PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

NOWCASTING DO DESEMPREGO COM *GOOGLE TRENDS*: EVIDÊNCIAS DO MERCADO DE TRABALHO BRASILEIRO

Nome do aluno: Raphael de Aquino Ludwig Pereira
Número da matrícula: 1311078
Orientador: Pedro Carvalho Loureiro de Souza

Outubro de 2016

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

NOWCASTING DO DESEMPREGO COM *GOOGLE TRENDS*: EVIDÊNCIAS DO MERCADO DE TRABALHO BRASILEIRO

Nome do aluno: Raphael de Aquino Ludwig Pereira
Número da matrícula: 1311078
Orientador: Pedro Carvalho Loureiro de Souza

"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".

_____

Raphael de Aquino Ludwig Pereira
Outubro de 2016

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

AGRADECIMENTOS

Ao meu pai, Ruy ‘Lula’ Ludwig, por ser minha eterna inspiração e motor de minha pequenina e fugaz vivência.

A minha mãe, Monica Ludwig, e irmã, Danielle Ludwig, por sustentarem a vida como um eterno pilar de apoio.

Ao meu avô, Sérgio de Aquino, pela sua paciência e confiança que jamais conseguirei retribuir a altura.

A toda minha família pelo apoio incondicional e por possibilitarem um espaço para a construção de quem eu sou.

Aos meus amigos, as melhores pessoas nesse mundo, com quem eu reaprendo quem eu sou todo dia.

Ao Departamento de Economia da PUC-Rio por me mostrar a beleza e a arte que a fria Economia pode possuir.

Ao meu orientador, Pedro Carvalho, pela ajuda e interesse, sempre solícito em me guiar pelos os inúmeros rumos e dobras de uma aprendizagem.

“Eu quase que nada não sei. Mas desconfio de muita coisa.”

João Guimarães Rosa

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

0. INTRODUÇÃO

A faculdade de realizar previsões é um traço chave do homem como ser humano. Seja no dia-a-dia ou em áreas especializadas, prever com razoável precisão acontecimentos futuros molda a vida em sociedade e afeta diretamente o bem-estar das pessoas. A capacidade de se antecipar um desastre natural iminente ou os desdobramentos que novos instrumentos financeiros podem ter nos mercados de ativos são exemplos claros.

A eficácia da previsão pode ser simplificada numa função que depende basicamente da quantidade e da qualidade de informação disponível. Mesmo usufruindo dos métodos estatísticos mais avançados, sem uma base de dados bem construída a previsão não será precisa.

Essa interação entre previsão e informação disponível ganha ainda mais importância no setor público. Os impactos de políticas públicas são capazes de afetar toda uma população, modificando complexas dinâmicas de interações sociais.

Um caso emblemático são as políticas ditas sociais que, muitas vezes, giram em torno de fornecer uma igualdade de oportunidades para todos, possibilitando uma melhor inserção da população no mercado de trabalho. O desenho dessas políticas depende diretamente de dados que, através de uma eficaz descrição da realidade, norteiam a tomada de decisão. A taxa de desemprego faz parte desse grupo de dados essenciais para melhor compreensão da conjuntura econômica nacional. No Brasil, sua estimação é feita e divulgada pelo Instituto Brasileiro de Geografia e Estatística ("IBGE") através da Pesquisa Nacional por Amostra de Domicílios Contínua ("PNAD Contínua"), sendo fruto de um processo custoso de coleta de dados.

Assim, as divulgações da PNAD Contínua sempre saem com um já conhecido *delay*. O *policymaker*, por sua vez, se vê obrigado a tomar decisões sem conhecer a situação corrente oficial do desemprego.

Varian e Choi (2009) propuseram a utilização de índices do *Google Trends* como uma maneira de conseguir contornar esse *delay* de divulgação para diversas variáveis, como índices de venda publicados pelo *U.S. Census Bureau*, ou pedidos de seguro desemprego publicados pelo *US Department of Labor.*

Tabela 0.1.: Calendário das divulgações da PNAD Contínua mensal e trimestral

Pesquisa	Mês de referência	Divulgação
PNAD contínua mensal	mai/16	29/jun/16
	jun/16	29/jul/16
	jul/16	30/ago/16
	ago/16	30/set/16
	set/16	27/out/16
	out/16	29/nov/16
	nov/16	29/dez/16
PNAD contínua trimestral	4° Trimestre 2015	15/mar/16
	1° Trimestre 2016	19/mai/16
	2° Trimestre 2016	17/ago/16
	3° Trimestre 2016	22/nov/16
	4° Trimestre 2016	23/fev/16

A ideia central é de que, sendo esses índices do *Google Trends* mensurados em tempo real, se tornaria possível melhorar a previsão do presente (*nowcasting*) desses dados com divulgações defasadas. Intuitivamente isso seria possível, pois essas variáveis possuem informações relevantes sobre o presente. Assim, ao adicioná-las de maneira contemporânea na especificação de regressões, estar-se-ia acrescentando informações sobre a trajetória temporal da variável defasada que, por definição, seriam ainda desconhecidas.

Nesse trabalho examino se os índices do *Google Trends* são capazes de melhorar o *nowcasting* da taxa de desocupação nacional. Primeiramente, modelo a sazonalidade e o ciclo da série da taxa de desocupação obtida na PNAD Contínua e escolho onze modelagens possíveis do ciclo. Essa escolha é feita com base em critérios *in-sample*. Em seguida, seleciono termos de busca relacionados ao mercado de trabalho e escolho focar em quatro especificamente: “vagas”, “vagas emprego”, “vagas de emprego” e “emprego”. Modelo esses quatro para expurgar sua sazonalidade.

Uma vez modeladas as séries, seleciono por critérios *out-of-sample* o melhor método de previsão entre *Rolling Window* (o escolhido) e *Expanding Window*.

Descubro que, para cada modelagem testada da série da taxa de desocupação sem índices do *Google Trends*, existem pelo menos cinco modelos que utilizam os índices e se saem melhores nos critérios *out-of-sample*. Entre as especificações de regressões envolvendo índices testadas, claramente a regressão que leva em conta apenas o índice contemporâneo à data da previsão possui uma hegemonia.

Em seguida, testo o poder preditivo dos índices por si só, isto é, sem adicionar nenhum componente ARIMA(p,d,q) na regressão. Como foi selecionado um número elevado de termos, utilizo métodos de *shrinkage regression*, como LASSO, adaLASSO e *Elastic Net* para previsão e seleção de variáveis.

Encontro resultados variados dependendo da finalidade da *shrinkage regression*. Para determinado desenho do critério de informação BIC, encontro resultados preditivos satisfatórios, porém selecionando uma quantidade grande de variáveis. Para outro formato da equação do BIC adaptado para seleção de variáveis, a previsão é a pior encontrada em todo esse trabalho, porém o *shrinkage* é satisfatório, selecionando em média quatro índices do *Google Trends* entre os 28 disponíveis.

Por fim, adiciono essas variáveis selecionadas via *shrinkage* à melhor modelagem do ciclo encontrada anteriormente e comparo seu desempenho na previsão *out-of-sample* com o melhor modelo achado até então no exercício. O desempenho deles é muito parecido e ambos realizam previsões melhores que os modelos que utilizam apenas componentes *ARIMA(p,d,q)* da taxa de desocupação.

O restante desse trabalho está organizado da seguinte maneira: Seção 1 apresenta os dados utilizados, explicitando suas fontes e aspectos mais relevantes. Nesse contexto, são expostos os termos de pesquisa escolhidos. Seção 2 faz uma revisão da recente literatura sobre *nowcasting* com índices do *Google Trends*. Seção 3 expõe toda metodologia de modelagem e escolha de modelos utilizada ao longo do trabalho. Nela também é feita a seleção das modelagens do ciclo estudadas. Seção 4 testa por critérios *out-of-sample* o poder dos índices de melhorar previsões. Também são feitas as regressões de *shrinkage* para seleção de variáveis e previsão com as selecionadas. Seção 5 resume a discussão do trabalho e os resultados encontrados. Seção 6 é a bibliografia.

1. BASE DE DADOS

Utilizei duas fontes de dados: a plataforma *online Google Trends* ("*Trends*") e a PNAD Contínua. O *Trends* é uma plataforma, na qual diariamente são computados em tempo real para níveis de desagregação mundial, nacional e estadual índices de relevância da pesquisa de determinados termos no site *Google.com.* O índice é apresentado de duas maneiras: (i) em relação a si próprio no tempo para diversas frequências e (ii) em relação a diferentes conjuntos nacionais e subnacionais para o corte temporal mais atual. Essa ferramenta existe desde 2009, porém disponibiliza dados que vão de janeiro de 2004 até o exato momento (Apêndice A).

Dados da relevância da busca *online* por diferentes termos relacionados ao mercado de trabalho tiveram, assim, o *Trends* como fonte. Nesse exercício usei os índices de relevância relativos a si mesmo no tempo. Isto é, para cada termo escolhido, existe uma série temporal em frequência mensal de índices do *Trends*. Esses índices são normalizados, de maneira que 100 é o valor do mês no qual o total de buscas pelo termo desejado sobre o total de buscas realizadas no período registra o maior montante. Os outros valores da série temporal são ajustados em relação a esse valor máximo, montando uma série histórica do índice na qual o valor máximo é 100. Mesmo esses dados sendo disponibilizados em níveis de desagregação estadual e nacional, nesse exercício foquei apenas no âmbito nacional.

Tabela 1.1.: Divisão dos termos utilizados no exercício por conjuntos semânticos

Termos	Conjunto semântico
vagas, vagas emprego, vagas de emprego, emprego, trabalho, salario, salário, desemprego, oportunidade de emprego, oportunidades de emprego, curriculo, currículo, remuneração	Relacionados diretamente com o mercado de trabalho
mais emprego, fgts, fgts caixa,	Relacionados com ações do governo
ensino medio completo, ensino fundamental completo, fundamental completo, ensino médio completo	Relacionados ao nível de escolaridade
infojobs, infojobs vagas, catho, catho vagas, indeed, indeed vagas, sine, sine vagas	Relacionados aos *sites* de *matching* entre empregado e empregador

Foram usadas as séries de todos os termos explicitados na Tabela 1.2.1. na frequência mensal de janeiro de 2012 até outubro de 2016 (valor parcial, pois o mês ainda não havia terminado).

A outra fonte utilizada foi a PNAD Contínua realizada pelo IBGE. Tendo começado em 2012 após o anúncio do fim da Pesquisa Mensal do Emprego (terminada de fato em março 2016), a PNAD Contínua é realizada por meio de uma amostra de domicílios, procurando garantir a representatividade dos níveis geográficos de divulgação. Dela retirei a série histórica da taxa de desocupação.

A taxa de desocupação mede o percentual de pessoas desocupadas - pessoas não ocupadas, que tomaram alguma providência efetiva para conseguir um trabalho no período de referência de 30 dias e que estavam disponíveis para iniciar um trabalho na semana de referência - em relação às pessoas na força de trabalho - a soma das pessoas ocupadas e desocupadas no período. "São classificadas como ocupadas na semana de referência as pessoas que, nesse período, trabalharam pelo menos uma hora completa em trabalho remunerado em dinheiro, produtos, mercadorias ou benefícios (moradia, alimentação, roupas, treinamento etc.) ou em trabalho sem remuneração direta em ajuda à atividade econômica de membro do domicílio ou, ainda, as pessoas que tinham trabalho remunerado do qual estavam temporariamente afastadas nessa semana".[1]

Para níveis estaduais, a taxa de desocupação está disponível apenas como média trimestral. Cada ano, assim, é composto por quatro taxas, uma para cada trimestre, por estado da federação. Sua divulgação também é trimestral. Para o nível federal, tem-se uma série de divulgação mensal de médias móveis trimestrais, contando, desse modo, com uma maior frequência de doze taxas nacionais divulgadas por ano.

A série destina-se a produzir informações contínuas sobre a inserção da população no mercado de trabalho. Para esse exercício, usei a variável taxa de desocupação das pessoas de 14 anos ou mais de idade, na semana de referência ("taxa de desocupação"), para o nível federal como a taxa de desemprego federal em termos mais gerais. Como ela é construída através de uma média móvel trimestral, retirei a média móvel trimestral dos índices do *Google Trends* que, por sua vez, foram obtidos na forma mensal.

[1] INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Pesquisa Nacional por Amostra de Domicílios Contínua. Série Notas Metodológicas, vol. 1, Coordenação de Trabalho e Rendimento, Diretoria de Pesquisas. Rio de Janeiro, 2014.

Devido aos fatos de que a PNAD Contínua é muito recente e de que a série estadual da taxa de desocupação possui apenas quatro observações anuais, optei por realizar todo exercício apenas para escala nacional.

2. REVISÃO BIBLIOGRÁFICA

O surgimento da literatura de *nowcasting* com índices do *Trends* é muito recente e está obviamente ligado ao surgimento da plataforma do *Google Trends*. O primeiro artigo a jogar luz sobre tal temática foi Varian e Choi (2009). Nele os autores basicamente focaram em explicar a estrutura dos dados do *Google Insights* (primeiro nome do *Trends*) e testar com critérios *out-of-sample* sua capacidade de melhorar previsões de pedidos por seguro desemprego e venda automotivas, entre outros exemplos. Artigos anteriores a esse, como Ettredge et al. (2005), já sugeriam que dados de pesquisa *online* poderiam ser úteis para previsão.

Estudos posteriores foram Asiktas e Zimmermann (2009), D'Amuri e Marcucci (2010) e Suhoy (2009) que examinaram as taxas de desemprego na Alemanha, nos Estados Unidos e em Israel respectivamente. Todos esses estudos se utilizaram de métodos de modelagem ARIMA e critérios de avaliação *out-of-sample* muito parecidos com aqueles inicialmente propostos em Varian e Choi (2009).

A evidência empírica encontrada em todos eles condiz com aquela achada nesse trabalho. Os índices do *Trends* foram sempre associados a uma melhora nos critérios *out-of-sample* de previsão, superando outros modelos sem índices em termos de precisão na previsão e de capacidade preditiva.

Carrière-Swallow e Labbé (2013) criou um índice de venda automotiva com a ferramenta do *Trends* e encontraram evidências de que sua utilização foi capaz de melhorar e eficiência tanto *in-sample* quanto *out-of-sample* do *nowcasting* das vendas do setor. Schmidt e Vosen (2009) comparou um indicador para consumo privado baseado em índices do *Trends* com indicadores baseados em pesquisas e ecnontrou uma eficácia preditiva maior tanto para critérios *out-of-sample* quanto *in-sample* ao utilizar o indicador criado a partir dos índices.

