Universidade Tecnológica Federal do Paraná

Departamento Acadêmico de Informática
Curso de Bacharelado em Sistemas de Informação

Relatório

Projeto

Bruno Wandrey
Daniel Vieira
Eduardo Darrazão

Curitiba
Maio, 2022

Universidade Tecnológica Federal do Paraná

Departamento Acadêmico de Informática
Curso de Bacharelado em Sistemas de Informação

Relatório

Projeto

Relatório referente à Disciplina Introdução à Ciência de Dados do Curso de Bacharelado de Sistemas de Informação da Universidade Tecnológica Federal do Paraná (UTFPR).

Curitiba

Maio, 2022

Sumário

1 Introdução

1.1 Tema e Equipe

A equipe *Dubruda* possuí três integrantes, todos cursando Bacharelado em Sistemas de Informação na UTFPR-Curitiba. Nosso trabalho tem como objetivo medir o impacto de políticas públicas na mortalidade por *COVID-19* nas capitais brasileiras e na maior cidade de cada estado além da capital.

2 Perguntas de Pesquisa

Esperamos encontrar possíveis indicadores de como as políticas públicas sobre COVID-19 afetam a quantidade de óbitos em razão do mesmo. Por conseguinte, construímos as seguintes perguntas que procuraremos responder:

1. A aplicação de políticas públicas reduziu a quantidade de novas mortes por COVID-19?
2. O fechamento de escolas reduziu a quantidade de óbitos causados pela COVID-19?
3. O fechamento de ambientes de trabalho reduziu a quantidade de óbitos causados pelo COVID-19?
4. O cancelamento de eventos públicos reduziu a quantidade de óbitos causados pelo COVID-19?
5. A imposição de restrições de aglomerações reduziu a quantidade de óbitos causados pelo COVID-19?
6. O fechamento dos transportes públicos reduziu a quantidade de óbitos causados pelo COVID-19?
7. A imposição de quarentena reduziu a quantidade de óbitos causados pelo COVID-19?
8. As restrições na movimentação entre cidades/estados reduziram a quantidade de óbitos causados pelo COVID-19?
9. O controle de viagens internacionais reduziu a quantidade de óbitos causados pelo COVID-19?
10. As campanhas de conscientização reduziram a quantidade de óbitos causados pelo COVID-19?

3 Hipóteses

1. A aplicação de políticas públicas reduziu a quantidade de novas mortes por COVID-19.
2. O fechamento de escolas reduziu a quantidade de óbitos causados pela COVID-19.
3. O fechamento de ambientes de trabalho reduziu a quantidade de óbitos causados pelo COVID-19.
4. O cancelamento de eventos públicos reduziu a quantidade de óbitos causados pelo COVID-19.
5. A imposição de restrições de aglomerações reduziu a quantidade de óbitos causados pelo COVID-19.
6. O fechamento dos transportes públicos reduziu a quantidade de óbitos causados pelo COVID-19.
7. A imposição de quarentena reduziu a quantidade de óbitos causados pelo COVID-19.
8. As restrições na movimentação entre cidades/estados reduziram a quantidade de óbitos causados pelo COVID-19.
9. O controle de viagens internacionais reduziu a quantidade de óbitos causados pelo COVID-19.
10. As campanhas de conscientização reduziram a quantidade de óbitos causados pelo COVID-19.

4 Dados e modelos

4.1 Visualização

Após realizada a análise exploratória, encontramos e organizamos os dados que serão utilizados para testar e demonstrar nossas hipóteses. Estes dados incluem:

- Quantidade de novas mortes (ponderada) por semana epidemiológica (semanas de 2020, 2021 e até março de 2022), em cada cidade. (1)
- Variação do Índice de Rigor por semana em cada cidade. O Índice de Rigor representa a média simples normalizada dos nove indicadores que estamos estudando, C1 a C8 e H1. (2 e 3)
- Relação do Índice de Rigor em cada cidade (nos anos de 2020 e 2021) com a quantidade de mortes (ponderada). (4 e 5)
- Agrupamento das 54 cidades em grupos com IDHM próximo. (6)

