



Boletim de Projetos 2023.2

Conteúdo

I Faixa Azul e segurança viária	2
II Crescimento da população em situação de rua e a capacidade de atendimento dos Centros POP em 2021 no município de São Paulo: um estudo descritivo.	18
III Qualidade Educacional e Valorização Imobiliária: Um Estudo sobre o Efeito das Escolas em São Paulo	29
IV Análise Comportamental em Plataformas Digitais: Aplicando Cadeias de Markov e Simulações de Monte Carlo	38
V Engenharia de dados para plataforma de consultoria educacional	45

Faixa Azul e segurança viária

working paper

Pesquisadores: Gustavo Theil, Esdras Gomes, Júlio Mugnol e Carlos Cabral

Orientado por: Adriano Dutra Teixeira

Um projeto em colaboração com Falconi

Resumo

Com o objetivo de melhorar a segurança viária na cidade de São Paulo, a CET implementou uma motofaixa exclusiva na Avenida dos Bandeirantes, que ao segregar os usos da via, pretendeu diminuir os conflitos de trânsito e levar a um cenário de tráfego mais seguro. Caso bem sucedida, a medida será implementada em diversos outros trechos da cidade. A teoria microeconômica indica que, apesar de poder haver uma melhora na segurança viária, os motoristas passam a ser menos cautelosos, causando efeitos adversos. Para testar o efeito resultante, utilizou-se o método de controle sintético. Desde a sua inauguração, em outubro de 2022 até dezembro de 2023, estima-se que foram evitados 32 sinistros que envolveriam motociclistas e ao menos uma pessoa ferida de forma grave ou leve, o que representa uma redução de 26.8% dos sinistros com este perfil registrados no período pós tratamento. Os resultados apresentam robustez com p-valor de 1.35%.

Palavras-chave: Segurança viária, motofaixa, faixa azul, controle sintético

1 Introdução

De acordo com a Secretaria de Vigilância em Saúde, “acidentes de trânsito” são a segunda principal causa de óbito no Brasil para pessoas entre 15 e 49 anos. Na literatura de segurança viária, o termo acidente não é empregado, visto que subentende uma ocorrência inesperada e não intencional. O termo sinistro é considerado mais adequado, pois reconhece a responsabilidade que os agentes de trânsito apresentam para impedir situações envolvendo danos físicos ou materiais (NBR, 2020).

Ao redor do planeta, diversas medidas para segurança viária são adotadas, algumas universalmente, outras com cunho experimental e pequena escala. Wang, Quddus e Ison (2013) qualificam dois principais fatores que interferem no risco de sinistro: engenharia e comportamento. Sinalização, iluminação, visibilidade, air bags e freio ABS são alguns dos fatores de engenharia que contribuem para a segurança no trânsito. A qualificação do motorista, seu nível de atenção ou se está alcoolizado são exemplos de componentes comportamentais determinantes para o risco de sinistro.

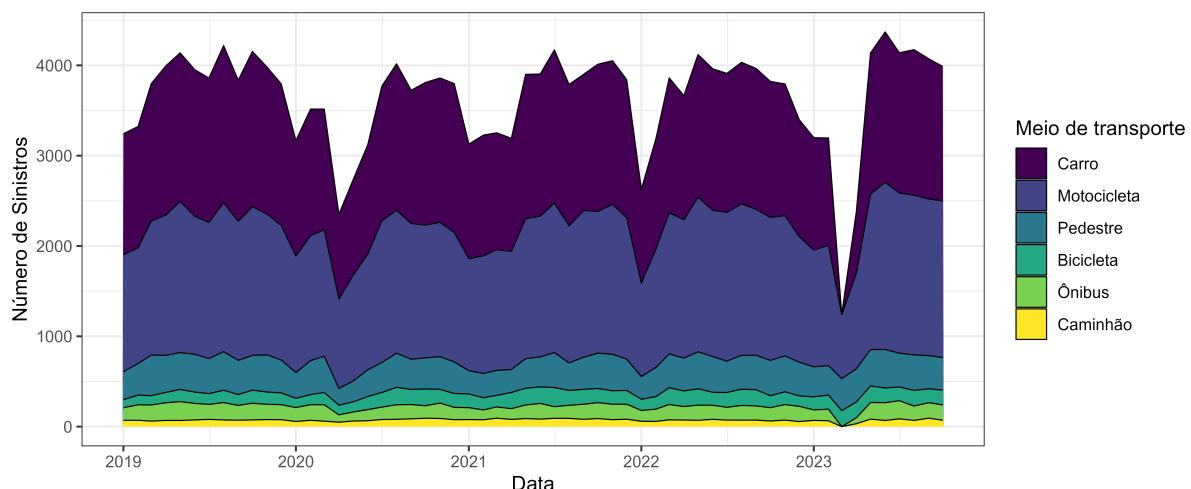
No Brasil, diversas instituições compartilham da responsabilidade de garantir o Art. 1 §2 do Código de Trânsito Brasileiro (CTB): “O trânsito, em condições seguras, é um direito de todos e dever dos órgãos e entidades componentes do Sistema Nacional de Trânsito, a estes cabendo, no âmbito das respectivas competências, adotar as medidas destinadas a assegurar esse direito”. Segundo Vasconcellos (2005), o CTB, que substituiu seu predecessor Código Nacional de Trânsito (1966), transferiu grande parte das

responsabilidades dos estados para os municípios. Apesar de apresentar desdobramentos preocupantes para municípios pequenos, como investigado em Bavoso (2014), possibilita também maior atenção para as especificidades de cada município.

O município de São Paulo está entre as áreas mais densas e urbanizadas do mundo, apresentando, portanto, um padrão de mobilidade característico. Além da extensa rede de transporte público, o município conta com uma frota de mais de 6 milhões de automóveis e mais de um milhão de motocicletas, segundo dados do IBGE. O uso em massa de motocicletas como meio de transporte é um fenômeno comum em grandes cidades na América Latina e Ásia, mas geralmente não está no radar de pesquisadores americanos e europeus.

No município de São Paulo, como é possível observar na Figura 1, uma parte considerável dos sinistros de trânsito apresentam um motociclista envolvido – grupo que é naturalmente mais vulnerável pelas características do veículo. Esse fenômeno também é observado em grandes cidades na Malásia, Colômbia, Tailândia, Índia, entre outras e motivou, inclusive, a pesquisa de Saini, Chouhan e Kathuria (2022). Os pesquisadores fizeram uma revisão sobre estudos nessas regiões, especificamente sobre uma nova ferramenta de segurança viária: as faixas exclusivas de motocicletas. As evidências apontam para o sucesso geral da medida, mas apenas na escala local. Também na Malásia, Radin Sohadi, Mackay e Hills (2000) analisaram um trecho expandido de uma motofaixa já existente e observaram uma redução de 39% nos sinistros mensais de motocicletas, reforçando evidência preliminar que uma faixa exclusiva diminui conflitos e aumenta a segurança para motociclistas.

Figura 1: Evolução mensal no número de sinistros no município de SP

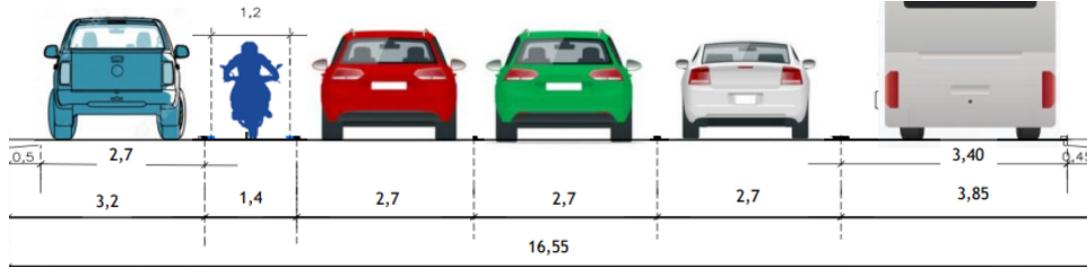


Essa intuição tem respaldo da literatura, afinal, a teoria de desenho urbano considera que em vias com usos mistos de automóveis, caminhões, motocicletas, ônibus e bicicletas, a segregação de usos é uma estratégia que pode melhorar o fluxo e segurança dos veículos. Jun e Heng (1992) analisam como a segregação de usos entre veículos motorizados e não motorizados ao longo das vias e em intersecções de cidades chinesas reduz o conflito entre eles e, portanto, é importante para garantir maior segurança.

Foi com essa mentalidade que a Companhia de Engenharia de Tráfego (CET), um órgão da Secretaria Municipal de Mobilidade e Trânsito de São Paulo, deu o primeiro passo para implementar essa medida

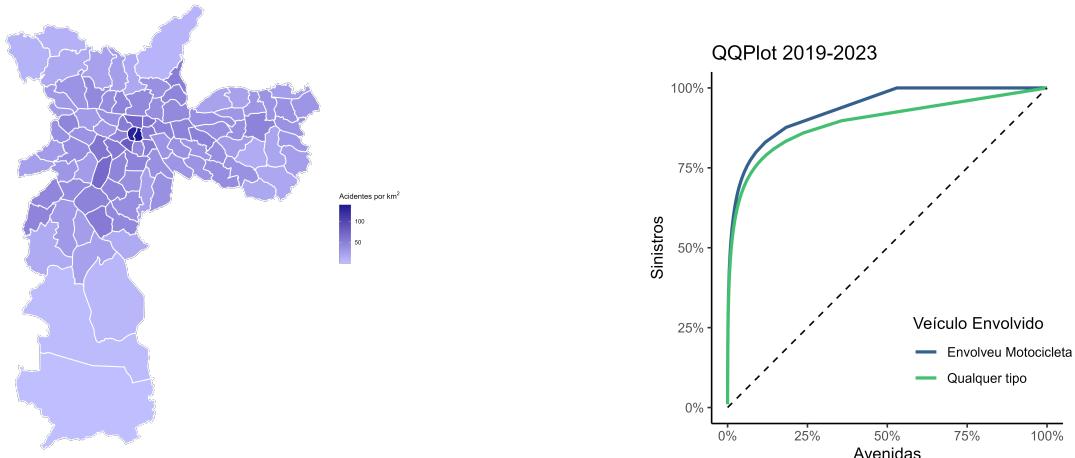
nas vias de SP. Segundo o Código de Trânsito Brasileiro (CTB Art. 80 §2), a implementação de uma medida não prevista no código segue a seguinte diretriz: “O órgão máximo executivo de trânsito da União poderá autorizar, em caráter experimental e por período prefixado, a utilização de sinalização e equipamentos não previstos neste Código”. Portanto, em janeiro de 2022, a CET implementa com cunho experimental a Faixa Azul, uma via exclusiva para motocicletas. A primeira via que recebe essa nova medida é a Avenida 23 de Maio, e é possível observar como fica a divisão das faixas na Figura 2.

Figura 2: Nova disposição da Avenida 23 de Maio



Entretanto, essa não é a primeira vez que São Paulo experienciou a segregação de usos das vias. A primeira motofaixa no Brasil foi implementada em 2006, na Avenida Sumaré e a segunda, em 2010 na Avenida Vergueiro. Todavia, ambas as faixas foram desativadas em 2013 e 2014, respectivamente. A CET afirma que a faixa foi extinta porque “não representou aumento da segurança dos motociclistas” e a decisão de desativar foi elogiada por especialistas (UOL 2013). De acordo com o Despachante DOK (2022), “[a motofaixa] era à esquerda e um dos motivos de não ter funcionado é que os motociclistas não gostam de andar perto da guia, pois são locais com resíduos e que podem enganchar a pedaleira”.

Figura 3: Perfil geográfico dos sinistros no município de São Paulo



(a) Distribuição espacial dos sinistros

(b) Concentração dos sinistros (QQ-Plot)

É importante destacar também o perfil geográfico dos sinistros. O que se pode concluir no caso de São Paulo é que eles estão muito concentrados em grandes avenidas, principalmente nas regiões centrais da cidade, como é possível observar na Figura 5. Na Figura 3b, observa-se que 50% dos sinistros estão

concentrados em menos de 1% das avenidas. Isso é um indicativo de que caso seja feita uma intervenção bem sucedida nessas poucas e relevantes avenidas, grande parte do problema será resolvido.

Considerando, portanto, tanto o caráter experimental do novo modelo de motofaixa, quanto seu fracasso no passado, é necessária uma análise de seu desempenho para avaliar se deve ser implementada em maior escala. A CET (2023) anunciou o sucesso da medida após um ano de faixa azul sem mortes e a iniciativa foi premiada pelo SENATRAN. No entanto, todas as análises feitas foram puramente descritivas e não conferem nenhum tipo de inferência causal, ou seja, não é possível saber se o efeito analisado foi causado pela faixa azul ou descreve uma relação espúria, uma coincidência. A Avenida 23 de Maio, que teve zero mortes durante o primeiro ano da medida, também apresentou zero mortes em 2019 (CET, 2021), além de que outros fatores podem interferir na dinâmica do trânsito em São Paulo e precisam ser controlados.

Dessa forma, o presente estudo tem como objetivo investigar se a implementação da faixa azul causou um aumento na segurança viária.

2 Modelagem Teórica

A literatura identificou várias estratégias para modelar o comportamento de agentes de trânsito. Entre as principais, destacam-se os modelos microeconômicos baseados em utilidade e os estocásticos, que geralmente estão associados a cadeias de Markov (CHEN et al., 2021). O comportamento dos usuários do trânsito pode ser considerado caótico e errático, portanto, os modelos estocásticos apresentam uma boa aderência à realidade, mas os modelos microeconômicos são de mais fácil interpretabilidade e contemplam melhor a intuição do problema. Assim, o problema da faixa azul será tratado através das lentes de um modelo de escolha racional, no qual o agente, de forma bem tradicional à literatura microeconômica, maximiza sua utilidade.

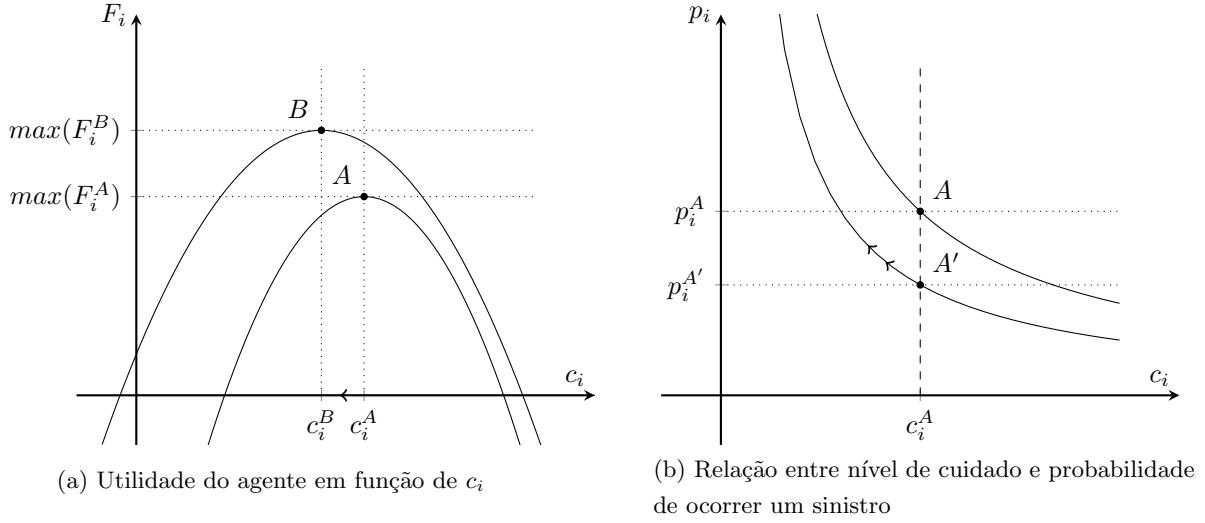
Para basear a modelagem teórica, a principal referência utilizada foi Blomquist (1986), na qual desenvolveu-se um modelo microeconômico de escolha racional apresentado na Equação 1, com F_i representando a utilidade líquida do agente de trânsito, que será maximizada conforme esse escolhe o seu nível ótimo de cuidado ou atenção no trânsito c_i . Outros pesquisadores deram continuidade a este modelo, principalmente incorporando um componente interação estratégica entre os agentes (PEDERSEN, 2003). Entretanto, para os fins deste estudo, optou-se pelo modelo simplificado, com pequenas adaptações.

$$\max_{c_i} F_i = U_i(c_i) - p_i(c_i, s)L_i(c_i, s) \quad (1)$$

O primeiro componente, $U_i(c_i)$, representa a utilidade direta de prestar atenção advinda do nível de cuidado escolhido pelo agente. Na medida em que é alocado maior cuidado, há uma piora no nível de utilidade, visto que o esforço pode envolver tempo, inconveniência, desconforto, energia ou dinheiro para garantir condições mais seguras no trânsito. Um agente pode, por exemplo, escolher um caminho mais longo ou caro, mas que considere mais seguro, ou investir mais dinheiro em medidas de segurança, fazendo revisões mais frequentes.

