PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

O USO DO GOOGLE PARA *NOWCASTING* DA ATIVIDADE COMERCIAL: APLICAÇÕES AO SETOR VAREJISTA

Júlio Cezar Monteiro de Barros
Número de Matricula: 1313322

Orientadores: Eduardo Zilberman &
Pedro Carvalho Loureiro de Souza

Junho de 2017

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

O USO DO GOOGLE PARA *NOWCASTING* DA ATIVIDADE COMERCIAL: APLICAÇÕES AO SETOR VAREJISTA

Júlio Cezar Monteiro de Barros
Número de Matricula: 1313322

Orientadores: Eduardo Zilberman &
Pedro Carvalho Loureiro de Souza

Junho de 2017

"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"

Júlio Cezar Monteiro de Barros

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

Agradecimentos

A todos aqueles que foram fundamentais para a minha trajetória acadêmica, bem como para a minha trajetória de vida.

A Deus, pela paz e felicidade de cada dia vivido.

Àquelas que lutaram cada dia de suas vidas e lançaram os alicerces para que hoje eu esteja aqui. Emilia, Maria e Marlene, obrigado.

Àqueles que me construíram como pessoa, me seguraram pelas mãos, caminharam comigo onde quer que eu fosse, me suportaram em minhas escolhas e me direcionaram para a vida. Fernando e Verônica, é tudo por vocês.

À minha família, Fernanda, Claudia, Carlos, Alessandra e Sérgio. A alegria de quem somos é meu bem mais precioso.

À Amanda Schutze, Dimitri Szerman e Eduardo Zilberman, por terem me tutoriado pelos caminhos da vida acadêmica, pelo conhecimento que me passaram e pelo amor à economia que desenvolvi.

Aos meus orientadores, que me conduziram por este processo.

Aos meus amigos, parte fundamental de quem sou e cujo agradecimento nominal tomaria mais de uma página. Em especial, Mariana, Thaissa, Marta, Marina, Felipe, Maitê, Isabela, Rebeca, Duda, Carol, Karen, Juliana, Luiza, João e Roni.

Sumário

Lista de Tabelas

1 Introdução

A vida contemporânea pode ser entendida a partir da presença de um dos seus principais planos de fundo: a rede mundial de computadores interconectados. A sociedade atual usa a internet não como mero acessório, mas como meio que permeia as relações sociais, interpessoais e comerciais. A abrangência do acesso permitiu que a globalização avançasse a passos largos, modificando em processo contínuo os padrões estabelecidos ao trazer informação de forma mais célere. Modificam-se as estruturas tradicionais de comunicação, de entretenimento e também as emoções intrínsecas ao ser-humano, os padrões de consumo e comportamentais, estendendo-se o mundo real ao virtual. Demonstramos que os dados de utilização de internet por indivíduos explicam de certa forma o comportamento dos mesmos na economia real.

É no campo virtual onde se consolida grande parte das interações humanas. Para além do simples entretenimento e dos relacionamentos sociais, também na rede novas estruturas comerciais e de serviços podem ser consolidadas. É habitual utilizar as plataformas para comprar e vender produtos e serviços, no que se nomea *e-commerce*, além de ser comum a utilização de novos canais de troca dos mesmos, na emergência de uma chamada economia colaborativa. Também o simples consumidor utiliza a grande quantidade de informação disponível para comparar preços, discutir com outros consumidores, avaliar a qualidade, prestar reclamação, analisar prós e contras de um bem, de forma que o *review* de produtos, videos de *unboxing* e sites de reclamação viralizaram. A internet não somente se caracteriza por uma fonte infinita de informação para o consumidor mais aguçado, mas também responde aos questionamentos mais simples, como inquirir sobre a pizzaria mais próxima de casa.

Milhões de pessoas utilizam esses recursos a todo instante: mais de quarenta mil pesquisas são performadas pelo Google por segundo, isto é, 3,5 bilhões de pesquisas diárias [1]. É uma quantidade de dados imensa que já é utilizada para fins comerciais nos diversos produtos de marketing e que passou a ser recentemente aplicada nas ciências econômicas, principalmente após estudo de Varian e Choi (2009), com foco no *nowcasting* de variáveis macroeconômicas, mas com potencial de aplicação também com outros enfoques, como economia comportamental.

Seja no mercado, no meio acadêmico ou no setor público, a análise de indicadores econômicos se baseia – além da essência interpretativa do pesquisador – nos dados utilizados. Desta forma, seja em projeções pontuais de indicadores, na análise de cenários, na elaboração de políticas públicas ou em estratégias comerciais, empresas, governos e acadêmicos vêm se debruçando sobre métodos de pesquisa mais rebuscados estatisticamente. Essa otimização, entretanto, encontra um empecilho: a defasagem nas divulgações de da-

[1] Dados do Internet Live Stats. Disponível em <http://www.internetlivestats.com/>

dos econômicos. Estes dados são muitas vezes de difícil mensuração, ou tomam tempo para um tratamento estatístico prévio à divulgação. Soma-se a isso o tempo particular que agências governamentais tomam, seja por questões burocráticas, por falta de recursos públicos para um trabalho mais produtivo ou ainda pela captura das instituições pela esfera política. Assim, popularizou-se o termo "olhar pelo retrovisor" em alusão ao referido *lag*, como se olhasse para o passado.

Um exemplo claro é a Pesquisa Mensal do Comércio, divulgada mensalmente pelo IBGE. A pesquisa investiga a atividade varejista de empresas com vinte ou mais pessoas ocupadas e cujas unidades estejam dentro do território nacional. Os dados de comércio trazem à luz o desempenho da demanda agregada da economia (via consumo das famílias), os ciclos e movimentos intrínsecos dos mercados, além de serem fundamentais para o *policy-making*, uma vez que o varejo, consumo, produção e inflação são altamente correlacionados. Além disso, esses dados são considerados antecedentes da atividade produtiva. As empresas que apresentam as caracteristicas supracitadas alimentam ao longo do mês a base de dados do IBGE, seja por via física ou por versão eletrônica. Entretanto, os dados só são disseminados pelo Instituto após seis semanas aproximadamente, vide o calendário disposto na Tabela 1.

Tabela 1: Calendario de Divulgação da PMC

Periodo	Divulgacao
Novembro/2016	10-jan-2017
Dezembro/2016	14-fev-2017
Janeiro/2017	30-mar-2017
Fevereiro/2017	12-abr-2017
Março/2017	11-mai-2017
Abril/2017	13-jun-2017
Maio/2017	12-jul-2017
Junho/2017	15-ago-2017
Julho/2017	12-set-2017
Agosto/2017	11-out-2017
Setembro/2017	14-nov-2017
Outubro/2017	13-dez-2017
Novembro/2017	09-jan-2018
Dezembro/2017	09-fev-2018

Fonte: IBGE (2017)

Se por um lado é possível elaborar uma crítica sobre tamanha defasagem, por outro, no setor privado, diversos segmentos desenvolveram outros indicadores com base em dados em tempo real. A exemplo do "Índice Cielo do Varejo Ampliado", uma empresa de cartões de crédito com considerável penetração no mercado nacional consegue utilizar

seus dados de venda para dimensionar as vendas no varejo [2]. Da mesma forma, o Google tem a capacidade de filtrar e categorizar o volume de pesquisas em sua plataforma, identificando tendências no comportamento dos indivíduos.

Assim, dados em tempo real têm grande potencial de "prever" a situação corrente, ou seja, trazer à luz o presente, superando o problema de defasagem antes mencionado. Neste estudo, pesquisa-se se dados de pesquisas no Google atuam como bons preditores das vendas no varejo no Brasil. Os motivos para tanto são *straightforward*: consumidores pesquisam online antes de fazerem compras ou contratarem um serviço. Deve-se considerar, ainda, que algumas pesquisas tomam tempo antes de se efetivarem em forma de vendas. É natural imaginar que um indivíduo que almeja comprar um carro pesquisará mais do que um consumidor que busca o *delivery* mais em conta para o jantar.

Fundamentanda nos dados do Google, a pesquisa demonstra que os mesmos podem trazer informações em tempo real sobre o comportamento dos consumidores, o que abre um grande espaço de utilização não só pelo setor privado, mas para análise da situação corrente da economia pelo setor público. O dispositivo oferecido pelas buscas online alcançou um progresso expressivo de mais de 40% para algumas categorias de comércio analisadas.

Esta monografia segue a estrutura a seguir: a seção dois expõe a motivação inicial para a pesquisa com os dados do Google e a possível aplicação ao comércio varejista brasileiro. Um compilado da produção acadêmica considerada mais relevante, com aplicação das mesmas ideias a outras economias, é disposto na seção três. A seção quatro trata dos dados utilizados, sua proveniência, histórico e forma de coleta. Seguidamente, estão dispostos na seção cinco o método empregado para avaliar os potenciais modelos econométricos com e sem a inclusão do Google Trends, assim como a decomposição da séries obtidas online. A seção seis clarifica os resultados obtidos para diferentes modelagens, previamente a uma conclusão na seção sete.

