UNIVERSIDADE FEDERAL DE MINAS GERAIS
School of Engineering
Graduate Program in Electrical Engineering

Hélder Seixas Lima

EXPLORING LOCAL FACTORS ASSOCIATED WITH COVID-19 MORTALITY IN BRAZILIAN MUNICIPALITIES: a computational epidemiology approach for the first three years

Belo Horizonte
2025

Hélder Seixas Lima

EXPLORING LOCAL FACTORS ASSOCIATED WITH COVID-19 MORTALITY IN BRAZILIAN MUNICIPALITIES: a computational epidemiology approach for the first three years

Thesis presented to the Graduate Program in Electrical Engineering at the Universidade Federal de Minas Gerais, as a partial requirement for obtainning the title of Doctor in Electrical Engineering.

Advisor: Prof. Dr. Frederico Gadelha Guimarães

Belo Horizonte
2025

Lima, Hélder Seixas.
L732e Exploring local factors associated with covid-19 mortality in brazilian municipalities [recurso eletrônico] : a computational epidemiology approach for the first three years / Hélder Seixas Lima. - 2025.
1 recurso online (175 f. : il., color.) : pdf.

Orientador: Frederico Gadelha Guimarães.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Inclui bibliografia.

1. Engenharia elétrica - Teses. 2. Mineração de dados (Computação) - Teses. 3. COVID-19 Pandemia, 2020-. - Teses. 4. Previsão - Teses. 5. Epidemiologia - Modelos matemáticos - Teses. 6. Análise de regressão - Teses. I. Guimarães, Frederico Gadelha. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 621.3(043)

Ficha catalográfica elaborada pelo bibliotecário Reginaldo César Vital dos Santos CRB/6 2165
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

ESCOLA DE ENGENHARIA

COLEGIADO DO CURSO DE GRADUAÇÃO / PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FOLHA DE APROVAÇÃO

"Exploring Local Factors Associated With Covid-19 Mortality In Brazilian Municipalities: A Computational Epidemiology Approach For The First Three Years"

Hélder Seixas Lima

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Engenharia Elétrica.

Aprovada em 26 de fevereiro de 2025.

Por:

Prof. Dr. Frederico Gadelha Guimarães
DCC (UFMG) - Orientador

Prof. Dr. Americo Barbosa da Cunha Junior
IME (UERJ)

Prof. Dr. Rodrigo Weber dos Santos
DCC (UFJF)

Prof. Dr. Joicymara Santos Xavier
ICA (UFVJM)

Prof. Dr. Wagner Meira Junior
DCC (UFMG)

Documento assinado eletronicamente por **Frederico Gadelha Guimaraes**, **Professor do Magistério Superior**, em 13/03/2025, às 22:04, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Joicymara Santos Xavier**, **Usuário Externo**, em 17/03/2025, às 15:26, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rodrigo Weber dos Santos**, **Usuário Externo**, em 27/03/2025, às 10:14, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wagner Meira Junior**, **Professor do Magistério Superior**, em 27/03/2025, às 11:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Americo Barbosa da Cunha Junior**, **Usuário Externo**, em 28/03/2025, às 13:28, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3969349** e o código CRC **53716D3F**.

Referência: Processo nº 23072.209139/2025-35 SEI nº 3969349

To my beloved wife, Kelly, and my dear son, Samuel.

ACKNOWLEDGEMENTS

First and foremost, I thank God and my parents for the precious gift of life and their unwavering love and support throughout my journey. I am profoundly grateful to my wife, Kelly, and my son, Samuel, whose patience, understanding, and encouragement have been a constant source of strength and inspiration during the completion of this thesis.

I sincerely thank my adviser, Prof. Dr. Frederico Gadelha Guimarães, for his guidance, insightful comments, and unwavering support. His expertise and mentorship have been instrumental in shaping my research.

I thank the Machine Intelligence and Data Science (MINDS) laboratory and FutureLab for supporting my research. I am particularly grateful to their members for their companionship throughout this academic journey and for their valuable suggestions, which enriched my work.

My appreciation goes to Prof. Dr. Petrônio Cândido de Lima e Silva, Prof. Dr. Wagner Meira Jr., Prof. Dr. Unaí Tupinambás, and Prof. Dr. Marcelo Azevedo Costa for their collaboration in the production of papers. I also thank the anonymous reviewers of my publications for their constructive critiques, which significantly enhanced the quality of my work.

I am grateful to the faculty members and professors of the Graduate Program in Electrical Engineering (PPGEE) at the Universidade Federal de Minas Gerais (UFMG) for their dedication, enthusiasm, and willingness to share their knowledge, significantly contributing to my academic development.

I extend my genuine thanks to the members of my examining committee, Prof. Dr. Americo Barbosa da Cunha Junior, Prof. Dr. Rodrigo Weber dos Santos, Prof. Dr. Joicymara Santos Xavier, and Prof. Dr. Wagner Meira Junior, for graciously accepting the invitation to evaluate my thesis.

I would also like to acknowledge the developers of Python libraries (Statsmodels, Epyestim, Scikit-learn, SciPy, TensorFlow, Matplotlib, and Seaborn) for creating such fantastic tools that made the journey through this thesis much smoother.

I am deeply grateful to the Instituto Federal do Norte de Minas Gerais (IFNMG) for fostering my academic growth and supporting my research. Special thanks go to the Scholarships for Employee Qualification Program and the financial support provided by IFNMG, which enabled me to pursue and complete this work. I also wish to thank my colleagues at IFNMG for their encouragement and motivation throughout this journey.

Lastly, I acknowledge the vital role of Brazillian research funding agencies, Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council for Scientific and Technological Development (CNPq), and the Minas Gerais Research Funding Foundation (Fapemig), for their support to science in Brazil.

"Ai de nós, se por culpa nossa, semente morrer semente" (Unknown author).

RESUMO

O Brasil foi severamente impactado pela pandemia de COVID-19, com mais de 36 milhões de casos e quase 694 mil mortes reportadas até o final de 2022. Notavelmente, o Brasil apresenta significativa desigualdade socioeconômica entre suas regiões e municípios. Esta tese analisa a interação entre fatores locais e a mortalidade por COVID-19 em municípios brasileiros ao longo dos três primeiros anos da pandemia. Mais especificamente, investiga como as desigualdades nacionais, representadas por fatores demográficos, sociais, econômicos e políticos, se correlacionam com a mortalidade por COVID-19. Também avalia os efeitos do isolamento social, da vacinação e do surgimento de variantes na dinâmica da pandemia. Utilizando abordagens de epidemiologia computacional, como análise de regressão, técnicas de agrupamento, análise de correlação cruzada e modelagem epidemiológica, esta tese fornece percepções sobre os determinantes da mortalidade e sua evolução ao longo do tempo. Os resultados sugerem que a urbanização desempenha um papel significativo no aumento das mortes por COVID-19 nos municípios. No entanto, o impacto da urbanização variou ao longo do tempo, refletindo as medidas de saúde pública adotadas em cada momento. No início da pandemia, os municípios implementaram isolamento social preventivo, reduzindo o número básico de reprodução (R_0) e mitigando a mortalidade em municípios urbanizados. Uma mudança para o isolamento social reativo em 2021, associada à disseminação das variantes Gama e Delta, correlacionou-se com um aumento do R_0 e com uma mortalidade desproporcionalmente maior em municípios urbanizados. Os resultados também mostram que os esforços de vacinação se correlacionaram significativamente com a redução da letalidade e ajudaram a controlar os riscos de mortalidade relacionados à urbanização em 2022. Outras variáveis, como fatores ligados à pobreza, população idosa, povos indígenas e preferência política, também desempenharam papéis relevantes na dinâmica da pandemia nos municípios. Esta tese contribui com métodos e evidências para que autoridades de saúde possam analisar e monitorar epidemias, enfatizando a importância de medidas proativas para controle. Ela introduz um novo modelo epidemiológico com transições nebulosas para analisar epidemias de múltiplos surtos, validado por meio de sua aplicação a dados nacionais e municipais, além de sua aplicação na previsão de mortes por COVID-19 e seus resultados contrastados com evidências sorológicas. Por fim, os resultados ressaltam o valor de análises dinâmicas e sensíveis ao tempo, além de destacar o papel da desinformação e das influências políticas no agravamento das crises de saúde.

Palavras-chave: pandemia de COVID-19; epidemiologia computacional; técnicas de mineração de dados; análise de regressão; modelagem epidêmica; modelos de previsão.

ABSTRACT

Brazil was severely impacted by the COVID-19 pandemic, reporting over 36 million cases and nearly 694 thousand deaths by the end of 2022. Notably, Brazil has pronounced socioeconomic disparities across its regions and municipalities. This thesis analyzes the interplay between local factors and COVID-19 mortality in Brazilian municipalities across the first three pandemic years. More specifically, it investigates how national inequalities denoted by demographic, social, economic, and political factors correlate with COVID-19 mortality. It also evaluates the effects of social isolation, vaccination, and the emergence of variants in the pandemic dynamic. Using computational epidemiology approaches, such as regression analysis, clustering techniques, cross-correlation analysis, and epidemiological modeling, this thesis comprehensively provides insights into the mortality determinants and their temporal evolution. The findings suggest that urbanization plays a significant role in increasing COVID-19 deaths in the municipalities. However, the impact of urbanization varied over time, reflecting the public health measures employed at each moment. In the early pandemic, the municipalities implemented preventive social isolation, which reduced the basic reproduction number (R_0) and mitigated the mortality in urbanized municipalities. A shift to reactive social isolation in 2021, associated with the spread of Gamma and Delta variants, correlates with higher R_0 and disproportionately increased mortality in urbanized municipalities. Findings show vaccination efforts significantly correlated with lethality reduction and controlled urban-related mortality risks in 2022. Other variables, such as related poverty factors, elderly population, Indigenous people, and political preference, also explain the COVID-19 mortality in the municipalities. This thesis contributes methods and insights for health authorities to analyze and monitor epidemics, emphasizing the importance of proactive measures to control. It introduces a novel epidemiological model with fuzzy transitions to analyze multi-outbreak epidemics, demonstrating generalization through application to national and municipal data. It demonstrated robustness by forecasting COVID-19 deaths and validation against serological evidence. The findings underscore the value of dynamic and time-sensitive analysis and also highlight the role of misinformation and political influences in exacerbating health crises.

Keywords: COVID-19 pandemic; computational epidemiology; data mining techniques; regression analysis; epidemic modeling; forecasting models.

LIST OF FIGURES

Figure 1 – Time series of COVID-19 in Brazil on the National Monitoring Panel. . 25
Figure 2 – COVID-19 mortality rate across Brazilian municipalities (2020-2022). . 26
Figure 3 – Illustration of the Susceptible-Infected-Recovered-Dead-Susceptible (SIRDS) model. 32
Figure 4 – Temporal dynamics of COVID-19 in Brazil sourced from different datasets. 41
Figure 5 – Time series of cumulative COVID-19 death rates in Brazilian regions. . 43
Figure 6 – Maps of Brazilian municipalities illustrating COVID-19 death rates. . . 45
Figure 7 – Boxplot illustrating COVID-19 death rate for the 41 largest Brazilian municipalities across 2020-2022. 46
Figure 8 – COVID-19 death rate for the 41 largest Brazilian municipalities. 47
Figure 9 – Effective reproduction number (R_t) for COVID-19 in Brazil. 48
Figure 10 – Effective reproduction number (R_t) estimated for the 41 largest Brazilian municipalities. 49
Figure 11 – General Case Fatality Rate (CFR). 50
Figure 12 – Case Fatality Rate (CFR) calculated for Severe Acute Respiratory Syndrome (SARS) patients. 50
Figure 13 – *Stay-at-home index* (Δ_H) reported for the Brazilian population during the COVID-19 pandemic. 52
Figure 14 – *Stay-at-home index* (Δ_H) reported for the 41 largest Brazilian municipalities. 52
Figure 15 – Cumulative time series of COVID-19 vaccination in Brazil. 53
Figure 16 – Maps of COVID-19 vaccination coverage in Brazilian municipalities (2021-2022). 54
Figure 17 – Cumulative time series showing the percentage of people fully vaccinated against COVID-19 across the 41 largest Brazilian municipalities. 55
Figure 18 – Time series of COVID-19 Intensive Care Unit (ICU) beds for adults in Brazil throughout the pandemic. 55
Figure 19 – Monthly relative frequency (%) of the coronavirus variants over time in Brazil. 56
Figure 20 – Colormaps illustrating 31 sociodemographic attributes of Brazilian municipalities analyzed in this thesis. 60
Figure 21 – Map of Brazilian municipalities highlighting the percentage of votes for Bolsonaro in the first round of the 2022 Election. 61
Figure 22 – Map of Brazilian municipalities highlighting sociodemographic clusters. 62
Figure 23 – SIRDS model simulations across three years for different basic reproduction numbers (R_0). 78

Figure 24 – Comparative boxplots of effective reproduction number (R_t) similarity distributions in synthetic SIRDS outbreaks. 79
Figure 25 – Boxplots illustrating key metrics of the model optimization process for infection periods ranging from 8 to 20 days. 83
Figure 26 – Comprehensive analysis of simulation results for COVID-19 in Brazil. . 86
Figure 27 – Time-varying model parameters fitted for COVID-19 in Brazil. 86
Figure 28 – Comparison of cumulative COVID-19 infections simulated by our model, cumulative reported cases by health authorities, and serological prevalence. 88
Figure 29 – Sensitivity analysis heatmap for perturbations of 1%, 10%, and 50% in optimized parameters with COVID-19 data in Brazil. 90
Figure 30 – COVID-19 death rate for the 41 largest Brazilian municipalities. 95
Figure 31 – Illustration of death data used to fit and evaluate the performance of the Fuzzy SIRDS model across nine analysis windows. 96
Figure 32 – Illustration of death data used to train and evaluate the performance of the LSTM model across nine analysis windows. 97
Figure 33 – Structure of the Hybrid LSTM model, highlighting the input and output windows. 98
Figure 34 – COVID-19 death rate forecasts per 100,000 inhabitants across nine forecasting windows for the São Paulo/SP, Rio de Janeiro/RJ, Brasília/DF, and Fortaleza/CE municipalities. 101
Figure 35 – Heatmaps illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across the 41 largest Brazilian municipalities, based on the Fuzzy SIRDS model predictions over nine forecasting windows. 102
Figure 36 – Boxplots illustrating the Root Mean Squared Error (RMSE) for the COVID-19 death forecasting models in the medium term. 103
Figure 37 – Boxplots illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for the COVID-19 death forecasting models in the medium term. 103
Figure 38 – Boxplots illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for the COVID-19 death forecasting models in the short term. 104
Figure 39 – Bar plots illustrating the number of times each COVID-19 death forecast model achieved the lowest Symmetric Mean Absolute Percentage Error (SMAPE) for a municipality across nine forecasting windows. 104
Figure 40 – Comprehensive analysis of simulation results for Cuiabá/MT, the municipality with the highest COVID-19 mortality in our sample. 113
Figure 41 – Model parameters varying by time (t) estimated for COVID-19 in the 41 largest Brazilian municipalities from 2020 to 2022. 114

Figure 42 – Scatter plots for the coefficients and lag values resulting from cross-correlation analysis between $\Delta_H^{''}$ (stay-at-home index) and COVID-19 $R_0^{'}$ (time-varying basic reproductive number) estimated for the Brazilian municipalities. 115
Figure 43 – Validation metrics for the sociodemographic clustering with different numbers of clusters (k). 150
Figure 44 – Scatter plot of the sociodemographic clusters considering the principal component 1 and principal component 2. 150
Figure 45 – Time series of correlations between mortality rates and sociodemographic variables. 156
Figure 46 – Time series of correlations between mortality rates and sociodemographic variables across clusters. 157
Figure 47 – Time series of COVID-19 in Spain at the reported date by the health authorities. 158
Figure 48 – Time series of COVID-19 in the United Kingdom at the reported date by the health authorities. 159
Figure 49 – Time series of COVID-19 in the United States at the reported date by the health authorities. 160
Figure 50 – Time series of COVID-19 Case Fatality Rate (CFR) in Spain. 161
Figure 51 – Time series of COVID-19 Case Fatality Rate (CFR) in the United Kingdom. 161
Figure 52 – Time series of COVID-19 Case Fatality Rate (CFR) in the United States. 162
Figure 53 – Effective reproduction number (R_t) for COVID-19 in Spain. 163
Figure 54 – Effective reproduction number (R_t) for COVID-19 in the United Kingdom. 163
Figure 55 – Effective reproduction number (R_t) for COVID-19 in the United States. 164
Figure 56 – Fuzzy variables fitted for smoothing transitions between epidemic periods in Brazil. 165
Figure 57 – Fuzzy variables fitted for smoothing transitions between epidemic periods in Spain. 165
Figure 58 – Fuzzy variables fitted for smoothing transitions between epidemic periods in the United Kingdom. 166
Figure 59 – Fuzzy variables fitted for smoothing transitions between epidemic periods in the United States. 166
Figure 60 – Comprehensive analysis of simulation results for COVID-19 in Spain. . 167
Figure 61 – Comprehensive analysis of simulation results for COVID-19 in the United Kingdom. 168
Figure 62 – Comprehensive analysis of simulation results for COVID-19 in the United States. 169
Figure 63 – Time-varying model parameters fitted for COVID-19 in Spain. 170

Figure 64 – Time-varying model parameters fitted for COVID-19 in the United Kingdom. 170
Figure 65 – Time-varying model parameters fitted for COVID-19 in the United States. 170
Figure 66 – Heatmaps of the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across Brazilian municipalities, based on the LSTM model. 172
Figure 67 – Heatmaps of the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across Brazilian municipalities, based on the Hybrid LSTM model. 173
Figure 68 – Heatmaps of the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across Brazilian municipalities, based on the Hybrid SIRDS model. 174
Figure 69 – Heatmaps of the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across Brazilian municipalities, based on the Ensemble model. 175