O segundo componente, $p_i(c_i, s)$, se refere à probabilidade de ocorrer um acidente, dado o nível de cuidado escolhido pelo indivíduo e as condições exógenas de segurança da via ou do veículo, s . Um agente que é mais cuidadoso apresenta menor probabilidade de se envolver em um sinistro, visto que pode adotar medidas que evitem colisões, como manter maior distância ao carro da frente ou sinalizar mudanças de

Figura 4: Estática comparativa do modelo microeconômico



faixa. As condições da via (s) afetam todos os agentes de trânsito, de forma a melhorar a sua segurança e diminuir a probabilidade de ocorrer um sinistro. A iluminação e sinalização da via, por exemplo, são fatores relevantes para garantir a segurança viária. A faixa azul também pode ser considerado um fator que contribui para as condições da via, visto que segregá os usos e diminui os conflitos.

O último componente, $L_i(c_i, s)$, é uma função de perdas, que computa a utilidade perdida quando ocorre um sinistro. Assim como no componente da probabilidade de ocorrer um acidente, o nível de cuidado e a segurança da via podem diminuir as perdas, caso haja um acidente. Uma via com menor velocidade máxima pode fazer com que, caso haja um sinistro, ele seja de menor severidade. Um carro com maior estabilidade, que costuma ser mais caro, pode permitir maior espaço de manobra, possibilitando que o motorista minimize as perdas.

Na Figura 4b é possível observar a dinâmica das variáveis no modelo. Na medida em que aumenta c_i , há uma redução na probabilidade de ocorrer um acidente, entretanto, o efeito marginal do cuidado é decrescente, visto que cada vez fica mais difícil de reduzir o risco. Quando há um choque exógeno que aumenta s , como a adoção da faixa-azul, a curva do *trade-off* cuidado e risco se desloca para baixo, o que significa que, para o mesmo nível de atenção, há uma probabilidade menor de ocorrer um acidente. Na Figura 4b se observaria então um deslocamento do ponto A para A'.

Todavia, os agentes do trânsito, ao observarem esse choque exógeno vão se adaptar e escolher um novo nível de cuidado c_i , que maximize sua utilidade. O modelo trata o nível de cuidado e a segurança como substitutos em certa magnitude, o que indica que, ao observarem uma via mais segura, os agentes se permitirão prestar menos atenção, afinal a segurança “substitui” uma parte dos benefícios da atenção no trânsito ao diminuir tanto a probabilidade quanto a perda causada pelo sinistro. O agente, então, maximizará sua utilidade F_i^B escolhendo um novo nível de cuidado c_i^B , que será inferior ao anterior, como pode ser observado na 4a. Portanto, observa-se um ganho de bem estar quando a escolha do indivíduo muda do ponto A para o ponto B.

Apesar do modelo indicar que o novo nível de cuidado será menor e, no geral, os agentes serão mais “hawkish”, ou agressivos (PEDERSEN, 2003), não é possível prever a partir desse modelo a magnitude

do efeito. No gráfico 4b, sabe-se que os agentes vão escolher um nível de $c_i^B < c_i^A$ e, portanto, a nova probabilidade de ocorrer um acidente será maior do que $p_i^{A'}$. Entretanto, não é possível identificar se será maior, menor, ou igual a p_i^A . O que se tem de resultado desse modelo é um efeito ambíguo, no qual ao mesmo tempo em que a segurança da via diminui a probabilidade de ocorrer um acidente, há também um *moral hazard*, no qual os agentes passarão a ser menos cuidadosos, aumentando a probabilidade de ocorrer um acidente.

Sob a premissa de que um efeito é maior do que o outro, a hipótese a ser testada empiricamente na Seção 3 é de que a faixa azul causa uma redução na probabilidade de ocorrer um acidente. Em outras palavras, caso seja verificada verdadeira a hipótese, a magnitude do efeito da segurança da via é maior do que o efeito da redução no nível de cuidado.

3 Modelagem Empírica

3.1 Dados

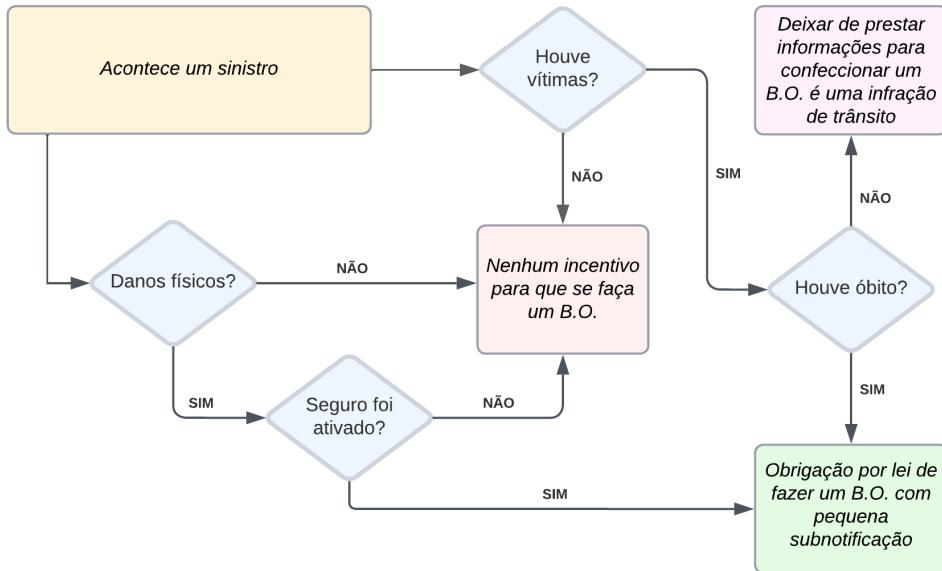
A principal fonte de dados públicos sobre sinistros de trânsito em São Paulo é o Infosiga (Sistema de Informações Gerenciais de Acidentes de Trânsito do Estado de São Paulo), um sistema criado em 2016 e gerenciado pelo Detran-SP. A proposta, quando foi criado, era de apresentar dados de óbitos e acidentes com vítimas, atualizados mensalmente (DOE-SP, 2016), mas o levantamento de sinistros sem vítimas fatais passou a ser realizado apenas a partir de 2019. Os dados do Infosiga são coletados a partir da base de dados do RDO (Registro de Ocorrências), sistema usado pela polícia para registrar os Boletins de Ocorrência (BOs) em escala estadual (DECRETO MUNICIPAL N°58.717, 2019). A partir da análise desses BOs, os dados são tabulados e registrados, dessa forma, caso não haja um BO sobre o acidente, ele não constará na base de dados.

É importante ressaltar que quando há vítimas envolvidas no sinistro, independentemente da natureza da lesão, registrar o BO é uma norma, cujo descumprimento leva à uma multa gravíssima, de acordo com o Art. 176 do Código de Trânsito Brasileiro (CTB). Ademais, mesmo quando não obrigatório – em cenários de sinistros sem vítimas –, o BO pode ser requerido em diversas situações, como pela seguradora, caso seu serviço seja necessário. Quando o sinistro resulta em óbito, é necessário que se faça um BO para enterrar a vítima. Dessa forma, é de se esperar que, para sinistros graves ou que envolvam óbitos, não haja subnotificações. Ao passo que a gravidade diminui, há menores chances de que seja feito o BO e que o evento conste na base de dados. A figura 5 ajuda a compreender os cenários em que os dados estão mais completos.

Há outras fontes de dados que devem também ser mencionadas, sendo que em escala federal há apenas três (ROMÃO; CAMPOS, 2011). O Sistema de Informação sobre Mortalidade (SIM), criado pelo DATASUS, do Ministério da Saúde, divulga informações sobre vítimas de sinistros de trânsito que passaram pelo SUS. O Registro Nacional de Acidentes e Estatísticas de Trânsito (Renaest), sob a coordenação do Departamento Nacional de Trânsito (Denatran) organiza e junta os dados dos Detrans de cada unidade federativa (NOGUEIRA, 2016). Por fim o DPVAT é um seguro obrigatório que existe desde 1974 para a cobertura de danos pessoais causados por sinistros de trânsito, cujos dados são divulgados trimestralmente através de relatórios. O comparativo entre as bases se encontra no Anexo 5.1, mas em suma, os dados do Infosiga são mais completos e são a escolha para este estudo.

Na seção do anexo 5.2 foi feita uma análise mais aprofundada da base dados, que apresenta algumas

Figura 5: Fluxograma de notificação de um sinistro



limitações dos dados, bem como sugestões de melhor organização dessa base. Em seguida se encontra a descrição apenas das variáveis utilizadas na pesquisa:

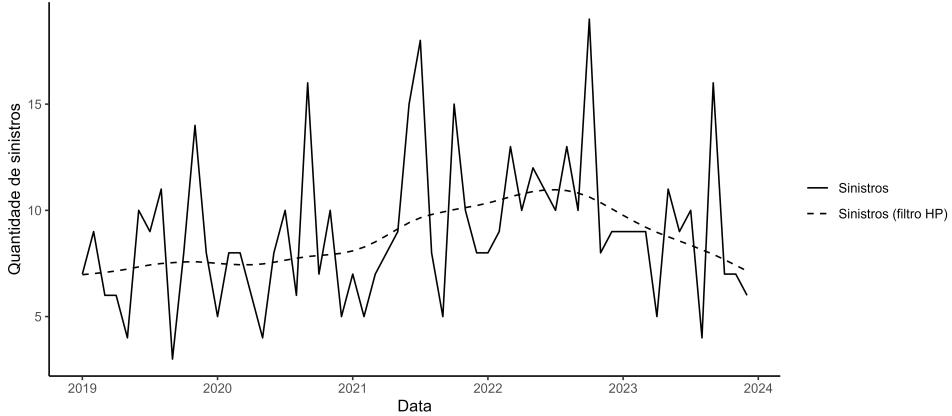
- Data do Sinistro: Inclui ano, mês, dia e horário do sinistro.
- Logradouro e número: Endereço e número do local em que o sinistro ocorreu, da forma como foi registrada no BO.
- Veículos Envolvidos: Existe uma coluna para cada tipo de veículo, em que os valores são binários (o veículo estava presente no sinistro ou não). As possibilidades são automóvel, motocicleta, pedestre, bicicleta, caminhão, ônibus e outros.
- Pessoas Envolvidas: Existem 3 colunas desse tipo, relacionadas a gravidade do quadro clínico, podendo ser ileso, leve ou grave. Em cada coluna é apresentada a quantidade de pessoas que apresentavam a gravidade em questão no sinistro.

Os dados então foram filtrados, agregados e organizados e receberam ajustes pontuais¹. Foram considerados apenas os sinistros no município de São Paulo, que envolveram ao menos um motociclista e que deixaram ao menos um ferido (grave ou leve), de forma a minimizar o viés de sub-notificação, como demonstrado na Figura 5. Depois, os dados foram agrupados por avenida e mês, resultando em um número de sinistros para esse par de chaves únicas. Entretanto, como comentado na Seção 2, o comportamento no trânsito é muito caótico, levando a série a apresentar muito ruído – que prejudica a análise. Para lidar com isso, aplicou-se a raiz quadrada e o filtro Hodrick-Prescott (HP) com o objetivo

¹Como é possível observar na Figura 1, o mês de março de 2023 apresenta zero sinistros de carros e caminhões no município de São Paulo, o que é impossível. Nesse sentido, considerando que isso é um erro na base de dados, os dados desse mês foram interpolados utilizando a média dos três meses anteriores.

de isolas as tendências de longo prazo dos componentes cíclicos de curto prazo nos dados de sinistros, além de estabilizar a variância e diminuir o efeito de *outliers*. O resultado desse procedimento para a Avenida Bandeirantes pode ser observado na Figura 6.

Figura 6: Sinistros na Avenida dos Bandeirantes



A Avenida 23 de Maio, apesar de ser a primeira a receber o tratamento, apresentou faixa azul apenas em um sentido, mas na base de dados do Infosiga não é discriminado em qual sentido da via ocorreu o sinistro, impossibilitando que a análise seja feita sobre ela. A Avenida dos Bandeirantes foi implementada poucos meses depois, em sua completude, sendo o caso perfeito para este estudo. Em 2023 outras faixas azuis começaram a surgir, mas dado que foram implementadas apenas recentemente, não há meses suficientes para que seja feita uma análise robusta. Dessa forma, a única unidade tratada na análise é a Av. dos Bandeirantes.

Em relação aos dados de fluxo de veículos, a CET-SP divulga anualmente um relatório contendo a lentidão de trechos de vias de grande circulação em São Paulo. Esses dados são coletados a partir dos radares espalhados pelas vias e a Avenida 23 de Maio, por exemplo, está categorizada junto com a Avenida Rubem Berta e Avenida Moreira Guimarães, apresenta diferentes trechos: da rua Aratás até o Viaduto João João Julião da Costa Águia, do Viaduto da Rua Pedroso até o Terminal Bandeira, do Viaduto General Euclides de Figueiredo até a Avenida Indianópolis, entre outros. Ademais, a lentidão dos trechos está separada para o sentido da via, e, no caso da Av. 23 de Maio, é separado por Santana ao Aeroporto ou o contrário.

Todavia, a base de dados apresenta uma série de problemas que impossibilitam sua utilização. Primeiramente, as ruas não apresentam um código identificador que possa ser cruzado com outra base e os nomes estão escritos de maneira inconsistente e que pode variar de ano para ano. A medida apresentada de “lentidão” não tem seu significado apresentado nem uma metodologia de como é calculada, tornando-a uma variável sem muito valor. Ademais, os valores estão incompletos, apresentando grandes períodos em que não há dados. Por fim, a separação de trechos é feita de tal forma que uma única avenida pode apresentar dezenas ou centenas de trechos, impossibilitando que se identifique um trecho que represente a avenida como um todo, dado que é necessário fazer esse processo para um grande contingente de avenidas.

3.2 Identificação

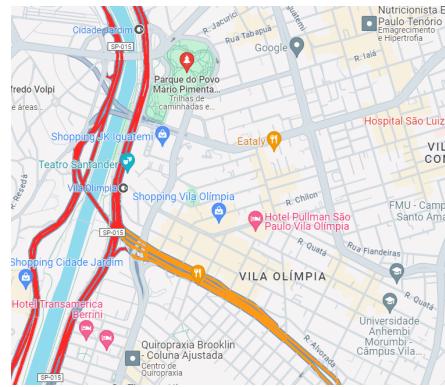
Para fazer a identificação causal do efeito da faixa azul, o método escolhido neste estudo foi o de controle sintético, apresentando como principal referência as contribuições de Abadie, Diamond e Hainmueller (2010). A escolha por esse método se encontra no fato de haver apenas uma unidade amostral tratada. Essa abordagem envolveu a criação de uma versão sintética da Avenida Bandeirantes, uma das principais vias de São Paulo. A Avenida Bandeirantes sintética foi modelada a partir de uma combinação linear de dezenas de outras grandes vias urbanas que não receberam a implementação da faixa azul, mas que possuem características semelhantes em termos de tráfego, estrutura urbana, e outros fatores relevantes.

Figura 7: Construção do controle sintético

(a) Componentes da Bandeirantes sintética

Nome da Via	Peso
Av. Mutinga	30.26%
Sp 015	20.23%
Ace. Estr. do M'Boi Mirim	05.59%
Sp 270	01.74%
Av. Sen. Teotônio Vilela	01.41%
Estr. De Itapecerica	01.14%
Av. Sapopemba	01.26%
Av. Aricanduva	01.17%
Av. Atlântica	01.06%
Av. Washington Luis	01.01%
Av. Interlagos	00.96%
Estr. Do Campo Limpo	00.82%
Outras avenidas	33.70%

(b) Av. Bandeirantes (laranja) e SP015 (vermelha)

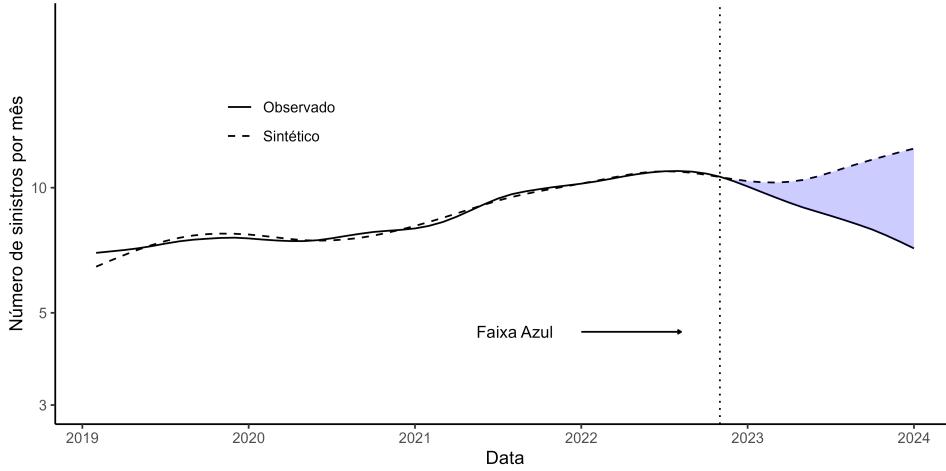


A escolha de quais avenidas e qual peso cada uma delas apresenta no controle sintético é calculado automaticamente pelo método, de forma a minimizar a diferença entre a Bandeirantes observada (real) e a sintética antes do período de tratamento. Os principais componentes da Bandeirantes sintética podem ser visualizadas na Figura 7a. Vale salientar que a SP 015, segundo componente mais relevante, é uma continuação física da Av. dos Bandeirantes (Figura 7b), que troca de nome depois de um viaduto. Isso é uma evidência qualitativa de que o método acabou por escolher uma avenida que compartilha de muitas características com a Av. dos Bandeirantes, a não ser o fato de receber o tratamento. Evidências quantitativas serão discutidas adiante.