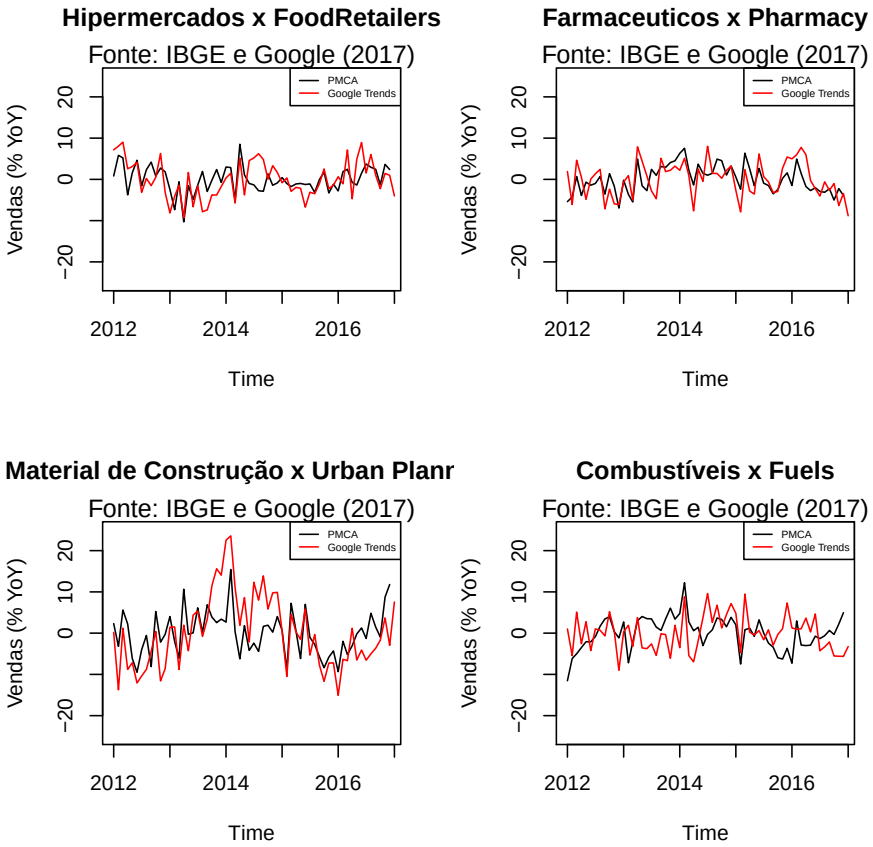
[2] Além do ICVA, a Cielo também comercializa um produto com base nos dados coletados sob a marca *Cielo Farol*. Ver: <http://www.releaseindicecielo.com.br/> e <https://www.cielo.com.br/venda-mais/cielofarol/>

2 Motivação

A motivação para este trabalho surgiu durante análise dos dados do Google Trends [3] para diversas categorias e diversos termos de pesquisa. Nesta análise, observei que algumas pesquisas seguiam padrões muito interessantes de tendência e sazonais, muito semelhantes a séries históricas de instituições públicas oficiais. Mais do que isso, algumas séries geradas pela máquina do Trends pareciam seguir padrões em linha com o contexto histórico, e.g. aumento das pesquisas por vagas de emprego *versus* o cenário interno de aceleração do desemprego.

Seguindo portanto a pesquisa de forma mais minunciosa, selecionei algumas categorias no Google Trends com a sua contraparte divulgada pelo governo brasileiro. A exemplo, pesquisas por supermercados na plataforma do Google (categoria "Grocery and Food Retailers") e seu par divulgado pela PMC, volume de vendas de "Hipermercados, supermercados, produtos alimentícios, bebidas e fumo". Outros exemplos incluem: "Vehicle Fuels" *versus* "Combustíveis e lubrificantes", "Pharmacy" *versus* "Artigos farmacêuticos, médicos, ortopédicos, de perfumaria e cosméticos" e "Urban and Regional Planning" *versus* "Materiais de Construção". Para todas estas séries, divulgadas em base mensal, transformei os dados em variação anual *year-over-year*, retirando o componente de tendência, a partir de uma regressão em uma variável de tempo, método detalhado na sessão de Metodologia a seguir. Além disso, em algumas séries, considerei um *lag* de um a seis meses para as séries provenientes do Google, considerando que é bastante natural que parte do comportamento dos consumidores é pesquisar antes de realizar qualquer compra. De outra forma, em alguns setores do comércio varejista, indivíduos comparam preços e pesquisam o bem a ser adquirido com alguma antecedência. Os resultados animadores deram ímpeto para o prosseguimento da pesquisa.

[3] Disponível em: `<https://trends.google.com/trends/>`


Hipermercados x FoodRetailers
Fonte: IBGE e Google (2017)
PMCA
Google Trends
Vendas (% YoY)
−20
0
10
20
2012
2014
2016
Time
Farmaceuticos x Pharmacy
Fonte: IBGE e Google (2017)
PMCA
Google Trends
Vendas (% YoY)
−20
0
10
20
2012
2014
2016
Time
Material de Construção x Urban Planr
Fonte: IBGE e Google (2017)
PMCA
Google Trends
Vendas (% YoY)
−20
0
10
20
2012
2014
2016
Time
Combustíveis x Fuels
Fonte: IBGE e Google (2017)
PMCA
Google Trends
Vendas (% YoY)
−20
0
10
20
2012
2014
2016
Time

3 Revisão Bibliográfica

O simples ato de realizar uma pesquisa no Google, checar as mídias sociais, ou ainda realizar uma pesquisa de preços online gera uma grande quantidade de dados, passíveis de análise e interpretação. Embora alguns autores, como Ginsberg (2009), já tenham utilizado o ferramental das buscas na internet na produção acadêmica, o ponto de partida para muitos trabalhos que utilizam um arcabouço econométrico para análise de dados de atividade virtual é comumente tido como Varian e Choi (2009). Em sua produção, destaca-se o potencial dos dados do Google e de outras empresas privadas para análise da atividade econômica em tempo real, uma vez que dados oficiais são geralmente divulgados com defasagem, e, não incomum, revisados *a posteriori*. Nele, os autores utilizam o Google Trends como "preditor do presente" para dados de vendas de automóveis, evolução do desemprego, confiança do consumidor e planejamento de viagens. Abre-se, assim, uma janela para futuras pesquisas de outras variáveis e em outras áreas de conhecimento e também para o *forecast* do futuro e uso de modelos mais sofisticados.

Já Bollen (2010), faz uma análise do mercado de ações baseado nos dados do Twitter: uma vez que o *price action* das ações é comumente guiado pelos sentimentos individuais dos agentes econômicos, o sentimento público mensurado na rede social se mostrou capaz de melhorar os critérios *in* e *out-of-sample* na predição do Dow Jones Industrial Average. Seguindo esta janela aberta, Artola e Galán (2012) foram capazes de construir um modelo mais bem específicado para o influxo de turistas britânicos para a economia espanhola utilizando dados de pesquisas no Google.

Regionalmente, outros passos foram dados à frente, principalmente para o *nowcasting* da atividade de países: McLaren (2011) aplica os dados de pesquisa à economia britânica enquanto Bughin (2011) explora a aplicabilidade dos dados do motor de pesquisas na Bélgica, com foco nas vendas do varejo e no desemprego, confirmando que as pesquisas conseguem explicar uma parte da variância das flutuações econômicas nas duas esferas. A autora ainda ressalta a relevância da internet como laboratório para testar e antecipar o comportamento dos indivíduos e que a abordagem do *nowcasting* deve se tornar material.

O trabalho de Labbé e Swallow (2010) também é comumente citado pela sua aplicação em economias emergentes. Os autores mostram que para o Chile, por mais que o acesso a internet não seja tão abrangente comparando-se a um país desenvolvido, modelos de previsão de compras de carros se adaptam melhor aos dados passados e têm eficiência aumentada, além de preverem bem *turning points*.

Quanto ao mercado de trabalho, alguns autores sondaram a prática, por parte de indivíduos, de pesquisar vagas de emprego online. Entre estes, Suhoy (2009) faz essa análise para Israel, além de análises em outras áreas como mercado imobilário, viagens, entre outros, destacando que o preditor mais forte foi para a área de recursos humanos (que engloba o recrutamento de candidatos). Baker e Fradkin (2011) montam um índice de

pesquisas por emprego na economia americana baseada nos dados do Trends, com foco nas flutuações nas pesquisas, com base nas mudanças nas políticas sociais voltadas a proteção do trabalho – a exemplo, o seguro desemprego. Os autores mostram que indivíduos tendem a procurar mais emprego quando seus benefícios são mais restritos, seja por exaustão dos mesmos ou por mudanças políticas. Outras abordagens para Itália e Alemanha são vistas em Francesco (2009) e Zimmermann e Askitas (2009), respectivamente. Além disso, Varian e Choi (2009,b) também abordam o assunto.

4 Dados

Serão duas as principais fontes de dados para a condução deste estudo. Primeiramente, os dados da Pesquisa Mensal do Comércio, divulgada mensalmente pelo IBGE. Para o *now-cast* serão utilizados os dados de volume de pesquisas do Google. Os dados da PMC utilizados são obtidos na plataforma do SIDRA - Sistema IBGE de Recuperação Automática [4]. A instituição oferece dois conceitos quando divulga os dados de atividade do varejo: a medida de núcleo e a medida ampliada. A diferença está na inclusão das categorias consideradas mais voláteis ("Materiais de Construção" e "Veículos Automotores") nesta segunda. No mais, os dados da PMC estão dispostos como "Volume de Vendas" e "Receita Nominal de Vendas". Para esta pesquisa, estamos interessados nos dados mensais do índice de volume de vendas considerando o conceito ampliado do comércio varejista ajustado sazonalmente.