LIST OF TABLES

Table 1 – Summary of COVID-19 data in Brazil (2020-2022). 42
Table 2 – List of the 41 largest Brazilian municipalities. 46
Table 3 – Descriptive statistics of COVID-19 vaccination metrics for Brazilian municipalities (2021-2022). 53
Table 4 – Goodness-of-fit statistics for the regression models across five analysis periods. 65
Table 5 – Relative gain of *Model 2* compared to the baseline. 66
Table 6 – Mean COVID-19 death rates for Brazilian municipalities grouped by sociodemographic clusters. 66
Table 7 – Estimated Rate Ratio (RR) of COVID-19 mortality for different clusters. 67
Table 8 – Rate Ratios (RR) of COVID-19 deaths by sociodemographic, vaccination, and political variables. 68
Table 9 – Model parameter bounds for optimization. 83
Table 10 – Results for COVID-19 simulation with data from Brazil, Spain, United Kingdom, and United States. 85
Table 11 – Results of the COVID-19 simulations for the 41 largest Brazilian municipalities. 112
Table 12 – Basic reproduction number (R_0) for each coronavirus variant during months when it was dominant in states from the 41 largest Brazilian municipalities. 116
Table 13 – Infection Fatality Rate (IFR) for different ranges of the fully vaccinated population against COVID-19 in the 41 largest Brazilian municipalities. 116
Table 14 – Descriptive statistics of variables for Brazilian municipalities. 149
Table 15 – Descriptive statistics of variables across the different sociodemographic clusters. 151
Table 16 – Estimated coefficients and goodness-of-fit statistics for *Model 1*. 152
Table 17 – Estimated coefficients and goodness-of-fit statistics for *Model 2*. 153
Table 18 – Comparison of goodness-of-fit statistics between *Model 2* using 30 bootstrap resamples and the reference dataset. 154
Table 19 – Comparison of the coefficients between *Model 2* with 30 bootstrap resamples and *Model 2* using the reference dataset. 155
Table 20 – Estimated coefficients and goodness-of-fit statistics for *Model 2*, calculated using the dataset excluding outliers and influential points. 155

LIST OF ALGORITHMS

1 SIRDS model with fuzzy epidemic period transitions. 81

LIST OF ABBREVIATIONS AND ACRONYMS

ABC Approximate Bayesian Computation

ADF Augmented Dickey Fuller

AIC Akaike Information Criterion

ARIMA Autoregressive Integrated Moving Average

BIC Bayesian Information Criterion

CEopt Cross-Entropy Optimization

CFR Case Fatality Rate

CI Confidence Intervals

DTW Dynamic Time Warping

GLM Generalized Linear Model

GMM Gaussian Mixture Model

ICD International Classification of Diseases

ICU Intensive Care Unit

IBGE Brazilian Institute of Geography and Statistics

IFR Infection Fatality Rate

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MCMC Markov Chain Monte Carlo

MSE Mean Squared Error

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

PSO Particle Swarm Optimization

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

RR Rate Ratio

SARS Severe Acute Respiratory Syndrome

SGD Stochastic Gradient Descent

SEIRS Susceptible, Exposed, Infected, Recovered, Susceptible

SEIHR Susceptible, Exposed, Infected, Hospitalized, Recovered

SIM Mortality Information System

SIR Susceptible, Infected, Recovered

SIRDS Susceptible, Infected, Recovered, Dead, Susceptible

SIS Susceptible, Infected, Susceptible

STF Supreme Federal Court

SMAPE Symmetric Mean Absolute Percentage Error

VOCs Variants of Concern

VIF Variance Inflation Factor

WHO World Health Organization

LIST OF SYMBOLS

β	*contact rate*
γ	*recovery rate*
f	*infection fatality probability*
ω	*immunity loss rate*
R_0	*basic reproduction number*
R_t	*effective reproduction number*
Δ_H	*stay-at-home index*
R^2	*coefficient of determination*
R^2_{CS}	*Cox & Snell pseudo-R^2*
R^2_{McF}	*McFadden pseudo-R^2*
χ^2	Pearson's chi-squared statistic

CONTENTS

1 INTRODUCTION 24
1.1 Research question 26
1.2 Objectives 27
1.3 Contributions 27
1.4 Thesis structure 28
2 EPIDEMIOLOGICAL CONCEPTS AND MODELING 30
2.1 Concepts and measures 30
2.2 Epidemiological modeling 31
2.2.1 Compartmental models 32
2.2.2 Data-driven models 35
2.2.3 Regression analysis models 36
3 THE COVID-19 IN BRAZIL 40
3.1 Data sources 40
3.1.1 Epidemiological data 40
3.1.2 Vaccination data 42
3.1.3 Variants data 42
3.1.4 Human mobility data 42
3.2 COVID-19 dynamics 43
3.2.1 Levels of analysis 43
3.2.1.1 National level 43
3.2.1.2 Regional level 43
3.2.1.3 Municipal level 44
3.2.1.4 Large municipalities level 44
3.2.2 Estimated reproduction number 47
3.2.3 COVID-19 outbreaks 49
3.2.4 Case Fatality Rate (CFR) 50
3.3 Public health measures 51
3.3.1 Reduction in human mobility patterns 51
3.3.2 COVID-19 vaccination campaign 53
3.3.3 Economic support and health infrastructure expansion 55
3.4 Prevalence of coronavirus variants 56
4 MULTIPLE REGRESSION ANALYSIS OF KEY FACTORS IN COVID-19 MORTALITY 57
4.1 Methodology 58
4.1.1 Study design 58

4.1.2 Data . . . 59
4.1.3 Clustering municipalities by sociodemographic variables . . . 59
4.1.3.1 Sociodemographic clusters . . . 62
4.1.4 Statistical analysis . . . 63
4.2 Results . . . 65
4.2.1 Associations between sociodemographic clusters and mortality . . . 66
4.2.2 Associations between sociodemographic, vaccination, and political variables and COVID-19 mortality . . . 68
4.2.3 Sensitivity analysis . . . 69
4.3 Discussion . . . 69
4.3.1 Related work . . . 73
4.3.2 Limitations . . . 73
5 A COVID-19 MODEL WITH FUZZY TRANSITIONS BETWEEN EPIDEMIC PERIODS . . . 75
5.1 Methodology . . . 76
5.1.1 Data . . . 76
5.1.2 SIRDS model with fuzzy epidemic period transitions . . . 76
5.1.2.1 Assumptions . . . 76
5.1.2.2 Model implementation . . . 79
5.1.2.3 Parameter optimization . . . 81
5.1.3 Experiments . . . 82
5.1.3.1 Fitting the recovery period . . . 82
5.1.3.2 Model application to other countries . . . 84
5.1.3.3 Parameter sensitivity assessment . . . 84
5.1.3.4 Estimating underreporting factor . . . 85
5.2 Results . . . 85
5.2.1 Retrospective analysis . . . 85
5.2.2 Model generalization . . . 87
5.2.3 Comparisons with serological research . . . 87
5.2.4 Sensitivity analysis . . . 89
5.3 Discussion . . . 89
5.3.1 Limitations . . . 93
6 MODELS FOR MEDIUM-TERM FORECASTING COVID-19 MORTALITY IN LARGE BRAZILIAN MUNICIPALITIES . . . 94
6.1 Methodology . . . 95
6.1.1 Data . . . 95
6.1.2 Forecasting horizon and analysis windows . . . 95
6.1.3 Forecasting models . . . 95
6.1.3.1 Fuzzy SIRDS model . . . 96

6.1.3.2 LSTM model . . . 97
6.1.3.3 Hybrid LSTM model . . . 98
6.1.3.4 Hybrid SIRDS model . . . 98
6.1.3.5 Ensemble model . . . 99
6.1.4 Evaluation of model performances . . . 99
6.2 Results . . . 100
6.3 Discussion . . . 105
6.3.1 Related work . . . 106
6.3.2 Limitations . . . 108
7 ANALYZING THE COVID-19 PARAMETERS FOR LARGE BRAZILIAN MUNICIPALITIES . . . 110
7.1 Methodology . . . 111
7.1.1 Data . . . 111
7.1.2 Simulations . . . 111
7.1.3 Data analysis . . . 111
7.2 Results and discussion . . . 112
7.2.1 Limitations . . . 117
8 CONCLUSIONS . . . 118
8.1 Limitations . . . 121
8.2 Future work . . . 121
8.3 Publications . . . 122

REFERENCES . . . 123

APPENDIX A Descriptive statistics of Brazilian municipalities . . . 148
APPENDIX B Sociodemographic clustering details . . . 150
APPENDIX C Detailed results of the regression models . . . 152
APPENDIX D Sensitivity analysis for regression model . . . 154
APPENDIX E Time series of correlations between mortality rates and sociodemographic variables . . . 156
APPENDIX F Epidemiological Time Series of COVID-19 for Spain, the United Kingdom, and the United States . . . 158
APPENDIX G Time series of COVID-19 Case Fatality Rate (CFR) for Spain, the United Kingdom, and the United States . . . 161
APPENDIX H Effective reproduction number for Spain, the United Kingdom, and the United States . . . 163
APPENDIX I Fuzzy variables fitted for smoothing transitions between epidemic periods . . . 165

APPENDIX J Comprehensive analysis of simulation results for COVID-19 in Spain, the United Kingdom, and the United States 167
APPENDIX K Time-varying model parameters fitted for COVID-19 in Spain, the United Kingdom, and the United States . . 170
APPENDIX L Heatmaps illustrating the model performance for COVID-19 death forecast across the 41 largest Brazilian municipalities 171

1 INTRODUCTION

The novel coronavirus outbreak began in Wuhan, China, in December 2019 (Huang et al., 2020) and, in the following months, spread to several countries worldwide. COVID-19 is an infectious disease caused by the SARS-CoV-2 virus (Velavan and Meyer, 2020). An infected person can spread the virus by small liquid particles when they cough, sneeze, speak, or breathe (WHO, 2020b). The contagion occurs when a susceptible person comes into contact with the virus through their eyes, nose, or mouth (WHO, 2020b). Most people will have mild effects of the disease, but older people and people with certain types of pre-existing conditions are more likely to have severe illness (Richardson et al., 2020).

The World Health Organization (WHO) declared the COVID-19 pandemic on 11 March 2020 (WHO, 2020d). According to the COVID-19 Data Explorer by the Our World Data project (Ritchie et al., 2020), nearly 729 million cases and 6.8 million deaths were registered worldwide until December 2022. Brazil notified nearly 36 million cases and 694 thousand deaths until December 2022, the second country worldwide with the most reported deaths, behind only the United States (Ritchie et al., 2020).

Many countries worldwide experienced recurrent waves of the COVID-19 pandemic (Hale et al., 2021). Figure 1 shows that Brazil faced four distinct waves by November 2022: a prolonged wave in 2020, another in 2021, and two smaller waves in 2022. Additionally, a new outbreak emerged in December 2022 but began to decrease by the end of the year. Scientists have studied the relationship between the rise of COVID-19 waves and factors such as coronavirus variants (Dutta, 2022; Thakur et al., 2021; El-Shabasy et al., 2022; Batistela et al., 2021), immunity loss (López and Rodó, 2020; Vinceti et al., 2021; Friston et al., 2020; Batistela et al., 2021), and the Peltzman effect (Iyengar et al., 2022; Juyal et al., 2021), which regards that people adjust their behavior based on perceived risk.

Beyond these factors, we examine how demographic, social, economic, political, vaccination, and population mobility variables correlate with COVID-19 mortality in Brazilian municipalities over the first three years of the pandemic. The fact is that Brazil experienced a flattened mortality peak of around 1,000 deaths per day in 2020. The second wave in 2021, however, was the deadliest, with a peak of 3,000 deaths per day. Although 2022 saw the highest number of COVID-19 cases, the Case Fatality Rate (CFR) dropped significantly compared to the second wave.

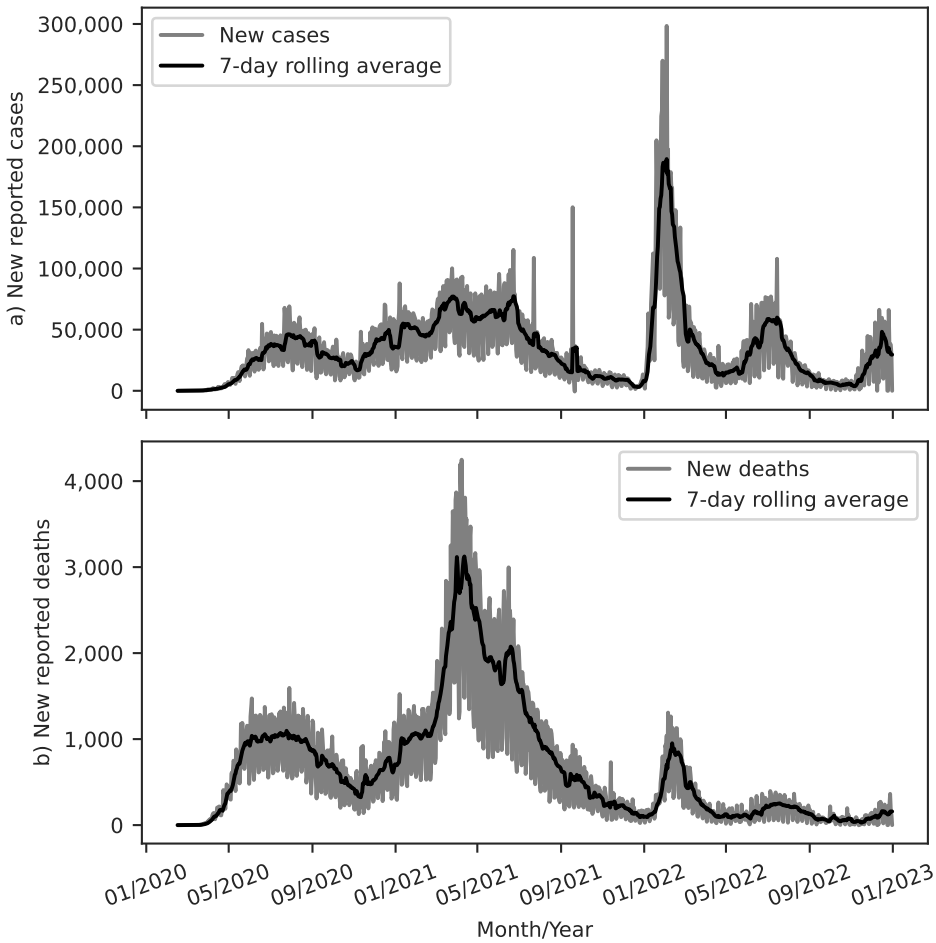
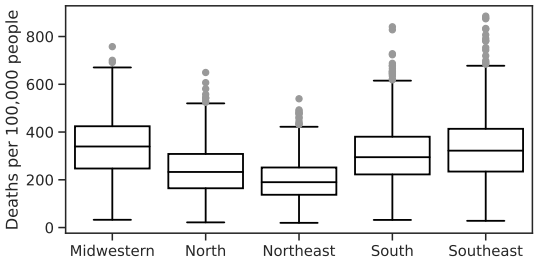


Figure 1 – Time series of COVID-19 in Brazil at the reported date on the National Monitoring Panel. (a) New reported cases and (b) new reported deaths.

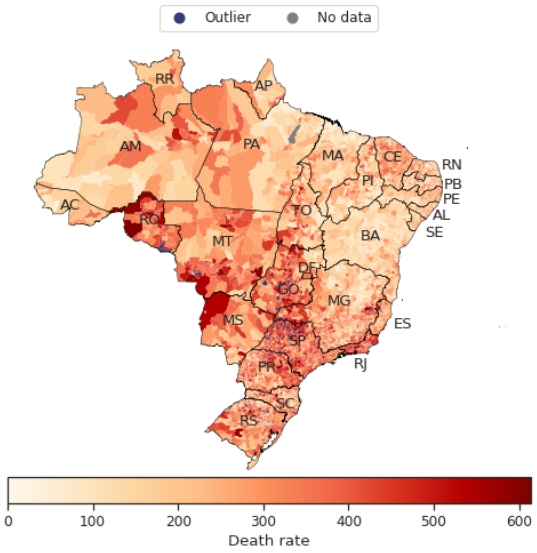
Source: COVID-19 National Monitoring Panel (DATASUS, 2020a)

We are motivated to investigate the COVID-19 pandemic at the municipal level in Brazil due to the pronounced socioeconomic disparities across its regions (Salata, 2020; Mourao and Junqueira, 2021; Cavalini and de Leon, 2008). This thesis aims to determine whether local characteristics may have contributed to the differences in COVID-19 mortality rates between municipalities. As shown in Figure 2, these rates varied widely, even within the same region or state, highlighting the variable impact of the pandemic throughout the country.

This thesis is a study in computational epidemiology, in which we apply data mining techniques, multiple regression analysis, and epidemiological modeling to analyze the COVID-19 pandemic in Brazilian municipalities. Our approach provides a comprehensive understanding of the factors related to the pandemic progression and mortality rates across different regions and moments.



(a) Boxplots of COVID-19 mortality rate per 100,000 inhabitants in Brazilian municipalities grouped by regions. The lower and upper bounds represent the first and third quartiles, respectively. The vertical line within the box indicates the median, while the whiskers extend to the minimum and maximum values within 1.5 times the interquartile range. The points represent the outlier municipalities.



(b) Map of Brazilian municipalities highlighting COVID-19 mortality rate per 100,000 inhabitants. Outliers represent municipalities with a death rate higher than the third quartile plus 1.5 times the interquartile range. The missing data pertain to 10 Brazilian municipalities excluded from this study. The maps segment the states of the Brazilian regions: North (AC: Acre, AM: Amazonas, AP: Amapá, PA: Pará, RO: Rondônia, RR: Roraima, TO: Tocantins), Northeast (AL: Alagoas, BA: Bahia, CE: Ceará, MA: Maranhão, PB: Paraíba, PE: Pernambuco, PI: Piauí, RN: Rio Grande do Norte, SE: Sergipe), Midwest (DF: Federal District, GO: Goiás, MS: Mato Grosso do Sul, MT: Mato Grosso), Southeast (ES: Espírito Santo, MG: Minas Gerais, RJ: Rio de Janeiro, SP: São Paulo), and South (PR: Paraná, RS: Rio Grande do Sul, SC: Santa Catarina).