Caso tenha sido bem sucedido o método, com essa combinação linear de avenidas, têm-se a Bandeirantes sintética, que é igual a observada, caso esta não tivesse recebido o tratamento. O resultado pode ser visto na Figura 8. Dessa forma, é possível fazer uma comparação entre ambas para identificar o efeito do tratamento. A métrica utilizada para isso é o *root mean squared prediction error* (RMSPE), que calcula soma da distância entre a unidade sintética e verdadeira para cada período de tempo antes do tratamento (RMSPE pré) e depois do tratamento (RMSPE pós). Essa comparação pode ser observada na Figura 9a. Caso o RMSPE pré seja alto, significa que o método não conseguiu formar uma unidade sintética que seja parecida com a observada, e no caso em que o RMSPE pós seja alto, isso pode ser indicativo de um efeito do tratamento. É importante destacar que ao minimizar o RMSPE pré, o método apresenta como variáveis disponíveis apenas o número de sinistros dos períodos pré tratamento.

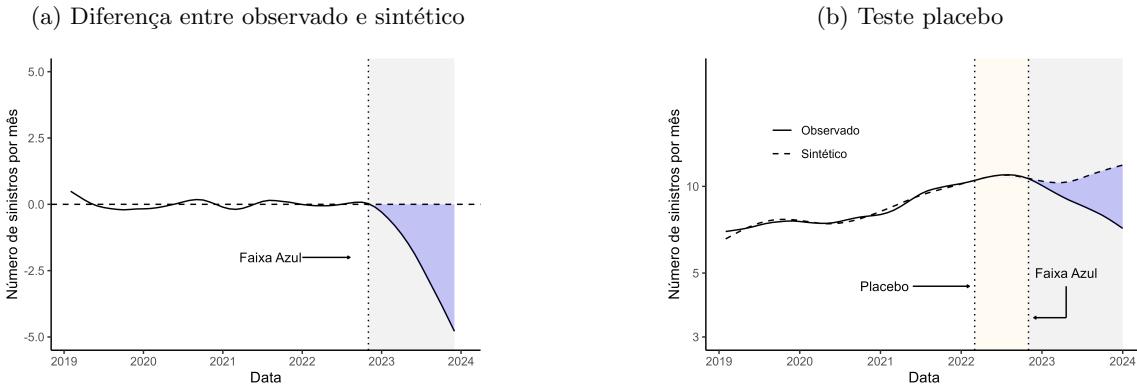
Como é de praxe na literatura de modelo sintético, foi feito um teste placebo para avaliar se o método identificaria efeito espúrio. Para tanto, foi considerado que o período de tratamento aconteceu 8 meses

Figura 8: Bandeirantes observada e sintética



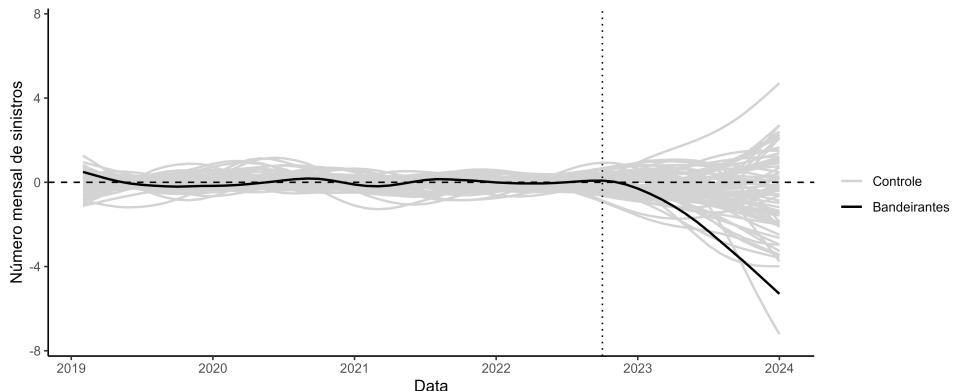
antes do real tratamento. Com isso, agora na minimação do RMSPE pré, há 8 meses de dados a menos disponíveis e é possível calcular um RMSPE pós', mas apenas para os 8 meses que foram removidos (sem contar com os meses de verdadeiro tratamento). Caso esse RMSPE pós' seja maior do que do RMSPE pré, o método está identificando efeito em períodos nos quais não houve tratamento – uma evidência de que a unidade sintética não é tão parecida com a verdadeira. Na Figura 9b é possível observar que no período de 8 meses mencionadas não houve efeito, uma evidência favorável à validade do método.

Figura 9: Análise do controle sintético



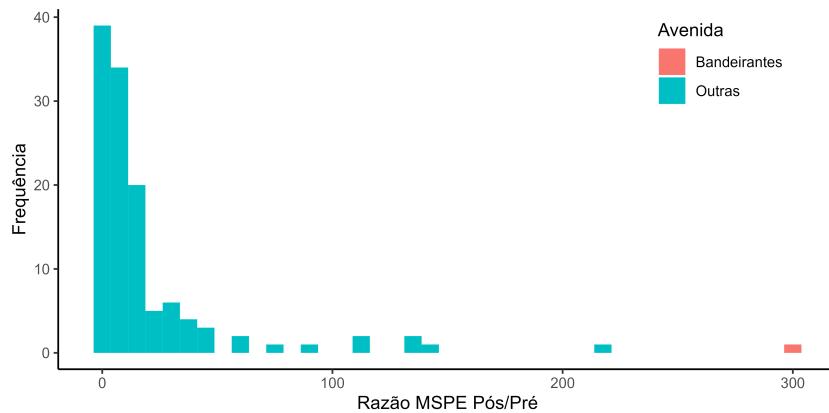
Um outro tipo de teste placebo pode ser feito, no qual ao invés de manipular qual data é considerada o começo do tratamento, é possível mudar qual avenida recebeu a faixa azul – procedimento que recebe o nome de teste da permutação. Caso se identifique efeito maior em avenidas que não receberam faixa azul, se comparadas à Bandeirantes, isso é uma evidência de que o efeito observado do tratamento na Bandeirantes pode ser espúrio, já que as outras não receberam tratamento. Para avaliar o efeito em cada avenida, é construído um controle sintético para cada uma delas que estão presentes no *pool* de opções – a mesma que foi utilizada para a construção do controle sintético da Bandeirantes. Os resultados podem ser vistos na figura 10. Nesse *pool* foram consideradas avenidas que apresentam mais de 100 sinistros desde o início da disponibilidade de dados, resultando em 122 avenidas.

Figura 10: Resultado da permutação de controle sintético



Depois, é utilizada uma métrica na qual se divide o RMSPE pós pelo RMSPE pré para cada avenida disponível no *pool*. Quanto maior for o resultado dessa razão, mais significante é o efeito do tratamento. Entretanto, seria problemático caso unidades não tratadas tenham uma razão mais alta do que a tratada, visto que elas não receberam o tratamento. Com isso, são ranqueadas todas as unidades do *pool* de maior à menor razão RMSPE pós/pré. Com base nesse ranqueamento, é calculado o p-valor do método.

Figura 11: Ranqueamento das razões RMSPE pós/pré



Como é possível observar na Figura 11, a Avenida dos Bandeirantes ficou em primeiro lugar no ranking, o que é um forte indicativo de que a medida foi bem sucedida. É possível fazer uma interpretação semelhante ao clássico p-valor das estatísticas construídas com regressões lineares. Caso fosse amostrada uma nova avenida a partir da permutação aleatória de avenidas da *pool*, a probabilidade de uma apresentar o mesmo ou maior efeito, se comparada à Bandeirantes é 1/122, que equivale a um “p-valor” de 0.82%. Entretanto, essa *pool* pode ser considerada muito grande e com avenidas que não tiveram um bom controle sintético formado. Nesse sentido, a literatura recomenda a remoção de unidades amostrais com erros pré-tratamento muito grandes.

Em Abadie, Diamond e Hainmueller (2010), é colocado como valor de corte as unidades da *pool* que apresentaram erro pré-tratamento 20 vezes maior do que a unidade tratada, sendo considerada uma *lenient cutoff* pelos autores. Quanto mais leniente for o valor de corte, mais significativo se torna o efeito

da faixa azul. Quando imposto este limite, o novo p-valor torna-se 1.35%, já com um limite de 10 vezes a unidade tratada (uma medida mais conservadora), têm-se um p-valor de 2.22%.

4 Conclusão

Diante da análise empírica realizada sobre os sinistros de trânsito na Avenida dos Bandeirantes, em São Paulo, após a implementação da faixa azul, algumas conclusões podem ser destacadas. As evidências apontam que a faixa azul causou uma redução no número de sinistros envolvendo pessoas feridas. Desde a sua inauguração, em outubro de 2022 até dezembro de 2023², estima-se que foram evitados 32 sinistros que envolveriam motociclistas e ao menos uma pessoa ferida de forma grave ou leve, o que representa uma redução de 26.8% dos sinistros registrados no período pós tratamento.

Apesar de demonstrar significância estatística, a análise apresenta uma série de limitações. Primeiramente e mais importante, foi identificado um efeito local, que não necessariamente pode ser extrapolado para outras avenidas. Em segundo, não foi estimado o custo de implementação da faixa azul ou uma análise de custo benefício dessa política pública. Além disso, pelos poucos períodos pós-tratamento disponíveis, a longevidade do efeito não foi estimada, sendo possível que o efeito da faixa azul se dissipe ou se fortaleça ao longo dos próximos meses. Por fim, não foi feita uma análise de transbordamentos da medida.

Seria interessante avaliar qual o impacto da faixa azul sobre o fluxo de veículos, bem como sua velocidade – especialmente para o caso das motocicletas. Possivelmente, a medida pode ter incentivado o uso de mais motocicletas, o que vai na contramão do que se propõe em cidades modernas de incentivar o transporte público coletivo, que é mais seguro e possui melhores externalidades.

Em suma, o estudo apresenta evidências que indicam o sucesso da faixa azul na Avenida dos Bandeirantes segundo o critério de redução do número de sinistros envolvendo vítimas com ferimentos leves ou graves. Entretanto, para inferir sobre os benefícios da faixa azul em escala municipal, estadual ou federal, é necessário analisar mais avenidas que receberam o tratamento. Diversas avenidas estão recebendo a implementação da faixa azul³ e na medida em que os dados mais recentes se tornarem disponíveis, serão viáveis estudos para analisar o impacto da medida.

²Os dados estão disponíveis até este instante, consultado em 18/01/2024. A divulgação dos dados é defasada em alguns meses

³A primeira faixa azul foi implementada na Avenida 23 de Maio, em janeiro de 2022; seguida da Avenida dos Bandeirantes, em outubro de 2022; Avenida Prestes Maia em outubro de 2023; Avenidas Sumaré, Paulo VI, Miguel Yunes e Av. das Nações Unidas em novembro de 2023; Avenidas Brigadeiro Faria Lima, Luiz Dumont Villares, Zaki Narchi, Jacu Pêssego e do Estado em dezembro de 2023.

Referências

- ABADIE, Alberto; DIAMOND, Alexis; HAINMUELLER, Jens. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. **Journal of the American statistical Association**, Taylor & Francis, v. 105, n. 490, p. 493–505, 2010.
- BAVOSO, Natália Couto. O sistema nacional de trânsito e os municípios de pequeno porte. **Universidade Federal de Minas Gerais**, 2014.
- BLOMQUIST, Glenn. A utility maximization model of driver traffic safety behavior. **Accident Analysis & Prevention**, Elsevier, v. 18, n. 5, p. 371–375, 1986.
- CET. Apresentação da Campanha Educativa FaixaAzul. **CETSP**, 2021. Acesso em: 11 set. 2023.
- _____. **Faixa Azul completa 1 ano sem registros de mortes onde foi implantada**. 2023. Disponível em: [jhttp://www.cetsp.com.br/noticias/2023/01/25/faixa-azul-completa-1-ano-sem-registros-de-mortes-onde-foi-implantada-\(1\).aspx](http://www.cetsp.com.br/noticias/2023/01/25/faixa-azul-completa-1-ano-sem-registros-de-mortes-onde-foi-implantada-(1).aspx);. Acesso em: 15 set. 2023.
- CHEN, Qinghong et al. Modeling accident risks in different lane-changing behavioral patterns. **Analytic methods in accident research**, Elsevier, v. 30, p. 100159, 2021.
- DECRETO MUNICIPAL N°58.717. Plano de segurança viária do município de São Paulo, 2019.
- DESPACHANTE DOK. **Faixa azul SP: entenda como funciona o projeto**. 2022. Disponível em: [jhttps://www.despachantedok.com.br/blog/veiculo/faixa-azul-sp/](https://www.despachantedok.com.br/blog/veiculo/faixa-azul-sp/);. Acesso em: 15 set. 2023.
- DOE-SP. Resolução SG-6, de 23-2-2016. **Diário Oficial do Estado de São Paulo**, Palácio dos Bandeirantes, Av. Morumbi 4.500, 24 fev. 2016. Acesso em: 13 out. 2023.
- JUN, Wang; HENG, Wei. Traffic segregation on spatial and temporal bases: the experience of bicycle traffic operations in China. **Transportation research record**, v. 1396, p. 11, 1992.
- MARANHÃO, Fabiana. **Prefeitura decide desativar motofaixas em São Paulo**. 2013. Disponível em: [jhttps://noticias.uol.com.br/cotidiano/ultimas-noticias/2013/11/18/prefeitura-decide-desativar-motofaixas-em-sao-paulo.htm](https://noticias.uol.com.br/cotidiano/ultimas-noticias/2013/11/18/prefeitura-decide-desativar-motofaixas-em-sao-paulo.htm);. Acesso em: 15 set. 2023.
- NBR. NBR 10697: Pesquisa de sinistros de trânsito — Terminologia. Rio de Janeiro, 2020.
- NOGUEIRA, André Fernando da Silva. **Morte no trânsito não é acidente: por que o registro nacional de acidentes e estatísticas de trânsito precisa sair do papel?** [S.l.], 2016.
- PEDERSEN, Pål Andreas. Moral hazard in traffic games. **Journal of Transport Economics and Policy (JTEP)**, Journal of Transport Economics e Policy, v. 37, n. 1, p. 47–68, 2003.
- RADIN SOHADI, Radin Umar; MACKAY, Murray; HILLS, Brian. Multivariate Analysis of Motorcycle Accidents and the Effects of Exclusive Motorcycle Lanes in Malaysia. **Journal of Crash Prevention and Injury Control**, Taylor & Francis, v. 2, n. 1, p. 11–17, mar. 2000. ISSN 1028-6586. DOI: 10.1080/10286580008902549.
- ROMÃO, Magaly Natália Pazzian Vasconcellos; CAMPOS, Cintia Isabel de. Análise comparativa dos bancos de dados disponíveis no Brasil sobre vítimas fatais em acidentes de trânsito. **18º Congresso Brasileiro de Transporte e Trânsito**, 2011.
- SAINI, Harish Kumar; CHOUHAN, Shivam Singh; KATHURIA, Ankit. Exclusive motorcycle lanes: A systematic review. **IATSS research**, Elsevier, v. 46, n. 3, p. 411–426, 2022.

VASCONCELLOS, Eduardo Alcântara de. A cidade, o transporte e o trânsito. **São Paulo: Prolivros**, 2005.

WANG, Chao; QUDDUS, Mohammed A; ISON, Stephen G. The effect of traffic and road characteristics on road safety: A review and future research direction. **Safety science**, Elsevier, v. 57, p. 264–275, 2013.

5 Anexo

5.1 Comparativo Bases de Dados

O DPVAT por muitos anos era uma responsabilidade da Seguradora Líder, mas a partir de 2021 teve administração instável, visto que Seguradora Líder foi dissolvida. Em 2023 a Caixa Econômica Federal torna-se responsável por gerenciar o DPVAT. Os dados do DPVAT sobre sinistros são divulgados em forma de relatório, sem que haja microdados para uma análise mais aprofundada. Enquanto sob a gestão da Seguradora Líder, havia anualmente um boletim estatístico⁴, no qual há uma separação por unidade federativa, o que não é mais feito sob a gestão da Caixa – encarregada pelos dados a partir de 2021.

Tabela 1: Comparativo dados de óbitos para diferentes fontes no estado de São Paulo

Ano	DPVAT	DATASUS	Infosiga	Infosiga*	Renaest
2014	9.093	7.444			
2015	6.884	6.270	6.493	6.168	
2016	5.248	5.846	5.966	5.651	
2017	6.103	5.462	5.649	5.350	
2018	5.462	4.730	5.464	5.169	5.233
2019	6.026	5.181	5.422	5.123	5.164
2020	4.972	5.326	4.950	4.647	4.838
2021		5.416	4.925	4.649	4.847
2022		4.818	5.738	5.455	3.475

* Removendo as rodovias federais

Com base na Tabela 1, é possível observar as diferenças entre as bases de dados disponíveis e alguns detalhes merecem atenção. Primeiramente, é curioso que os números do DPVAT sejam maiores do que do Infosiga e Renaest, visto para ativar o seguro, é necessário entregar um Boletim de Ocorrência, que então constaria nessas bases. Uma possível explicação para essa inconsistência é de que o DPVAT é registrado quando o seguro é exercido, não quando ocorre o acidente, resultando em uma defasagem temporal.