Goel et. al (2010), por sua vez, descreve algumas limitações dos dados de busca *online*, explicitando que tais dados podem não providenciar uma melhorar tão drástica assim na *predictability*, mesmo melhorando de fato as previsões. Contudo, acaba por exaltar suas qualidades de fácil acesso e alta frequência temporal (tempo real).

Shimshoni et al. (2009), finalmente, comparou a eficácia de se prever os próprios índices do *Trends*, chegando à conclusão de que diversas categorias, principalmente as categorias que agregam diversos termos, possuem uma alta *predictability*, muitas vezes por possuírem um padrão sazonal claro.

3. METODOLOGIA

3.1. MODELAGEM DO CICLO

O primeiro passo da análise será a modelagem da série das taxas de desocupação mensais de janeiro de 2012 até agosto de 2016. Como a série é construída como uma sucessão de médias móveis trimestrais, o primeiro valor da série se refere à média da taxa de desocupação entre janeiro, fevereiro e março de 2012, enquanto o último se refere à média entre junho, julho e agosto de 2016. Essa primeira modelagem será baseada na metodologia Box-Jenkins de (BOX; JENKINS, 1970). A escolha dos melhores modelos será feita através dos critérios de informação Akaike ("AIC"), sua correção para amostras pequenas em relação ao número de regressores ("AICc") e Schwarz ("BIC").

$$AIC = -2ln(L_p) + 2K \tag{1}$$

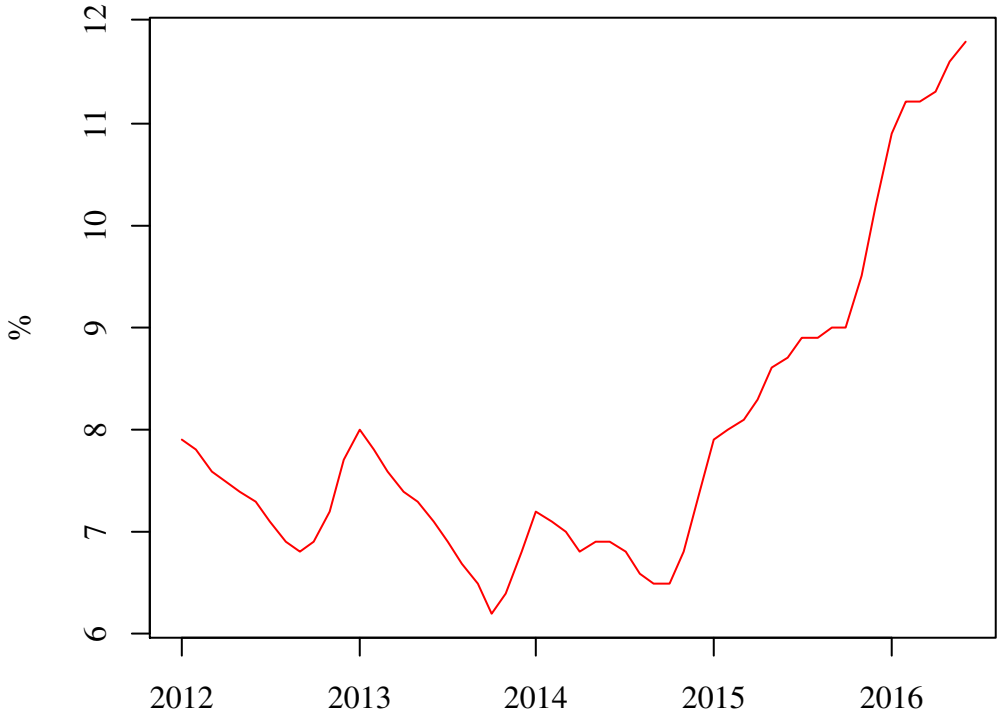
$$AICc = AIC + \frac{2K(K+1)}{n-K-1} \tag{2}$$

$$BIC = -2ln\left(L_p\right) + ln(n) \cdot K \tag{3}$$

Onde $ln(L_p)$ é o logaritmo natural da função de verossimilhança do modelo estimado; K é o número de parâmetros do modelo e n é o número de observações. O melhor modelo é aquele que apresenta o melhor balanceamento entre *fit* e complexidade e que minimiza os critérios de informação.

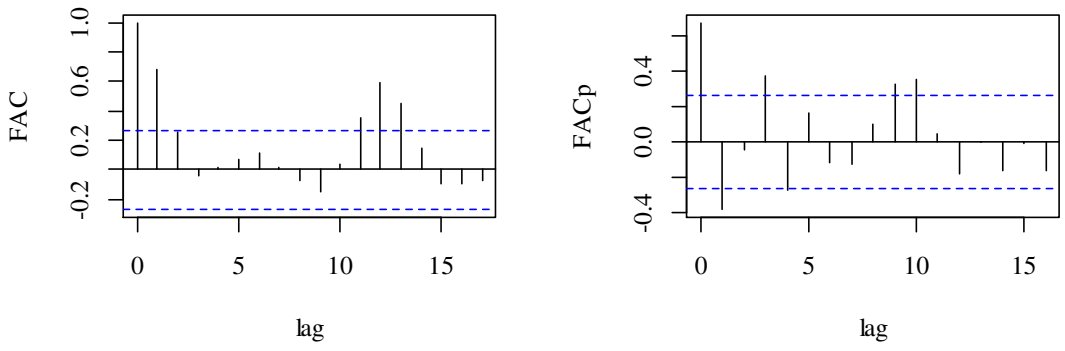
Como primeira etapa, analisei a evolução da taxa de desocupação ao longo do tempo. A série temporal consiste de 54 observações no espaço de janeiro de 2012 até agosto de 2016.

Figura 3.1.1.: Evolução da taxa de desocupação nacional



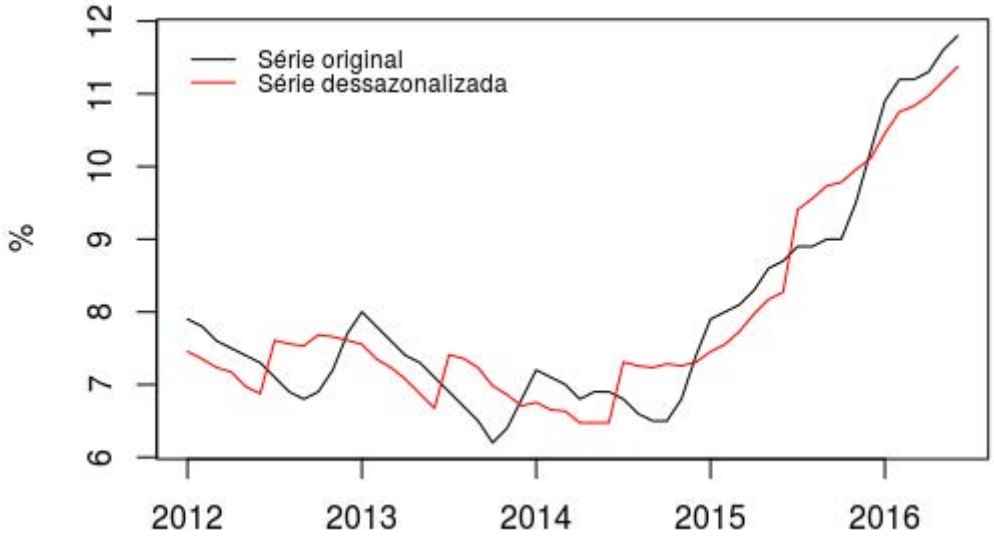
A série não aparenta apresentar nenhum tipo de sazonalidade. Ela parece, contudo, indicar uma quebra estrutural no começo de 2015. Mais precisamente, sua dinâmica começa a mudar a partir da média trimestral entre dezembro de 2014, janeiro e fevereiro 2015, configurando uma tendência de subida praticamente ininterrupta. Essa tendência se acelera ainda mais em 2016. Desse modo, a série poderia ser modelada apenas pós-quebra estrutural, buscando alguma tendência linear ou quadrática. Contudo, optei por trabalhar com ela inteiramente, devido à pequena quantidade de observações. Assim sendo, trata-se de um processo claramente não estacionário e, então, tirei a primeira diferença para analisar os correlogramas das funções de autocorrelação e autocorrelação parcial.

Figura 3.1.2.: Correlogramas da FAC e FACp da primeira diferença da série da taxa de desocupação



Os correlogramas não dão nenhuma indicação muito clara sobre o processo gerador dos dados da série, exceto a de que tanto a parte autoregressiva quanto a parte de média móvel são importantes na modelagem do processo ARIMA(p,d,q) que melhor descreve a evolução da taxa de desocupação. O argumento de presença de sazonalidade ganha força com o correlograma da FAC, pois as defasagens de número 12 e 13 se mostram relevantes.

Figura 3.1.3.: Evolução da taxa de desocupação original e dessazonalizada



Retirei a sazonalidade regredindo a série original em variáveis *dummy* relativas a cada trimestre móvel. A série dessazonalizada apresenta um desenho muito parecido com a original, porém sempre reagindo de maneira antecipada, i.e., quando a série dessazonalizada tem uma queda ou subida, o mesmo ocorre na série original, porém no período seguinte. Devido à semelhança da dinâmica, optei por trabalhar com a série original ao longo do exercício.

O próximo passo foi testar de fato se a série é estacionária, logo se não há presença de raiz unitária. As Tabelas 3.1.1. e 3.1.2. apresentam as estatísticas t e os valores críticos para os níveis 0.01, 0.5 e 0.1 de significância dos testes Dickey-Fuller (4), Dickey-Fuller Aumentado (5) e Dickey-Fuller Aumentado com *drift*, i.e., adição do intercepto (6).

$$\Delta Y_t = \theta Y_{t-1} + \varepsilon_t \tag{4}$$

$$\Delta Y_t = \theta Y_{t-1} + \sum_{p=1}^{n} \beta_p \Delta Y_{t-p} + \varepsilon_t \tag{5}$$

$$\Delta Y_t = \theta Y_{t-1} + \alpha_t + \sum_{p=1}^{n} \beta_p \Delta Y_{t-p} + \varepsilon_t \tag{6}$$

Onde,

$$H_0 : \theta = 0$$

$$H_1 : \theta < 0$$

Y_t é a série temporal que desejo testar, no caso, a série da taxa de desocupação; α_t é o intercepto da regressão e ε_t o erro da regressão. O p – o número de defasagens usadas nos modelos (5) e (6) – é escolhido através da minimização do critério BIC. Em todos os casos, o número de *lags* escolhido foi apenas um.

Tabela 3.1.1.: Resultados dos testes DF e ADF na variável em nível

	Estatística t	1pct	5pct	10pct
DF	2.44	-2.60	-1.95	-1.61
ADF	0.98	-2.60	-1.95	-1.61
ADF com drift	-0.12	-3.51	-2.89	-2.58

Tabela 3.1.2.: Resultados dos testes DF e ADF na primeira diferença da variável

	Estatística t	1pct	5pct	10pct
DF	-2.92	-2.60	-1.95	-1.61
ADF	-3.93	-2.60	-1.95	-1.61
ADF com drift	-4.21	-3.51	-2.89	-2.58

As Tabelas 3.1.1. e 3.1.2. parecem indicar a existência de uma raiz unitária, pois, enquanto para a série em nível a estatística t é incapaz de rejeitar a hipótese nula de existência de raiz unitária a níveis clássicos de significância estatística, para a série em primeira diferença a hipótese nula é rejeitada ao nível de 0.01 em todos os casos. Entretanto, mesmo na presença de raiz unitária, escolhi na próxima etapa de modelagem do ciclo utilizar tanto modelos ARMA(p,q) quanto modelos ARIMA(p,1,q). Devido à presença de raiz unitária, os modelos foram estimados por máxima verossimilhança.

Foram estimados modelos do formato ARIMA:

$$\left(1 - \sum_{k=1}^{p} \gamma_k L^k\right)(1 - \mathrm{L})^d Y_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right)\varepsilon_t \tag{7}$$

Onde,

$$p, q = \{0,1,2,3\}; d = \{0,1\}$$

Para cada modelo estimado, foram calculados os critérios de Akaike - equações (1) e (2) - e de Schwartz - equação (3).

Tabela 3.1.3.: Resultados dos critérios de informação para modelos escolhidos, em nível

	ARMA(0,0)	ARMA(0,1)	ARMA(0,2)	ARMA(0,3)
AIC	198.34	137.68	80.33	46.90
BIC	198.57	138.16	81.14	48.15
AICc	202.32	143.65	88.28	56.85

	ARMA(1,0)	ARMA(1,1)	ARMA(1,2)	ARMA(1,3)
AIC	21.43	-4.13	-28.67	-26.89
BIC	21.91	-3.31	-27.42	-25.10
AICc	27.40	3.83	-18.72	-14.95

	ARMA(2,0)	ARMA(2,1)	ARMA(2,2)	ARMA(2,3)
AIC	-13.54	-19.39	-26.96	-26.81
BIC	-12.72	-18.14	-25.17	-24.37
AICc	-5.58	-9.44	-15.02	-12.88

	ARMA(3,0)	ARMA(3,1)	ARMA(3,2)	ARMA(3,3)
AIC	-17.90	-10.96	-26.16	-25.96
BIC	-16.65	-9.18	-23.73	-22.76
AICc	-7.96	0.97	-12.24	-10.05

Nas Tabelas 3.1.3. e 3.1.4. foram buscados os menores valores dos critérios de informação. Assim, formei um grupo de melhores modelos estimados. Esses melhores modelos foram usados de diferentes maneiras no *nowcasting* da taxa de desocupação. Exceções foram abertas para os casos do AR(1), ARIMA(1,1,0), MA(1) e ARIMA(0,1,1) que, mesmo não figurando entre os modelos que apresentaram os menores critérios Akaike/Schwarz, foram utilizados no restante do exercício.