Figure 2 – COVID-19 mortality rate per 100,000 inhabitants across Brazilian municipalities between 2020 and 2022.

Source: Mortality Information System (SIM) (DATASUS, 2022b)

1.1 Research question

What are the key factors correlating with COVID-19 mortality rates across Brazilian municipalities during the first three years of the pandemic, and how do these factors interact with epidemiological dynamics?

1.2 Objectives

The main objective of this thesis is to measure, analyze, and explain the factors correlating COVID-19 mortality across Brazilian municipalities during the different phases of the pandemic, spanning its first three years. To achieve this, we set the specific objectives:

- Develop and organize a comprehensive database covering the first three years of the COVID-19 pandemic in Brazilian municipalities.
- Analyze the relationships between demographic, social, economic, political, and vaccination coverage factors and the COVID-19 mortality rate in these municipalities.
- Estimate the parameters of an epidemiological model to describe COVID-19 mortality in Brazilian municipalities.
- Evaluate the performance of models in forecasting the COVID-19 mortality rate in the medium-term for these municipalities.
- Investigate the associations between estimated epidemiological parameters and factors such as social isolation measures, vaccination efforts, and the emergence of new variants in Brazilian municipalities.

1.3 Contributions

This thesis makes significant contributions to the fields of computational epidemiology, public health, and bioinformatics, which are outlined as follows:

- A detailed analysis of the COVID-19 pandemic in Brazil over three years reveals that, in the early stages, municipalities with lower human development indices experienced a higher correlation with COVID-19 mortality. The impact shifted toward municipalities with larger urban populations as the pandemic progressed.
- Investigates the relationship between political preferences and COVID-19 mortality in Brazilian municipalities throughout the pandemic revealed a positive correlation between the percentage of votes for Bolsonaro and mortality.
- Demonstration that assessing COVID-19 mortality risk in Brazilian municipalities requires a dynamic, time-sensitive approach due to the changing correlations between sociodemographic factors and mortality throughout the pandemic.
- Insights into the need for targeted interventions in more urbanized areas, municipalities with high proportions of elderly residents, and Indigenous populations to better mitigate the effects of future epidemics.

- Development of a modified epidemiological model for analyzing multi-outbreak epidemics, incorporating fuzzy transitions between epidemic phases to account for variations over time.
- Comparison of model simulations with official reports, uncovering significant disparities and revealing potential underreporting of COVID-19 cases in official data.
- Recommendations for the use of diverse models in epidemiological surveillance to improve the reliability and comprehensiveness of future forecasts.
- Identification of challenges associated with forecasting beyond a four-week horizon, underscoring the limitations of longer-term predictions.
- Correlation of model outcomes with data on social isolation, vaccination rates, and the emergence of new variants, providing a holistic understanding of factors shaping the pandemic trajectory.
- Implications for public health policies and interventions aimed at mitigating the effects of infectious disease outbreaks, contributing to more informed decision-making in future public health crises.

1.4 Thesis structure

We organized the remainder of this thesis as follows:

- **Chapter 2 - EPIDEMIOLOGICAL CONCEPTS AND MODELING:** This chapter introduces key epidemiological concepts and measures necessary for understanding this thesis. It covers epidemic modeling techniques, including mathematical, data-driven, and regression-based models. Additionally, it reviews related work that has utilized these models to analyze the COVID-19 pandemic.
- **Chapter 3 - THE COVID-19 IN BRAZIL:** This chapter provides an in-depth analysis of the COVID-19 dynamics in Brazil over the first three years of the pandemic. It highlights key epidemiological data, COVID-19 outbreaks, the effective reproduction number (R_t), CFR, public health measures, human mobility, vaccination efforts, and the emergence of coronavirus variants.
- **Chapter 4 - MULTIPLE REGRESSION ANALYSIS OF KEY FACTORS IN COVID-19 MORTALITY:** This chapter presents an ecological analysis at the municipal level to investigate sociodemographic clusters and variables associated with COVID-19 mortality from 2020 to 2022. Using the Gaussian Mixture Model (GMM), we clustered Brazilian municipalities into five sociodemographic groups and applied

negative binomial regression models to estimate correlations between these factors and COVID-19 mortality. The analysis reveals that municipalities with lower human development indices experienced higher mortality early in the pandemic, while later stages saw higher mortality in municipalities with higher levels of urbanization and elderly populations. Our findings show that the political preference of the municipalities plays a significant role in the regression analysis. We also discuss the temporal exposure and other factors correlating with COVID-19 mortality, such as service workers, Indigenous populations, and vaccination coverage.

- **Chapter 5 - A COVID-19 MODEL WITH FUZZY TRANSITIONS BETWEEN EPIDEMIC PERIODS:** This chapter provides a thorough analysis of the COVID-19 pandemic in Brazil, covering five distinct waves over three years. It introduces a novel SIRDS model with fuzzy transitions between epidemic periods, designed to estimate key epidemiological parameters, such as the basic reproduction number (R_0), underreporting factors, Infection Fatality Rate (IFR), and the immunity period. Using this model, we accurately assessed the extent of case underreporting and the pandemic dynamics. We validated our results through comparison with serological studies, and the model demonstrated its versatility by successfully simulating epidemic scenarios in other countries.

- **Chapter 6 - MODELS FOR MEDIUM-TERM FORECASTING COVID-19 MORTALITY IN LARGE BRAZILIAN MUNICIPALITIES:** This chapter offers a retrospective evaluation of COVID-19 mortality forecasting models across the 41 largest Brazilian municipalities. The performance of compartmental, data-driven, hybrid, and ensemble models is analyzed over nine forecasting windows. While our compartmental model provided the most stable results, the study underscores the value of using multiple models for reliable predictions.

- **Chapter 7 - ANALYZING THE COVID-19 PARAMETERS FOR LARGE BRAZILIAN MUNICIPALITIES:** This chapter analyzes the evolution of COVID-19 in Brazilian largest municipalities between 2020 and 2022. It uses the model with fuzzy transitions to estimate key epidemiological parameters and correlates them with social isolation, vaccination efforts, and the emergence of new variants. The chapter highlights the role of social isolation in reducing R_0 in 2020 and the impact of mass vaccination on lowering the IFR in 2022.

- **Chapter 8 - CONCLUSIONS:** This final chapter synthesizes the findings presented in the previous chapters, discussing the factors correlating with COVID-19 mortality in Brazilian municipalities. It also provides insights and directions for future research.

2 EPIDEMIOLOGICAL CONCEPTS AND MODELING

Epidemiology investigates the emergence and spread of health-related events in human populations, examines their causes, and evaluates the impact of proposed interventions to control them (Carr et al., 2007). The events studied in epidemiology may include disease, mortality, recovery, and the use of health services (The Open University, 2016).

There are two main categories of epidemiological studies: analytical and interventional. Analytical studies include ecological, cross-sectional, case-control, and cohort studies, while intervention studies refer to clinical and community trials (Carr et al., 2007).

This thesis uses the ecological study design, which only considers aggregated data. Ecological studies do not use individuals as units of analysis but data from a group of people, usually defined by geographic regions. These studies help formulate hypotheses but have limited robustness in checking causality (Bonita et al., 2006).

2.1 Concepts and measures

Below, we present concepts and measures of epidemiology used in this thesis.

- Prevalence: a metric used to determine the proportion or frequency of a health condition within a specific population at a particular point in time (Ceylan, 2020).
- Incidence: a metric denoting the frequency of new cases regarding a health condition that developed within a specific population during a particular period (Ceylan, 2020).
- Mortality rate: $M = \frac{d}{P} \times 100,000$, where M is the death rate per 100,000 inhabitants, d is the number of deaths during a specific period, and P is the population.
- Case rate: $C = \frac{c}{P} \times 100,000$, where C is the case rate per 100,000 inhabitants, c is the number of cases during specific period, and P is the population.
- Case Fatality Rate (CFR): $CFR = \frac{d}{c} \times 100$, where CFR is the disease lethality expressed as a percentage, d is the number of deaths during specific period, and c is the number of reported cases in same period (Mahase, 2020).

- Infection Fatality Rate (IFR): $IFR = \frac{d}{i} \times 100$, where IFR is a refined disease lethality expressed as a percentage, d is the number of deaths during a specific period, and i is the number of infections (reported cases or not) in same period (Mahase, 2020).
- Primary case: an individual who first brings a disease into a group of individuals (a school class, community, or country); this term is applied only when it is an infectious disease that spreads from human to human (Giesecke, 2014).
- Secondary case: "a person who gets a disease from exposure to a diseased person, or primary case, rather than the epidemic source itself" (El-Gilany, 2021).
- Index case: "the patient in an outbreak who is first noticed by the health authorities, and who makes them aware that an outbreak might be emerging" (Giesecke, 2014).
- Incubation period: corresponds to the period between infection and the onset of symptoms (Xiang et al., 2021).
- Infectious period: "the time interval during which the infected individuals could transmit the disease to any susceptible contacts" (Xiang et al., 2021).
- Latent period: corresponds to the time between infection and the start of the infectious period (Xiang et al., 2021).
- Protected period: the duration of time in which individuals who have recovered from a previous infection are immune to the same disease (Bjørnstad et al., 2020a).
- Serial interval: "the time from the onset of symptoms in the primary case to the onset of symptoms in secondary case" (Xiang et al., 2021).
- Generation time: "the time the onset of infectiousness in the primary case to the onset of infectiousness in the secondary case" (Xiang et al., 2021).

2.2 Epidemiological modeling

Epidemiological modeling comprises mathematical and computational techniques to deepen our understanding of epidemics and support policymakers' decision processes. Researchers have applied different methods to study COVID-19, including compartmental models, agent-based models, regression analysis, and data-driven approaches (Rahimi et al., 2021; Shankar et al., 2021). These models serve various purposes, such as estimating the actual magnitude of an epidemic, forecasting future trends, identifying correlated factors, evaluating the impact of public health interventions, assessing healthcare system demands, predicting potential outbreak hotspots, and guiding resource allocation to mitigate the spread of disease (Rahimi et al., 2021; Shankar et al., 2021).

The following subsections provide an overview of the modeling approaches relevant to this thesis: compartmental models (Section 2.2.1), data-driven models (Section 2.2.2), and regression analysis models (Section 2.2.3).

2.2.1 Compartmental models

Second Newman (2003), the Susceptible, Infected, Recovered (SIR) model was first formulated by Lowell Reed and Wade Hampton Frost in the 1920s. This is a simple compartmental model for modeling an epidemic, which divides the population into three compartments: *susceptible* (S), *infected* (I), and *recovered* (R). The dynamic of this model consists in the process where infected individuals can infect susceptible individuals, and after the infectious period, the infected individuals recover, which implies the individual getting immunity (Newman, 2003; Easley and Kleinberg, 2010; Bjørnstad et al., 2020a).

Many works have been utilized the compartmental models in the context of COVID-19 for a variety of purposes, including the estimation of epidemiological parameters (Bastos and Cajueiro, 2020; Davarci et al., 2023; Kamrujjaman et al., 2022; Massard et al., 2022), forecasting future trends (Tang et al., 2020; Martins et al., 2020; Davarci et al., 2023; Sarkar et al., 2020), and simulating potential scenarios (Abolpour et al., 2021; Volpatto et al., 2023; Shah et al., 2022; Jung et al., 2023; Gao et al., 2023; Ramos et al., 2021).

There are many specializations of SIR model, such as, the Susceptible, Infected, Susceptible (SIS) model (Newman, 2003), the Susceptible, Exposed, Infected, Recovered, Susceptible (SEIRS) model (Bjørnstad et al., 2020a), the Susceptible, Infected, Recovered, Dead, Susceptible (SIRDS) model (Gallos and Fefferman, 2015), and the Susceptible, Exposed, Infected, Hospitalized, Recovered (SEIHR) model (Choi and Ki, 2020).

Figure 3 shows the SIRDS model. This model is appropriate for modeling disease epidemics in that the immunity after infection is temporal, and a recovered individual becomes susceptible after some time. SIRDS model also adds the new compartment *dead* (D), which allows specifying the outcome of an infected individual as recovered or dead.

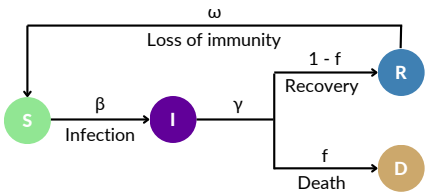


Figure 3 – Illustration of the Susceptible-Infected-Recovered-Dead-Susceptible (SIRDS) model. Each compartment is denoted by a corresponding letter: S for Susceptible, I for Infected, R for Recovered, and D for Deceased. The model parameters include contact rate (β), recovery rate (γ), infection fatality probability (f), and immunity loss rate (ω).

The SIRDS model, with S representing susceptible individuals, I representing infected individuals, R representing recovered individuals, and D representing deceased individuals, follows this dynamics:

- The *contact rate* (β) at each time unit (t) defines the rate at which the I infects the S.
- The I become either R or D at each t according to the *recovery rate* (γ) and the *infection fatality probability* (f).
- The R become S at each t due to the *immunity loss rate* (ω).

The reproduction number (R_0) is a key parameter in epidemic models, representing the average number of new infections generated by each infected individual introduced into a population with no prior immunity (Bjørnstad et al., 2020a). The others SIRDS parameters γ, β, f, and ω are specified in the equations below:

$$\gamma = \frac{1}{\textit{infectious period}}, \tag{2.1}$$

$$\beta = \frac{R_0}{\textit{infectious period}} = \gamma R_0, \tag{2.2}$$

$$f = \frac{IFR}{100}, \tag{2.3}$$

$$\omega = \frac{1}{\textit{protected period}}. \tag{2.4}$$

Being $N = S(0) + I(0) + R(0)$, the compartments change over time following the four equations presented below (Bjørnstad et al., 2020b; Bastos and Cajueiro, 2020):

$$\frac{dS}{dt} = -\frac{\beta IS}{N} + \omega R, \tag{2.5}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \tag{2.6}$$

$$\frac{dR}{dt} = (1 - f)\gamma I - \omega R, \tag{2.7}$$

$$\frac{dD}{dt} = f\gamma I. \tag{2.8}$$

The *effective reproduction number* (R_t) is a crucial metric used to track the progression of an epidemic, calculated by (Nishiura and Chowell, 2009; Gostic et al., 2020; Cintrón-Arias et al., 2008):

$$R_t = R_0 \frac{S}{N}. \tag{2.9}$$

R_t measures whether the number of new infections is increasing ($R_t > 1$) or decreasing ($R_t < 1$) in a population (Arroyo-Marioli et al., 2021). In essence, the R_t indicates the epidemic trend, whether it is growing or declining, or if it is transitioning into an endemic state, where the R_t remains stable around one over time (Choi and Ki, 2020).

We can also estimate the R_0 from the R_t and S, as:

$$R_0 = R_t \frac{N}{S}. \tag{2.10}$$

After defining the compartmental model, the modeler must estimate the relevant epidemiological parameters. According to Shankar et al. (2021), parameter estimation is a significant challenge for simulating infectious disease outbreaks using compartmental models. When performing these estimations, it is crucial to regard scientific evidence on factors such as transmission rates, infectious periods, lethality, and the duration of immunity. Using appropriate methods to estimate these parameters ensures that simulations are robust and reliable.

There are several methods available for estimating epidemiological parameters in compartmental models. Generally, these methods involve fitting the model to available data and minimizing the deviance between observed and simulated epidemic trajectories (Keeling and Rohani, 2008). In the context of COVID-19, various approaches have been employed, including Markov Chain Monte Carlo (MCMC) (Suchoski et al., 2022; Acuña-Zegarra et al., 2020; Mbuvha and Marwala, 2020), least squares fitting (Cantó et al., 2017; Mpinganzima et al., 2023; Batistela et al., 2021), gradient-based optimization (Fan et al., 2021; Millevoi et al., 2024), genetic algorithms (Fanelli and Piazza, 2020; Xavier et al., 2022a; Reis et al., 2021; Pinto Neto et al., 2021), Particle Swarm Optimization (PSO) (Putra et al., 2019; Massard et al., 2022), curve fitting (Pereira et al., 2020), and Approximate Bayesian Computation (ABC) combined with Monte Carlo simulations (Asher et al., 2023; Cunha Jr et al., 2023; Ritto et al., 2021).

In summary, compartmental models are deterministic frameworks that simulate epidemics by splitting the population into distinct groups based on disease status. These models rely on parameter estimation to align simulations with observed data, often involving stochastic optimization techniques, such as MCMC or genetic algorithms. These methods introduce an element of uncertainty, enabling the calculation of Confidence Intervals (CI) for the parameter estimates.

2.2.2 Data-driven models

While compartmental models use the optimization of ordinary differential equations to simulate the epidemic dynamic, data-driven models focus on predicting patterns directly from the available data, often without explicitly modeling underlying epidemic processes. These models encompass approaches such as autoregressive models, machine learning algorithms, and deep learning frameworks. While data-driven epidemiological models excel in short-term forecasting, they tend to lack reliability for long-term predictions due to their limited ability to capture epidemic dynamics (Shankar et al., 2021). Moreover, they are less effective during the early stages of an outbreak when historical data is sparse.

In the context of COVID-19, many studies utilized time series models like Autoregressive Integrated Moving Average (ARIMA) to forecast confirmed cases. For example, Ribeiro et al. (2020) applied ARIMA to predict cases over a six-day horizon, Hernandez-Matamoros et al. (2020) used it for a 15-day horizon, and Alzahrani et al. (2020) extended the prediction horizon to four weeks. ARIMA is a statistical approach that captures temporal dependencies in the data by optimizing three key parameters: autoregressive terms (p), differencing order (d), and moving average terms (q).

