INSPERDATA

BOLETIM DE PROJETOS

2023.1

Conteúdo

Carta ao leitor

INSPERDATA

Passe Livre e Voto: como a mobilidade pode influenciar o comparecimento eleitoral

Working paper

Pesquisadores: Gustavo Theil, Fábio Cantom
Orientado por: Adriano Borges Costa e Bianca Tavolari

Resumo

Nas eleições de 2022, cerca de 400 municípios adotaram o passe livre para os modais de transporte público no dia da eleição com o objetivo de democratizar o acesso ao voto. A literatura e teoria microeconômicas indicam que, caso uma medida de redução de "custos de votar" seja implementada de maneira eficiente, é esperado um efeito positivo no comparecimento. Por outro lado, caso seja implementada com ineficiência, pode haver efeitos adversos, que se configuram como um *moral hazard*. Os resultados empíricos deste estudo validam uma série de resultados teóricos discutidos, mas se demonstraram inconclusivos na identificação da magnitude do efeito do passe livre no comparecimento eleitoral.

Palavras-chave: Comparecimento eleitoral, Microeconomia, Mobilidade, Ciência Política

1 Introdução

Nas eleições brasileiras de 2022, houve um grande movimento conhecido como Passe Livre pela Democracia[1], que lutou pela adoção do transporte público gratuito no dia da votação, para que mais pessoas votassem. O movimento enfrentou bastante resistência política e foi grande alvo de discussão. Inclusive, a coligação Pelo Bem do Brasil, composta por PL – o partido do ex-presidente da República, Bolsonaro –, Republicanos e Progressistas, solicitou ao TSE que tomasse providências para impedir a concessão do passe livre[2].

Instituir o passe livre tem seu mérito divido em dois aspectos, o primeiro deles sendo normativo, considerando a mobilidade como um direito e a dificuldade de uma pessoa se locomover à urna como uma possível barreira monetária ao exercício democrático de votar. O segundo aspecto se refere aos incentivos econômicos do voto, sendo que um estímulo monetário influencia a decisão das pessoas e pode ser decisivo para "convencer" um cidadão a votar. A literatura microeconômica indica que, mesmo sendo um pequeno incentivo, é esperado um aumento também pequeno no comparecimento.

> "À primeira vista, todos esses custos podem parecer triviais, e os viéses na capacidade de suportá-los parecem irrelevantes. No entanto, os retornos do ato de votar geralmente são tão baixos que variações mínimas em seu custo podem ter efeitos enormes na distribuição

[1] Mais informações disponíveis em `https://www.passelivrepelademocracia.org`

[2] Processo disponível em `https://static.poder360.com.br/2022/10/Pedido-PL-TSE-transporte-gratuito-30-09-22.pdf`

> do poder político. Esse fato explica por que práticas tão simples como realizar eleições em feriados, manter os locais de votação abertos até mais tarde, revogar pequenos impostos eleitorais[3] e oferecer transporte gratuito para os locais de votação podem afetar de maneira marcante os resultados das eleições.", Downs (1957), página 266, traduzido pelo autor.

Pesquisas empíricas como a conduzida por Haspel e Knotts (2005) demonstram que a mobilidade é um elemento crucial para o comparecimento eleitoral. No estudo, identificam que quanto mais longe a residência de uma pessoa se encontra de sua respectiva urna, menor é sua probabilidade de votar. Os autores também concluem que acesso a um veículo atenua esse efeito da distância. Utilizando esse arsenal e outras referências da literatura, outro estudo conduzido por Konishi e Murata (2010) simula quanto seria o comparecimento eleitoral em uma cidade no Japão, dados diferentes valores para o número de urnas, considerando simulações com e sem acesso a ônibus. O artigo conclui que oferecer ônibus de graça pode ter impacto significativo no dispêndio de votar na cidade japonesa estudada.

Entretanto, adotar o passe livre pode ser um tanto quanto custoso e da maneira como foi implementada na maioria dos municípios em 2022, não houve garantias de que as pessoas que aproveitaram o subsídio do passe livre o fizeram para votar ou para realizar outras atividades. Pereira et al. (2023) conduziram um estudo na tentativa de identificar essa hipótese para as eleições brasileiras de 2022, e seus resultados serão discutidos no tópico 3.1. Ademais, o conturbado cenário político complicou a execução da medida, visto que a comunicação dos governos municipais pode não ter sido eficiente e possivelmente a população não estava plenamente ciente da adoção do passe livre e como ele funcionaria.

O objetivo deste estudo é identificar se a adoção do passe livre causa uma redução na abstenção. Na seção 2 é discutido o que é estabelecido na teoria de ciência política com um olhar microeconômico, bem como é feita uma adaptação dos modelos da literatura para o caso do passe livre. Na seção 3 é feita uma tentativa empírica de medir o efeito do passe livre nas eleições de 2022 a partir de um *diff-in-diff* pareado com efeitos fixos e um *event study* para testes de robustez. Por fim, é feita uma discussão dos, resultados principais e possíveis agendas de pesquisa para futuros estudos.

2 Modelagem Teórica

2.1 Revisão Literária

A discussão sobre abstenção é extensiva no campo da economia política e diversos autores contribuíram com o debate de seus determinantes. Downs (1957) é um dos precursores do modelo microeconômico que é continuado por uma série de autores, que serão discutidos a seguir. Sua premissa é simples: os eleitores buscam maximizar suas utilidades, que são definidas pela Equação 1. A partir disso, é considerado racional o eleitor que vota quando sua utilidade esperada é maior que zero e, caso contrário, se abstém.

$$\mathbb{E}(U_i) = B \cdot P - C \tag{1}$$

Na equação, $\mathbb{E}(U_i)$ representa a utilidade esperada do cidadão i apto, ao votar. Esta utilidade é impactada negativamente por C, o custo de votar, que se refere aos dispêndios de tempo, recursos e custos de oportunidade. Exemplos de fatores que podem afetar o custo são longas distâncias até as

[3] No período que foi escrito este livro, havia um imposto para votar nos Estados Unidos, medida que hoje é considerada uma ferramenta discriminatória que viola os direitos civis.

urnas, filas e demora para votar, o custo financeiro do transporte, fortes chuvas, etc. Os custos de não votar, como a necessidade de justificar o voto e pagar uma multa no caso brasileiro, são considerados nesse custo também afetando positivamente a utilidade de votar. A variável B representa o benefício líquido de seu candidato favorito, caso seja eleito. Na análise de Downs, o benefício é dado por quanto o eleitor julga que sua primeira opção de candidato (A) seja melhor que a segunda opção (B), caso seja eleita: $B_{it} = \mathbb{E}(U^A_{t+1}) - \mathbb{E}(U^B_{t+1})$. P representa a probabilidade de seu voto ser decisivo. O voto decisivo é considerado aquele que desempata o resultado da eleição, caso o número de votantes seja ímpar, ou defina o resultado da votação, caso seja par. Em termos práticos, o benefício do candidato preferido apenas será importante para o eleitor, caso o seu voto seja a causa dele ser eleito.

O modelo apresenta uma série de complicações, que são inclusive discutidas de forma descritiva por Downs. Um eleitor indiferente entre os candidatos neste modelo apresenta benefício zero e não votaria para qualquer valor de C positivo, então não seria racional votar branco. Além disso, em eleições de grande escala, como é o caso das eleições federais que apresentam mais de 150 milhões de votantes, a probabilidade do voto ser decisivo é muito próxima a zero. Nesses casos, caso o custo seja positivo, a utilidade do voto seria sempre negativa e ninguém votaria.

Para endereçar estes problemas, Riker e Ordeshook (1968) introduzem, entre outras contribuições, mais um componente autônomo da utilidade[4] D. Este se refere ao dever cívico de votar, que não está relacionado à probabilidade do voto ser decisivo. O autor argumenta que há um ganho de utilidade ao contribuir com o sistema democrático, bem como pode haver o sentimento de arrependimento caso o sujeito se abstenha. Além disso, considera votar um ato político, relacionado à tradição. Por outro lado, reconhece que para a maior parte das pessoas $D < C$ e, portanto, ainda não seria racional votar em grandes eleições.

$$\mathbb{E}(U_i) = B \cdot P - C + D \tag{2}$$

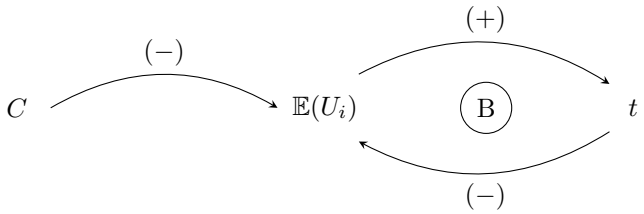
Os pesquisadores da escolha racional geralmente são muito focados no interesse individual das escolhas, mas no final do século XX uma série de estudos e experimentos passaram a questionar e contribuir com evidências de comportamentos racionalmente altruístas. Alguns modelos de escolha racional foram revisitados e Fowler (2006) e Edlin, Gelman e Kaplan (2007) são autores que incorporam este componente no modelo de forma a dividir o benefício líquido em duas partes (Equação 3). A variável B_p se refere ao benefício líquido que o candidato proporciona ao indivíduo, enquanto B_s se refere ao benefício social. A questão é que enquanto B_p beneficia apenas o sujeito, B_s beneficia a todos no país na visão do eleitor, então este benefício é multiplicado pela população N. Entretanto, o benefício social não necessariamente apresenta a mesma relevância ao eleitor, se comparado ao individual e α controla por este fator, de forma que para a maioria das pessoas $0 < \alpha < 1$.

$$B_i = B_p + \alpha \cdot NB_s \tag{3}$$

Outra contribuição de Edlin, Gelman e Kaplan (2007) foi de consolidar o que se considera a probabilidade do voto ser decisivo (Equação 4). O coeficiente K representa o grau de competitividade da eleição e quanto mais próximas forem os votos do candidato A em relação ao B, maior é a probabilidade do voto

[4]É considerado um componente autônomo aquele que não está relacionado à abstenção. Considerando que a probabilidade do voto ser decisivo é dada por $P = 1/v$, no qual v é o número de votantes, quanto maior for a abstenção, maior é a probabilidade do voto ser decisivo

Figura 1: Paradoxo de Downs: ciclo de balanceamento.



ser decisivo. Em uma votação em que a diferença de votos entre os candidatos está em torno de $10p.p.$, $K = 5$, enquanto uma diferença de $\pm 2p.p.$ resulta em um $K = 25$. Portanto, a probabilidade do voto ser decisivo P é dada pela competitividade dividida pelo número de votantes, que equivale ao total de eleitores elegíveis E vezes taxa de comparecimento t.

$$\begin{aligned} K &= \frac{1}{\log(V_A) - \log(V_B)} \\ P &= K/(tE) \end{aligned} \tag{4}$$

Ao substituir as equações 3 e 4 em 2, têm-se a equação em sua forma final (Equação 5). É interessante observar que na medida em que o número de votantes V aumenta, o benefício pessoal se aproxima a zero, enquanto o benefício social não, pois quando o número de votantes aumenta, o número de pessoas beneficiadas N também aumenta. Considerando os fatores agregados, as ideias primordiais de Downs (1957) tornam-se mais robustas, mas o esqueleto se mantém igual: $\alpha N B_s$ representam o benefício líquido do voto, $\frac{K}{tE}$ é a probabilidade do voto ser decisivo e $C + D$ é o componente autônomo do voto.

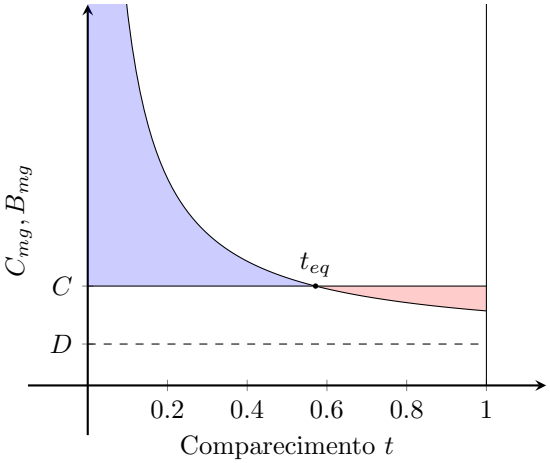
$$\begin{aligned} \mathbb{E}(U_i) &= (B_p + \alpha \cdot NB_s) \cdot \frac{K}{tE} - C + D \\ &= \left(\frac{B_p}{\cancel{V}} + \frac{\alpha \cdot NB_s}{tE} \right) \cdot K - C + D \\ &= \underbrace{\frac{\alpha N B_s}{1}}_{B} \cdot \underbrace{\frac{K}{tE}}_{P} - \underbrace{C + D}_{A} \end{aligned} \tag{5}$$

Um desdobramento peculiar desse modelo é uma espécie de paradoxo, que inclusive foi discutido por Downs (1957). Nessa relação paradoxal, quanto maior é a utilidade de votar $\uparrow \mathbb{E}(U_i)$, mais pessoas votam $\uparrow t$, o que diminui a utilidade de votar $\downarrow \mathbb{E}(U_i)$, reduzindo o número de pessoas que votam $\downarrow t$. Em outras palavras, a relação entre a utilidade de voto e comparecimento é um ciclo de balanceamento. Quando uma variável exógena, como o custo, altera a utilidade do voto, há um efeito multiplicador que atenua a mudança no comparecimento.

