UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE ECONOMIA

MONOGRAFIA DE BACHARELADO

PREVISÃO DO VOLUME DE VENDAS DO COMÉRCIO VAREJISTA COM O USO DO *GOOGLE TRENDS*

MATHEUS PASCHE AZEVEDO

Matrícula: 116194745

ORIENTADORA: Prof[a] Susan Schommer

Rio de Janeiro

Fevereiro/2021

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE ECONOMIA

MONOGRAFIA DE BACHARELADO

PREVISÃO DO VOLUME DE VENDAS DO COMÉRCIO VAREJISTA COM O USO DO *GOOGLE TRENDS*

MATHEUS PASCHE AZEVEDO

Matrícula: 116194745

ORIENTADORA: Prof[a] Susan Schommer

Rio de Janeiro

Fevereiro/2021

As opiniões expressas neste trabalho são de exclusiva responsabilidade do(a) autor(a)

AGRADECIMENTOS

Esta nota destina-se a lembrar e reconhecer o tamanho apoio e esforço que recebi para a criação deste trabalho, sobretudo pela complexidade econômica, psicológica e social imposta pela pandemia do coronavírus, que assolou o mundo em 2020, ano de criação deste trabalho.

Agradeço, sobretudo, aos brasileiros que, por meio de trabalho, suor e esforço, mantém uma universidade como a UFRJ. É deles pelos recursos e é para eles que todo trabalho desenvolvido deve retornar.

A Alexandra Elbakyan por permitir o acesso livre ao conhecimento científico.

Aos meus pais, Renata e Alexandre, por terem batalhado tanto para que eu pudesse ter a melhor educação que poderiam custear e que me fez chegar à maior Universidade Federal do país. Não foi fácil e nem simples, mas valeu e valerá à pena. Agradeço também por todo o incentivo para que eu prosseguisse visando sempre a excelência que eu posso alcançar. Cheguei na faculdade que nem eu mesmo saberia que seria possível chegar e me despeço com o melhor que pude dar em um trabalho ao longo de 12 longos meses.

Ao meu tio Ronaldo e minha irmã, Larissa por terem sido importantes colaboradores em toda a trajetória.

A Mariana, meu amor, por conceder indefinida e voluntariamente o notebook em que foi feita a maior parte do processo. Agradeço pela confiança, suporte, crítica, correção, incentivo e até mesmo pela lembrança de que estava na hora de parar de digitar código e ir dormir.

A toda equipe do IFec RJ, pelo apoio e confiança. Que este trabalho sirva para o enriquecimento da análise macroeconômica do estado do Rio de Janeiro. Agradeço, em especial, a Rafael Zanderer, pelos inúmeros ensinamentos, por atuar como co-orientador e participar ativamente na construção do código e do próprio entendimento do modelo.

A Susan Schommer, minha orientadora, por toda a sabedoria e conhecimento prático e teórico que me passou. O interesse por econometria vem, em boa parte, porque tive uma excelente professora.

A Gabriel Vasconcellos, criador do pacote *HD Econometrics*, de onde partiu o início e parte do desenvolvimento do trabalho. Agradeço pelas dicas pontuais e por colaborar com a comunidade com os frutos do seu trabalho de doutorado.

Por fim, mas não menos importante aos amigos do ETZ, que contribuíram com dicas de artigos e com o código.

"E o futuro é uma astronave que tentamos pilotar,
Não tem tempo nem piedade, nem tem hora de chegar"
(Aquarela - Toquinho)

RESUMO

A tentativa de análise macroeconômica das entidades subnacionais no Brasil enfrenta diversos percalços em termos de disponibilidade de dados, como séries incompletas, com divulgação irregular ou de difícil acesso para não familiarizados com o tema. Uma alternativa possível e acessível para a compreensão parcial dos movimentos conjuntural é observar as pesquisas sobre o volume de vendas do comércio varejista, da indústria e dos serviços, todos produzidos pelo IBGE. Estas, no entanto, são divulgadas com dois meses de defasagem frente ao período corrente. Este trabalho visa criar um instrumental de seleção de variáveis e modelos preditivos de curtíssimo prazo a fim de reduzir a diferença temporal entre os dados disponíveis e o mês em curso com uso das estatísticas providas pelo *Google Trends*, base que constitui índices de pesquisas no buscador Google de acordo com palavras-chave identificadas como *proxies* da atividade econômica, com aplicação à pesquisa do volume de vendas do comércio varejista do Rio de Janeiro. Foram confrontados diversos algoritmos de seleção de variáveis presentes na literatura recente, que posteriormente foram utilizados como insumo para os modelos econométricos. O melhor resultado foi obtido pelo modelo *Complete Subset Regressions* com variáveis selecionadas pelo Critério de Informação de Akaike modificado para amostras pequenas.

Palavras-Chave:modelos de previsão, Pesquisa Mensal do Comércio, volume de vendas do comércio, *Google Trends*, *Complete Subset Regressions*, *Least Absolute Shrinkage and Selection Operator*, *Nowcasting*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PMC	Pesquisa Mensal do Comércio
CSR	*Complete Subset Regressions*
ASM	Algoritmo de seleção de modelos
CI	Critério de informação
AIC	*Akaike Information Criterion*
AICc	*Akaike Information Criterion corrigido para pequenas amostras*
BIC	*Bayesian Information Criterion*
LASSO	*Least Absolute Shrinkage and Selection Operator*
RMSE	*Root mean square error*
MQO	Mínimos Quadrados Ordinários

SUMÁRIO

I INTRODUÇÃO

A análise da conjuntura macroeconômica nas entidades subnacionais do Brasil pode ser uma tarefa árdua do ponto de vista da obtenção dos dados. Desde séries incompletas e defasadas em um horizonte temporal distante do momento presente alguns dados não estão disponíveis em formato editável ou de complexa obtenção, a disponibilidade e a transparência variam de estado para estado. A partir dessa problemática, surgem duas alternativas: i) a previsão de variáveis já existentes de modo a reduzir a defasagem temporal, ou; ii) criar novos indicadores a partir dos dados já disponíveis. Este trabalho se debruça na primeira solução e cria um instrumental que pode ser utilizado para a segunda alternativa.

No entanto, se o problema é justamente a ausência de dados, qualquer previsão multivariada enfrentaria um grande obstáculo de reunião de séries que sirvam como insumo para os modelos econométricos. Tendo como base o trabalho de Scott e Varian (2013), que preveem a venda de carros por meio dos dados do *Google Trends*, ferramenta que cria um índice de pesquisas no Google por palavra-chave, objetiva-se elevar este potencial à possibilidade de previsão de todo o comércio varejista.

A variável a ser prevista neste trabalho é o volume de vendas da Pesquisa Mensal do Comércio (PMC) para o estado do Rio de Janeiro. No entanto, o instrumental desenvolvido pode ser facilmente replicado e adaptado para outras pesquisas equivalentes à PMC tais como os setores de serviços e indústria.

No capítulo seguinte será discutida a metodologia de análise, em que são detalhadas as estruturas e características da PMC e do *Google Trends*, bem como os Algoritmos de Seleção de Modelos (ASM), métricas que auxiliam na especificação dos modelos de forma mais parcimoniosa e com o menor erro possível. Em seguida, serão apresentados os modelos utilizados, ARIMA, *Complete Subset Regressions* (CSR), *Least Absolute Srinkage and Selection Operator* (LASSO) e Regressão Ridge.

Ao fim das comparações e testes, o método *Complete Subset Regressions* se mostrou o mais eficiente em prever com o menor valor de erro quadrático médio fazendo uso da seleção de variáveis por *Akaike Information Criterion*. A Regressão Ridge apresentou desempenho semelhante com o uso de variáveis selecionadas por LASSO-AICc. Ambos apresentaram diferença considerável em relação à previsão com uso do processo Autorregressivo de ordem 3.

.

