, namorada, irmãs e família
pelo apoio e encorajamento.

Agradecimentos

Ao meu orientador Professor Gustavo Gonzaga pela parceria e paciência na confecção desse trabalho.

Ao meu co-orientador Henrique Pires por ser muito mais do que um chefe e co-orientador. Tenho orgulho de ter aprendido tanto em tão pouco tempo. Obrigado.

Aos meus amigos Caio, Thiago, Renzo, Luca, Gustavo e Leonardo por todo apoio, paciência e compreensão.

Aos meus pais pelo apoio, paciência e amor.

Aos meus colegas de turma da PUC-Rio. Sem vocês, aprender seria um pouco menos divertido.

Aos meus colegas de trabalho da Gávea Investimentos e Gap Asset pelo auxílio e aprendizado.

A todos os professores e funcionários do Departamento pelos ensinamentos e pela ajuda.

A todos os amigos e familiares que sempre se fizeram presentes em minha vida.

À minha namorada Ana Duarte, pela compreensão e apoio nas horas mais complicadas. Amo-te.

Ao meu pai Arlindo, por me ensinar e suportar durante 23 anos. Sem você, nada disso faria sentido, és minha inspiração em todos os momentos.

Sumário

Lista de figuras

Lista de tabelas

Minha inspiração

Fernando Sabino, *De tudo, ficaram três coisas: a certeza de que ele estava sempre começando, a certeza de que era preciso continuar e a certeza de que seria interrompido antes de terminar. Fazer da interrupção um caminho novo. Fazer da queda um passo de dança, do medo uma escada, do sono uma ponte, da procura um encontro.*

1
Resumo

Essa monografia pretende gerar uma estimação robusta a partir de dados de pesquisa na *search engine* da Google, disponibilizados na API Google Trends, para a série de Desocupação da PNAD Contínua. Para computar as previsões será utilizada uma grande base de dados, contendo diversos termos de pesquisa associados a emprego, vagas, programas governamentais, remuneração e outras variáveis macroeconômicas tradicionais. Além disso, seguindo a metodologia do *paper* Medeiros e Pires (2021) usaremos um *framework* de estimações diretas com o objetivo de termos uma avaliação fora da amostra confiável e independente de cenários sobre as covariadas. Nesse *setup*, conseguiremos comparar a performance de todos os modelos estimados, de modo a averiguar se os nossos melhores seriam capazes de superar os modelos *benchmark* por algumas das métricas de erro mais usuais.

2
Introdução

No âmbito das previsões econômicas, a construção de modelos eficazes demanda não apenas a adoção de uma metodologia robusta, mas também o acesso a uma base de dados sólida e de alta qualidade. Em setores nos quais os dados governamentais disponíveis, divulgados mensalmente, apresentam defasagens consideráveis (Choi e Varian (2009)), torna-se imperativo buscar fontes alternativas de informação para manter-se atualizado diante das dinâmicas do mercado. É precisamente nesse contexto que se destaca o conceito de *Nowcasting*.

O termo *Nowcasting* é uma combinação das palavras *Now* (agora) e *Forecasting* (previsão), refletindo a ideia de previsão do presente ou de um futuro iminente. A relevância do Nowcasting para a economia, conforme abordado por Choi e Varian (2009), é novamente evidenciada em uma passagem do artigo Bánbura et al. (2013):

> Now-casting is relevant in economics because key statistics on the present state of the economy are available with a significant delay.

Nesse contexto, o Google, como a principal ferramenta de pesquisa global, surge como uma fonte potencialmente rica de dados em alta frequência. As pesquisas realizadas por usuários no Google Trends refletem diretamente o interesse atual em diversos tópicos. A partir dessa premissa, emerge a perspectiva de que uma abordagem promissora para contornar a defasagem nas variáveis macroeconômicas seja a utilização dos dados da API do Google (Google Trends, s.d.) como complemento às variáveis tradicionais da teoria macroeconômica, visando aprimorar os modelos de previsão, especialmente no que tange ao mercado de trabalho.

Assim, a utilização de dados de alta frequência é apenas uma das metades essenciais do trabalho em questão. O uso de modelos simples, tais quais o OLS (*Ordinary Least Squares*), têm sua eficácia reduzida quando aplicados em problemas de **alta dimensão**. Nesse contexto, surge uma série de modelos de *Machine Learning*, que, via projeção direta, poderiam gerar estimações mais consistentes.

Os mais famosos, ex-OLS, talvez sejam os modelos de regularização, que focaram em solucionar o *trade-off* **Viés-Variança** ao atribuir restrições e/ou penalizações aos parâmetros. Dentre esses, alguns são: *Least Absolute Shrinkage and Selection Operator* (**LASSO**) proposto pela primeira vez em Tib-

shirani (1996) e o **RIDGE** citado em Hoerl e Kennard (1970). No entanto, no contexto de *Big Data*, métodos não-lineares também produzem estimações consistentes. Exploraremos esses e mais modelos comparando-os com os *benchmarks* escolhidos - **AR** e **Random Walk**.