Tabela 3.1.4.: Resultados dos critérios de informação para modelos escolhidos, primeira diferença

	ARIMA(0,1,0)	ARIMA(0,1,1)	ARIMA(0,1,2)	ARIMA(0,1,3)
AIC	13.43	-12.04	-36.21	-34.34
BIC	13.51	-11.80	-35.72	-33.51
AICc	15.40	-8.10	-30.30	-26.46

	ARIMA(1,1,0)	ARIMA(1,1,1)	ARIMA(1,1,2)	ARIMA(1,1,3)
AIC	-20.44	-22.84	-34.37	-32.99
BIC	-20.20	-22.35	-33.54	-31.71
AICc	-16.50	-16.93	-26.49	-23.14

	ARIMA(2,1,0)	ARIMA(2,1,1)	ARIMA(2,1,2)	ARIMA(2,1,3)
AIC	-25.50	-23.50	-33.40	-30.99
BIC	-25.01	-22.67	-32.13	-29.17
AICc	-19.58	-15.62	-23.55	-19.17

	ARIMA(3,1,0)	ARIMA(3,1,1)	ARIMA(3,1,2)	ARIMA(3,1,3)
AIC	-23.52	-28.76	-32.88	-31.74
BIC	-22.69	-27.49	-31.05	-29.25
AICc	-15.64	-18.91	-21.05	-17.95

Foram selecionados os modelos: AR(1), MA(1), ARMA(1,2), ARMA(2,2) e ARMA(1,3), lidando com a variável em nível; e ARIMA(1,1,0), ARIMA(0,1,1), ARIMA(0,1,2), ARIMA(0,1,3), ARIMA(1,1,2) e ARIMA(2,1,2), representando os modelos que levam em conta a variável em primeira diferença.

Para checar se os resíduos dos modelos escolhidos são ruídos brancos independente- e identicamente distribuídos, escolhi o teste de Ljung-Box (LJUNG; BOX, 1978). O teste foi realizado até a 12ª defasagem.

Tabela 3.1.5.: p-valores dos testes de Ljung-Box até 12ª defasagem

	1o lag	2o lag	3o lag	4o lag	5o lag	6o lag
AR(1)	0.000	0.000	0.000	0.000	0.000	0.000
MA(1)	0.000	0.000	0.000	0.000	0.000	0.000
ARMA(1,2)	0.621	0.704	0.798	0.830	0.912	0.640
ARMA(2,2)	0.854	0.907	0.880	0.811	0.902	0.605
ARMA(1,3)	0.798	0.860	0.870	0.816	0.906	0.607
ARIMA(1,1,0)	0.064	0.123	0.001	0.002	0.004	0.002
ARIMA(0,1,1)	0.029	0.006	0.003	0.004	0.008	0.005
ARIMA(0,1,2)	0.873	0.944	0.825	0.903	0.956	0.808
ARIMA(0,1,3)	0.971	0.991	0.812	0.866	0.926	0.754
ARIMA(1,1,2)	0.936	0.994	0.799	0.853	0.916	0.744
ARIMA(2,1,2)	0.846	0.934	0.599	0.750	0.777	0.830

	7o lag	8o lag	9o lag	10o lag	11o lag	12o lag
AR(1)	0.000	0.000	0.000	0.000	0.000	0.000
MA(1)	0.000	0.000	0.000	0.000	0.000	0.000
ARMA(1,2)	0.687	0.727	0.673	0.721	0.659	0.058
ARMA(2,2)	0.662	0.699	0.650	0.694	0.688	0.059
ARMA(1,3)	0.660	0.700	0.656	0.703	0.689	0.058
ARIMA(1,1,0)	0.005	0.009	0.001	0.001	0.001	0.000
ARIMA(0,1,1)	0.008	0.013	0.006	0.010	0.008	0.000
ARIMA(0,1,2)	0.775	0.820	0.724	0.760	0.683	0.054
ARIMA(0,1,3)	0.729	0.779	0.687	0.723	0.690	0.051
ARIMA(1,1,2)	0.723	0.772	0.677	0.710	0.685	0.051
ARIMA(2,1,2)	0.896	0.936	0.753	0.738	0.632	0.073

A Tabela 3.1.5. demonstra os p-valores do Teste de Ljung-Box em relação a doze defasagens. Para todas as modelagens selecionadas via critérios de informação Akaike/Schwarz – todas aquelas presentes na Tabela 3.1.5. exceto AR(1), MA(1), ARIMA(1,1,0) e ARIMA(0,1,1) – a hipótese nula do Teste Ljung-Box de ausência de um padrão de autocorrelação nos resíduos do modelo estimado não é rejeitada ao nível de 0.05 para as defasagens propostas, i.e., não se pode rejeitar que os resíduos desses modelos sejam i.i.d. Tal aspecto configura-se como um bom sinal para a estimação e previsão com esses modelos, pois uma correlação entre termos de erro tende a desviar os valores dos coeficientes de seus valores verdadeiros, fazendo com que os preditores pareçam significantes ou insignificantes quando, na realidade, isso pode não ser verdade.

Uma vez escolhidos os modelos, a próxima etapa da metodologia consistiu na análise e modelagem das séries temporais de índices específicos retirados do *Google Trends*.

3.2. *GOOGLE TRENDS* E SAZONALIDADE

Como uma primeira análise, selecionei quatro termos dos 28 antes citados na Tabela 1.1.: "vagas", "vagas emprego", "vagas de emprego" e "emprego". Esses quatro termos são vistos como essenciais e são os principais a serem estudados nesse exercício.

Tabela 3.2.1.: Regressões MQO dos termos escolhidos na taxa de desocupação

	Variável dependente				
	Taxa de Desocupação				
	(1)	(2)	(3)	(4)	(5)
Constante	1.526*	2.736***	2.629***	0.141	3.915**
	(0.795)	(0.618)	(0.633)	(1.056)	(1.651)
vagas	0.104***				−0.066
	(0.013)				(0.115)
vagas de emprego		0.088***			0.076
		(0.010)			(0.187)
vagas emprego			0.089***		0.078
			(0.010)		(0.203)
emprego				0.112***	−0.015
				(0.015)	(0.061)
Observations	54	54	54	54	54
R^2	0.562	0.587	0.585	0.516	0.593
Adjusted R^2	0.554	0.579	0.577	0.507	0.560
Residual Std. Error	0.986	0.958	0.960	1.037	0.980
F Statistic	66.838***	73.927***	73.402***	55.409***	17.840***
Note:					*p<0.1; **p<0.05; ***p<0.01

A importância desses termos pode ser vislumbrada quando são realizadas regressões de mínimos quadrados ordinários que possuem a taxa de desocupação como variável dependente e os termos tanto individualmente quanto conjuntamente como variáveis explicativas. Nas regressões individuais (1), (2), (3) e (4) da Tabela 3.2.1. todos os termos registraram alta significância estatística, rejeitando a hipótese nula de o coeficiente ser igual a zero individualmente ao nível de 0.01. Quando os quatro termos

foram usados conjuntamente na regressão, contudo, eles se mostraram individualmente insignificantes.

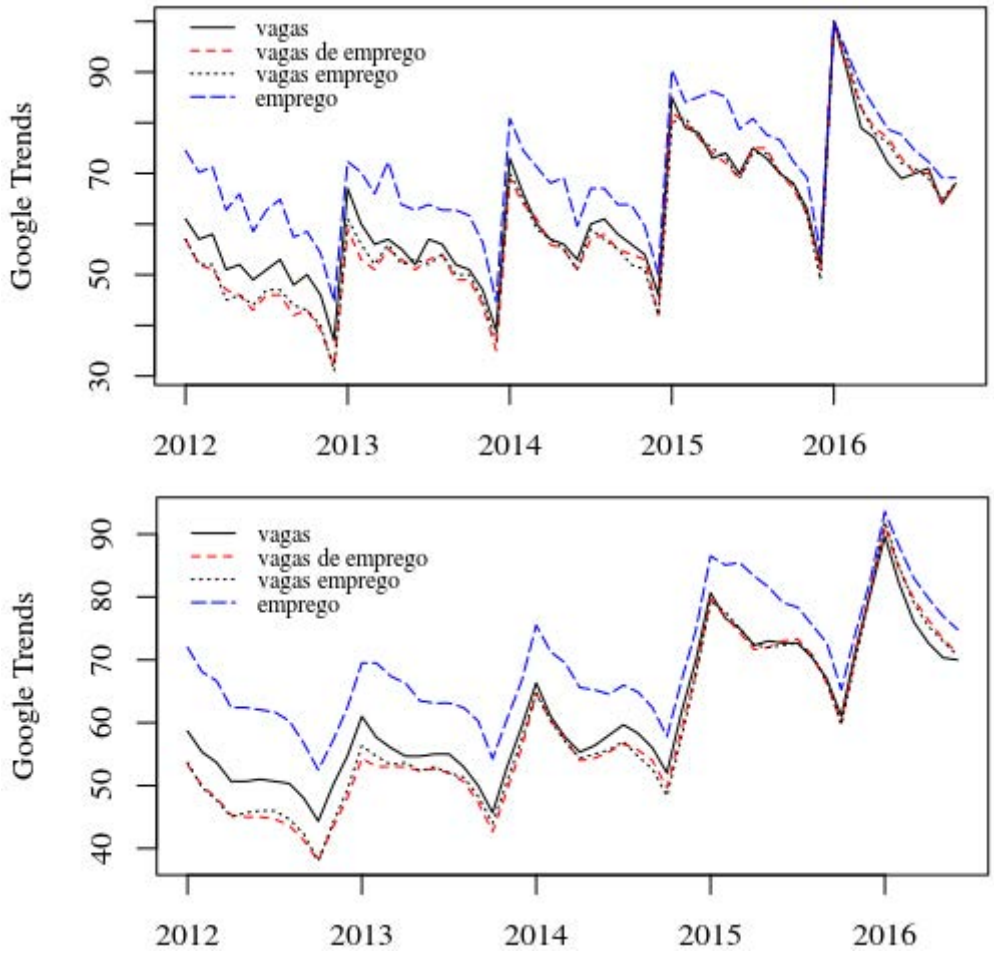
Tabela 3.2.2.: Matriz de correlações entre os termos escolhidos

	emprego	vagas	vagas de emprego	vagas emprego
emprego	1	0.970	0.956	0.961
vagas	0.970	1	0.991	0.992
vagas de emprego	0.956	0.991	1	0.998
vagas	0.961	0.992	0.998	1

Porém esse fenômeno se dá por causa da alta correlação entre eles, como mostra a Tabela 3.2.2. O teste de significância conjunta F, por sua vez, registrou um p-valor próximo de zero, rejeitando a hipótese nula dos quatro coeficientes serem iguais a zero conjuntamente ao nível de 0.01.

A escolha desses quatro termos também teve um componente de evidência anedótica. Busquei pensar – e entrevistei conhecidos – em quais seriam os primeiros termos de pesquisa que surgiriam como potenciais buscas caso estivesse desempregado e buscando um emprego. Esses foram os quatro termos mais recorrentes.

Figura 3.2.1.: Evolução dos índices do *Google Trends* na frequência mensal e em média trimestral



Analisando a trajetória temporal de cada um dos termos escolhidos, percebi claramente uma semelhança muito forte nas dinâmicas das séries e, principalmente, presenças de sazonalidade, tanto quando é calculada a média móvel trimestral da série, quanto quando ela é analisada em sua frequência mensal. A sazonalidade configura-se como picos muito elevados no mês de janeiro, seguido de uma queda razoavelmente constante até metade do ano, na qual há mais uma subida menos brusca. Após essa subida, a série continua a decair até alcançar a mínima do ano por volta de novembro e dezembro.

Expurgar a sazonalidade dessas quatro séries do *Google Trends* é um passo fundamental para o processo de *nowcasting*. O método escolhido foi o uso de variáveis *dummy* relativas a cada mês (ou a cada média trimestral). A especificação da regressão é:

$$Y_t = \alpha_t + X_{meses}\beta + \varepsilon_t \quad (8)$$

Onde Y_t corresponde às séries temporais uma vez em frequência mensal e outra em média móvel trimestral dos termos do *Google Trends* escolhidos; X_{meses} é uma matriz de variáveis *dummy* relativas a todos os meses menos Janeiro; e β é um vetor de coeficientes relativos a cada *dummy*. Desse modo, no final dessa etapa foram constituídas duas séries dessazonalizadas diferentes: (i) uma correspondente à série dos termos em frequência mensal que, uma vez expurgada a sazonalidade, foi transformada em média móvel trimestral; e (ii) outra na qual a sazonalidade foi expurgada diretamente na média móvel trimestral.