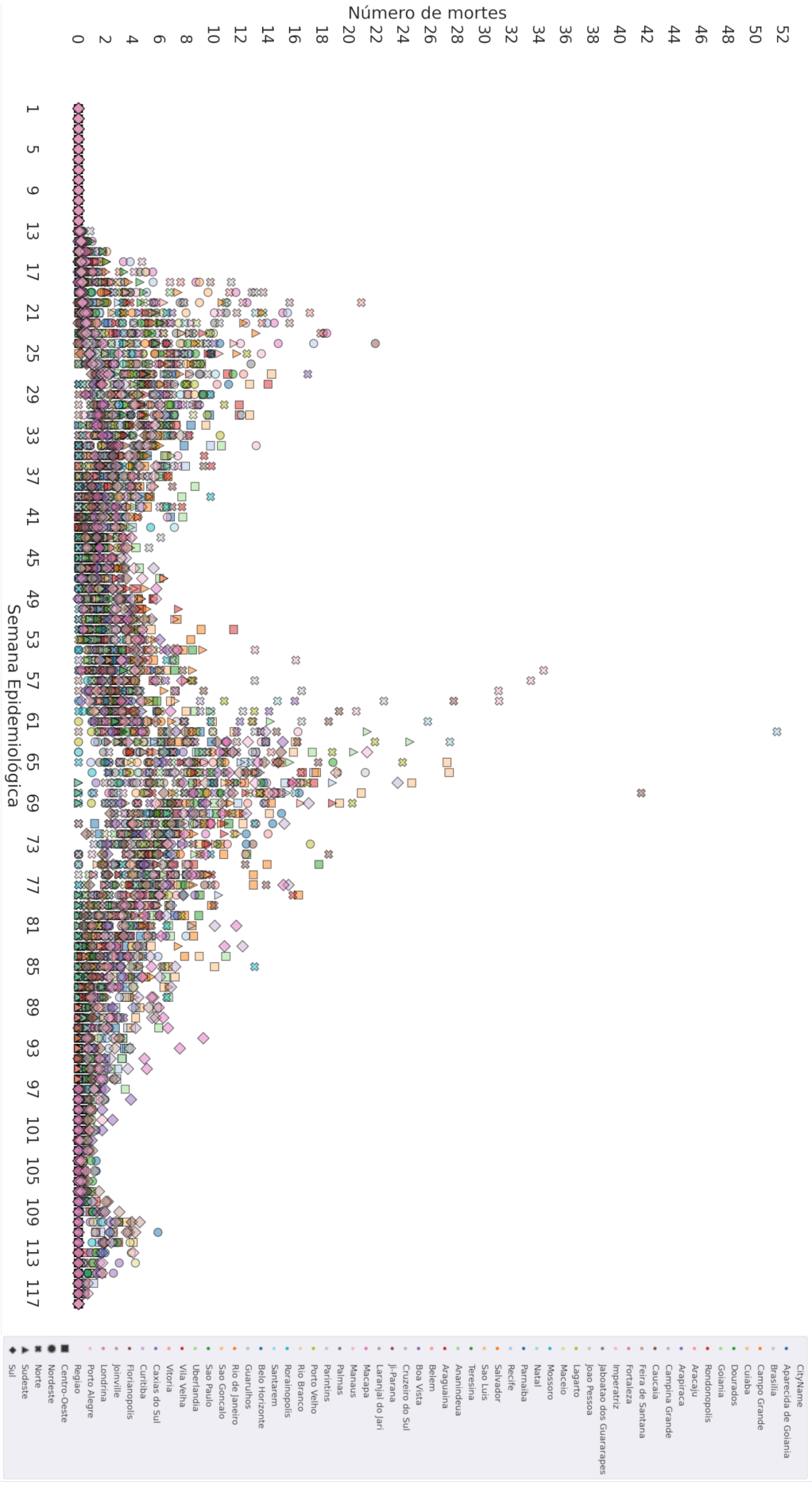

Figura 1 – Número ponderado de mortes por semana epidemiológica para cada cidade. A legenda está ordenada por região para realizar a identificação dos ícones mais facilmente. É possível perceber duas grandes ondas completas no período.