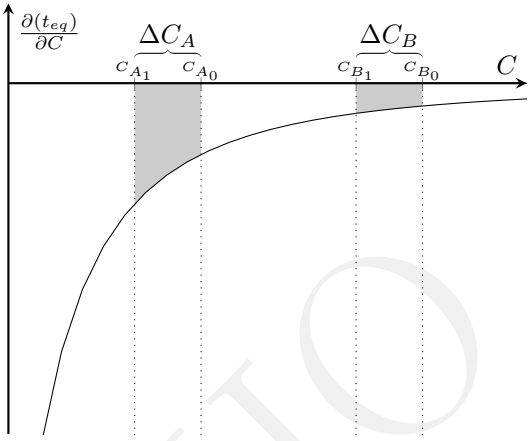
2.2 Extensão do modelo e resultados teóricos

A partir do modelo apresentado é possível intuir o que aconteceria a nível agregado dada uma mudança em alguma das variáveis – no caso deste estudo, a variável de interesse é o custo de votar. Para tanto, é necessário estabelecer um sistema que determine a abstenção de equilíbrio – a partir dele será possível analisar como o equilíbrio é deslocado. O equilíbrio de comparecimento t_{eq} é dado quando todos os cidadãos elegíveis escolhem de forma a maximizarem suas utilidades, ou seja, aqueles que apresentam

Figura 2: Visualização do modelo microeconômico



(a) Oferta e demanda de votos



(b) Derivada parcial do comparecimento de equilíbrio em relação ao custo

utilidade esperada maior que zero votam, caso contrário, se abstêm. Esse ponto se encontra na intersecção entre a função de benefício marginal do voto e custo marginal do voto (Equação 6).

$$\begin{aligned} C_{mg} &= B_{mg} \\ C &= \left(\frac{\alpha \cdot B_s \cdot N \cdot K}{E \cdot t^e_{eq}}\right) + D \\ t^e_{eq} &= \left(\frac{\alpha \cdot B_s \cdot N \cdot K}{E \cdot (C - D)}\right) \end{aligned} \tag{6}$$

Na figura 2a é possível observar para qualquer nível de comparecimento t menor do que o comparecimento de equilíbrio t_{eq}, haveria eleitores que apresentariam utilidade esperada positiva de votar, o que torna a ação de votar uma escolha racional e, portanto, mais pessoas votariam deslocando o comparecimento ao equilíbrio. De maneira semelhante, caso fosse observado um comparecimento acima do equilíbrio, eleitores apresentam utilidade esperada de votar negativa, sendo racional que se abstenham. Com este modelo, é possível identificar de que forma o equilíbrio se desloca, dada uma variação em alguma das variáveis.

$$\frac{\partial(t^e_{eq})}{\partial C} = -\left(\frac{\alpha \cdot B_s \cdot N \cdot K}{E \cdot (C - D)^2}\right) \tag{7}$$

A equação 7 apresenta a derivada parcial do equilíbrio de comparecimento em relação ao custo do voto. Na medida em que o custo aumenta, *ceteris paribus*, o comparecimento reduz. Na figura 2b é possível observar qual é o impacto de um ΔC no comparecimento, visto que a área pintada abaixo do gráfico representa a quantidade de pessoas que foram "convencidas" a votar por conta da redução do custo e, caso houvesse um aumento, representaria a quantidade de pessoas que desistiram de votar. Um ponto importante é que um ΔC impacta de maneira heterogênea localidades com níveis de comparecimento diferentes: em municípios cujo comparecimento é baixo, um ΔC surte efeito diminuto, enquanto em um município com comparecimento alto, um ΔC pequeno pode causar diferenças significativas. Isso

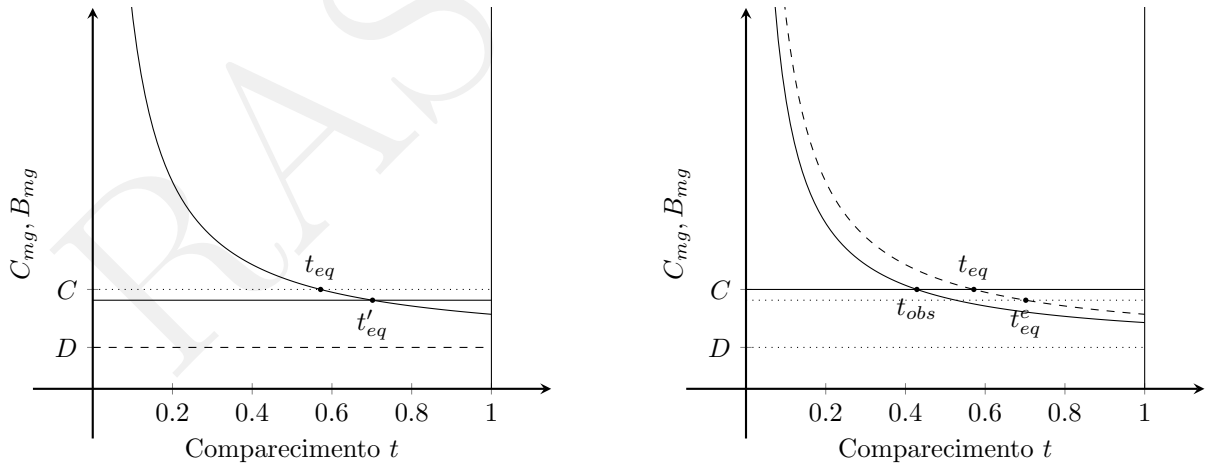
ocorre, pois a função de benefício marginal é recíproca[5] e assintótica em D, ou seja, na medida em que o comparecimento aumenta, o benefício marginal se aproxima de D, o dever cívico do voto. Nesse sentido, quando o comparecimento é muito alto, o principal fator que define o equilíbrio é a diferença entre C e D.

Entretanto, um fator que não pode ser ignorado é de que estas variáveis todas se referem às expectativas das pessoas. A competitividade da eleição, por exemplo, é dada pela percepção dos eleitores com base nas pesquisas eleitorais, conteúdo da mídia e interações com seus conhecidos. Nesse sentido, o comparecimento de equilíbrio é dado por expectativas prévias à eleição, que não necessariamente se confirmam no dia da votação. Na equação 8, tem-se que o comparecimento esperado de equilíbrio é dado pelo verdadeiro equilíbrio t_{eq}, mais um erro de expectativas oriundo de assimetria de informações ε. Substituindo a equação 8 em 6, incorpora-se um componente de expectativas na estática comparativa.

$$t^e_{eq} = t_{eq} + \varepsilon \tag{8}$$

Este componente permite analisar o efeito que a assimetria de informações pode causar. Na figura 3a é possível observar como o comparecimento de equilíbrio se desloca quando há uma redução no custo sem que haja assimetria de informações. Entretanto, na figura 3b, é hipoteticamente anunciada uma redução no custo para todos os eleitores, mas no dia da eleição esta redução não ocorre, sem que as pessoas fossem avisadas. Neste caso, há uma assimetria de informações e os eleitores antecipam um número maior de eleitores por causa da redução do custo, que reduz a utilidade de seus votos via menor probabilidade dele ser decisivo, criando desincentivos para o voto. Dessa forma, há um aumento em ε, que desloca a curva de benefício marginal para baixo e se observa o comparecimento na intersecção da curva de custo marginal, que efetivamente não se desloca, com a nova curva de benefício marginal. Este equilíbrio, inclusive, é de um comparecimento menor do que tinha se não fosse anunciada a redução no custo.

Figura 3: Estática comparativa e *moral hazard*



(a) Sem assimetria de informações

(b) Com assimetria de informações

[5] A função do benefício marginal ($B_{mg} = \frac{Z}{t_{eq}} + D$), na qual Z representa todas as variáveis que multiplicam t_{eq} é recíproca, já que depende inversamente do comparecimento de equilíbrio

Portanto, do ponto de vista teórico, uma medida de redução do custo de votar com a intenção de aumentar o comparecimento, por conta da assimetria de informações pode acabar por diminuí-lo. Esse fenômeno se configura como um *moral hazard*. No caso do passe livre, um município pode cair nessa armadilha caso adote o passe livre, mas as pessoas não efetivamente utilizem o transporte público para votar ou a redução no custo não convença o número esperado de pessoas a votar. Apesar desse risco existir, ainda é possível que mesmo com assimetria de informações o comparecimento aumente, dependendo do grau da assimetria.

3 Modelagem Empírica

3.1 Revisão da literatura empírica

Há apenas um estudo que analisa o impacto do passe livre adotado em 2022 na abstenção, conduzido por Pereira et al. 2023. O estudo traz uma contribuição inovadora utilizando dados de celular para identificar padrões de mobilidade no dia da eleição, com base na localização do usuário. Segundo os autores:

> "Não encontramos qualquer efeito da gratuidade no transporte público sobre o comparecimento às urnas ou sobre os resultados eleitorais, mas encontramos um efeito positivo, entre 7,2% e 17,5% de aumento, nos níveis de mobilidade no dia das eleições. Embora a redução dos custos monetários de transporte possa melhorar o acesso das pessoas aos locais de votação, nossos resultados sugerem que apenas políticas de redução desses custos não são suficientes para aumentar o comparecimento dos eleitores."

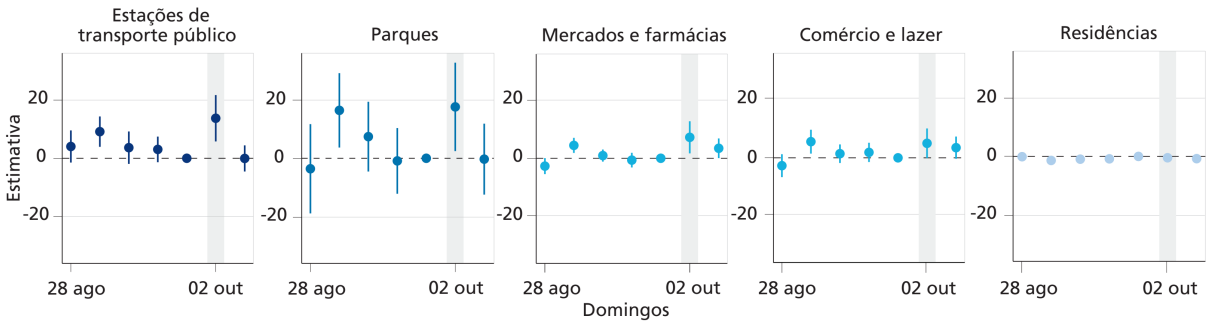
> "Assim, os governos podem justificar a adoção de políticas de isenção de tarifas por motivos normativos [...], mas sabendo que tais políticas podem não efetivamente trazer mais eleitores às urnas."

Entretanto, há algumas pontos a serem levantados em relação à interpretação dos resultados. O primeiro comentário se refere à afirmação sobre o aumento nos níveis de mobilidade. Os autores identificam que houve aumento de mobilidade nos municípios que adotaram o passe livre em relação aos que não adotaram – efeito não encontrado para aumento do comparecimento –, mas no *event study* apresentado (Figura 4), entre os quatro domingos anteriores, observou-se efeito significativo em um deles, uma evidência contrária à hipótese de identificação.

Em segundo, diferentemente do *event study* para identificar efeito na abstenção, este *event study* não compara o domingo de eleição com outros domingos de eleição, mas sim, com outros domingos do mês anterior. Isso é uma limitação, visto que o domingo de eleição funciona de maneira muito diferente de outros domingos e possivelmente os municípios dos grupos diferentes apresentam distintas tradições de mobilidade no dia da eleição. Nesse sentido, inclusive, não é ideal comparar os resultados de mudança na mobilidade com mudança na abstenção, já que foram utilizados *designs* diferentes no *event study*.

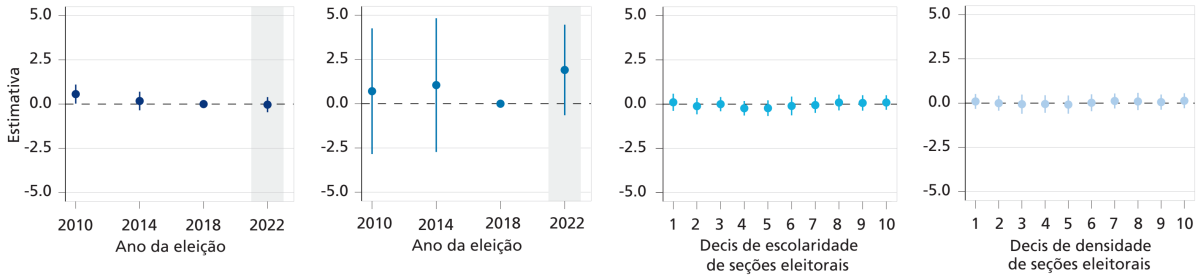
Em relação a conclusão do artigo de que não houve efeito do passe livre na abstenção, faltou uma discussão sobre a hipótese de identificação. No gráfico 5A, é mensurado se nos grupos de tratamento há um aumento na diferença de comparecimento do segundo e primeiro turnos em relação ao grupo de controle. Todavia, há apenas as eleições de 2010 e 2014 para fazer testes de robustez e é identificado um efeito placebo em 2010, evidência de que a hipótese de identifacação é inválida. Portanto, o efeito da medida pode ser considerado inconclusivo ao invés de nulo.

Figura 4: Mudança nos níveis de mobilidade nos municípios tratados nos domingos anterior e posterior e no dia do primeiro turno da eleição de 2022 em comparação aos níveis de mobilidade dos municípios do grupo de controle



Fonte: Pereira et al. 2023

Figura 5: Efeitos da política de passe livre no transporte público sobre: comparecimento eleitoral (A), parcela de votos para o PT (B), comparecimento em seções eleitorais com diferentes níveis socioeconômicos (C) e comparecimento em seções eleitorais em áreas com menor e maior densidade populacional (D)



Fonte: Pereira et al. 2023

De maneira geral, é necessário tomar cuidado com a narrativa de que o passe livre aumentou a mobilidade, mas as pessoas ao invés de votar foram aos parques, tendo em vistas as limitações na hipótese de identificação e dificuldade de comparar os diferentes designs de *event study*.

3.2 Dados

O passe livre foi uma medida adotada em nível municipal, então os dados foram coletados também nessa escala. Para aumentar a robustez da análise, os dados foram organizados em painel, com todas as eleições nos últimos 20 anos, desde 2002. Para fins de comparação de comparáveis, foram apenas analisadas os anos de eleições presidenciais. Algumas variáveis são diferentes para o primeiro e segundo turno, como a abstenção e passe livre, mas outras como o PIB *per capita* não. Os dados foram coletados para todos os municípios brasileiros. Na tabela 1 é possível observar os níveis de abstenção para os municípios de tratamento e de controle para cada turno, bem como o número de municípios em cada categoria.

Smets e Van Ham (2013) conduziram uma meta análise de 90 estudos empíricos que foram publicados

Tabela 1: Abstenção e tamanho amostral para cada grupo em 2022

		Primeiro Turno			Segundo Turno		
	Passe Livre	Média	Desvio	N	Média	Desvio	N
Abstenção	Não Houve (contole)	0.21	0.05	5437	0.21	0.05	5148
	Houve (tratamento)	0.20	0.03	81	0.20	0.03	370

em grandes jornais durante a década de 2000 que estudaram a abstenção de votos. Como variáveis a nível individual que foram consideradas relevantes, estão a idade, escolaridade e renda. Os dados do TSE são segmentados por faixa etária, gênero e escolaridade, mas o dado da escolaridade não é de qualidade, visto que não é atualizado regularmente - apenas quando há uma atualização do título de eleitor. Para contornar este problema, seria ideal analisar pela PNAD a média de anos de estudo na população, mas dessa forma não haveria cobertura de todos os municípios do Brasil. A *proxy* adotada foi a nota do IDEB do município, um índice de educação básica calculado com base no Censo Escolar e desempenho no Sistema de Avaliação da Educação Básica (Saeb). Quanto aos dados de renda, os dados de PIB *per capita* a nível municipal foram coletados do IBGE.

