Universidade Tecnológica Federal do Paraná

Departamento Acadêmico de Informática
Curso de Bacharelado em Sistemas de Informação

Relatório

Resultados Finais

Bruno Wandrey
Daniel Vieira
Eduardo Darrazão

Curitiba
Junho, 2022

Universidade Tecnológica Federal do Paraná

Departamento Acadêmico de Informática
Curso de Bacharelado em Sistemas de Informação

Relatório

Resultados Finais

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

Junho, 2022

Sumário

1 Problema, Materiais e Métodos

Nosso trabalho tem como objetivo medir o impacto das políticas públicas na mortalidade por *COVID-19* nas capitais brasileiras e na maior cidade de cada estado além da capital. Com esta análise, podemos sugerir como melhor lidar com situações análogas, como por exemplo, ênfase nas medidas que tiveram melhor resposta na redução da quantidade de mortes. Sendo assim nossa pergunta de pesquisa é: A aplicação de políticas públicas impactou na quantidade de novas mortes por COVID19? Para respondê-la, realizamos uma regressão linear múltipla para projetar o impacto das políticas públicas e das variáveis de controle na quantidade de mortes ponderadas por faixa etária e por 100 mil habitantes 45 dias após a implantação das políticas. Com isso conseguimos analisar os coeficientes e perceber que as políticas obtiveram impactos variados.

Os principais dados foram obtidos no repositório *Brazil-covid-policy* [*]. O dataset é um conjunto de indicadores sobre políticas públicas implantadas durante a pandemia de COVID-19. Os dados são indexados por dia, cidade, estado e jurisdição. Cada um desses indicadores funciona de forma similar, onde o valor zero corresponde a não implantação da política, e o maior valor corresponde a implantação total da política. Uma descrição mais ampla de cada indicador pode ser encontrada no *codebook* do repositório *Brazil-covid-policy*.

Usamos dados da mortalidade por Covid-19 oriundos da *National Center for Health Statistics*[†] para gerar uma distribuição estatística dos números brutos de mortes para projetar o perfil das mortes em cada faixa etária. Consideramos que a doença produziu um efeito na população brasileira similar ao efeito produzido na população estadunidense, em relação a taxa de mortalidade por idade. Para realizar a adaptação dos dados para o nosso caso, foi necessário criar uma relação dos dados que já possuíamos com os dados demográficos do Brasil, obtivemos parte desses dados na página do IBGE e parte foi disponibilizada pelo orientador, sendo essa parte obtida através do portal DataSUS.

Utilizamos dados sobre *IDHM - Índice de Desenvolvimento Humano Municipal* do censo feito em 2010, encontrado no *Atlas do Desenvolvimento Humano de 2013* [‡]. Estes dados são utilizado como variável de controle no modelo, dando ênfase nas características locais de cada região. Nota-se que os dados utilizados de 2010 são os mais atuais, dado que não ocorreu censo no Brasil em 2020. A ponderação do número de mortes foi feita considerando um ajuste por idade e tamanho da população, esse é um processo estatístico aplicado a taxas de mortalidade que permite comparar comunidades com diferentes

[*]disponível em https://github.com/OxCGRT/Brazil-covid-policy
[†]Disponível em: https://data.cdc.gov/d/9bhg-hcku
[‡]Disponível em:https://censo2010.ibge.gov.br/resultados.html

estruturas etárias. Uma vez que quase todas as doenças possuem mortalidade diferentes em diferentes faixas etárias. Uma maneira de examinar os resultados em comunidades de diferentes tamanhos é calcular uma taxa de incidência ou mortalidade, que é o número de novos casos ou mortes dividido pelo tamanho da população (HEALTH, 1999).

					Grupo de risco				
	0 a 4	5 a 14	15 a 24	25 a 34	35 a 44	45 a 54	55 a 64	65 a 74	>=75
Média	6,86%	14,48%	16,60%	16,78%	15,91%	12,04%	8,96%	5,29%	3,08%
Média Acumulada	54,72%				27,94%		17,34%		
DP	1,18	2,25	1,90	1,21	1,01	1,28	1,77	1,45	1,08
Mortes	0,04%	0,03%	0,26%	1,11%	2,77%	6,68%	14,70%	23,00%	51,41%
Mortes Acumuladas	1,45%				9,45%		89,11%		
Mínimo	5,18%	10,88%	13,22%	14,48%	12,53%	8,61%	5,15%	2,56%	1,15%
25%	6,09%	12,83%	14,78%	16,10%	15,60%	11,65%	7,82%	4,43%	2,44%
Mediana	6,51%	14,30%	16,34%	16,70%	16,03%	12,46%	9,14%	5,31%	3,00%
75%	7,48%	15,44%	18,03%	17,23%	16,45%	12,87%	10,24%	6,06%	3,70%
Máximo	10,74%	20,86%	20,95%	21,31%	18,07%	13,59%	11,95%	8,68%	5,98%

Figura 1 – Tabela com resumo das distribuições de população por faixa etária do Brasil e a distribuição dos Estados Unidos de mortes por faixa etária.