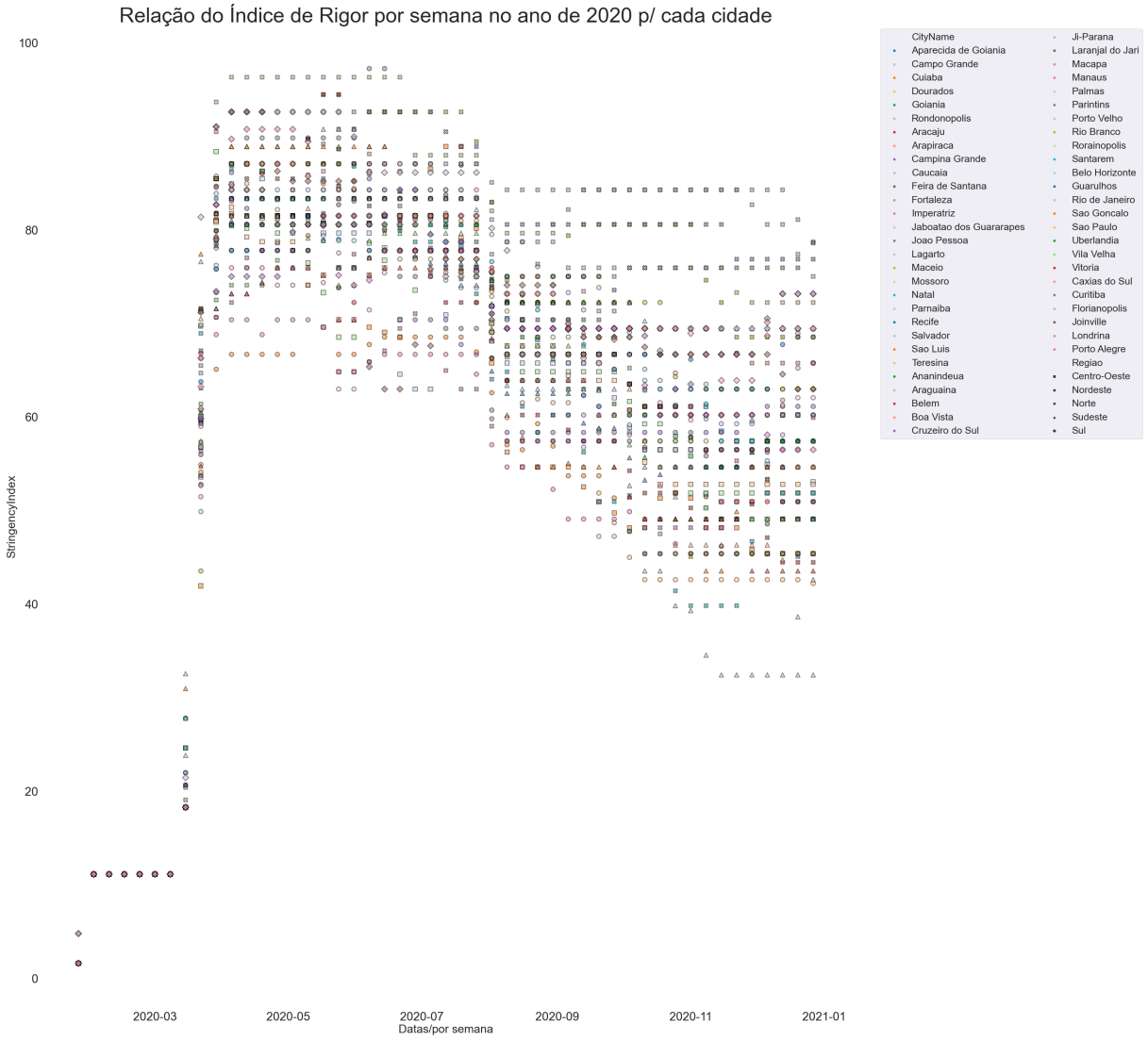

Figura 2 – Podemos perceber aqui como os indicadores variaram desde o início da pandemia até o final do ano de 2020. Por exemplo, percebemos que no início cidades do Norte do Brasil já começaram bem rigorosas, e a maioria se manteve assim. Já cidades do Nordeste foram bem discrepantes, algumas sendo bem rigorosas e outras pouco.