Machine learning models, on the other hand, depend on a training process to discover complex patterns within the data. These models offer flexibility but present challenges such as hyperparameter tuning, particularly critical for artificial neural networks. In the COVID-19 literature, machine learning models have been widely applied, including random forest (Watson et al., 2021; Dansana et al., 2022; Ribeiro et al., 2020), gradient boosting (Rahman and Chowdhury, 2022; Shrivastav and Jha, 2021), and Long Short-Term Memory (LSTM) architectures (Arora et al., 2020; Chimmula and Zhang, 2020; Devaraj et al., 2021).

In Chapter 7, we implemented a LSTM model to forecast COVID-19 mortality in Brazilian municipalities. LSTM is a type of Recurrent Neural Network (RNN), specifically designed for time series prediction tasks, and is equipped with extended long-term memory capabilities (Hochreiter and Schmidhuber, 1997). The essential components of an LSTM architecture include the input layer, where each input feature corresponds to a neuron; the output layer, where neurons represent the predicted values of the time series; and one or more hidden layers, each consisting of multiple neurons (Hochreiter and Schmidhuber,

1997). Within these hidden layers, the LSTM algorithm manages short-term and long-term memory using mechanisms such as forget gates, input gates, and output gates (Zaki and Meira Jr, 2020). These gates operate with weights and activation functions optimized during training (Zaki and Meira Jr, 2020). This design helps mitigate vanishing and exploding gradients and facilitates recurrent feedback loops, enabling effective learning of temporal dependencies (Hochreiter and Schmidhuber, 1997).

Training an LSTM model consists of preparing a time series dataset and structuring it into sequences compatible with the model input layer. During training, the LSTM processes input data through its layers, using mechanisms like forget, input, and output gates to manage internal states and capture temporal dependencies (Zaki and Meira Jr, 2020). A loss function, such as Mean Squared Error (MSE), computes the error between predicted and actual values. Parameters are then optimized using backpropagation through time and gradient-based methods, such as Stochastic Gradient Descent (SGD) (Bottou, 2010) or Adam (Kingma and Ba, 2017). This process is repeated over multiple epochs, adjusting weights and biases to minimize the loss (Zaki and Meira Jr, 2020). Validation data is used to fine-tune hyperparameters and prevent overfitting, with testing data employed to evaluate the model generalization.

2.2.3 Regression analysis models

Regression techniques are fundamental statistical methods used to model and understand the relationship between a dependent variable (the response variable) and one or more explanatory variables (independent or predictor variables, covariates, or risk factors) (Bender, 2009). These techniques help quantify how changes in the explanatory variables influence the response variable, making them essential for predictive modeling and inference in various fields (Bender, 2009). When the model includes only one explanatory variable, it is called simple regression. Including multiple explanatory variables, by contrast, results in multiple or multifactorial regression (Bender, 2009).

Multiple regression models are widely used in epidemiology to explore the impact of various risk factors on outcomes like mortality or disease incidence (Bender, 2009). These models allow researchers to estimate adjusted effects, controlling for potential confounders and delivering more accurate and unbiased estimates of relationships (Bender, 2009). Depending on the nature of the outcome variable, different regression models are chosen, including linear regression for continuous outcomes, logistic regression for binary outcomes, Cox regression for time-to-event data, and Poisson regression for count data (Bender, 2009).

In the context of COVID-19, the works commonly apply the Poisson regression, particularly for modeling the number of cases or deaths (Hastenreiter Filho and Cavalcante, 2022; López-Bazo, 2024; Unruh et al., 2022; Lorenz et al., 2021; Morrissey et al., 2021;

Hoebel et al., 2021; Bermudi et al., 2021; Clouston et al., 2021b; Das et al., 2020; Fantin et al., 2023; Yoshikawa and Kawachi, 2021). This model is part of the Generalized Linear Model (GLM) framework, where the response variable Y is linked to its mean μ through the logarithmic link function (Bender, 2009). It assumes a linear relationship between the explanatory variables, $X_1, \ldots, X_k$, and the logarithm of the expected value of Y (Bender, 2009). Specifically, in Poisson regression, Y is assumed to follow a Poisson distribution, where the mean and variance are equal, i.e., $\mu = \text{var}(Y)$. The general form of the Poisson GLM is given by:

$$\begin{cases} Y \sim \text{Poisson}(\mu), \\ \ln(\mu) = a_0 + a_1 x_1 + \cdots + a_k x_k, \end{cases} \tag{2.11}$$

where a_0 is the intercept, a_j are the regression coefficients, and x_j are the values of the explanatory variables (Bender, 2009; Dunn et al., 2018). The model assumes that a one-unit increase in X_j results in a multiplicative change of $\exp(a_j)$ in the expected frequency μ, holding other variables constant (Bender, 2009; Dunn et al., 2018).

However, researchers have noted that COVID-19 data often exhibit overdispersion, where the variance exceeds the mean. They have used the negative binomial regression model to account for this (Hastenreiter Filho and Cavalcante, 2022; Johnson and Owusu, 2024; López-Bazo, 2024; Unruh et al., 2022; Yoo et al., 2022; Das et al., 2020; De Angelis et al., 2021). The negative binomial model provides additional flexibility to account for overdispersed data (Dunn et al., 2018). This model can be formulated as a Poisson-Gamma mixture model, where:

$$Y|\lambda \sim \text{Poisson}(\lambda) \text{ and } \lambda \sim \text{Gamma}(\mu, \psi), \tag{2.12}$$

with λ following a Gamma distribution with mean μ and coefficient of variation ψ. The parameter ψ represents the dispersion, related to the Gamma distribution shape parameter α as $\psi = \text{Var}(\lambda)/E(\lambda)^2 = 1/\alpha$. Consequently, the variance of Y becomes $\text{Var}(Y) = \mu + \mu^2/\alpha$, where the term μ^2/α accounts for the observed overdispersion (Dunn et al., 2018).

Collinearity among explanatory variables is a significant challenge in interpreting regression models (Dunn et al., 2018). Collinearity occurs when some covariates are highly correlated, potentially destabilizing the regression coefficients (Dunn et al., 2018; Dobson and Barnett, 2018). We can detect collinearity for each explanatory variable X_j by calculating the Variance Inflation Factor (VIF):

$$\text{VIF}_{X_j} = \frac{1}{1 - R_j^2}, \tag{2.13}$$

where R_j^2 is the coefficient of determination got from regressing X_j against all other variables. A VIF of 1 indicates no collinearity. Montgomery et al. (2021) suggest that VIFs exceeding 5 indicate poorly estimated regression coefficients. Additionally, Dunn et al. (2018) note that pairwise correlations greater than 0.7 are a collinearity concern. Methods to address collinearity include omitting explanatory variables, combining explanatory variables, collecting more data, and using ridge regression (Montgomery et al., 2021; Dunn et al., 2018).

Once the model specification is defined, including the link function, explanatory variables, and the response distribution, the next step is the parameter estimation (Dobson and Barnett, 2018). Maximum likelihood estimation and least squares are parameter estimation methods widely used for GLMs (Dobson and Barnett, 2018). Statistical software, such as *R* and *statsmodels* (Python library), provides convenient tools for estimating these parameters without requiring manual calculations.

Fitting a model is an iterative process, encompassing defining the model, estimating its parameters, and performing diagnostics (Bender, 2009; Dunn et al., 2018; Dobson and Barnett, 2018). Crucial diagnostics include evaluating the normality of residuals and assessing goodness-of-fit using various statistics (Dunn et al., 2018; Dobson and Barnett, 2018).

Deviance (D) is a key statistic for evaluating the goodness-of-fit in regression models. It compares the log-likelihood of a saturated model, one that perfectly fits the data, with the log-likelihood of the fitted model (Hilbe, 2011; Dunn et al., 2018). For a Poisson-Gamma mixture model, the deviance, $D(Y, \hat{\mu})$, is calculated as (Hilbe, 2011; Dunn et al., 2018):

$$D(Y, \hat{\mu}) = 2 \sum_{i=1}^{n} \left(y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i + \alpha) \ln \left(\frac{y_i + \alpha}{\hat{\mu}_i + \alpha} \right) \right), \tag{2.14}$$

where $\hat{\mu}$ are the outcomes estimated by the model.

Another widely used statistic is Pearson's chi-squared (χ^2), which evaluates the sum of squared standardized residuals. For a Poisson-Gamma mixture model, this statistic is given by (Hilbe, 2011):

$$\chi^2(Y, \hat{\mu}) = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i + \frac{\hat{\mu}_i{}^2}{\alpha}}. \tag{2.15}$$

In addition to D and χ^2, pseudo-R^2 metrics, such as Cox & Snell (R^2_{CS}) and McFadden (R^2_{McF}), offer insight into the model fit relative to the null model, a model with only an intercept and offsets (Smith and McKenna, 2013). These metrics are valuable for comparing model performance and evaluating improvements in predictive accuracy (Smith and McKenna, 2013). Unlike the traditional R^2 in linear regression, which measures the proportion of variance explained, pseudo-R^2 values in GLMs reflect relative model fit. R^2_{CS} is calculated as:

$$R^2_{\text{CS}} = 1 - \exp\left(\frac{\text{LL}_{\text{null}} - \text{LL}_{\text{model}}}{n}\right), \tag{2.16}$$

where LL_{null} is the log-likelihood of the null model, LL_{model} is the log-likelihood of the fitted model, and n is the number of observations (Smith and McKenna, 2013).

R^2_{McF}, derived from the log-likelihood ratio between the fitted and null models, is given by (Smith and McKenna, 2013):

$$R^2_{\text{McF}} = 1 - \frac{\text{LL}_{\text{model}}}{\text{LL}_{\text{null}}}. \tag{2.17}$$

For a more comprehensive evaluation, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) provide a balanced measure of model fit and complexity, being especially useful for non-nested models (Dobson and Barnett, 2018; Dunn et al., 2018). The AIC is calculated as:

$$\text{AIC} = 2k - 2\text{LL}_{\text{model}}, \tag{2.18}$$

where k represents the number of estimated parameters in the model (Dunn et al., 2018). Similarly, the BIC is calculated as (Dunn et al., 2018):

$$\text{BIC} = \ln(n)k - 2\text{LL}_{\text{model}}. \tag{2.19}$$

As emphasized by Dunn et al. (2018), diagnosing potential outliers is essential when fitting a GLM, as they can significantly influence model estimates. One method for detecting outliers is by assessing Pearson residuals (r_P), which for a negative binomial model is defined as:

$$r_P = \frac{y - \hat{\mu}}{\sqrt{\hat{\mu} + \alpha\hat{\mu}^2}}. \tag{2.20}$$

In addition to residual diagnostics, we should evaluate influential observations that can substantially alter the fitted model. Dunn et al. (2018) recommend using Cook's distance to diagnose such observations. Cook's distance is given by:

$$C_D = \frac{(r'_P)^2}{p}\frac{h}{1-h}, \tag{2.21}$$

where r'_P is the standardized Pearson residual, p is the number of predictors in the model (including the intercept), and h is the leverage of the observation. Thus, C_D helps identify observations that combine large residuals with high leverage (Dunn et al., 2018).

3 THE COVID-19 IN BRAZIL

In this chapter, we present essential data for understanding the COVID-19 pandemic in Brazil. Section 3.1 introduces the primary data sources related to COVID-19, while Section 3.2 explores the dynamics of the pandemic at different granularity levels. Sections 3.3 and 3.4 provide overviews of public health measures and significant coronavirus variants, respectively.

3.1 Data sources

This thesis utilizes a range of databases relevant to COVID-19, including epidemiological data, vaccination records, variant tracking, population mobility, and health infrastructure. To support comparative analysis across different locations, we used population data from the 2022 Demographic Census (IBGE, 2022), provided by the Brazilian Institute of Geography and Statistics (IBGE).

3.1.1 Epidemiological data

The initial source of COVID-19 data in Brazil was the Monitoring Panel (DATASUS, 2020a), which provided timely updates throughout the pandemic, incorporating daily information on new cases and deaths. However, it is essential to acknowledge that this data reflects the date of reporting by health authorities rather than the onset of symptoms or death. This characteristic has resulted in delayed information and time series marked by artificial weekly seasonality, as depicted in Figure 1.

To address these limitations, we utilized the Mortality Information System (SIM) (DATASUS, 2022b) database from the Brazilian Health Ministry, which offers comprehensive details on all reported deaths in the country. Focusing on COVID-19 deaths, we filtered the dataset using the International Classification of Diseases (ICD) code B34.2 as the primary cause. Unlike the Monitoring Panel, this database reports the actual date of death occurrence, leading to a time series free of artificial seasonality, as illustrated in Figure 4b.

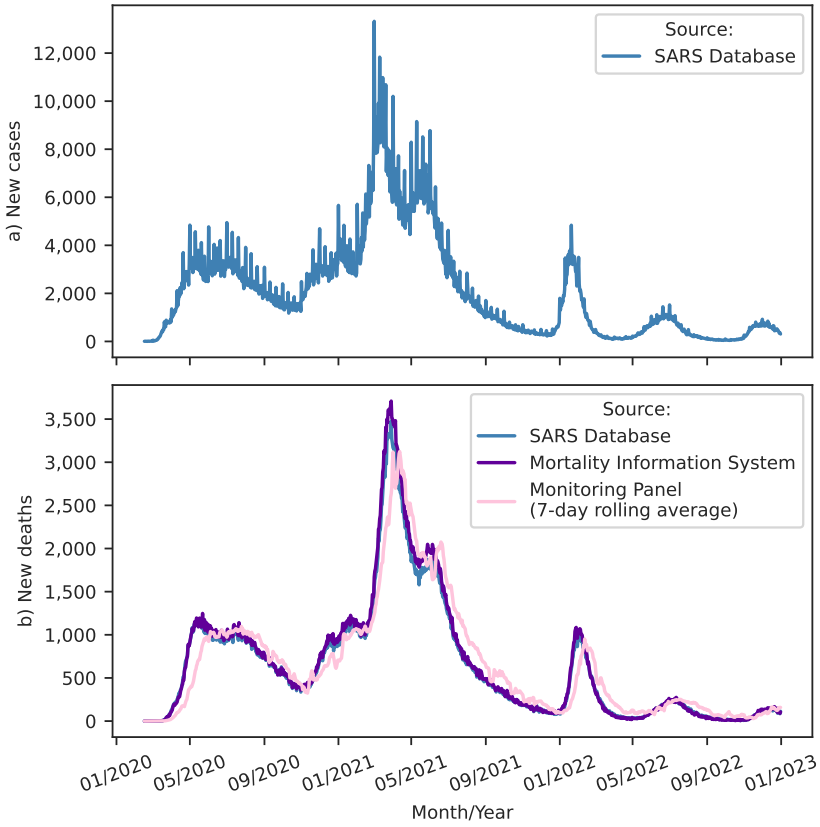


Figure 4 – Temporal dynamics of COVID-19 in Brazil sourced from different datasets. (a) New COVID-19 cases reported by onset symptoms date of Severe Acute Respiratory Syndrome (SARS) patients. (b) New deaths of SARS patients reported by death date, new COVID-19 deaths reported in Mortality Information System (SIM) by death date, and a 7-day rolling average of new COVID-19 deaths reported in Monitoring Panel by reporting date.

Source: SARS database (DATASUS, 2022a), SIM (DATASUS, 2022b), and COVID-19 Monitoring Panel (DATASUS, 2020a).

Furthermore, the Brazilian Health Ministry provides the Severe Acute Respiratory Syndrome (SARS) database (DATASUS, 2022a), containing severe cases of COVID-19. In Figure 4b, we present COVID-19 deaths reported for SARS patients, with this dataset recording the event date and avoiding artificial seasonality in death data. Regarding severe COVID-19 cases, Figure 4a shows that only a tiny fraction of COVID-19 cases escalated in severity compared to the reported cases in Figure 1a. The SARS database reports the onset date of symptoms, a crucial contribution to analyzing disease spread dynamics. Lastly, we observe a weekly seasonality in the onset date of symptoms, although this is lower than the seasonality observed for cases reported in the Monitoring Panel (DATASUS, 2020a).

This thesis analyzes the COVID-19 pandemic in Brazil across the first three years, from 2020 to 2022. Table 1 summarizes the data obtained from the investigated databases.

Table 1 – Summary of COVID-19 data in Brazil (2020-2022).

Database	Data	2020	2021	2022	Total
Monitoring Panel (DATASUS, 2020a)	Cases	7,675,973	14,611,548	14,043,760	36,331,281
Severe Acute Respiratory Syndrome (SARS) (DATASUS, 2022a)	Cases	708,141	1,219,288	238,679	2,166,108
Monitoring Panel (DATASUS, 2020a)	Deaths	194,949	424,107	74,797	693,853
Severe Acute Respiratory Syndrome (SARS) (DATASUS, 2022a)	Deaths	206,468	397,498	63,952	667,918
Mortality Information System (SIM) (DATASUS, 2022b)	Deaths	212,704	424,461	65,392	702,557

3.1.2 Vaccination data

We used vaccination data from the Brazilian Health Ministry (Brasil, 2022). In Brazil, up to 2022, the national health service administered COVID-19 vaccines produced by AstraZeneca, Janssen, Pfizer/BioNTech, and Sinovac. The standard vaccination protocol required two doses for all vaccines except for the Janssen vaccine, which required only a single dose (WHO, 2023). Consequently, we consider an individual fully vaccinated after receiving two doses of AstraZeneca, Pfizer/BioNTech, or Sinovac vaccines or a single dose of the Janssen vaccine.

3.1.3 Variants data

We collected state-level data on COVID-19 genomic surveillance in Brazil reported monthly by the Fundação Oswaldo Cruz (Fiocruz, 2020). We focused on identifying the Variants of Concern (VOCs) with significant prevalence in the country. These include Gamma, Delta, and Omicron, with specific attention to the Omicron sublineages BA.1, BA.2, BA.4, and BA.5. We grouped other lineages with low prevalence in Brazil, providing a comprehensive overview of the genomic landscape in the country.