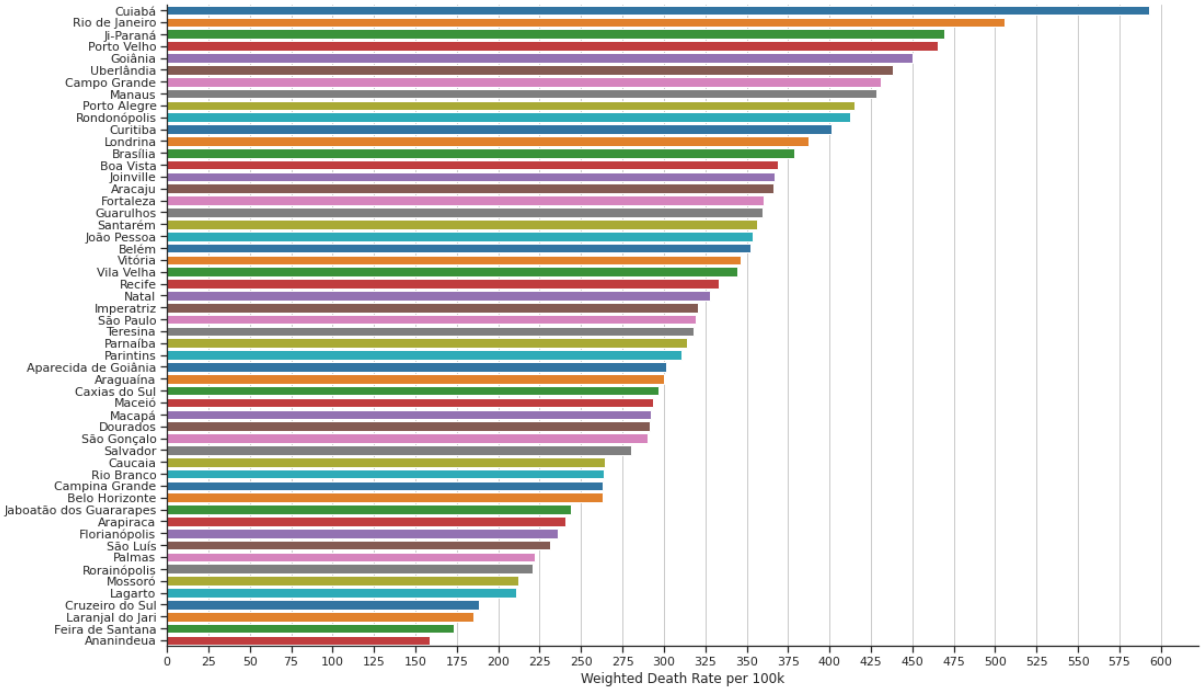

Figura 2 – Cidades avaliadas ordenadas por número ponderado de mortes.

2 Resultados

Nosso modelo é um OLS onde a variável dependente é a quantidade de mortes ponderadas por faixa etária de cada cidade e por 100 mil habitantes, esta quantidade é referente às mortes de 45 dias após a data da entrada. Aplicamos uma média móvel nos valores para deixá-los mais uniformes e melhorar a capacidade de projeção do modelo.

As variáveis independentes que foram utilizadas para predizer as mortes são: o índice de rigorosidade de cada política pública (C1-8, H1), o IDHM, a quantidade de novos casos de infecção por COVID-19 ponderado por 100 mil habitantes, e 54 variáveis que representam cada cidade, de modo categorizado. Esta *dummyzação* das cidades é uma simplificação do *FixedEffects*, onde buscamos capturar diferenças intrínsecas a cada cidade, como cultura, estilo de vida e características regionais.