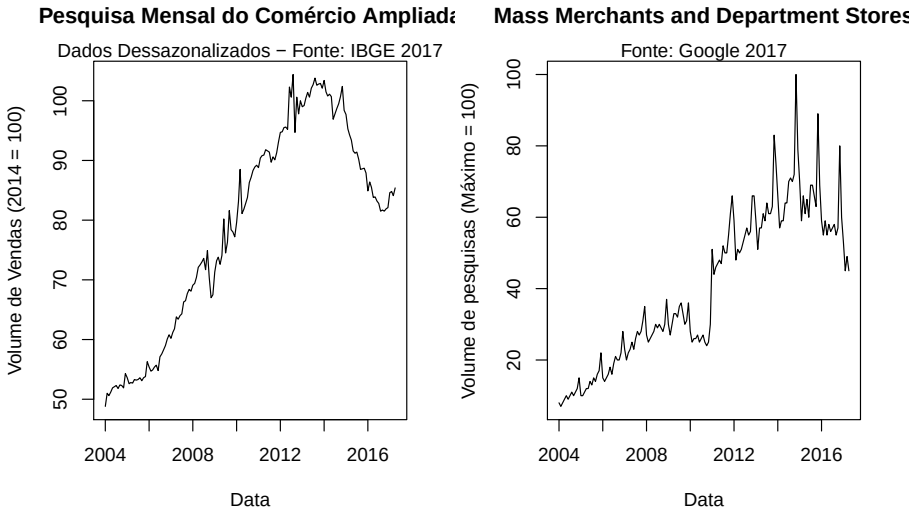
Embora o IBGE tenha iniciado a pesquisa do comércio em 1995, no ano 2000 uma revisão do Sistema de Índices do Comércio Varejista foi realizada levando a um ajuste nos parâmetros da pesquisa e expandindo o caráter da mesma para esfera nacional. O dado mais recente possui uma defasagem de cerca de seis semanas em relação ao tempo atual. A amostra utilizada no IBGE para a PMC utiliza os dados do Cadastro Nacional de Empresas (CEMPRE) e a unidade de investigação é a empresa individual, com foco na receita bruta de revenda. A partir deste conceito, o instituto elabora os indicadores supracitados de volume de vendas e receita nominal. Os grupos pesquisados pela PMC tem como base a Classificação Nacional de Atividades Econômicas (CNAE). A partir de 2004, uma nova revisão na pesquisa se consolidou de forma a ampliar a cobertura de um dos grupos pesquisados, criar indicadores para outros grupos e fazer-se a diferenciação entre os conceitos restrito e ampliado. Esta revisão de 2004 também permitiu que novos procedimentos fossem adotados para a PMC, visando a utilização dos novos adventos tecnológicos. Uma outra revisão da PMC foi realizada pelo IBGE em 2011 e, entre as mudanças, destaca-se a utilização da CNAE 2.0 para classificação dos grupos, além da desagregação do grupo "Móveis e Eletrodomésticos" em duas categorias independentes ("Móveis" e "Eletrodomésticos").

Já em relação aos dados do Google, o volume de pesquisas pode ser obtido na plataforma *online* do Google Trends. Embora haja dados em termos diários e semanais, acreditamos que os dados mensais são mais adequados ao que pesquisamos. O Google divulga uma série de categorias montadas pela máquina, agregando termos de várias pesquisas. Por exemplo, a categoria 45 (*"Health"*), agrupa os termos pesquisados ligados à área de saúde. Nos debruçamos sobre estas categorias, e iremos adiante escolher as que mais combinam com as categorias da PMC. O Google disponibiliza dados desde 2004 e

[4] Disponível em: <https://sidra.ibge.gov.br/>

em novembro de 2011 percebe-se que há uma quebra estrutural na maioria das pesquisas consultadas. Isso decorre, segundo uma nota divulgada no *site* do mecanismo, de uma melhoria na atribuição geográfica do mecanismo.

Desta forma, focaremos esta pesquisa no período de interseção entre a PMC e os dados do Google, isto é, utiliza-se os dados de 2004 até a última divulgação da PMC.



5 Metodologia

A análise será baseada na série da Pesquisa Mensal do Comércio, divulgada pelo IBGE, considerando o conceito amplo da pesquisa, focando no índice de volume de vendas. O IBGE divulga mensalmente a série já dessazonalizada, a qual será aqui investigada. A seguir, a metodologia observada é a mesma apresentada na metodologia Box-Jenkins (Box e Jenkins,1970), com estimação de modelos no formato $ARIMA(p,d,q)$, sendo d o número de raízes do processo, p um hiperparâmetro que identifica a defasagem máxima da variável Y_t (a parte auto-regressiva) e q a defasagem máxima do componente de distúrbio (a parte de médias móveis). A representação geral deste tipo de modelo pode ser dada por:

$$\phi(B)(1-B)^d Y_t = \delta + \theta(B)u_t \tag{1}$$

Sendo B um operador de defasagem (*backward shift*) e $\phi(B)$ e $\theta(B)$ são os polinômios que representam as defasagens p e q.

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - ... - \phi_p B^p \tag{2}$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - ... - \theta_q B^q \tag{3}$$

Em uma análise visual das séries, observa-se que as mesmas não parecem ser estacionárias, enquanto suas relativas primeiras diferenças sim. No mais, outros testes estatísticos devem ser performados para confirmar tal hipótese.

A partir dos gráficos das séries temporais, podemos perceber que talvez haja presença de raízes unitárias no processo gerador de dados. Assim, testes formais são realizados objetivando uma conclusão mais sólida e uma análise mais precisa. Utilizamos o teste Dickey-Fuller aumentado (ADF). Este teste é uma expansão do teste Dickey-Fuller mais básico, que se baseia numa especificação autoregressiva de ordem 1, para ordens mais elevadas, diga-se p, incluindo-se as defasagens de ΔY_t.

$$\Delta Y_t = \alpha_1 + \alpha_2 + \gamma Y_{t-1} + u_t \tag{4}$$

$$\begin{aligned} \Delta Y_t = \alpha_1 + \alpha_2 t + \gamma Y_{t-1} + \beta_1 \Delta Y_{t-1} + \\ \beta_2 \Delta Y_{t-2} + ... + u_t \end{aligned} \tag{5}$$

De todas as formas, as hipóteses testadas são:

$$\begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma \neq 0 \end{cases} \tag{6}$$

Isto é, testa-se a hipótese nula de que o processo contém uma raiz unitária contra a hipótese alternativa de que este é estacionário ou possui tendencia determinística. Caso o processo possua d raízes unitárias, ele pode ser diferenciado d vezes, transformado-se em um processo I(0). Desta forma, caso a hipótese nula não seja rejeitada, sabe-se que o processo possui ao menos uma raiz unitária e pode-se refazer o teste para a sua primeira diferença. Este processo é repetido até que a hipótese nula seja rejeitada e a série final seja estacionária.

Para todas as séries análisadas, o teste ADF teve a hipótese nula rejeitada considerando a primeira diferença e confirmando a suspeita citada anteriormente de que o processos teriam raiz unitária mas suas respectivas primeiras diferenças seriam estacionárias, conforme observa-se na Tabela 2.

Tabela 2: Teste Dickey-Fuller Aumentado (ADF)

Categoria	ADF	ADF na 1a diferença
PMC Ampliada	1.40	-9.78
Combustíveis e Lubrificantes	0.16	-8.34
Supermercados, Hipermercados, produtos alimentícios, bebidas e fumo	2.67	-9.60
Supermercados, Hipermercados	2.38	-10.69
Tecidos, vestuário e calçados	0.73	-12.31
Móveis e Eletrodomésticos	1.46	-8.66
Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos	5.09	-7.42
Livros, jornais, revistas e papelaria	0.37	-10.90
Equipamentos e materiais para escritório, Informática e de comunicação	0.84	-12.26
Outros artigos de uso pessoal e doméstico	2.36	-10.16
Veículos e motocicletas, partes e peças	0.03	-11.25
Material de construção	0.71	-13.16

5.1 Critérios *In-Sample*

Uma vez que sabemos que devemos utilizar a primeira diferença das séries analisadas, i.e. estamos tratando de processos $ARIMA(p,d,q)$ com $d = 1$, o próximo passo é postular qual o melhor modelo para os processos. Assim, devemos nos ater agora a encontrar os melhores valores p e q. Para realizar esta identificação, devemos analisar as funções de autocorrelação e autocorrelação parcial dos processos. A função geral para autocorrelações é dada por

$$\rho_k = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t) \times Var(Y_{t-k})}} \tag{7}$$

Podemos escolher os modelos baseados nos critérios de informação de Akaike e Schwartz, descritos a seguir. Os melhores modelos são aqueles que minimizam tais critérios. Para cada uma das séries, os modelos com valores mínimos de AIC e BIC estão dispostos na Tabela 3.

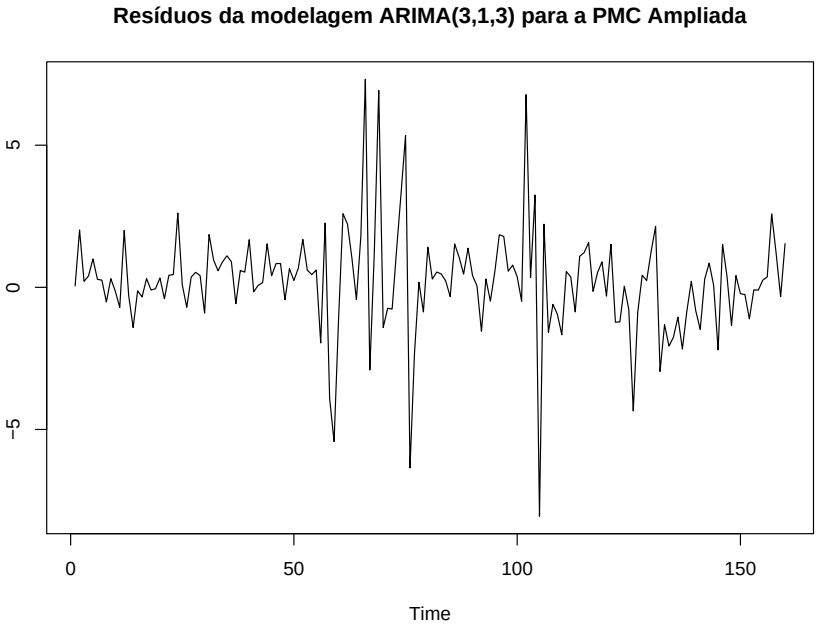
$$AIC = -2\ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + \frac{2(k+1)}{n} \tag{8}$$

$$BIC = -2\ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + \frac{\ln n(k+1)}{n} \tag{9}$$