3.1.4 Human mobility data

We utilized data from the COVID-19 Community Mobility Report (Google, 2020) produced by Google to monitor human mobility during the COVID-19 pandemic. This report provided a measure of the percent change in time spent in residential places by Google users at the municipal level, which we referred to as *stay-at-home index* (Δ_H). To calculate this measure, Google compared the time spent in residential places on a specific date to a baseline period of January 3, 2020, to February 6, 2020. For all days of the week, Google defines the baseline as the median time spent in residential places on that day during the baseline period. The data were aggregated at the national, state, and municipal levels, covering the period from February 15, 2020, to October 15, 2022 (Google, 2020).

3.2 COVID-19 dynamics

3.2.1 Levels of analysis

A comprehensive understanding of COVID-19 progression requires examining the pandemic from multiple perspectives. While national-level analysis provides a valuable overview, detailed insights emerge from examining data in increased granularity, such as at the municipal level. In this thesis, we explore COVID-19 dynamics across various levels, presented in the following subsections.

3.2.1.1 National level

Figure 1 shows the national progression of COVID-19 in Brazil, as discussed in Chapter 1. In summary, Brazil experienced five distinct waves from 2020 to 2022. The highest peak in mortality occurred in 2021, while 2022 saw the highest number of cases, although with a lower CFR. We provide an epidemic simulation at the national level in Chapter 5.

3.2.1.2 Regional level

At the onset of the COVID-19 pandemic in Brazil, the North region experienced the highest mortality rates, as shown in Figure 5. This trend continued in early 2021, when a severe health crisis in Manaus, the largest city in that region, led to a sharp increase in deaths (Barreto et al., 2021; Lalwani et al., 2021). During the second wave, however, the Southeast, Midwestern, and South regions emerged as the most impacted, while the North and Northeast regions saw comparatively lower mortality rates since that time.

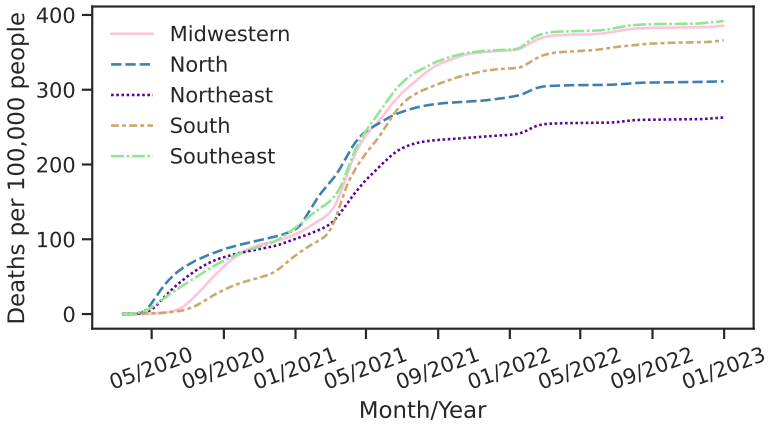


Figure 5 – Cumulative COVID-19 deaths per 100,000 people in Brazilian regions.

Source: Mortality Information System (SIM) (DATASUS, 2022b).

While we did not conduct experiments specifically at the regional level, this level provided valuable context for analyzing and interpreting the findings presented in this thesis.

3.2.1.3 Municipal level

Brazil comprises 5,569 municipalities and one Federal District, which we treat as a municipality in this thesis. We excluded five municipalities founded after 2010 due to missing data. Additionally, we removed five municipalities that experienced a reduction in area and population following the establishment of new municipalities. Consequently, this thesis includes data from 5,560 Brazilian municipalities.

Figure 6 illustrates the temporal and municipal progression of COVID-19 mortality across Brazil over three years. Initially, high mortality rates were observed in the North region, coastal Northeast, and Rio de Janeiro state. This map aligns with the trends in Figure 5, highlighting at the municipal level that the Midwest, Southeast, and South regions recorded the highest mortality rates in the later stages of the studied period.

The 2022 year, Figure 6d, was the year of the pandemic, in the studied period, when fewer municipalities had an elevated mortality rate. This year, the Northeast and North regions were again the least impacted. However, we note a reduction in impact difference between those regions and the Midwestern, Southeast, and South regions. The cumulative of the first three COVID-19 years, Figure 6e, resembles the pattern observed in 2021.

In Chapter 4, we conducted a municipal-level study analyzing the correlation between local factors and COVID-19 mortality throughout the first three pandemic years.

3.2.1.4 Large municipalities level

We selected a sample of large municipalities for a more in-depth analysis at a municipal level. To conduct successful epidemic simulation at a municipal level, we must ensure that the epidemiological time series of these locations has a significant frequency of events. Additionally, a reduced sample is necessarily due to the computational cost of fitting epidemiological parameters into mathematical models. So this sampling is crucial for the experiments detailed in Chapters 6 and 7. Consequently, our analysis concentrates on municipalities with populations exceeding 500,000, representing a sample of the 41 largest municipalities in Brazil. Table 2 lists these selected municipalities.

The average death rate per 100,000 inhabitants in our sample of municipalities on December 31, 2022, is 413 ($\pm$ 89). Figure 7 shows the existence of outlier municipalities in this sample. Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO recorded the highest death rates at 580, 564, and 558, respectively. Conversely, Feira de Santana/BA, São Luís/MA, and Florianópolis/SC reported the lowest rates at 189, 234, and 245, respectively. We highlighted these two outlier groups in our investigations in Chapter 7.

Figure 8 shows the time series of death rates for the 41 largest Brazilian municipalities, illustrating the contrast between municipalities with the highest and lowest death rates.

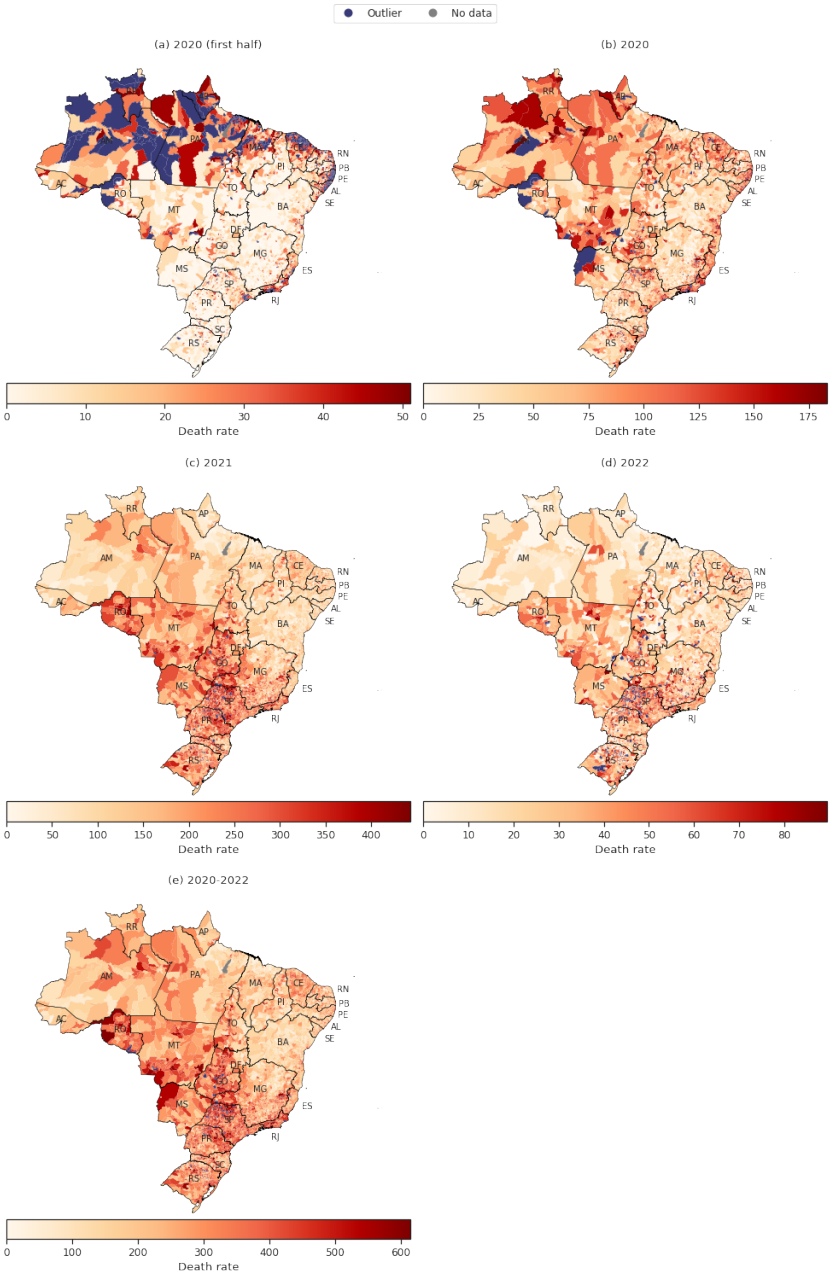


Figure 6 – Maps of Brazilian municipalities illustrating COVID-19 death rates per 100,000 inhabitants for: (a) the first half of 2020, (b) the year 2020, (c) the year 2021, (d) the year 2022, and (e) the cumulative period. Outliers represent municipalities with a death rate higher than the third quartile plus 1.5 times the interquartile range for each period. The missing data pertain to 10 Brazilian municipalities excluded from this thesis. The maps segment the states of the following regions of Brazil: North (AC: Acre, AM: Amazonas, AP: Amapá, PA: Pará, RO: Rondônia, RR: Roraima, TO: Tocantins), Northeast (AL: Alagoas, BA: Bahia, CE: Ceará, MA: Maranhão, PB: Paraíba, PE: Pernambuco, PI: Piauí, RN: Rio Grande do Norte, SE: Sergipe), Midwest (DF: Federal District, GO: Goiás, MS: Mato Grosso do Sul, MT: Mato Grosso), Southeast (ES: Espírito Santo, MG: Minas Gerais, RJ: Rio de Janeiro, SP: São Paulo), and South (PR: Paraná, RS: Rio Grande do Sul, SC: Santa Catarina).

Source: Mortality Information System (SIM) (DATASUS, 2022b).

Table 2 – List of the 41 largest Brazilian municipalities with populations over 500,000, forming the dataset of large municipalities. Source: 2022 Demographic Census (IBGE, 2022).

Municipality	**Population**	**Municipality**	**Population**
São Paulo/SP	11,451,245	Duque de Caxias/RJ	808,152
Rio de Janeiro/RJ	6,211,423	Nova Iguaçu/RJ	785,882
Brasília/DF	2,817,068	Natal/RN	751,300
Fortaleza/CE	2,428,678	Santo André/SP	748,919
Salvador/BA	2,418,005	Osasco/SP	743,432
Belo Horizonte/MG	2,315,560	Sorocaba/SP	723,574
Manaus/AM	2,063,547	Uberlândia/MG	713,232
Curitiba/PR	1,773,733	Ribeirão Preto/SP	698,259
Recife/PE	1,488,920	São José dos Campos/SP	697,428
Goiânia/GO	1,437,237	Cuiabá/MT	650,912
Porto Alegre/RS	1,332,570	Jaboatão dos Guararapes/PE	643,759
Belém/PA	1,303,389	Contagem/MG	621,865
Guarulhos/SP	1,291,784	Joinville/SC	616,323
Campinas/SP	1,138,309	Feira de Santana/BA	616,279
São Luís/MA	1,037,775	Aracaju/SE	602,757
Maceió/AL	957,916	Londrina/PR	555,937
Campo Grande/MS	897,938	Juiz de Fora/MG	540,756
São Gonçalo/RJ	896,744	Florianópolis/SC	537,213
Teresina/PI	866,300	Aparecida de Goiânia/GO	527,550
João Pessoa/PB	833,932	Serra/ES	520,649
São Bernardo do Campo/SP	810,729		

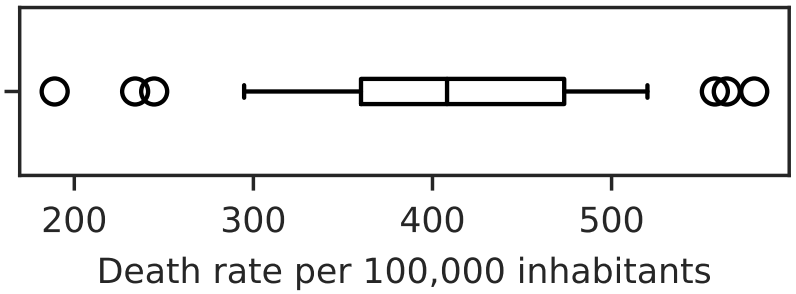


Figure 7 – Boxplot illustrating COVID-19 death rate per 100,000 for the 41 largest Brazilian municipalities across 2020-2022. The lower and upper bounds represent the first and third quartiles, respectively. The vertical line within the box indicates the median, while the whiskers extend to the minimum and maximum values within 0.7 times the interquartile range. The points represent the outlier municipalities.

Source: Mortality Information System (SIM) (DATASUS, 2022b).

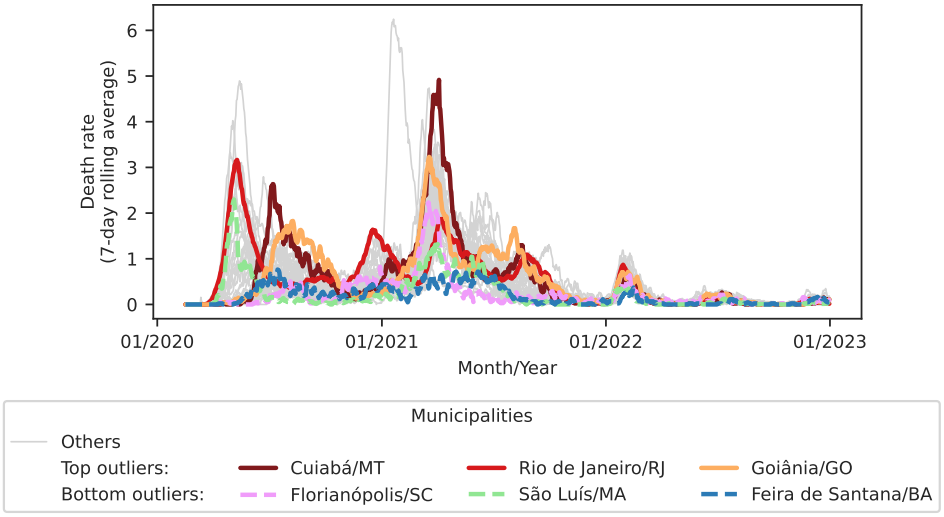


Figure 8 – COVID-19 death rate per 100,000 inhabitants in a 7-day rolling average for the 41 largest Brazilian municipalities. The plot highlights cities with notable deviations from the average death rate, categorized as top outliers (Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO) and bottom outliers (Feira de Santana/BA, São Luís/MA, and Florianópolis/SC).

Source: Mortality Information System (SIM) (DATASUS, 2022b)

3.2.2 Estimated reproduction number

We estimated the *effective reproduction number* (R_t) using the time series of new cases from SARS patients (DATASUS, 2022a). This estimation involved the use of the *epyestim* library[1], a Python toolkit implementing the methodology proposed by Cori et al. (2013). To apply this method, we initially specified the distribution for the COVID-19 generation time, approximating it by the serial interval reported by Bi et al. (2020) as $\sim$ Gamma$(2.29, 0.36)$, with an average of 6.36 days. Additionally, we defined the delay between infection and the onset of symptoms, representing the incubation period, as $\sim$ Lognormal$(1.57, 0.42)$, with an average time of 5.93 days (Bi et al., 2020).

Other parameters set in *epyestim* included a window size of 28 days to smooth the cases time series, a window size of 14 days for the final rolling average, one hundred bootstrap samples for estimating R_t, and a prior $R_t \sim$ Gamma$(9.90, 9.28)$.

Fig 9 illustrates the time-varying R_t of COVID-19 in Brazil, showing that R_t exceeds one in different moments of the study period. The highest R_t values occurred in the initial pandemic phase. The period from late 2020 to mid-2021, often identified as the second wave in Brazil, is marked by recurrent R_t peaks, closely spaced and with lower amplitude than other periods of this pandemic. In contrast, 2022 exhibits three prominent R_t peaks with substantial amplitude and spaced widely in time.

[1] https://github.com/lo-hfk/epyestim

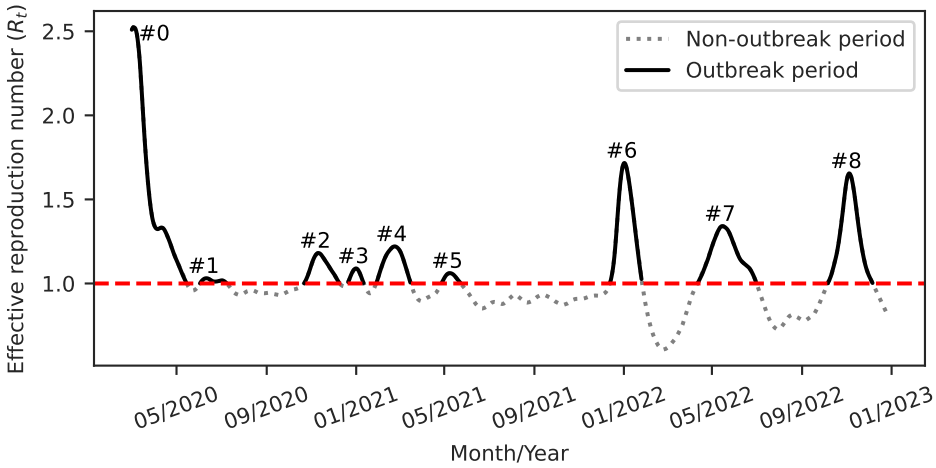


Figure 9 – Effective reproduction number (R_t) for COVID-19 in Brazil. The dashed horizontal line represents the reference value ($R_t = 1$) used to monitor epidemics. The R_t time series alternates between outbreak periods (solid line) and non-outbreak periods (dotted line). Nine outbreak periods, labeled from #0 to #8, were identified.