II METODOLOGIA

As previsões deste trabalho serão de curtíssimo prazo, isto é, apenas um período à frente do último dado observado na série temporal, com o uso de *nowcasting*. O termo, é uma diferenciação à palavra *forecasting* no que se refere ao horizonte de previsão. Enquanto este caracteriza previsões *stricto sensu*, isto é, com base em valores ainda não conhecidos, *nowcasting* tem o objetivo de "prever o presente"e antecipar o cenário macroeconômico.

Dessa forma, *nowcasting* e *forecasting* são complementares e ambos os horizontes devem ser observados para uma construção adequada do cenário macroeconômico, de tal forma que as previsões de curto prazo podem ser usadas como base de previsões de longo prazo. Como exemplo poder-se-ia citar o resultado do uso de uma técnica de *nowcasting* assumido como "verdadeiro"para sustentar um modelo de Vetores Autorregressivos, de modo que seja possível prever mais observações antes da convergência à média.

II.1 Pesquisa Mensal do Comércio

A Pesquisa Mensal do Comércio é uma sondagem realizada junto a 6.157 empresas cuja principal atividade é o comércio varejista, sediadas em território nacional e com quantidade de empregados maior ou igual a 20 em caso de operação em apenas um estado ou com menos de 20 funcionários, desde que com atuação em mais de uma Unidade da Federação. A pesquisa teve início em janeiro de 1995 na Região Metropolitana do Rio de Janeiro e teve sua ampliação para todo o país em 2000, período usado como referência neste trabalho.

A partir da consulta às empresas, a receita bruta de revenda das mercadorias de fabricação não própria, sem considerar os impostos são construídos os indicadores de receita nominal e volume de vendas, esta última a ser considerada neste trabalho. A série é composta por oito grupos, acrescidos de *Veículos e motocicletas, partes e peças* e *Material de Construção* em sua versão ampliada. A razão pela opção da alternativa restrita se deu pela praticidade em poder comparar de desempenho contra instituições de consultoria, corretoras, bancos e demais empresas do mercado financeiro.

Além disso, apesar da PMC, bem como as demais pesquisas do IBGE serem divididas em receita nominal e volume de vendas, o habitual para o acompanhamento conjuntural é o uso do volume de vendas. No entanto, tendo em vista que existe uma correlação alta entre receita e volume de vendas, é possível utilizar o mesmo instrumental para a previsão da receita sem consideráveis problemas.

Tabela 1 – Grupos da Pesquisa Mensal do Comércio

Instituto	Grupo
IBGE	Combustíveis e lubrificantes
	Hipermercados e supermercados
	Produtos alimentícios, bebidas e fumo
	Tecidos vestuário e calçados
	Móveis e eletrodomésticos
	Artigos farmacêuticos, médicos, ortopédicos, perfumaria e cosméticos
	Equipamentos e Material para escritório, informática e comunicação
	Livros, jornais, revistas e papelaria
	Outros artigos de uso pessoal e doméstico

Fonte: IBGE e cálculos do autor.

II.2 *Google Trends*

Google Trends consiste em uma ferramenta de consulta às pesquisas feitas por meio do buscador Google. Segundo a empresa, a divulgação em números absolutos dificilmente seria rápida, acessível e em tempo real, como se dá hoje por meio do site Google Trends, uma vez que são realizadas bilhões de pesquisas todos os dias e em quase todo o mundo. Como forma de contornar esse problema, os dados são divulgados em um formato de índice que varia de 0 a 100, em que o valor mínimo é atingido somente quando não há amostra significativa para gerar qualquer valor comparativo e o máximo está sempre presente e servirá de referencial para os demais valores.

O índice é normalizado dividindo o valor absoluto pelo número de pesquisas na região e tempo, reduzindo o viés do acesso à internet e celulares. É preciso ter como base que em 2004, data de início dos registros do *Google Trends*, as tecnologias e a confiança dos consumidores no comércio eletrônico eram drasticamente diferentes, de modo que é esperada uma tendência de alta nas pesquisas na maior parte dos itens pesquisados. Se em 2004 comprar uma televisão pela internet poderia ser algo praticamente inexistente, atualmente esse tipo de compra já está integrada em nossa sociedade. Acima de tudo, ainda considerando transformações da economia e da sociedade, as variações, e não o nível, serão importantes para prever os próximos valores.

Uma questão relevante advinda da atualização em tempo real é a não constância dos índices, que consequentemente afeta as variações e eventualmente pode incluir ou excluir variáveis analisadas sob o mesmo algoritmo de seleção de modelos. Esta é uma questão ainda sem solução e que pode ser problemática sobretudo em previsões que utilizem um número baixo de variáveis. Neste trabalho foram realizados alguns testes e a repetição da captura do mesmo conjunto de palavras em dias diferentes não gerou mudanças relevantes na qualidade preditiva. Esta conclusão não é definitiva e pode mudar drasticamente em eventos atípicos. Por exemplo, um evento em uma empresa pode ser tão polêmico a ponto

Tabela 2 – Palavras-chave utilizadas e os grupos da PMC

Combustíveis e lubrificantes	Gasolina, Diesel, Petrobras, Shell
Hipermercados, supermercados, produtos alimentícios, bebidas e fumo	Supermercado, Extra, Carrefour, Guanabara,
	Pizza, Sorvete, Coca Cola, Skol, Cerveja, Brahma
Tecidos, vestuário e calçados	Shopping, Cinema, Ingresso, Calçados, Vestido,
	Camisas, Tecidos, Cueca, Tenis, Mochila, Marisa, Leader
Móveis e eletrodomésticos	Móveis, Fogão, Televisão, Ar Condicionado, Forno,
	Câmera, Mesa, Violão, Cortina, Pratos, Casas Bahia,
	Magazine Luiza, Ponto Frio, Casa E Video, Brastemp, Lojas Americanas
Artigos farmacêuticos, médicos, ortopédicos, de perfumaria e cosméticos	Farmácia, Homeopatia, Maquiagem, Natura
Livros, jornais, revistas e papelaria	Papel, Livraria, Saraiva, Livraria Cultura, Gráfica
Equipamentos e materiais para escritório, informática e comunicação	Softwares, Computador, Teclado,
	Monitor, Asus, LG, Samsung, Motorola, HP, Notebook
Outros artigos de uso pessoal e doméstico	Antena, Brinquedos, Boneca, Playstation,
	Xbox, Nintendo, Lego, Barbie

Fonte: Elaboração do autor.

de destinar um número de pesquisas muito maior que qualquer valor já observado, ainda que sem relação com a atividade econômica. Espera-se que nesses casos os algoritmos de seleção de modelos não incluam tal palavra chave. De todo modo, o analista deve estar sempre atento ao conjunto de palavras a fim de eliminar as que se comportem de maneira anômala.

A escolha de variáveis neste trabalho foi baseada nos códigos das atividades da Classificação Nacional de Atividades Econômicas (CNAE) que compõem a Pesquisa Mensal do Comércio. Ainda assim, a distribuição da quantidade de palavras por grupos não segue o peso que desempenham na PMC, dado que há inúmeros casos em que não há amostra significativa para gerar valor de índice. Em produtos ou serviços recentes, como "iFood", que não existiam em 2004, a matriz assume valor zero. Como as séries em sua totalidade não são estacionárias, o processo de variação contra o igual mês do ano anterior seria infinito, impossibilitando a estimação. As palavras-chave nessa situação foram eliminadas.

Importa mencionar que o processo de construção por meio das CNAEs se deu pela tentativa de encontrar palavras que pudessem ser utilizadas pelos consumidores em buscas no Google. Dessa forma, este trabalho visa apresentar o instrumental e os resultados obtidos fazendo uso do conjunto apresentado abaixo. É plenamente possível que um leitor ao tentar replicar este trabalho obtenha resultados mais satisfatórios por conseguir identificar um grupo mais adequado de palavras-chave.

É possível observar na tabela acima que foram utilizados os nomes de produtos como "computador"e nomes próprios, como "Magazine Luiza". Tal mescla surge a partir de testes e tentativas. Espera-se que com a expansão esperada do comércio eletrônico nos próximos anos, a busca pelas grandes marcas seja ainda mais importante na previsão do volume de vendas.