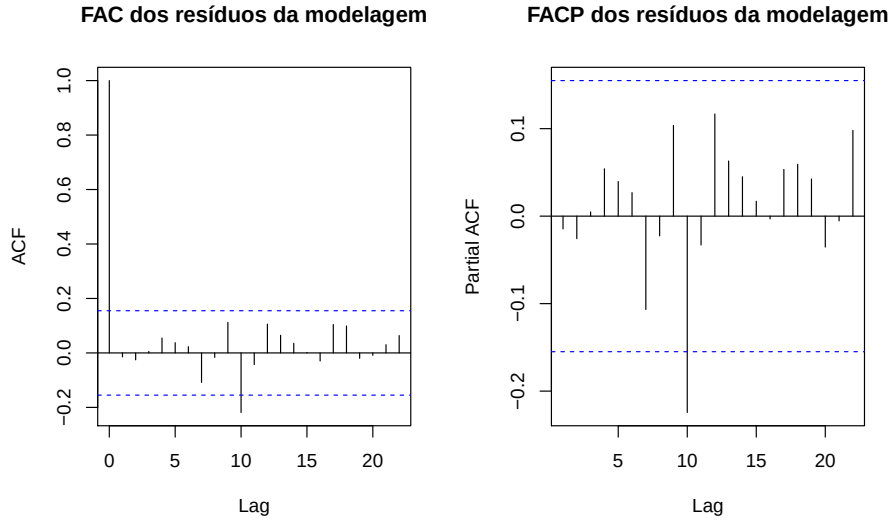
Tabela 3: Modelos que minimizam criteirios AIC/BIC

Categoria	Melhor modelo AIC	Melhor modelo BIC
PMC Ampliada	ARIMA(3,1,3)	ARIMA(1,1,0)
Combustíveis e Lubrificantes	ARIMA(0,1,0)	ARIMA(0,1,0)
Supermercados, Hipermercados, produtos alimentícios, bebidas e fumo	ARIMA(1,1,2)	ARIMA(1,1,2)
Supermercados, Hipermercados	ARIMA(1,1,2)	ARIMA(1,1,2)
Tecidos, vestuário e calçados	ARIMA(2,1,2)	ARIMA(0,1,1)
Móveis e Eletrodomésticos	ARIMA(1,1,2)	ARIMA(0,1,0)
Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos	ARIMA(1,1,2)	ARIMA(1,1,2)
Livros, jornais, revistas e papelaria	ARIMA(3,1,1)	ARIMA(1,1,0)
Equipamentos e materiais para escritório, Informática e de comunicação	ARIMA(3,1,3)	ARIMA(0,1,1)
Outros artigos de uso pessoal e doméstico	ARIMA(1,1,2)	ARIMA(1,1,2)
Veículos e motocicletas, partes e peças	ARIMA(3,1,2)	ARIMA(0,1,1)
Material de construção	ARIMA(3,2,3)	ARIMA(3,2,3)

Considerando os modelos identificados na Tabela 3, deve-se performar um diagnóstico da adequação dos mesmos. Busca-se examinar aqui a autocorrelação dos resíduos gerados, que, caso seja encontrada, significa que nem toda informação foi capturada pelo modelo. De outra forma, o resíduo deve possuir um padrão ruído-branco. A exemplo, pode-se observar o resíduo produzido pelo modelo da série da PMC que minimiza o critério de Akaike.



Além disso, plota-se também a função de aucorrelação dos resíduos, ratificando que a série gerada se assemelha a um ruído-branco.



Entretanto, um teste estatístico mais robusto deve ser performado para conclusões mais rigorosas e precisas. Dessa forma, o teste Ljung-Box é útil por testar a significância conjunta das autocorrelações residuais para certa defasagem, diga-se k. Este teste é descrito como *Portmanteau Lack-of-Fit Test* (Box, Jenkins & Reinsel, 2008). As auto-

correlações são tomadas conjuntamente, de modo a indicar uma possível inadequação do modelo. Isto é, caso os resíduos evoluam com certo padrão, o modelo selecionado não captura toda a informação da série analisada. A estatística analisada (*modified Ljung-Box-Pierce statistic*), é a seguinte:

$$\tilde{Q} = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}r_k^2 \tag{10}$$

Onde, para um processo $ARIMA(p,d,q)$, $\tilde{Q}$ é distribuido conforme uma χ^2 com $(K-p-q)$ graus de liberdade e $n = N-d$. Assim, testa-se a seguinte hipótese:

$$\begin{cases} H_0 : \text{Ausência de } \textit{lack-of-fit} \\ H_1 : \text{Presença de } \textit{lack-of-fit} \end{cases} \tag{11}$$

Rejeita-se a hipótese nula, H_0, caso $\tilde{Q} > \chi^2$. Da mesma forma, considerando-se o nível de significância de 5%, não se rejeita a hipótese nula para p-valores acima de 0.05. Os resultados estão dispostos na tabela 4, onde utilizou-se um *lag* (k = 10). Observa-se que não se rejeita a hipótese nula para qualquer uma das séries, isto é, pode-se afirmar que, considerando o nível de significância de 5%, não há autocorrelação entre os resíduos e o teste não rejeita a hipótese de ausência de *lack-of-fit*.

Tabela 4: Estatística do Teste Ljung-Box

Categoria da PMC	Teste LB no modelo AIC	Teste LB no modelo BIC
PMC Ampliada	0.2	0.11
Combustíveis e Lubrificantes	0.23	0.23
Supermercados, Hipermercados, produtos alimentícios, bebidas e fumo	0.83	0.83
Supermercados, Hipermercados	0.85	0.85
Tecidos, vestuário e calçados	0.98	0.48
Móveis e Eletrodomésticos	0.79	0.02
Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos	0.89	0.89
Livros, jornais, revistas e papelaria	0.96	0.42
Equipamentos e materiais para escritório, Informática e de comunicação	0.84	0.34
Outros artigos de uso pessoal e doméstico	0.12	0.12
Veículos e motocicletas, partes e peças	0.88	0.8
Material de construção	0.74	0.74

5.2 Critérios *Out-of-Sample*

A partir dos três melhores modelos selecinados por cada modelo AIC e BIC, realiza-se agora o *out-of-sample*, utilizando uma janela de $\frac{3}{4}$ da série para previsão dos próximos $\frac{1}{4}$ da mesma, já realizados. Os melhores modelos serão aqueles capazes de minimizar o erro quadrático médio (MSE) e o erro médio absoluto (MAE). Além disso, faz-se uma comparação direta com a capacidade de previsão do modelo mais simples, um $AR(1)$, isto é, utiliza-se o $AR(1)$ como *benchmark*. Parte-se dos critérios *in-sample* para os critérios *out-of-sample* na necessidade de utilizar medidas que englobem tanto a média quanto a variância do erro de previsão $\hat{e}$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_{n+h,n}^2 \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid e_{n+h,n} \mid \tag{13}$$

$$\hat{e}_{n+h,n} = y_{n+h} - \hat{y}_{n+h,n} \tag{14}$$

Uma vez que não se conhece os valores futuros de y_i, para comparação com os valores previstos $\hat{y}_i$, utiliza-se para os critérios *out-of-sample* o método *Fixed Window*. Neste, um período fixo para cálculo do erro de previsão é utilizado dentro da janela descrita acima ($\frac{1}{4}$ da série): toda a informação disponível até n é utilizada para previsão de $n+h$. Desta forma, para cada observação já conhecida em $n+h$ (y_{n+h}), há a respectiva previsão $\hat{y}_{n+h}$, sendo possível assim observar o erro de previsão.

Assim, constrói-se para cada variável uma tabela, conforme exemplificado abaixo nas tabelas 5 e 6, onde se observa o AR(1), os seis melhores modelos descritos acima (3 melhores AIC e 3 melhores BIC), e seus respectivos MAE e MSE. A comparação é feita a partir da razão entre os erros dos seis modelos e o erros relativos ao AR(1). Assim, observou-se que, para as doze séries analisadas, o AR(1) pode ser um modelo melhor em termos de minimização do MAE e do MSE em relação aos modelos extraídos na subseção anterior. Testaremos a frente como os erros se modificam com a inclusão dos dados do *Google Trends*, objetivando encontrar os modelos que minimizem o MSE e MAE.

Tabela 5: Comparação de modelos: PMC Ampliada

Modelo	MSE	MSE/benchmark	MAE	MAE/benchmark
AR(1)	2.45	1	1.15	1
ARIMA(0,1,0)	2.28	0.93	1.11	0.96
ARIMA(2,1,2)	2.35	0.96	1.11	0.96
ARIMA(1,1,0)	2.36	0.96	1.12	0.97
ARIMA(0,1,0)	2.28	0.93	1.11	0.96
ARIMA(1,1,0)	2.36	0.96	1.12	0.97
ARIMA(0,1,1)	2.36	0.96	1.12	0.97

Tabela 6: Comparação de modelos: Supermercados e Hipermercados

Modelo	MSE	MSE/benchmark	MAE	MAE/benchmark
AR(1)	7.95	1	1.89	1
ARIMA(2,1,2)	9.54	1.20	2.19	1.16
ARIMA(0,1,3)	9.07	1.14	2.15	1.13
ARIMA(2,1,0)	9.00	1.13	2.16	1.14
ARIMA(0,1,1)	9.05	1.14	2.13	1.13
ARIMA(2,1,0)	9.00	1.13	2.16	1.14
ARIMA(0,1,3)	9.07	1.14	2.15	1.13

5.3 Inclusão do *Google Trends*

Os dados do Google estão dispostos *online* em termos mensais e são de simples acesso pela sua plataforma *Trends*. Para cada categoria da PMC Ampliada, busca-se a categoria do Google Trends que mais se aproximaria. Utiliza-se também termos de pesquisa considerados próximos das devidas categorias da PMC. Por fim, pode-se ainda realizar uma combinação de termos de pesquisa dentro da categoria do Google Trends. Na Tabela 7 está disposto o dado do Google Trends utilizado para cada categoria englobada pela pesquisa do IBGE.

Tabela 7: Relação entre Grupos da PMC e pesquisas no Google

Grupo da PMC	Categoria/Termo de Pesquisa no Trends
PMC Ampliada	Mass Merchants and Department Stores
Combustíveis e Lubrificantes	Vehicle Fuels and Lubricants
Supermercados, Hipermercados, produtos alimentícios, bebidas e fumo	Search query: 'encarte'
Supermercados, Hipermercados	Search query: 'encarte'
Tecidos, vestuário e calçados	Search query: 'Marisa S.A.'
Móveis e Eletrodomésticos	Home and Garden
Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos	Pharmarcy
Livros, jornais, revistas e papelaria	Books and Literature
Equipamentos e materiais para escritório, Informática e de comunicação	Office Supplies
Outros artigos de uso pessoal e doméstico	Online shopping
Veículos e motocicletas, partes e peças	Car
Material de construção	Construction and Maintenance

Pode-se, assim, adicionar estas categorias nos modelos ARIMA, de forma a minimizar os erros de previsão, encontrando-se o melhor modelo. Ao utilizar destes regressores, admitimos que os modelos escolhidos pelos critérios *in-* e *out-of-sample* poderão diferir dos originais vistos acima uma vez que o objetivo principal é encontrar modelos que possuam critérios MSE e MAE menores do que os originais, o que resultaria em regressões mais precisas para o *nowcasting*.