We inferred the basic reproduction number (R_0) by considering the R_t peak during the early pandemic phase. Our analysis estimated R_0 for Brazil at 2.52 (95% CI: 2.34 - 2.71). Our estimated R_0 aligns with other studies estimating R_0 for Brazil, such as $R_0 = 3.10$ (95% CI: 2.40 – 5.50) reported by de Souza et al. (2020), $R_0 = 2.13$ (95% CI: 0.81 - 3.04) estimated by Arroyo-Marioli et al. (2021), and $R_0 = 2.78$ (95% CI: 1.9 - 3.81) for a serial interval mean of 4.8 days calculated by Nouvellet et al. (2021). Additionally, Nouvellet et al. (2021) estimated $R_0 = 3.81$ (95% CI: 2.73 - 4.81) for a serial interval mean of 6.48 days, which approximates of our estimation.

Figure 10 shows the time-varying behavior of R_t estimated for the 41 largest Brazilian municipalities. Notably, in the first year, epidemic periods exhibited some lack of synchronization among municipalities; however, by early 2021 and across 2022, four synchronized epidemic periods became apparent.

The lack of synchronization in the R_t estimates for the municipal dataset over a significant portion of the study period (Figure 10), compared to the R_t observed in the national-level estimates (Figure 9), supports the findings of Birello et al. (2024). Their study indicates that reproduction number estimates derived from surveillance data may be unreliable when applied to spatially structured populations, as the interaction between space and mobility can obscure the actual dynamics of the epidemic. These observations reinforce the insight that R_t estimates are more informative when calculated for specific localities rather than for large and heterogeneous regions.

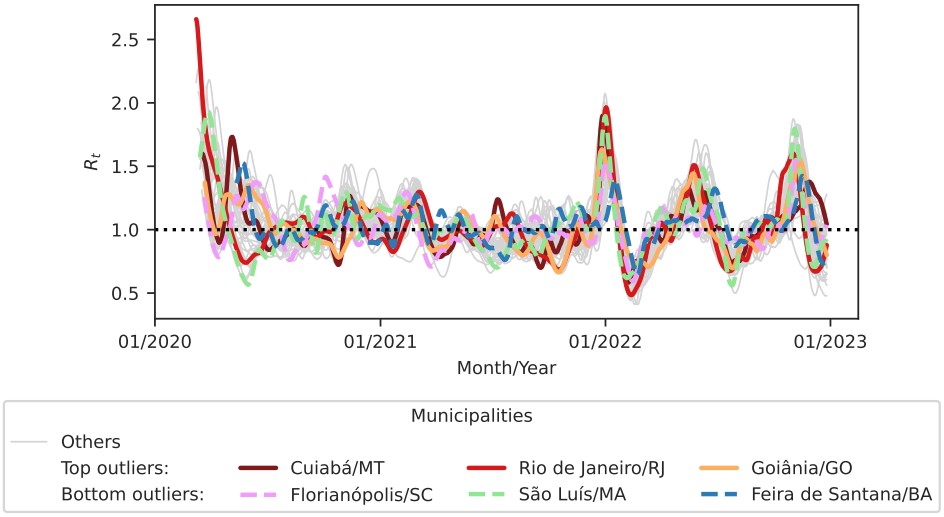


Figure 10 – Effective reproduction number (R_t) estimated for the 41 largest Brazilian municipalities. The dotted horizontal line represents the reference value ($R_t = 1$) used to monitor epidemics. The plot highlights cities with notable deviations from the average death rate, categorized as top outliers (Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO) and bottom outliers (Feira de Santana/BA, São Luís/MA, and Florianópolis/SC).

3.2.3 COVID-19 outbreaks

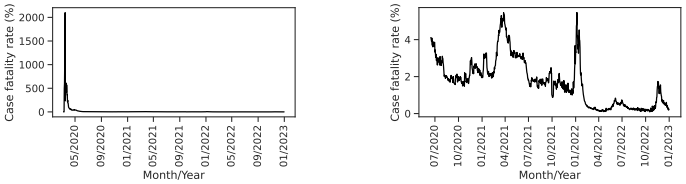
Successive waves of cases and deaths characterize the dynamic of the COVID-19 pandemic. Usually, we describe that the pandemic begins in Brazil with an initial wave in 2020, followed by a second wave spanning the end of 2020 through 2021, and by three distinct waves throughout 2022. However, for enhanced modeling in this thesis, we opt for a more fine approach by segmenting the epidemic periods based on the emergence of outbreaks. Here, we define an outbreak as when the disease spreads actively, signifying a phase when the epidemic is out of control.

To identify these outbreaks, we analyze periods where the R_t remains above one for a minimum duration of seven consecutive days, allowing for a maximum of seven days below this threshold during the identified outbreak period.

Consequently, our analysis identified nine COVID-19 outbreaks in Brazil over the study period, as illustrated in Fig 9. There are two outbreaks during the initial wave, four during the second wave, and three in 2022. We also identified outbreaks for the 41 largest Brazilian municipalities when we verified that each municipality experienced approximately 10.6 outbreaks ($\pm$ 1.5).

3.2.4 Case Fatality Rate (CFR)

We calculated the Case Fatality Rate (CFR) by dividing the number of deaths on a given day by the number of COVID-19 reported cases 12 days before, based on the methodology suggested by Baud et al. (2020). Figure 11a presents the time series of COVID-19 CFR in Brazil. Notably, in the early days of the pandemic, the SIM (DATASUS, 2022b) reported more deaths than the number of cases reported by Panel Monitoring (DATASUS, 2020a), indicating significant underreporting of cases during that period. Figure 11b shows the CFR time series excluding the first 120 days, where the CFR fluctuates between 1% and 5% during the first and second years. However, the CFR remains consistently below in 2022, with a median of 0.62%.



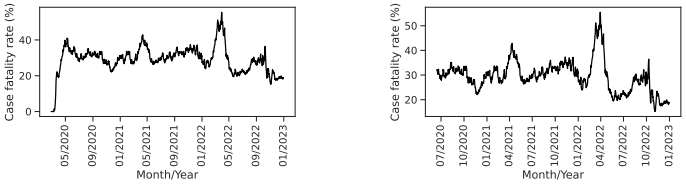
(a) CFR time series

(b) CFR time series excluding the first 120 days

Figure 11 – General Case Fatality Rate (CFR) calculated from cases reported in the Monitoring Panel and deaths reported in the Mortality Information System (SIM).

Source: Monitoring Panel (DATASUS, 2020a) and SIM (DATASUS, 2022b).

We also calculated the CFR for SARS patients, as shown in Figure 12. Unlike the general CFR, which revealed significant underreporting of cases during the early stages of the pandemic, the SARS CFR exhibited a lower proportion of deaths reported in the initial period compared to later stages. Both general and SARS CFR metrics were potentially influenced by case underreporting, particularly during pandemic peaks. Nevertheless, despite these reporting challenges, our analysis indicates a gradual decline in COVID-19 lethality throughout the pandemic.



(a) SARS CFR time series

(b) SARS CFR excluding the first 120 days

Figure 12 – Case Fatality Rate (CFR) calculated for Severe Acute Respiratory Syndrome (SARS) patients.

Source: SARS database (DATASUS, 2022a).

3.3 Public health measures

The first doses of COVID-19 vaccines began to be administered globally in December 2020 (Dong et al., 2020). Prior to this, the WHO had recommended non-pharmacological measures to prevent the spread of the virus, such as hand washing, physical distancing, suspension of mass gatherings, closure of non-essential workplaces and schools, reduced public transport, enhanced screening, quarantine, and mask-wearing (WHO, 2020a,c). However, the federal government did not follow the WHO recommendations in Brazil and failed to coordinate a national response to the pandemic (Lancet, 2020; Azevedo et al., 2020; Asano et al., 2021). The scientific community heavily criticized President Jair Bolsonaro and his administration for promoting denialism (Ferrante et al., 2021a; Barberia and Gómez, 2020; Lasco, 2020; Ricard and Medeiros, 2020; Lotta et al., 2020), opposing containment measures (Ferrante et al., 2021a; Barberia and Gómez, 2020; Ricard and Medeiros, 2020), advocating for unproven COVID-19 treatments (Ferrante et al., 2021a; Barberia and Gómez, 2020; Lasco, 2020; Ricard and Medeiros, 2020; Lotta et al., 2020), discouraging vaccination (Ferrante et al., 2021a), and downplaying the importance of wearing masks (Ferrante et al., 2021a; Lasco, 2020).

In response to the COVID-19 pandemic, the Supreme Federal Court (STF) granted autonomy to states and municipalities to take non-pharmacological measures[2] in order to prevent a collapse of the healthcare system in Brazil. With the federal government neglecting its responsibilities, local governments implemented various measures, such as school closures, which were widely adopted (Halpern, 2021; Bartholo et al., 2022). Additionally, municipalities encouraged using face masks, remote work, and the closure of non-essential businesses to reduce human mobility in the early pandemic (Petherick et al., 2020).

3.3.1 Reduction in human mobility patterns

Figure 13 illustrates how the COVID-19 pandemic affected the mobility patterns of the Brazilian population. The first notable change occurred on March 17, 2020, when the country reported its first COVID-19 death. The first outbreak in 2020 was characterized by the highest peak in Δ_H, even though the death peak occurred during 2021. It is important to note that although Δ_H tends to converge to the baseline, there were periods when this trend was interrupted. For instance, during the January months in 2021 and 2022, corresponding to the school vacation period, and in mid-2021, coinciding with the peak of COVID-19 deaths, Δ_H remained high.

[2] https://g1.globo.com/politica/noticia/2020/04/15/maioria-do-supremo-vota-a-favor-de-que-estados-e-municipios-editem-normas-sobre-isolamento.ghtml

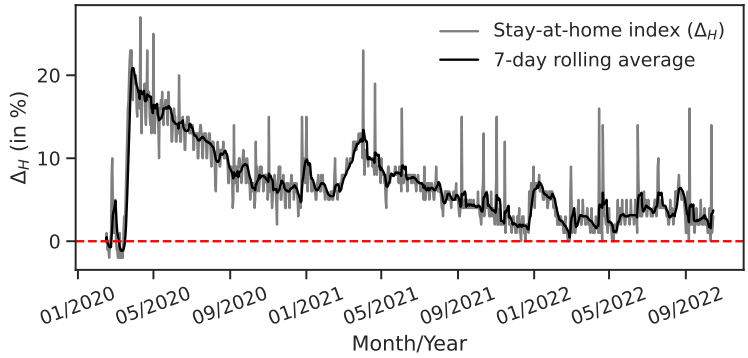


Figure 13 – *Stay-at-home index* (Δ_H) reported in Brazil during the COVID-19 pandemic. The dashed horizontal line shows the baseline ($\Delta_H = 0\%$).

Source: Google COVID-19 Mobility Report (Google, 2020).

We also observed a weekly seasonality in Δ_H, with lower values on weekends than on weekdays. This pattern is consistent with pre-pandemic behavior, where people typically spent more time at home during weekends. Additionally, the Δ_H trends for the 41 largest Brazilian municipalities closely mirrored the national patterns, as illustrated in Figure 14.

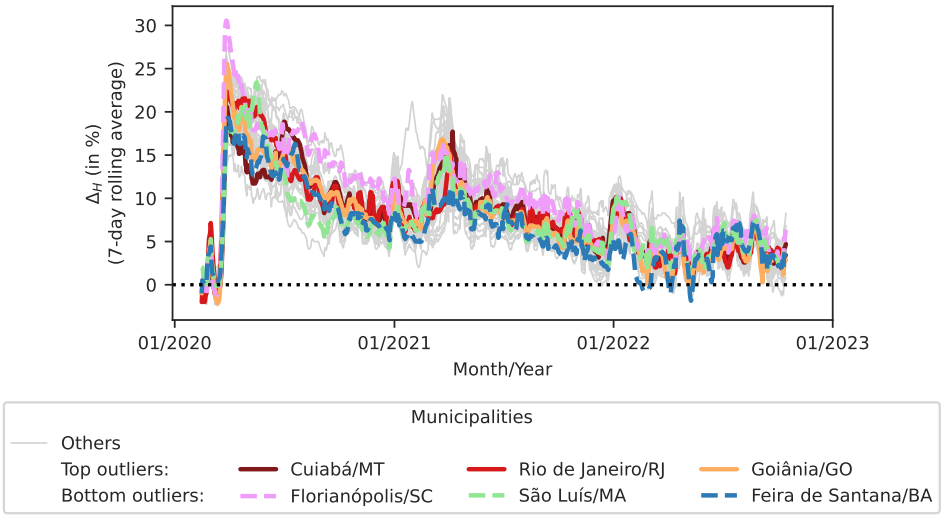


Figure 14 – *Stay-at-home index* (Δ_H) reported in a 7-day rolling average for the 41 largest Brazilian municipalities. The dotted horizontal line highlights the baseline ($\Delta_H = 0\%$). The plot highlights cities with notable deviations from the average death rate, categorized as top outliers (Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO) and bottom outliers (Feira de Santana/BA, São Luís/MA, and Florianópolis/SC).

Source: Google COVID-19 Mobility Report (Google, 2020).

3.3.2 COVID-19 vaccination campaign

Figure 15 illustrates the progress of the COVID-19 vaccination campaign in Brazil, which administrated the first dose on January 17, 2021. The data reveal that only by October 2021, half of the population received the total dosage required for complete vaccination. By the end of 2022, 86% of Brazilians were fully vaccinated, reflecting a significant public interest in immunization. Furthermore, the metric indicating vaccine doses per 100 people exceeding 200 underscores the extensive interest by the population in booster doses.

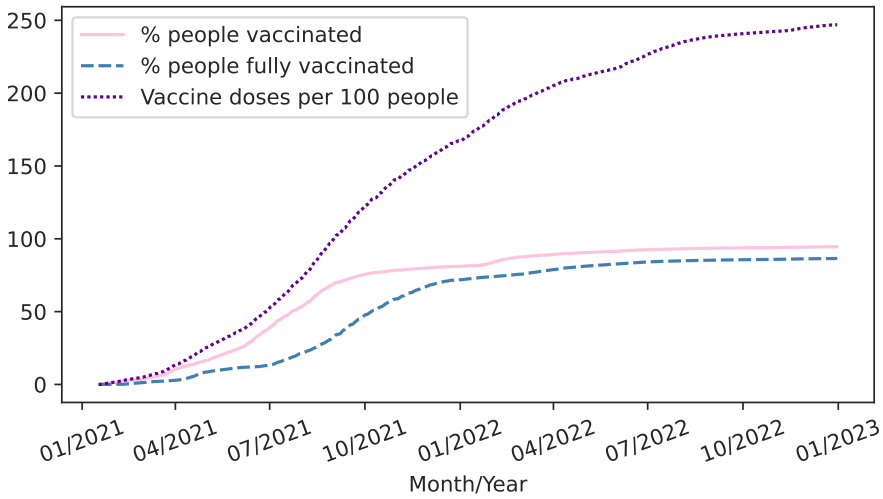


Figure 15 – Cumulative time series of COVID-19 vaccination in Brazil, showing the percentage of vaccinated individuals, percentage of fully vaccinated individuals, and doses administered per 100 inhabitants.

Source: Brazilian Health Ministry (Brasil, 2022).

Table 3 reveals that some municipalities reported vaccination rates exceeding 100%, indicating potential data inconsistencies. The database provided by the Brazilian Health Ministry (Brasil, 2022) includes patient identification and the municipality of residence; inaccuracies in either of these attributes could result in overreporting. Therefore, we advise caution when interpreting the vaccination data.

Table 3 – Descriptive statistics of COVID-19 vaccination metrics for 5,560 Brazilian municipalities during 2021-2022, highlighting the percentages of vaccinated individuals, fully vaccinated individuals, and doses administered per 100 people. Source: Brazilian Health Ministry (Brasil, 2022).

	Mean	SD	Min	Q1	Median	Q3	Max
% people vaccinated	93	14	29	86	94	101	272
% people fully vaccinated	85	15	21	77	86	95	185
Doses per 100 people	252	52	55	221	254	285	576

SD: Standard deviation. Min: Minimum. Q1: First quartile. Q3: Third quartile. Max: Maximum.

Figure 16 illustrates that municipalities in the North region reported lower vaccination coverage than those in other regions of Brazil. However, due to uncertainties surrounding the data provided by the Brazilian Health Ministry (Brasil, 2022), we again advise caution in drawing definitive conclusions from this observation.