II.3 Escolha de variáveis

Apesar de todas as séries escolhidas apresentarem relação baseada na teoria econômica com a variável a ser prevista, ter um critério de seleção por algum marco estatístico se mostra necessário, uma vez que se observou que a exclusão de algumas séries teria o condão de apresentar menores erros de previsão, além de consequentemente possibilitar agilização do processo de estimação. Castle et al. (2009) justifica a necessidade de uso um algoritmo objetivo tendo como base as possibilidades de, no melhor caso, que a especificação mais eficiente não seja utilizada – o que em última instância leva a um modelo com maiores erros – e no pior dos cenários poderia reforçar vieses do pesquisador.

Existe uma vasta literatura recente a respeito dos Algorítimos de Seleção de Modelos, sobretudo pelo crescente uso de técnicas de *machine learning*. Seguindo Castle et al. (2009) e Scott e Varian (2013), as técnicas utilizadas podem ser agrupadas do seguinte modo:

- Critérios de Informação
- Significância estatística
- Regressão com penalizações

II.3.1 Critérios de informação

A seleção de variáveis por critérios de informação (CI) como o de Akaike (AIC) (AKAIKE, 1974) ou o critério de Schwarz, também conhecido como critério bayesiano (SCHWARZ, 1978) são amplamente utilizados e estão relacionados, haja vista que ambos seguem a estrutura $-2\mathcal{L} + penalização$, em que $\mathcal{L}$ é o máximo valor do logarítimo da função de verossimilhança e a penalidade cresce monotonicamente com o objetivo de garantir um modelo parcimonioso. Portanto, o menor valor possível é desejado.

Ambos os CIs tendem a incluir muitas variáveis quando as amostras são pequenas, o que se mostra o caso neste trabalho Hurvich e Tsai (1989) criam uma variação ao AIC com correção para tamanho de amostra, que será denotado por AICc.

$$AIC = -2\mathcal{L} + 2k \tag{II.1}$$

$$AICc = -2\mathcal{L} + 2k\left(\frac{n}{n-k-1}\right) \tag{II.2}$$

$$BIC = -2\mathcal{L} + k * ln(n) \tag{II.3}$$

Em que n representa o número de observações e k, o número de variáveis no modelo.

Idealmente, um CI que funcione como ASM deve comparar todas as combinações possíveis de modelos e escolher o melhor ou os melhores modelos. Poskitt and Tremayne

(1987) argumentam que talvez seja interessante não considerar apenas a equação que apresentou o melhor valor para o CI adotado, mas um portfólio com algumas equações que estejam entre as melhores. Entretanto, como demonstrado no anexo, a soma de combinações possíveis é 2^K, o que se mostrou computacionalmente inviável. A alternativa sub-ótima adotada se deu por um algorítimo que realiza a regressão de y contra todas as candidatas e escolhe a com menor; dentre as candidatas restantes, adiciona, se existir, aquela que apresentar o menor valor do CI. O processo é repetido até que o chegue ao seu mínimo.

Input: Variaveis

Output: Selecionadas

Algoritmo 1: Seleção por Critério de Informação

```
for j in 1:ncol(Variaveis) do
    /* Seja 'Variaveis' todo o conjunto de séries previsoras disponível, com i
       linhas e j colunas;                                                    */
    /* Seja 'ValorCriterio' uma matriz linha de zeros e j colunas;            */
    /* Seja 'Candidatas' = Variaveis                                          */
    /* Seja 'Selecionadas' uma matriz [i,j] de zeros                          */
    /* Seja 'CI' uma função dos os Critérios de Informação                    */
    ValorCriterio[1, j] = CI(lm(y ~ Variaveis[1, j]))
    Candidatas = Variaveis[, −which.min(ValorCriterio)]
    Selecionadas = Variaveis[, which.min(ValorCriterio)]
/* Seja 'CI.Min' um vetor que receberá o menor valor de CI obtido             */
/* Seja 'CI.Loop' uma matriz [1,j] que receba todos os valores dos CIs.       */
1
CI.Min = abs(CI(lm(y ~ Selecionadas)))
while CI.Min ≤ min(CI.Loop) do
    for h in 1:ncol(Candidatas) do
        Selecionadas = cbind(Selecionadas, Candidatas[, h])
        CI.Loop[, h] = abs(CI(lm(y ~ Selecionadas)))
        Selecionadas = Selecionadas[, −ncol(Selecionadas)]
    if min(CI.Loop) < CI.Min then
        Selecionadas = cbind(Selecionadas, Candidatas[, which.min(CI.Loop)])
        Candidatas = Candidatas[, −which.min(CI.Loop)]
        CI.Min = min(CI.Loop)
    else
        CI.Loop = 0
```

II.3.2 Significância estatística

II.3.2.1 p-valor

Uma alternativa semelhante à descrita no ASM por CI é realizar uma série de eliminações até que a regressão possua um determinado p-valor desejado, como descrito no algoritmo abaixo

Input: Variaveis

Output: Selecionadas

Algoritmo 2: Seleção por p-valor

```
/* Seja 'Variaveis' todo o conjunto de séries previsoras disponível, com i
   linhas e j colunas;                                                    */
/* Seja 'ValorCriterio' uma matriz linha de zeros e j colunas;            */
/* Seja 'maximo' um objeto que receberá o maior valor dentre os p-valores da
   regressão. Define-se inicialmente como 100%                            */
maximo = 100%
while maximo > p.value do
    ValorCriterio = summary(lm(y ~ ., Variaveis))$coefficients[, "Pr(>|t|)"]
     maximo = max(ValorCriterio)
    if maximo ≥ p.value then
        Variaveis = Variaveis[, −which.max(ValorCriterio)] maximo = 100%
```

II.3.2.2 Estatística T

O artigo de Garcia et. al (2017), por sua vez, propõe um algorítimo de seleção de variáveis que consiste em:

1. Para cada i = 1,...,K, realize-se uma regressão individual de y_t contra X_i, t.
2. Selecione-se as k variáveis com maiores |t|.

O número de variáveis k escolhidas dentre as K disponíveis é arbitrário e, tendo em vista que o número de equações cresce com trajetória exponencial, deve ser pensado de modo que seja computacionalmente possível realizar estimação. Todavia, tal arbitrariedade mantém em aberto que se tenha um método que garanta que as séries em uso serão as que possuem capacidade preditiva mais acurada dentre o conjunto disponível. Foram realizados testes com as top-10, top-15, top-20, top-25 e top-30.

II.3.3 Regressão com penalizações

O método LASSO cumpre nesta análise um papel duplo de algoritmo de seleção de modelos e um dos métodos de previsão. Como será mais profundamente analisada na seção Modelo, o LASSO possui a capacidade de zerar coeficientes pouco significativos nas regressões. Adotou-se a possibilidade de coletar as variáveis em que os valores dos coeficientes eram diferentes de zero e utilizá-las, por exemplo, para realizar previsões com o método *Complete Subset Regressions*.

III MODELOS

Uma vez selecionadas as variáveis, esta seção contempla a apresentação dos modelos utilizados para a previsão. Dos quatro modelos descritos, apenas um é monovariado: ARIMA. Os demais utilizarão as variáveis selecionadas pelos ASMs como insumo para as previsões que farão.

III.1 Descrição dos modelos

III.1.1 ARIMA

O método *autorregressive integrated moving average*(ARIMA), amplamente conhecido pelos economistas, é uma generalização do modelo ARMA, proposto por Box e Jenkins (1994), que une a estimação.

Processo autorregressivo (AR). Neste modelo, os valores do período corrente t são expressos como um agregado linear finito dos períodos anteriores $t-1, t-2, ..., t-p$ por $y_t, y_{t-1}, y_{t-2}, ...$ Seja $\bar{y_t}$ o desvio de y_t em relação à média μ. Assim,

$$\bar{y} = \phi_1 \bar{x}_1 + \phi_2 \bar{x}_2 + ... + \phi_p \bar{x}_p + a \tag{III.1}$$

é conhecido como processo autorregressivo (AR) de ordem p. Defina-se um operador autorregressivo de ordem p por:

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

o modelo autorregressivo pode ser escrito economicamente como:

$$\phi(B)\hat{y}_t = a_t \tag{III.3}$$

Médias Móveis (MA). O processo autorregressivo expressa o desvio $\hat{y}_t = y_t - \mu$ como uma soma ponderada de ordem p e sujeita ao choque a_t, além de expressar y_t como uma soma ponderada infinita de a's.