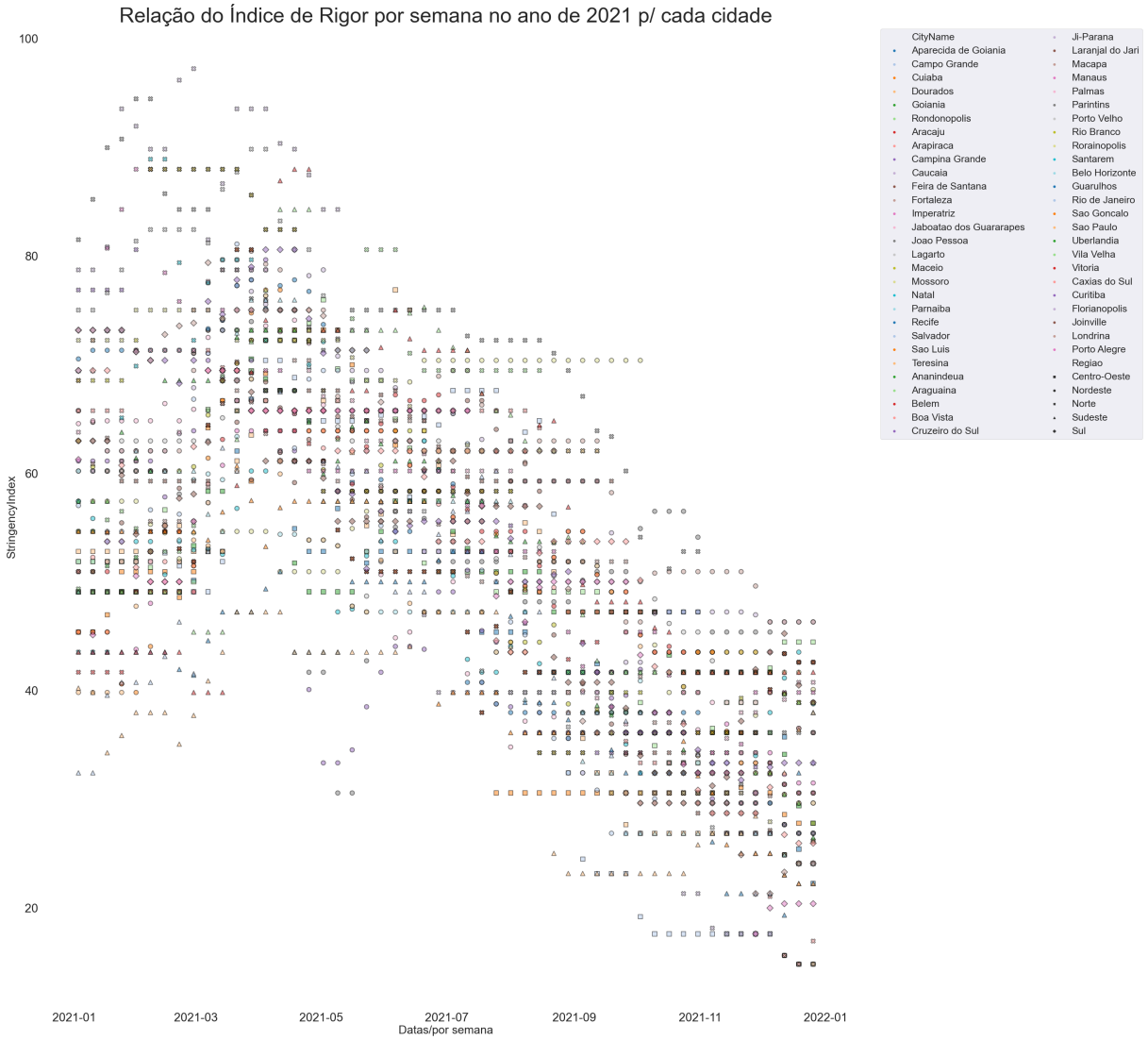

Figura 3 – No ano de 2021 houveram muito mais mortes devido a pandemia, e é nítido que o Norte se manteve rigoroso quanto as políticas. No geral o Sudeste foi condescendente, com as médias mais baixas de rigor. As cidades do Nordeste se mantiveram discrepantes, umas relativamente rigorosas e outras com políticas bem brandas.