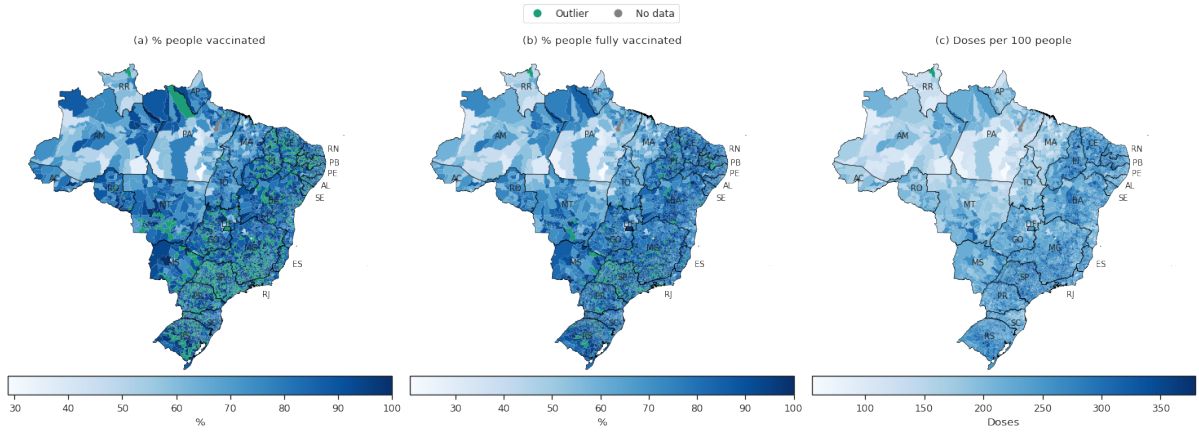


Figure 16 – Maps of COVID-19 vaccination coverage in Brazilian municipalities (2021-2022): Panels show key vaccination metrics across Brazilian municipalities: (a) percentage of population vaccinated, (b) percentage fully vaccinated, and (c) doses administered per 100 people. Outliers in panels (a) and (b) denote municipalities with vaccination rates exceeding 100%, while outliers in panel (c) reflect municipalities with doses above the third quartile plus 1.5 times the interquartile range. The missing data pertain to 10 Brazilian municipalities excluded from this thesis. The maps segment the states of the Brazil: AC: Acre, AL: Alagoas, AM: Amazonas, AP: Amapá, BA: Bahia, CE: Ceará, DF: Federal District, ES: Espírito Santo, GO: Goiás, MA: Maranhão, MG: Minas Gerais, MS: Mato Grosso do Sul, MT: Mato Grosso, PA: Pará, PB: Paraíba, PE: Pernambuco, PI: Piauí, PR: Paraná, RJ: Rio de Janeiro, RN: Rio Grande do Norte, RO: Rondônia, RR: Roraima, RS: Rio Grande do Sul, SC: Santa Catarina, SE: Sergipe, SP: São Paulo, TO: Tocantins.

Source: Brazilian Health Ministry (Brasil, 2022).

Additionally, Figure 17 shows the vaccination in the largest Brazilian cities. These locations achieved 50% full vaccination coverage between September and November 2021, mirroring the national trends shown in Figure 15. There is no evident distinction in vaccination efforts between municipalities classified as bottom outliers and those as top outliers.

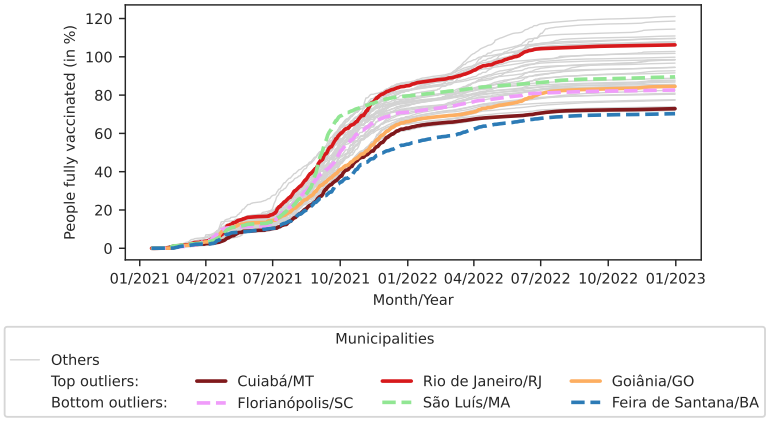


Figure 17 – Cumulative time series showing the percentage of people fully vaccinated against COVID-19 across the 41 largest Brazilian municipalities. The plot highlights cities with notable deviations from the average death rate, categorized as top outliers (Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO) and bottom outliers (Feira de Santana/BA, São Luís/MA, and Florianópolis/SC).

Source: Brazilian Health Ministry (Brasil, 2022).

3.3.3 Economic support and health infrastructure expansion

The *Auxílio Emergencial* socioeconomic program was crucial in promoting social isolation and mitigating the spread of the coronavirus throughout 2020 (Albani et al., 2022; de Leon et al., 2023). This program reached approximately 22% of the Brazilian population, paying a monthly mean of USD 154.86 from April to August 2020, continuing with reduced amounts after that period (Albani et al., 2022). In parallel, Brazil also responded to the COVID-19 challenge by progressively increasing the number of Intensive Care Unit (ICU) beds dedicated to COVID-19 patients during the pandemic, as depicted in Figure 18 (DATASUS, 2020b). Although it does not directly impact disease transmission dynamics, this measure is a significant intervention that can potentially alleviate the IFR.

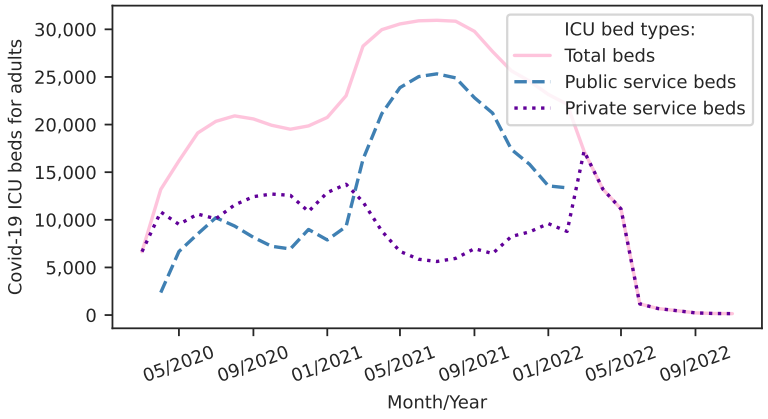


Figure 18 – Time series of COVID-19 Intensive Care Unit (ICU) beds for adults in Brazil.

Source: Brazilian Health Ministry (DATASUS, 2020b).

3.4 Prevalence of coronavirus variants

The prevalence of coronavirus variants is a relevant topic for understanding COVID-19 in Brazil. Figure 19 shows that the emergence of the Gamma variant coincides with the second wave of the virus in the country. This variant was initially identified in the Brazilian state of Amazonas and has been associated with increased transmissibility, higher mortality, and immune evasion, resulting in reinfections and potentially reduced efficacy of vaccines and neutralizing antibodies (Zimerman et al., 2022).

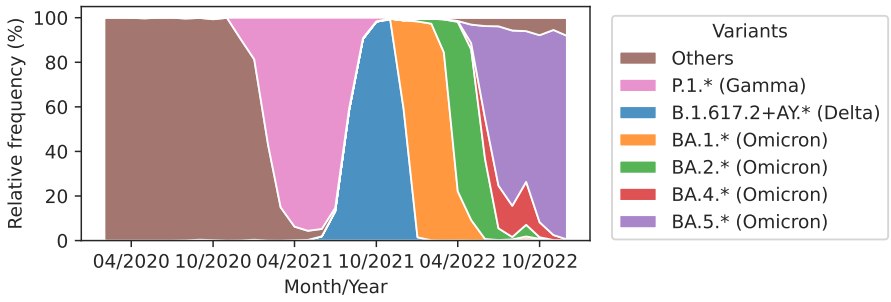


Figure 19 – Monthly relative frequency (%) of the coronavirus variants over time in Brazil.

Source: Fiocruz (2020)

During the middle of the second COVID-19 wave in Brazil, the Delta variant became the dominant variant. Like the Gamma variant, the Delta variant exhibits increased transmissibility, mortality, and immune evasion (Tian et al., 2021). Subsequently, coinciding with the onset of the third wave, the Omicron (BA.1) variant swiftly became the prevailing strain. Throughout 2022, other Omicron subvariants, including BA.2, BA.4, and BA.5, have emerged as the most dominant strains in the country. The Omicron variant has many mutations compared to its predecessors, which exhibit elevated transmissibility and the ability to evade the immune response (Rana et al., 2022; Nyberg et al., 2022). However, works indicate Omicron with lower severity regarding disease outcomes (Beppu et al., 2024; Rana et al., 2022; Nyberg et al., 2022; Ward et al., 2022; Lorenzo-Redondo et al., 2022).

4 MULTIPLE REGRESSION ANALYSIS OF KEY FACTORS IN COVID-19 MORTALITY

Brazil exhibits significant socioeconomic inequality among its regions (Salata, 2020; Mourao and Junqueira, 2021; Cavalini and de Leon, 2008). The coronavirus pandemic raised concerns about its impact on areas with pronounced poverty, particularly the North and Northeast regions. Studies investigating sociodemographic factors and COVID-19 outcomes in Brazil have addressed these concerns in the first pandemic year (Rocha et al., 2021; Castro-Alves et al., 2022; Bermudi et al., 2021; Martines et al., 2021; Demenech et al., 2020; Figueiredo et al., 2020; Raymundo et al., 2021).

Misinformation was another issue of concern during the pandemic. Influential figures on the Internet and political leaders promoting fake news and conspiracy theories can damage public adherence to health authority guidelines (Loomba et al., 2021; Bridgman et al., 2020). In Brazil, for instance, former President Jair Bolsonaro displayed controversial conduct in pandemic management, as discussed in Section 3.3. This behavior stimulated studies investigating the correlation between political preferences in Brazilian municipalities and COVID-19 outcomes (Lima et al., 2024a; Xavier et al., 2022b; Hastenreiter Filho and Cavalcante, 2022). Additionally, misinformation can contribute to vaccine hesitancy, affecting a public health strategy for combating infectious disease outbreaks (Loomba et al., 2021; Wilson and Wiysonge, 2020).

We advocate for a deep analysis to investigate these concern factors in an extended analysis period. Our motivation is because the most affected regions changed over time, as presented in Figures 5 and 6. Furthermore, given that Brazil is a large country, encompassing 5,569 municipalities and one Federal District, it is essential to conduct rigorous analyses at the municipal level using comprehensive sociodemographic datasets that capture the characteristics of these locations.

In this chapter, we formulated four research questions to guide our analysis:

1. What profiles of Brazilian municipalities correlate most strongly with COVID-19 mortality?
2. What sociodemographic factors are associated with COVID-19 mortality across Brazilian municipalities?
3. Is the political preference manifested by the municipal populations correlated with COVID-19 mortality?
4. Did municipalities with higher COVID-19 vaccination coverage experience lower mortality rates?

To address these research questions, we conducted an ecological study examining the association between covariates and COVID-19 mortality across Brazilian municipalities during the first three years of the pandemic (2020–2022). Our regression analysis controls for municipality groups, sociodemographic variables, political preferences, vaccination coverage, and temporal exposure to the virus. We emphasize temporal dynamics and their influence on the relationships between these variables and the observed outcomes throughout the study period.

We initially selected 31 sociodemographic variables to characterize Brazilian municipalities. We applied a Gaussian Mixture Model (GMM) (Dempster et al., 1977) to identify sociodemographic clusters. We then employed negative binomial regression models to estimate the correlations between key factors and COVID-19 mortality across these municipalities. Section 4.1 details our methodology. We present the results and discussion in Sections 4.2 and 4.3, respectively.

4.1 Methodology

4.1.1 Study design

In this study, we employed a data mining process encompassing the steps: database collection, attribute pre-processing, attribute selection, execution of a clustering algorithm, and evaluation of the resulting clusters. From an epidemiological view, the research adopts an ecological study design, using Brazilian municipalities as the primary unit of analysis, following a multiple-group approach with an analytic dimension (Morgenstern, 1995).

4.1.2 Data

The outcome variable for this study is the number of COVID-19 deaths per municipality in Brazil, sourced from the Mortality Information System (SIM) database (DATASUS, 2022b), as detailed in Sections 3.1.1 and 3.2.1.3. The analysis covers the period from March 12, 2020, to December 31, 2022.

On the side of explanatory variables, we collected data on municipalities from various public sources to describe their social, economic, and demographic characteristics. We obtained data from the 2022 Demographic Census provided by IBGE (2022). Since the 2022 Demographic Census data had not yet been fully published, we also utilized data from the 2010 Demographic Census (IBGE, 2010). Additionally, we gathered demographic and economic data from Atlas Brasil, an online platform offering databases with indicators derived from the 2010 Census (PNUD et al., 2013) and administrative records from public bodies (PNUD et al., 2017). In total, we analyzed 31 sociodemographic attributes, illustrated across Brazilian municipalities in Figure 20.

We also regarded the vaccination coverage for each municipality, expressed by the percentage of individuals who are fully vaccinated against COVID-19 (Brasil, 2022), as discussed in Sections 3.1.2 and 3.3.2.

Additionally, we collected electoral data to characterize the political preferences in the Brazilian municipalities. Specifically, we computed the percentage of votes received by Jair Bolsonaro in the first round of the 2022 Brazilian Presidential Election (TSE, 2022). This measure enables us to explore potential correlations between political affiliation and COVID-19 mortality, particularly given the controversial public health policies advocated by former President Jair Bolsonaro during the pandemic, as discussed in Section 3.3. This approach is consistent with similar studies that have investigated the correlation between COVID-19 impacts and political preferences (Hastenreiter Filho and Cavalcante, 2022; Xavier et al., 2022b; Ajzenman et al., 2023; Lima et al., 2024a). Figure 21 represents this attribute across Brazilian municipalities.

Appendix A recaps the statistical characteristics of the data used in this chapter.

4.1.3 Clustering municipalities by sociodemographic variables

We conducted a clustering analysis using the 31 sociodemographic variables listed in Table 14 (Appendix A). First, we standardized the data to ensure that each variable contributed equally to our analysis. Then, we applied Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) to reduce the dataset to its first two principal components, enhancing both the outcomes and performance of the clustering algorithm. These two principal components accounted for 50% of the original data variance.

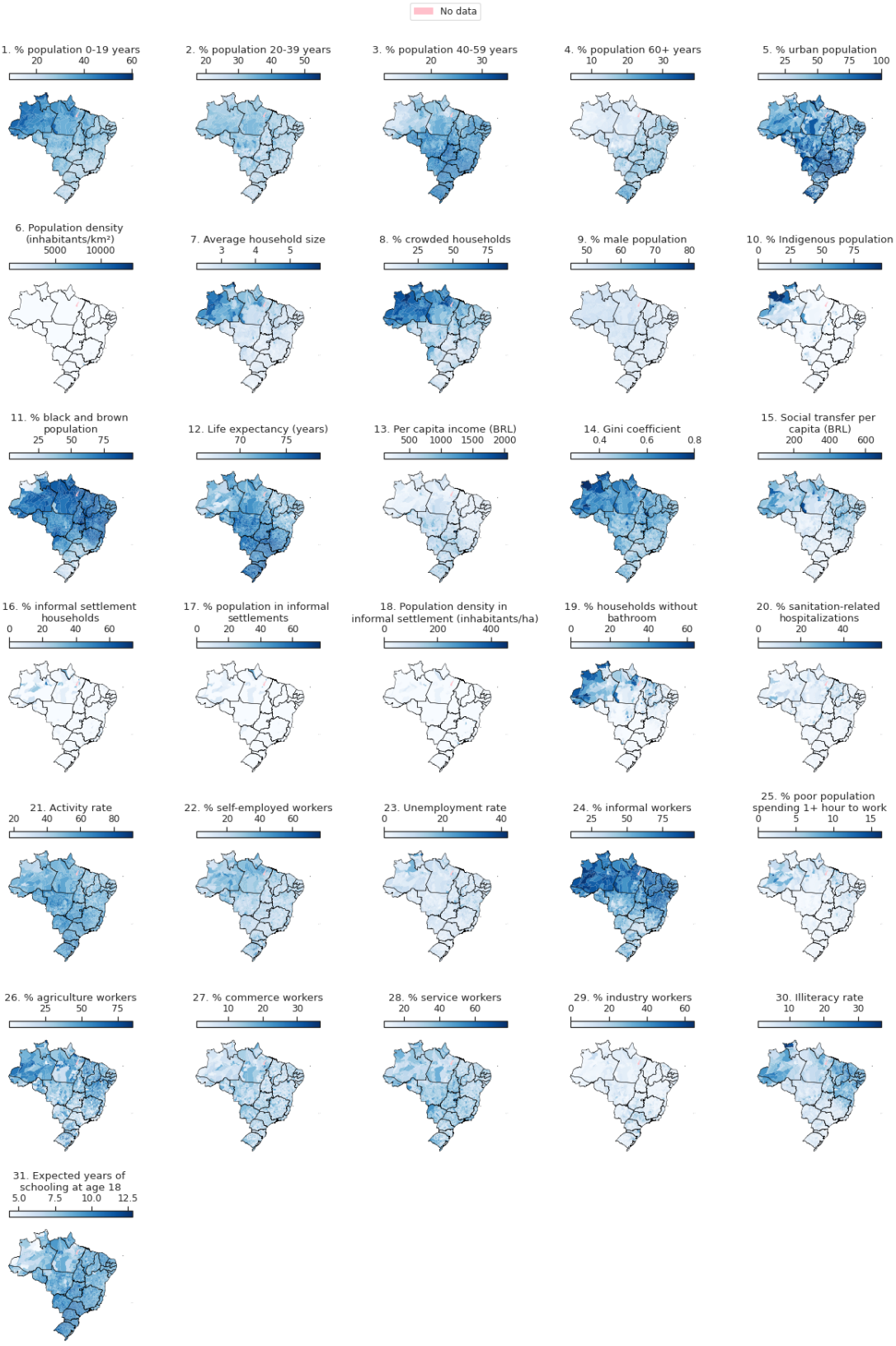


Figure 20 – Colormaps illustrating the 31 sociodemographic attributes of Brazilian municipalities analyzed in this thesis. The missing data pertain to 10 Brazilian municipalities excluded from this study.