O processo de médias móveis se baseia no estabelecimento da dependência linear de y_t em relação a um número finito de a's. Assim,

$$\bar{y_t} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \tag{III.4}$$

é um processo de médias móveis de ordem q. Apesar da nomenclatura, os pesos de θ's não necessitam seguir as propriedades de qualquer média, tais como serem positivos e apresentarem somatório resultando na unidade. O operador será definido como:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q \tag{III.5}$$

que pode ser escrito economicamente como

$$\bar{y_t} = \theta(B)a_t \tag{III.6}$$

ARMA. Em muitas séries temporais é possível unir os dois modelos previamente apresentado conhecido como modelo ARMA.

$$\bar{y_t} = \phi_1\bar{y}_{t-1} + ... + \phi_p\bar{y}_{t-p} + a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-p} \tag{III.7}$$

$$\varphi(B) = \phi(B)(1-B)^d \tag{III.8}$$

em que $\phi(B)$ é o operador estacionário. Portanto, um modelo que represente o comportamento não estacionário seria

$$\varphi(B)y_t = \phi(B)(1-B)^d y_t = \theta(B)a_t \tag{III.9}$$

isto é,

$$\varphi(B)w_t = \theta(B)a_t \tag{III.10}$$

em que

$$w_t = \nabla^d y_t \tag{III.11}$$

ARIMA. A união do processo ARMA com a generalização sobre estacionaridade é chamada de *autorregressive integrated moving average* (ARIMA), que assume ordem (p, d, q), definido por:

$$w_t = \varphi_1 w_{t-1} + ... + \varphi_p w_{t-p} + a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q} \tag{III.12}$$

III.1.2 Regressão Ridge

A Regressão Ridge e o LASSO fazem parte de um conjunto de modelos conhecidos como *shrinkage methods*. Ambos minimizam a soma do quadrado dos erros penalizados.

$$\hat{\beta}^{ridge} = \arg\min_{\hat{\beta}} \left[\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{N} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{N} {\beta_j}^2 \right] \tag{III.13}$$

Em que $\lambda \geq 0$ é um parâmetro sobre o tamanho do encolhimento realizado pela penalização. Quanto maior λ, maior será o peso da penalização. Existem inúmeros critérios para determinação de λ na literatura de *machine learning*, tais como *cross-validation*, critérios de informação e o estabelecimento por parte das hipóteses do analista. Neste caso, optou-se pelo uso de critérios de informação.

III.1.3 LASSO

Tibshirani (1996) propõe o LASSO, que é bastante semelhante à Regressão Ridge, mas com uma fundamental diferença.

$$\hat{\beta}^{LASSO} = \arg\min_{\hat{\beta}} \left[\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{N} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{N} |\beta_j| \right] \tag{III.14}$$

Nas análises deste trabalho, o valor de λ foi escolhido com base nos CIs AIC, AICc e BIC. Dessa forma, "LASSO-AICc", por exemplo, significa a estimação de um modelo LASSO, com λ escolhido de acordo com o menor valor de AICc[1].

Importa ressaltar que tanto o LASSO quanto a Regressão Ridge já apresentam métodos de seleção por penalização. Enquanto o primeiro possui a propriedade de zerar coeficientes, a Regressão Ridge minimiza os coeficientes de algumas variáveis em seu processo de estimação. A opção pelo uso de algorítimos de seleção de variáveis previamente a esses modelos não é habitual. O exercício se deu apenas pela padronização do procedimento para todos os modelos. É plenamente possível especificar um modelo LASSO ou Regressão Ridge contra todas as variáveis e averiguar que no processo de estimação é realizado um processo de seleção.

III.1.4 Complete Subset Regressions (CSR)

Seja $\mathbf{x}_t = \{x_1, \ldots, x_{K_x}\}$ um conjunto de K_x variáveis previsoras. O modelo conhecido como *Complete Subset Regressions* (CSR) proposto por Elliott et. al(2013) realiza as combinações de regressões lineares e a previsão para cada k é dada pela média das previsões de todas as combinações. É possível e opcional extrair o conjunto $\mathbf{z}_t = \{z_1, \ldots, z_{K_x}\}$ de variáveis que operarão como controles fixos, isto é, sempre especificadas. Dessa forma, a quantidade de elementos em $\mathbf{z}_t$ e $\mathbf{x}_t$ é necessariamente igual ao número de vetores pertencentes a K_x.

Seja a equação abaixo um exemplo de *subset regression*.

$$\hat{y}_{t+1} = \delta' z_t + \beta' x_t + u_t \qquad \forall t = 1...T \tag{III.15}$$

Sejam $\hat{\delta}$ e $\hat{\beta}$ os valores estimados de δ e β, respectivamente. A previsão será definida por:

$$\hat{y}_{T+1} = \frac{\sum_{i=1}^{Q} \hat{\delta} z'_T + \sum_{i=1}^{K} \hat{\beta} x'_T}{Q + k} \tag{III.16}$$

Em que z_t é um vetor $Q \times 1$ de variáveis sempre presentes nas equações e x_t, as que estarão sujeitas ao processo de combinações que caracterizam o modelo, um vetor $K \times 1$. Além

[1] O código para estimação do LASSO por meio de critérios de informação foi amplamente baseado no pacote *HD Econometrics*

disso, se faz necessário que $dim(\delta) + dim(\beta) = Q + K < T$, uma vez que o número de parâmetros deve ser menor que o número de observações para que o método dos Mínimos Quadrados Ordinários (MQO) possa ser estimado.

III.1.4.1 Questões computacionais

O número de combinações possíveis é dado por ${}_kC_K = K!/(k!(K-k)!)$. Como é possível ver abaixo, a curva do número de combinações para um K fixo possui forma aproximada de um sino. Por exemplo, para K = 24 e k = 12, tem-se um número de 2.704.156 para que seja possível prever um período, o que pode acarretar em dias para estimação em um sistema local de apenas um processador, sobretudo no período de treinamento, em que se multiplicará a quantidade de equações pelo número de meses necessários para o treinamento. Dessa forma, o uso do CSR pode ser inviabilizado de forma prática quando o número de variáveis é suficientemente grande.

Como se trata de uma combinação, temos, pelo Binômio de Newton, que[2]

$$\sum_{k=1}^{K} \binom{K}{k} = 2^k - 1 \qquad 1 \leq k \leq K \tag{III.17}$$

Existem algumas formas para tentar reduzir o custo computacional de realizar milhões de equações, como a substituição de programação condicional padrão – *for, while* por programação funcional, sendo a principal o uso do pacote *purrr*, que realiza as iterações de forma consideravelmente mais rápida que a programação padrão.

[2] Mais comumente encontra-se que a equação resulte em 2^k, desde que o valor mínimo de k seja zero, impossível para um CSR.

IV RESULTADOS E COMPARAÇÕES

Neste capítulo serão analisados os resultados obtidos após tentativa de prever a informação subsequente por meio dos modelos ARIMA, LASSO, Regressão Ridge e CSR.

Como mencionado previamente, a série do volume de vendas da PMC do Rio de Janeiro tem seu início em Janeiro de 2000 e é atualizada mensalmente. O início dos dados do *Google Trends* se dá em Janeiro de 2004, sendo esta data, portanto, o início da análise deste trabalho.

Optou-se por restringir a data final de corte a Fevereiro de 2020, posto que a excepcionalidade do pandêmico nos meses seguintes de do mesmo ano poderia levar a afirmações sobre a qualidade preditiva que não correspondem à média em condições ordinárias, com especial agravo ao modelo ARIMA, que não consegue captar grandes deslocamentos pela própria natureza de se basear nos valores passados.