Para fazer o balanceamento dos grupos, foram utilizadas algumas variáveis do censo de 2010, descritas na tabela 3, que se encontra no Anexo (Seção 6). Apesar do censo estar bastante desatualizado, as variáveis se referem a fatores estruturais dos municípios, que não mudam muito ao longo do tempo. Além das variáveis do senso, foi considerado o valor médio das variáveis de controle do modelo principal, considarando dados até 2018.

Em relação às variáveis do modelo microeconômico, nem todos os componentes são observáveis. O α e B_s, por exemplo, são extremamente subjetivos, o que os torna praticamente imensuráveis. A literatura identificou alguns fatores que podem contribuir com o dever cívico do voto, D, como a participação da população em organizações políticas, filiação partidária, número de funcionários públicos, etc. A maioria desses fatores são invariantes no tempo e são relacionados à cultura do município. Em relação ao custo de votar, muitos componentes estão envolvidos. Entre eles, o custo do transporte, a distância às urnas, demora e filas, chuvas fortes, etc. Pela dificuldade de mensurar esses fatores[6], foi adotada apenas uma variável mensurável, que é o número de eleitores por urna - pode indicar se há grandes filas. Já os componentes K, E, N, t são obtidos ou calculados através dos dados do TSE e IBGE. A descrição mais detalhada das variáveis adotadas se encontra na tabela 4, no Anexo (Seção 6).

Segundo um levantamento do IDEC, no Brasil, 52 municípios apresentam transporte público gratuito, que se configurariam como municípios que receberam tratamento. Entretanto, estes municípios receberam tratamento também em ao longo do ano e eleições anteriores, então foram removidos da base de dados. Entretanto, é importante destacar uma preocupação levantada pelo próprio IDEC na coleta dos dados:

> "A prática de conceder subsídio, sem mencionar qual o valor do recurso, se mostrou recorrente nos municípios analisados. No universo total pesquisado (100 maiores cidades + cidades menores em que há informações sobre subsídios), há 122 cidades em que foi implementado

[6]Entre outras tentativas de adotar mais variáveis, foram coletados os dados históricos pluviométricos disponibilizados pelo INMET para identificar se choveu no dia da eleição, mas há muitos dados faltantes, o que pode prejudicar a análise, inviabilizando a utilização desses dados.

algum tipo de subsídio. Em 46 cidades, ou 38% do universo pesquisado, não houve divulgação de valores específicos, seja em sites oficiais ou entrevistas na imprensa."

Essa dificuldade enfrentada pelo IDEC na pesquisa de municípios que já adotavam o passe livre antes da eleição se aplica também para o passe livre na eleição. Como a medida foi adotada em nível municipal, cada município teve liberadade para executá-lo de formas diferentes, como foi o caso da prefeitura de Rio Branco, que "decidiu conceder a gratuidade apenas na volta do eleitor da zona eleitoral. Para evitar o pagamento da passagem, o usuário do transporte público deverá apresentar o comprovante do voto"[7]. Todos os municípios que tiveram qualquer tipo de subsídio no dia da eleição foram considerados municípios tratados.

3.3 Metodologia

A estratégia adotada foi de realizar um *diff-in-diff* com efeitos fixos de tempo e município no qual o grupo de tratamento é composto pelos municípios que adotaram o passe livre e o de controle, pelos que não adotaram. A hipótese de identificação é de que os grupos apresentam trajetórias de abstenção paralelas na ausência de tratamento ao longo das eleições e, portanto, caso a diferença entre os grupos mude apenas no ano de tratamento, isso é consequência do tratamento.

$$\log(y_{it}) = \beta \text{Passe Livre}_{it} + X_{it}\gamma + \alpha_t + \delta_i + \epsilon_{it} \tag{9}$$

Na equação 9, y_{ti} representa a abstenção no ano t para o município i e X_{it} representa um vetor de covariantes. O coeficiente β captura o efeito médio do passe livre no log da abstenção e α_t e δ_i representam os efeitos fixos de tempo e município, respectivamente. A variável Passe Livre assume o valor 1 em 2022 para os municípios que adotaram o passe livre, caso contrário, assume zero. As estimações foram feitas separadamente para o primeiro e segundo turnos.

A variável passe livre é endógena, já que não foi decidido de forma aleatória qual município adotaria o passe livre. Portanto, a hipótese de identificação não se sustenta, pois os grupos de controle e tratamento são diferentes e não podem ser comparados diretamente. Para lidar com essa questão, foi utilizado o método do *Propensity Score Matching*, no qual calcula-se a probabilidade do passe livre ser adotado e compara-se municípios com propensões a ser tratados parecidas. O PSM foi calculado a partir do *Nearest Neighbor* ou "vizinho mais próximo" com reposição.

Como discutido e observado na Figura 2, mudanças no custo afetam de forma diferente municípios que apresentam diferentes faixas de abstenção. Nesse sentido, a modelagem da abstenção em termos logarítmicos lineariza essa relação.

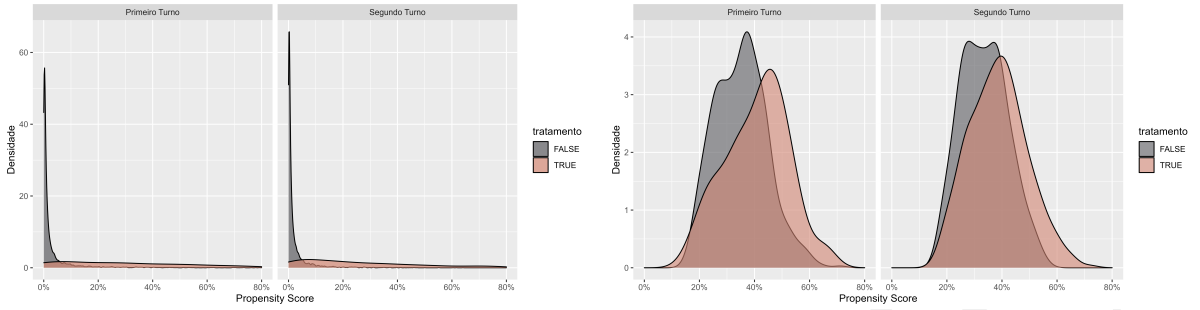
3.4 Resultados empíricos

Considerando a endogeneidade discutida, o primeiro passo adotado foi fazer o *Propensity Score Matching* (PSM). Na figura 6 é possível observar o suporte comum antes e depois de aplicar o método do PSM. O eixo x dos gráficos se refere à probabilidade do município receber o tratamento, enquanto o y se trata da densidade de probabilidade. Na figura 6b identifica-se uma grande sobreposição nas probabilidades dos grupos de receber o tratamento, uma evidência de que os municípios que estão sendo comparados são muito semelhantes, menos em relação a receber o tratamento. O suporte comum é menos significativo

[7] Trecho retirado da notícia da Uol `http://bit.ly/3YzDykk`,

no primeiro turno porque menos municípios foram tratados, então é mais difícil de encontrar vizinhos próximos.

Figura 6: Balanceamento dos grupos antes e depois do *Propensity Score Matching* (PSM)



(a) Antes do PSM

(b) Depois do PSM

Depois do balanceamento dos grupos, foi conduzido um *event study* (Tabela 2), para que se verifique a paralelidade das tendências dos diferentes grupos a partir de testes placebos. Foi escolhido como referência o ano de $t-1$, que neste caso é 2018. Caso se identifique efeito no tratamento em um ano que não teve tratamento, há uma forte evidência de que as trajetórias não são paralelas. O *event study* foi conduzido com variáveis controle e sem. A descrição detalhada das variáveis se encontra na tabela 4.

A partir dos resultados da tabela 2, a hipótese de identificação não se sustenta para o primeiro turno, visto que foi observado um efeito do passe livre na abstenção em 2014, para o nível de significância de 5%, o que é impossível, visto que não houve passe livre em 2014. Para o segundo turno não há evidências que rejeitem a hipótese de paralelismo das tendências, mas também não há evidências que apontem para um efeito do passe livre na abstenção.

É relevante destacar que as variáveis de controle selecionadas se mostraram bastante relevantes e apresentam o sinal esperado, o que aponta na mesma direção da literatura teórica e outros estudos empíricos. Os coeficientes do primeiro turno são semelhantes aos do segundo turno, com exceção da competitividade e eleitores por seção, que se mostram menos relevantes no segundo turno. Entretanto, a variável de interesse a partir dessa estratégia não se demonstra relevante.

Com isso, foi feita uma análise de efeito heterogêneo de renda. Como descrito com melhor detalhe na seção 2.1 e na figura 2b, uma redução no custo de votar afeta de formas diferentes cada município e uma variável importante que define essa diferença é o nível de renda do município. Portanto, foram separados os municípios em 4 quantis de renda e foi conduzido novamente o *event study* separadamente para cada um desses grupos (Figura 7).

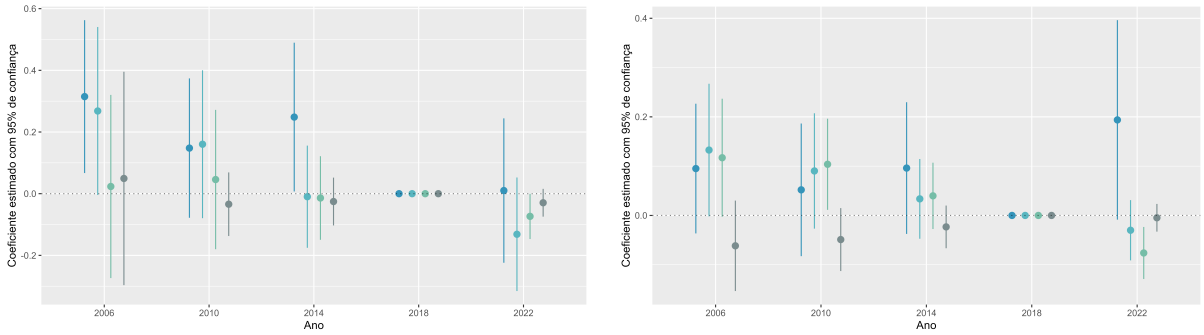
A partir da análise da Figura 7, fica evidente a dificuldade de validar a hipótese de identificação, visto que com 95% de confiança observa-se mais de uma vez efeito placebo em anos de eleição nas quais não houve tratamento, principalmente quando a amostra é dividida em quatro e perdem-se muitos graus de liberdade. Entretanto, para o terceiro quartil de renda foi identificada uma redução na abstenção do segundo turno com 5% de significância. Com base nos resultados do *event study* como um todo, o efeito do passe livre na abstenção é inconclusivo utilizando a estratégia empírica escolhida, dada a fragilidade da hipótese de identificação.

Por fim, é importante destacar alguns resultados secundários relevantes. Uma variável que não é

Tabela 2: Resultado do *event study* com efeito fixo de tempo e município, estimações por MQO

	Sem Controles		Com Controles	
	Primeiro Turno	Segundo Turno	Primeiro Turno	Segundo Turno
tratamento:2002	−0.069 (0.290)	0.044 (0.216)		
tratamento:2006	−0.044 (0.379)	0.024 (0.451)	−0.049 (0.273)	0.027 (0.238)
tratamento:2010	−0.068 (0.196)	0.003 (0.904)	−0.069 (0.109)	0.008 (0.681)
tratamento:2014	−0.146 (0.005) *	−0.010 (0.627)	−0.090 (0.023) *	−0.001 (0.956)
tratamento:2022	−0.010 (0.692)	0.016 (0.182)	−0.026 (0.282)	0.003 (0.803)
log(Competitividade)			−0.024 (<0.001) *	−0.010 (0.018) *
log(PIB per capita)			−0.022 (0.043) *	−0.016 (<0.001) *
log(Beneficiados)			−0.865 (<0.001) *	−0.857 (<0.001) *
IDEB			−0.063 (0.002) *	−0.062 (<0.001) *
log(População)			0.176 (0.287)	0.164 (0.089)
log(PIB governo)			−0.018 (0.005) *	−0.011 (<0.001) *
log(Eleitores por seção)			0.711 (<0.001) *	0.298 (<0.001) *
Num.Obs.	972	4440	802	3656
R2	0.567	0.590	0.706	0.709
R2 Adj.	0.483	0.526	0.631	0.651
R2 Within	0.021	0.004	0.276	0.184
R2 Within Adj.	0.015	0.002	0.263	0.181

Figura 7: Análise de efeito heterogêneo através de um *event study*



(a) Primeiro Turno

(b) Segundo Turno

O quartil à esquerda se refere ao quartil mais pobre e o da direita, o mais rico

muito explorada na literatura que foi adotada é o número de beneficiados. Foi observado que em média, um aumento de 1% na proporção de beneficiados para eleitores aptos está associado a uma redução na abstenção em aproximadamente 0,86%, *ceteris paribus*. Isso significa que em um município que há poucas pessoas votando em relação ao tamanho populacional, os eleitores apresentam uma probabilidade maior de votar. Este é um resultado esperado, como discutido na seção 2, visto que aumenta a percepção de importância do voto para o eleitor.

Outro resultado importante é que para um aumento de 1% na quantidade de eleitores alocados em uma seção eleitoral, é observado em média um aumento de 0.711% na abstenção para o primeiro turno e um aumento de 0.298% para o segundo turno, *ceteris paribus*. Esse resultado faz sentido, pois um número maior de seções possibilita distribuí-las melhor geograficamente, bem como reduz as filas. Entretanto, um detalhe interessante é que esta variável se demonstra muito mais decisiva no primeiro turno do que no segundo. Partindo do pressuposto de que as unas no primeiro turno são as mesmas do que no segundo, isso pode ser uma evidência de que as pessoas no segundo turno estão dispostas a passar por maiores custos de voto no segundo turno – possivelmente porque atribuem um benefício maior para o voto no segundo turno.

A participação do governo no PIB municipal também demonstrou ser um fator que está associado a um comparecimento maior. Possivelmente os eleitores atribuam maior importância para o voto e tenham maior sentimento de dever cívico em municípios nos quais o governo está mais presente e é responsável por uma parcela mais significativa do desenvolvimento do município.

3.5 Limitações

A principal limitação do estudo empírico conduzido, como mencionado anteriormente, foi a dificuldade de validar a hipótese de identificação da metodologia escolhida. Como o intervalo de tempo de quatro anos entre as eleições é muito grande, muitos fatores em um município mudam de uma votação para outra e depois de apenas 5 eleições, já se passaram 20 anos, tornando eleições distantes praticamente incomparáveis do ponto de vista empírico. Nesse sentido, seria importante conduzir um estudo com um método que não precise comparar os resultados de 2022 com a eleição anterior.

Outra dificuldade é oriunda da endogeneidade na decisão de adotar ou não o tratamento. Mesmo

depois do balanceamento dos grupos, eles se mostraram não tão comparáveis, o que indica que o perfil de município que adotou o passe livre é bastante diferente do perfil dos que não adotaram. Informações como se o município escolheu adotar o passe livre ou foi obrigado a fazê-lo podem ajudar a entender o perfil do município tratado e balancear melhor os grupos.