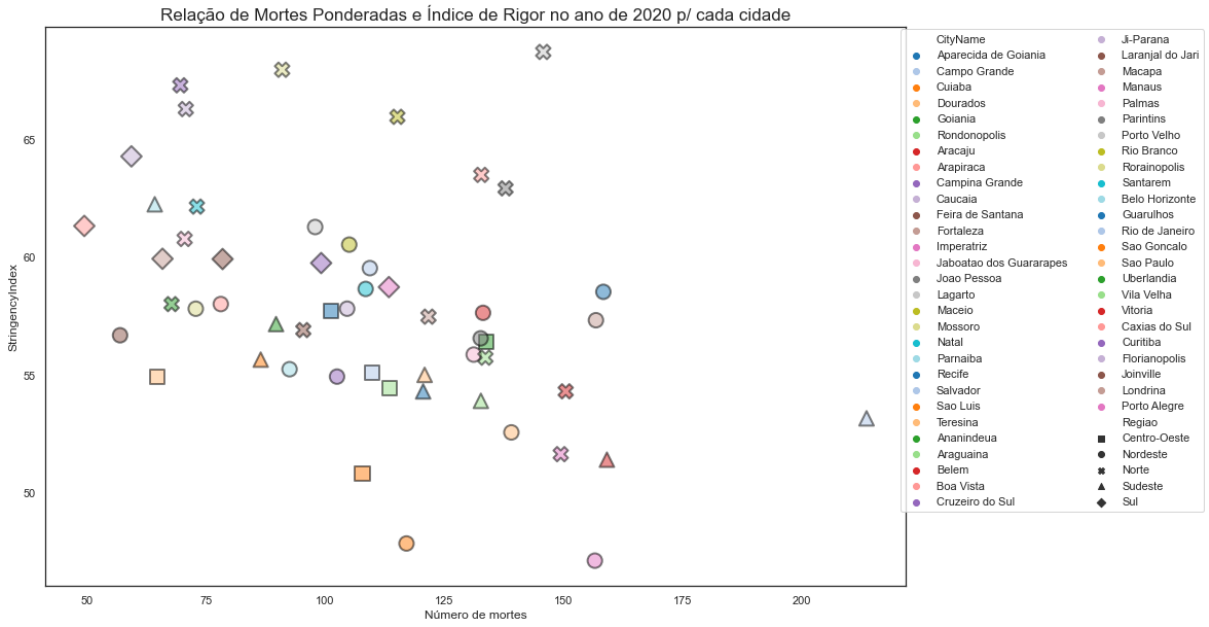

Figura 4 – Nota-se que neste primeiro ano de pandemia a cidade que foi menos rigorosa disparadamente foi Cuiabá, Centro-Oeste, contudo não teve muitas mortes. Rio de Janeiro, Sudeste, apesar de ter sido bem rigorosa morreu muitas pessoas. É possível perceber também que na parcela de cidades que morreram menos pessoas (0-80) há uma tendência a uma maior concentração na parte superior do eixo Y, ou seja, mais concentrado em cidades mais rigorosas.