Observou-se ainda que a maioria das séries obtidas do Google apresentava uma quebra estrutural no mês de novembro de 2011. Tal ponto, antes não esclarecido, foi posteriormente clarificado com a inclusão pela empresa de uma nota explicativa, onde se eviden-

ciou uma melhoria na atribuição geográfica do motor de buscas a partir de novembro de 2011. Para lidar com tal quebra estrutural nos dados, incluiu-se uma variável *dummy* na regressão. Tal variável recebeu valor 1 (um) caso o dado individual fosse relativo a data e 0 (zero) caso contrário. Assim, chegamos a seguinte especificação da modelagem ARIMA:

$$\phi(B)(1-B)^d Y_t = \delta + Trends_{t-i} + \xi D + \theta(B)u_t \tag{15}$$

$$D = \begin{cases} 1 & \text{se } t = novembro/2011 \\ 0 & \text{c.c.} \end{cases} \tag{16}$$

E da mesma forma anterior, B é o operador de defasagem (*backward shift*) e $\phi(B)$ e $\theta(B)$ são os polinômios que representam as defasagens p e q. O regressor do Trends é defasado para valores de $i \in [0,6]$, isto é: é usado de forma contemporânea ou com até seis meses de defasagem. Numa segunda especificação, utiliza-se a séries da Pesquisa Mensal do Comércio e os dados do Google a partir de 2012. Neste caso, a especificação da regressão de reduz a:

$$\phi(B)(1-B)^d Y_t = \delta + Trends_{t-i} + \theta(B)u_t \tag{17}$$

5.4 Decomposição das Séries do Google Trends

Um representação tradicional de uma série temporal tal qual o Google Trends é a partir da sua decomposição em elementos de tendência e sazonalidade, além de um componente de ruído (Box, Jenkins & Reinsel, 2008). Ignorar um componente tendencial pode levar a uma conclusão errônea acerca dos efeitos de uma série sobre outra, se ambas crescem na mesma direção ou em direção oposta ao longo do tempo (Wooldridge, 2010). Em termos da sazonalidade, a decomposição por um método aditivo é mais apropriada em casos nos quais a amplitude da mesma não varia com o nível da série. Em outros casos, uma decomposição multiplicativa é mais pertinente (Montgomery, Jennings & Kulahci, 2007).

$$Y_t = f(T_t, S_t, N_t) \tag{18}$$

Observou-se que algumas das séries teriam um componente tendencial determinístico. É necessário considerar esta tendência temporal, evitando-se a indução à uma conclusão falsa de que as alterações ocorridas em uma variável guardariam relação com as alterações em uma outra variável, quando ambos os processos somente estão correlacionados por evoluirem temporalmente de forma parecida por fatores não observados, acarretando

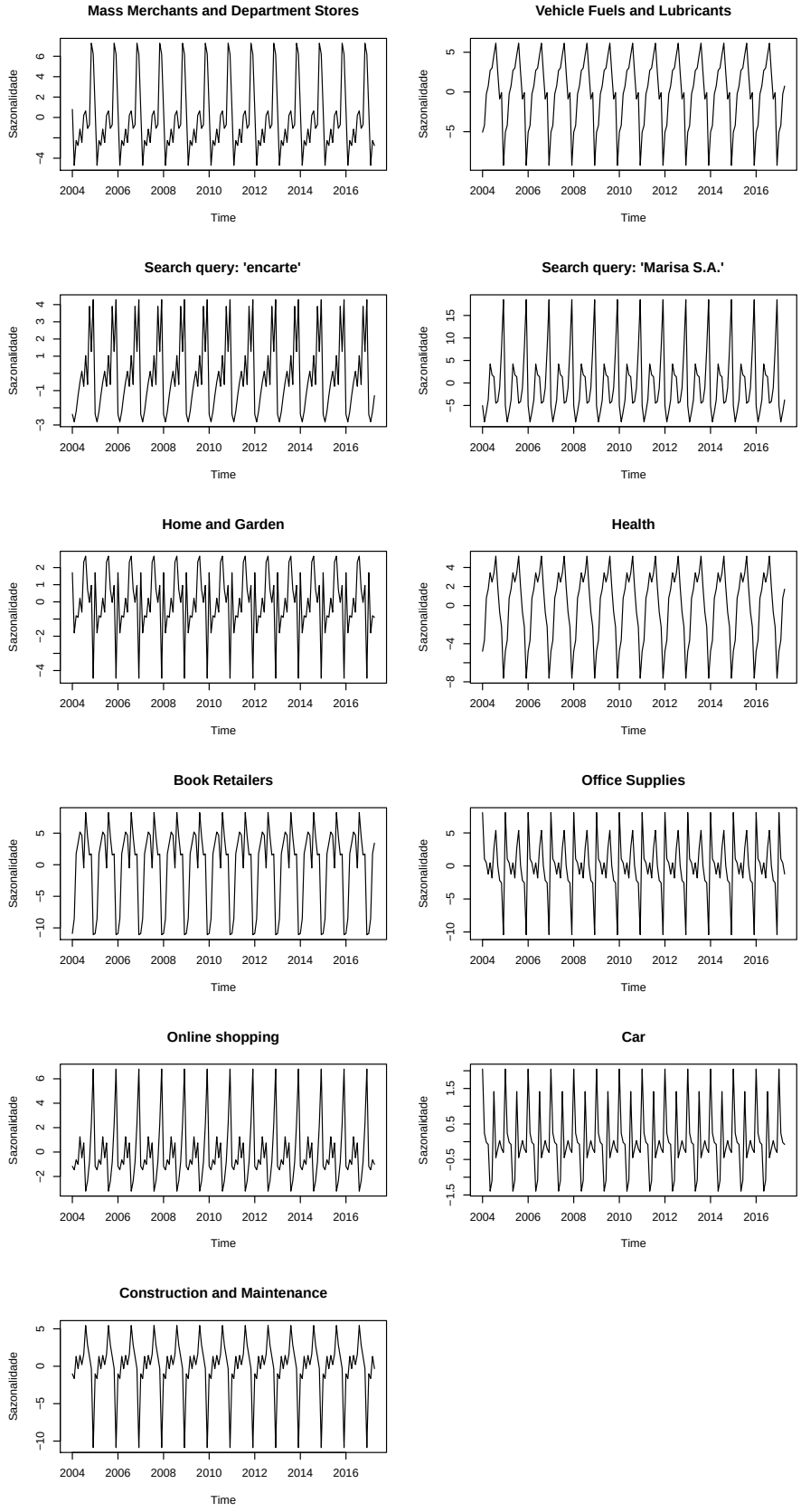
em um problema de regressão espúria. Argumenta-se que a tendência temporal nas séries obtidas do Google muito provavelmente seja proveniente do alargamento do acesso à internet no Brasil, e tem por consequência o aumento das pesquisas *online*. Estas séries podem ser caracterizadas, portanto, como não-estacionárias, e o tratamento deste componente de tendência pode-se dar a partir de um modelo de regressão simples que descreve o componente tendencial (ou seja, a evolução com o tempo), para depois removê-lo. Esta regressão pode ser disposta por uma função quadrática ou exponencial do tempo, mas aqui verifico que uma tendência linear é capaz de expurgar este elemento. Assim, o valor esperado da série deve se alterar de forma linear ao longo do tempo (Montgomery, Jennings & Kulahci, 2007):

$$E(y_t) = \alpha_0 + \alpha_1 t \tag{19}$$

Nos casos analisados, observa-se que o coeficiente de inclinação é positivo para a maioria das séries, i.e. $\alpha_1 > 0$, indicando tendência ascendente – em linha com a argumentação anterior. Para uma melhor análise, pode-se estimar os parâmetros do modelo acima pelo método de Mínimos Quadrados Ordinários, a fim de retirar-se o componente de tendência da série original.

Outro componente observado nas séries obtidas a partir da plataforma do Google é a sazonalidade. Uma série possui um componente sazonal quando apresenta um comportamento periódico em um período s e similaridades ocorrem sempre após s intervalos de tempo (Box, Jenkins & Reinsel, 2008).

A partir da utilização do pacote `fpp` disponível para o `R`, pode-se realizar a decomposição das séries, a fim de observar-se o componente sazonal. Para os fins utilizados neste trabalho, o método utilizado para expurgar a sazonalidade foi o multiplicativo, em linha com o método de decomposição utilizado pelo IBGE para dessazonalizar as séries da Pesquisa Mensal do Comércio (Instituto Brasileiro de Geografia e Estatística, 2017).


Mass Merchants and Department Stores
Vehicle Fuels and Lubricants
Search query: 'encarte'
Search query: 'Marisa S.A.'
Home and Garden
Health
Book Retailers
Office Supplies
Online shopping
Car
Construction and Maintenance
Sazonalidade
Time
2004
2006
2008
2010
2012
2014
2016

6 Resultados

6.1 *Google Trends* como regressor

Os resultados da inclusão dos dados do Google Trends nos modelos, conforme especificação anterior, são, em primeira análise, não muito promissores, uma vez que níveis de MAE e MSE são na maior parte dos casos maiores em relação a modelagem inicial. De uma outra forma, os erros quadrático médio e médio absoluto são maiores quando incluídos como regressores contemporâneos na maioria dos modelos. Os resultados estão dispostos nas tabelas 8 e 9. Alguns resultados positivos são vistos para algumas séries como "Tecidos, vestuário e calçados" e para a própria PMC Ampliada, onde o MSE cai notáveis 31% no primeiro caso e 4% no segundo. Ainda, a série relacionada ao varejo de materiais de construção apresenta uma melhora de apenas 2% e "Combustíveis e Lubrificantes", 1%. Considerando o erro absoluto médio, o progresso maior obtido foi novamente para a série relacionada a Vestuário, enquanto outros resultados diminutos são vistos para "Combustíveis e Lubrificantes", "Veículos e motocicletas, partes e peças" e "Material de Construção".