Tabela 1 – Metadados *OLS*

Dep. Variable:	Morte Ponderada (t+45)	R-squared:	0.298
Model:	OLS	Adj. R-squared:	0.294
Method:	Least Squares	F-statistic:	74.51
Date:	Wed, 22 Jun 2022	Prob (F-statistic):	0.00
Time:	21:35:06	Log-Likelihood:	-3664.2
No. Observations:	10767	AIC:	7452.
Df Residuals:	10705	BIC:	7904.
Df Model:	61		
Covariance Type:	nonrobust		

Tabela 2 – Indicadores de políticas públicas e variáveis de controle IDHM e NewCases-Ponderado

INDICADOR	coef	std err	t	P>|t|	[0.025	0.975]
C1_School closing	-0.0107	0.013	-0.839	0.401	-0.036	0.014
C2_Workplace closing	-0.0200	0.008	-2.405	0.016	-0.036	-0.004
C3_Cancel public events	0.1666	0.016	10.335	0.000	0.135	0.198
C4_Restrictions on gatherings	-0.0161	0.005	-3.089	0.002	-0.026	-0.006
C5_Close public transport	0.0893	0.011	8.374	0.000	0.068	0.110
C6_Stay at home requirements	-0.1433	0.011	-12.993	0.000	-0.165	-0.122
C7_Restrictions on internal movement	0.0388	0.008	4.835	0.000	0.023	0.055
C8_International travel controls	0.1141	0.003	36.097	0.000	0.108	0.120
H1_Public information campaigns	-0.2109	0.023	-9.145	0.000	-0.256	-0.166
IDHM	0.3234	0.027	12.082	0.000	0.271	0.376
Novos casos ponderado	0.0001	0.000	1.035	0.301	-0.000	0.000

O modelo que propomos para a regressão linear múltipla está apresentado na expressão 2.1, com t sendo a data, M_p^{t+45} a quantidade de mortes ponderadas 45 dias

após a data, αZ o *intercept* e seu coeficiente α, βN^t o número de novas infeções na data t e seu coeficiente, γH o IDHM e coeficiente, $\delta_i P_i$ o indicador de política pública i e seu coeficiente e $\sigma_j C_j$ os valores categorias das cidades e respectivos coeficientes. Como pra cada entrada só uma das cidades terá valor 1 e o resto 0, será um somatório que representa somente a cidade da entrada e seu coeficiente.

$$M_p^{t+45} = \alpha Z + \beta N^t + \gamma H + \sum_{i=1}^{9} (\delta_i P_i) + \sum_{j=1}^{54} (\sigma_j C_j) \tag{2.1}$$

Analisando os coeficientes, vemos que algumas medidas possuem mais peso na projeção da mortalidade, entretanto não podemos afirmar que o efeito é positivo ou negativo, pelo fato da possibilidade de correlações cíclicas entre a variável dependente com as independentes (políticas públicas). Sendo assim todas as políticas influenciaram de algum modo a quantidade de óbitos, menos o fechamento das escolas que não teve um resultado estatisticamente significativo. Percebemos também que utilizar as infecções por COVID-19 como variável de controle não obteve efeito/não teve impacto significativo.