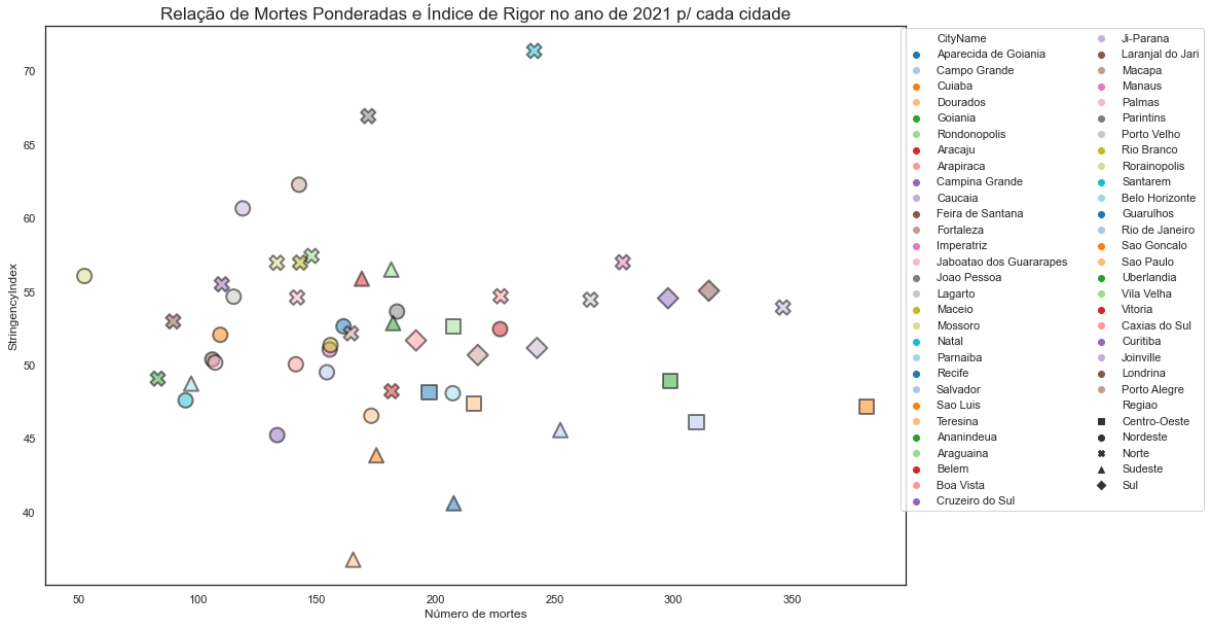

Figura 5 – No ano de 2021, onde houve mais óbitos por COVID-19, a possível correlação entre cidades mais rigorosas tiveram menos mortes fica mais evidente. No geral as cidades que mais morreram pessoas foram bem lenientes no rigor das políticas públicas, enquanto as mais rigorosas no geral se sairam bem melhor.

Figura 6 – Separamos as cidades em 5 grupos de IDHM, para utilização como variável de controle.

4.2 Estatística

O conjunto de dados trabalhado, *Brazil-covid-policy*, é um conjunto de dados que pode ser considerado *Panel Data*, que consiste em um conjunto de dados ao longo do tempo referente a diferentes entidades , no caso deste trabalho: dados de municípios ao longo do tempo (dias). (SMOLSKI, 2019; NAGATA; LEITE, 2019; TORRES-REYNA, 2007)

Existem três tipos de modelagem de dados baseado em *Panel Data Regression*, escolhemos utilizar o *Pooled* (BRUGGER, 2021), por ser a forma mais simples e ainda produzir um resultado satisfatório, dessa forma é o tipo que melhor se adéqua as condições de trabalho que dispomos dado os limites da disciplina.

Construiremos um modelo Panel OLS (*Ordinary Least Squares*) com os dados das 54 cidades avaliadas, no período da 10ª semana de 2020 (02/03/2020) até a 45ª semana de 2020 (02/11/2020), que é o período referente a primeira onda de intensificação da propagação do COVID-19 no Brasil, como pode ser verificado na imagem 1.

Serão criados dois modelos, cada um com um conjunto próprio de variáveis dependentes: o primeiro conjunto consistirá das entradas dos indicadores $C1, C2, ..C8$ e $H1$ (para cada dia e cada cidade), e o IDH de cada cidade, que será usado como variável de controle. Estes indicadores dizem respeito ao fechamento das escolas, fechamento dos locais de trabalho, cancelamento de eventos públicos, restrições de aglomerações, fechamento de transportes públicos, imposição de quarentena domiciliar, restrições na movimentação entre cidades/estados, restrições em viagens internacionais e campanhas de conscientização quanto a pandemia; o segundo conjunto de variáveis dependentes será composto pela variável de controle IDH para cada cidade e pelo Índice de Rigor (*StringencyIndex*), que é uma variável que representa os nove indicadores trabalhados. A variável independente do modelo será a quantidade de novas mortes, que é calculada por dia, por cidade, ponderada por 100 mil habitantes e levando em consideração a pirâmide etária da população de cada cidade.