Além disso, há uma grande defasagem nas variáveis. O PIB *per capita*, por exemplo não foi divulgado ainda para o ano de 2022, e foi necessário utilizar o dado desatualizado de 2020. Ademais, as variáveis utilizadas para o balanceamento estão defasadas em 12 anos. Como já foi discutido, essa defasagem não é tão problemática quanto a dos dados para a regressão principal, mas ainda merece ser destacada.

A cobertura do sistema de transporte público do município determina quantas pessoas efetivamente conseguem usufruir do benefício do passe livre e para resultados mais robustos isso teria que ser considerado. Por fim, como discutido na seção teórica, a assimetria de informação é uma variável extremamente relevante para determinar o sucesso ou fracasso do passe livre e isso não foi contabilizado ou considerado na análise empírica. Além disso, não foi considerado se os municípios que anunciaram o passe livre efetivamente o aplicaram da forma como prometeram.

4 Conclusão

Pelas limitações e dificuldades apresentadas, os resultados empíricos em relação ao efeito do passe livre na abstenção são inconclusivos. Nesse sentido, o único resultado que pode ser validado é o teórico. É importante destacar que todas as premissas teóricas que foram mensuradas empiricamente tiveram sua intuição validada, o que vai de acordo com a literatura amplamente aceita.

O que a teoria microeconômica indica é que, assumindo as premissas do modelo, uma redução no custo do voto leva a um maior comparecimento quando não há assimetria de informação, mesmo que muito pequeno. Entretanto, não há evidências de que o passe livre foi adotado de maneira eficiente, então pode ser que a medida realmente não tenha apresentado efeito de reduzir a abstenção em 2022. Eficiente se refere à uma maneira que todos os habitantes estão cientes da medida e que ela não gere assimetria de informação. Neste caso, os eleitores anteciparam de maneira correta a quantidade aproximada de pessoas que seriam "convencidas" a votar, mesmo que sejam poucas.

Na medida em que a adoção é mais ineficiente, têm-se uma assimetria de informação e, como discutido na figura 3b, o efeito do passe livre pode ser atenuado ou pode até acabar sendo o efeito contrário em um caso extremo. De maneira realística, é impossível que não haja nenhum grau de assimetria de informação, o ponto é que sua magnitude pode ser determinante do fracasso ou não de uma medida de redução de custo de votar.

Portanto, caso seja adotada a medida, há alguns pontos importantes para serem levados em conta. Primeiramente, é muito importante garantir que os canais de comunicação do governo trabalhem de forma a fazer a notícia chegar em todos, principalmente aqueles que usufruem do sistema de transporte público. Em segundo, no dia da eleição a capacidade e qualidade do transporte entregues devem ser igual ou melhor do que o prometido, para evitar que aconteça o *moral hazard*.

Por fim, uma boa estimativa de quantas pessoas usariam o transporte público para votar é relevante não apenas para diminuir a assimetria de informação, mas também para ajudar na decisão de adotar ou não o passe livre, considerando que pode ser uma política muito custosa que apenas poucas pessoas usem. Nesse sentido, há outras formas, talvez menos custosas, de incentivar o comparecimento eleitoral que podem ser consideradas. Como discutido na seção 3.4, os resultados empíricos indicam que criar mais

seções eleitorais, por exemplo, pode também ser uma forma de incentivar o comparecimento eleitoral.

5 Notas Finais

5.1 Agenda de pesquisa

Como mencionado nas limitações, adotar a estratégia empírica de comparar a eleição de 2022 com as anteriores dificulta muito a validação da hipótese de identificação do método, então seria interessante explorar outros métodos econométricos que não necessitem dos dados de eleições anteriores. Ferramentas de GIS que utilizam os dados para seção eleitoral ao invés de nível municipal pode ser uma estratégia viável, que inclusive possibilita o uso de variáveis novas, que apenas existem nessa escala. A dificuldade de seguir este caminho é de georreferenciar estes locais, visto que os dados do TSE não apresenta muita qualidade[8].

Além disso, a modelagem teórica indicou uma série de variáveis importantes de serem consideradas no modelo empírico de forma a investigar efeitos heterogêneos. Uma exemplo disso seria uma métrica de conhecimento da população de que o passe livre tenha sido adotado. Para tanto, considerar dados de cobertura de mídia e efetividade dos canais de comunicação governamentais em cada município pode ser relevante. Entretanto, caso a estratégia empírica continue sendo de comparar a eleição com as anteriores, estes dados devem também ser coletados na mesma cobertura temporal, o que pode ser desafiador.

Outro fator relevante que foi deixado de fora é um estudo sobre os padrões de mobilidade de cada município e a cobertura e efetividade do sistema de transporte público. O passe livre apenas pode ter efeito se os eleitores aptos se deslocam pelos modais públicos de transporte. Possivelmente em cidades mais densas as pessoas preferem se deslocar à pé até as urnas, o que diminui a eficácia de uma medida de incentivo monetário para transporte de alta capacidade.

Uma outra agenda de pesquisa possível é identificar melhor os fatores que levam o município a adotar ou não o passe livre. Pode ser interessante investigar se o alinhamento político do prefeito com o governador ou com algum candidato à presidência é relevante para a decisão de adotar a medida. Outro caminho que pode ser investigado é, caso se identifique aumento no comparecimento por consequência do passe livre, se é possível identificar um perfil político dos eleitores que foram "convencidos" a votar pela medida.

Por fim, estimar o custo da política do passe livre possibilita compará-la com outras medidas que objetivam também estimular o comparecimento como aumentar o número de zonas eleitorais. Inclusive, pode haver uma sinergia de, caso seja adotado o passe livre, que crie-se mais seções eleitorais perto de estações de transporte público. Este tipo de interação pode ser apenas investigada caso mude a escala da análise para seção eleitoral.

5.2 Agradecimentos

Gostaríamos de fazer um agradecimento aos professores orientadores do trabalho Adriano Borges Costa e Bianca Tavolari, que não apenas ajudaram no encaminhamento do projeto, como também inspiraram seu

[8] Os dados de latitude e longitude do TSE são extremamente defasados, então alguns pesquisadores estão trabalhando em cruzar os dados que o TSE oferece, como CEP e endereço, com outras bases, como a do Google Maps ou a do Censo Escolar

desenvolvimento. Ao professor Adriano Dutra Teixeira também dedicamos nossa gratidão por colaborar ao longo do projeto com suas valiosas contribuições e pela revisão do relatório.

A organização `basedosdados.org` também teve papel importante no trabalho, visto que coletaram e trataram dados do TSE de forma a possibilitar uma consulta de SQL, poupando o trabalho de vasculhar o site do TSE. Como foram coletados dados para diversos anos, tanto do TSE, quanto de outras fontes para obter as variáveis de controle, o trabalho da organização foi um grande facilitador.

Agradecemos também pelo apoio dos integrantes do Insper Data, bem como o Insper por tornar possível este projeto.

5.3 Reprodutibilidade

O código inteiro utilizado para gerar todos os resultados apresentados se encontra no github `https://github.com/gustavo-tm/passe-livre` e estão em estado reprodutível. Os dados utilizados se encontram na pasta "data", mas no script é possível reproduzir a coleta dos dados utilizando a API da `basedosdados.org`. Os gráficos e tabelas da seção empírica foram confeccionados no R, mas os gráficos da seção teórica foram produzidos no LaTeX.

6 Anexo

Tabela 3: Variáveis utilizadas no *Propensity Score Matching* (2010)

Variável	Descrição
Razão Dependência	Percentual da população de menos de 15 anos e da população de 65 anos e mais em relação à população de 15 a 64 anos
Taxa Envelhecimento	Taxa de envelhecimento
Expectativa estudo	Expectativa de anos de estudo aos 18 anos de idade
Taxa analfabetismo	Taxa de analfabetismo da população de 18 anos ou mais de idade
Índice Gini	Índice de Gini
Prop. pobreza extrema	Proporção de extremamente pobres
PIB *per capita*	Renda per capita média
IDHM	Índice de Desenvolvimento Humano Municipal
Taxa desocupação	Taxa de desocupação da população de 18 anos ou mais de idade
Taxa água encanada	Percentual da população que vive em domicílios com água encanada
População	População
População Urbana	População urbana

Tabela 4: Variáveis utilizadas nos modelos econométricos

Variável	Descrição
Tratamento	Variável binária que assume 1 caso o município tenha recebido passe livre e 0 caso contrário
Competitividade	Quão proximos foram os resultados entre primeiro e segundo candidatos. Calculado por $1/[\log(V_A/V_B)]$, na qual V_A representa o número de votos recebidos pelo candidato mais votado e V_B, pelo segundo candidato mais votado.[9]
População	População do município
PIB per capita	PIB per capita do municipio, disponível até 2020. Para 2022, foram utilizados os últimos dados disponíveis, de 2020.
Beneficiados	Número de pessoas no municipio dividido pela quantidade de eleitores aptos.
IDEB	Nota da educacao dos anos finais do ensino fundamental nas escolas publicas. A nota é apenas calculadas nos anos ímpares, com o primeiro dado disponível em 2005, então foram utilizados dados defasados em um ano.
PIB governo	É definido como valor adicionado bruto a preços correntes da administração, defesa, educação e saúde públicas e seguridade social dividido pelo PIB municipal.
Eleitores por seção	Média do número de eleitores aptos por seção eleitoral no município

[9] Esta variável é calculada com base em dados póstumos à votação, mas a premissa adotada para incluir esta variável é de que as pesquisas eleitorais e o sentimento dos votantes em relação à competitividade não são significativamente diferentes dos resultados observados *a posteriori*

INSPERDATA

Estudo e Predição da Exposição de Liquidez de Capital

Pesquisadores: Gabriel Villaça, Giovanna Spirandelli , Thiago Hampl
Projeto de consultoria com DAO Capital

Resumo
O projeto desenvolvido em parceria com o fundo Dao Capital tem como objetivo principal responder à pergunta motivadora: é possível prever a tendência do Ibovespa a partir de indicadores macroeconômicos? Para responder essa pergunta, foi desenvolvido um backtesting robusto que balanceia uma carteira entre índice Ibovespa e a taxa DI (livre de risco). Dessa forma, a simulação da carteira varia a exposição da carteira entre esses dois ativos a partir de diferentes sinais macroeconômicos. Ao fim, não foi possível chegar a um resultado que previa perfeitamente as tendências do Ibovespa, porém houve um cuidado para que o código fosse o mais agnóstico possível, o que significa que não é muito difícil testar outras variáveis e estratégias.

Palavras-chave: Simulação, Fundo de Investimento, Macroeconomia, Momentum

1 Introdução

A equipe de Consultoria 1 do Insper Data trabalhou em conjunto com a gestora Dao Capital a fim de praticar um estudo prático sobre estratégias de variação da exposição de um fundo. A empresa é um fundo quantitativo que utiliza, majoritariamente, estratégias de Momentum na alocação do patrimônio. O orientador do projeto foi o Caio Castro, que auxiliou na execução do projeto em todas as etapas. O objetivo principal do projeto é pesquisar formas de balancear a exposição de uma carteira genérica a fim de prever momentos de queda e de subida. A equipe não teve acesso à carteira da Dao Capital, a exposição era balanceada entre títulos do Ibov ou taxa livre de risco diária, o DI. Em conversas iniciais com o orientador e outras pesquisas preliminares, uma decisão do grupo foi aplicar a variação de indicadores macroeconômicos como estratégia do projeto, em detrimento de outras opções, como o uso de dados micro empresariais ou técnicos.

A empresa, no entanto, utiliza de prognósticos fundamentalistas para a aplicação de seus métodos quantitativos de gestão. Por isso, não bastava simplesmente desenvolver um código que testa qualquer variável à escolha. Seria necessária uma prova preliminar de que o uso daquela variável para prever o retorno do Ibov era admissível. Por isso, o projeto se dividiu em duas etapas principais, em que a primeira as variáveis eram testadas e uma segunda em as estratégias eram realmente desenvolvidas e, a partir de um backtesting, os resultados de performance da estratégia eram desenvolvidos.

2 Regressão

O teste de variáveis foi feito por um modelo de regressão linear entre o sinal testado e o retorno do Ibovespa, diariamente. Dessa forma, a partir dos resultados da regressão, haveria uma base para afirmar que faz sentido manipular a exposição ao Ibovespa a partir de tal variável. Isso seria justificativa suficiente para que o fundo pudesse aplicar algum possível resultado positivo do projeto na estratégia quantitativa. Ao final dessa etapa de validação, as variáveis escolhidas para serem usadas na estratégia foram: Índice futuro de commodity, taxa de câmbio dólar real, índice de ações de mercados emergentes, índice S&P500, juros de longo prazo (5 anos) e futuros de títulos do tesouro dos EUA.

3 Simulação da Carteira

As estratégias se resumiram à variação dos sinais macroeconômicos, assim como já descrito. Para calcular a exposição a partir desse sinal, foi utilizado um balanceamento de pesos para calcular a melhor carteira num mesmo período com a mesma variável. Esse peso define o quanto o sinal será relevante para a janela testada, e como unidade de comparação entre carteiras foi utilizado um Sharpe simplificado (retorno da carteira dividido pelo desvio padrão do retorno).

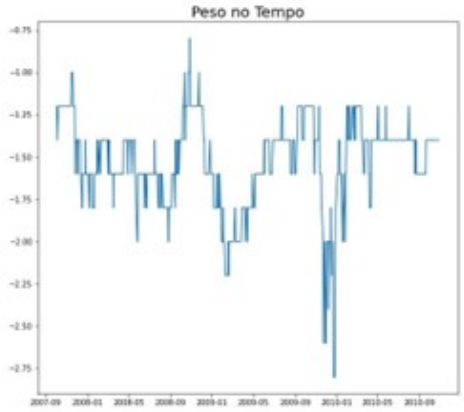
$$Exp = 1 - K \cdot \text{Sinal}$$

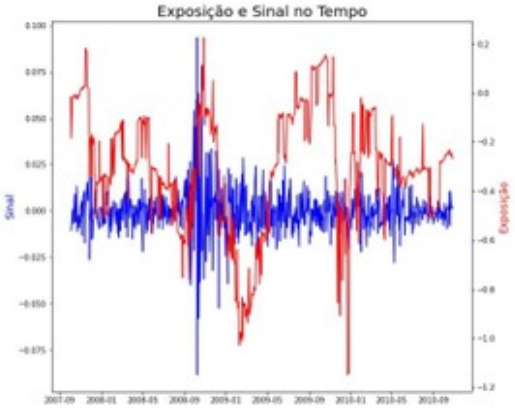
Além disso, o cálculo do retorno depende do valor da exposição. Em exposições maiores do que 1, o retorno diário é alavancado em Ibov e perde em DI. Em exposições entre 0 e 1, o retorno é calculado pela soma das duas partes. Já em exposições menores do que 0, o retorno é análogo ao maior do que 1: alavancado em DI e perdendo em Ibov.