Figura 3.2.2.: Sazonalidade expurgada da série mensal dos termos escolhidos e série dessazonalizada mensal transformada em média móvel trimestral

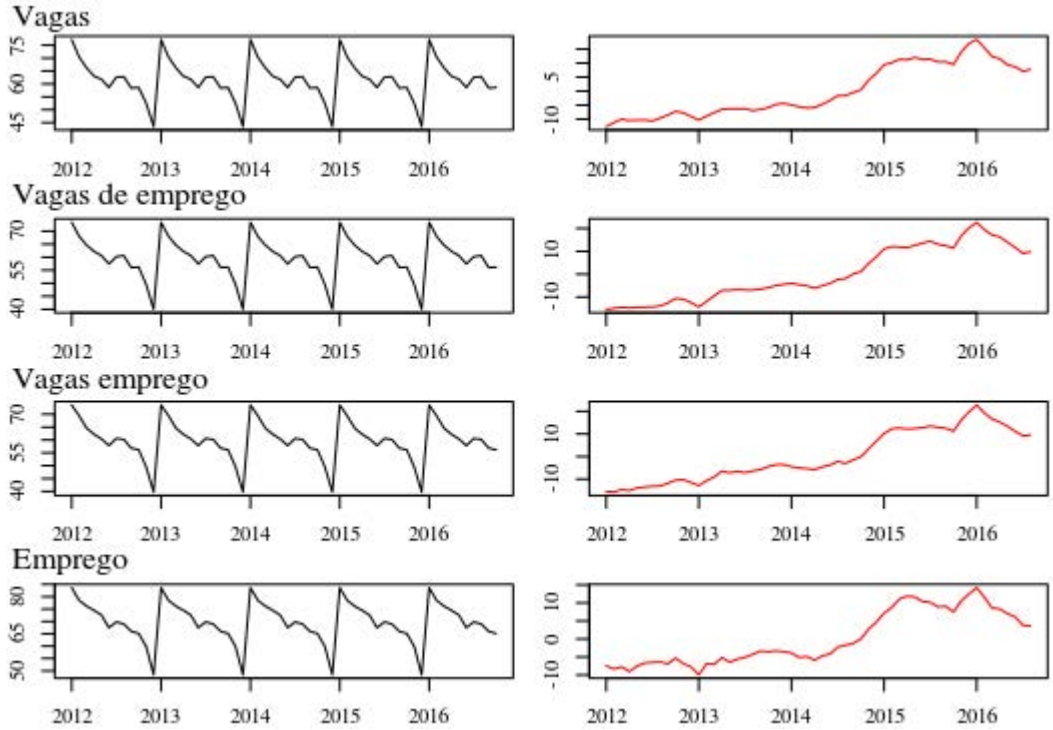
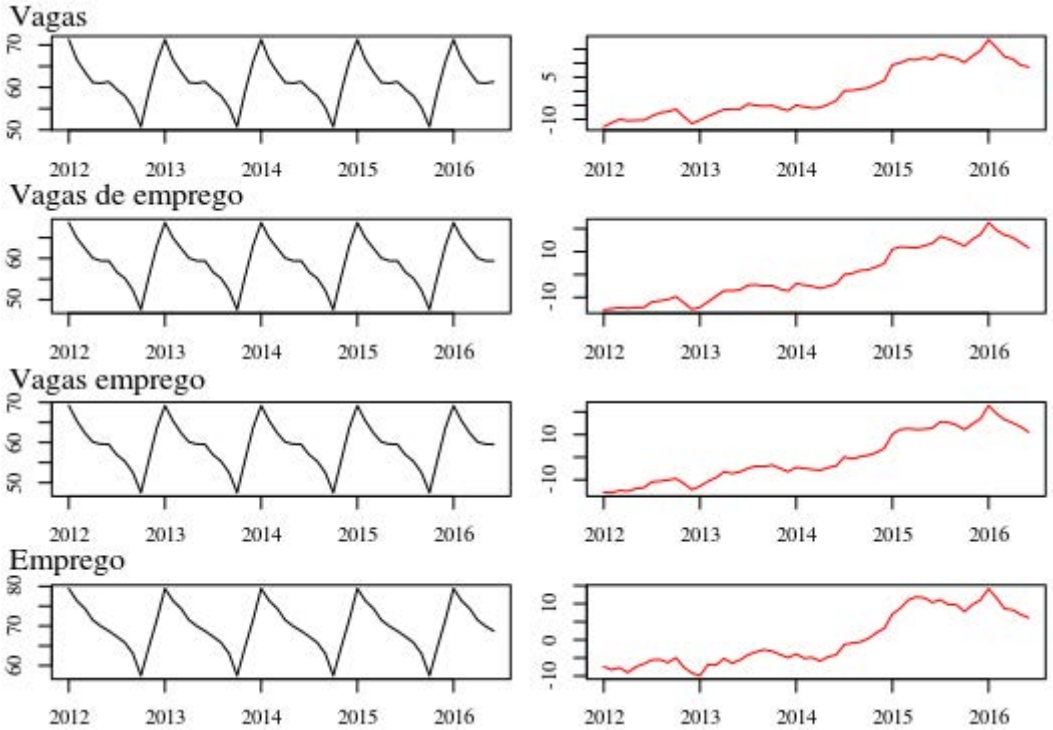


Figura 3.2.3.: Sazonalidade expurgada da série de média móvel trimestral dos termos e série da média móvel trimestral dessazonalizada



Como pode ser observado, uma vez retirada a sazonalidade, a dinâmica e formato da série de todos os termos escolhidos passaram a se assemelhar bastante entre si e com a série da taxa de desocupação. Como anteriormente não foi expurgada uma tendência,

seguiremos com as séries do *Google Trends* dessazonalizadas também sem expurgar uma tendência delas. Ademais, as séries dessazonalizadas também parecem indicar uma abrupta elevação no mesmo corte temporal da quebra estrutural da série da taxa de desocupação.

3.3. *EXPANDING WINDOW* E *ROLLING WINDOW*

Antes de testar as previsões com *Google Trends,* fez-se necessário decidir que método de previsão seria utilizado. A escolha do método baseou-se em selecionar aquele que minimizava o RMSE (*root-mean-square error*) e o MAE (*mean absolute error*) para previsão *out-of-sample* de 12 passos à frente da série da taxa de desocupação.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|Y_j - \hat{Y}_j| \tag{10}$$

Onde o *n* representa o número de passos utilizados na previsão *out-of-sample* (12, nesse caso específico); Y_j o resultado de fato observado no período *j* e $\hat{Y}_j$ a previsão do modelo para o período *j*. Os melhores modelos foram aqueles que obtiveram os menores resultados para ambos os tipos de erro. O melhor método de previsão foi aquele que apresentou os melhores modelos. Foram testados no total três métodos de previsão que podem ser englobados em duas metodologias diferentes: *Expanding Window* e *Rolling Window*.

O modelo de série completa (ou *Expanding Window)* utiliza todas as observações disponíveis em *t* para gerar um modelo de previsão para *t+1*. Quando for gerar um modelo de previsão para *t+2* são utilizadas todas as informações disponíveis, inclusive o valor de fato observado em *t+1* e não o previsto, e assim por diante. Desse modo, a janela de previsão sempre está se expandindo, i.e., sempre se aumenta o número de observações utilizadas para chegar ao melhor modelo preditivo.

Já no caso do *Rolling Window*, o período utilizado para previsão se comporta, como seu próprio nome demonstra, como uma janela móvel, i.e., a janela de previsão tem sempre um tamanho fixo, porém a cada novo modelo gerado ela se locomove um

passo a frente. Assim, o modelo utiliza toda informação disponível entre *t-n* e *t*, sendo *n* o tamanho da janela, para gerar um modelo de previsão para *t+1*. Em *t-n+1*, utilizar-se-ia toda a informação disponível entre e *t-n+1* e *t*+1 para calcular o melhor modelo para *t+2* e assim por diante. Nesse tipo de modelagem é necessário estabelecer *a priori* o *n*, o tamanho da janela. Por mais que esse tipo de modelo conte com menos observações do que o anterior, ele pode retornar melhores resultados por acabar dando mais peso a certas dinâmicas de curto prazo ou por ignorar outras de longo prazo.

Tabela 3.3.1.: Comparação do RMSE e MAE pra previsão *out-of-sample* 12 passos a frente dos modelos de *Expanding* e *Rolling Window* de 30 observações

	RMSE			*MAE*		
	Expanded	Rolling 30	Roll/Exp	Expanded	Rolling 30	Roll/Exp
AR(1)	0.363	0.362	0.997	0.281	0.271	0.965
MA(1)	1.614	1.532	0.949	1.529	1.460	0.955
ARMA(1,2)	0.224	0.208	0.928	0.180	0.180	0.998
ARMA(1,3)	0.226	0.199	0.884	0.177	0.170	0.964
ARMA(2,2)	0.235	0.226	0.963	0.182	0.183	1.009
ARIMA(1,1,0)	0.232	0.234	1.009	0.191	0.193	1.008
ARIMA(0,1,1)	0.271	0.272	1.004	0.209	0.208	0.999
ARIMA(0,1,2)	0.194	0.187	0.961	0.158	0.158	1.001
ARIMA(0,1,3)	0.196	0.192	0.981	0.160	0.163	1.019
ARIMA(1,1,2)	0.197	0.196	0.995	0.161	0.166	1.027
ARIMA(2,1,2)	0.197	0.194	0.984	0.156	0.157	1.006

Os resultados encontrados indicaram dois caminhos diferentes dependendo da maneira escolhida para mensurar o erro. Caso for escolhido o RMSE, temos que os modelos de *Rolling Window* tendem a se sair melhores, enquanto no caso do MAE os modelos de *Expanding Window* obtêm melhores resultados. Isso ocorre quando são utilizadas 30 observações na janela móvel. Se forem utilizadas apenas 20 observações, contudo, os resultados achados passam a ser parecidos para ambos os tipos de erros.

Tabela 3.3.2.: Comparação do RMSE e MAE pra previsão *out-of-sample* 12 passos a frente dos modelos de *Expanding* e *Rolling Window* de 20 observações

	RMSE			*MAE*		
	Expanded	Rolling 20	Roll/Exp	Expanded	Rolling 20	Roll/Exp
AR(1)	0.363	—	—	0.281	—	—
MA(1)	1.614	1.274	0.789	1.529	1.215	0.795
ARMA(1,2)	0.224	0.206	0.916	0.180	0.167	0.925
ARMA(1,3)	0.226	0.207	0.920	0.177	0.164	0.927
ARMA(2,2)	0.235	0.234	0.998	0.182	0.183	1.008
ARIMA(1,1,0)	0.232	0.232	1.001	0.191	0.192	1.006
ARIMA(0,1,1)	0.271	0.275	1.013	0.209	0.208	0.998
ARIMA(0,1,2)	0.194	0.185	0.949	0.158	0.153	0.965
ARIMA(0,1,3)	0.196	0.187	0.953	0.160	0.151	0.946
ARIMA(1,1,2)	0.197	0.189	0.958	0.161	0.152	0.944
ARIMA(2,1,2)	0.197	0.189	0.955	0.156	0.154	0.991

Nesse caso, tirando alguns modelos específicos, o método de janela móvel com 20 observações se saiu melhor do que aqueles de série completa em quase todos os

casos. Além disso, os resultados se mostraram melhores do que aqueles encontrados para o modelo de *Rolling Window* com 30 observações.

Mesmo os valores do RMSE e MAE sendo menores em termos absolutos, se fez ainda necessária a realização de testes Diebold-Mariano (DIEBOLD; MARIANO, 1995) para identificar se de fato esses valores são diferentes entre si. As hipóteses do teste são:

$$H_0 : E\left[L\left(\varepsilon^1_{t+h|t}\right)\right] = E\left[L\left(\varepsilon^2_{t+h|t}\right)\right]$$
$$H_1 : E\left[L\left(\varepsilon^1_{t+h|t}\right)\right] \neq E\left[L\left(\varepsilon^2_{t+h|t}\right)\right] \quad (11)$$

Onde H_0 pode ser reescrita como,

$$H_0 : E[d_t] = 0$$

L(x) é uma função perda, sendo nesse exercício o RMSE ou MAE; $\varepsilon_{t+h|t}$ é um vetor dos erros de previsão relativos aos dois modelos a serem testados; e d_t é a diferença entre as funções perdas dos dois erros de previsão.

Tabela 3.3.3.: Resultados dos testes Diebold-Mariano para RMSE e MAE comparando as metodologias de *Rolling Window 20* e *Expanding Window*

	RMSE		*MAE*	
	Estatística DM	p-valor	Estatística DM	p-valor
MA(1)	-4.844	0.001	-4.481	0.001
ARMA(1,2)	-2.650	0.023	-2.629	0.023
ARMA(1,3)	-1.008	0.335	-0.917	0.379
ARMA(2,2)	-0.660	0.523	-0.639	0.536
ARIMA(1,1,0)	-2.405	0.035	-2.385	0.036
ARIMA(0,1,1)	0.170	0.868	-0.001	1.000
ARIMA(0,1,2)	-0.690	0.504	-0.708	0.494
ARIMA(0,1,3)	-0.086	0.933	-0.138	0.893
ARIMA(1,1,2)	0.036	0.972	-0.037	0.971
ARIMA(2,1,2)	-1.799	0.100	-1.833	0.094

A *H_0 não* foi rejeitada em quase todos os casos para ambos os critérios de erro escolhidos ao nível de 0.1, exceto nas modelagens ARIMA(1,1,0), MA(1) e ARMA(1,2) nas quais ela foi rejeitada ao nível de 0.05. A conclusão tirada, contudo, é ambígua, pois enquanto o modelo ARIMA(1,1,0) se saiu melhor com *Expanding Window*, os modelos ARMA(1,2) e MA(1) obtiveram melhores resultados com a modelagem de *Rolling Window*. Escolhi continuar o restante do exercício, portanto, utilizando a modelagem de *Rolling Window* com janela de 20 observações.

3.4. ESPECIFICAÇÕES DAS REGRESSÕES

Como demonstrado na introdução, a defasagem da divulgação dos resultados para as taxas de desocupação (Tabela 1.1.) é relevante. Assim sendo, caso fosse desejado prever a taxa de desocupação em *t+1*, o modelo seria restrito apenas às observações de *t* e de períodos anteriores. A situação se torna especialmente mais problemática caso fosse desejado prever a taxa vigente (aquela relativa ao final do mês imediatamente anterior ao dia atual) que, dado o atraso de dois meses, configurar-se-ia então como *t+2*. Essa previsão estaria restrita apenas às observações de dois períodos anteriores.

Os índices do *Google Trends* surgem, então, como uma maneira de aprimorar estas previsões, porque, sendo disponibilizados em tempo real, existiriam dados para *t+1* e *t+2* já consolidados e, ainda por cima, para *t+3* divulgados como a parcial do mês em questão. A intuição leva a acreditar que, ao adicionar nos modelos escolhidos variáveis contemporâneas à previsão desejada, os critérios escolhidos RMSE e MAE diminuiriam. Por consequência, o *nowcasting* se tornaria mais preciso. Escolhi, ao longo de todo esse exercício, realizar previsões de um passo a frente *out-of-sample* para testar a hipótese de que o *Google Trends* é capaz de melhorar previsões de presente.

Existem diversas maneiras de adicionar os índices do *Trends* aos modelos. Um primeiro aspecto a se levar em conta é a temporalidade dos termos adicionados. Supondo que a previsão buscada é a da taxa de desocupação em *t* (Y_t), pode-se adicionar à especificação da regressão o índice defasado, contemporâneo à Y_t ou um período à frente (GT_{t-1}, GT_t e GT_{t+1}, respectivamente).

Outra escolha a ser feita é o tipo de dessazonalização a ser utilizada. Como exposto na Seção 3.2., o mesmo método de dessazonalizar via variáveis *dummy* foi realizado de duas maneiras diferentes: direto na série trimestral (GT^q) e na série mensal para depois tirar a média móvel trimestral (GT^m).