Outro ponto que merece ser investigado é o fato de que em 2020 e 2021 os óbitos registrados pelo SUS superam os do Infosiga. Em teoria, todos os óbitos em decorrência de acidente de trânsito devem ser registrados no RDO (fonte de dados do Infosiga) para possibilitar o enterro da vítima, então se uma pessoa morreu no SUS, essa informação deveria estar disponível também no Infosiga. Portanto, os dados do DATASUS em teoria nunca deveriam superar os do Infosiga, principalmente se considerar que a vítima pode ser encaminhada para um hospital que não faz parte da rede SUS e, neste caso, seria contabilizado apenas no Infosiga.

Por fim, uma última inconsistência é observável no comparativo do Infosiga* com o Renaest. O Renaest não contabiliza sinistros em rodovias federais, então as BRs 381, 459, 101, 153, 116, 383, 158 e 488 foram removidas da base do Infosiga para fins de comparação. Como ambas as bases apresentam as mesmas fontes, os dados deveriam ser idênticos, mas nos anos de 2018 a 2021 há uma diferença de cerca de 500 óbitos.

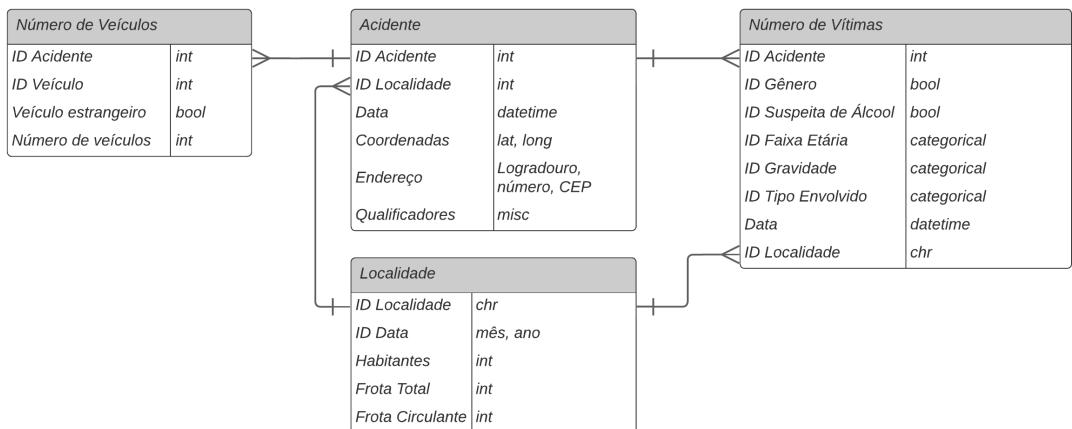
⁴Disponível em <https://seguradoralider.com.br/Sala-de-Imprensa/Boletim-Estatistico>, acesso em 15/10/2023

5.2 Descrição base Renaest

Tanto a base de dados do Infosiga quanto a do Renaest são alimentadas pela mesma fonte: os boletins de ocorrência da polícia. Entretanto, a forma como são divulgados é diferente, pelo Infosiga, de forma agregada em uma base de óbitos, no qual cada linha é uma vítima fatal, e uma base de não fatais, na qual cada linha é um acidente. É importante destacar que a forma como estes dados são agregados gera uma série de limitações, visto que com os microdados é possível investigar perguntas mais específicas e com maior profundidade.

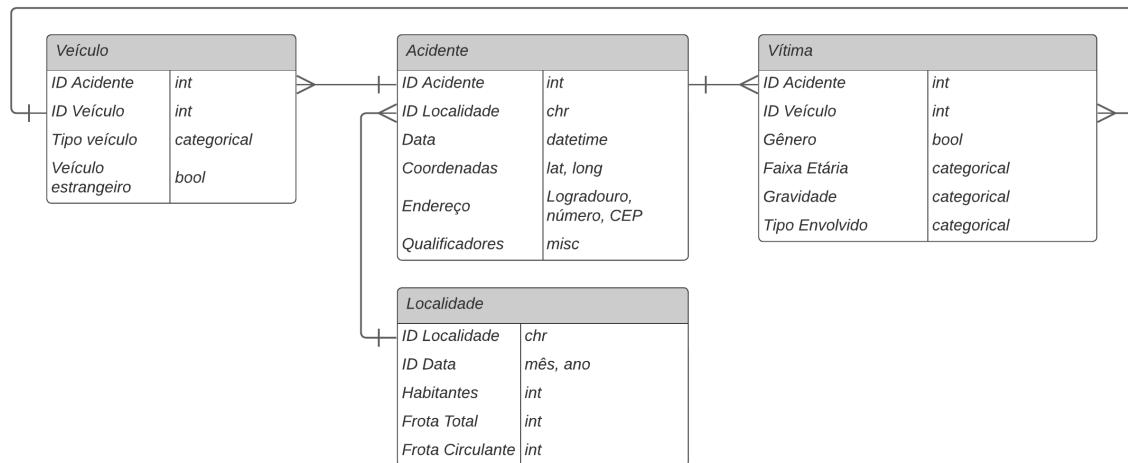
Já a base disponibilizada pelo Renaest, apresenta um banco relacional, descrito na Figura 12, que apresenta dados menos agregados, mas ainda com limitações. Da forma como foram organizados os dados, se por exemplo uma motocicleta, um carro, um ônibus e um pedestre se envolvem em um sinistro, é possível saber se havia suspeita de álcool e quantas pessoas foram feridas. Entretanto, se houve uma fatalidade, não é possível saber em qual veículo essa vítima estava, bem como suas respectivas informações – idade, gênero, se era motorista, se estava embriagada, etc.

Figura 12: Diagrama entidade relacionamento da base do Renaest



Considerando isso e a importância dessas informações para possíveis pesquisas dentro do universo da segurança viária, foi construído um diagrama (Figura 13) que contém uma proposta de como estes dados deveriam estar dispostos. Com a estrutura apresentada, os problemas apontados anteriormente não aconteceriam mais.

Figura 13: Diagrama entidade relacionamento, sugestão para dados do Renaest



Crescimento da população em situação de rua e a capacidade de atendimento dos Centros POP em 2021 no município de São Paulo: um estudo descritivo.

Políticas Públicas

Pesquisadores: Ana Beatriz Parra Ferreira, Arthur Sóter Assis, Giulia Beatriz Brombine Alves Rodrigues
Orientadora: Laura Abreu

Resumo

O presente estudo investiga a capacidade de atendimento dos Centros POP em 2021 em relação ao crescimento da população em situação de rua em São Paulo nos últimos anos. Para isso, utilizam-se dados de Registros Mensais de Atendimentos (2017-2021) e Censos da população em situação de rua (2015-2021) a fim de se analisar a oferta e demanda desse equipamento de políticas públicas. A partir desse trabalho, é possível concluir que a capacidade de atendimento dos Centros POP não foi suficiente para suprir a demanda advinda do aumento significativo da população em situação de rua em São Paulo no ano de 2021.

Palavras-chave: População em situação de rua, Centro POP

1 Introdução

O Artigo 5 da Constituição Federal brasileira é frequentemente considerado a espinha dorsal dos direitos fundamentais e liberdades individuais do país. Este conjunto de princípios e garantias constitui a base para uma sociedade mais justa e igualitária, na qual os cidadãos devem gozar de proteção legal e respeito aos seus direitos. No entanto, existe uma parcela vulnerável da população que parece não gozar das mesmas garantias constitucionais, pois, enfrenta constantemente a violação flagrante de todos os direitos previstos pela Carta Magna. Trata-se da população em situação de rua.

Definida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) como um “Grupo populacional heterogêneo constituído por pessoas que possuem em comum a garantia da sobrevivência por meio de atividades produtivas desenvolvidas nas ruas, os vínculos familiares interrompidos ou fragilizados e a não referência de moradia regular”, a população em situação de rua vivencia diariamente manifestações extremas da desigualdade e da vulnerabilidade social. Pessoas nessa condição enfrentam desafios diáriamente para satisfazer necessidades básicas, como abrigo, alimentação, saúde e segurança (SERAFINO; LUZ, 2015). De modo síncrono, este grupo é regularmente vítima de discriminação, violência e negligência, tendo violados os seus direitos fundamentais, como a igualdade perante a lei, a integridade física e moral, a dignidade e a inviolabilidade da vida.

De tal modo, planos de ações e políticas públicas voltados a Pop. Rua são construídos e executados, com intuito de fornecer condições básicas de sobrevivência, como acesso a refeições, higiene, atividades de lazer e convivência, majoritariamente, por meio de um equipamento de assistência social. Em um estudo sobre a relação Estado - Pop. Rua brasileira, (BARBOSA, 2018) faz uso de fontes documentais e históricas para demonstrar como este segmento populacional foi inserido, apenas, em meados dos anos 2000 na agenda do governo federal no que se refere à estruturação de políticas públicas de inclusão e proteção social, divergindo das anteriores iniciativas de repressão e controle. De tal forma, verifica-se a importância de se compreender o funcionamento e estado atual das estratégias e equipamentos oferecidos à esta parcela, historicamente à margem das prioridades dos poderes públicos.

Consoante à esta realidade, este grupo sofreu um crescimento populacional expressivo nos últimos anos (SILVA; NATALINO; PINHEIRO, 2020). Devido à falta de abundância em dados oficiais sobre a população nas ruas, modelos lineares são utilizados para estimar essas pessoas. Em pesquisas realizadas pelo Ipea, estima-se que a população brasileira em situação de rua era composta por aproximadamente 102 mil pessoas em 2015 (NATALINO, 2016). Em 2022, utilizando o mesmo modelo linear, a população em situação de rua era composta por aproximadamente 281,5 mil pessoas, de forma que em uma década o crescimento foi de 211% .

Uma vez que o Brasil não conta com dados oficiais sobre a população em situação de rua, majoritariamente devido a dificuldade de obtenção de dados, e destacando como tal fato reproduz e invisibiliza socialmente a Pop. Rua no âmbito das políticas sociais, tornando marginal e ou inexistente o planejamento, a inclusão e o acolhimento destes, enfatiza-se a relevância do uso das informações disponíveis sobre a mesma em prol da produção de uma análise descritiva capaz de gerar evidências sobre as relações de oferta e demanda de equipamentos de assistências sociais que a atendem.

A população em situação de rua se refere a um grupo cuja assistência social requer um equipamento de nível de média a alta complexidade. O principal equipamento para atender esse grupo é o Centro de Referência Especializado para População em Situação de Rua (Centro POP), ou, na ausência de uma unidade de Centro POP na região ou município, o Centro de Referência Especializado de Assistência Social (CREAS). Inicialmente, para a análise da cidade de São Paulo, as unidades do CREAS presentes nas zonas municipais que não contam com a presença de um Centro POP foram consideradas como objeto de estudo. Entretanto, após uma devida análise dos dados, utilizar o CREAS como ferramenta de pesquisa foi uma abordagem desconsiderada, assim como será detalhado na seção 5.

Desta maneira, o seguinte estudo teve como principal propósito desenvolver uma análise descritiva de Centros POP na cidade de São Paulo, examinando as relações de oferta e demanda dos atendimentos nas unidades presentes no município.

1.1 Centros POP

Os Centros de Referência Especializado para População em Situação de Rua, popularmente conhecidos como Centros POP, representam um elemento fundamental na estrutura de políticas públicas voltadas para indivíduos em condições de vulnerabilidade. Estes centros são projetados para fornecer uma gama de serviços que vão além do atendimento das necessidades básicas, desempenhando um papel crucial na reintegração social e na garantia de direitos dessa população. O funcionamento das unidades é garantido a partir do aporte do Governo Federal, juntamente com estados e municípios.

Os serviços oferecidos pelos Centros POP incluem acesso a alimentação, higiene pessoal e fornecimento

de roupas, essencial para a promoção da dignidade dessas pessoas. Além disso, o atendimento psicosocial oferecido nesses centros aborda questões de saúde mental e fornece suporte emocional, considerando os desafios únicos enfrentados por aqueles que vivem nas ruas. A assistência para regularização de documentação é outra funcionalidade de alta relevância, pois é um passo fundamental para o acesso a outros direitos e serviços públicos. Vale também destacar que os Centros POP realizam o encaminhamento dos indivíduos que atendem para serviços de saúde e oportunidades educacionais, além de promoverem atividades voltadas para a reintegração social e profissional, como oficinas, cursos e atividades culturais voltadas à Pop. Rua.

Nesse estudo, a escolha de focar nos Centros POP se deu pela sua relevância direta e significativa para a população em situação de rua. Isso porque estes centros não foram pensados apenas para atenderem às suas necessidades imediatas de subsistência, mas também para oferecerem atendimentos com profissionais de assistência social, psicologia, sociologia, terapia ocupacional entre outros, a fim de construir caminhos para a superação de seus desafios complexos diários.

Sendo assim, vale destacar que no município de São Paulo existem seis unidades desse equipamento em diferentes bairros:

1. Centro POP BELA VISTA
2. Centro POP SANTA CECÍLIA
3. Centro POP MOOCA
4. Centro POP VILA MARIA
5. Centro POP SANTANA
6. Centro POP SANTO AMARO

Todas foram incluídas no estudo. De cada uma, analisou-se os registros mensais de atendimento (RMA's) e a quantidade de profissionais em seu quadro de colaboradores por profissão.

Por fim, vale dizer que os Centros POP representam um ponto de intersecção entre as necessidades imediatas da população em situação de rua e as políticas públicas destinadas a atendê-la. Ao analisar a relação de oferta e demanda desse equipamento, pode-se compreender de que maneira essa iniciativa governamental de fato chega "na ponta" e se a quantidade de atendimentos registrados variou tanto quanto os números de pop rua no ano de 2021 na cidade de São Paulo.

2 Revisão Literária

Por meio de entrevistas, Raupp e Adorno (2015) demonstram, a prevalência da violência e da violação de direitos pela parcela entrevistada que correspondia à população em situação de rua. Esse dado deixa claro que o acesso à políticas públicas de qualidade são essenciais para a proteção e garantia de dignidade à Pop. Rua do município de São Paulo.

Em outro paper, Serafino e Luz (2015) discutem o fenômeno da maior concentração de população sem-abrigo nos centros dos municípios devido à presença de trabalhos informais que representam possibilidades de renda e à disponibilidade de infraestruturas urbanas, como praças, pontes e vias nas quais essa população pode estabelecer sua moradia improvisada.

Vale destacar que a regulamentação da população de rua no Brasil ocasionou, não apenas a criação de equipamentos públicos e planos de articulação de combate ao crescimento da população de rua, mas também trouxe-os a um maior nível de visibilidade política, o que, em tese garantiria a continuidade de programas como o Centro POP, além de ser apontado como um marco da cidadania por estudiosos do tema (MEDEIROS et al., 2020).

Contudo, sabe-se que as políticas sociais não necessariamente atingem todos os objetivos para os quais foram elaboradas e, pensando nisso, este estudo foi norteado pela seguinte questão: A capacidade de atendimento dos Centros POP acompanhou a demanda crescente do município de São Paulo em 2021?.

Vale destacar por fim, que parte da metodologia do presente artigo foi baseada no trabalho da Secretaria de Desenvolvimento Social da Criança e Juventude do Governo do Estado de Pernambuco (SDSCJ), principalmente porque os seus pesquisadores também realizaram o cruzamento entre a oferta dos serviços oferecidos pelos Centro POP dos municípios de Pernambuco e a demanda, representada pelo censo da população em situação de rua. (TRABALHO; PERMANENTE, s.d.).

3 Metodologia

Dada a relevância dos Centros Pop enquanto equipamentos de assistência social voltado à população em situação de rua, é crucial que esse público tenha acesso aos atendimentos ofertados por esses centros a fim de ter assegurados alguns de seus direitos mais básicos como alimentação e regularização de documentos. Partindo desse ponto, esse trabalho buscou compreender a relação entre oferta e demanda por tais atendimentos nesses equipamentos.

Para isso, foram analisadas as bases de dados de Registros Mensais de Atendimentos dos Centros Pop dos anos de 2017 a 2021 - Dado considerado a oferta de atendimentos desse equipamento para fins desse estudo, além do Censo da população em situação de rua do município de São Paulo dos anos de 2015 a 2021 - Considerado a demanda total pelos serviços oferecidos pelo Centro Pop.

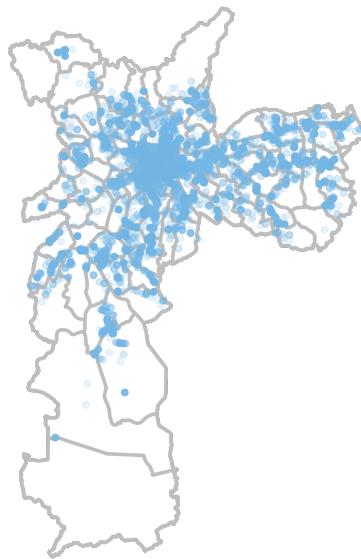
Vale destacar que a opção metodológica por esses dados para oferta e demanda se deram principalmente, porque o Censo se mostrou a melhor base para descrever a pop rua e seu crescimento dos últimos anos e, além disso, porque o Registro mensal de atendimentos já foi analisado em estudos de oferta e demanda de Centros Pop. (TRABALHO; PERMANENTE, s.d.).