Assim, este trabalho tem como objetivo central investigar se a incorporação de diferentes séries temporais do Google Trends como covariáveis adicionais às variáveis macroeconômicas convencionais pode efetivamente aprimorar os modelos de previsão no contexto do mercado de trabalho. Para alcançar tal objetivo, este estudo se estrutura da seguinte maneira: Apresento minha motivação para confecção do trabalho no Capítulo 3. No Capítulo 4, realizamos uma revisão da literatura relevante para nossa proposta. No Capítulo 5, detalhamos o método proposto para a análise. No Capítulo 6, apresentamos os resultados obtidos com base em nossa pesquisa. Por fim, no Capítulo 7, destacamos as conclusões derivadas das análises realizadas e delineamos as implicações de nossas descobertas para o campo do Nowcasting no mercado de trabalho.

3
Motivação

A motivação para a realização deste trabalho surge da observação da demora na disponibilização dos dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), que pode prejudicar a tomada de decisões que demandam informações rápidas e precisas.

Nesse contexto, a manipulação de dados que apresentam defasagem em relação às condições econômicas pode impactar substancialmente a eficácia de utilização dos modelos de previsão. Isso ocorre devido à rápida evolução das condições, o que torna as previsões baseadas em dados desatualizados inadequadas para refletir a realidade econômica. Portanto, torna-se imperativo buscar fontes de dados alternativas, igualmente confiáveis e atualizadas de acordo com as demandas do mercado.

A utilização de dados de alta frequência, em especial o Google Trends - objeto de foco deste estudo - em conjunto com variáveis macroeconômicas, emerge como uma potencial ferramente para obter estimativas que capturem com precisão as condições atuais. Esta abordagem, alinhada ao conceito de *Nowcasting*, se revela fundamental para atender às exigências de previsão no cenário econômico atual.

Durante minha pesquisa, deparei-me com uma monografia de conclusão de curso intitulada **"Nowcasting do desemprego com Google Trends: evidências do mercado de trabalho brasileiro"** de Ludwig e Souza (2016). Embora essa monografia tenha oferecido uma análise abrangente, identifiquei a oportunidade de aprimorá-la com novos modelos, métodos e dados mais recentes, principalmente sobre o contexto de *Big Data*. Além disso, a pandemia de COVID-19, ocorrida após a conclusão da monografia em 2016, não havia sido considerada em seus modelos, tornando este trabalho uma contribuição importante para atualizar as questões já exploradas e aumentar a capacidade preditiva, levando em conta os impactos da pandemia.

4
Revisão de Bibliografia

Neste capítulo, realizaremos uma revisão da literatura relevante sobre o uso do Google Trends como fonte de dados para modelos econométricos. Abordaremos os estudos pioneiros, críticas e pesquisas mais recentes que utilizam-se dessa mesma base de dados.

4.1
Primeiros Estudos

Os primeiros defensores da utilização do Google Trends como base de dados para modelos econométricos foram Choi e Varian (2009). Eles empregaram modelos ARIMA para investigar se o uso de consultas de pesquisa para termos como **seguro-desemprego** (destacado por sua proximidade com o mercado de trabalho) e outros, poderia melhorar o desempenho fora da amostra de algumas variáveis macroeconômicas.

Apesar de ter sido uma inovação na área de *Nowcasting* utilizando dados do Google Trends e ter servido de inspiração para diversos outros pesquisadores, como Zimmermann e Askitas (2009), D'Amuri e Marcucci (2023) e Suhoy (2009), o estudo de Choi e Varian (2009) está desatualizado. Desde então, a versão da API mudou (anteriormente chamada de Google Insight) e as pesquisas se tornaram ainda mais frequentes devido à disseminação dos *smartphones*. Nesse contexto, nos concentraremos em estudos mais recentes.

4.2
Abordagens e Desafios

Como crítica ao uso inadequado das séries do Google Trends, Nagao, Takeda e Tanaka (2019) procuram mostrar que há limitações quanto à frequência das consultas de pesquisa em contribuir para a precisão dos modelos. Eles testaram essas consultas em relação a modelos AR, tendo como alvo a variável de desemprego nos Estados Unidos da América.

Dimpfl e Bleher (2022) criaram um índice chamado IPSO (Index of Prices Searched Online) para comparar diferentes amostragens temporais das consultas de pesquisa. O estudo relatou melhorias na previsão da inflação mensal e do consumo mensal tanto nos Estados Unidos quanto na Europa ao utilizar dados do Google Trends.

Fondeur e Karamé (2013) chegaram a conclusão que os dados de *query's* do *Google Trends* melhoram a previsão de desemprego na França (para pessoas

de 15 anos até 24 anos).