Portanto, foram testados diferentes formatos de regressões para cada um dos quatro termos selecionados anteriormente (“vagas de emprego”, “vagas emprego”, “vagas” e “emprego”):

$$Y_t = \alpha_t + ARIMA + GT_t^d + \varepsilon_t \tag{12}$$

$$Y_t = \alpha_t + ARIMA + GT_{t-1}^d + GT_t^d + \varepsilon_t \tag{13}$$

$$Y_t = \alpha_t + ARIMA + GT_t^d + GT_{t+1}^d + \varepsilon_t \tag{14}$$

Onde *d* simboliza o tipo de dessazonalização escolhida (*m*, *q* ou série com sazonalidade); *ARIMA* representa o processo ARIMA(p,d,q) escolhido; e *GT* representa cada um dos quatro termos escolhidos para a análise individualmente.

A análise dos impactos dos índices do *Trends* na previsão da taxa de desocupação, por conseguinte, se deu da seguinte forma:

(i) foram escolhidas as três modelagens do ciclo na Seção 3.3. que minimizaram o RMSE da previsão 12 passos à frente pelo método de *Rolling Window* com janela de 20 observações. Estas foram comparadas cada uma com cinco modelos que utilizavam a mesma modelagem do ciclo e *Rolling Window* com 20 observações, porém possuíam algum dos formatos (12), (13) ou (14). Basicamente os melhores modelos com e sem índices do *Trends* foram postos lado a lado;

(ii) o mesmo foi feito para o MAE;

(iii) foram expostos modelos com índices do *Trends* que seguem modelagens do ciclo não escolhidas nas seleções anteriores (Seção 3.1. e 3.3.) que, porém, retornaram critérios de avaliação (RMSE e MAE) relativamente baixos; e

(iv) foram realizadas *shrinkage regressions* para: testar o desempenho de modelos que utilizavam apenas índices do *Trends* e para seleção dos termos relevantes entre os 28 termos possíveis.

4. PRINCIPAIS RESULTADOS

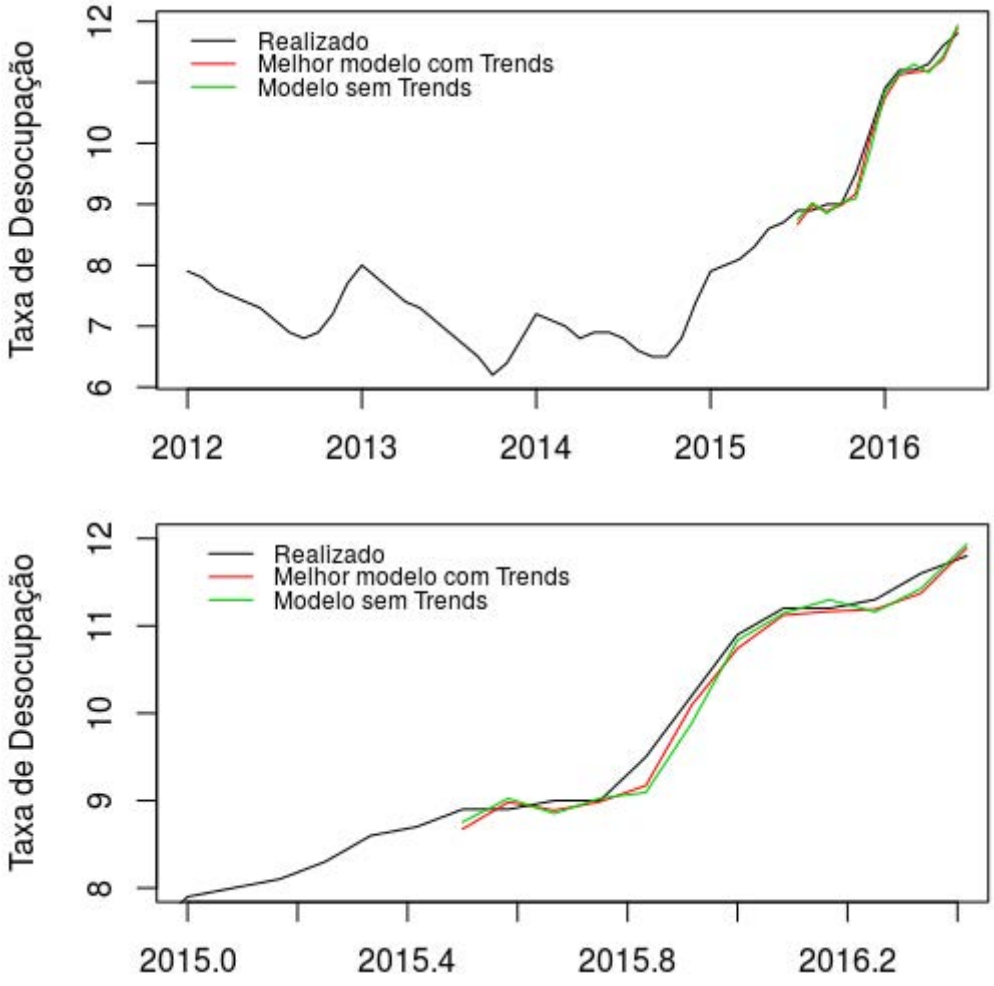
4.1. ESCOLHA DOS MODELOS PARA ERRO QUADRÁTICO (RMSE)

Nessa Seção, expus as melhores modelagens encontradas ao longo de diferentes metodologias e testes. Os três melhores modelos encontrados dado o critério RMSE e a metodologia de *Rolling Window* com 20 observações foram: ARIMA(0,1,2), ARIMA(0,1,3) e ARIMA(1,1,2).

Tabela 4.1.1.: RMSE da modelagem ARIMA(0,1,2) e cinco melhores modelagens que utilizam índices do *Google Trends*

	Absoluto	Relativo
ARIMA(0,1,2)	0.185	1.000
(12), emprego, m	0.156	0.848
(12), vagas, m	0.161	0.873
(12), vagas emprego, m	0.166	0.898
(12), emprego, com Sazonalidade	0.173	0.938
(12), vagas de emprego, m	0.174	0.943

Figura 4.1.1.: Realização da série e suas previsões *out-of-sample* da melhor modelagem ARIMA(0,1,2) com índices do *Google Trends* e sem

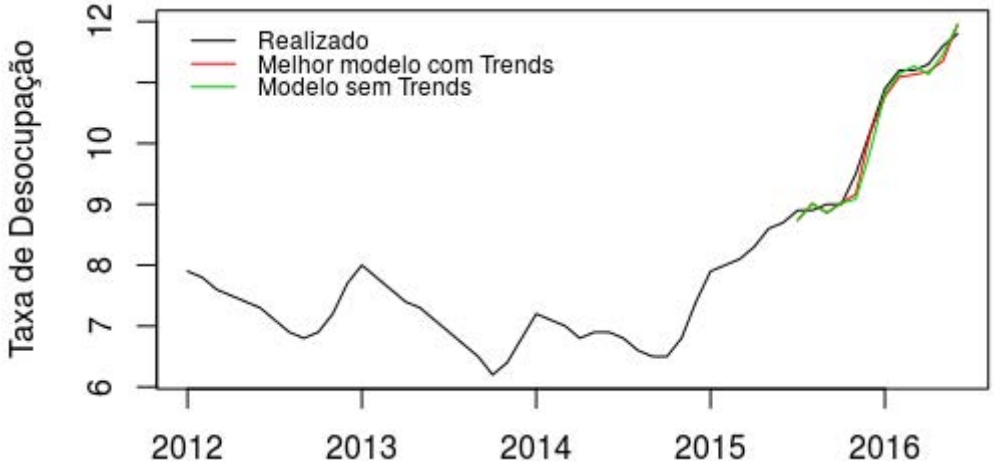


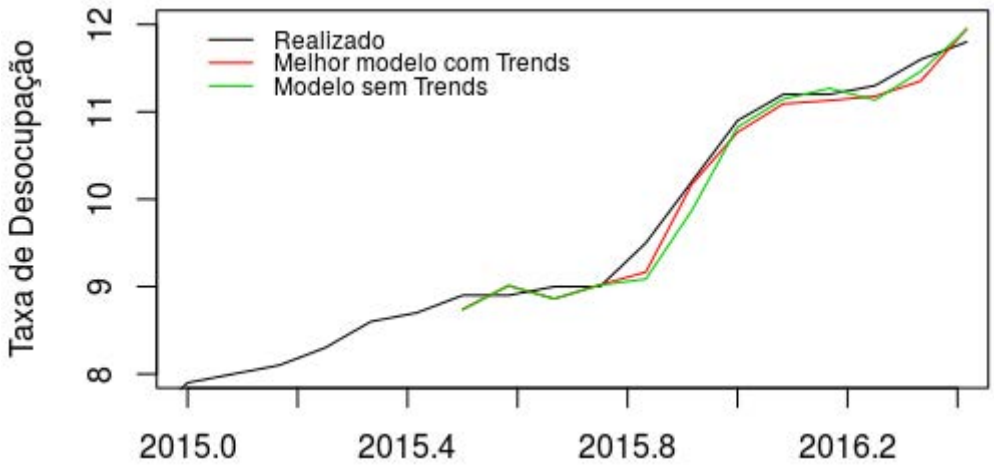
Como explicitado na Tabela 4.1.1., as cinco modelagens com índices do *Trends* que minimizaram o RMSE possuem um erro menor do que o mesmo modelo sem índices. É interessante notar que, para todos os resultados, o formato escolhido foi o (12), i.e., aquela que leva em conta apenas o índice contemporâneo à previsão. A escolha da forma de expurgar a sazonalidade, por sua vez, não foi homogênea. Os três melhores modelos selecionados usaram índices com suas sazonalidades expurgadas pelo mesmo processo: retirá-la na série mensal para, uma vez feito isso, transformar em média móvel trimestral. Outro modelo escolhido levou em conta a sazonalidade do índice. No Gráfico 4.1.1. nota-se que a previsão sem índices do *Google Trends* e a previsão adicionando o termo "emprego" dessazonalizado mensalmente possuem uma dinâmica bastante semelhante.

Tabela 4.1.2.: RMSE da modelagem ARIMA(0,1,3) e cinco melhores modelagens que utilizam índices do *Google Trends*

	Absoluto	Relativo
ARIMA(0,1,3)	0.187	1.000
(12), vagas, m	0.159	0.852
(12), vagas emprego, m	0.165	0.885
(12), emprego, com Sazonalidade	0.170	0.910
(12), vagas, com Sazonalidade	0.173	0.926
(12), vagas de emprego, m	0.174	0.931

Figura 4.1.2.: Realização da série e suas previsões *out-of-sample* da melhor modelagem ARIMA(0,1,3) com índices do *Google Trends* e sem



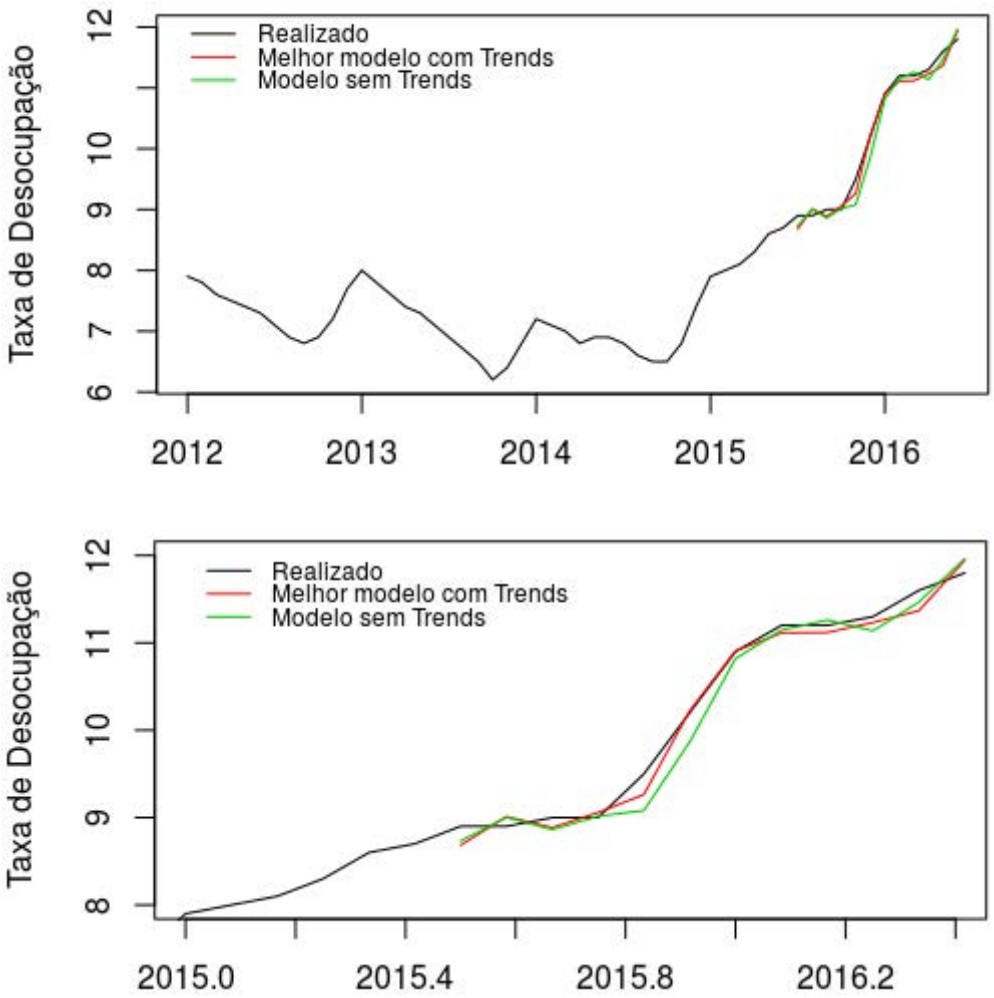


Como no caso anterior, as modelagens escolhidas pertencem ao formato (12) e os melhores modelos tiveram a sazonalidade do índice do *Trends* expurgada ainda na série mensal, como no caso anterior. A série da previsão com índices, nesse caso, do final de 2015 até março de 2016 subestima a série original. A dinâmica captada por ela, entretanto, segue um formato extremamente parecido com o da série original no período citado. De meados de 2015 até o final desse ano a previsão com índices captou de maneira mais precisa em relação à previsão sem índices a dinâmica de crescimento da série da taxa de desocupação de fato realizada.