Ao refletirmos sobre o fato que existe um período entre o momento que uma política pública é implantada até o momento que ela atinge o ápice da sua eficácia e por conseguinte reduz a quantidade de contaminação por COVID-19, e que ainda existe outro período entre o momento que a redução de infecções atua na redução de mortes por COVID-19, inseriremos um *lag* na variável independente (novas mortes).

Junto com conversas com o orientador, decidimos retardar a variável independente em 37 dias após o momento da implantação das políticas públicas. Esta quantidade foi escolhida considerando que: levam 7 dias para uma política pública chegar no ápice da eficácia, depois levam mais 10 dias para reduzir a quantidade de infecções, e após mais 20 dias para atuar na redução de mortes. Vale citar que este *lag* inserido acontecerá

separadamente para cada cidade.

Com o modelo pronto para os dois grupos de variáveis dependentes, construiremos uma visualização do tipo *ScatterPlot* dos resíduos de cada grupo, para analisar a distribuição dos resíduos e descobrir se estão coerentes com o esperado de um modelo de regressão.

Por fim, algo que poderá ser estudado e analisado no modelo com o primeiro grupo de variáveis dependentes, é o fenômeno denominado *Granger Causality*, devido a aparente correlação entre alguns dos indicadores, como pode ser visualizado nas imagens 7 e 8.

	C1_School closing	C2_Workplace closing	C3_Cancel public events	C4_Restrictions on gatherings	C5_Close public transport
C1_School closing	1.0000	0.6342	0.6217	0.5075	0.4277
C2_Workplace closing	0.6342	1.0000	0.6858	0.5422	0.3779
C3_Cancel public events	0.6217	0.6858	1.0000	0.5699	0.3591
C4_Restrictions on gatherings	0.5075	0.5422	0.5699	1.0000	0.4062
C5_Close public transport	0.4277	0.3779	0.3591	0.4062	1.0000
C6_Stay at home requirements	0.5099	0.4934	0.4805	0.3702	0.3424
C7_Restrictions on internal movement	0.4035	0.3845	0.3372	0.3592	0.3164
C8_International travel controls	0.3507	0.3364	0.3037	0.3448	0.2958
H1_Public information campaigns	0.3606	0.3483	0.3798	0.2695	0.2849

Figura 7 – Correlação *Pearson* entre as variáveis C1-C8 e H1

	C6_Stay at home requirements	C7_Restrictions on internal movement	C8_International travel controls	H1_Public information campaigns
C1_School closing	0.5099	0.4035	0.3507	0.3606
C2_Workplace closing	0.4934	0.3845	0.3364	0.3483
C3_Cancel public events	0.4805	0.3372	0.3037	0.3798
C4_Restrictions on gatherings	0.3702	0.3592	0.3448	0.2695
C5_Close public transport	0.3424	0.3164	0.2958	0.2849
C6_Stay at home requirements	1.0000	0.4007	0.2388	0.2303
C7_Restrictions on internal movement	0.4007	1.0000	0.4205	0.1570
C8_International travel controls	0.2388	0.4205	1.0000	0.2523
H1_Public information campaigns	0.2303	0.1570	0.2523	1.0000

Figura 8 – Segunda parte da tabela de correlação *Pearson* entre as variáveis C1-C8 e H1

4.3 Aprendizado de Máquina

Usaremos a sessão de *Aprendizado de Máquina* como suporte para a análise da verificação das hipóteses: agrupando cidades em função dos atributos mais relevantes dentro do escopo do trabalho (tamanho percentual do grupo de risco e IDHM). Esse agrupamento é proposto com o objetivo de oferecer uma forma visual de compreender o contexto inicial das cidades e também oferecer uma base de comparação direta para cidades mais similares (parte do mesmo grupo).