Outros pesquisadores, como Kohns e Bhattacharjee (2023), buscaram uma análise Bayesiana dos sentimentos das consultas de pesquisa para prever o crescimento do PIB nos Estados Unidos da América. Como resultado, eles apresentaram que essas consultas representam informações importantes para antecipar a variável PIB, que possui uma frequência trimestral. Segundo o estudo, os termos de pesquisa podem sinalizar *Wealth Effects* e ansiedade econômica.

Outros, como Borup, Rapach e Schütte (2023), ainda abordam o uso *daily data* do *Google Trends* para melhorar as previsões de Initial Claims (dado semanal) nos Estados Unidos. O resultado é que os modelos que incorporam os dados diários *outperformam* substancialmente aqueles que não incorporam. A capacidade preditiva ainda aumenta quanto mais *daily Google Trends datapoints* são adicionados na base. Ainda, como resultado, mostram um *link* relevante entre a relevância do uso de *Google Trends* para previsões e a pandemia de COVID-19.

Alguns pesquisadores, como Woloszko (2020), apesar de usar apenas um único modelo (*neural network*), ainda citam o método de fazer médias simples de amostras para gerar uma base de dados mais fidedigna

Por fim, Medeiros e Pires (2021) demonstraram que o tratamento dos dados das consultas de pesquisa deve ser realizado por meio da coleta de amostras diárias, seguida pela aplicação da média simples entre as amostras.

O estudo também mostrou que, como a série do Google Trends é construída através de um índice variável (conceito a ser detalhado no Capítulo 5), a média de várias amostras torna a estimação mais robusta.

5
Metodologia

Neste capítulo, descreveremos a metodologia utilizada neste trabalho. Começaremos com a composição da base de dados, que é fundamental para as análises subsequentes. Em seguida, explicaremos o funcionamento do Google Trends e como essa ferramenta será empregada na pesquisa. No final, mostraremos como as previsões serão feitas.

5.1
Composição da Base de Dados

A base de dados deste projeto é composta por informações da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua). A variável alvo a ser prevista é a taxa de desocupação. Além disso, utilizaremos consultas ao Google Trends e séries temporais de variáveis macroeconômicas clássicas como variáveis explicativas para os modelos. A tabela "Base de Dados" que contém as variáveis está no final do capítulo.

No entanto, dentro do modelo também consideraremos os *lags* das variáveis e alguns componentes principais da base como covariadas, assim como destacado em Medeiros et al. (2021). O número de componentes principais é tal que a variação explicada da base como um todo seja maior que 70%. Vale resaltar que, como estamos lidando com series *real time*, foi tomado o cuidado de só usar informações disponíveis no momento das previsões. Segui o calendário do IBGE e Bloomberg.

5.2
Google Trends

Para compreendermos o Google Trends e sua relevância neste estudo, é crucial entender como essa ferramenta funciona e quais informações ela fornece.

O aspecto relevante do Google Trends para esta pesquisa é a série conhecida como *interest over time* aplicada ao Brasil para cada um dos termos selecionados. Essa série representa a frequência de pesquisas de cada termo em uma escala que varia de 0 a 100. Em resumo, o Google mede o volume de pesquisas para cada termo e o compara com o volume total de pesquisas, criando o índice mencionado anteriormente. Vale ressaltar que uma diminuição no índice não necessariamente indica uma redução no número absoluto de pesquisas para o termo, mas sim que uma parcela menor das pesquisas totais está relacionada a esse termo.

Os termos selecionados seguem quatro categorias específicas, parecidas com as existentes na monografia de conclusão de curso de Ludwig e Souza (2016)). São elas: termos relacionados a ações governamentais, termos relacionados a procura de vagas, termos relacionados a remuneração e termos relacionados aos sites de correspondência entre empregados e empregadores. Os termos utilizados estão na tabela 5.1, na última página do capítulo.

É importante destacar uma limitação ao utilizar os dados do Google Trends, conforme apontado por Medeiros e Pires (2021). Essa limitação decorre do fato de que as séries do Google Trends são baseadas em uma amostra pequena da população total de pesquisas feitas no Google para cada termo, em cada momento do tempo. Além disso, a amostra da qual cada termo é derivado é atualizada aproximadamente três vezes ao dia. Isso implica que consultas idênticas (mesmo termo e período) em dias diferentes podem resultar em séries diferentes.

Portanto, a metodologia recomendada por Medeiros e Pires (2021), que será adotada neste trabalho, é o método de amostragem. Esse método envolve a realização de múltiplas consultas ao longo de vários dias e, posteriormente, o cálculo da média simples para cada um dos termos. Neste estudo, optamos por coletar amostras ao longo de um período de 10 dias, acreditando que esse procedimento aprimora a consistência das previsões.