Tabela 8: Comparação do MSE dos melhores modelos

Categorias	Modelo sem Trends	MSE sem Trends (1)	Modelo com Trends	MSE com Trends (2)	Razão (2)/(1)
PMC Ampliada	ARIMA(1,1,0)	2.43	ARIMA(1,1,0)	2.33	0.96
Combustíveis	ARIMA(0,1,0)	2.28	ARIMA(0,1,0)	2.26	0.99
Mercados e Outros	ARIMA(2,1,1)	2.70	ARIMA(2,1,2)	3.12	1.15
Hipermercados	ARIMA(0,1,1)	3.44	ARIMA(0,1,1)	3.61	1.05
Vestuário	ARIMA(2,1,0)	9	ARIMA(2,1,0)	6.21	0.69
Moveis e Eletrodomésticos	ARIMA(0,1,0)	8.05	ARIMA(0,1,0)	8.20	1.02
Farmaceuticos	ARIMA(2,1,1)	1.58	ARIMA(3,1,2)	1.70	1.08
Livros	ARIMA(0,1,2)	7.98	ARIMA(0,1,2)	8.70	1.09
Materiais de Escritório	ARIMA(0,1,1)	27.63	ARIMA(0,1,1)	27.75	1.00
Outros Artigos de Uso Pessoal	ARIMA(1,1,3)	5.03	ARIMA(2,1,2)	5.84	1.16
Veículos	ARIMA(3,1,3)	12.37	ARIMA(1,1,0)	12.99	1.05
Materiais de Construção	ARIMA(3,1,0)	8.33	ARIMA(2,1,3)	8.15	0.98

Tabela 9: Comparação do MAE dos melhores modelos

Categorias	Modelo sem Trends	MAE sem Trends (1)	Modelo com Trends	MAE com Trends (2)	Razão (2)/(1)
PMC Ampliada	ARIMA(1,1,0)	1.22	ARIMA(1,1,0)	1.22	1.00
Combustíveis	ARIMA(2,1,2)	1.11	ARIMA(2,1,3)	1.08	0.97
Mercados e Outros	ARIMA(2,1,2)	1.03	ARIMA(2,1,2)	1.07	1.04
Hipermercados	ARIMA(0,1,1)	1.09	ARIMA(0,1,1)	1.12	1.03
Vestuário	ARIMA(0,1,1)	2.13	ARIMA(2,1,0)	1.84	0.86
Moveis e Eletrodomésticos	ARIMA(0,1,0)	2.14	ARIMA(0,1,0)	2.19	1.02
Farmaceuticos	ARIMA(2,1,1)	1.01	ARIMA(3,1,2)	1.07	1.06
Livros	ARIMA(0,1,2)	2.16	ARIMA(0,1,2)	2.27	1.05
Materiais de Escritório	ARIMA(0,1,1)	3.97	ARIMA(0,1,1)	3.95	1.00
Outros Artigos de Uso Pessoal	ARIMA(1,1,3)	1.50	ARIMA(1,1,3)	1.75	1.17
Veículos	ARIMA(1,1,0)	2.68	ARIMA(1,1,0)	2.61	0.98
Materiais de Construção	ARIMA(3,1,0)	2.24	ARIMA(0,1,1)	2.14	0.96

É necessário ainda observar o padrão de evolução dos resíduos dos modelos propostos acima com a inclusão do Google Trends, garantindo que não haja autocorrelação residual. Performa-se o teste de Ljung-Box já descrito na seção anterior. Os resultados estão dispostos na tabela 10 e confirmam que a hipótese nula do teste – de que não há ausência de *fit* dos modelos – não deve ser rejeitada para a maior parte dos casos, considerando-se um número de defasagens de 1 até 12. A tabela dispõe apenas alguns dos níveis de defasagem, para melhor disposição gráfica. Para sua melhor interpretação, poder-se-ia rejeitar a hipótese nula, em prol da hipótese alternativa de que o modelo não apresenta bom *fit*, para p-valores menores de 0.05, considerando um nível de significância de 5%. Para a série da PMC Ampliada, no entanto, já há rejeição da hipótese nula a partir da décima defasagem. O mesmo ocorre para "Móveis e Eletrodomésticos" a partir da oitava defasagem.

Tabela 10: P-Valores do Teste de Ljung-Box para 12 defasagens

Série	k = 2	k = 6	k = 8	k = 10	k = 12
PMC Ampliada	0.92	0.39	0.44	0.03	0.03
Combustíveis e Lubrificantes	0.46	0.48	0.17	0.24	0.30
Supermercados, Hipermercados, produtos alimentícios, bebidas e fumo	0.95	0.75	0.85	0.84	0.88
Supermercados, Hipermercados	0.49	0.33	0.40	0.54	0.63
Tecidos, vestuário e calçados	1.00	0.77	0.90	0.86	0.89
Móveis e Eletrodomésticos	0.05	0.06	0.02	0.03	0.01
Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos	1.00	0.84	0.81	0.92	0.92
Livros, jornais, revistas e papelaria	0.99	0.92	0.81	0.90	0.45
Equipamentos e materiais para escritório, Informática e de comunicação	0.61	0.16	0.13	0.25	0.26
Outros artigos de uso pessoal e doméstico	0.78	0.96	0.61	0.16	0.25
Veículos e motocicletas, partes e peças	0.29	0.43	0.63	0.59	0.38
Material de construção	0.80	0.86	0.85	0.68	0.81

Os dados do Google Trends parecem ter algum potencial na capacidade de predição dos dados das vendas no varejo no Brasil, embora na análise contemporânea a melhoria seja, na maior parte dos casos, nula ou negativa. Faz-se necessário, no entanto, análisar a questão considerando que consumidores podem realizar pesquisas não somente de forma contemporânea às suas compras, mas também com alguma defasagem, realizando pesquisas e comparação *online* meses antes de adquirir um produto ou serviço. Explora-se este comportamento na subseção abaixo.

6.2 Análise dos resultados com *lags*

Um consumidor pode utilizar o Google certo tempo antes de efetivar uma transação envolvendo um bem ou serviço – é natural que haja, para certas categorias de consumo, alguma pesquisa anterior à compra. Nesta seção analisa-se se o quanto a capacidade preditiva dos modelos modifica-se quando os regressores utilizados são os dados das pesquisas online com defasagens. A intuição para tanto é bastante simples: um consumidor por vezes realiza uma busca online, comparando preços ou ainda comparando bens ou serviços, para somente após este processo adquirir o bem ou serviço. Este tipo de comportamento do consumidor é evidente dado os diversos sites de comparação de preços e de *reviews* de produtos.

Utiliza-se os dados do Google para seis níveis de defasagens. Cada defasagem significa um mês de diferença em relação à série original, de forma que o mesmo dado foi defasado, portanto, de um até seis meses. Um consumidor estaria buscando pelo produto ou serviço com antecedência de meses igual ao número de desafasagens. A exemplo, a defasagem quatro ($lag = 4$) significa que o consumidor performa uma pesquisa quatro meses antes de realizar uma compra.

Os resultados estão dispostos nas tabelas 11 e 12, onde se observa o erro quadrático médio (MSE) e o erro médio absoluto (MAE) do melhor modelo escolhido na defasagem da coluna respectiva em relação ao erro (MSE ou MAE) do melhor modelo sem defasagens ($lag = 0$). Observa-se que, na maioria dos casos, há uma assertiva melhoria na capacidade preditiva dos modelos ao considerar-se tal efeito. A redução do MSE chega a 34% para a série de “Livros, jornais, revistas e papelaria”, variando entre 10% a 15% para “Veículos e motocicletas, partes e peças”, “Outros artigos de uso pessoal e doméstico”,“Material de construção”,“Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos”,“Supermercados, Hipermercados, produtos alimentícios, bebidas e fumo” e “Combustíveis e Lubrificantes”.O uso de dados passados apresentou uma melhoria de até 5% para as demais séries, sendo que em apenas uma (“Tecidos, vestuário e calçados”) houve um aumento do erro para qualquer defasagem utilizada. Na análise do MAE, os *lags* levam a uma atenuação deste erro entre 5% e 9% para metade das séries. A inclusão de dados defasados foi novamente melhor para “Livros, jornais, revistas

e papelaria", enquanto outras cinco séries os ganhos foram nulos ou até 5%.

Tabela 11: MSE para diferentes defasagens (razão com lag 0)

Categoria da PMC	lag=0	lag=1	lag=2	lag=3	lag=4	lag=5	lag=6
PMC Ampliada	1	1.03	0.93	1.15	1.30	1.06	1.03
Combustíveis	1	0.96	1.15	0.96	1.12	1.03	0.90
Mercados e Outros	1	0.85	0.88	0.93	0.91	0.88	0.89
Hipermercados	1	0.99	0.96	1.03	1.06	0.97	0.97
Vestuário	1	1.64	1.39	1.22	1.39	1.61	1.46
Moveis e Eletrodomésticos	1	0.98	0.97	1.03	0.97	1.09	1.05
Farmaceuticos	1	0.93	0.88	1.01	0.94	0.89	0.88
Livros	1	1.39	1.65	1.00	0.66	0.76	0.79
Materiais de Escritório	1	1.00	0.99	1.01	0.95	1.00	1.08
Outros Artigos de Uso Pessoal	1	0.85	0.95	1.04	1.20	0.90	0.91
Veículos	1	0.85	1.65	0.92	0.91	0.91	1.01
Materiais de Construção	1	0.99	1.08	0.85	0.98	1.03	1.10

Tabela 12: MAE para diferentes defasagens (razão com lag 0)

Categoria da PMC	lag=0	lag=1	lag=2	lag=3	lag=4	lag=5	lag=6
PMC Ampliada	1	1.00	0.95	1.05	1.11	1.00	0.98
Combustíveis	1	0.98	1.17	0.98	1.07	1.01	0.91
Mercados e Outros	1	0.91	1.00	1.01	0.99	1.02	1.02
Hipermercados	1	1.02	0.97	1.06	1.08	1.03	1.01
Vestuário	1	1.26	1.16	1.02	1.10	1.19	1.11
Moveis e Eletrodomésticos	1	0.98	0.98	1.01	0.96	1.07	1.02
Farmaceuticos	1	0.93	0.93	0.98	0.96	0.95	0.91
Livros	1	1.25	1.39	1.05	0.77	0.86	0.89
Materiais de Escritório	1	1.00	1.00	1.03	1.01	1.03	1.06
Outros Artigos de Uso Pessoal	1	0.91	0.95	0.95	1.08	0.92	0.93
Veículos	1	0.98	1.29	0.96	1.07	0.96	0.97
Materiais de Construção	1	1.03	1.10	0.95	1.05	1.05	1.09

Partindo destes resultados, pode-se reconstruir as tabelas elaboradas na subseção anterior, comparando os melhores modelos (aqueles que minimizam os erros) com e sem a inclusão do Google Trends, considerando agora que estes primeiros podem apresentar *lags*. Os resultados estão dispostos nas Tabelas 13 e 14.