Nossa expectativa é poder avaliar um recorte relevante das condições de cada cidade, uma vez que o recorte foi projetado para representar simultaneamente a suscetibilidade da população à morte causada pelo vírus SARS-CoV-2 (representada pelo grupo de risco) e a capacidade de implementar medidas de redução de danos (representada pelo IDHM).

4.3.1 Algoritmo

Pretendemos usar o algoritmo *DBSCAN* para agrupar cidades de acordo com os atributos previamente mencionados, sendo que cada cidade será representada por um vetor bidimensional, na forma:

$$Cidade_k = [R_k, I_k]$$

Onde R_k representa o tamanho percentual do grupo de risco para a cidade k e I_k representa o IDHM da cidade k.

O DBSCAN usa a ideia de densidade de distribuição dos dados para criar grupos, de forma que para cada ponto de um grupo, a vizinhança no raio r contém n pontos ou mais, ou seja, a densidade na vizinhança tem que exceder o limiar, por isso, em geral, a vizinhança de um ponto de borda contém significativamente menos pontos do que a vizinhança de um ponto central do grupo.

Como a utilização do DBSCAN não exige a definição prévia do número de grupos resultantes, essa definição é baseada na densidade da distribuição e vai além da análise visual da dispersão dos dados de cada cidade no espaço bidimensional, o que acreditamentos que poderia ser uma fonte de problemas durante o processo de agrupamento, pois trata-se de uma análise subjetiva de um conjunto de dados que pode não ser claramente separável. O algoritmo escolhido também trata possíveis cidades que não se encaixam bem em um grupo, por serem casos específicos demais, dessa forma não há prejuízo necessariamente quando comparado com a observação e tratamento de um *outlier* no agrupamento por K-Means, por exemplo.

O DBSCAN depende de parâmetros definidos pelo usuário: o tamanho do raio que delimita a vizinhança e a quantidade mínima de pontos para que a vizinhança seja considerada densa. Para nós é esperado que a definição e avaliação desses parâmetros serão os maiores desafios para a aplicação do algoritmo na nossa base de dados.

5 Cronograma

1. Separar os dados com apenas as entradas e informações que serão utilizadas no modelo, (indicadores C1, C2, ..., C8, H1; somente entradas do período da primeira onda; índice de rigor; e possíveis dados auxiliares).
2. Agregar o IDH de cada cidade para cada entrada de cada cidade.
3. Construir a ponderação de novas mortes por cidade com o intervalo diário.
4. Realizar o *lag* das mortes em 37 dias.
5. Separar os dados para cada modelo.
6. Criar o modelo que contém os indicadores como variáveis dependentes.
7. Construir o *ScatterPlot* dos resíduos do primeiro modelo.
8. Criar o modelo que contém o Índice de Rigor como variável dependente.
9. Construir o *ScatterPlot* dos resíduos do segundo modelo.
10. Integrar dados de IDHM e faixa etária
11. Aplicar DBSCAN para obter grupos
12. Avaliar grupos
13. Corrigir agrupamento
14. Avaliar modelos
15. Avaliar possível correlação das variáveis do primeiro modelo (*Granger Causality*).

Referências

BRUGGER, B. **A Guide to Panel Data Regression: Theoretics and Implementation with Python**. 2021. Disponível em: <https://towardsdatascience.com/a-guide-to-panel-data-regression-theoretics-and-implementation-with-python-4c84c5055cf8>. Citado na página 12.

NAGATA, M. H.; LEITE, A. W. **Dados em Painel**. 2019. Disponível em: <https://lamfo-unb.github.io/2019/09/11/Dados-em-painel/>. Citado na página 12.

SMOLSKI, F. M. da S. **Capítulo 6 Regressão com Dados em Painel**. 2019. Disponível em: <https://smolski.github.io/livroavancado/regressao-com-dados-em-painel.html>. Citado na página 12.

TORRES-REYNA, O. **Panel Data AnalysisFixed and Random Effects using Stata**. 2007. Disponível em: <https://www.princeton.edu/~otorres/Panel101.pdf>. Citado na página 12.