5.3
Modelos de Previsão

Este estudo utilizará uma variedade de modelos lineares e não-lineares no contexto de previsão direta e *Big Data*, em comparação com modelos de referência (*benchmark*). A abordagem de *Big Data* nos permite incluir um grande número de termos nas análises, diferentemente das regressões lineares ordinárias (OLS). Os modelos serão estimados tanto em *rolling windows* quanto em *expanding windows*. A seleção do melhor modelo, no *backtest*, será baseada em uma métrica de erro: Erro Quadrático Médio (RMSE).

Devido à PNAD Contínua ser composta por uma média móvel trimestral, o mesmo procedimento poderá ser aplicado às covariadas, a depender do *backtest*.

$$RMSE_{m,h} = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \hat{e}_{t,m,h}^2} \tag{5-1}$$

5.3.1
AR

Usaremos um AR de ordem p como um dos *benchmarks*. O parâmetro p será escolhido pelo teste BIC (*Bayesian information criterion*) de modo que, poderá existir um modelo para cada horizonte de previsão.

$$\hat{y}_{t+h} = \hat{\theta}_{0,h} + \hat{\theta}_{1,h} y_t + ... + \hat{\theta}_{p,h} y_{t-p-1} \tag{5-2}$$

5.3.2
Random Walk (RW)

O modelo de *Random Walk* sugere que a previsão para $h = 1, \ldots, 12$ será o valor prévio mais um termo de média da série.

$$X_{t+h} = X_t + \varepsilon_t \tag{5-3}$$

5.3.3
LASSO

Proposto pela primeira vez em Tibshirani (1996), a função de penalização segue a seguinte equação:

$$\sum_{i=1}^{n} p(\beta_{h,i}; \lambda, w_i) := \lambda \sum_{i=1}^{n} |\beta_{h,i}| \tag{5-4}$$

Por definição, o estimador de *Shrinkage* é:

$$\hat{\beta}_h = arg\min_{\beta_h} [\sum_{t=1}^{T-h} (y_{t+h} - \beta_h^{'} x_t)^2] + \sum_{t=1}^{n} p(\beta_{h,i}; \lambda, w_i)] \tag{5-5}$$

Logo, o estimador de Lasso é:

$$\hat{\beta}_h = arg\min_{\beta_h} [\sum_{t=1}^{T-h} (y_{t+h} - \beta_h^{'} x_t)^2 + \sum_{t=1}^{n} \lambda |\beta_{h,t}|] \tag{5-6}$$

5.3.4
RIDGE

Proposto pela primeira vez em Hoerl e Kennard (1970), a função de penalização segue a seguinte equação:

$$\sum_{i=1}^{n} p(\beta_{h,i}; \lambda, w_i) = \lambda \sum_{i=1}^{n} \beta_{h,i}^2 \tag{5-7}$$

Por definição, o estimador de *Shrinkage* é representado pela equação 5-5. Logo, o estimador de Ridge é:

$$\hat{\beta}_h = arg\min_{\beta_h} [\sum_{t=1}^{T-h} (y_{t+h} - \beta_h^{'} x_t)^2 + \sum_{t=1}^{n} \lambda \beta_{h,t}^2] \tag{5-8}$$

5.3.5
Adaptive Lasso (AdaLasso)

Também faz parte do ambiente de algoritmos de *Shrinkage*. Leva o nome de *Adaptive Least Absolute Shrinkage and Selection Operator* por ser uma adaptação do estimador de lasso. A função de penalização do adaLasso é igual a do Lasso, porém leva a adição de um parâmetro de *weighting*. A função de penalização é a seguinte:

$$\sum_{i=1}^{n} p(\beta_{h,i}; \lambda, w_i) = \lambda \sum_{i=1}^{n} w_i |\beta_{h,i}| \tag{5-9}$$

Onde,

$$w_t = |\beta_{h,i}^*|^{-1} \tag{5-10}$$

5.3.6
Elastic Net (ElNet)

Elastic Net é um modelo de generalização que mistura LASSO e RIDGE. A função de penalização é definida como:

$$\sum_{i=1}^{n} p(\beta_{h,i}; \lambda, w_i) = \alpha\lambda \sum_{i=1}^{n} \beta_{h,i}^2 + (1-\alpha)\lambda \sum_{i=1}^{n} w_i |\beta_{h,i}| \tag{5-11}$$

$$\alpha \in [0, 1] \tag{5-12}$$

5.3.7
Adaptive Elastic Net (AdaNet)

Proposto pela primera vez em Zou e Zhang (2009), o modelo tenta combinar os modelos Elastic Net (ElNet) e Adaptive Lasso (AdaLasso). Utilizando o método de Zou e Zhang (2009), inicialmente, o β do ElNet é calculado (con-

forme equação 5-11). Em seguida são construídos pesos adaptativos obedecendo a fórmula, considerando $-\gamma$ como uma constante positiva.

$$\hat{w}_j = (|\hat{\beta}_j(elnet)|)^{-\gamma} \tag{5-13}$$

$$j = 1, 2, \ldots, p \tag{5-14}$$

Para achar o $\hat{\beta}(AdaNet)$ o problema de otimização abaixo é resolvido:

$$arg\min_{\beta} \left\{ (1 + \frac{\lambda_2}{n}) \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1^* \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right\} \tag{5-15}$$

5.3.8
Complete Subset Regression (CSR)

Originalmente proposto por Elliott, Gargano e Timmermann (2013), o modelo foi desenvolvido para resolver o problema de escolher o conjunto ideal de variáveis x_t para prever $y_t + h$, uma tarefa que se torna computacionalmente inviável ao considerar todas as combinações possíveis de regressores.