Por fim, realizou-se uma comparação entre as taxas de crescimento tanto da oferta (RMA dos Centros Pop) quanto da demanda (Número de pessoas em situação de rua) a fim de compreender se a capacidade de atendimento do Centro POP acompanhou a demanda crescente do município de São Paulo no ano de 2021.

4 Dados

A escassez de dados oficiais atualizados é uma realidade que permeia o cenário de estudos sobre a Pop. Rua, de forma que a disponibilidade de dados teve influência na extensão geográfica avaliada no estudo. De modo relacionado, o município de São Paulo possui a maior concentração de pessoas em situação de rua no país e dispõe uma quantidade de dados e detalhamentos sobre a Pop. Rua maior do que os outros distritos brasileiros. O setor de pesquisa de SPGEO é responsável pelo eixo da Vigilância Socioassistencial, produzindo análises das informações territorializadas sobre as situações de risco e vulnerabilidade. De tal forma, uma vez que censos e relatórios sobre a população em situação

Figura 1: Distribuição da população em situação de rua em São Paulo



de rua são disponibilizados pela prefeitura de São Paulo, o acesso à tais bases de dados favoreceu uma análise em escala municipal.

Analogamente, o mesmo fator influenciou a análise temporal do estudo. Para a população de São Paulo presente nas ruas, a disponibilidade de informações censitárias se limitam aos anos de dentro do período de 2015 até 2021, o ano do último censo publicado até o momento da pesquisa.

Para o estudo, 2 principais fontes foram utilizadas: o Censo da População em situação de rua e o Censo SUAS. O primeiro é disponibilizado pela prefeitura de São Paulo, enquanto o Censo SUAS é um processo de monitoramento que coleta dados por meio de um formulário eletrônico preenchido pelas Secretarias e Conselhos de Assistência Social dos Estados e Municípios, sendo disponibilizado pela Secretaria Nacional de Assistência Social (SNAS).

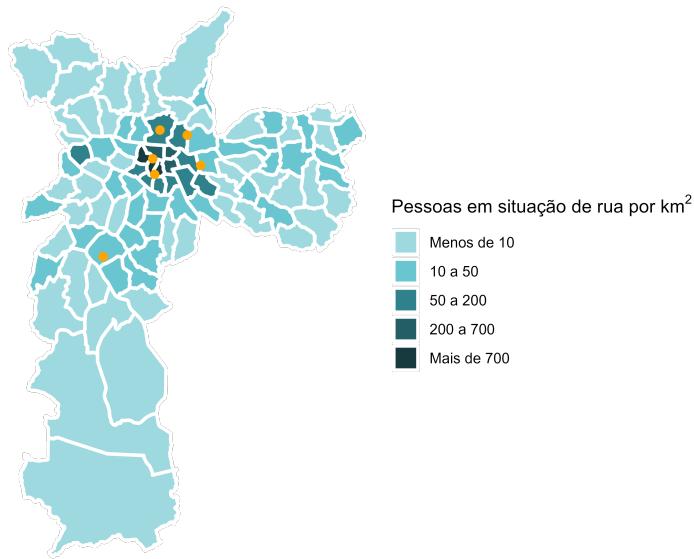
As bases adquiridas por meio do Censo SUAS e utilizadas foram as bases gerais sobre os RMAs (Registro Mensal de Atendimentos) e as bases de RH sobre as equipes que casa unidade de serviço compunha em seu respectivo ano. Igualmente, o acesso temporal dessas informações se limitou aos anos de 2017 até 2021.

5 Análise descritiva

A medida inicial da análise foi entender as características principais da Pop. Rue presente no Censo de São Paulo. De tal forma, buscou-se compreender a maneira como essa população estava distribuída pelo município. Assim como demonstra a figura 1, a Pop. Rue se concentra de forma expressiva no centro da cidade. Ulteriormente, buscou-se traçar os vínculos entre as localizações das 6 unidades de Centros POP com a distribuição geográfica da população em situação de rua. A partir da figura 2, percebe-se que 5 das unidades tem presença na região central da cidade, com concentrações a partir de mais de 200 pessoas por km^2 .

Inicialmente, considerou-se utilizar as informações de RMAs dos CREAS (Centro de Referência Es-

Figura 2: Distribuição de Centros POP e População em situação de rua



pecializado de Assistência Social) na zona oeste de São Paulo. A região se trata da única zona com ausência de unidades de Centro POP, de forma que, nestes casos, as unidades do CREAS passam a prestar determinados serviços a Pop.Rua, visando prevenir agravamentos das situações de risco pessoal e social. Vale ressaltar que o CREAS não funcionará como substitutivo do trabalho social desenvolvido no Centro POP, mas poderá oferecer acompanhamento especializado, na localidade, às essas pessoas em situação de rua, possibilitando a construção do processo de saída das ruas.

De tal forma, incluir as unidades CREAS Pinheiros e CREAS Butatã foi uma abordagem para avaliar a capacidade de atendimento. Todavia, ao analisar os dados de atendimentos direcionados à população em situação de rua nestas unidades, viu-se que estas unidades apresentaram uma quantidade de atendimentos muito pequena para a Pop. Rua, assim como a tabela ?? apresenta. Sendo assim, as duas unidades não foram mais incluídas em análises futuras.

Tabela 2: Nome

Unidade do CREAS	Atendimentos totais	Atendimentos à Pop. Rua
Pinheiros	74	0
Butantã	655	10

hipótese de uma determinada relação entre profissionais, índice de recurso como oferta, e atendimentos foi avaliada a partir das taxas de variação dos mesmos. Examinando, portanto, a variação anual total de profissionais, apresentada na figura 3, e comparando-a com a variação anual de atendimentos nos Centro POP, figura 4, verifica-se que a quantidade de registros mensais apresenta semelhança no comportamento de variação com a quantidade de pessoas trabalhando na unidade de assistência social. De tal forma, parece existir uma alta correlação entre os dois.

Por fim, analisando os valores absolutos de crescimento da população em situação de rua com os

Figura 3: Variação anual do total de profissionais

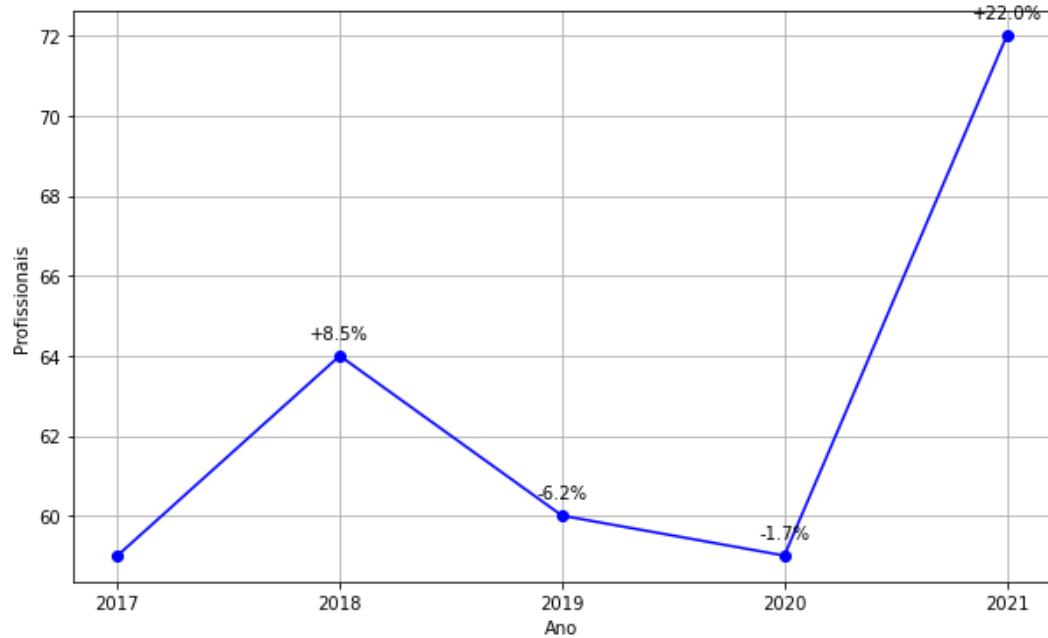


Figura 4: Variação anual de atendimentos

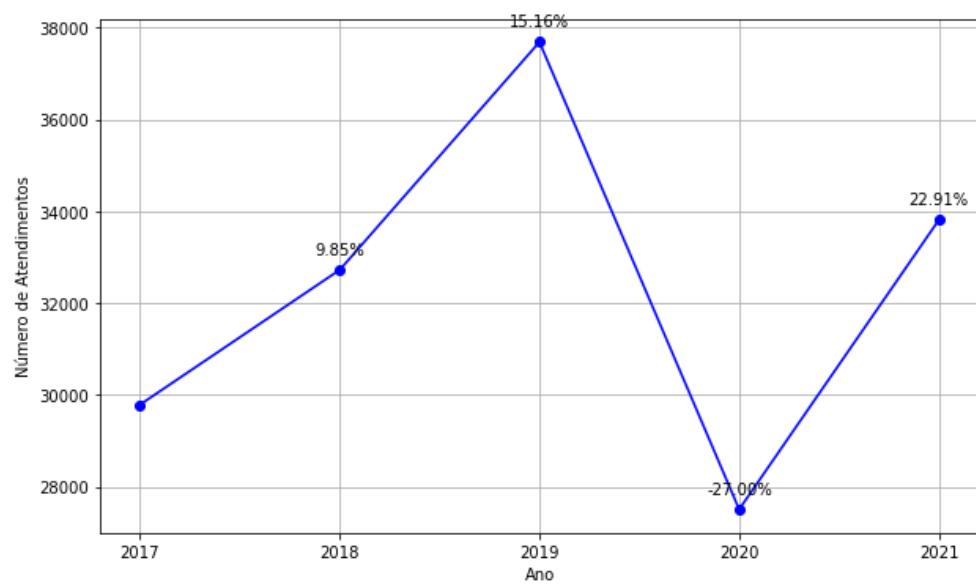
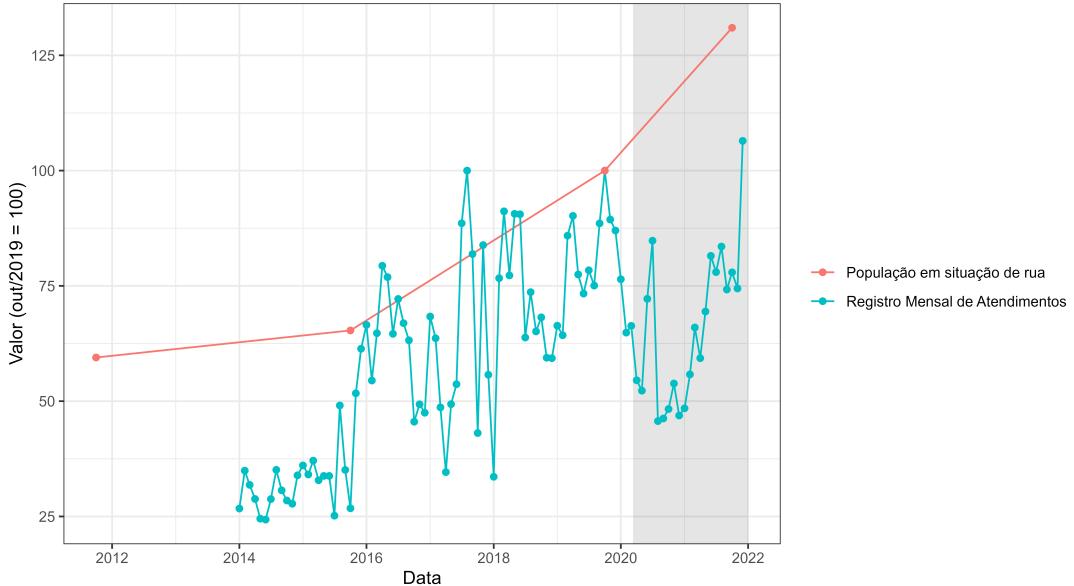


Figura 5: Variação cumulativa da população em situação de rua e RMA



registros mensais de atendimento, a figura 5 demonstra as relações entre as duas variáveis. Para a construção do gráfico, escolheu-se o início da pandemia de Covid-19 como a data designada como valor de referência, devido ao seu impacto na aceleração do crescimento da Pop. Rua. Logo, outubro de 2019 atua como referência (100), de modo que as variáveis População em situação de rua e Registro Mensal de Atendimento se localizam no mesmo ponto. Sendo este valor estabelecido, cada ponto no gráfico representa um comportamento proporcional à esta referência. Por exemplo, estando estabelecido que em 2019 houveram 100 atendimentos anuais, caso em 2021 a variável que representa Pop.Rua esteja acima de RMA, sabe-se que houveram menos do que 100 atendimentos, uma vez que a população aumentou mais do que o número de atendimentos.

6 Conclusões e limitações

A partir das análises realizadas neste estudo, é possível afirmar que a população em situação de rua questionado cresceu mais do que o número de atendimentos dos Centros Pop no município de São Paulo em 2021. Além disso, de acordo com os dados, cada pessoa em situação de rua foi atendida em média 1 vez por um Centro Pop ao longo de todo o ano de 2021.

Esse número é alarmante, porque significa que, em média, uma pessoa sem-abrigo acessa alimentação, banho e acolhida noturna uma vez a cada 365 dias em um Centro Pop. Partindo desse ponto, pode-se dizer que existem meios alternativos utilizados pela Pop. Rua para acessar seus direitos básicos, meios como o trabalho de organizações da sociedade civil e comunidades solidárias organizadas pela própria população sem-abrigo, que, apesar da relevância social, não oferecem estabilidade e capacidade de atender todos os que precisam, o que significa mais insegurança para uma população tão vulnerável.

Vale destacar ainda, que o apoio à Pop. Rua tem muitos obstáculos e dentre eles, destaca-se a lacuna nos dados estatísticos que fomentam as políticas públicas como os Centros Pop, dificuldades de crescimento e articulação de redes de apoio das políticas e obstáculos enfrentados pela população em

situação de rua no sector profissional que exigem uma resposta interdisciplinar.

Além disso, as bases de dados do Ministério do Desenvolvimento Social que dispõem de informações sobre a população em situação de rua como Registros Mensais de Atendimento não são suficientes para dimensionar o tamanho e perfil dessa população e também não esclarecem de que forma a política chega na ponta, por exemplo, se um morador recebeu acolhimento mais de uma vez ou se sequer chegou aos centros.

Dessa forma, apesar das limitações, este estudo buscou compreender o funcionamento de uma política social voltada à Pop. Rua a fim de analisar a maneira pela qual essa população acessa seus direitos e como o Estado - Mais especificamente, na figura do município de São Paulo - Enquanto agente implementador de políticas públicas atua para garantir acesso à direitos básicos aos cidadãos mais vulneráveis de seu território.

Referências

BARBOSA, JC. Implementação das políticas públicas voltadas para a população em situação de rua: desafios e aprendizados. **Programa de Pós-Graduação em Políticas Públicas e Desenvolvimento, Instituto de Pesquisa Econômica Aplicada.** Brasília, 2018.

MEDEIROS, Fernanda Cavalcanti de et al. Entre a benesse e o direito: as políticas de atendimento à população em situação de rua na América Latina. **Psicologia em Estudo**, SciELO Brasil, v. 25, e45025, 2020.

NATALINO, Marco Antonio Carvalho. **Estimativa da população em situação de rua no Brasil.** [S.l.], 2016.

RAUPP, Luciane; ADORNO, Rubens de Camargo Ferreira. Territórios psicotrópicos na região central da cidade de Porto Alegre, RS, Brasil. **Saúde e Sociedade**, SciELO Public Health, v. 24, p. 803–815, 2015.

SERAFINO, Irene; LUZ, Lila Cristina Xavier. Políticas para a população adulta em situação de rua: questões para debate. **Revista Katálysis**, SciELO Brasil, v. 18, p. 74–85, 2015.

SILVA, Tatiana Dias; NATALINO, Marco Antônio Carvalho; PINHEIRO, Marina Brito. População em situação de rua em tempos de pandemia: um levantamento de medidas municipais emergenciais. Instituto de Pesquisa Econômica Aplicada (Ipea), 2020.

TRABALHO, Gerência de Gestão do; PERMANENTE, Educação. Secretaria de Desenvolvimento Social, Criança e Juventude.