Tabela 4.1.3.: RMSE da modelagem ARIMA(1,1,2) e cinco melhores modelagens que utilizam índices do *Google Trends*

	Absoluto	Relativo
ARIMA(1,1,2)	0.189	1.000
(12), emprego, m	0.137	0.726
(12), vagas, m	0.156	0.828
(12), vagas emprego, m	0.168	0.891
(12), emprego, com Sazonalidade	0.173	0.913
(12), vagas, com Sazonalidade	0.176	0.930

Figura 4.1.3.: Realização da série e suas previsões *out-of-sample* da melhor modelagem ARIMA(1,1,2) com índices do *Google Trends* e sem



Mais uma vez, foram apenas selecionados modelos de formato (12) e entre os melhores estão aqueles que levam em conta as séries do índice do *Trends* dessazonalizada de maneira mensal. Os dois melhores modelos escolhidos nessa etapa apresentaram os menores RMSE's entre todos os modelos expostos nessa seção. Isso ocorreu, em parte, pelo fato de ambos terem sido capazes de captar de maneira bastante precisa a dinâmica ascendente da série no final de 2015.

Dados esses resultados e levando em consideração como critério de análise o RMSE, a melhor modelagem é um ARIMA(1,1,2) utilizando a série do *Google Trends* do termo "emprego" dessazonalizada mensalmente. Essa modelagem conseguiu prever com elevada precisão o crescimento da taxa de desocupação entre agosto e dezembro de 2015, além de apresentar resultados satisfatórios e bastante próximos da séria realizada para o ano de 2016.

4.2. ESCOLHA DOS MODELOS PARA ERRO ABSOLUTO (MAE)

Os três melhores modelos dado o critério MAE e a metodologia de *Rolling Window* com 20 observações foram: ARIMA(0,1,2), ARIMA(0,1,3) e ARIMA(1,1,2). Nesse exercício, mesmo utilizando um tipo de erro diferente para a avaliação *out-of-sample*, as mesmas modelagens do ciclo utilizadas para o RMSE foram escolhidas.

Tabela 4.2.1.: MAE da modelagem ARIMA(0,1,2) e cinco melhores modelagens que utilizam índices do *Google Trends*

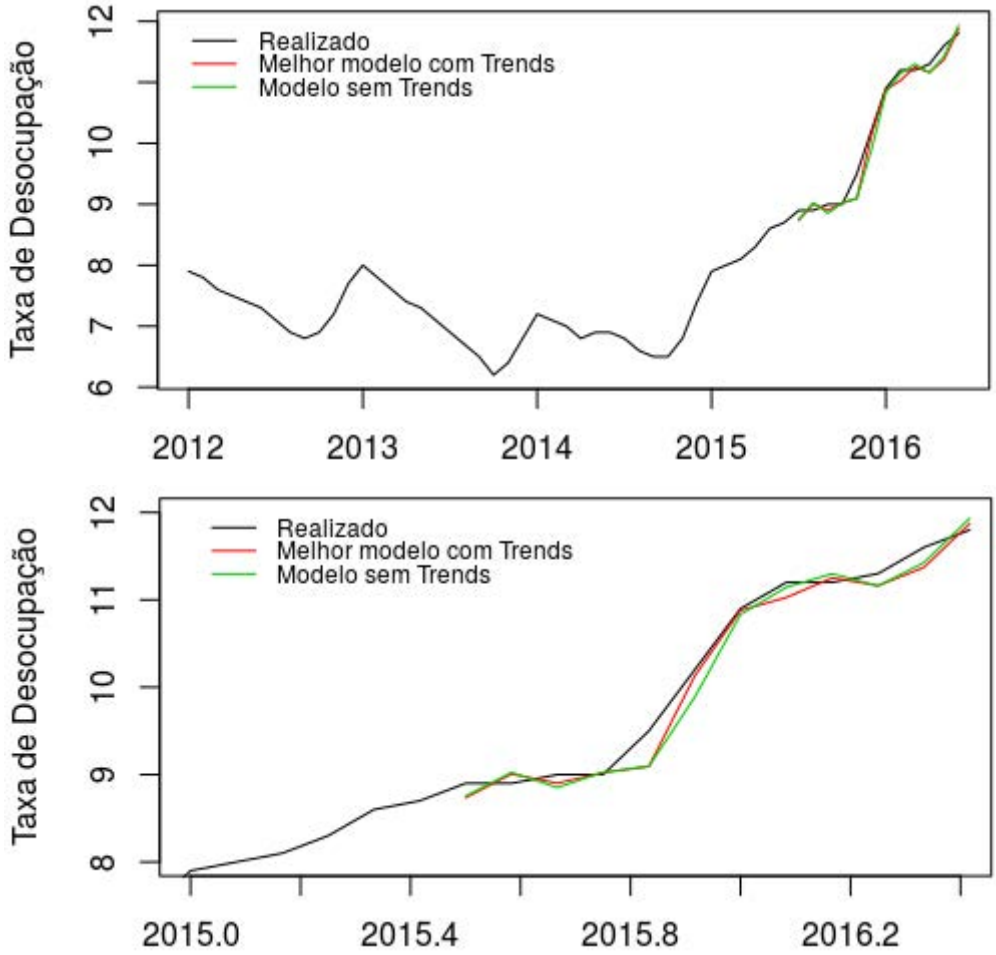
	Absoluto	Relativo
ARIMA(0,1,2)	0.153	1.000
(12), emprego, m	0.131	0.856
(12), vagas, m	0.139	0.909
(12), vagas emprego, m	0.142	0.932
(14), emprego, m	0.143	0.934
(12), emprego, com Sazonalidade	0.146	0.958

Quanto à seleção de modelos para a modelagem ARIMA(0,1,2) usando como critério de avaliação o MAE houve algumas divergências em relação ao mesmo exercício usando RMSE. Dessa vez foram selecionados não apenas modelos com formato (12), como também um de formato (14), i.e., levando em conta o índice do *Trends* contemporâneo e de um passo a frente em relação à variável dependente. O melhor modelo, em contrapartida, continuou sendo o mesmo selecionado quando utilizado o RMSE como critério de seleção. Assim, os gráficos e suas análises são os mesmos (Gráfico 4.1.1).

Tabela 4.2.2.: MAE da modelagem ARIMA(0,1,3) e cinco melhores modelagens que utilizam índices do *Google Trends*

	Absoluto	Relativo
ARIMA(0,1,3)	0.151	1.000
(12), vagas emprego, m	0.129	0.855
(12), vagas, m	0.136	0.897
(14), emprego, m	0.136	0.902
(12), vagas de emprego, m	0.139	0.921
(12), emprego, com Sazonalidade	0.140	0.927

Figura 4.2.1.: Realização da série e suas previsões *out-of-sample* da melhor modelagem ARIMA(0,1,3) com índices do *Google Trends* e sem



No caso da modelagem ARIMA(0,1,3), ao utilizar a minimização do MAE como critério, escolheu-se o termo "vagas emprego" ao invés de "vagas" para a especificação do melhor modelo. O formato, contudo, continua sendo o (12) e a dessazonalização continua sendo feita na série mensal.

Tabela 4.2.3.: MAE da modelagem ARIMA(1,1,2) e cinco melhores modelagens que utilizam índices do *Google Trends*

	Absoluto	Relativo
ARIMA(1,1,2)	0.152	1.000
(12), emprego, m	0.115	0.754
(12), vagas emprego, m	0.124	0.814
(12), vagas, m	0.129	0.847
(14), vagas, m	0.139	0.914
(12), emprego, com Sazonalidade	0.141	0.925

Na modelagem ARIMA(1,1,2), foram obtidos resultados qualitativamente muito parecidos com aqueles da modelagem desse mesmo processo tendo o RMSE como critério de avaliação. Como na Seção 4.1., o formato (12) com o termo "emprego" (MAE, nesse caso) entre todos os modelos testados.

4.3. OUTRAS MODELAGENS DO CICLO

Buscando testar a capacidade dos índices do *Google Trends* de melhorar modelos, selecionei todas as modelagens do ciclo citadas anteriormente nesse trabalho que não foram utilizadas nas Seções 4.1. e 4.2., pois ou não possuíam os menores critérios de informação AIC e BIC ou não minimizavam os erros quadrático e/ou absoluto da previsão *out-of-sample* de 12 passos a frente. Assim, foram rodadas previsões de método *Rolling Window* de 20 observações com modelos de formato (12), (13) e/ou (14) que utilizavam essas modelagens do ciclo antes deixadas de lado. Os cinco modelos que minimizaram os critérios de erro RMSE e MAE foram reportados.

É importante frisar que todas as modelagens do ciclo que acabaram sendo selecionadas nessa Seção, com exceção do ARIMA(2,1,2), não estiveram nem entre as onze modelagens escolhidas no primeiro corte, no qual foi utilizado como preceito os critérios de informação AIC e BIC. (Seção 3.1.).

Tabela 4.3.1.: RMSE com e sem índices do *Google Trends* dos modelos selecionados

	RMSE	RMSE sem Trends
ARIMA(2,1,3), (12), emprego, m	0.144	0.184
ARIMA(1,1,3), (12), vagas, m	0.145	0.171
ARIMA(1,1,3), (12), emprego, m	0.147	0.171
ARMA(2,1,3), (12), vagas, m	0.148	0.184
ARIMA(3,1,0), (12), emprego, m	0.149	0.233

Algumas dessas modelagens, como ARIMA(3,1,0) e ARIMA(1,1,3) sem índices computaram erros menores do que as outras modelagens selecionadas na Seção 4.2. também sem índices. Quando adicionados os índices do *Trends*, os modelos escolhidos que minimizaram o RMSE sempre foram de formato (12) utilizando os termos "vagas" e "emprego". Esse resultado do formato escolhido é o mesmo encontrado nas Seções 4.1. e 4.2. É interessante exaltar que os RMSE's obtidos nessa etapa figuram entre os melhores desse exercício inteiro.

O poder explicativo dos índices do *Trends* se fez especialmente presente no caso da modelagem ARIMA(3,1,0). Quando não utilizados os índices, seu RMSE é um dos piores entre os modelos ARIMA(p,1,q) no geral. Adicionando-os, entretanto, há uma melhora de aproximadamente 36%.

Tabela 4.3.2.: MAE com e sem índices do *Google Trends* dos modelos selecionados

	MAE	MAE sem Trends
ARIMA(2,1,3), (12), emprego, m	0.117	0.157
ARIMA(2,1,3), (12), vagas, m	0.117	0.157
ARIMA(1,1,3), (12), vagas emprego, m	0.118	0.133
ARMA(1,1,3), (12), vagas, m	0.122	0.133
ARIMA(2,1,2), (12), emprego, m	0.123	0.154

Os resultados achados anteriormente a respeito dos formatos praticamente se repetem quando utilizei o critério de erro absoluto (MAE). Todos os modelos que minimizaram o MAE tiveram o formato (12) com os termos "emprego" e "vagas" tendo um papel predominante. Mais uma vez todas as modelagens que levaram em conta índices do *Trends* figuraram entre as melhores achadas nesse exercício inteiro. Ademais, nesse caso é importante ressaltar que as modelagens sem *Trends*, além de não haverem sido selecionadas pelos critérios de informação AIC e BIC, devolveram valores para seus erros absolutos muito parelhos ou menores do que os achados anteriormente na Seção 4.2. para as modelagens sem índices.

4.4. LASSO, *ADAPTIVE* LASSO E *ELASTIC NET*

Uma vez compreendido o poder dos índices de melhorar modelos de previsão que já levam em conta alguma modelagem ARMA, busquei testar o poder explicativo deles por si só. Portanto, nessa subseção foram utilizados todos os 28 termos expostos na Tabela 1.1. Devido ao número relativamente alto de regressores perante o número de observações da série da taxa de desocupação (56 observações), fez-se necessário uma *shrinkage estimation* dos dados, i.e., diminuir o valor de potenciais coeficientes irrelevantes da regressão para zero. Isso é feito através da adição de um termo de penalidade na estimação que varia de acordo com a modelagem desejada. Foram utilizados três métodos diferentes de *shrinkage estimation*: LASSO (TIBSHIRANI, 1996), *Adaptive* LASSO ou adaLASSO (ZOU, 2006) e *Elastic Net* (ZOU; HASTIE, 2005).

Como já explicitado, tais métodos de *shrinkage estimation* consistem em calcular um estimador com base em alguma função de perda convexa – como a minimização do quadrado dos resíduos – adicionando um termo de penalidade.

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=1}^{T} \ell\big((Y_t, X_t), \beta\big) + \lambda P(\beta) \tag{15}$$

Onde $\ell\big((Y_i, X_i), \beta\big)$ é uma função perda convexa que depende da variável dependente, das variáveis explicativas e dos coeficientes estimados de uma regressão; λ é um parâmetro não negativo regularizador do *shrinkage*; e $P(\beta)$ é uma função de penalidade dependente dos coeficientes β estimados que permite atribuir alguns componentes do vetor $\hat{\beta}$ para zero dado valor de λ.

Nesse exercício, a equação de estimação (15) assumiu o formato específico:

$$\hat{\beta}^{(n)} = \arg\min_{\beta} \ \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j \left[(1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right], \qquad 0 < \alpha \leq 1 \tag{16}$$

Onde β é um vetor de parâmetros; Y é um vetor transposto de resultados realizados; X é uma matriz de dados; n são os modelos possíveis (LASSO, adaLASSO e *Elastic Net*) α é um parâmetro que pondera entre o uso dos modelos LASSO (α = 1) e modelos *Elastic Net* (0 < α < 1); e $\widehat{w}$ é o vetor adaptativo de pesos ponderados pela base de dados. Dependendo do tipo de modelo escolhido, $\widehat{w}$ assume valores diferentes.

$$\widehat{w} = \begin{cases} 1, & modelo = \{LASSO, Elastic\ Net\} \\ \left|\hat{\beta}_j^{ini}\right|^{-\gamma}, & modelo = \{adaLASSO\} \end{cases} \tag{17}$$

O vetor adaptativo $\widehat{w}$ depende, por sua vez, de um valor inicial para os betas. Portanto, foram usados como vetor de pesos para os modelos adaptados os coeficientes encontrados na regressão LASSO, fazendo com que o modelo adaLASSO tivesse que escolher entre todas as variáveis ainda não cortadas do LASSO. O algoritmo desses métodos (FRIEDMAN; HASTIE; TIBSHIRANI, 2010) computa um total de 100 λ's , no qual o valor máximo de λ é aquele que possui o menor valor possível e, ao mesmo tempo, torna zero o valor de todos os coeficientes relativos às variáveis na matriz X. Cada λ é associado a um modelo com um determinado grau de liberdade referente ao número de variáveis cujos coeficientes não foram zerados excluindo o intercepto. O melhor modelo escolhido para cada um dos métodos foi aquele que minimizou os respectivos critérios BIC de informação. A matriz X de variáveis explicativas é composta por todas as 28 palavras explicitadas na Tabela 1.1., com destaque especial para os quatro termos ("vagas", "vagas de emprego", "vagas emprego" e "emprego")

usados anteriormente que entraram apenas na sua forma dessazonalizada mensalmente. O exercício foi realizado para $\alpha = \{0.25, 0.5, 1\}$ e $\gamma = 1$. A previsão de 12 passos a frente foi calculada através do método de *Expanding Window* explicitado na Seção 3.3.