Portanto, a abordagem proposta consiste em selecionar um número limitado de variáveis, por definição $q \leq n$, e realizar regressões usando todas as combinações q das n variáveis. A previsão final é obtida calculando a média de todas as previsões geradas.

5.3.9
Random Forest (RF)

O modelo foi inicialmente proposto por Breiman (2001). É um modelo não-linear que se baseia no *bootstrap aggregation (bagging)* de árvores construídas de forma randômica. Essas árvores são chamadas de *regression trees*. Utilizando a metodologia de Medeiros et al. (2021), utilizaremos B amostras de *bootstrap*, de modo que para cada amostra b, $b = 1, ..., B.$, uma árvore com K_b regiões - definido para obter um número mínimo de observações em cada região - é estimada com um *subset* selecionado randomicamente dos regressores. A

previsão final é dada pela a equação abaixo.

$$\frac{1}{B}\sum_{b=1}^{B}[\sum_{k=1}^{K_b}\hat{c}_{k,b}I_{k,b}(x_t;\hat{\theta}_{k,b})] \tag{5-16}$$

5.4 Modelagem

Trabalharemos sobre preceito de estimação direta. Isso é, não faremos qualquer tipo de tentativa de previsão das covariadas. Além disso, iremos testar tanto *rolling-window* quanto *expanding-window* no *backtest*. Uma parte interessante do *framework* a ser usado é a robustez considerada no tratamento das variáveis. Ao serem adicionadas *raw* na base, o próprio *framework* decide qual melhor tratamento a ser feito, exceto quando o próprio hiperparâmetro é especificado no *backtest*. No contexto de *Big Data*, esse passo é fundamental para tratar a base da melhor forma possível.

Por conta da PNAD contínua ter começado apenas em março de 2012, nosso *in-sample* começará nessa data de início da série. A estrutura de estimação foi construída através de alguns hiper-parâmetros, falarei sobre eles na lista a seguir.

1. with.Trends: Se o modelo incorpora Trends ou não.

2. NBenchSelected: Número de vezes que algum dos benchmarks apareceu nos 12 horizontes.

3. to.pretest: Pré-Teste ou não. Objetivo é simples, testar o quão importantes são as covariadas para a previsão de y. Teste utilizado foi o teste-t. As variáveis são, então, ordenadas conforme o valor absoluto do teste-t.

4. fixed.n.to.keep: Número de variáveis a serem mantidas pelo Pré-Teste. Por conta da demora computacional na realização dos *backtests*, os valores variaram de 10 a 15.

5. to.test.cor: Teste de correlação entre as variáveis ou não. Com intuito de evitar a questão da multicolinearidade, caso o teste esteja especificado

como hiperparâmetro, as variáveis com mais de 95% de correlação são retiradas.

6. to.control.SA: Controle de sazonalidade ou não.

7. to.mm3: Média móvel de 3 meses nas variáveis que não são da PNAD.

Tabela 5.1: Categorias de termos e palavras-chave

Categorias de Termos e Palavras-chave	
Ações governamentais	– Seguro desemprego – seguro desemprego – fgts – fgts caixa – jovem aprendiz – Jovem aprendiz – mais emprego
Procura de vagas	– vagas – vagas emprego – vagas de emprego – emprego – trabalho
Remuneração	– salario – salário – salário mínimo – remuneração – salário motorista – salário uber – salario uber
Sites de correspondência entre empregados e empregadores	– infojobs – infojobs vagas – catho – catho vagas – indeed – indeed vagas – sine – sine vagas