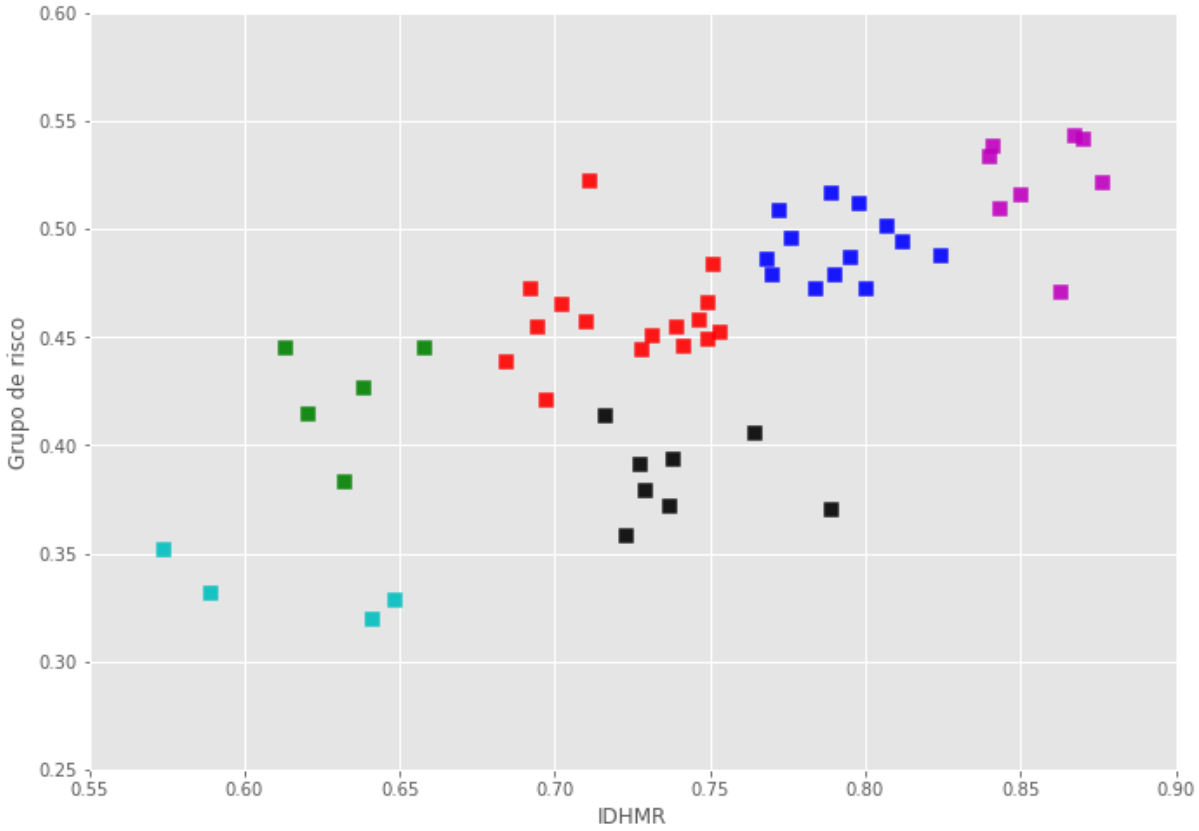

Figura 3 – Cidades agrupadas por IDHMR e tamanho percentual do grupo de risco.

Usamos o algoritmo *K-Means* para realizar *Aprendizado de Máquina* como suporte para a análise da verificação da hipótese: agrupando cidades em função de atributos impor-

tantes para o 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). Assim conseguimos 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).

Como apenas os grupos 2 e 3 apresentaram comportamentos parecidos na evolução do número ponderado de mortes, concluímos que os participantes de cada grupo formado não possuem grande similaridade na forma como lidaram com a pandemia. Partindo da nossa hipótese inicial, os membros de um mesmo grupo que apresentam comportamento muito diferente entre si são objetos interessantes para uma investigação comparativa mais aprofundada para descobrir novos elementos definitivos para trabalhos futuros, pois as condições iniciais eram similares (o que permite comparação nesses termos) e ao mesmo tempo não foram suficientes para determinar a trajetória daquela cidade (o que nos leva a querer compreender o que separa essas cidades no espaço das nossas variáveis), provavelmente porque as cidades tiveram diferentes graus de dificuldade e problemas decorrentes da COVID-19.

3 Limitações e Trabalhos Futuros

Dentre as limitações, o conjunto de dados *Brazil-covid-policy* possui muitas inconsistências e dados faltantes, como o caso dos dados referentes à Brasília estão todos vazios. Os dados que utilizamos como *features* para capturar o problema provavelmente não são diretamente relacionadas com as mortes. Por exemplo, não é o fato de as escolas fecharem que faz com que a quantidade de infecções diminua, mas sim, o fato de que com isso, as pessoas fiquem mais em casa. O modelo também sofre de dependências cíclicas: existe uma grande possibilidade de o número de mortes ser o motor das políticas públicas, e existe também a possibilidade da existência de *Correlações Espúrias*, ou seja, o fato de que as variáveis não possuem uma relação de causa e efeito diretamente, mas sim, dependem de um terceiro fator não explorado no modelo.

O modelo utilizado é altamente complexo, e sofreu várias simplificações até chegarmos no modelo final. Provavelmente deveríamos ter utilizado modelagem com o *PanelOLS* (visto que nossos dados são em painel), que foi simplificado para *OLS*. Também simplificamos a utilização de *Fixed Effects* utilizando no lugar um coeficiente para cada cidade.

No futuro, poderíamos utilizar de fato *Fixed Effects* e *Time Effects* de maneira completa, não simplificada. Também precisamos melhorar as variáveis utilizadas no modelo, como por exemplo, aplicar logaritmo em exponenciais e adicionar *lags* mais variados e em diferentes *features*. Seria interessante inclusive realizar analises mais profundas dos resíduos, com intuito de compreender padrões que dificultam a modelagem (e.g. variações intensas nos valores).