O processo de previsão com esses modelos se deu da seguinte forma: (i) foram estimados os coeficiente dos modelos relativos a cada λ computado; (ii) foi calculado o *fit* e o critério BIC de informação de cada um desses modelos; (iii) o modelo que apresentou o menor critério BIC foi selecionado para continuar o processo; (iv) caso o método escolhido tivesse sido LASSO ou *Elastic Net*, a previsão foi feita para um passo a frente com base nesse modelo e o processo recomeçou adicionando uma observação (*Expanding Window*); caso o método escolhido tivesse sido adaLASSO, os coeficientes do LASSO compuseram o termo de penalidade e estimou-se um novo modelo - com o qual a previsão foi feita - através desse mesmo processo, porém adicionando esse termo de penalidade. A previsão, depois, se deu da mesma maneira dos processos LASSO e *Elastic Net*.

O BIC, por sua vez, foi calculado de duas maneiras diferentes e, para cada maneira de calculá-lo, resultados muito discrepantes foram encontrados.

Uma primeira maneira de calcular o BIC foi exatamente como expus anteriormente na Equação (3). Desse modo, para cada janela das modelagens via *Expanding Window* foi selecionado o melhor modelo por algum dos métodos de *shrinkage* expostos e feita a previsão de um passo a frente.

Figura 4.4.1.: Comparação dos erros de previsão entre modelos

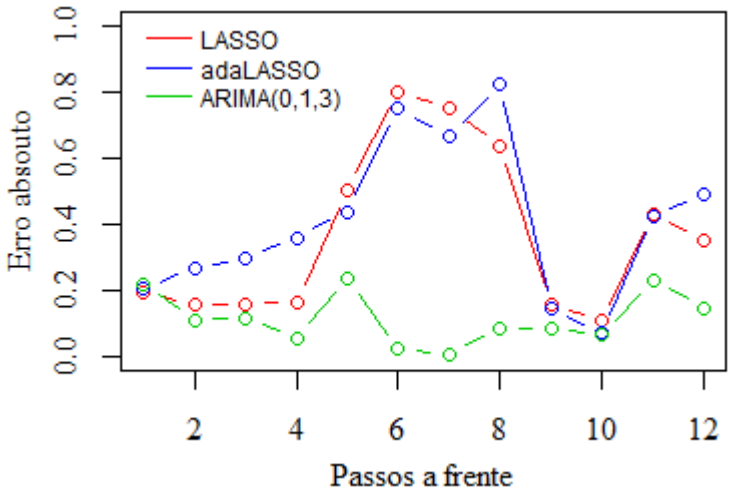
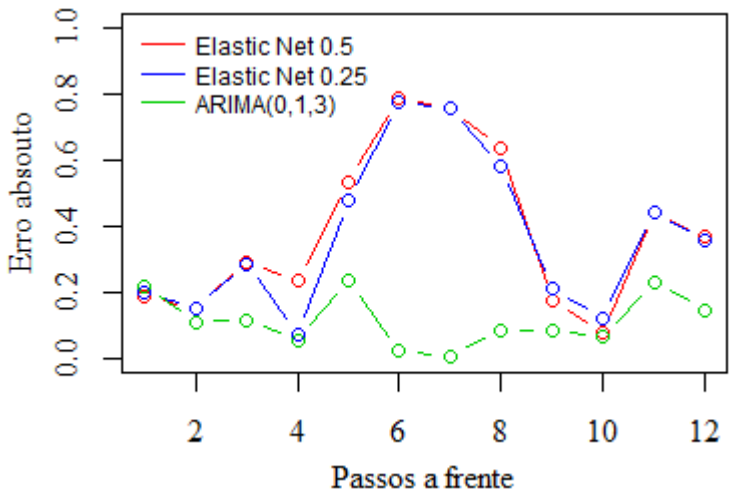


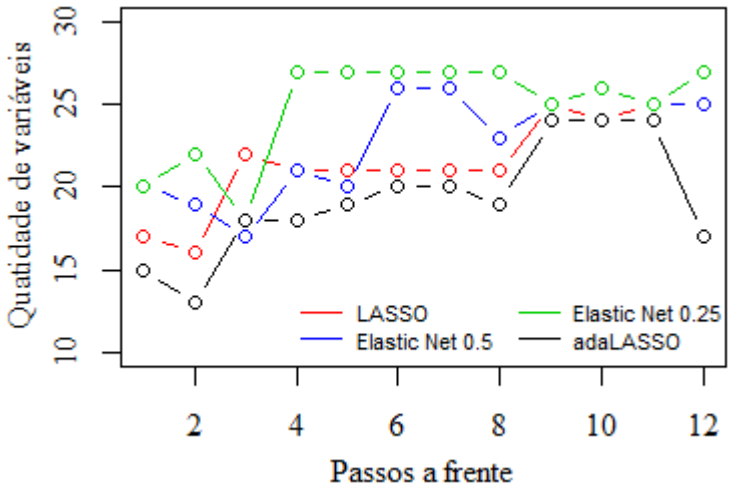
Figura 4.4.2.: Comparação dos erros de previsão entre modelos



Como comparação foi usada a melhor previsão da taxa de desocupação encontrada na Seção 4.2. sem usar índices do *Google Trends*. A previsão escolhida foi aquela gerada pela modelagem ARIMA(0,1,3) através do método de *Rolling Window* com janela de tamanho 20. Os diferentes métodos de *shrinkage estimation* obtiveram erros de previsão bastante parecidos entre si. Para os mesmos períodos, esses erros das *shrinkage regressions* se aproximaram – como nas primeiras quatro previsões ou nas 3 últimas – dos erros de previsão do modelo ARIMA(0,1,3). As previsões dos períodos intermediários, por sua vez, se afastaram consideravelmente. Tal resultado, contudo, não é negativo, pois, tendo em vista a mudança de dinâmica da série da taxa de desocupação a partir do final de 2015 e a ausência de defasagens da série entre os regressores, a previsão se mostrou relativamente precisa.

Quanto à seleção de variáveis, o uso do BIC como descrito na Equação (3) teve menor sucesso.

Figura 4.4.3.: Comparação da quantidade de variáveis selecionadas para cada método de *shrinkage*



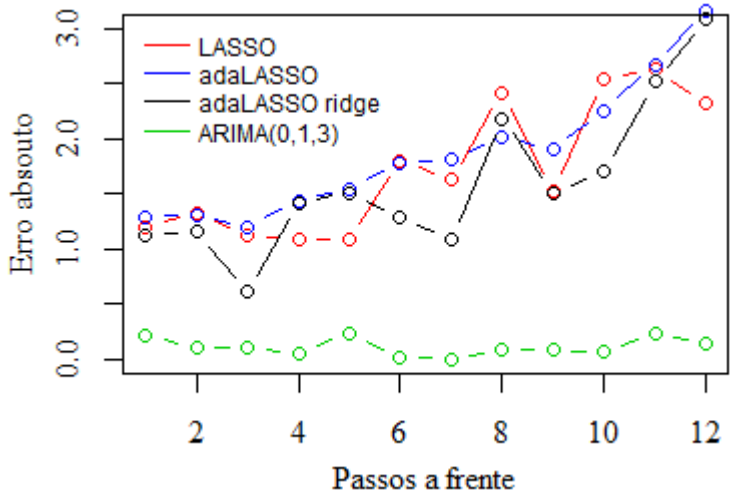
O adaLASSO, como esperado, se mostrou o menos parcimonioso dos métodos de *shrinkage regression*. Mesmo assim, o número de variáveis selecionadas em cada modelo para diferentes previsões se mostrou ainda muito elevado. No caso do *Elastic Net* com α = 0.25, para cinco previsões seguidas fora usadas 27 das 28 variáveis disponíveis. Tal resultado pode mostrar a incapacidade da utilização desse formato do BIC para selecionar variáveis quando elas são muito correlacionadas entre si, que é o caso da matriz X de índices do *Google Trends*. Vale ressaltar que os valores para λ selecionados e, logo, os coeficientes de cada modelo preditivo utilizado foram muito próximos daqueles encontrados via *cross-validation*.

A outra maneira de calcular o BIC foi proposta por H. Zou (ZOU; HASTIE; TIBSHIRANI, 2007).

$$BIC^{LASSO} = \frac{\|Y - \hat{\mu}\|^2}{n\sigma^2} + \frac{\log(n)}{n}\widehat{df}(\hat{\mu}) \tag{18}$$

Onde *Y* é a série realizada da variável dependente; $\hat{\mu}(\lambda_m)$ é um *fit* estimado via *shrinkage estimation* que depende de cada um dos λ's calculados; *n* é o número de observações; σ^2 é a variância de *Y;* e *df* são os graus de liberdade relativos a cada $\hat{\mu}(\lambda_m)$. É importante frisar que, nesse caso, os graus de liberdade foram calculados como no algoritmo utilizado ao longo desse exercício. Basicamente, os graus de liberdade são aproximados para o número de coeficientes – menos o intercepto – diferentes de zero. Ao adicionar $\widehat{df}(\hat{\mu})$ na equação, ela se torna mais punitiva e, então, era esperado encontrar um menor número de variáveis da matriz X selecionadas.

Figura 4.4.4.: Comparação dos erros de previsão entre modelos

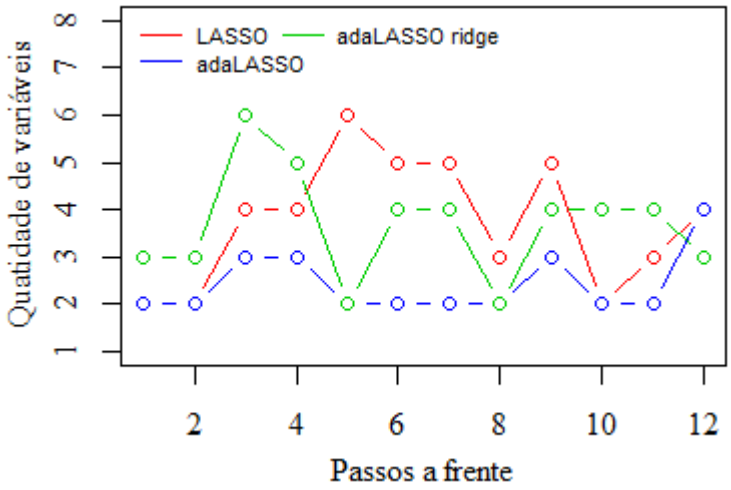


Nessa parte do exercício, foram estimados apenas modelos de formato LASSO e adaLASSO. Contudo, dessa vez foram usados como vetor de pesos para os modelos

adaptados os coeficientes encontrados na regressão LASSO e na regressão *ridge*. A regressão *ridge* é semelhante à LASSO, porém, ao invés de penalizar o valor absoluto dos coeficientes, penaliza o quadrado deles. Desse modo, ela acaba por não excluir nenhuma variável, atribuindo um valor muito baixo para os coeficientes das irrelevantes. O que acabei fazendo, no fundo, foi forçar o adaLASSO a fazer uma seleção com todas as variáveis disponíveis. No caso da Equação (16), a regressão *ridge* é obtida quando $\alpha = 0$.

As previsões, como expostas na Figura 4.4.4., pioraram significativamente. Explicações para tal fato podem ser a alta correlação entre as variáveis, fazendo com que esse desenho do BIC exclua muitas delas do modelo selecionado, ou a incapacidade dessas variáveis explicarem a taxa de desocupação. Contudo, esse método de BIC-Lasso deve ser usado primariamente para seleção de variáveis, como argumentado em Zou et al. (2007).

Figura 4.4.5.: Comparação da quantidade de variáveis selecionadas para cada método de *shrinkage*



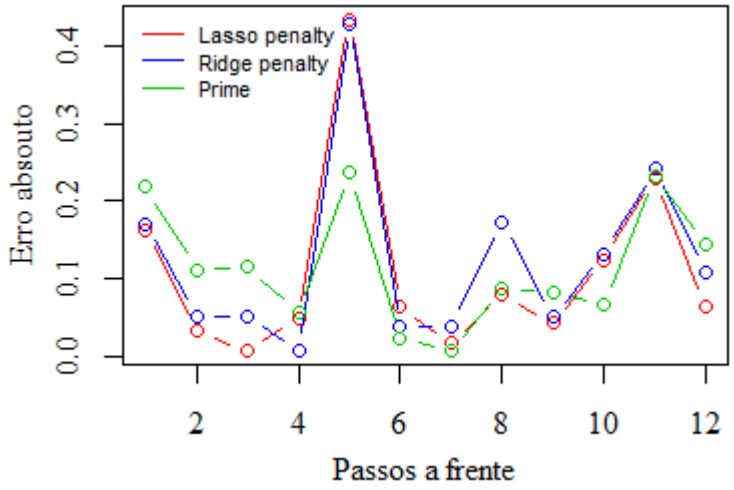
De fato, o *shrinkage* teve muito mais sucesso com essa nova fórmula para o BIC. Nenhum dos modelos em nenhuma das previsões selecionou mais do que seis variáveis, diferentemente dos modelos anteriores que selecionaram em média 21 variáveis. Também é interessante notar como o adaLASSO ponderado pelas estimativas do LASSO, como previsto, sempre selecionou menos do que os outros dois modelos, exceto na previsão do último período. Isso indica sua característica de corrigir parte da parcimônia associada aos modelos LASSO.