Tabela 5.2: Base de Dados

PIM (Pesquisa Industrial Mensal)	
pim_total	Pesquisa Industrial Mensal Total
pim_ext	Pesquisa Industrial Mensal Extrativa
pim_transf	Pesquisa Industrial Mensal por Transformação
pim_bens_K	Pesquisa Industrial Mensal de Bens de Capital
pim_bens_interm	Pesquisa Industrial Mensal de Bens Intermediários
pim_bens_cons	Pesquisa Industrial Mensal de Bens de Consumo
pim_cons_dur	Pesquisa Industrial Mensal de Consumo Durável
pim_cons_semi_ndur	Pesquisa Industrial Mensal de Consumo Semi Durável Não Durável
pim_cons_semi	Pesquisa Industrial Mensal de Consumo Semi Durável
pim_cons_ndur	Pesquisa Industrial Mensal de Consumo Não Durável
PMS (Pesquisa Mensal de Serviços)	
pms	Pesquisa Mensal de Serviços
pms_serv_fam	Pesquisa Mensal de Serviços de Serviços Familiares
pms_tic	Pesquisa Mensal de Serviços de Tecnologia da Informação e Comunicação
pms_serv_adm	Pesquisa Mensal de Serviços Administrativos
pms_serv_trans	Pesquisa Mensal de Serviços de Transporte
pms_serv_outros	Outros Serviços da Pesquisa Mensal de Serviços
PMC (Pesquisa Mensal de Comércio)	
pmc_restr	Pesquisa Mensal de Comércio Restrita
pmc_ampl	Pesquisa Mensal de Comércio Ampliada
pmc_mercado	Pesquisa Mensal de Comércio do Mercado
pmc_eletrodom	Pesquisa Mensal de Comércio de Eletrodomésticos
pmc_veic	Pesquisa Mensal de Comércio de Veículos
pmc_constr	Pesquisa Mensal de Comércio de Materiais de Construção
PNAD (Pesquisa Nacional por Amostra de Domicílios)	
pnad_forc_trab	Taxa de Força de Trabalho
pnad_po	População Ocupada
pnad_pd	População Desocupada
pnad_tx_des	Taxa de Desocupação
pnad_po_ind	População Ocupada na Indústria
pnad_po_constr	População Ocupada na Construção
pnad_po_comerc	População Ocupada no Comércio
pnad_po_massa	População Ocupada nos Serviços de Manutenção e Reparação
pnad_po_serv	População Ocupada nos Serviços
pnad_formal	Empregados Formais
pnad_informal	Empregados Informais
Anfavea (Associação Nacional dos Fabricantes de Veículos Automotores)	
car_prod_total	Produção Total de Carros
car_lev_prod	Produção de Carros Leves
truck_prod	Produção de Caminhões
bus_prod	Produção de Ônibus
Fenabrave (Federação Nacional da Distribuição de Veículos Automotores)	
Fenbr_automoveis	Vendas de Automóveis
Fenbr_comerciais_leves	Vendas de Veículos Comerciais Leves
Fenbr_caminhoes	Vendas de Caminhões
Fenbr_onibus	Vendas de Ônibus
Fenbr_total	Total de Vendas

6
Resultados

Nesse capítulo comentarei um pouco sobre os resultados. Primeiramente, há de se ressaltar que dos 10 melhores modelos, 8 são com uso do Google Trends. Isso confirma nossa tese inicial de que o uso do Google Trends ajuda na geração de previsões mais robustas e consistentes.

Segundamente, os modelos superam os *backtests* que escolhemos e, que, também são padrões na academia - RW e AR. Dentre os 10 melhores *backtests*, cada um com 12 horizontes, os benchmarks são selecionados somente em 21 ocasiões, ou seja, 17.5% das vezes. Além disso, apesar da série ser relativamente curta (começa em março de 2012), fato que impediu ter uma grande janela de *in-sample*, os modelos apresentaram bons RMSEs.

Por último, percebemos que nossa opção por transformar as variáveis, que não são da PNAD, aplicando Médias Móveis de 3 meses em prol de aproximar do modelo de divulgação da Taxa de Desocupação esteve presente em todos os modelos do Top 10.

Na tabela 6.1 podemos ver o resultado dos 10 melhores modelos selecionados por RMSE de H1 e também se usam Trends ou não (parâmetro with.trends). Já na tabela 6.2 podemos ver as especificações desses modelos, alguns já mencionados no sub-capítulo 5.4.

Analisando pré e pós pandemia na tabela 6.3 podemos ver que o modelo *Elastic-Net* é selecionado em ambas. Interessante notar que o RMSE pré-pandemia é consideravelmente melhor do que o pós-pandemia. A explicação recaí sobre a dificuldade de testar a série logo após um choque idiossincrático, como o da pandemia de Covid-19, já que a primeira janela do *out-of-sample* começa em janeiro de 2019.

Como exemplo, o melhor modelo de H1 foi selecionado para demonstrar o *fit* com a Taxa de Desocupação. Abaixo, a imagem 6.1 exemplifica tal *fit*.