Foi realizada uma separação entre dois períodos: treino e teste. O período de treino compreende de Janeiro de 2004 a março de 2017, um total de 159 observações, cerca de 82% do período disponível. As 35 observações restantes, compreendidas de Abril de 2017 a Fevereiro de 2020, serão o período em que as previsões serão realizadas e analisadas.

IV.1 Bases de dados

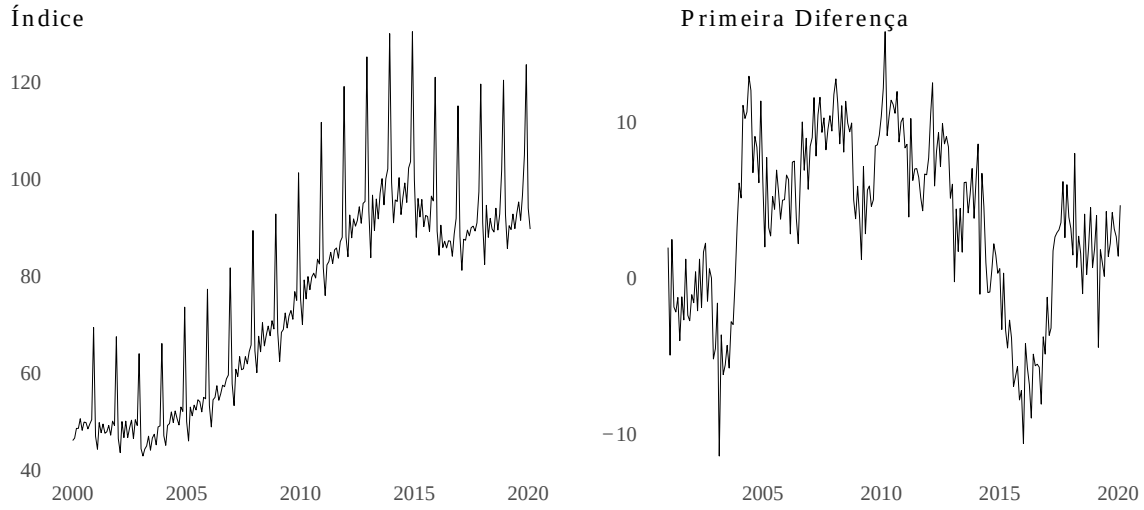
A observação a série do volume de vendas da PMC em nível, é possível constatar forte componente sazonal nos meses de dezembro por decorrência das festas de fim de ano, além de flutuações decorrentes de datas comemorativas como páscoa, dia das mães e dia dos pais. A série apresenta claro comportamento não-estacionário e o teste *Augmented Dickey-Fuller* (ADF) (DICKEY; FULLER, 1979) não pode rejeitar a hipótese nula de presença de raiz unitária.

A partir dessa constatação, surgem duas possibilidades de diferenciação: a previsão em relação ao mês anterior com o uso da série dessazonalizada já disponível pelo IBGE, ou a variação em relação ao mesmo mês do ano anterior. Optou-se pela última alternativa pelo fato de, considerando que diversas outras séries também apresentariam componente sazonal, a diferenciação interanual não tornaria necessária quaisquer ajustes sazonais. É possível obter a previsão dessazonalizada seguindo o ajuste recomendado pelo IBGE [1].

A série foi transformada em uma variação $y_t = \frac{\Delta_1 2 Y_t}{Y_{t-12}}$ de modo a garantir estacionariedade. Após transformação, testou-se a hipótese de raiz unitária por meio do teste ADF.

[1] Ver Nota metodológica do IBGE

Figura 1 – Volume de vendas da Pesquisa Mensal do Comércio



Fonte: IBGE. Elaboração do autor.

O resultado foi de **-2,29**, que aponta rejeição à hipótese H_0 de existência de raiz unitária a 5%, sendo, portanto, estacionária.

Tabela 3 – Valores críticos

	1%	**5%**	**10%**
Valores críticos	-2,58	-1,95	-1,62

Fonte: Elaboração do autor.

O mesmo tratamento foi aplicado para as variáveis do *Google Trends*. Todas apresentavam comportamento não-estacionário. Após a variação em relação ao igual mês do ano anterior, todas puderam rejeitar H_0 de existência de raiz unitária. A tabela com os valores críticos após a variação pode ser encontrada no Anexo.

IV.2 Avaliação de modelos

A avaliação da qualidade preditiva será dada pela raiz do erro quadrático médio (RMSE) multiplicado por 100 para facilitar a visualização, definida por:

$$RMSE = 100 * \left(\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T} \right)^{1/2} \tag{IV.1}$$

Apesar de existir uma série de critérios de avaliação, a razão da utilização do RMSE se deu pela sua fácil e imediata compreensão

IV.2.1 ARIMA

Seguindo a metodologia de Box & Jenkins(1994), deve-se estabelecer as ordens do modelo por meio da Função de Autocorrelação (ACF), que estabelece a ordem MA. A autocorrelação amostral é definida por

$$\hat{\rho}_j = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(y_{t-j} - \bar{y})/T}{\sum_{t=1}^{T}(y_t - \bar{y})^2/T} \tag{IV.2}$$

Em que $\bar{y}$ é a média amostral de y_t e j=1,2... A Função de Autocorrelação Parcial (PACF), que estabelece a ordem AR ao realizar regressões contra os valores defasados.

$$y_t = \hat{\phi}_{11} y_{t-1} + e_t \tag{IV.3}$$

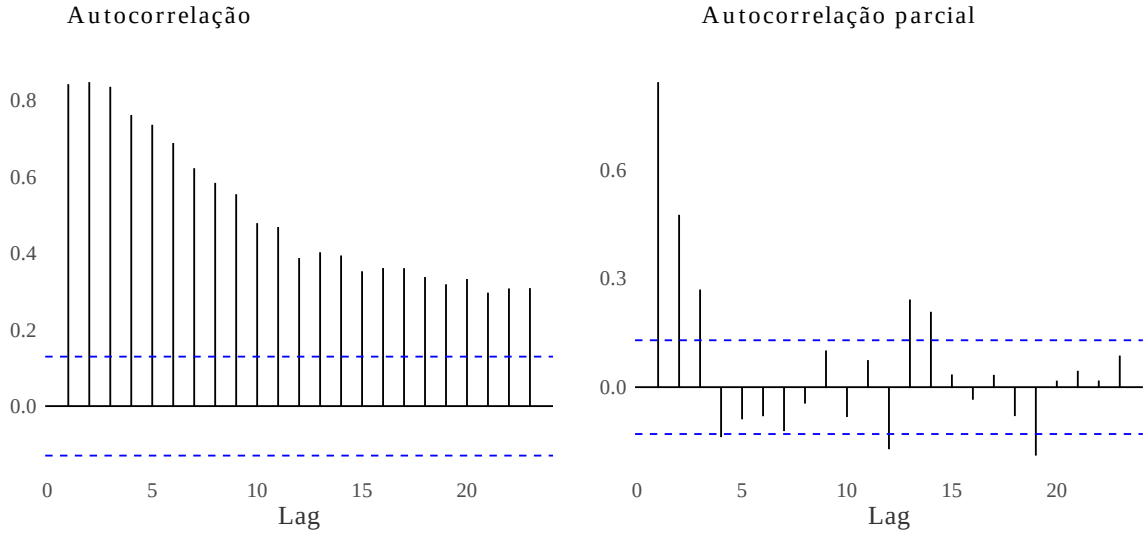
Em que $\hat{\phi}_{11}$ é a autocorrelação e autocorrelação parcial entre y_t e y_{t-1}. Posteriormente, estima-se:

$$y_t = \hat{\phi}_{11} y_{t-1} + \hat{\phi}_{22} y_{t-2} + e_t \tag{IV.4}$$

Em que $\hat{\phi}_{22}$ é a autocorrelação e autocorrelação parcial entre y_t e y_{t-2}. O processo é repetido para as s defasagens. Portanto, a PACF é $\hat{\phi}_{11}$,$\hat{\phi}_{22}$,...,$\hat{\phi}_{ss}$.

A análise gráfica da PACF sugere um processo autorregressivo. O comportamento de ordem infinita da ACF se deve à possibilidade de escrever um processo AR(p) como um MA(∞).