No final, o objetivo do projeto é calcular a exposição do próximo dia usando a estratégia numa janela de dias anteriores. Foi definido que a última exposição ideal do backtesting seria a exposição do dia sendo testado e, dessa forma, seria possível calcular uma exposição ideal. Abaixo segue um exemplo prático de cálculo da exposição para um dia:

Utilizando pesos 1, 2, 3 e uma janela de 10 dias, para calcular a exposição ideal do dia 8 de Junho utilizando o índice S&P500 é calculado as exposições dos últimos 10 dias utilizando os valores da variação do S&P500 nesses dias com peso 1, peso 2 e peso 3. Com as exposições, é possível calcular o retorno de cada dia utilizando a variação do Ibov nos últimos dias (com shift de 2 dias para simular o atraso de informação). Depois disso, basta calcular o Sharpe entre as 3 séries de performance e escolher a última exposição da série com melhor Sharpe. Com esse método, foi possível calcular a melhor exposição para o dia 8 de Junho para os parâmetros de testes descritos inicialmente. Para um backtesting da carteira utilizando essa estratégia, basta repetir esse método para cada dia da análise e depois calcular a performance a partir das exposições ideiais.

A simulação feita no projeto utilizou uma janela de 180 dias e 30 pesos diferentes (de -3 a 3 com passo de 0.2). Existiam dados de 2007 a 2022 disponíveis para o executar o backtesting. Abaixo segue um exemplo de como os pesos são balanceados para calcular a exposição ideal com os parâmetros descritos acima, utilizando a variação do Câmbio dólar real como estratégia.





4 Resultados

Os resultados do backtesting foram separados por período para facilitar o entendimento da performance da carteira em diferentes ciclos econômicos. Abaixo há um exemplo de resultado utilizando a estratégia do indicador de juros a longo prazo:

O gráfico de resultado possui duas curvas: a amarela representa a performance do Ibovespa no período, que serve de benchmark para esse projeto. A outra curva representa a performance da carteira

Estratégia	Retorno Acumulado	Meses Exp<=0	Meses Exp>0	Sharpe Anual Médio
Câmbio (DOL-BRL)	0.403	160	33	-0.039
S&P Futures	0.778	39	154	0.020
US Treasury Bond Fut	0.632	144	49	-0.015
Commodity Index	0.663	48	145	0.013
Taxa Selic Longo Prazo	0.933	104	89	-0.019
Emerging Markets Fut	0.733	28	165	0.011
Ibov (Benchmark)	2.487	-	-	-

variando a exposição a partir da estratégia. Para essa curva, a cor representa o valor da exposição: para exposições menores do que 0, vermelho; entre 0 e 1, azul; maior do que 1, cor verde.

Foram gerados esses 3 gráficos nestes 3 períodos diferentes para todas as 6 variáveis, a fim de entender se a estratégia está realmente conseguindo prever o comportamento do Ibovespa e balanceando a exposição de forma a gerar uma boa performance, entendendo seu comportamento nesses períodos diversos. No entanto, para ter uma resposta plausível sobre a qualidade das estratégias é necessário simular para todo o período, de 2007 a 2022. Abaixo segue a tabela resumo das 6 variáveis, comparando-as com o benchmark.

É possível perceber que o retorno acumulado das estratégias não supera o próprio Ibovespa e, por isso, as estratégias não estão conseguindo prever o comportamento do índice. Mesmo que os resultados não sejam positivos, a ferramenta criada para a simulação do backtesting possui código versátil e adaptável, possibilitando pequenas mudanças para testar outras possibilidades de calcular a exposição. Uma das possibilidades pensada pela equipe seria adicionar uma informação no cálculo da exposição: a última exposição calculada. Essa informação não é irrealista para o backtesting, já que ao calcular a exposição do dia seguinte, é esperado que a do dia anterior já tenha sido calculada. Com a última exposição, é possível tomar decisões em relação ao cálculo das exposições na janela. Um exemplo seria utilizar dois sinais em uma estratégia, e escolher qual sinal será usado para a janela a partir do valor da última exposição.

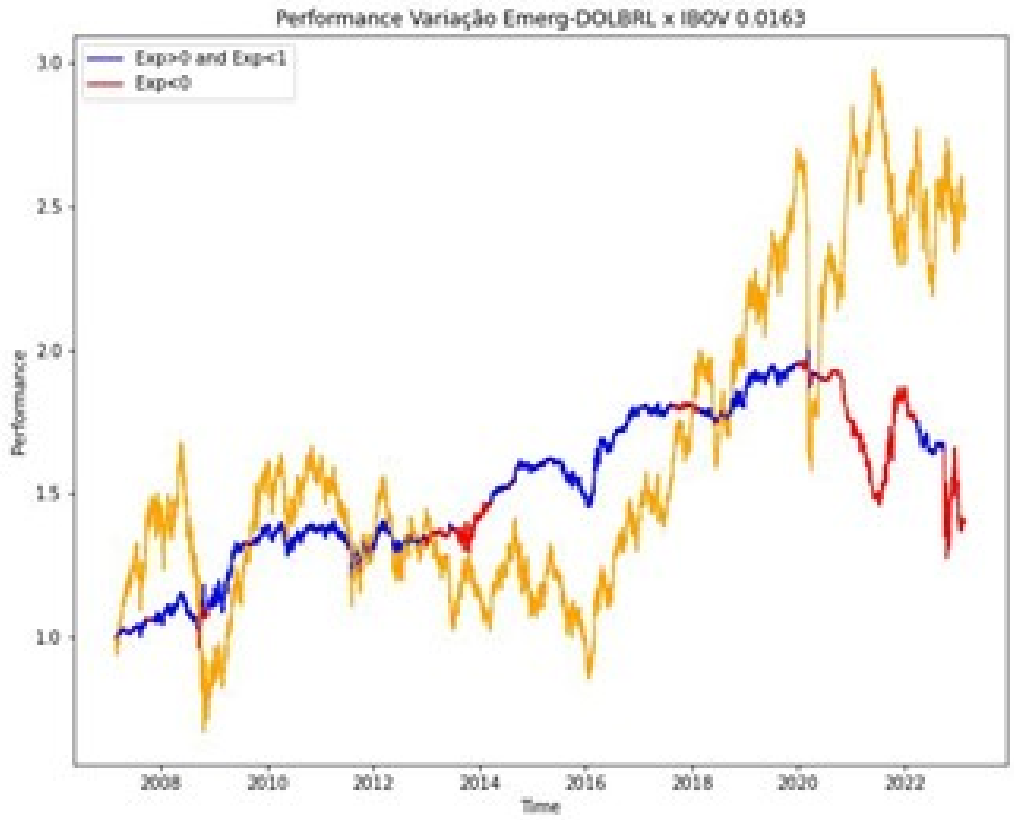
Um exemplo dessa possibilidade de carteira é utilizar o sinal do índice de ações de mercados emergentes para exposições anteriores menores do que 0 e o câmbio dólar real para exposições anteriores maiores que 0. Abaixo segue os resultados dessa estratégia para os períodos estudados anteriormente:

Além disso, foi construído o gráfico também da performance para todo o período de estudo simulando como seria investir nessa carteira a longo prazo.

Para essa estratégia, que muda o sinal da janela dependendo da última exposição calculada, o retorno acumulado foi melhor que qualquer estratégia que utiliza apenas um sinal.

5 Conclusão

Mesmo que o último resultado ainda seja inconsistente e com retorno menor que o próprio Ibovespa, a possibilidade de mudar o código original de forma a testar várias possibilidades de agrupamento de variáveis ou outras estratégias semelhantes evidencia a capacidade da ferramenta desenvolvida. Considerando que o objetivo inicial do projeto era estudar as possibilidades do balanceamento da exposição ao partir de sinais macroeconômicos, é possível concluir que esse estudo foi realizado no escopo definido pela equipe. Além da ferramenta de simulação desenvolvida, ao todo, foram gerados 32 Gráficos do retorno acumulado com mais 13 estratégias diferentes.


Performance Variação Emerg-DOLBRL x IBOV 0.0163
Exp>0 and Exp<1
Exp<0
Performance
3.0
2.5
2.0
1.5
1.0
2008
2010
2012
2014
2016
2018
2020
2022
Time

INSPERDATA

Predição de performance de modelos de crédito a partir de indicadores de estabildiade

Pesquisadores: André Corrêa Santos, Rafael Coca Leventhal
Projeto de consultoria com Serasa

Resumo
Neste projeto, propõe-se o desenvolvimento de um modelo preditivo para estimar a performance de modelos de crédito com base em métricas de instabilidade disponíveis mensalmente. Foram utilizados três tipos de modelos (regressão logística, regressão linear e rede neural) e métricas como erro absoluto médio (MAE) e erro quadrático médio (MSE) para avaliação. A limitação dos dados trabalhados tornou-se evidente, destacando a necessidade de abordagens futuras que considerem a inclusão de variáveis macroeconômicas para explicar as variações do KS2 ao longo do tempo.

Palavras-chave: -

1 Introdução

Um modelo de crédito consiste em uma ferramenta estatística que relaciona variáveis cadastrais com um resultado, chamado de score, capaz de informar a chance de determinado indivíduo pagar o crédito. Modelos de crédito são treinados com dados de segmentos específicos da população e para que um modelo produza resultados coerentes é necessário que seja utilizado com dados cadastrais de indivíduos que pertençam ao mesmo segmento da população com o qual o modelo foi treinado. Dessa forma, para garantir que o modelo continue performando como esperado, é necessário calcular métricas de instabilidade. Essas métricas avaliam se a distribuição das variáveis cadastrais da população sob a qual o modelo atua está próxima da distribuição das variáveis cadastrais da população de treino.

Caso um modelo opere em uma população muito diferente daquela com a qual ele foi treinado, é esperado que haja uma queda de performance. Contudo, esse não é sempre o caso. Existem instâncias em que as métricas de instabilidade acusam uma diferença muito grande entre a população recebida e a população esperada, mas, ainda assim, o modelo continua capaz de distinguir bons e maus pagadores. A métrica de performance que avalia a capacidade de distinção do modelo é chamada de KS2.

O objetivo do projeto consiste em tentar estimar a variável de performance (KS2) em função das variáveis de instabilidade disponíveis. Isso é interessante uma vez que só se tem acesso à variável de performance após os credores pagarem ou ficarem inadimplentes. Contudo, as variáveis de instabilidade estão disponíveis mensalmente. Dessa forma, seria possível estimar uma possível queda de performance antes dela ocorrer e assim mitigá-la.

Para estimar o KS2, primeiramente os dados referentes à instabilidade e performance de modelos de crédito foram tratados e, em seguida, três tipos de modelos diferentes foram treinados com as variáveis de instabilidade como entrada e a variável de performance como *"target"*.

2 determinação das variáveis de instabilidade

Essa etapa consiste na extração das métricas de instabilidade que serão utilizadas nas etapas de treinamento e validação do modelo. Existem duas métricas de instabilidade principais, o KS1 e o PSI. O KS1 (também conhecido como KS one) é calculado a partir dos "scores" produzidos pelo modelo de crédito para cada segmento separadamente e para o modelo de crédito inteiro. O PSI é calculado para as mesmas instâncias do KS1 e também para cada variável do modelo de crédito. Também calcula-se um PSI geral que considera todas variáveis e segmentos. Por fim, o KS2 (o *"target"*) é calculado a partir de 90 dias do início da execução do modelo e ele é obtido em função de todos os segmentos. Vale ressaltar que esses indicadores foram calculados por safra.

As "safras"referem-se aos períodos mensais em que os dados de instabilidade e performance dos modelos de crédito são coletados e analisados. Cada safra representa um mês específico, em que as métricas de instabilidade são calculadas.

O KS1 (Kolmogorov-Smirnov One) e o PSI (Population Stability Index) são métricas fundamentais no contexto da modelagem de crédito para avaliar a estabilidade e a performance dos modelos. O KS1 é calculado a partir dos scores produzidos pelo modelo de crédito para cada segmento individualmente e para o modelo como um todo. Essa métrica mede a distância máxima entre a distribuição esperada, presente no grupo de treino do modelo de crédito, e a população observada durante o uso do modelo de crédito.

$$KS1 = \max |\text{Esperada}(i) - \text{Obtida}(i)| \tag{10}$$

O cálculo do PSI é semelhante ao cálculo do KS1 no sentido de que ambos comparam as distribuições das variáveis cadastrais na amostra de treino e na população recebida pelo modelo de crédito. Para realizar o cálculo do PSI, a amostra de dados é dividida em categorias ou bins, e as proporções dessas categorias são comparadas entre a amostra de desenvolvimento (treino) e a amostra atual, bem como os logaritmos da razão dessas proporções.

$$PSI = \sum_{i=1}^{n} \left((Obtida(i) - Esperada(i)) \cdot \ln \left(\frac{Esperada(i)}{Obtida(i)} \right) \right) \tag{11}$$

Finalmente, o KS2 é uma métrica de performance que avalia a capacidade do modelo em distinguir entre bons e maus pagadores. Ele é calculado a partir das distribuições de scores do modelo para os dois grupos: bons pagadores e maus pagadores. O KS2 mede a maior diferença entre essas duas distribuições, indicando quão bem o modelo pode separar os dois grupos.

$$KS2 = \max |\text{Bons Pagadores}(i) - \text{Maus Pagadores}(i)| \tag{12}$$

Nas equações acima o termo *"Obtida"* diz respeito à distribuição do "score" do modelo em um determinado segmento e safra do modelo de crédito. Já o termo "Esperada" é a distribuição para esse mesmo segmento nos dados de treinamento do modelo de crédito.

3 treinamento e avaliação dos modelos preditivos

Uma vez calculadas as métricas de instabilidade supracitadas, uma série de modelos foram treinados e avaliados visando estimar KS2 geral de cada safra em função dos dados de cada segmento de modelo.

Porque cada modelo de crédito recebido possui um número variável de segmentos e variáveis cadastrais, não foi possível a construção de um modelo agnóstico, isto é, um modelo capaz de receber métricas de instabilidades oriundas de qualquer modelo de crédito. Dessa forma, os dados para treinamento e avaliação dos modelos de modelos foram produzidos a partir da primeira base de dados recebida pelo grupo. Essa base contém dados de um modelo de crédito baseado em regressão logística.