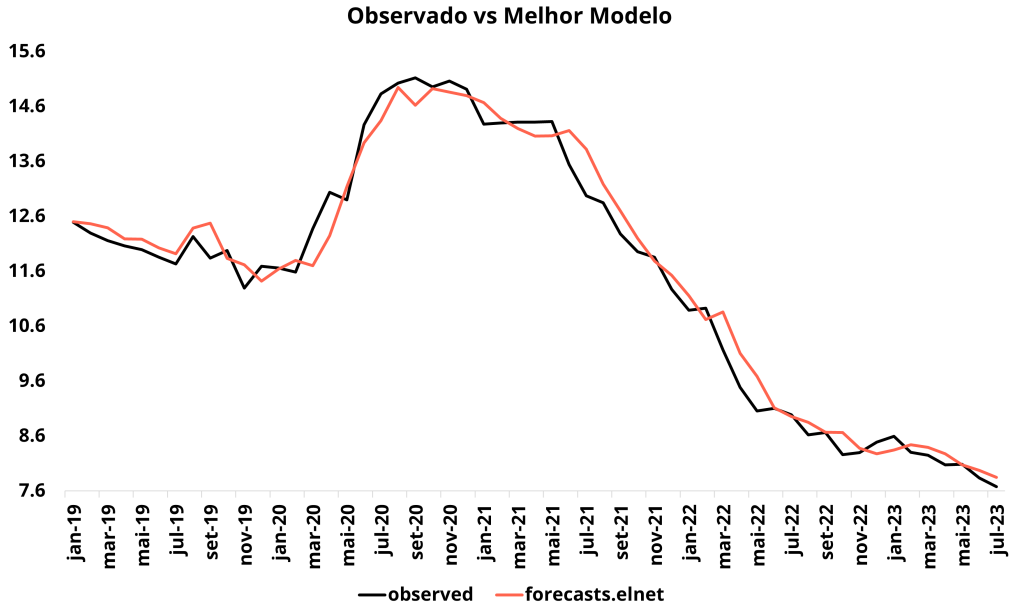

Figure 6.1: Melhor Modelo vs Observado

Tabela 6.1: Resultados dos melhores modelos

backtest	with_Trends	best.rmse.H1	best.model.H1	best.rmse.H2	best.model.H2	best.rmse.H3	best.model.H3
66	TRUE	0.340575387	elnet	0.477048184	lasso	0.671249668	lasso
65	TRUE	0.348353924	lasso	0.579299478	lasso	0.691444588	adanet
67	TRUE	0.349756259	lasso	0.570900681	lasso	0.725188908	lasso
83	TRUE	0.354259122	lasso	0.501523326	lasso	0.685637552	lasso
5	FALSE	0.357674671	lasso	0.514537451	lasso	0.746795014	lasso
69	TRUE	0.360018566	lasso	0.493100619	lasso	0.70522585	lasso
81	TRUE	0.366052529	lasso	0.562631564	adanet	0.687204554	lasso
75	TRUE	0.369201494	adanet	0.600192546	adanet	0.75418061	elnet
91	TRUE	0.375811887	adanet	0.561191906	adanet	0.776902174	ada
3	FALSE	0.375906772	lasso	0.577726259	lasso	0.709568933	lasso

Tabela 6.2: Especificações dos melhores modelos

backtest	with.Trends	NBenchSelected	expand.window	to.pretest	fixed.n.to.keep	to.test.cor	to.control.SA	to.mm3
66	TRUE	0	TRUE	FALSE	15	TRUE	TRUE	TRUE
65	TRUE	2	TRUE	TRUE	15	TRUE	TRUE	TRUE
67	TRUE	2	TRUE	TRUE	15	FALSE	TRUE	TRUE
83	TRUE	1	TRUE	TRUE	15	FALSE	FALSE	TRUE
5	FALSE	0	TRUE	TRUE	10	TRUE	TRUE	TRUE
69	TRUE	2	TRUE	TRUE	10	TRUE	TRUE	TRUE
81	TRUE	1	TRUE	TRUE	15	TRUE	FALSE	TRUE
75	TRUE	5	FALSE	TRUE	15	FALSE	TRUE	TRUE
91	TRUE	7	FALSE	TRUE	15	FALSE	FALSE	TRUE
3	FALSE	1	TRUE	TRUE	15	FALSE	TRUE	TRUE

Tabela 6.3: H1 do melhor backtest - Pré vs Pós Pandemia

Modelo	RMSE Pré Pandemia	Ranking Pré Pandemia	RMSE Pós Pandemia	Ranking Pós Pandemia	RMSE	Ranking Final
lasso	0.289	5	0.391	2	0.368	2
ridge	2.015	10	2.484	10	2.374	10
elnet	0.259	1	0.364	1	0.341	1
ada	0.292	6	0.415	4	0.387	4
adanet	0.286	3	0.408	3	0.380	3
RW	0.288	4	0.426	5	0.395	5
ar	0.379	8	0.443	6	0.428	7
RF	0.377	7	0.956	9	0.847	9
csr	0.557	9	0.690	8	0.659	8
median	0.274	2	0.455	7	0.416	6

7
Conclusão

Neste capítulo, apresentaremos as conclusões e os resultados obtidos a partir da condução de 128 *backtests*. O objetivo do início do trabalho era avaliar o impacto do uso do Google Trends na previsão da Taxa de Desocupação da PNAD Contínua. Durante este processo, sempre fixando janeiro de 2019 como início do *Out-of-sample* observamos que a maioria dos 10 melhores *backtests* (8/10), selecionados por RMSE do H1, incluiam termos do Google Trends na base de dados.