7 Anexos

Figura 6: Tempo em situação de rua por etnia

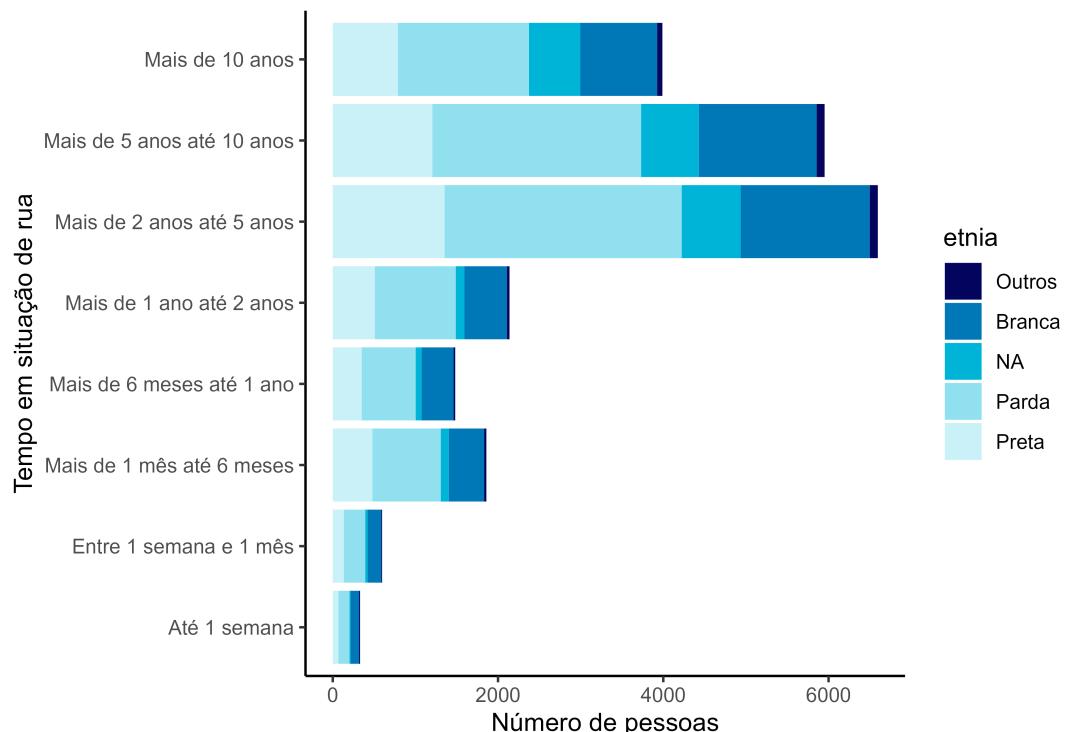


Figura 7: Evolução do RMA Anual por Centro POP

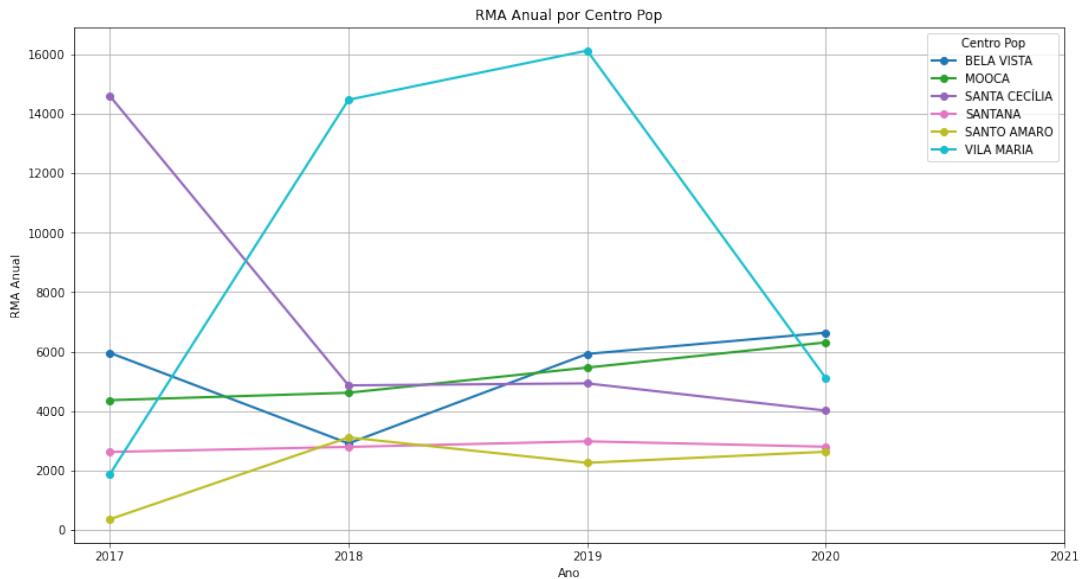
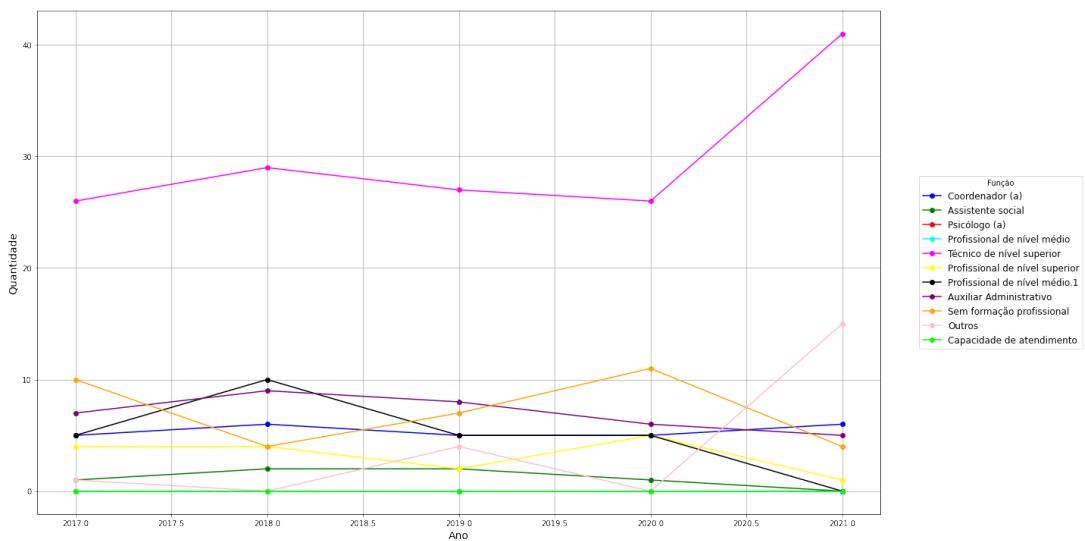


Figura 8: Evolução do número de profissionais por função ao longo dos anos



Qualidade Educacional e Valorização Imobiliária: Um Estudo sobre o Efeito das Escolas em São Paulo

Pesquisadores: Gabriel de Araújo Alves, Phillippe Libreti Dias e Tiago Milani

Resumo

Este relatório investiga o impacto da localização de escolas sobre os preços de imóveis na cidade de São Paulo. A pesquisa examina a relação entre a proximidade de escolas e o valor médio dos imóveis, buscando identificar se a presença de escolas de alta qualidade influencia de forma significativa o mercado imobiliário na cidade. O estudo utiliza dados de preços de imóveis e informações sobre a localidade de escolas para analisar esta relação, com o objetivo de fornecer insights valiosos para compradores, vendedores e formuladores de políticas.

Palavras-chave: Microeconomia, Modelagem preditiva, Previsão, Impacto, Localização, Valorização imobiliária.

1 Introdução

A cidade de São Paulo é uma metrópole em constante transformação, onde a busca por um lar ideal envolve inúmeras considerações. Uma dessas considerações, particularmente importante para muitos cidadãos, é a proximidade de escolas, dado que uma grande parte dos trabalhadores se deslocam grandes distâncias até o trabalho e não tem tanto tempo para seguir com a rotina escolar de seus filhos. Dessa forma, esse artigo busca analisar o efeito das escolas sobre o preço do metro quadrado, especificamente se a existência de escolas próximas à residências eleva seus preços.

O assunto de como a educação afeta o mercado imobiliário já é amplamente estudado, com diversos artigos abordando tanto a situação nacional como de outros países. Um deles, Valuing primary schools in urban China (CHAN et al., 2020), busca analisar, no contexto chinês, o valor dos imóveis relacionados com a qualidade das escolas primárias. Apesar das diferenças com o contexto brasileiro, o artigo traz diversos *insights* sobre como a medição do valor dos imóveis dependem de diversos fatores, tanto das características internas de cada um quanto de seus arredores.

Uma das principais metodologias, aplicada inclusive no artigo citado, é a de preços hedônicos. A análise hedônica considera que o preço de um imóvel é determinado não apenas pela sua localização, tamanho e idade, mas também por uma série de características intrínsecas, como o número de quartos e banheiros, a presença de áreas de lazer, a vista, a orientação solar e muitas outras variáveis. Cada uma dessas características tem um impacto no preço final por metro quadrado do imóvel, podendo inclusive ser capaz de quantificar o quanto cada uma dessas peculiaridades afeta o valor encontrado.

Dessa forma, o conhecimento e o estudo dos caminhos que definem os preços dos imóveis, em especial relacionado à educação, uma das necessidades essenciais de todo cidadão, será capaz de guiar tanto os

estudos de qualidade de ensino e deslocamento urbano, por parte da formulação e aplicação de políticas públicas, além de assistir a tomada de decisão de compradores e vendedores que desejam auferir os preços de seus imóveis com base em aspectos escolares.

2 Revisão da Literatura

Antes de explorar nossos próprios dados, é relevante destacar pesquisas semelhantes em outros contextos. Por exemplo, o estudo Valuing primary schools in urban China (CHAN et al., 2020) investigou a valorização do metro quadrado de áreas residenciais segundo a qualidade das escolas primárias em áreas urbanas chinesas, com a particularidade específica do país de que a escolha da família sobre a alocação de filhos à escola não é arbitrária, sendo limitada pela região onde vivem, assim tornando a escolaridade uma alocação ao invés de uma escolha.

Além desse, outro artigo que explora a precificação de imóveis com base na existência de escolas ao seu redor é When do better schools raise housing prices? Evidence from Paris public and private schools (FACK; GRENET, 2010), que analisa os valores de compra e venda de imóveis que estejam a algumas distâncias de referência de escolas na cidade de Paris, indicando que eles serão os mais afetados pela existência do serviço, comparados com imóveis mais distantes.

Por fim, o último artigo utilizado para embasar nosso trabalho é o Hedonic prices and implicit markets: product differentiation in pure competition (ROSEN, 1974), que explora, de forma totalmente teórica, a formação de uma teoria de preços hedônicos, que seria formulada a partir tanto de preços implícitos dos compradores e vendedores, quanto da característica de equilíbrio do próprio mercado em que os bens estariam inseridos.

3 Motivação

A motivação para o problema descrito neste relatório parte da curiosidade dos pesquisadores em entender um tema altamente relevante para a sociedade paulistana. A cidade de São Paulo é uma metrópole em constante transformação, onde a busca por um lar ideal envolve inúmeras considerações. Uma dessas considerações, particularmente importante para muitos cidadãos, é a proximidade de escolas. Reforçando que uma grande parte dos trabalhadores deslocam grandes distâncias até o trabalho e não tem tanto tempo para seguir com a rotina escolar de seus filhos.

Nesse sentido, é importante destacar que questões relacionadas ao mercado imobiliário já têm sido amplamente debatidas tanto no campo da microeconometria quanto em diversos problemas de modelagem preditiva. No entanto, a combinação dessas duas abordagens de estudo, no contexto da análise do impacto da localização de escolas sobre os preços de imóveis, se revela particularmente relevante e motivadora para os pesquisadores.

A convergência dessas ferramentas analíticas permite abordar um problema de grande interesse prático de maneira abrangente e fundamentada em dados. Além disso, oferece a oportunidade de aprofundar nossa compreensão das complexas interações entre o mercado imobiliário e o setor educacional.

4 Modelo Microeconômico

O modelo definido, baseado nas literaturas previamente vistas de Chan et al. (2020), Fack e Grenet (2010) e Rosen (1974) , define a utilidade do indivíduo como sendo

$$U = U\{z_1, z_2, \dots, z_n, x\} \quad (2)$$

Com z_n definindo a percepção de valor do imóvel, que possui um conjunto de características y , e x representando uma variável agregada de todos os outros bens de consumo desejados. Além disso, há uma restrição orçamentária para cada um dos indivíduos, definida pela equação

$$I = x + p(z) \quad (3)$$

Sendo $p(z)$ a variável que define o preço do imóvel.

A percepção de valor do imóvel, z_n , é definida por

$$z_n = f(\phi_n, \epsilon_n) \quad (4)$$

Com ϕ_n representando a distância até a escola mais próxima ao imóvel, e ϵ_n sendo um vetor de outras características relevantes que ele possui, como, por exemplo, número de banheiros e a existência de ar condicionado no local. No modelo, é assumido que ϕ_n é negativamente relacionado à percepção de valor dada pelo indivíduo. Quanto mais afastado o imóvel está da escola, para a maioria das pessoas, menor é a utilidade percebida pelos moradores. Consequentemente, a propriedade torna-se menos atrativa e, por conseguinte, é atribuído um valor menor a ela.

Dessa forma, ao compararmos dois imóveis, um mais distante de uma escola, e outro menos, observaríamos que, considerando todas as outras variáveis constantes, para aquele mais longe, há uma percepção de valor menor, o que, portanto, impactaria em um preço comparativamente menor àquele com uma distância menor à escola, que, dessa forma, teria um $p(z)$ maior, tudo o mais constante. Assim, podemos concluir que, segundo o modelo, a distância de imóveis até a escola, sejam elas tanto públicas quanto privadas, leva a uma valorização do valor do metro quadrado dos imóveis mais próximos em São Paulo.

5 Base de Dados

Este projeto combina dois conjuntos de dados essenciais para sua execução: O primeiro dataset, obtido através do site de dados da Prefeitura de São Paulo (2019) (clique [aqui](#) para acessar os dados de cada ano), refere-se ao cadastro de escolas municipais, conveniadas e privadas na metrópole paulista. Este conjunto de dados inclui informações como descrição regional da escola, código da escola, diretoria e subprefeitura onde está localizada, detalhes sobre a localização (CEP, endereço, latitude e longitude) e outras variáveis pertinentes. Dessa forma, variáveis relevantes para o modelo incluem o nome da escola, o tipo de escola (pública ou privada) e, como mencionado anteriormente, as coordenadas de localização (latitude e longitude) do ambiente escolar.

O segundo conjunto de dados, disponibilizado pelo **Kaggle**, é proveniente de um business case da empresa de imóveis Loft e oferece uma perspectiva detalhada sobre características específicas de propriedades, como metragem quadrada e número de quartos, além do tipo de contrato (aluguel ou venda) e do preço médio de lançamento do imóvel correspondente para cada tipo mencionado anteriormente.

Combinados, esses conjuntos de dados possibilitarão uma análise integrada do mercado imobiliário e escolar, explorando a relação entre fatores urbanos e características específicas de imóveis.

5.1 Resumo Estatístico Loft dataset

Tabela

Variável	Descrição	Média	Desvio padrão	Mínimo	Máximo
Rooms	Número de quartos na propriedade	2.32	0.715	1	6
Toilets	Quantidade de banheiros na propriedade	2.038	0.918	1	7
Suites	Número de suítes na propriedade	0.9315	0.773	0	6
Parking	Quantidade de vagas de estacionamento	1.326	0.752	0	7
Elevator	Quantidade de elevadores	0.415	0.493	0	1
Furnished	Se é mobiliado ou não	0.117	0.322	0	1
Swimming Pool	Se há piscina ou não no imóvel	0.54	0.498	0	1
New	Indicação de propriedade nova	0.03	0.177	0	1
District	Nome do bairro	-	-	-	-
Price per sqm	Preço por metro quadrado	6891.64	3182.7	755.55	46212.17
Coordenadas	Latitude e Longitude	-	-	-	-

Tabela 3: Descrição das Variáveis da Base de Dados Da Loft filtrada

A tabela acima oferece uma visão geral das variáveis relevantes presentes na Base de Dados Loft de São Paulo. É importante destacar que os valores apresentados na tabela são aproximados e referem-se especificamente a imóveis caracterizados como tendo preço de venda. Embora a base contenha imóveis com preço de aluguel, esses foram desconsiderados para fins de análise. Essa base de dados contém informações sobre transações imobiliárias, incluindo detalhes sobre as propriedades, valores e características relevantes. As variáveis listadas nesta tabela são fundamentais para entender o conteúdo e a estrutura dos dados, além disso houve filtragem e modificação de algumas variáveis, por exemplo, a variável *Price per sqm* é fruto da divisão entre as variáveis *Price* e *Size*, ou seja, houve a retirada de algumas variáveis e modificação de outras para constituir as variáveis desta tabela.

Nesse sentido, excetuando-se as variáveis District e Coordenadas, todas as outras colunas possuem

medidas-resumo que foram apresentadas na tabela anterior afim de depreender e fornecer *insights* iniciais de forma estatística.

Além disso, pode-se ver uma análise descritiva rápida com a imagem a seguir:



Figura 1: Preço por metro quadrado amostra cidade São Paulo

Na imagem anterior, a quantidade de imóveis foi reduzida, representando uma amostra de 100 unidades, composta por 50 dos imóveis mais caros e 50 dos mais baratos da base total da Loft. Isso se deve à inviabilidade computacional de gerar imagens para os mais de 6000 imóveis disponíveis na base completa.