Figura 2 – Funções de autocorrelação



Fonte: Elaboração do autor.

Tabela 4 – Comparações entre modelos AR

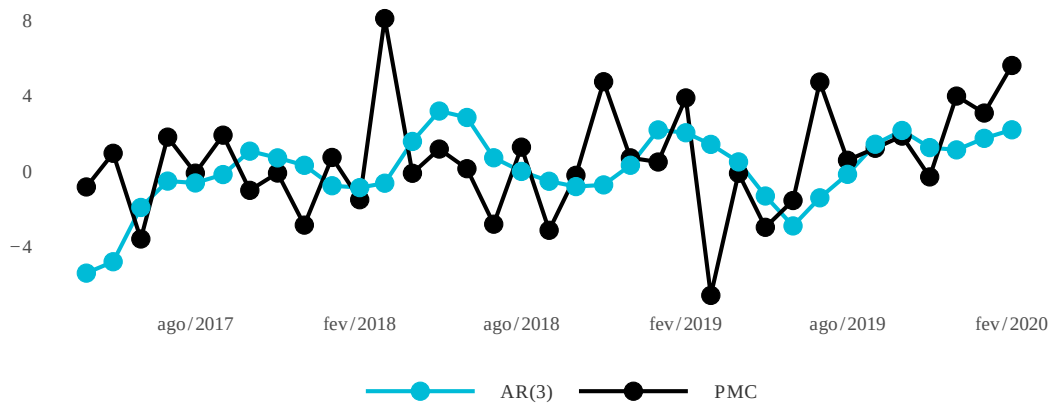
	AR(2)	AR(3)
Ljung-Box	7,471e-07	1.382e-05
AIC	428,5	415,6
BIC	438,7	429,2

Fonte: Elaboração do autor.

Foram propostos dois modelos concorrentes: AR(2) e AR(3). Em ambos os casos os erros não são autocorrelacionados segundo o teste de Ljung-Box. Os critérios de informação sugerem a escolha do modelo AR(3), em conformidade com o observado graficamente.

A partir da escolha do modelo AR(3) foram realizadas previsões para o período de 04/2017 a 02/2020. O RMSE da previsão é **3,22**.

Figura 3 – Previsão utilizando AR(3) x valor original (%)



Fonte: IBGE e cálculos do autor.

Apesar da previsão captar as principais variações e captar corretamente a tendência e o sinal na maior parte das situações, o modelo não foi capaz de ser preciso. Como se trata de um modelo que opera de apenas com as defasagens da própria PMC, é esperado que não desempenhe adequadamente em picos e vales maiores que a média, como nos meses de março de 2018 e março de 2019.

IV.2.2 CSR

Dentre os modelos multivariados, em todos o conjunto de palavras é o mesmo, assim como os ASMs utilizados. Dessa forma, a previsões com CSR ou LASSO utilizando AIC como critério de informação possuem as mesmas variáveis.

A união entre palavras-chave advindas de marcas e produtos gerou um resultado interessante, uma vez que das seis palavras-chave figuraram nas seleções por todos os ASMs, duas são produtos genéricos – tênis e papel –, três são empresas – Saraiva, Leader e Asus – e uma é um produto específico – Xbox. O gráfico abaixo demonstra o número de seleções dentre os treze ASMs utilizados.

16 palavras-chave não foram selecionadas por nenhum modelo. São elas: Gasolina, Diesel, Shell, Carrefour, Sorvete, Coca-Cola, Skol, Cerveja, Ingresso, Calçados, Vestido, Camisas, Mochila, Fogão, Ar-Condicionado, Forno, Câmera, Mesa, Cortina, Pratos, Farmácia, Livraria Cultura, Gráfica, Softwares, Computador, Teclado, Lg, Hp, Boneca, Nintendo, Lego, Barbie

O modelo *Complete Subset Regressions* apresentou o melhor desempenho dentre todos os testados, com RMSE de **2,57**, utilizando a seleção por meio do critério de informação de Akaike. A exclusão por meio da regressão utilizando LASSO também apresentou resultados próximos aos encontrados com AICc.

Na tabela abaixo é possível observar que os algoritmos que contavam com critérios de informação e LASSO desempenharam consideravelmente melhor que os advindos da exclusão por p-valor e estatística T, como indicado no artigo de Garcia et al. (2017). No gráfico é possível observar o melhor valor obtido por meio do modelo CSR com seleção por AICc. Em comparação com o obtido por meio do AR(3), o incremento de performance é notável graficamente, especialmente nos picos e vales em períodos anômalos. Em seu uso cotidiano, uma das principais vantagens práticas dos modelos de curto prazo é a contraposição da análise macroeconômica e dos modelos.

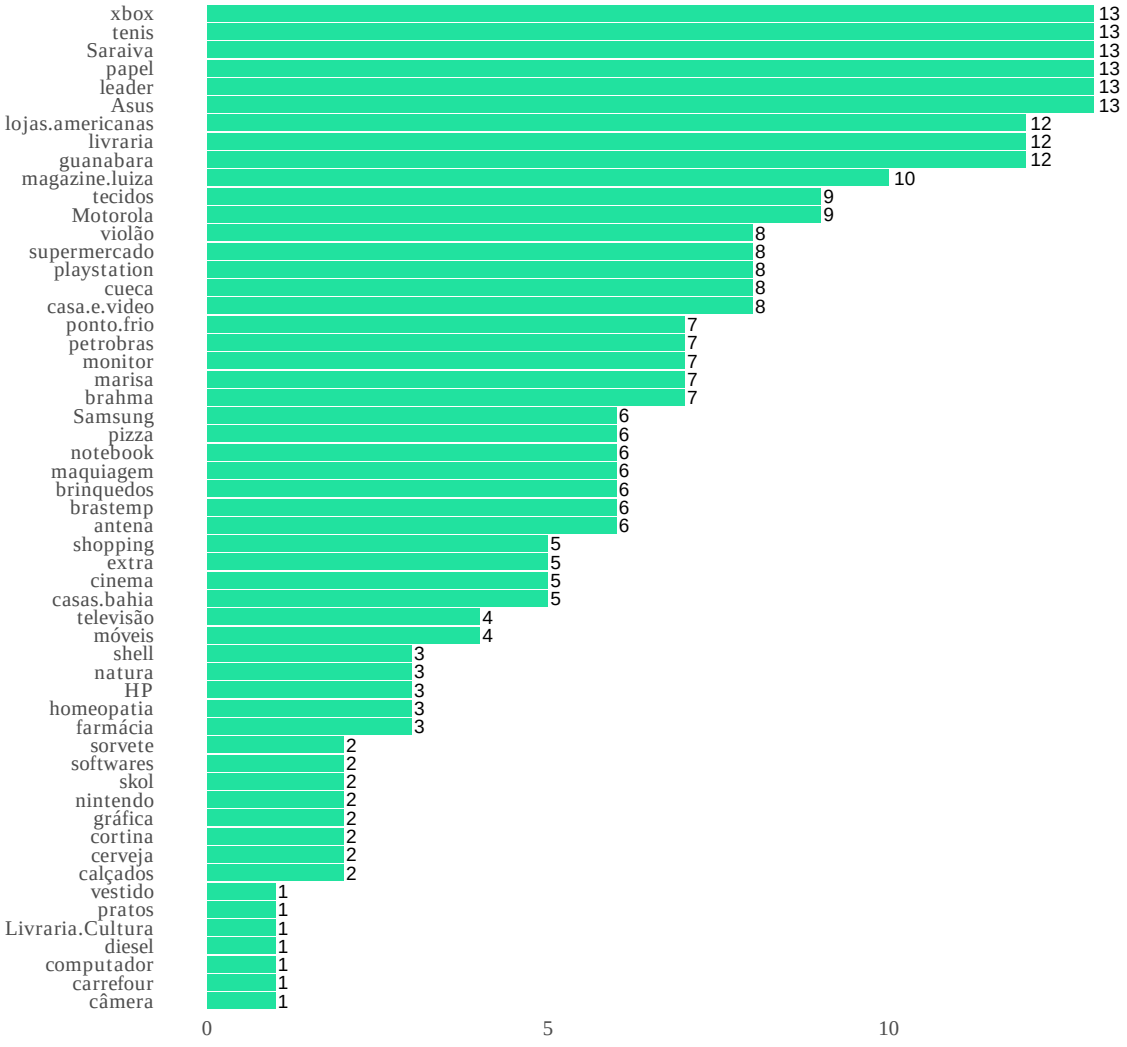
Nos modelos que fizeram uso da seleção por AIC e tstat30 não foi possível realizar a estimação para todas as combinações por uma limitação computacional. Dessa forma, o resultado na tabela abaixo em ambos os casos foi o melhor dentre as combinações que puderam ser realizadas.