Referências

HEALTH, D. of. **Age-Adjusted Rates - Statistics Teaching Tools**. 1999. Disponível em: <https://www.health.ny.gov/diseases/chronic/ageadj.htm>. Citado na página 4.

4 Anexo

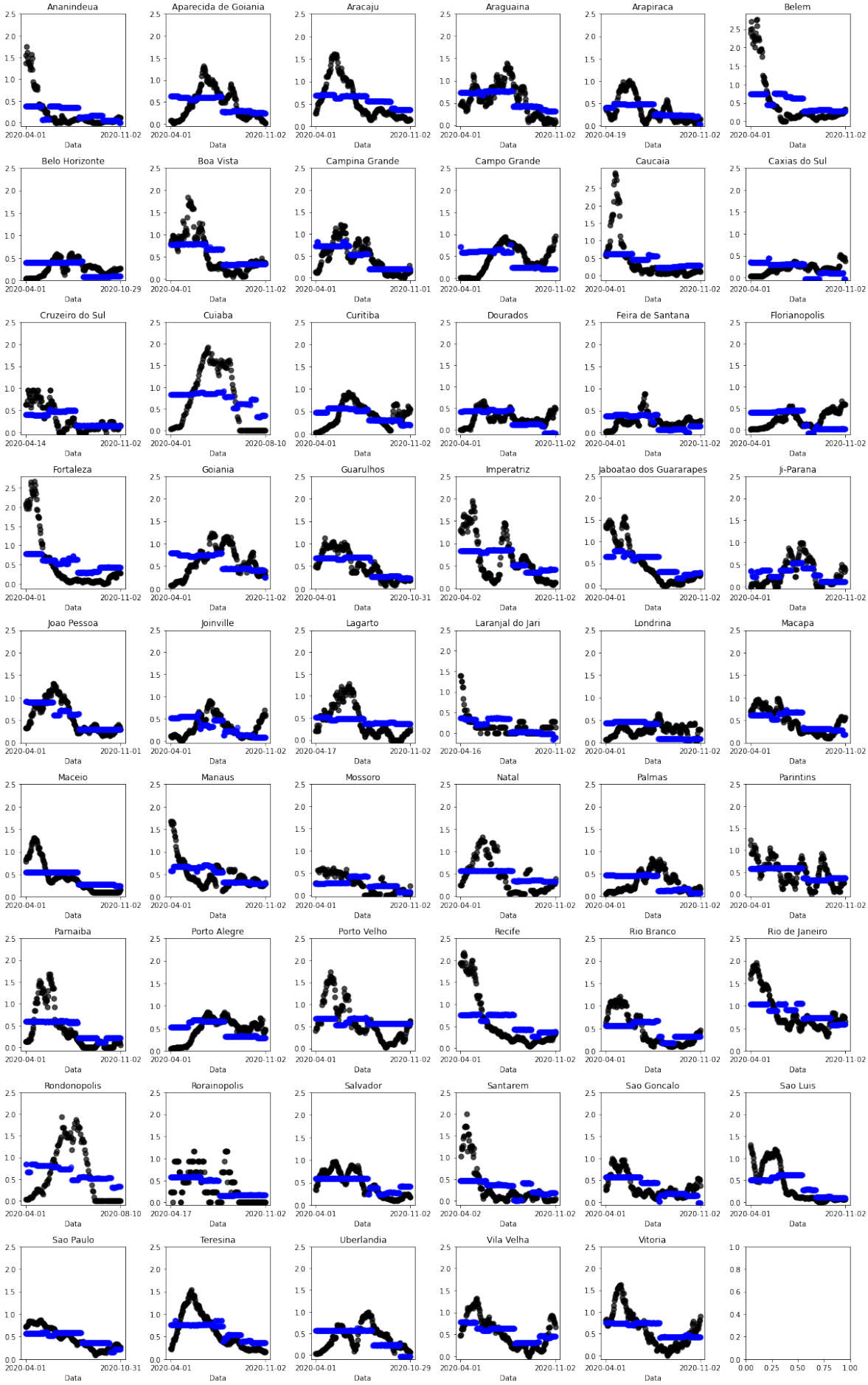

Figura 4 – Apresentação visual do resíduo por cidade, o eixo-y é o número ponderado de mortes (válido para os dados reais (preto) e para a projeção (azul)).