Além disso, o uso das técnicas de tratamento de dados, que foram adaptadas dos estudos de Medeiros e Pires (2021), melhorou a consistência dos modelos em todos os horizontes de previsão. Em particular, a inclusão da média móvel de 3 meses nas covariadas, uma estratégia discutida no trabalho de Ludwig e Souza (2016), mostrou-se eficaz, estando presente em todos os modelos do Top 10.

Ao concluir este trabalho, podemos afirmar que atingimos com sucesso nosso objetivo de superar os *benchmarks* selecionados em vários *backtests*. Isso sugere que os modelos de ML utilizados apresentaram robustez e consistência na previsão da taxa de desocupação. Além disso, infere-se igualmente que a utilização do Google Trends incrementou a eficácia desses modelos.

8
Referências bibliográficas

BÁNBURA, M. et al. Now-casting and the real- time data flow. **ECB Working Paper**, European Central Bank, v. 1, n. 1565, p. 2, 2013. Citado na página 10.

BORUP, D.; RAPACH, D.; SCHüTTE, M. C. E. Mixed-frequency machine learning: Nowcasting and backcasting weekly initial claims with daily internet search volume data. **International Journal of Forecasting**, v. 39, n. 3, 2023. Citado na página 14.

BREIMAN, L. Random forests. **Machine Learning**, p. 5–32, 2001. Citado na página 19.

CHOI, H.; VARIAN, H. Predicting the present with google trends. **Economic Record**, v. 88, 2009. Citado 2 vezes nas páginas 10 e 13.

DIMPFL, T.; BLEHER, J. Knitting multi-annual high-frequency google trends to predict inflation and consumption. **Econometrics and Statistics**, v. 24, p. 1–26, 2022. Citado na página 13.

D'AMURI, F.; MARCUCCI, J. 'google it!' forecasting the us unemployment rate with a google job search index. **International Journal of Forecasting**, n. 31, p. 1031–1492, 2023. Citado na página 13.

ELLIOTT, G.; GARGANO, A.; TIMMERMANN, A. Complete subset regressions. **Journal of Econometrics**, v. 177, n. 2, p. 357–373, 2013. ISSN 0304-4076. Dynamic Econometric Modeling and Forecasting. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304407613000948>. Citado na página 19.

FONDEUR, Y.; KARAMé, F. Can google data help predict french youth unemployment? **Economic Modelling**, v. 30, p. 117–125, 2013. Citado na página 13.

HOERL, E. A.; KENNARD, W. R. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, v. 12, n. 1, p. 55–67, 1970. Citado 2 vezes nas páginas 11 e 17.

KOHNS, D.; BHATTACHARJEE, A. Nowcasting growth using google trends data: A bayesian structural time series model. **International Journal of Forecasting,**, p. 1031–1492, 2023. Citado na página 14.

LUDWIG, R.; SOUZA, P. Nowcasting do desemprego com google trends: evidências do mercado de trabalho brasileiro. **Trabalho de Conclusão de Curso. Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, Rio de Janeiro**, 2016. Citado 3 vezes nas páginas 12, 16 e 29.

MEDEIROS, M. C.; PIRES, H. F. The proper use of google trends in forecasting models. **Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, Rio de Janeiro.**, n. 683, 2021. Citado 4 vezes nas páginas 9, 14, 16 e 29.

MEDEIROS, M. C. et al. Forecasting inflation in a data-rich environment: The benefits of machine learning methods. **Journal of Business & Economic Statistics**, Taylor Francis, v. 39, n. 1, p. 98–119, 2021. Citado 2 vezes nas páginas 15 e 19.

NAGAO, S.; TAKEDA, F.; TANAKA, R. Nowcasting of the u.s. unemployment rate using google trends. **Finance Research Letters**, v. 30, 2019. Citado na página 13.

SUHOY, T. Query indices and a 2008 downturn: Israeli data. **Bank of Israel Working Papers, Bank of Israel**, n. 2009.06, 2009. Citado na página 13.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society**, v. 58, n. 1, p. 267–288, 1996. Citado 2 vezes nas páginas 11 e 17.

WOLOSZKO, N. Tracking activity in real time with google trends. **ECD Economics Department Working Papers**, n. 1634, 2020. Citado na página 14.

ZIMMERMANN, K. F.; ASKITAS, N. Google econometrics and unemployment forecasting. **German Council for Social and Economic Data (RatSWD) Research Notes**, n. 41, 2009. Citado na página 13.

ZOU, H.; ZHANG, H. H. On the adaptive elastic-net with a diverging number of parameters. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 37, n. 4, p. 1733–1751, 2009. Disponível em: <https://doi.org/10.1214/08-AOS625>. Citado na página 18.