As variáveis utilizadas para prever a resposta incluíram: 'PSI médio das variáveis', 'KS1 do segmento', 'KS1 GERAL', 'PSI GERAL' e 'KS1 segmentos geral'. Essas métricas são calculadas em função dos indicadores de instabilidade já mencionados. A variável 'PSI médio das variáveis' é calculada como a média dos PSI das variáveis de um segmento em determinado mês. 'KS1 Geral' e 'PSI Geral' representam os KS1 e PSI gerais para determinado mês. Enquanto 'KS1 segmentos geral' indica o KS1 calculado para todos os segmentos em uma dada safra. Essas características foram selecionadas por serem indicadores de instabilidades de fácil acesso para a empresa e pois sintetizam todos os outros indicadores de instabilidade que pudemos calcular.

Para filtrar as variávis explicativas uma matriz de correlação foi produzida:

	Media_PS_vars	KS_segmento	KS_GERAL	PS_GERAL	KS_SEGMENTOS_GERAL
Media_PS_vars	1.000000	0.614133	-0.063322	-0.027665	0.165005
KS_segmento	0.614133	1.000000	0.157603	0.188356	0.043072
KS_GERAL	-0.063322	0.157603	1.000000	0.984639	0.067978
PS_GERAL	-0.027665	0.188356	0.984639	1.000000	0.121905
KS_SEGMENTOS_GERAL	0.165005	0.043072	0.067978	0.121905	1.000000

À partir dessa matriz foi possível concluir que o KS Geral e o PS geral são muito fortemente correlacionados e portanto foi decidido que o somente o PS geral seria usado no treinamento dos modelos.

Três modelos diferentes foram treinados: uma regressão logística, uma regressão linear e uma rede neural simples com três camadas internas. A rede neural foi construída com 10 neurônios por camada e reLU por função de ativação. Os dados processados anteriormente foram estratificados por segmento e separados aleatoriamente em treino e teste na proporção 70% para treino e 30% para teste. As métricas escolhidas para avaliação dos modelos foram o erro absoluto médio (MAE) e o erro quadrático médio (MSE) Seguem os resultados:

	Rede Neural	Regressão Logística	Regressão Linear
MAE	0.033	0.036	0.037
MAE/avg	13.56%	16.12%	16.53%
MSE	0.0023	0.0019	0.0020

Seguem os p-valores obtidos para a regressão linear:

	Media_PS_vars	KS_segmento	PS_GERAL	KS_SEGMENTOS_GERAL
p>[t]	0.206	0.591	0.255	0.000

O R^2 ajustado obtido equivale a 0.279.

Para simular um possível caso de uso para a empresa, os modelos foram treinados com os primeiros 8 meses avaliados nos últimos 3. Os resultados e as previsões obtidas são observados a seguir:

	Rede Neural	Regressão Logística	Regressão Linear
MAE	0.035	0.054	0.052
MAE/avg	13.81%	20.07%	19.51%
MSE	0.0014	0.0031	0.0029

seguem os p-valores calculados para a regressão linear com esse novo treinamento:

	Media_PS_vars	KS_segmento	PS_GERAL	KS_SEGMENTOS_GERAL
p>[t]	0.506	0.951	0.009	0.000

O R^2 ajustado obtido equivale a 0.337 e isso indicaria uma melhor *"fit"* do modelo, contudo, o erro mais significativo observado poderia indicar alguma outra variável importante para a determinação da variável resposta não foi considerada ou ainda que há endogeneidade.

Nota-se que os resultados obtidos para o segundo treinamento e teste foram significativamente piores. Por conta do tamanho reduzido do grupo de teste não é possível ter grande certeza do erro apresentado pelos modelos, visto que esse grupo não pode ser considerado representativo.

Ainda que a rede neural tenha performado bem em ambas as rodadas de teste e as regressões tenham sido satisfatórias nos testes iniciais, não é possível concluir que a estratégia de estimar o KS2 safra a safra seja a ideal. Entendemos que esse seja o caso, pois, para treinar e avaliar os modelos utilizados de forma robusta é necessário centenas de linhas de dados e, mesmo com a estratégia de expandir o conjunto de dados ao estimar o KS2 em função das métricas de instabilidade para cada segmento, ainda assim não é possível atingir uma ordem de magnitude desejável. No caso, ao estimar o KS2 geral em função de cada segmento, foi possível produzir 88 instâncias de dados a partir das 11 instâncias originais. Entretanto, é impossível atingir uma ordem de magnitude desejável para esse tipo de modelagem - na ordem de centenas ou milhares de instâncias de dados, visto que cada instância de dado depende de uma safra (um mês) registrada por um modelo de crédito. Além disso, é importante observar que dentre as variáveis explicativas selecionadas, as variáveis oriundas de segmentos específicos, "Média_PS_Vars"e "KS_Segmento"possuem os maiores p-valores e, portanto, as variáveis mais importantes para determinação do KS2 geral estão contidas nas variáveis de instabilidade gerais, variáveis essas que possuem um número muito reduzido de instâncias de dados em relação às variáveis específicas aos segmentos.

O trabalho indica que há evidências sugerindo que o KS2 pode ser estimado a partir de métricas de instabilidade. Mas seria interessante considerar também diferentes variáveis macroeconômicas para tentar explicar as variações do *"target"* ao longo do tempo, uma vez que a performance dos modelos apresentou uma queda significativa quando treinados e avaliados temporalmente. Finalmente, o maior desafio enfrentado nesse tipo de modelagem é a obtenção de uma quantidade suficiente de dados para o treinamento desses modelos.

INSPERDATA

Transparência no cenário político: coleta e tratamento de dados do STF

Pesquisador: Esdras Gomes Carvalho
Orientador: Ivar Alberto Glasherster Lange Hartmann

Resumo

O presente projeto visa auxiliar o entendimento e análise do padrão de votação dos ministros do Supremo Tribunal Federal (STF) do Brasil. Dada a relevância do judiciário e a complexidade em acessar os dados dos processos, a pesquisa concentra-se em criar um banco de dados público abrangente, compreendendo pelo menos 95% dos processos do STF. O projeto foi executado a partir da coleta dos processos disponíveis no site do STF, concentrando esforços na computação paralela com 15 máquinas virtuais. A coleta foi concluída com um total de 2.356.676 processos coletados, e serão disponibilizados de online. O projeto repr-esenta um passo fundamental na compreensão das decisões judiciais no Brasil e promete fornecer insights sobre os fatores que influenciam as decisões dos ministros, como ideologia, influências políticas e confiança no relator.

Palavras-chave: Supremo Tribunal Federal, Análise de Dados Judiciais, Web scrapping

1 Introdução (Motivação)

A tripartição dos poderes é uma das características centrais no sistema político brasileiro, e presente em boa parte do mundo, com eleições regulares para os representantes do legislativo e do executivo. O terceiro poder, o judiciário, apresenta uma importância tão grande como os demais, entretanto, diferente dos dois primeiros os interesses e padrões de decisão não são tão claros.

Aliado a isso, os dados acerca de processos são públicos, porém de difícil acesso para realizar análises de grande porte, haja vista que a suprema corte brasileira é ímpar em relação ao número de processos anuais, como pode ser visualizado na figura 1. A única suprema corte com mais processos é a Índia, porém existem 30 ministros e a população é quase sete vezes superior, ao passo que o Brasil tem apenas 11 ministros. Até então, pesquisas sobre os padrões de decisão de ministros (clustering) eram realizadas com uma pequena parcela do total.

Fig. 1: Número de processos de supremas cortes em 2019.

Além disso, não há dados estruturados sobre quais foram as decisões, de modo que, a análise de como cada ministro votou se dá em um processo artesanal de leitura do texto da decisão e saber quais ministros estavam presentes à época. Isso torna uma análise geral inviável. Contudo, dado que a informação de quais ministros compunham o rol de magistrados do supremo na data da decisão esteja disponível, o texto da decisão possibilita a utilização de um algoritmo que se baseia em expressões regulares (RegEx), e tem por objetivo a classificação dos votos dos ministros em cada decisão analisada.

Nesse sentido, o presente projeto tinha como objetivo inicial responder a pergunta "Como votam os ministros do STF?", tentando compreender se existem agrupamentos de ministros e quais os fatores que levam as suas decisões (ideologia, influências políticas, confiança no relator, etc.). O projeto focou na criação de um banco de dados público, com pelo menos 95% dos dados dos processos do STF, presentes no site (https://portal.stf.jus.br/), sendo uma primeira etapa de 3 para responder a pergunta proposta.

2 Metodologia aplicada

Para a coleta dos dados, foi necessário seguir as seguintes etapas: levantar quais dados do processo seriam salvos, analisar a estrutura do site, estudar requisições HTTPS realizadas, criar código de web scrap, realizar testes, selecionar uma proposta de arquitetura e, por fim, a limpeza e manipulação dos dados.

A partir de consultas ao orientador, o levantamento dos dados relevantes para análise foi realizado. Abaixo são apresentados cada dado e uma breve descrição:

- Classe: Denomina o tipo de processo dentre um rol que é de competência do supremo tribunal Federal, por exemplo: Arguição de Descumprimento de Preceito Fundamental, Ação Direta de Inconstitucionalidade, Habeas Corpus, Recurso Extraordinário, entre outros).
- Número: Apenas atribuído ao processo quando ele chega ao STF. Cada classe possui uma série de processos cujos números são ordenados de maneira sequencial. É possível que existam processos de classes diferentes com o mesmo número justamente por conta da sequência que é continuada a cada novo processo destinado àquela classe.
- Número único: número identificador do processo.
- Origem: local geográfico de origem do processo, se for no Brasil é apresentado o estado.
- Relator: ministro do supremo que recebe o caso, faz as primeiras análises e resume o caso aos demais, caso o processo vá para decisão em turma ou plenário.
- Relator do último incidente: Define o relator da última movimentação do processo. Pode ser diferente do inicial porque nos casos em que a decisão da turma/plenário diverge do relator inicial ou há ausência deste, o processo é distribuído para um novo relator.
- Partes: Interessados no processo e envolvidos diretamente na lide judicial (discussão) que tramita por meio do processo.
- Andamentos: Atualiza o leitor sobre o status atual do processo e cada uma das etapas pelas quais ele passou dentro do Supremo.
- Assunto: tema relacionado ao processo
- Número de origem: número do processo original como foi gerado no tribunal de origem, ou seja, o tribunal no qual o processo foi inicialmente protocolado e julgado.
- Órgão de origem: o órgão que deu início ao processo, seja público ou privado.
- Tamanho físico do processo: quantidade de volumes, folhas e apensos do processo físico, se houver.

Todas essas informações estão presentes no site do processo. Em seguida, foi realizada uma análise da estrutura do site, bem como de quais requisições HTTP eram necessárias para coletar as informações desejadas. Com isso, foi possível a elaboração de um código, feito em python e com auxílio da biblioteca "beautifulsoup4", para realizar a coleta das informações relevantes, caso haja uma classe e número de processo (significando que é um link com processo válido).

Ao realizar alguns testes para verificar se as informações eram coletadas e ajustar o código, seguiu-se para a seleção de uma proposta de arquitetura. Nesse ponto, é importante ressaltar uma característica desafiadora desse projeto: o bloqueio temporário de acesso ao site. A saber, após uma série de requisições, o servidor do STF bloqueia a conexão (ligada ao número ip da máquina) como uma forma de evitar ataques de negação de serviço - uma série de acessos simultâneos intencionais que inviabilizam o acesso ao site.

Ciente do desafio, duas propostas foram arquitetadas, a primeira solução envolve a criação de uma máquina virtual (no inglês, virtual machine comumente abreviado para VM) ec2 que realiza chamadas para uma função lambda, que realiza o web scrap e salvamento dos dados no dynamoDB (banco não relacional da amazon). Essa primeira arquitetura foi aplicada na primeira coleta e a sua vantagem seria a mudança de ip de forma dinâmica, já que o aws lambda é usado em arquiteturas servless, de modo que, quando a função é chamada, uma VM com o código é criada e realiza a execução do código. Esse processo ocorre "debaixo dos panos" e permite a execução do código em máquinas diferentes a cada execução. Para a primeira coleta foram utilizadas 3 dessa arquitetura (uma VM, uma lambda e um banco dynamoDB em cada conta).

Contudo, a primeira coleta apresentou algumas falhas, elencadas a seguir: A escolha de como considerar um processo válido, que na primeira coleta era a presença de número único, contudo há processos válidos sem número único; Descobriu-se que a lambda não realiza esse processo de subir uma nova máquina a cada requisição, somente após alguns minutos (tipicamente uns 10 a 20). Desse modo, elas também sofriam com o bloqueio temporário; Além disso, um certo intervalo que continha processos válidos (do número de incidente 1 até 1465222). Devido a essas falhas e somado a questão do preço, tendo em vista que o tempo de computação numa VM ec2 é bem mais barato que o de uma lambda, uma nova arquitetura foi proposta para a segunda coleta.

Antes de apresentar a proposta, é válido ressaltar que foi realizado um estudo sobre a presença de processos válidos e sua relação com o número de incidentes. Esse processo se deu a partir de uma verificação de existência de processos válidos a cada 1000 números incidentes, para buscar por "buracos" na sequência, relatados pelo professor Fernando (autor da primeira coleta realizada semelhante a essa). Com isso, foi determinado um intervalo a ser buscado de 1 a 15.000 e de 1.400.000 a 6.605.876, usado na próxima coleta.

A partir disso, a segunda arquitetura foi posta em prática, que consiste em concentrar os esforços na computação paralela. A segunda proposta contém 15 VMs e 3 bancos dynamoDB e dessa vez a VM realiza o processo completo de coleta, manipulação e salvamento dos dados. Para essa segunda versão, alguns cuidados extras foram adotados, como a criação de cláusulas de erro no código, de modo que os processos que falharam tem um log escrito no arquivo "errors.txt" e a presença de um "checkpoint.txt" para saber em que ponto da coleta a máquina se encontra. Após criar o código a ser utilizado nas VMs, foi criado um service do systemd (para mais informações sobre acesse o link https://www.baeldung.com/linux/create-remove-systemd-services), para permitir que o código continuasse rodando mesmo que não esteja conectado a máquina, bem como facilitar o monitoramento.

3 Resultados

Na primeira coleta foram levantados cerca de 1.490.986. Os valores por ano dos processos coletados são apresentados na imagem a seguir:

A segunda coleta coletou 2.356.676 processos. A partir da contagem de processos que falharam chega-se à conclusão que esse valor representa cerca de 99,96% do total de processos existentes no intervalo considerado. Os dados serão disponibilizados para acessos online, com auxilio do time de Ciência de dados do Insper, a partir da disponibilização de uma máquina virtual para hospedar os dados.

INSPERDATA

Taxa de juros neutra para o Brasil e Estados Unidos

Pesquisadores: Giovanni Vescovi Filho, Camila Vaz

Resumo

Devido às recentes notícias e observações de elevados juros reais ex-Ante, buscamos entender o comportamento do juros e se o mesmo encontra-se em um valor alto demais ou não seguindo abordagens mais simples de mensuração, como a regra de taylor e um modelo VAR. Após os modelos e seus resultados, algumas das justificativas para os valores são variáveis de difícil mensuração e de modelagem mais difícil ainda.