Tabela 13: Comparação do MSE entre melhores modelos considerando defasagens

Categorias	Modelo sem Trends	MSE sem Trends (1)	Modelo com Trends	Lag	MSE com Trends (2)	Razão (2)/(1)
PMC Ampliada	ARIMA(1,1,0)	2.43	ARIMA(3,1,3)	2	2.17	0.90
Combustíveis	ARIMA(0,1,0)	2.28	ARIMA(0,1,0)	6	2.03	0.89
Mercados e Outros	ARIMA(2,1,1)	2.70	ARIMA(1,1,0)	1	2.64	0.97
Hipermercados	ARIMA(0,1,1)	3.44	ARIMA(0,1,1)	2	3.47	1.01
Vestuário	ARIMA(2,1,0)	9	ARIMA(2,1,0)	0	6.21	0.69
Moveis e Eletrodomésticos	ARIMA(0,1,0)	8.05	ARIMA(0,1,0)	2	7.98	0.99
Farmaceuticos	ARIMA(2,1,1)	1.58	ARIMA(1,1,2)	2	1.50	0.95
Livros	ARIMA(0,1,2)	7.98	ARIMA(3,1,0)	4	5.71	0.71
Materiais de Escritório	ARIMA(0,1,1)	27.63	ARIMA(2,1,0)	4	26.33	0.95
Outros Artigos de Uso Pessoal	ARIMA(1,1,3)	5.03	ARIMA(3,1,1)	1	4.96	0.99
Veículos	ARIMA(3,1,3)	12.37	ARIMA(0,1,1)	1	11.02	0.89
Materiais de Construção	ARIMA(3,1,0)	8.33	ARIMA(3,1,0)	3	6.90	0.83

Tabela 14: Comparação do MAE entre melhores modelos considerando defasagens

Categorias	Modelo sem Trends	MAE sem Trends (1)	Modelo com Trends	Lag	MAE com Trends (2)	Razão (2)/(1)
PMC Ampliada	ARIMA(1,1,0)	1.22	ARIMA(1,1,0)	2	1.16	0.95
Combustíveis	ARIMA(2,1,2)	1.11	ARIMA(0,1,0)	6	0.98	0.88
Mercados e Outros	ARIMA(2,1,2)	1.03	ARIMA(1,1,0)	1	0.98	0.95
Hipermercados	ARIMA(0,1,1)	1.09	ARIMA(0,1,1)	2	1.08	1.00
Vestuário	ARIMA(0,1,1)	2.13	ARIMA(2,1,0)	0	1.84	0.86
Moveis e Eletrodomésticos	ARIMA(0,1,0)	2.14	ARIMA(0,1,0)	4	2.10	0.98
Farmaceuticos	ARIMA(2,1,1)	1.01	ARIMA(2,1,1)	6	0.98	0.97
Livros	ARIMA(0,1,2)	2.16	ARIMA(3,1,0)	4	1.74	0.81
Materiais de Escritório	ARIMA(0,1,1)	3.97	ARIMA(0,1,1)	1	3.93	0.99
Outros Artigos de Uso Pessoal	ARIMA(1,1,3)	1.50	ARIMA(1,1,3)	1	1.60	1.07
Veículos	ARIMA(1,1,0)	2.68	ARIMA(0,1,1)	3	2.50	0.93
Materiais de Construção	ARIMA(3,1,0)	2.24	ARIMA(2,1,2)	3	2.03	0.91

A modelagem com defasagens leva a resultados positivos em termos de queda dos erros quadrático e absoluto médio para todas as séries análisadas, exceto uma – a ligada a venda de roupas, que foi a que melhor desempenhou. Os resultados mais incisivos são para a categoria que analisa a volume de vendas de livros e relacionados, cujo erro quadrático médio contraiu-se quase 30% em um modelo que inclui pesquisas na categoria "*Books and Literature*" no Google com uma diferença de 4 meses. A pesquisa contemporânea (i.e., *lag = 0*) pela maior varejista em moda feminina gera um queda do MSE da categoria relacionada a vestuário de 31%.

Houve queda do MSE em 17% para as séries de "Material de Construção", cerca de 10% para a PMC Ampliada, para a série de combustíveis e relacionados e veículos. Na faixa de 5% estão "Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos" e "Equipamentos e materiais para escritório, Informática e de comunicação". O MSE contraiu-se menos de 5% nas séries de "Móveis e Eletrodomésticos" e naquelas relacionadas às compras em supermercados, onde uma de suas medidas teve um resultado negativo. Em termos do MAE, a razão entre o erro absoluto dos modelos com e sem a inclusão do Google Trends variou entre 0.81 ("Livros, jornais, revistas e papelaria") e 0.98 ("Móveis e Eletrodomésticos"). O modelo com regressor do Google Trends foi pior em termos do erro absoluto médio para "Outros artigos de uso pessoal e doméstico". Argumenta-se que este item abrange oficialmente uma gama grande de atividades difíceis de serem localizadas em um único termo de pesquisa ou categoria. Estão sob a mesma classificação as atividades da CNAE ligadas a "Artigos de Joalheria", "Frutos Ornamentais Naturais","Produtos de Limpeza para Óculos", "Peças para Geladeiras", entre outros.

6.3 Investigação a partir de 2012

Com a melhoria na atribuição geográfica do sistema do Trends em novembro de 2011, os dados a partir de então apresentam uma dinâmica mais próxima da realidade uma vez que capturam melhor o volume de pesquisas no país. Assim, podemos observar os efeitos da inclusão deste regressor na modelagem das vendas no varejo a partir de 2012 até os dados mais recentes.

A metodologia é semelhante a da subseção imediatamente anterior. Compara-se os modelos que minimizam os erros quadrático médio e médio absoluto com e sem a inclusão dos dados do Google Trends como regressor. Permite-se também que os dados do regressor sejam adicionados com defasagens (*"lags"*) mensais para explicar a variável dependente de forma contemporânea.

Na exposição de resultados a seguir, dois fatos são interessantes a se notar: (i) mesmo sem a inclusão do Google Trends, os erros dos modelos foram majoritariamente menores do que utilizando os dados desde 2004 e; (ii) a capacidade de predição alcança níveis

ainda maiores, reduzindo os erros de forma mais drástica – em alguns casos mais de 40%.

Tabela 15: MSE dos melhores modelos considerando defasagens - Séries a partir de 2012

Categorias	Modelo sem Trends	MSE sem Trends (1)	Modelo com Trends	Lag	MSE com Trends (2)	Razão (2)/(1)
PMC Ampliada	ARIMA(2,1,0)	1.24	ARIMA(1,1,2)	4	0.90	0.73
Combustíveis	ARIMA(1,1,0)	0.76	ARIMA(0,1,1)	6	0.64	0.85
Mercados e Outros	ARIMA(0,1,1)	4.70	ARIMA(0,1,1)	5	4.27	0.91
Hipermercados	ARIMA(0,1,1)	6.65	ARIMA(0,1,1)	5	6.19	0.93
Vestuário	ARIMA(0,1,0)	13.22	ARIMA(1,1,2)	0	8.76	0.66
Moveis e Eletrodomésticos	ARIMA(1,1,0)	5.05	ARIMA(0,1,0)	4	2.91	0.58
Farmaceuticos	ARIMA(0,1,0)	1.84	ARIMA(0,1,0)	1	1.87	1.01
Livros	ARIMA(0,1,3)	3.76	ARIMA(0,1,2)	6	2.15	0.57
Materiais de Escritório	ARIMA(0,1,1)	16.48	ARIMA(0,1,1)	1	16.86	1.02
Outros Artigos de Uso Pessoal	ARIMA(1,1,0)	2.81	ARIMA(0,1,1)	6	2.55	0.91
Veículos	ARIMA(1,1,0)	3.64	ARIMA(3,1,3)	4	2.11	0.58
Materiais de Construção	ARIMA(0,1,1)	6.05	ARIMA(0,1,1)	3	5.07	0.84

Tabela 16: MAE dos melhores modelos considerando defasagens - Séries a partir de 2012

Categorias	Modelo sem Trends	MAE sem Trends (1)	Modelo com Trends	Lag	MAE com Trends (2)	Razão (2)/(1)
PMC Ampliada	ARIMA(1,1,1)	0.84	ARIMA(1,1,2)	4	0.75	0.89
Combustíveis	ARIMA(1,1,0)	0.69	ARIMA(0,1,1)	6	0.57	0.83
Mercados e Outros	ARIMA(0,1,1)	1.33	ARIMA(0,1,1)	0	1.42	1.07
Hipermercados	ARIMA(0,1,1)	1.53	ARIMA(0,1,1)	0	1.50	0.98
Vestuário	ARIMA(0,1,0)	2.43	ARIMA(0,1,0)	4	2.25	0.93
Moveis e Eletrodomésticos	ARIMA(1,1,0)	1.61	ARIMA(0,1,1)	4	1.11	0.69
Farmaceuticos	ARIMA(0,1,0)	0.95	ARIMA(0,1,0)	1	0.99	1.04
Livros	ARIMA(0,1,3)	1.59	ARIMA(0,1,2)	6	1.14	0.71
Materiais de Escritório	ARIMA(0,1,1)	3.46	ARIMA(0,1,1)	6	3.42	0.99
Outros Artigos de Uso Pessoal	ARIMA(1,1,0)	1.22	ARIMA(0,1,1)	6	1.17	0.95
Veículos	ARIMA(2,1,0)	1.25	ARIMA(3,1,3)	4	1.19	0.95
Materiais de Construção	ARIMA(0,1,1)	1.98	ARIMA(0,1,1)	3	1.64	0.83