Tabela 5 – RMSE das previsões utilizando CSR

Método de Seleção	**AIC**	**AICc**	**BIC**	**LASSO AIC**	**LASSO AICc**	**LASSO BIC**	**p-value 1%**
RMSE	2,91	2,57	2,8	2,85	2,69	2,66	3,1
Método de Seleção	**p-value 5%**	**tstat 10**	**tstat 15**	**tstat 20**	**tstat 25**	**tstat 30**	
RMSE	3,01	3,15	2,99	2,99	3,04	2,99	

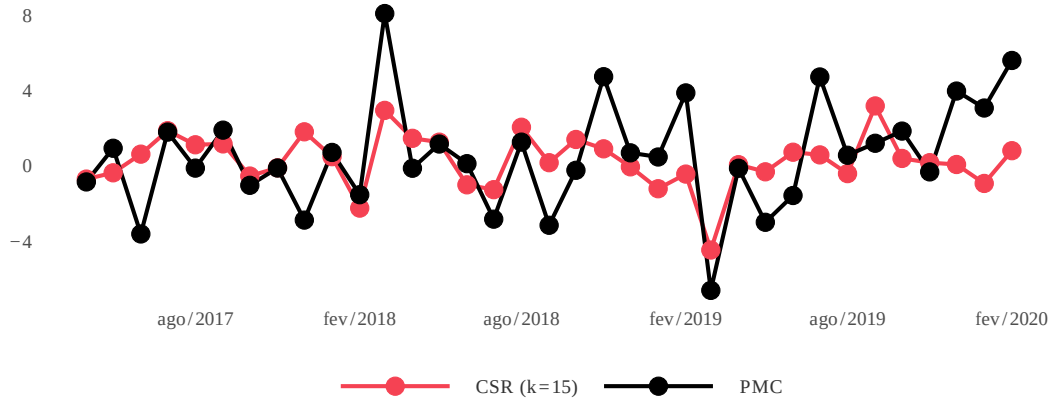
Fonte: Elaboração do autor.

Figura 4 – Número de seleções pelos ASMs



Fonte: Elaboração do autor.

Figura 5 – CSR x valor original (%)

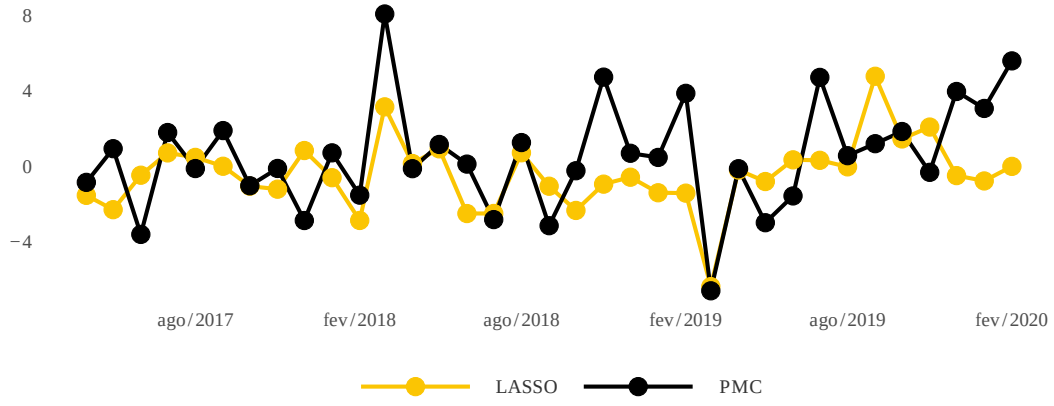


Fonte: IBGE e cálculos do autor.

IV.2.3 LASSO

Os resultados com o uso do LASSO como meio de previsão foram divididos de acordo com o λ utilizado para a previsão. Dessa forma, LASSO-AIC significa a previsão por LASSO utilizando a penalização sugerida pelo critério de informação AIC. Apesar de ser comum que os diferentes critérios de informação sugiram o mesmo grau de penalização, o que justifica, por exemplo, que o RMSE seja igual para os modelos selecionados por p-valor 1%.

Figura 6 – LASSO x valor original (%)



Fonte: IBGE e cálculos do autor.

Dentre todas as possibilidades, o menor RMSE foi obtido pelo modelo selecionado previamente por meio de um LASSO-AIC. Apesar de aparentar certa redundância, o resultado significa que partindo de um modelo selecionado por LASSO-AIC, a previsão fazendo uso de LASSO com penalização escolhida por BIC (ou LASSO AICc, que sugeriu o mesmo grau de λ), obteve RMSE igual a **2,76**.

Tabela 6 – RMSE das previsões utilizando LASSO-AIC

Método de Seleção	**AIC**	**AICc**	**BIC**	**LASSO AIC**	**LASSO AICc**	**LASSO BIC**	**p-value 1%**
RMSE	2,93	2,84	3,12	2,91	2,78	2,93	3,61
Método de Seleção	**p-value 5%**	**tstat 10**	**tstat 15**	**tstat 20**	**tstat 25**	**tstat 30**	
RMSE	3,32	3,28	3,08	3,11	3,11	3	

Fonte: Elaboração do autor.

Tabela 7 – RMSE das previsões utilizando LASSO-AICc

Método de Seleção	**AIC**	**AICc**	**BIC**	**LASSO AIC**	**LASSO AICc**	**LASSO BIC**	**p-value 1%**
RMSE	2,93	2,84	3,12	2,76	2,78	2,93	3,61
Método de Seleção	**p-value 5%**	**tstat 10**	**tstat 15**	**tstat 20**	**tstat 25**	**tstat 30**	
RMSE	3,34	3,32	3,04	3,06	3,13	2,97	

Fonte: Elaboração do autor.

Tabela 8 – RMSE das previsões LASSO-BIC

Método de Seleção	**AIC**	**AICc**	**BIC**	**LASSO AIC**	**LASSO AICc**	**LASSO BIC**	**p-value 1%**
RMSE	2,93	2,84	3,12	2,76	2,78	2,93	3,61
Método de Seleção	**p-value 5%**	**tstat 10**	**tstat 15**	**tstat 20**	**tstat 25**	**tstat 30**	
RMSE	3,34	3,32	3,04	3,06	3,13	2,97	

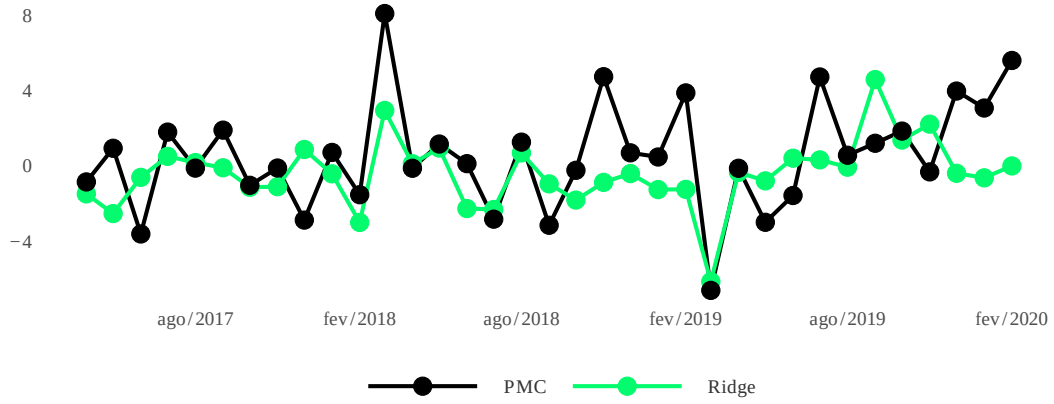
Fonte: Elaboração do autor.