Palavras-chave: Juros, Política Monetária, Brasil, Estados Unidos, VAR e Regra de Taylor

Introdução e motivação

No presente trabalho, foi desenvolvido um método para determinar a taxa de juros neutra de um país através da Regra de Taylor (*Taylor Rule*). A determinação da taxa de juros neutra ainda é um tema teórico com aplicações diferentes no mundo atual por seus diferentes métodos de determinação. Tal taxa consiste naquela capaz de fazer com que o PIB de um país convirja ao seu potencial simultaneamente ancorando a política monetária. Visto isso, sua aplicação apresentada nesse trabalho, consistirá na comparação da taxa de juros real efetiva de um país com a taxa neutra a fim de determinar distorções presentes nas economias apresentadas.

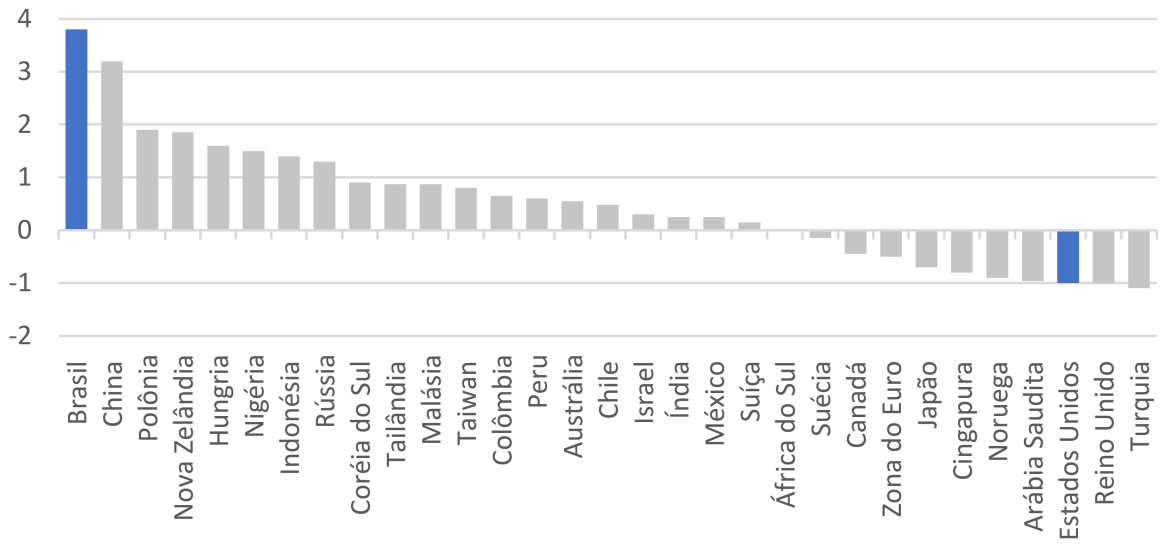
A maior motivação para o desenvolvimento de dada pesquisa foi a disparidade entre a taxa de juros efetiva brasileira comparada internacionalmente.

Como pode ser observado no Gráfico 1, nesta comparação, o Brasil é o país com a maior taxa de juros real no período de 2012 a 2016. Dessa forma, o objetivo deste trabalho será, através de uma análise histórica de dados, observar distorções nas taxas de juros brasileiras e determinar hipóteses para tais ao compararmos com um país com uma economia estável, como os Estados Unidos. Analogamente, será utilizado este país como um "grupo de controle" a fim de estipularmos distorções.

Dados e *Taylor Rule*

Primeiramente, necessita-se determinar uma fórmula padrão a fim de encontrarmos duas taxas de juros neutras, tanto para o Brasil como para os Estados Unidos. Para isso, foi utilizada a Regra de Taylor que consiste em um cálculo para política macroeconômica capaz de determinar a taxa de juros ideal para um país em função da inflação e do volume da atividade econômica, com isso, um de seus *inputs* é a taxa de juros real de equilíbrio, somada à um prêmio de risco que reflete as expectativas futuras de inflação e outros riscos. Para este trabalho em questão, tal taxa real de equilíbrio, definida por Taylor como

Figura 11: Taxa de Juros Real (Média Taxa Ex-Post 2012-2016 - %a.a.)



(a) Fonte: Itaú Asset Management 2017

tal por representar o nível da taxa de juros necessário para manter a economia em seu pleno emprego e estabilidade de preços, será a nossa taxa neutra.

Analogamente com uma simples manipulação da fórmula obtemos:

Taylor Rule:

$$i = r + \ \pi + 0,5 * \left(\pi - \pi^T\right) + 0,5 * \left(y - y^{PE}\right)$$

Isolar r:

$$r = \ i - \pi - 0,5 * \left(\pi - \pi^T\right) - 0,5 * \left(y - y^{PE}\right)$$

- $i = Juros\ real$
- $r = Juros\ nominal\ de\ equilíbrio$
- $\pi = Inflação$
- $\pi^T = meta\ de\ inflação$
- $y = PIB$
- $y^{PE} = PIB\ potencial$

A taxa de juros real definida acima por i, consiste na taxa efetiva de juros de determinado país descontada a inflação, já a taxa de juros nominal consiste na taxa de juros sem o desconto da inflação.

Modelagem

Para realização do cálculo temporal da taxa neutra pela fórmula, foi importando cada variável dos sites dos bancos centrais respectivos de cada país, bastou apenas colá-los respectivamente na fórmula

estipulada e, através de uma planilha do Excel, construiu-se uma nova variável temporal que pode ser incorporada nos gráficos a seguir. É importante ressaltar que para a meta inflação dos EUA foi utilizada a série *10-Year Breakeven Inflation Rate* que consiste na expectativa de inflação futura baseada no rendimento até o vencimento de títulos do Tesouro de 10 anos.

Como inputs para as fórmulas supracitadas temos que para o Brasil, os dados utilizados como taxa de juros, inflação e hiato do produto (diferença entre produto observado e potencial) foram a Selic [432], IPCA [433] e IBC-br[24364].

Já para os Estados Unidos, o que foi adquirido foram as séries temporais de: "Gross Domestic Product, Billions of Dollars, Quarterly, Seasonally Adjusted Annual Rate" (GDP), "Real Potential Gross Domestic Product, Billions of Chained 2012 Dollars, Quarterly, Not Seasonally Adjusted" (GDPPOT), "Federal Funds Effective Rate, Percent, Quarterly, Not Seasonally Adjusted" (DFF), "Sticky Price Consumer Price Index less Food and Energy, Percent Change from Year Ago, Quarterly, Seasonally Adjusted" (CORESTICKM159SFRBATL), "10-Year Breakeven Inflation Rate, Percent, Quarterly, Not Seasonally Adjusted" (T10YIE).

R Studio

Diferentemente do que foi feito no Excel, no RStudio, o método de estimação da variável de interesse será de Vetor Autorregressivo (VAR).

Para isso, as séries temporais de interesse, similarmente ao Excel, serão taxa de juros, inflação e hiato do produto também tendo como códigos, da base de dados de séries temporais do Banco Central brasileiro, os valores 432, 433 e 24364.

Após a coleta do Brasil, foi feita a dos Estados Unidos, desta vez tendo como fonte dos dados o banco de dados do FRED de St. Louis. Referente à inflação, taxa de juros e PIB, os códigos da base foram CPIAUCSL, DFF e GDP.

Em termos de instrumentalização prática, para ambos os países foram feitos os mesmos procedimentos, sejam de passo a passo seja de determinação de frequência desejada, trimestral.

Com isso, primeiro para se extrair o hiato do produto, foi utilizado o filtro Hodrick-Prescott (HP), para poder extrair a tendência da série, simbolizando o produto potencial, e os desvios da tendência, os ciclos, que representam os hiatos do produto.

Primeiros passos estimação

Após a consolidação das variáveis de interesse, e realizando os devidos testes, entendeu-se que tomar a diferença do logaritmo da série, para os índices de preço, transformando-os em inflação, seria o mais apropriado.

Após a estimação do VAR, entende-se que, após a verificação da matriz de correlação dos resíduos, como os resíduos de cada equação são correlacionados no mesmo espaço de tempo, que os valores das matrizes de coeficientes não são interpretáveis. Porém, como o intuito do trabalho não é verificar a intensidade das relações das variáveis entre si e sim apenas o nível dos juros isso não é um problema.

Figura 12: Backtesting BRA

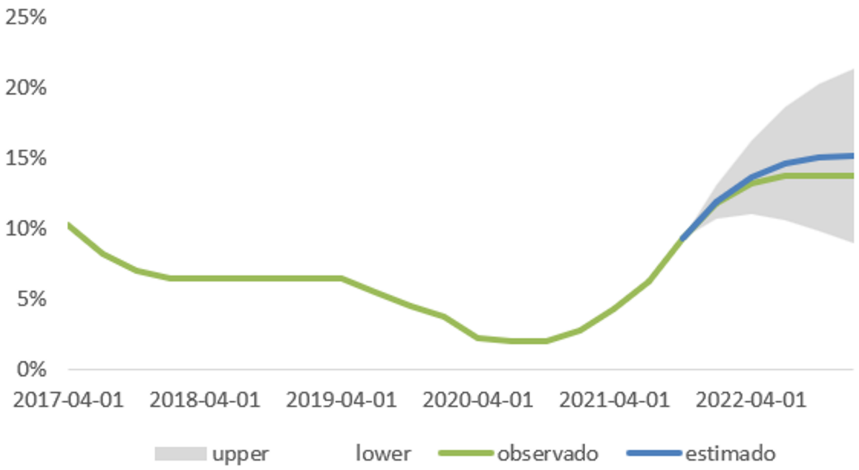
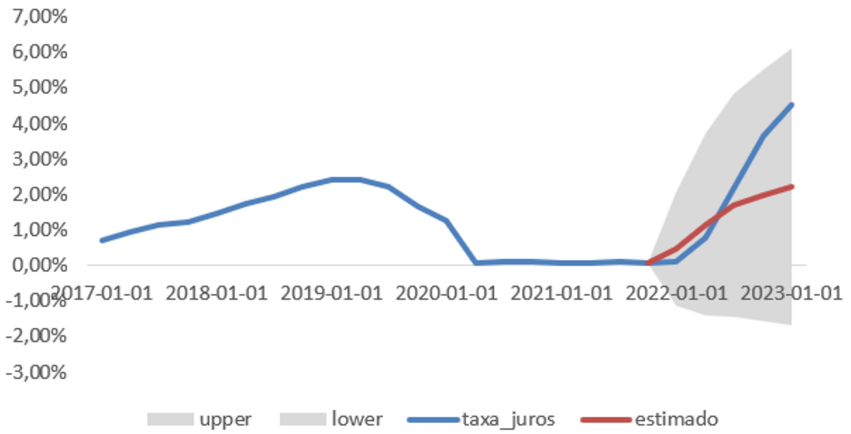


Figura 13: Backtesting USA



Conclusões e resultados

R Studio

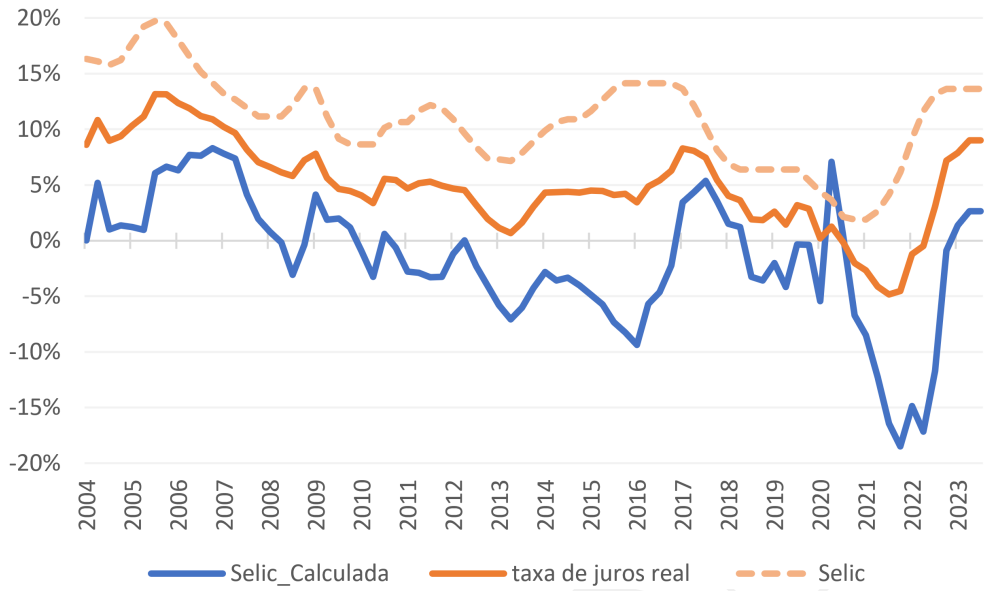
A respeito dos resultados do método VAR, tem-se que o nível de juros dos últimos períodos está dentro do intervalo de confiança observado, porém, como os intervalos de confiança possuem bandas muito largas, estes resultados são pouco conclusivos, tanto de maneira positiva quanto negativamente.

Tendo como visualização gráfica dos resultados Brasil do R Studio:

Já os resultados dos USA do R Studio:

Como é possível ver, graficamente, os intervalos de confiança são muito grandes, tanto para os USA quanto para o Brasil, neste sentido, algumas das explicações é o impacto da recente crise da pandemia, algo que modelos não são capazes de representar a real discricionariedade necessária para resolver as questões correntes. Sendo assim, os parâmetros e as estimativas aproximam-se mais de ruídos do que de sinal para a nossa análise. Sendo assim, deixamos em segundo plano as análises e resultados gerados a

Figura 14: Taxas de juros brasileiras



partir do uso do RStudio e do Vetor Autorregressivo.

Excel

Dada as interpretações e resultados gerados pelo RStudio, optou-se por dar mais destaque e maior peso para os resultados do Excel para o cálculo da taxa de juros e suas consequentes interpretações.