6 Modelo

Este estudo visa analisar o impacto da proximidade de escolas públicas e privadas nos preços dos imóveis na cidade de São Paulo. Utilizando um modelo de regressão linear, avaliaremos como a distância até a escola mais próxima e o tipo de escola (pública ou privada) afetam o valor de mercado dos imóveis.

6.1 Modelo Empírico

A modelagem empírica leva em consideração características intrínsecas ao imóvel (número de quartos, suítes, banheiros e características da área comum), o bairro onde o imóvel está localizado, a distância à escola mais próxima e a área do imóvel e, com isso, uma regressão linear é proposta como forma de encontrar o valor do metro quadrado para determinado imóvel, deste modo também isolando as características e similaridades entre imóveis que pudessem justificar a variação do seu preço, e isolar o efeito das escolas e si.

Régressao Linear :

$$\frac{Preco}{m_i^2} = \beta_0 + \beta_1 \cdot distancia_{escola_i} + \beta_2 \cdot D_{pub} + \beta_3 \cdot Area_i + \beta_4 \cdot Area_i^2 + \sum_{j=0}^k \gamma_j \cdot D_{bairro,j} + \sum_{p=0}^N \delta_p \cdot W_p + \epsilon_i \quad (5)$$

6.2 Descrição

A seguir uma secção com a explicação das variáveis:

- $distanciaescola_i$:Distância do Imóvel à Escola mais próxima
- D_{pub} :Dummy se a escola é pública ou não (valor 1 se pertence e 0 se não pertence)
- $Area_i$:Área do Imóvel
- $D_{bairro,j}$:Dummy de Bairros para identificar se determinado pertence ao bairro (valor 1 se pertence e 0 se não pertence)
- W_p :Vetor de Características
 - Número de Quartos
 - Banheiros
 - Elevador
 - Mobiliado ou Não
 - Piscina
 - Novo ou Não

6.3 Metodologia

A distância até as escolas será calculada utilizando a fórmula de euler, que considera a distância entre dois pontos para estimar distâncias entre pontos definidos por coordenadas de latitude e longitude. O modelo será estimado usando o método de Mínimos Quadrados Ordinários (MQO), e a significância estatística dos coeficientes será testada para determinar quais variáveis têm um efeito significativo no preço dos imóveis. Ademais, vale ressaltar que o vetor de características e a Dummy de Bairros serão consideradas variáveis de controle afim de evitar possíveis viéses relacionados ao modelo proposto de regressão linear.

7 Resultados

Chegando aos resultados obtidos pela regressão estimada, observa-se, como mostrado abaixo nos principais coeficientes de análise, que se confirma os sinais demonstrados pelo modelo teórico (efeito marginal positivo para escola privada e negativo para distância à escola mais próxima).

Portanto se observa que quanto maior a distância da escola mais próxima, menor tende a ser o valor do metro quadrado do imóvel, e do mesmo modo, sendo a escola mais próxima privada, se adiciona um efeito em nível de 147.17 reais no metro quadrado. Demonstrando assim a propensão do comprador. Ademais, é notório observar que o efeito marginal de maior metragem é negativo negativo para imóveis abaixo de 118 metros quadrados, e positivo para maiores. Como segue a derivada a seguir:

$$\frac{\partial \frac{\text{Preço}}{m^2}}{\partial \text{Área}} = -5.73 + 0.0476 \times \text{Área}$$

	Coeficientes	P-Valor	
Constante	+5308.98 [167.43]	0.00	***
Distância escola mais próxima	-0.16 [0.568]	0.777	
Dummy Escola Privada	+285.95 [111.044]	0.010	**
Área	-5.73 [1.925]	0.003	**
Área²	+0.0238 [0.005]	0.000	***

Tabela 4: Resultados da regressão

Em busca de validar os resultados, realizando o teste de homocedasticidade dos resíduos se observou Homocedasticidade com 99% de confiança. Além disso, ao testar o mesmo modelo, porém para Preço de Aluguel por metro quadrado, se observou pouca significância nos parâmetros de interesse como demonstrado a seguir.

	Coeficientes	
Constante	+37.05	***
Distância escola mais próxima	0.0056	**
Dummy Escola Privada	-1.3019	*
Área	-0.1138	***
Área²	+0.0002	***

Tabela 5: Resultados da regressão para Aluguel por metro quadrado

8 Limitações

A omissão de uma variável relevante é um ponto crítico a ser considerado na análise de transações imobiliárias. A presença ou ausência de Shoppings, Centros Culturais, Restaurantes e Comércio nas proximidades pode exercer uma influência significativa na utilidade percebida pelos agentes envolvidos. A falta de inclusão dessas variáveis pode, assim, restringir a compreensão abrangente dos fatores que impactam o cenário imobiliário.

Além disso, a heterogeneidade das preferências entre diferentes tipos de compradores e em relação a diversas características de bairros, ruas ou regiões é um aspecto crucial a ser contemplado. Considerar essas preferências distintas é fundamental para uma análise mais precisa, alinhada com a diversidade de agentes presentes no mercado imobiliário.

A detecção e tratamento de outliers são igualmente cruciais. A presença de imóveis com preços anômalos e as possíveis interferências de legislações urbanas devem ser cuidadosamente examinadas. Outliers têm o potencial de distorcer a análise estatística e levar a conclusões equivocadas, destacando a importância de uma investigação meticulosa para uma compreensão mais precisa do mercado.

Por fim, ao realizar uma análise baseada em anúncios, é imperativo reconhecer as limitações inerentes a essa abordagem. A discrepância entre os preços anunciados e os valores efetivamente transacionados pode comprometer a fidedignidade da análise. A consideração dessas variações é essencial para evitar conclusões incorretas sobre a dinâmica do mercado imobiliário.

9 Conclusão

Com base nos resultados obtidos, observa-se as seguintes tendências:

- A longitude às escolas, em geral, sem considerar distinção entre pública e privada, está associada a uma redução de 0,16 real nos preços dos imóveis por metro quadrado para cada quilometro de distância adicional.
- Contrariamente, a presença de escolas privadas está associada a um aumento significativo de 285.95 reais nos preços por metro quadrado dos imóveis em relação às escolas públicas. Este aumento sugere que a proximidade a escolas privadas é um fator gerador de valor no mercado imobiliário.

Essas conclusões ressaltam a importância de considerar o tipo de instituição educacional ao avaliar os preços de imóveis em determinadas áreas. A diferenciação entre escolas públicas e privadas revela nuances significativas nas preferências dos compradores e na dinâmica do mercado imobiliário.

No entanto, é crucial observar que, na amostra analisada, não se observou uma significância expressiva em relação à mudança no valor do metro quadrado para as escolas públicas. Em contraste, para as escolas privadas, a associação positiva entre sua presença e a valorização imobiliária é notável, como mencionado anteriormente.

Além disso, é relevante apontar que o desbalanceamento no número de escolas, favorecendo as privadas no conjunto de dados, pode introduzir um viés no modelo, levando a uma sugestão potencialmente inflada de valorização imobiliária, especialmente ao considerar escolas privadas em detrimento das públicas. Essa ponderação é crucial para uma interpretação precisa dos resultados e destaca a importância de abordar possíveis desequilíbrios na representação das escolas na amostra.

Referências

- CHAN, Jimmy et al. Valuing primary schools in urban China. **Journal of Urban Economics**, Elsevier, v. 115, p. 103183, 2020.
- FACK, Gabrielle; GRENET, Julien. When do better schools raise housing prices? Evidence from Paris public and private schools. **Journal of public Economics**, Elsevier, v. 94, n. 1-2, p. 59–77, 2010.
- PAULO, Prefeitura São. Dataset prefeitura. Elsevier, v. 94, n. 1-2, p. 59–77, 2019.
- ROSEN, Sherwin. Hedonic prices and implicit markets: product differentiation in pure competition. **Journal of political economy**, The University of Chicago Press, v. 82, n. 1, p. 34–55, 1974.

Análise Comportamental em Plataformas Digitais: Aplicando Cadeias de Markov e Simulações de Monte Carlo

Pesquisadores: Antônio Vicente Fernandes de Andrade, Rafael Albuquerque, João Pazotti

Resumo

Este artigo apresenta um trabalho de consultoria realizado pela organização estudantil Insper Data para a empresa FRST - Falconi. O objetivo da análise foi entender os padrões de comportamento dos usuários da plataforma online do cliente. As ferramentas utilizadas para esse fim foram a teoria das Cadeias de Markov e, paralelamente, simulações de Monte Carlo. Como resultado da nossa análise conseguimos identificar os estados (áreas do site) mais relevantes para entender a dinâmica de engajamento do usuário, isto é, as páginas que mais reforçam ou diminuem o engajamento. Esse trabalho é muito relevante em paralelo ao time de produtos da FRST, que focarão no redesenho dos estados indicados.

Palavras-chave: Cadeia de Markov, Simulação de Monte Carlo

1 Introdução

Dado a nova dinâmica que a era digital trouxe, cada clique e interação geram dados, o que transforma cada janela de um site em uma rica fonte de informações. Esse ambiente gera, portanto, grandes oportunidade para entender o comportamento e as decisões dos usuários. Neste contexto, nosso estudo propõe-se a explorar esses padrões de comportamento na plataforma da FRST, um ambiente digital com um grande tráfego de usuários, através da aplicação da teoria das Cadeias de Markov, que nos dão um amparo robusto para modelar processos estocásticos e, sobretudo, para entender o próximo estado que cada um dos clientes da FRST escolherão baseado no estado em que estão agora, e Simulações de Monte Carlo, que permitem a simulação de interações de usuários reais na plataforma, o que nos possibilitará identificar tendências que não são facilmente perceptíveis. O objetivo deste estudo é, então, empregar essas metodologias para identificar os estados que mais impactam os padrões de comportamento do usuário e sugerir para o time de produtos do parceiro um "ranking" de estados que mais necessitam de um redesenho.

1.1 Sobre a FRST

A FRST, uma filial do grupo Falconi, é uma plataforma focada no aprendizado através do fomento a uma comunidade de resolução de problemas. A empresa, com mais de 1000 clientes e mais de 30 mil usuários tem como modelo de negócio promover um espaço no qual os usuários consigam tirar suas

dúvidas através da criação de "Desafios" e da recomendação de trilhas de conteúdo educativo com base no perfil do sujeito. Cada Desafio segue o método científico e, uma vez que é publicado, a ideia é que outros usuários consigam interagir e compartilhar como resolveriam ou resolveram problemas semelhantes ao divulgado. Já as trilhas de conteúdo são recomendadas pela plataforma após o onboarding na plataforma, onde o usuário tem que fazer uma redação de "auto conhecimento".

2 Modelagem Teórica

2.1 Propriedade Markoviana

Processos Markovianos são todos aqueles que, além de estocásticos, respeitam a chamada 'Condição Markoviana', segundo a qual o futuro da série analisada independe de qualquer evento anterior ao estado atual em que ela se encontra.

$$P(X_{n+1} = x | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = P(X_{n+1} = x | X_n = x_n) \quad (6)$$

A propriedade Markoviana pode ser representada pela equação 6, (ROSS, 1997), na qual X_n representa um estado no passo n e $x, x_n, x_{n-1}, \dots, x_0$ representam as probabilidades de transição desse estado. Não é difícil enxergar, portanto, que a probabilidade de ir ao estado X_{n+1} não é alterada pelos estados anteriores (X_{n-1}) uma vez que depende apenas do estado atual (X_n).

2.2 Cadeias de Markov

Nesse sentido, quando uma série estocástica, além de respeitar a condição markoviana , é discreta, podemos classificar esse processo Markoviano como uma Cadeia de Markov, que pode ser representada, em uma versão simplificada, pela figura 1. Toda cadeia é composta por 3 elementos principais: Estados (S_i), ações (as setas de um estado a outro) e as probabilidades de transição (a_{ij}), que são as probabilidades de cada ação acontecer.

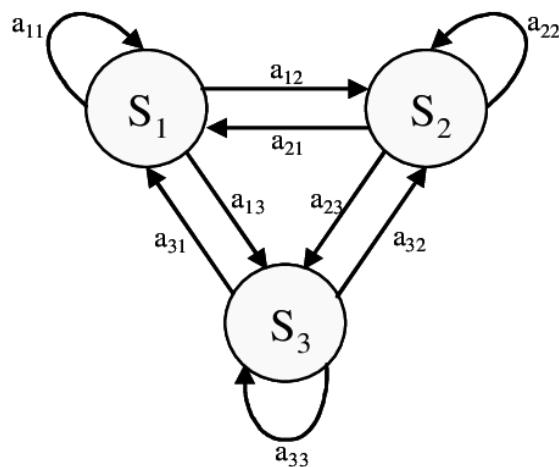


Figura 1: Cadeia de Markov com 3 estados

As cadeias de Markov já são amplamente utilizadas em diversas áreas, como econometria financeira (VERHOFEN, 2005), previsões temporais (KHIATANI; GHOSE, 2017) e mecanismos de pesquisa e recomendação de conteúdo (RAI; LAL, 2016). Isso ocorre porque a, apesar de ser conceitualmente simples, a metodologia oferece uma grande adaptabilidade e robustez matemática. Em termos do nosso problema, podemos pensar em cada estado do site como a página de um site, com S_1 sendo o login, S_2 o menu e S_3 a área de desafios da FRST, por exemplo. Ainda, as ações possíveis são as transições de uma página do site para outra.

2.2.1 Probabilidades de Transição

Um benefício central das cadeias são as probabilidades de transição em si. A perspectiva que as cadeias markovianas trazem, de que o passo seguinte depende apenas do presente e das probabilidades de transição atreladas a ele, nos ajuda a observar de forma intuitiva um problema de grande complexidade, que é a "previsão" das escolhas de terceiros.

Uma forma de complementar a representação é montar uma matriz com as probabilidades de transitar entre os estados, que recebe o nome de matriz de transição ???. A partir dela, podemos aplicar ferramentas de álgebra linear e da própria teoria Markoviana para extrair informações valiosas do problema.

Tabela 6: Matriz de Transição

$$P = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

2.2.2 Probabilidades Estacionárias

Um ponto que apesar de muito simples agrupa muito à análise, são as probabilidades estacionárias. Essa probabilidades são um número fixo que representa a probabilidade de um usuário estar em um respectivo estado da plataforma após n interações. Para chegar a essa matriz de probabilidades, elevamos a matriz de transição 6 a enesima potência 8 e, em seguida, a pré-multiplicamos por uma matriz que representa o estado inicial do sujeito, ou (ROSS, 1997) "Matriz de Decisão". Desse modo, chega-se ao vetor de probabilidades estacionárias π 9. (ROSS, 1997)

$$S_0 = \begin{pmatrix} s_{01} \\ s_{02} \\ s_{03} \end{pmatrix} \quad (7)$$

$$\lim_{n \rightarrow \infty} E = \lim_{n \rightarrow \infty} P^n \times \begin{pmatrix} s_{01} \\ s_{02} \\ s_{03} \end{pmatrix} = \pi \quad (8)$$

$$\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} \quad (9)$$

2.3 Modelo MCMC (Markov Chain - Monte Carlo)

Por fim, as cadeias possibilitam a criação de uma "Zona de Testes" que pode ser modificada da forma que o usuário preferir seja criando estados ou simulando a iteração de milhares de pessoas na plataforma com simulações de Monte Carlo.

As simulações de Monte Carlo são uma forma de resolver problemas utilizando uma "população hipotética" (JAMES, 1980). A estratégia será utilizada aqui para simular os reais estados de saída dos usuários para buscar identificar as páginas que prejudiquem a experiência do cliente.

3 Dados e Modelagem Empírica

3.1 Adaptação de Estados

As probabilidades de transição, representadas pelas linhas, indo de um estado a outro, foram calculadas com base em uma amostra com dados de navegação de clientes ao longo de 3 meses, somando aproximadamente 5000 sessões. Além disso, na base, temos o ID do usuário, o caminho que ele fez na plataforma e a data e o horário no qual ele entrou naquele estado.

A primeira etapa para começar a trabalhar com a base foi traduzir cada estado em um número de 0 a 22, ficando da seguinte forma:

- 0 - Pesquisa Semântica.
- 1 - Opções Onboarding.
- 2 - (Modal) Cadastro de Desafio - Clique.
- 3 - (Modal) Cadastro de Desafio - Atualização
- 4 - Step 2 para o 3 do Onboarding
- 5 - (Modal) Cadastro de desafio - Mudança de Step
- 6 - Hall de Desafios
- 7 - (Modal) Cadastro de desafio - Listagem de Hipóteses
- 8 - Clique no botão para um desafio específico
- 9 - (Modal) Cadastro de desafio - Upload de Arquivo
- 10 - (Modal) Cadastro de desafio - Ir de um formulário de desafio para outro.
- 11 - Hall de Desafios do usuário.
- 12 - Estado de Gerenciamento (Admin)
- 13 - Step 1 para o 2 do Onboarding
- 14 - (Modal) Cadastro de desafio - Cria um Indicador
- 15 - (Modal) Cadastro de desafio - Criar Desafio

16 - (Modal) Cadastro de desafio - Função para criar um Payload e cria um desafio

17 - Step 1 do Onboarding

18 - Desabilita o Botão de Seguir

19 - Clica nas notificações

20 - Clique no Menu Inicial

21 - (Modal) Cadastro de desafio - Função que inicia o desafio

22 - Login na Plataforma

Além disso, representando nossa cadeia em um grafo, como foi feito anteriormente, na Figura 1, temos a visualização da Figura 2. Claramente a complexidade do problema é muito maior, dada a quantidade de relações entre cada estado.