IV.2.4 Regressão Ridge

A regressão Ridge apresentou menor RMSE igual a **2,71**, proveniente da seleção realizada por LASSO-AICc. Os ASMs de critérios de informação e LASSO obtiveram os melhores resultados.

A Regressão Ridge também conta com diferentes penalizações com grau λ. No entanto, em todos os cenários os critérios de informação sugeriram o uso da mesma

Figura 7 – Ridge x valor original (%)



Fonte: IBGE e cálculos do autor.

penalização, possibilitando a exibição dos resultados em uma tabela. Apenas o critério de informação de Hannan-Quinn, não abordado neste trabalho, sugeriu penalizações levemente diferentes das obtidas por AIC, AICc e BIC.

Tabela 9 – RMSE das previsões utilizando Regressão Ridge

Método de Seleção	**AIC**	**AICc**	**BIC**	**LASSO AIC**	**LASSO AICc**	**LASSO BIC**	**p-value 1%**
RMSE	2,83	2,76	3,01	2,78	2,71	2,88	3,44
Método de Seleção	**p-value 5%**	**tstat 10**	**tstat 15**	**tstat 20**	**tstat 25**	**tstat 30**	
RMSE	3,19	3,27	3,05	3,08	3,2	3	

Fonte: Elaboração do autor.

V CONSIDERAÇÕES FINAIS

Os resultados obtidos demonstram a viabilidade do uso do *Google Trends* como fonte de dados para previsão. As 69 variáveis utilizadas foram obtidas por meio de conhecimento e informação do autor, o que não necessariamente representam o melhor conjunto de palavras-chave para a predição. Em uma finalidade de uso aplicado, é possível que se encontre outros substantivos, comuns ou próprios, que sejam *proxies* mais eficientes que as utilizadas neste trabalho.

Além disso, é preciso ressaltar a importância do uso dos algoritmos de seleção de modelos, que se provaram decisivos para a qualidade preditiva. Se analisarmos, por exemplo, os resultados utilizando a Regressão Ridge como modelo, enquanto o melhor resultado para o RMSE, 2,71 utilizando LASSO AICc como ASM, erra substancialmente menos que um modelo AR, o pior RMSE, 3,44 com p-valor 1% como ASM, apresenta desempenho inferior ao obtido pelo modelo AR, o que pode ser decisivo para a conclusão de um trabalho.

Tabela 10 – Resumo dos resultados

Modelos	**AR(3)**	**CSR**	**LASSO**	**Regressão Ridge**
RMSE	3,22	2,57	2,76	2,71

Fonte: Elaboração do autor.

O modelo *Complete Subset Regressions* mostrou-se bastante promissor em prever com menos erros que o AR, presente nos livros-texto desde o fim da década de 1970 e LASSO e Regressão Ridge, criados no fim da década de 1990. O modelo, descrito por Elliott et al. (2013) é eficiente e de fácil compreensão para alunos iniciantes em Econometria.

No entanto, é preciso destacar que, apesar do desempenho levemente superior à Regressão Ridge, o custo computacional da estimação do CSR deve ser considerado, sobretudo quando o conjunto de variáveis selecionadas é consideravelmente grande. Enquanto os resultados para o LASSO e Regressão Ridge puderam ser gerados em poucos minutos, o uso do CSR levou dias de computação. Em termos práticos, é possível que o uso da Regressão Ridge seja mais conveniente, posto que seu resultado é instantaneamente visualizável, ainda que tenha apresentado desempenho pior neste trabalho.

Nesse sentido, surge a necessidade de criação de um pacote na comunidade de R com os meios mais avançados de programação funcional e em paralelo a fim de minimizar a não praticidade do uso do CSR de forma cotidiana. Este é um projeto longo, complexo e já em desenvolvimento.

Por fim, a versatilidade do *Google Trends* cria a possibilidade de usar a ferramenta não só como meio de prever o comércio, mas como um indicador em si. É viável, após um longo sistema de adoção de ponderações, a criação de um índice do comércio virtual, por exemplo, que possua a intenção de observar as variações nas pesquisas por vestuário ou eletrodomésticos.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716–723, 1974. Citado na página 14.

BOX, G.; GWILYM, M.; JENKINS, G. R. *Time series analysis: Forecasting and control*. 3rd. ed. [S.l.]: Prentice Hall, 1994. ISBN 0130607746,9780130607744. Citado 2 vezes nas páginas 18 e 24.

CASTLE, J. L.; QIN, X.; REED, W. R. How to pick the best regression equation: A review and comparison of model selection algorithms. *Working Papers in Economics*, 2009. Citado na página 14.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, v. 74, n. 366, p. 427–431, 1979. Citado na página 22.

ELLIOTT, G.; GARGANO, A.; TIMMERMANN, A. Complete subset regressions. *Journal of Econometrics*, v. 177, n. 17, p. 357–373, 2013. Citado 2 vezes nas páginas 20 e 31.

GARCIA, M. G.; MEDEIROS, M. C.; VASCONCELOS, G. F. Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, v. 33, n. 3, p. 679–693, 2017. Citado 2 vezes nas páginas 16 e 26.

HURVICH, C. M.; TSAI, C.-L. Regression and time series model selection in small samples. *Biometrika*, v. 76, n. 2, p. 297–307, 1989. Citado na página 14.

POSKITT, D. S.; TREMAYNE, A. R. Determining a portfolio of linear time series models. *Biometrika*, v. 74, p. 125–137, 1987. Citado na página 15.

SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics*, v. 6, n. 2, p. 461–464, 1978. Citado na página 14.

SCOTT, S. L.; VARIAN, H. R. Bayesian variable selection for nowcasting economic time series. *NBER Working Paper*, n. 19567, 2013. Citado 2 vezes nas páginas 10 e 14.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996. Citado na página 20.

Anexos

ANEXO A – TESTE *AUGMENTED DICKEY-FULLER* NAS VARIÁVEIS DO *GOOGLE TRENDS*

Tabela 11 – Valores críticos para o teste *Augmented Dickey-Fuller*

Gasolina	Diesel	Petrobras	Shell	Supermercado	Extra
-8,07	-7,27	-4,92	-6,59	-4,88	-2,87

Carrefour	Guanabara	Pizza	Sorvete	Coca.cola	Skol
-6,07	-4,40	-8,51	-8,23	-6,77	-6,56

Cerveja	Shopping	Cinema	Ingresso	Calçados	Vestido
-4,64	-4,76	-6,11	-6,46	-4	-5,93

Camisas	Tecidos	Uniforme	Moda.feminina	Cueca	Tenis
-5,66	-9,81	-5,57	-6,64	-8,92	-4,63

Mochila	Marisa	Móveis	Geladeira	Televisão	Ar.condicionado
-8,27	-5,28	-7,03	-5,99	-4,50	-6,04

Forno	Câmera	Mesa	Violão	Casas.bahia	Ponto.frio
-6,29	-7,06	-5,72	-5,35	-3,41	-3,61

Casa.e.video	Brastemp	Consul	Lojas Americanas	Farmácia	Homeopatia
-3,95	-7,35	-7,29	-5,08	-4,61	-8,47

Avon	Natura	Softwares	Computador	Teclado	Monitor
-3,91	-6,47	-5,22	-3,76	-5,70	-8,10

Asus	Lg	Samsung	Motorola	Hp	Kalunga
-3,73	-4,16	-4,22	-3,56	-4,37	-5,24

Notebook	Antena	Brinquedos	Boneca	Playstation	Xbox
-6,56	-5,31	-7,85	-6,93	-4,46	-4,98

Nintendo	Barbie
-5,35	-3,10

Fonte: Elaboração do autor.

ÍNDICE