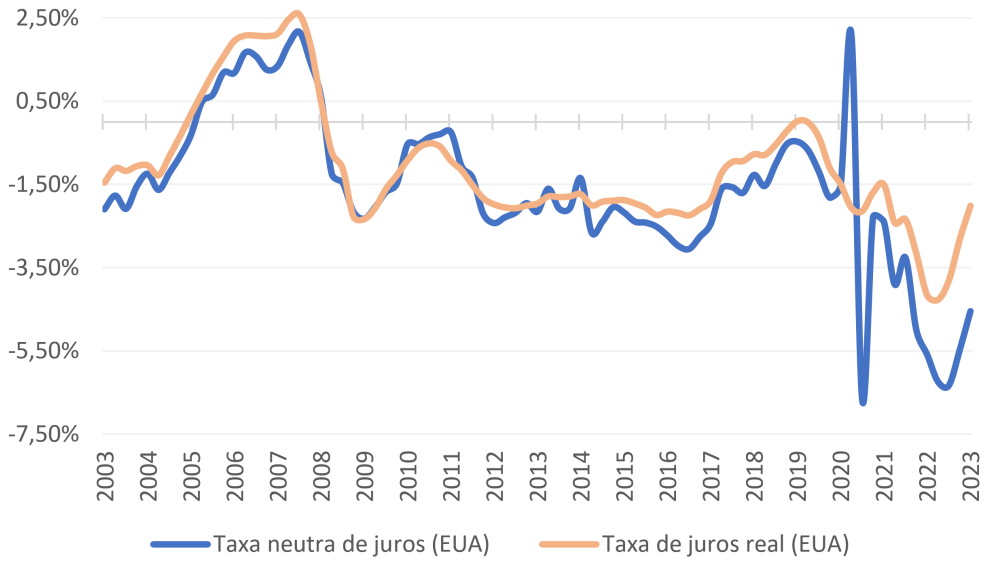
Portanto, obtivemos os seguintes resultados gráficos:

Tem-se como resultado uma clara disparidade entre a distância da taxa de juros neutra com as taxas de juros reais respectivas dos países. É importante destacar que, nos EUA, por se tratar de um país com uma economia desenvolvida e consequentemente mais estável, estes abordam o determinado Zero Lower Bound, um método de determinação da taxa de juros efetiva o qual consiste em manter suas taxas nominais acima de zero, uma vez que em muitos períodos de baixíssima inflação, cálculos para a taxa de juros ideal indiquem que uma adequada seria abaixo de zero, porém a Zero Lower Bound, também abordada por diversos países europeus, impede que as taxas fiquem negativas, pois tal contradiz os princípios econômicos da taxa de juros. Com isso, para melhor observação dos dados, não foi incluído no gráfico a taxa de juros nominal dos EUA, uma vez que esta se distanciaria da taxa neutra por imposições de valores alterados externamente.

Tendo em vista os resultados observados nos gráficos 4 e 5, e sabendo que é a partir da diferença entre a taxa de juros real e a taxa real neutra de juros que conseguimos avaliar a quão apertada (ou relaxada) está a política monetária, entendemos que o Brasil possui a tendência de ter sua política muito mais apertada do que os Estados Unidos, ou seja, o Banco Central do Brasil impões políticas para restringir a quantidade de moeda em circulação muito mais do que qualquer outro país.

Para identificarmos as razões das políticas monetárias brasileiras se diferenciarem das do restante do globo, seria necessário entender quais os fatores exógenos que influenciam a necessidade de impor uma taxa de juros tão elevada para que uma política contracionista entre efetivamente em vigor na

Figura 15: Taxas de juros estado-unidenses



sociedade brasileira. As hipóteses mais frequentes seriam a confiabilidade no banco central, a quantidade e divulgação de intrigas políticas e a influência de variáveis externas ao país, tendo em vista o grupo de controle, podemos considerar que uma das disparidades da sociedade americana da brasileira seria o apoio ou confiança por parte da população nas autoridades públicas, o que pode afetar as decisões financeiras dos indivíduos impactando a economia como um todo.

Em contrapartida, analisando as oscilações presentes na taxa de juros neutra ao longo dos anos, sabemos que esta possui uma tendência favorável à queda quando se trata de um período com maior grau de compromisso com metas fiscais, aprovação de reformas estruturais, publicação de matérias relacionadas à maior seguridade de trabalho, entre outros, como o período de 2009 à 2016. Já as medidas que podem ser relacionadas ao movimento de aumento da taxa neutra são a elevação da dívida pública e o fator exógeno do mercado internacional que sensibiliza o modelo nacional no mesmo sentido, como ao longo de 2016 e 2017.

Limitações

No trabalho descrito realizado pelo grupo, reconhece-se que o método de determinação da taxa neutra de juros pode estar equivocado no sentido que, dado por se tratar de uma variável utilizada de diferentes formas pelas economias ao redor do globo, é ineficiente determinar uma fórmula capaz de sumarizar todos os fatores influentes para essas economias, como a natureza discricionária dos *policy makers*, a fim de encontrar uma taxa neutra comparável entre todos. Porém, métodos aproximados que generalizam esses fatores como a Regra de Taylor, ainda podem ser aceitos dados que o objeto deste trabalho é uma comparação simples das variações históricas de duas taxas.

Visto isso, é importante destacar também as variáveis exógenas que foram desconsideradas na estipulação da fórmula, como a confiabilidade no BC, já que se trata de uma variável qualitativa sendo impraticável sua introdução em uma fórmula. São estas e outras limitações que podem ser reconhecidas

no modelo descrito no presente trabalho já que se trata de dois países com sociedades diferentes com fatores psicológicos e ambientais distintos, porém elas não são vistas como impeditivas para a análise apresentada.

INSPERDATA

Modelagem preditiva sobre Preços de imóveis em São Paulo

Pesquisadores: Rafael Albuquerque, Ricardo Wurzmann e Enzo Luidge
Projeto de consultoria com Lounge 161

Resumo

Em um primeiro momento, buscou-se informações para iniciar a extração de dados e sua metodologia, utilizando o Pythom como principal ferramenta. Logo no início da modelagem, após poucas extrações, quando resultados iniciais foram obtidos e métodos de modelagem estavam sendo discutidos, houve uma mudança de foco, de acordo com as vontades da empresa. Mudando nosso planejamento, buscamos obter uma ferramenta que pudesse ser utilizada rotineiramente de modo que ela fornecesse dados com variáveis produtivas diante do objetivo do projeto. E assim foi feito, mesmo diante de dificuldades e mudanças repentinas, obtivemos bons resultados diante das circunstâncias apresentadas.

Palavras-chave: Web Scrapping, Comunicação, Adaptação, Densidade de dados, ROI, Previsão

Introdução

Considerações iniciais

Um projeto base da entidade Insper Data consiste em uma análise profunda, desenvolvida durante um semestre, que agregue valores de análise de dados acerca da temática do grupo que o membro é direcionado, seja ela acadêmica ou feita a partir de parcerias com empresas, onde nosso projeto se encaixa.

A Louge 161 é uma empresa de empreendimentos de negócios que, entre seus diversos serviços, oferece a investidores de imóveis possíveis apartamentos para compra com foco em retorno financeiro, sendo ela a empresa com a qual fechamos parceria para a realização do projeto.

Motivação

A frase motivadora do projeto é basicamente a síntese de todos os objetivos que o cercam, “Analisar e extrair dados para que a Lounge consiga fazer um modelo de previsão sobre studios em São Paulo” cumpre essa funcionalidade.

É de grande interesse para a Lounge entender como funcionam os preços dos apartamentos em São Paulo para oferecer segurança e conforto a seus clientes e, para isso, seria necessário construir um modelo preditivo para conseguir atingir esse objetivo principal.

Metodologia

Planejamento

O planejamento iniciou-se a partir do primeiro contato com a Lounge 161 a partir da indicação de sites para buscar referência de dados para uma análise, para que as primeiras análises pudessem ser realizadas com o intuito de buscar as melhores metodologias para uma boa modelagem, porém o plano inicial não foi seguido durante todo semestre.

O que de fato ocorreu foi que ao realizar algumas extrações de dados, por meio de ferramentas de programação em Pythom, e obter análises preliminares, a Lounge 161 nos contatou que eles haviam contratado um grupo de modelagem para fazer esse mesmo trabalho que havíamos nos planejado para fazer, porém de uma maneira mais profissional e a longo prazo, não apenas visando um projeto semestral como o Insper Data sugere. Portanto, a partir daí nosso foco passou a ser a extração de dados apenas, porém com um foco maior na densidade de informações que poderíamos obter para potencializar o trabalho de modelagem que seria feito por essa outra parceria feita pela Lounge que utilizaria da nossa ajuda para cumprir o objetivo inicial.

É importante destacar que a partir desse momento de mudança de rumo do projeto a Lounge criou uma conta no Air DNA, um site que funciona como setor mais profissional do Air BNB, possibilitando nosso grupo a atingir os novos objetivos propostos com maior facilidade. Assim, nos aprofundamos na extração de dados dessas duas plataformas para obter dados e informações úteis e claras para a Lounge 161.

Obtenção de dados

Como já dito, a linguagem de programação utilizada foi Pythom, já a biblioteca principal adotada foi a Selenium, que é a biblioteca mais indicada para realizar web scrapping, possibilitando o usuário a automatizar "cliques" dentro de um determinado site e extrair as informações dentro dele que forem indicadas pelo código.

Dentro das próprias plataformas do Air BNB e do Air DNA existiam filtros que possibilitavam ambas as plataformas apresentarem os dados dos imóveis alvos do estudo, que são aqueles localizados em São Paulo e que estivessem dentro da descrição de studios. O resultado ideal é um código que poderá ser usado com frequência pela Lounge, para que a densidade de dados seja cada vez maior com o monitoramento das variáveis através do tempo para potencializar a modelagem em si.

Resultados

Uma vez que a metodologia está clara, podemos apresentar aquilo que obtivemos através dela, veja como cada uma das plataformas nos ajudou e quais aspectos cada uma se destacou e quais limitações cada uma apresentou.

Air BNB

Para o site do Air BNB a obtenção de dados foi mais direta, consistia na obtenção de uma base que fornece as principais informações disponíveis para cada imóvel. Possui a limitação de que alguns imóveis não

fornecem algumas dessas informações, porém isso não deverá ser um problema, uma vez que a utilização cotidiana do código trará a densidade de dados suficiente para uma boa análise, veja um exemplo dos resultados da execução do código em uma determinada data.

Apartamento em São Paulo ★ 4,0 (8)	Apartamento em São Paulo ★ 4,4 (5)	Apartamento em São Paulo ★ 4,63 (8)
Pinheiros - Studio - 28m² - Cama de Casal	Studio - Frei Caneca - 18m² - Geladeira -...	Consolação - Studio - 18m²- Varanda -...
1 cama de casal	1 cama de casal	1 cama de casal
16 – 21 de jun.	15 – 20 de jun.	15 – 21 de jun.
R$225 noite	**R$143** noite	~~R$180~~ **R$162** noite

Bairro	Tamanho	Preço/noite	Preço/m²	Avaliação	Número de Avaliações	Data
Pinheiros	28	225	8,035714	4	8	14/06/2023
Frei Caneca	18	143	7,944444	4,4	5	14/06/2023
Consolação	18	162	9	4,63	8	14/06/2023
Moema	25	170	6,8	4,27	48	14/06/2023
Luz	22	128	5,818182	4,7	254	14/06/2023
Consolação	25	252	10,08	5	3	14/06/2023
Brooklin	22	206	9,363636	4,21	89	14/06/2023
Frei Caneca	18	139	7,722222	5	4	14/06/2023
Predizes	17	159	9,352941	4,38	13	14/06/2023
Luz	22	122	5,545455	4,75	789	14/06/2023
Expo center	22	161	7,318182	3,67	3	14/06/2023

Para a plataforma do Air DNA foi preciso nos aprofundar mais ao utilizar das ferramentas do site, pois ele é bastante denso e tem diversas funções que poderiam nos ajudar na obtenção dos dados. Após explorar e discutir em grupo, decidiu-se criar algo com informações mais regionais e visuais dentro da cidade de São Paulo, justamente para complementar as informações vindas das extrações do Air BNB, veja exemplos dos resultados.

Desafios

As mudanças de planos em si foi uma dificuldade para nós, afinal quando ocorreu, o rumo do projeto era completamente diferente e, com o tempo passando, a cobrança por resultados começou a ficar mais pesada.

A metodologia para a extração de dados foi u desafiadora também, afinal, apesar da biblioteca utilizada ter sido a mesma, os códigos foram bem diferentes, pois o Air DNA funciona por assinatura, diferente do Air BNB.

Entretanto, ao longo do semestre, o grupo conseguiu se adaptar bem ao que foi proposto e superou as dificuldades, em geral os resultados obtidos foram satisfatórios e agora cabe a Lounge 161 utilizar as ferramentas fornecidas pelo Insper Data para realizar uma modelagem eficiente tendo em vista o tema central do projeto.

Agradecimentos

Dentro do Insper Data, recebemos diversas ajudas para desenvolver nosso projeto, porém, de forma especial gostaríamos de agradecer ao Presidente Thiago Rocha e o Diretor de Projetos André Correa, eles nos deram bons direcionamentos em momentos chaves do projeto.

Gostaríamos de agradecer também à Lounge 161 pela confiança depositar em nossos esforços para agregar em seus serviços profissionais, de modo especial, Enzo Pravatta e Leonardo Franuci que acompanharam o projeto dede o início e mantiveram contado e ciência em nossos avanços e desenvolvimentos

Referências

Berument, Hakan, Zubeyir Kilinc e Umit Ozlale (2005). «The missing link between inflation uncertainty and interest rates». Em: *Scottish Journal of Political Economy* 52.2, pp. 222–241.

Downs, Anthony (1957). «An economic theory of democracy». Em: *Harper and Row* 28.

Edlin, Aaron, Andrew Gelman e Noah Kaplan (2007). «Voting as a rational choice: Why and how people vote to improve the well-being of others». Em: *Rationality and society* 19.3, pp. 293–314.

Fowler, James H (2006). «Altruism and turnout». Em: *The Journal of Politics* 68.3, pp. 674–683.

Haspel, Moshe e H Gibbs Knotts (2005). «Location, location, location: Precinct placement and the costs of voting». Em: *The Journal of Politics* 67.2, pp. 560–573.

Istrefi, Klodiana e Anamaria Piloiu (2014). «Economic policy uncertainty and inflation expectations». Em.

Itaú Asset Management (2017). «Taxa neutra de juros no Brasil.» Em.

Konishi, Kenta e Tadahiko Murata (2010). «Examination of effect of free bus service in election using computer simulation». Em: *2010 2nd International Symposium on Aware Computing*. IEEE, pp. 263–268.

Laubach, T e J Williams (2001). *Measuring the Natural Rate of Interest. Board of Governors of the Federal Reserve System (US)*.

Pereira, Rafael HM et al. (2023). «Transporte público gratuito e participação eleitoral». Em.

Riker, William H e Peter C Ordeshook (1968). «A Theory of the Calculus of Voting». Em: *American political science review* 62.1, pp. 25–42.

Smets, Kaat e Carolien Van Ham (2013). «The embarrassment of riches? A meta-analysis of individual-level research on voter turnout». Em: *Electoral studies* 32.2, pp. 344–359.