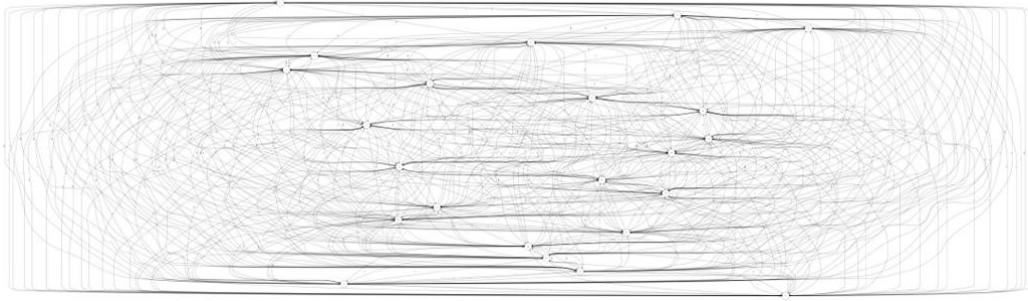


Figura 2: Cadeia de Markov FRST

3.2 Alterações na Base

Foram feitas então duas modificações essenciais para o desenvolvimento da análise. A primeira foi adicionar, ao final de cada caminho de usuário, um novo estado, o 23º. Que representou o estado de saída do site. Essa modificação nos permite identificar quais são os estados que mais prejudicam o engajamento do usuário, dada a probabilidade de transição de cada estado para o fim do trajeto. A outra alteração foi garantir que cada caminho começará pelo estado 22, isto é, pela página de login do site. Essa alteração nos garantiu uma perspectiva mais clara das probabilidades de transição dado que todos "escolheram" começar pelo início do site.

Id	Caminho Original	Caminho Novo
0	[22][11][6][2]	[22][11][6][2][23]
1	[7][2]	[22][7][2][23]
2	[2][11][7][20]	[22][2][11][7][20][23]

Tabela 7: Id e Trajetórias - Exemplo

4 Resultados

A abordagem aqui será muito simples. A ideia é observar as principais probabilidades estacionárias do site e compara-las com o fluxo efetivo de ida até o estado 23, encontrado através da simulação de Monte Carlo. Desse modo, se uma página do site tem uma alta probabilidade de presença e baixa probabilidade de transição para a saída, temos que essa página mantém o engajamento do usuário e abre espaço para mudanças e anúncios mais efetivos. Por outro lado, uma página que tem grande probabilidade estacionária e também um grande fluxo de saída, podemos indicar para a FRST que esse estado contribui negativamente para a experiência do usuário e necessita de um redesenho ou, eventualmente, menos possibilidades de transição para ele ao longo do percurso na plataforma.

Nessa perspectiva, as principais probabilidades estacionárias das páginas do site foram os 5 estados presentes na tabela 8, somando aproximadamente 90% de probabilidade das probabilidades de presença.

Estado	Probabilidade Estacionária (%)
7	30,1
2	26,4
20	19,7
6	7,4
11	4

Tabela 8: Principais probabilidades estacionárias

Como podemos observar, apesar das grandes probabilidades estacionárias dos estados na tabela 8, apenas o estado 20 apresenta também uma grande representatividade na probabilidade de transição para o estado 23 na simulação. O que pode indicar para a FRST que o Menu Inicial pode prejudicar o engajamento.

5 Próximos Passos

Apesar de esse trabalho ter focado apenas na transição de um estado a outro, idealmente ele servirá de base para uma futura análise dos eventos criados por cada transição. Isto é, cada clique e interação dentro de cada página. De modo que buscaremos entender com maior profundidade o que cada usuário busca fazer na plataforma.

Referências

JAMES, Frederick. Monte Carlo theory and practice. **Reports on progress in Physics**, IOP Publishing, v. 43, n. 9, p. 1145, 1980.

KHIATANI, Diksha; GHOSE, Udayan. Weather forecasting using hidden Markov model, p. 220–225, 2017.

RAI, Prerna; LAL, Arvind. Google pagerank algorithm: Markov chain model and hidden Markov model. **International Journal of Computer Applications**, Foundation of Computer Science, v. 138, n. 9, p. 9–13, 2016.

Estado	Percentual de Representatividade (%)
20	29.01
19	23.70
11	19.95
9	13.20
12	5.17
6	2.26
0	1.84
15	1.44
2	1.38
18	1.12
10	0.37
21	0.22
7	0.22
1	0.08
5	0.03
4	0.01

Tabela 9: % de ida ao Estado 23 - Monte Carlo

ROSS, Sheldon M. **Introduction to Probability Models**. Sixth. San Diego, CA, USA: Academic Press, 1997.

VERHOFEN, Michael. Markov chain monte carlo methods in financial econometrics. **Financial Markets and Portfolio Management**, Springer, v. 19, n. 4, p. 397–405, 2005.

Engenharia de dados para plataforma de consultoria educacional

Pesquisadores: Gabriel Araújo Alves, Gustavo Cavaletti e Felipe Bakowski Nantes de Souza

Orientador: José Renato Garcia Braga

Resumo

Este trabalho aborda a Engenharia de Dados no contexto do processo de qualidade de dados. Além disso, representa uma colaboração com a empresa FRST Falconi, que proporcionou suporte integral em termos de orientação e disponibilização de material (dataset). O objetivo dessa parceria é explorar soluções que aprimorem a eficiência do processo de qualidade de dados na empresa, proporcionando aos tomadores de decisão um ambiente de trabalho facilitado e dados de alta qualidade.

Palavras-chave: Engenharia de dados, Qualidade de dados, Big Data

1 Introdução

Na era digital, onde os dados se tornaram ativos essenciais para o sucesso corporativo, a Engenharia de Dados desponta como uma disciplina-chave para moldar o futuro das organizações. Este trabalho se dedica à aplicação da Engenharia de Dados para aprimorar o processo de qualidade de dados, em parceria com a empresa FRST Falconi.

O objetivo é compreender e catalisar melhorias que fortaleçam a eficiência do processo de qualidade de dados, proporcionando à FRST Falconi uma base sólida de dados para tomada de decisões.

Além disso, é notório inferir que o trabalho apresentado será de grande valia e servirá como alicerce para o processo de ciência de dados realizado por outro projeto, também em parceria com a FRST Falconi.

2 Motivação

A formalização do ambiente digital pelo mercado corporativo dá luz ao problema de organização de informações úteis para as empresas, em que o custo de acesso à dados de qualidade pode ser reduzido ao fazer uso de ferramentas de engenharia de dados. O desperdício de tempo dos cientistas de dados, por ano, com desafios de limpeza e organização de dados é de cerca de 60%. Além disso, anualmente US\$ 3.1 Trilhões são desperdiçados anualmente com tarefas oportunas à prática de engenharia de dados por causa de dados recebidos com baixa qualidade.

3 Problema

Compreender o funcionamento e falhas do processo de ingestão de dados da FRST é o objetivo principal desse trabalho, em vista de possibilitar a decodificação do caminho dos dados para elencar onde e quando as falhas ocorrem e assim possibilitar a correção. O resultado final será um dashboard de visualização das falhas com uso de gráficos de tempo e contagem de erros para demonstrar quando a falha ocorreu. A forma como a falha é representada também pode servir de indicativo para a parte do processo que resultou no erro. A exemplo, um dos gráficos demonstrou triplicação dos dados enviados à camada raw, assim, encontramos uma falha que após cópia dos dados da base de produção, fora enviado três vezes a mesma cópia.

4 Qualidade de dados

Qualidade de dados é o processo pelo qual se desenvolvem ferramentas de avaliação e tratamento das informações de uma base de dados para adequá-la à métricas de qualidade. A avaliação engloba análise da precisão dos dados, ou seja, quanto esse dado representa a realidade, a exemplo avalia-se a base de uso do cientista de dados se possui informações corretas em relação à quantidade de desafios completada pelos usuários. Com informações precisas, exige-se também a consistência desse trânsito correto de informações entre as diferentes fontes de dados, por exemplo analisa-se essa informação de quantidade de desafios completada se é exibida de maneira igual entre as diferentes fontes de uso desses dados. Integridade e confiabilidade também são aspectos relevantes para a engenharia de dados, visto que uma base de dados precisa estar completa, sem dados faltantes e ter origem de fontes confiáveis, na FRST, as informações são coletadas da própria base de produção, essa que poucas pessoas possuem acesso cujo propósito é propor confiabilidade aos dados da empresa. Além disso, a métrica de atualidade é relevante ao proporcionar as informações em tempo hábil para empresa, por exemplo, leva-se cerca de um dia para viabilizar a quantidade de dados necessária para tomadas de decisão na FRST, enquanto em corridas de Fórmula 1 a informação deve ser repassada do piloto à equipe em milésimos de segundo e em bancos a informação pode levar até mesmo semanas para se tornar hábil ao uso.

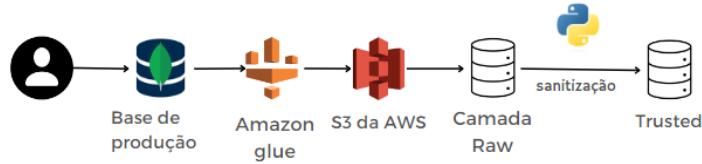
5 Processo de qualidade de dados

De início, todo comportamento de um usuário é registrado em uma base de dados de produção, ademais, o comportamento que nos interessa é a conclusão ou começo de desafios, quando um desafio é iniciado ele é adicionado na base e quando ele é deletado, atribui-se uma característica de 'deletado', mas, sem efetivamente o deletar. Em seguida, essa base é copiada, utilizando Amazon glue, e armazenada na camada raw da Amazon S3. Então, ocorre o processo de qualidade de dados na cópia e armazenamento dos dados da camada raw para a trusted.

Nessa medida, esse processo consiste em garantir que os dados de ambas as camadas estejam congruentes, ao verificar se não houve alguma repetição desnecessário ou corrompimento. Tal procedimento é feito em Python e pode ser reduzido à contagem de elementos em ambas os dados e na validação dessa contagem (verificar se estão iguais)

Por fim, pode-se visualizar esse processo completo na imagem abaixo:

Figura 1: Qualidade dados



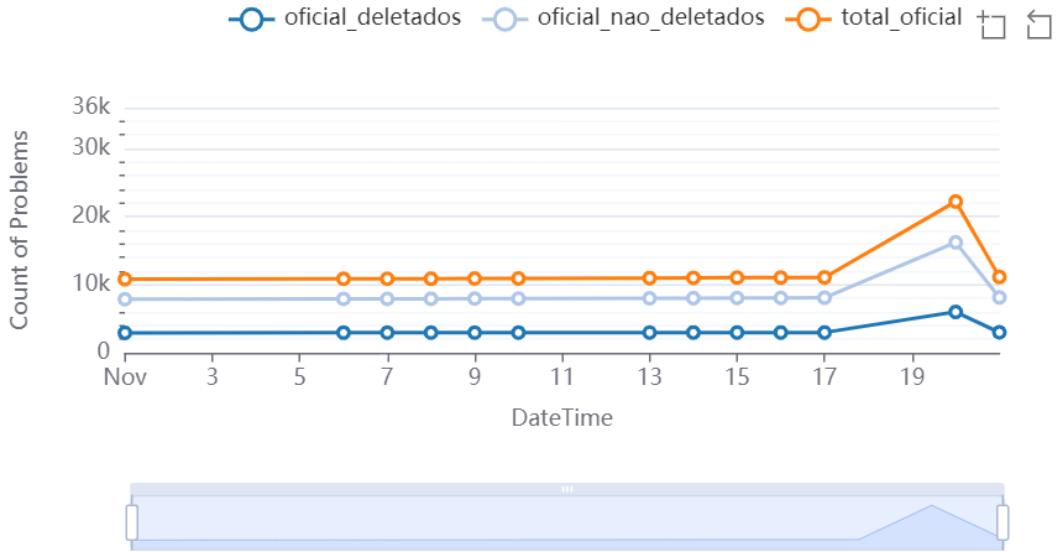
6 Hipótese e Objetivos

Tendo como a base a intuição construída, é possível formular a hipótese de que o erro ocorre durante o processo de cópia da base de produção para a raw ou no processo de qualidade dos dados. Assim, para validarmos essa hipótese iremos construir gráficos, utilizando superset, que comparam as informações da base de produção e na camada trusted. Dessa maneira, é possível validar a existência desse erro empíricamente.

7 Resultados

Na figura 2, pode-se observar, na base de produção, a contagem de desafios deletados, atuais e o total (atuais + deletados).

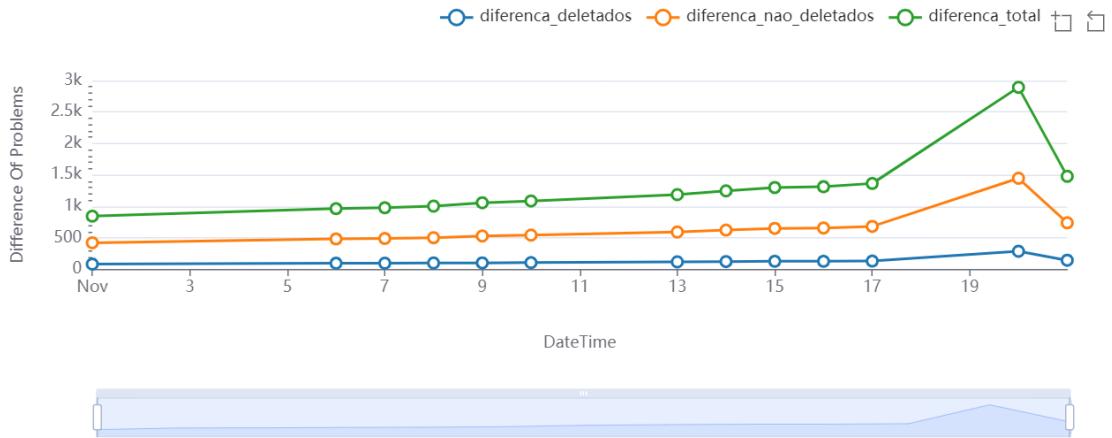
Figura 2: Total Desafios



Na figura 3, pode-se observar a diferença de desafios deletados, atuais e totais entre a base de produção e a camada trusted.

Analizando essas 2 figuras, foi possível perceber uma anomalia, na qual, a base teve um erro de cópia que resultou nela tendo suas informações duplicadas. Fato que foi possível observar graças a esses gráficos. Ademais, pode-se perceber que a diferença entre as duas bases continua a crescer.

Figura 3: Diferença Desafios



A seguir, podemos visualizar uma tabela (figura 4) que mostra cada um dos criterios da figura 3, mas, para o dia atual.

Figura 4: Diferença Desafios - Dia Atual

metric	Total (Sum)
diferenca_total	739
diferenca_nao_deletados	595
diferenca_deletados	144
Total (Sum)	1.48k

Por fim, a figura 5 é uma sequência de imagens, em que, a mais à esquerda representa a porcentagem de empresas com uma incongruência na base, a central é a porcentagem de empresas sem uma incongruência e a última é um histograma da contagem de erros em cada empresa.

Figura 5: Análise de erros



Assim, fica mais eficiente achar em que parte da base ocorreu um erro, pois, aponta diretamente para a empresa. Nessa medida, salva-se muito tempo diagnosticando o erro.

8 Conclusão

Em resumo, a integração de Engenharia de Dados e Qualidade de Dados apresenta-se como um diferencial estratégico para a FRST Falconi. Este trabalho não só destaca as capacidades transformadoras dessas disciplinas, mas também sinaliza um futuro no qual dados de qualidade impulsionarão decisões corporativas mais informadas.

Ao consolidar essa abordagem, firma-se um compromisso de contribuir para fortalecer o ecossistema de ciência de dados na parceria contínua com a FRST Falconi.

Neste sentido, há um cenário onde os dados não apenas informam, mas capacitam a tomada de decisões empresariais mais precisas e estratégicas.

Logo, é de fundamental importância entender o sentido do processo de tornar os dados confiáveis, eficientes, precisos, integros e atuais para a melhor tomada de decisão.

9 Referências

- V. Mayer-Schonberger and K. Cukier, *Big Data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.
- J. R. GalbRaith, *Organizational Design Challenges Resulting From Big Data*, *Journal of Organisation Design*, Apr 2014
- <https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year>

10 Ferramentas

- Amazon Web Services: S3 e Athena são os serviços da AWS utilizados neste projeto e que servem para armazenar os dados que foram trabalhados.
- MongoDB: Banco de dados NoSQL também responsável por armazenar os dados, no entanto, neste caso, o MongoDB é a base original de produção da empresa onde estão todos os desafios propostos por clientes.
- Python : linguagem em que houve o desenvolvimento da aplicação.
- Airflow: ferramenta responsável por construir pipelines de dados que são etapas afim de avaliar quaisquer erros do código desenvolvido.
- Superset: responsável pela construção do produto final, ou seja, o dashboard de visualização de dados.