É interessante comparar os modelos analisados com base nos erros, observando que há um padrão de queda dos mesmos para esta modelagem ainda sem a inclusão dos dados do Google na regressão. O MSE de alguns modelos caiu consideravelmente: na série destinada a veículos e afins, o erro sem Trends estava em 12.37, seguindo para 3.64 na análise da série a partir de 2012 – uma queda de mais de 70%. Em combustíveis, a queda foi de 67%, para a PMC Ampliada, cerca de 50%. Também na faixa de 40% a 50%, as séries relacionadas ao varejo de livros, jornais e afins, outros artigos de uso pessoal e materiais de escritório. Há ainda um recuo considerável naquelas de móveis e eletrodomésticos e materiais de construção. Para vestuário e correlatos, mercados e farmacêuticos o erro é aumentado, devendo-se manter para *nowcasting*, as modelagens das seções anteriores.

O declínio do MAE também é verdade para quase todas as categorias da PMC, novamente excetuando-se "Tecidos, vestuário e calçados", "Supermercados, Hipermercados, produtos alimentícios, bebidas e fumo" e sua versão de núcleo "Supermercados, Hipermercados", para estas a melhor modelagem é a aque considera os dados desde 2004. Nos resultados positivos, o erro se contraiu mais de 30% em alguns casos, chegando até 53% ("PMC Ampliada", "Combustíveis e Lubrificantes" e "Veículos e motocicletas, partes e peças", respectivamente). Nas outras séries, observa-se uma retração de 6 até 27%.

A inclusão do Google Trends como regressor nos modelos amplifica o declínio dos erros. O erro quadrático médio de algumas séries apresenta um decréscimo de até 43%, como no caso do varejo de livros e afins, e 42% para a série de veículos e móveis e eletrodomésticos. A série da PMC Ampliada também tem o MSE retraído em consideráveis 27%. Nestes casos, destaca-se a defasagem dos dados do Google escolhidos, entre 4 e 6 meses, o que mostra o comportamento do consumidor em utilizar o Google meses antes de efetivar uma compra. A categoria "Tecidos, vestuário e calçados" foi melhor modelada na subseção anterior, mas aqui o seu MSE também apresentou um comportamento semelhante, com um declínio de 34%. Nesta série é interessante notar a defasagem da série, que é zero. Isto mostra que as pesquisas no Google explicam melhor as compras de forma contemporânea, isto é, os indivíduos não tendem a apresentar um padrão de pesquisas antecipadas ou comparação online com meses de antecedência neste quesito, o que faz algum sentido para o comércio de roupas – o que seria, por outro lado, questionável para as vendas de automóveis. Outras séries continuaram mostrando quedas menores, embora ainda expressivas. Da mesma forma, a melhor modelagem para "Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos" e "Equipamentos e materiais para escritório, Informática e de comunicação" não considera a regressão com o Trends, embora para esta primeira, já se tenha argumentado que a melhor modelagem é aquela disposta na subseção anterior.

Os avanços no MAE quando considerando os dados do Google Trends como regressor também são notáveis. As categorias "Móveis e Eletrodomésticos", "Material de Constru-

ção" e "Combustíveis e Lubrificantes" registraram retração do MAE de, respectivamente, 31%, 29% e 17% para as duas últimas. Retrações menores são observadas nas outras séries, e duas registraram um aumento do erro ("Artigos farmacêuticos, médicos, Ortopédicos, de perfumaria e cosméticos" e uma das medidas de supermercado, ambas já discutidas anteriormente).

7 Conclusão

Os dados de utilização da internet fornecem hoje uma das formas mais inovadoras de se conhecer o comportamento dos indivíduos e também o estado atual e direção de uma economia. O grande avanço é a possibilidade de obtê-los em tempo real, aumentando o conjunto de informação necessário para melhores tomadas de decisão, seja no setor público ou no setor privado.

Foi possível ao longo desta pesquisa selecionar pesquisas e categorias de pesquisa do Google que mais se aproximavam da dinâmica das vendas no varejo no Brasil. Partindo de modelos do tipo *ARIMA(p,d,q)* e analisando o erro quadrático médio (MSE) e o erro médio absoluto (MAE), alcançou-se melhoras para a maioria das categorias de pesquisa. Alguns destes avanços superaram a marca de 40%. Um ponto base do argumento proposto são as defasagens entre a pesquisa e a efetivação da compra, um comportamento natural do consumidor que pesquisa e compara antes de comprar.

Destacamos ao longo dos resultados obtidos diversos avanços em termos de modelagem, para diversas categorias diferentes do varejo, todos possibilitados pelo uso de dados gratuitos do Google. Uma vez que infinitos termos de pesquisa podem ser utilizados, abre-se a possibilidade de que para todas as categorias de consumo pesquisadas – mesmo aquelas que não apresentaram grande avanço neste estudo – possam ter um termo de pesquisa melhor especificado em trabalhos futuros, permitindo que modelos ainda melhores possam ser construídos e aumentando nossa capacidade de "prever o presente".

O grande de volume de dados gerado por usuários nas diversas plataformas da *web* alargou as fronteiras do possível e a são incontáveis as possíveis utilidades dos mesmos pelos diversos segmentos da sociedade. Um grande desafio talvez seja a capacidade de filtrar tais dados, mas é fato que estes permitem um maior conhecimento do comportamento dos indivíduos. Isto é aplicável não só para a previsão de curto prazo de variáveis macroeconômicas, mas há um relevante escopo para o desenvolvimento de pesquisas de análise de políticas públicas. É ainda natural que se discuta quais os limites do uso de dados, o que levou a recentes debates na questão da privacidade nos meios virtuais.

Entender a sociedade e o comportamento de seus agentes é tarefa árdua mas que as ciências sociais vêm se debruçando há muito. A tecnologia permitiu com que a vida cotidiana real se estendesse ao campo virtual alterando nossos padrões de relacionamento, interação e consumo, tornando-os de alguma forma mais complexos. Em contrapartida, um grande volume dessas relações complexas é passível de ser análisado pelas pegadas de comportamento que deixamos como reação aos diversos estímulos recebidos. Resta-nos utilizar estes dados com os melhores objetivos a fim de não só desvelar padrões sociais mas também compreender nossas mais complexas idiossincrasias.

Referências

[1] ARTOLA, C.; GALÁN, E. *Tracking the future on the web: construction of leading indicators using internet searches*. Banco de España Documentos Ocasionales., n. 1203, 2012.

[2] ASKITAS, N.; ZIMMERMANN, K. F. Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, v. 55, n. 2, p. 107 - 120, 2009.

[3] BAKER, S.; FRADKIN, A. *What Drives Job Search? Evidence from Google Search Data*. SIEPR Discussion Paper, n. 10-020, 2011.

[4] BOLLEN, J.; MAO, H.; ZENG, X. *Twitter mood predicts the stock market*. 2010. Disponível em: <https://arxiv.org/pdf/1010.3003>. Acesso em: 05/05/2017.

[5] BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis: Forecasting and Control*. 4. ed. [S.l.]: Wiley, 2008.

[6] BUGHIN, J. *Nowcasting the Belgian Economy*. Disponível em: <https://goo.gl/Uogzoj>. Acesso em: 05/05/2017.

[7] CARRIÈRE-SWALLOW, Y.; LABBÉ, F. Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting*, v. 32, n. 4, p. 289 - 298, 2013.

[8] CHOI, H.; VARIAN, H. *Predicting Initial Claims for Unemployment Benefits*. Disponível em: <https://research.google.com/archive/papers/initialclaimsUS.pdf>. Acesso em: 05/05/2017.

[9] CHOI, H.; VARIAN, H. Predicting the Present with Google Trends. *The Economic Record*, v. 88, n. s1, p. 2 - 9, 2011.

[10] FRANCESCO, D. *Predicting unemployment in short samples with internet job search query data*. MPRA Paper, n. 18403, 2009.

[11] GINSBERG, J. et al. Detecting influenza epidemics using search engine query data. *Nature*, n. 457, p. 1012 - 1014, novembro 2008.

[12] BRASIL. Instituto Brasileiro de Geografia e Estatística. *Indicadores IBGE – Pesquisa Mensal do Comércio*, fevereiro 2017.

[13] KOOP, G.; ONORANTE, L. *Macroeconomic Nowcasting Using Google Probabilities*. In: First International Conference on Advanced Research Methods and Analytics. [S.l.: s.n.], 2013.

[14] MCLAREN, N.; SHANBHOGUE, R. *Using internet search data as economic indicators*. Bank of England Quarterly Bulletin, v. 51, n. 2, p. 134 - 140, 2011.

[15] MONTGOMERY, D. C.; JENNINGS, C. L.; KULAHCI, M. *Introduction to Time Series Analysis and Forecasting*. [S.l.]: Wiley, 2007.

[16] SUHOY, T. *Query Indices and a 2008 Downturn: Israeli Data*. Bank of Israel Discussion Paper, n. 06, 2009.

[17] WOOLDRIDGE, J. M. *Introdução à econometria: uma abordagem moderna*. 4. ed. São Paulo: Cengage Learning, 2013. ISBN 978-85-221-0446-8.

[18] WU, L.; BRYNJOLFSSON, E. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In: NATIONAL BUREAU OF ECONOMIC RESEARCH. *Economic Analysis of the Digital Economy*. [S.l.]: University of Chicago Press, 2015. p. 89 - 118.