Como última parte do trabalho, foram realizadas regressões adaLASSO na série da taxa de desocupação com todas as observações com termos de penalidade baseados nas estimativas dos coeficientes obtidos na regressão LASSO e *ridge* para selecionar os

índices que melhor produzem um *fit* pra série completa. Uma vez escolhidos os índices do *Trends* para cada um dos termos de penalidade, foram feitas previsões com a modelagem ARIMA(1,1,2) através do método de *Rolling Window* com janela de 20 observações e formato (12) adicionando esses termos escolhidos pelo adaLASSO. Essa modelagem foi escolhida, pois foi aquela que obteve tanto o menor MAE quanto o menor RMSE na Seção 4.2. Dessas previsões foram retirados os erros absolutos relativos para cada passo previsto. Esse erros foram comparados com os erros da modelagem de formato (12) de processo ARIMA(1,1,2) e utilizando os índices dessazonalizados do termo "emprego". Essa modelagem foi chamada de *Prime*.

Para a penalidade baseada na regressão *ridge* foi escolhido o termo "vagas emprego" – o que acaba por corroborar com a evidência anedótica antes explicitada na Seção 3.2. – e para a penalidade baseada na regressão LASSO foram escolhidos os termos "vagas emprego", "fgts caixa" e "infojobs vagas". Essa última seleção acabou por ser interessante, pois cada termo pertence a um conjunto semântico da Tabela 1.1. diferente.

Figura 4.4.6.: Comparação dos erros de previsão entre modelos



Devido à alta correlação entre os termos "emprego" e "vagas emprego", todos os erros acabam tendo uma dinâmica muito parecida. Tirando a previsão para o passo $t+5$, na qual a diferença entre os erros foi relativamente maior, os modelos que utilizaram as variáveis escolhidas pelas regressões adaLASSO tiveram uma performance muito satisfatória, obtendo previsões para alguns períodos de tempo mais precisas que aquelas do modelo *Prime*.

5. CONCLUSÃO

O *delay* na divulgação de dados oficiais e a necessidade cada vez maior de poder contar com estimativas precisas da realidade para a tomada de decisão sobre políticas públicas são temas centrais em muitos países. Assim, a capacidade de melhorar o *nowcasting* de variáveis tidas como essenciais para um *policymaker,* como a taxa de desemprego, utilizando dados *online* em tempo real de domínio público*,* se tornou um tópico muito recente de estudo na área das Ciências Econômicas.

Nesse trabalho, examinei através de métodos *out-of-sample* se os índices do *Google Trends* são capazes de melhorar a previsão e, logo, o *nowcasting* da taxa de desocupação brasileira. Obtive evidências de que o acréscimo dos índices nas especificações de regressões de fato melhora o *nowcasting*. Utilizando a metodologia Box-Jenkins para previsão de séries temporais, modelei a série da taxa de desocupação para mais de 30 modelagens do ciclo, selecionei algumas via critérios *in-sample* e realizei previsões *out-of-sample* através de diferentes métodos. Explorei os resultados adicionando os índices do *Trends* e sem os índices, usando RMSE e MAE como critério. Em ambos os casos, foi possível encontrar para todas as modelagens do ciclo escolhidas pelo menos cinco modelos com índices na sua especificação que obtiveram melhores resultados para previsão *out-of-sample*. Algumas melhoras chegaram a alcançar a casa dos 30 pontos percentuais.

Os resultados se mostraram razoavelmente robustos na medida em que a mesma especificação da regressão – índice do *Trends* dessazonalizado adicionado de maneira contemporânea à taxa de desocupação – e os mesmos termos – "vagas" e "emprego" – figuraram sempre entre os melhores modelos possíveis. Esses resultados sugerem que houve melhoras consideráveis no *nowcasting* da taxa de desocupação ao utilizar os índices do *Trends* e que, num mundo cada vez mais propenso a geração de *Big Data* e dados em tempo real, pode se tornar cada vez mais fácil a estimação de variáveis fundamentais para a tomada de decisão econômica, possibilitando até a diminuição do *delay* de tempo das divulgações oficiais.

Ademais, testei a eficácia também através de critérios *out-of-sample* de uma previsão utilizando apenas os índices selecionados por métodos de *shrinkage regression* (LASSO, adaLASSO e *Elastic Net*). Os resultados se mostraram ligeiramente inferiores daqueles encontrados anteriormente e os métodos utilizados retornaram um número muito elevado de variáveis relevantes para a previsão. Ao aumentar a punição dos

critérios de seleção associados a esses métodos, pude usá-los não para previsão, mas para seleção das variáveis relevantes. Ao usá-los dessa forma, foi selecionado um grupo pequeno de três variáveis e as utilizei em conjunto com modelagens ARIMA(p,d,q) (como havia feito) para testar se elas seriam capazes de bater os melhores modelos encontrados anteriormente. Os resultados *out-of-sample* dos melhores modelos se mostraram extremamente parecidos com os resultados gerados pelos modelos utilizando variáveis escolhidas por *shrinkage regressions*.

Interpreto meus resultados como uma evidência de que os índices do *Trends* são realmente capazes de melhorar não só o *nowcasting*, como também a estimação de diversas variáveis econômicas. Consistente com esse resultado está a intuição de que variáveis contemporâneas, i.e., variáveis geradas em tempo real têm um poder preditivo enorme a oferecer. Não só porque ajudam a diminuir os problemas que o *delay* de divulgações oficiais promove para a tomada de decisões, como também porque são de fácil acesso, rápida atualização e conseguem condensar importantes informações sobre a dinâmica da sociedade.

Finalmente, dados os resultados achados, é natural se questionar qual o limite dos benefícios e malefícios que a rápida expansão dos dados em tempo real e do *Big Data* pode trazer para a vida em sociedade e todas suas dimensões. Essa discussão foge do escopo desse trabalho, porém, ao fim desse exercício, tem-se a sensação de que foi mostrado um grande benefício que essa nova forma de informação pode trazer.

6. REFERÊNCIAS

CARRIÈRE-SWALLOW, Y.; LABBÉ F. *Nowcasting with Google Trends in an Emerging Market*. Journal of Forecasting, 32, p. 289-298, 2013.

CHOI, H.; VARIAN, H. *Predicting Initial Claims for Unemployment Benefits*. Technical Report, Google, 2009a.

CHOI, H.; VARIAN, H. *Predicting the Present with Google Trends*. Technial Report, Google, 2009b

GOEL, S.; HOFMAN, J.M.; LAHAIE, S.; PENNOCK, D.M.; WATTS, D.J. *Predicting Consumer Behaviour with Web Search",* Yahoo! Research, 2010.

ASIKTAS, N.; ZIMMERMAN, K. *Google Econometrics and Unemployment Forecasting*. Applied Economics Quaterly, 55(2), p. 107-120, 2009.

SUHOY, T. *Query Indices and a 2008 Downturn: Israeli Data.* Discussion Paper No. 2009.06, Research Department, Bank of Israel, 2009.

D'AMURI, F.; MARCUCCI, J. '*Google It!' Forecasting the US Unemployment Rate with a Google Job Search Index.* SSRN, 2010.

SHIMSHONI, Y.; EFRON, N.; MATIAS Y. *On the Predictability of Trends*. Technical Report, Google, 2009.

ETTREDGE M.; GERDES J.; KARUGA G. *Using Web-based Search Data to Predcit Macroeconomic Statistics.* Communications of the ACM, Vol. 48, p. 87-92, 2005.

SCHMIDT T.; SIMEON V. *Forecasting Private Consumption: Survey-based Indicators vs. Google Trends*. Ruhr Economic Papers #155, RWI, 2009.

AKAIKE H. *A new looke at the statistical model identification*. IEEE, Transactions on Automatic Control, 19 (6), p. 716-723, 1974.

BURNHAM, K.; ANDERSON D. *Model Selection and Multimodel Inference*. 2nd Edition, Springer-Verlag, 2002.

SCHWARZ G. *Estimating the Dimension of a Model*. The Annuals of Statistics, Vol. 6, No. 2, p. 461-464, 1978.

POSADA D.; BUCKLEY T. *Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests*. Syst. Biol. 53 (5), p. 793-808, 2004.

BOX G.; JENKINS G. *Time Series Analysis: Forecasting and Control.* 4th Edition, Wiley, 2008/1970.

LJUNG G.; BOX G. *On a measure of lack of fit in time series models.* Biometrika, 65, 2, p. 297-303, 1978.

DIEBOLD F.; MARIANO R. *Comparing Predictive Accuracy.* Journal of Business and Economic Statistics, 13, p. 253-265, 1995.

TIBSHIRANI R. *Regression Shrinking and Selection via the Lasso*. Journal of the Royal Statistical Society Series B, v. 58, Issue 1, p. 267-288, 1996.

ZOU H. *The Adaptive Lasso and its Oracle Properties*. Journal of the American Statistical Society, vol. 101, 476, p. 1418-1429, 2006.

ZOU H.; HASTIE T. *Regularization and variable selection via the elastic net.* J. R. Statistical Society Bulletin, 67, Part 2, p. 301-320, 2005.

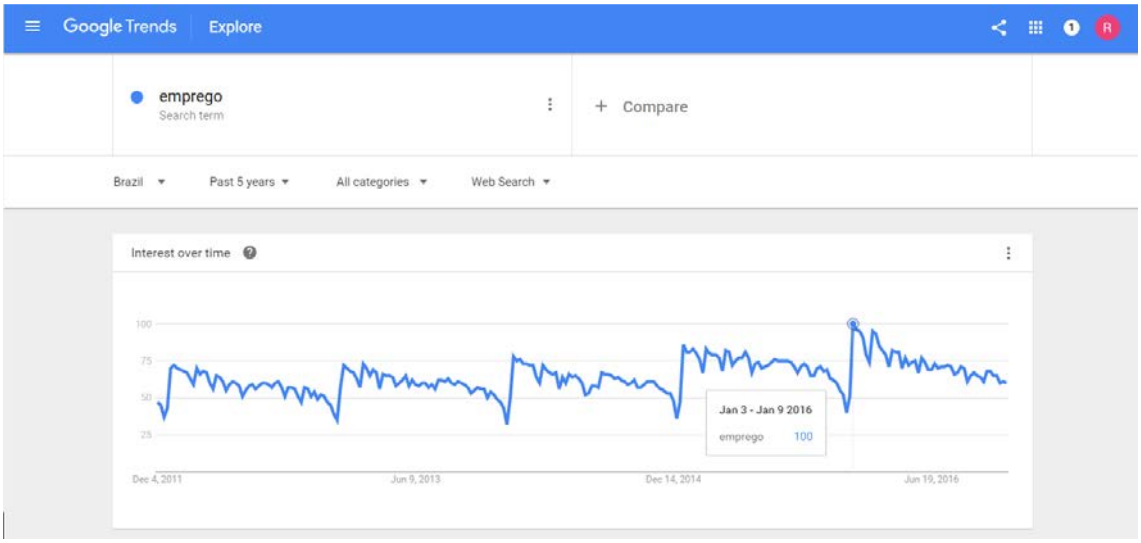
ZOU H.; HASTIE T.; TIBSHIRANI R. *On the "Degrees of Freedom" of the Lasso*. The Annals of Statistics, vol. 35, 5, 2007.

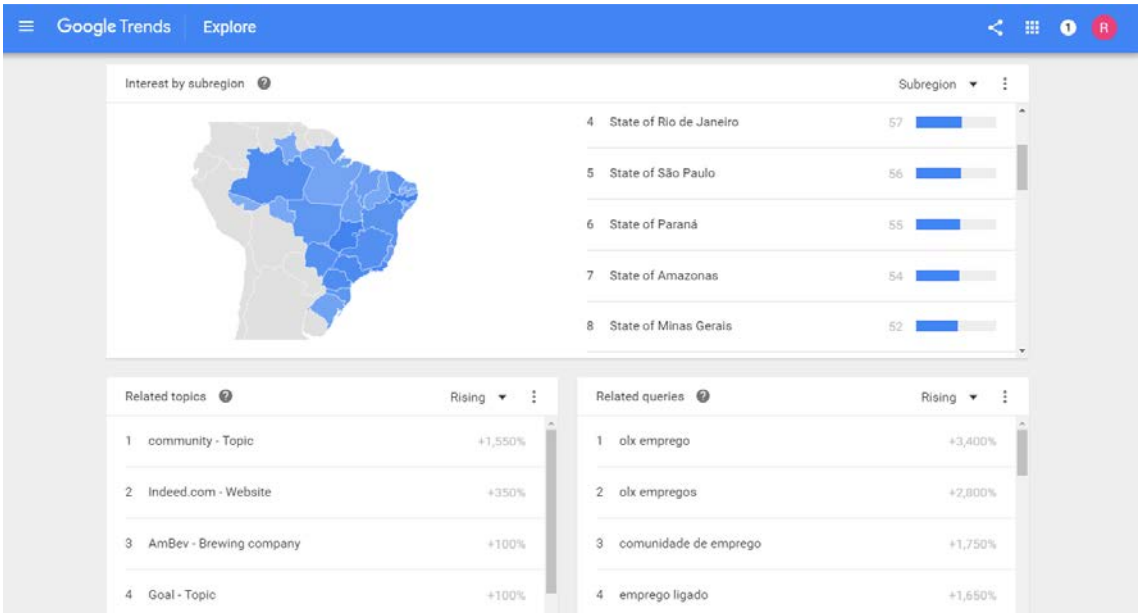
FRIEDMAN J.; HASTIE T; TIBSHIRANI R. *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software, 33(1), 1-22, 2010.

MAREK H. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. R package version 5.2. http://CRAN.R-project.org/package=stargazer, 2015.

DAHL D. *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2., http://CRAN.R-project.org/package=xtable, 2016.

APÊNDICE A


Google Trends
Explore
emprego
Search term
Compare
Brazil
Past 5 years
All categories
Web Search
Interest over time
100
75
50
25
Jan 3 - Jan 9 2016
emprego 100
Dec 4, 2011
Jun 9, 2013
Dec 14, 2014
Jun 19, 2016


Google Trends
Explore
Interest by subregion
Subregion
4 State of Rio de Janeiro 57
5 State of São Paulo 56
6 State of Paraná 55
7 State of Amazonas 54
8 State of Minas Gerais 52
Related topics
Rising
1 community - Topic +1,550%
2 Indeed.com - Website +350%
3 AmBev - Brewing company +100%
4 Goal - Topic +100%
Related queries
Rising
1 olx emprego +3,400%
2 olx empregos +2,800%
3 comunidade de emprego +1,750%
4 emprego ligado +1,650%