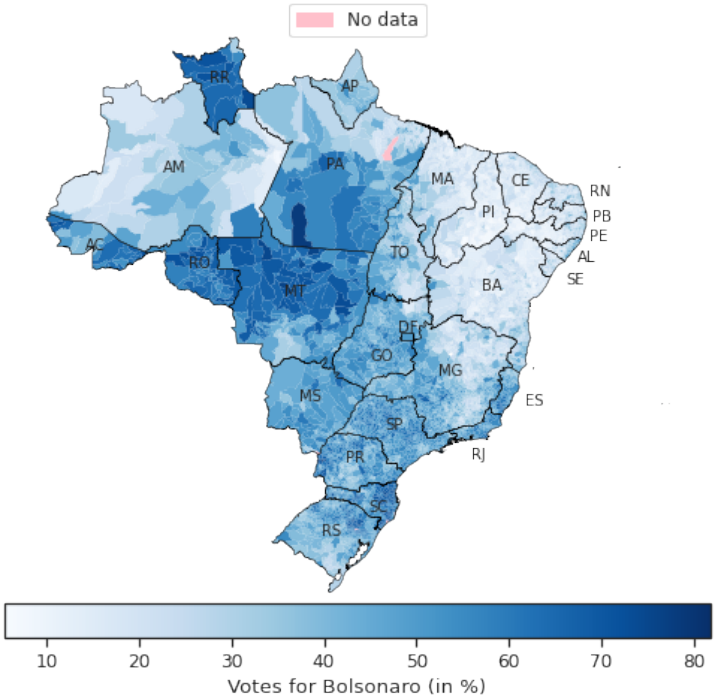


Figure 21 – Map of Brazilian municipalities highlighting the percentage of votes for Jair Bolsonaro in the first round of the 2022 Brazilian Presidential Election. The missing data pertain to 10 Brazilian municipalities excluded from this study. The maps segment the states of the Brazilian regions: North (AC: Acre, AM: Amazonas, AP: Amapá, PA: Pará, RO: Rondônia, RR: Roraima, TO: Tocantins), Northeast (AL: Alagoas, BA: Bahia, CE: Ceará, MA: Maranhão, PB: Paraíba, PE: Pernambuco, PI: Piauí, RN: Rio Grande do Norte, SE: Sergipe), Midwest (DF: Federal District, GO: Goiás, MS: Mato Grosso do Sul, MT: Mato Grosso), Southeast (ES: Espírito Santo, MG: Minas Gerais, RJ: Rio de Janeiro, SP: São Paulo), and South (PR: Paraná, RS: Rio Grande do Sul, SC: Santa Catarina).

Source: TSE (2022).

This work used the GMM clustering algorithm (Dempster et al., 1977) to perform clustering models with 2 to 7 clusters. For each model with k clusters, we performed it 100 times using spherical covariance, meaning that each cluster has its own single variance. We performed all clustering with the Python 3 library *sklearn* version 1.5.

For each k-clustering, we selected the one model that maximized the log-likelihood (Zaki and Meira Jr, 2020), as defined by:

$$\ln(P(D|\mu, \Sigma, P(C))) = \sum_{j=1}^{n} \ln\left(\sum_{i=1}^{k} \mathcal{N}(x_j|\mu_i, \Sigma_i)P(C_i)\right), \tag{4.1}$$

where D represents our dataset, consisting of n points x_j in a 2-dimensional space $\mathbb{R}^2$, corresponding to the first two principal components. Let X_a denote the random variable

associated with the a-th principal component. Let $X = (X_1, X_2)$ represent the vector random variable across the two principal components, with x_j being a sample from X. Furthermore, k denotes the number of clusters C_i, μ_i is the mean of cluster C_i, Σ_i is the covariance of cluster C_i, $P(C_i)$ is the probability of cluster C_i, and $\mathcal{N}$ denotes the normal probability density at x_j attributable to cluster C_i.

This algorithm provides a soft assignment of each dataset point x_j to each cluster C_i, indicating the probability that each x_j belongs to each C_i (Zaki and Meira Jr, 2020). Additionally, we conducted a hard assignment analysis, which involved linking each x_j to a single cluster C_i based on $max(P(C_i|x_j))$.

4.1.3.1 Sociodemographic clusters

In this study, we selected a model with five sociodemographic clusters for the statistical analysis. We labeled these clusters as *Urbanized*, *Urbanized with Informal Settlements*, *Semi-urbanized*, *Rural with High Human Development*, and *Rural with Low Human Development*. Figure 22 depicts these sociodemographic clusters across the Brazilian territory.

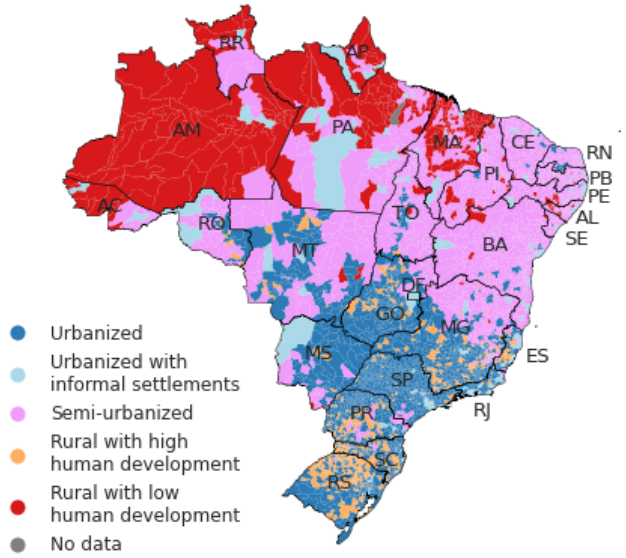


Figure 22 – Map of Brazilian municipalities highlighting sociodemographic clusters by colors. The missing data pertain to 10 Brazilian municipalities excluded from this study. States and Federal District of Brazil by regions: North (AC: Acre, AM: Amazonas, AP: Amapá, PA: Pará, RO: Rondônia, RR: Roraima, and TO: Tocantins), Northeast (AL: Alagoas, BA: Bahia, CE: Ceará, MA: Maranhão, PB: Paraíba, PE: Pernambuco, PI: Piauí, RN: Rio Grande do Norte, and SE: Sergipe), Midwestern (DF: Federal District, GO: Goiás, MS: Mato Grosso do Sul, and MT: Mato Grosso), Southeast (ES: Espírito Santo, MG: Minas Gerais, RJ: Rio de Janeiro, and SP: São Paulo), and South (PR: Paraná, RS: Rio Grande do Sul, and SC: Santa Catarina).

The *Urbanized* cluster comprises 1,994 municipalities concentrated in Southeast, South, and Midwestern Brazilian regions. This cluster shows a high proportion of the population living in urban areas with high per capita income. Another 261 municipalities fall into the *Urbanized with Informal Settlements* cluster, similar to the *Urbanized* cluster but with a significant population living in informal settlements. These municipalities are generally located in metropolitan areas and are dispersed throughout the country.

The *Semi-urbanized* cluster is the largest in our analysis, encompassing 2,136 municipalities characterized by a mix of urban and rural areas. This cluster shows a notable reduction in human development indicators compared to the urbanized clusters. Most municipalities in the Northeast region, along with those in the North of Minas Gerais and parts of the Midwestern and North regions, fall into this cluster.

The *Rural with Low Human Development* cluster includes 312 municipalities that experience the highest socioeconomic deprivation in the country, characterized by very low income, reduced life expectancy, and poor educational indicators. These municipalities are predominantly located in the North region. In contrast, the *Rural with High Human Development* cluster comprises 857 municipalities, primarily situated in the South region, with indicators reflecting a good quality of life. A distinguishing feature of the *Rural with High Human Development* cluster is the high proportion of the elderly population; 22% of the population in this cluster is aged 60 years or older, which is more than twice the proportion in the *Rural with Low Human Development* cluster.

We provide a detailed view of the clustering selection process and descriptive cluster statistics in the Appendix B.

4.1.4 Statistical analysis

Before we define the regression models, we identified collinearity among the sociodemographic variables, which are listed in Table 14 (Appendix A). So, we took feature selection to mitigate confounding effects. Resulting in the selection of a subset with 15 variables: (i) the percentage of the population aged 60 years or more, (ii) the percentage of the urban population, (iii) the population density, (iv) the percentage of the male population, (v) the percentage of the Indigenous population, (vi) the Gini coefficient, (vii) the percentage of informal settlement households, (viii) the population density in informal settlements, (ix) the percentage of sanitation-related hospitalizations, (x) the percentage of self-employed workers, (xi) the unemployment rate, (xii) the percentage of commerce workers, (xiii) the percentage of service workers, (xiv) the percentage of industry workers, and (xv) the expected years of schooling at age 18. The maximum values of VIF and Pearson correlation observed for all selected variables are 3.4 and 0.64, respectively.

We assessed the association between the explanatory variables and COVID-19 mortality in Brazilian municipalities across five distinct periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period from 2020 to 2022. We employed negative binomial regression models to account for the overdispersion in the outcomes, as the Poisson regression model is inadequate for such cases (Dunn et al., 2018).

We employed two models in our analysis. The first model, referred to as *Model 1*, incorporates dummy variables representing four of our five sociodemographic clusters. We defined the *Urbanized* cluster as the reference group. To account for additional factors, the model includes the 15 selected sociodemographic variables, the cumulative percentage of the population fully vaccinated at a specified period, and the percentage of votes cast for Jair Bolsonaro in the 2022 Presidential Election. The model is expressed as follows:

$$\ln(\mu_t) = a_{0,t} + \sum_{k=1}^{4} a_{k,t} C_k + \sum_{j=1}^{15} a_{4+j,t} x_{4+j} + a_{20,t} v_t + a_{21,t} b + \ln(P), \tag{4.2}$$

where μ_t denotes the estimated number of COVID-19 deaths during period t, a represents the vector of model coefficients estimated for t, C_k refers to the sociodemographic clusters, x represents the sociodemographic variables, v is the percentage of the population fully vaccinated at t, b is the percentage of votes for Bolsonaro, and $\ln(P)$ serves as the offset to adjust for population size variation.

The second model, referred to as *Model 2*, extends *Model 1* by introducing a temporal exposure term, Δ_t, as an additional offset. This term represents the days from the first recorded COVID-19 death in a municipality to the last day of period t. The inclusion of Δ_t aims to account for the temporal influence of the outbreak onset on μ_t during each period t. The model is formulated as follows:

$$\ln(\mu_t) = a_{0,t} + \sum_{k=1}^{4} a_{k,t} C_k + \sum_{j=1}^{15} a_{4+j,t} x_{4+j} + a_{20,t} v_t + a_{21,t} b + \ln(P) + \ln(\Delta_t). \tag{4.3}$$

We standardized all explanatory variables to enable direct comparison of their coefficients, providing insight into their relative importance within the model. We analyzed the results using exponentiated coefficients, i.e., Rate Ratio (RR), with 95% CI. We systematically tuned the negative binomial distribution dispersion by fitting models across a range of the dispersion parameter α values and selecting the one that maximizes the log-likelihood function. We also evaluated model fit using deviance, χ^2, R^2_{CS}, R^2_{McF}, log-likelihood, AIC and BIC.

Additionally, we evaluated the potential gain of *Model 2* relative to *Model 1*. This comparison involved assessing the relative improvement of the *null Model 2* over *null Model 1* and examining the relative gain of the *fitted Model 2* compared to *null Model 1*. To quantify this gain, we used the following metric:

$$G(\text{ext}, \text{base}) = \frac{\text{LL}_{\text{ext}} - \text{LL}_{\text{base}}}{|\text{LL}_{\text{base}}|}, \tag{4.4}$$

where *base* represents a baseline model, and *ext* represents an extended model, and *LL* corresponds to the log-likelihood of a model.

We performed two sensitivity analyses on *Model 2* to evaluate its stability and robustness. First, we generated 30 bootstrap resamplings for each period to assess the variability of the model parameters. Additionally, we examined model robustness by re-estimating it after excluding outliers and influential points to determine their impact on the results.

All analyses were performed using the Python 3 library *statsmodels* version 0.14.

4.2 Results

Table 4 presents the goodness-of-fit statistics for *Models 1* and *2*, evaluating their performance in predicting COVID-19 mortality across Brazilian municipalities during five analysis periods. Across all periods, *Model 2*, which introduces an additional temporal exposure offset, consistently outperforms *Model 1*.

Table 4 – Goodness-of-fit statistics for the regression models across five analysis periods.

Statistic	2020 (first half)		2020		2021		2022		2020-2022	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
LL_{null}	-11,271	**-8,813**	-17,079	**-15,795**	-22,418	**-21,911**	-13,642	**-12,599**	-23,744	**-23,614**
LL_{model}	-10,213	**-8,457**	-16,423	**-15,534**	-19,829	**-19,443**	-12,378	**-11,162**	-21,455	**-21,179**
AIC	20,468	**16,956**	32,888	**31,109**	39,702	**38,930**	24,799	**22,368**	42,953	**42,402**
BIC	20,607	**17,095**	33,027	**31,248**	39,848	**39,076**	24,945	**22,514**	43,099	**42,548**

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
M1: Model 1, the baseline model for the analysis.
M2: Model 2, an extension of M1 that incorporates an offset to account for temporal exposure.
LL_{null}: Log-likelihood of the null model (model with only an intercept and offsets).
LL_{model}: Log-likelihood of the fitted model.
AIC: Akaike Information Criterion.
BIC: Bayesian Information Criterion.
Bold: The best statistic per analysis period.

Table 5 provides the relative gains achieved by *Model 2* compared to the baseline, *null Model 1*. The results reveal that temporal exposure contributes to the model performance, particularly during the first year of the pandemic. In 2022, this offset accounted for nearly half of the observed gains. However, its impact reduced in 2021 and over the cumulative period, suggesting that other factors played a more prominent role during these times.

Table 5 – Relative gain of *Model 2* compared to the baseline.

Statistic	2020 (first half)	2020	2021	2022	2020-2022
G(*Null Model 2*, *Null Model 1*)	0.22	0.08	0.02	0.08	0.01
G(*Fitted Model 2*, *Null Model 1*)	0.25	0.09	0.13	0.18	0.11

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
G(*Null Model 2*, *Null Model 1*): Relative gain of *null Model 2* over *null Model 1*.
G(*Fitted Model 2*, *Null Model 1*): Relative gain of *fitted Model 2* over *null Model 1*.

These results underscore the improved predictive capability of *Model 2*, so we focus our analysis on the factors correlated with COVID-19 deaths using the *Model 2*. For a detailed analysis of coefficient estimates and additional model statistics, refer to Appendix C.

4.2.1 Associations between sociodemographic clusters and mortality

Table 6 shows the observed COVID-19 death rates in Brazilian municipalities, grouped by sociodemographic clusters, across five different analysis periods. Initially, the clusters *Urbanized with Informal Settlements* and *Rural with Low Human Development* were the most impacted by the pandemic. By the end of 2020, the cluster *Urbanized with Informal Settlements* had the highest mortality rate for the year, while the initially high mortality rate in the *Rural with Low Human Development* cluster was matched by other clusters.

Table 6 – Mean COVID-19 death rates per 100,000 inhabitants (95% Confidence Intervals) for Brazilian municipalities, grouped by sociodemographic clusters, across five analysis periods.

Cluster	2020 (first half)	2020	2021	2022	2020-2022
Urbanized	9.24 (8.56, 9.92)	68.86 (66.9, 70.83)	**243.29** (239.23, 247.35)	41.82 (40.58, 43.07)	353.97 (348.7, 359.25)
Urbanized with informal settlements	47.19 (42.74, 51.64)	115.8 (110.31, 121.3)	**192.68** (184.22, 201.14)	27.89 (26.19, 29.6)	336.37 (323.17, 349.58)
Semi-urbanized	17.88 (16.83, 18.93)	63.41 (61.57, 65.26)	**123.63** (120.71, 126.54)	22.93 (22.13, 23.74)	209.97 (206.09, 213.85)
Rural with high human development	4.58 (3.72, 5.43)	52.26 (48.76, 55.77)	**193.64** (186.67, 200.62)	37.97 (35.51, 40.43)	283.88 (275.03, 292.72)
Rural with low human development	30.19 (27.0, 33.39)	64.4 (59.79, 69.01)	**83.94** (78.38, 89.51)	11.35 (10.2, 12.5)	159.69 (150.64, 168.75)

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
Bold: The highest COVID-19 death rate per cluster and year.

From 2021 onward, the most impacted clusters changed. The clusters *Urbanized*, *Urbanized with Informal Settlements*, and *Rural with High Human Development* had the highest death rates, whereas *Semi-urbanized* and *Rural with Low Human Development* had the lowest. Notably, the *Rural with Low Human Development* cluster had the lowest death rates over the cumulated period, despite initially presenting an elevated COVID-19 mortality.

The regression analysis provides deeper insights into the associations between sociodemographic clusters and COVID-19 mortality by controlling for correlated factors and temporal exposure. Table 7 displays the RR estimates obtained from *Model 2*. The *Urbanized* cluster showed the lowest RR at the pandemic onset. However, from 2021 onward, the *Urbanized* cluster exhibited the highest RR. By 2022, significant correlations were less apparent, with the only notable finding for the *Rural with Low Human Development* cluster, which had 13% fewer deaths than the reference cluster (*Urbanized*).

Table 7 – Estimated Rate Ratio (RR) (95% Confidence Intervals (CI)) of COVID-19 mortality for different clusters across five periods based on *Model 2*.

Cluster	2020 (first half)	2020	2021	2022	2020-2022
Urbanized	1 (Reference)	1 (Reference)	1 (Reference)	1 (Reference)	1 (Reference)
Urbanized with informal settlements	**1.34** (1.17, 1.54)	**1.15** (1.06, 1.25)	**0.87** (0.82, 0.92)	1.02 (0.95, 1.08)	**0.92** (0.87, 0.97)
Semi-urbanized	**1.88** (1.66, 2.12)	**1.31** (1.24, 1.39)	**0.86** (0.83, 0.90)	0.98 (0.92, 1.04)	**0.95** (0.92, 0.99)
Rural with high human development	**1.66** (1.35, 2.05)	**1.30** (1.20, 1.40)	0.96 (0.91, 1.00)	1.02 (0.94, 1.10)	1.01 (0.96, 1.05)
Rural with low human development	**1.85** (1.52, 2.25)	**1.28** (1.16, 1.42)	**0.75** (0.70, 0.81)	**0.87** (0.78, 0.98)	**0.87** (0.82, 0.93)

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
Bold: Statistical significance within the 95% CI.

When comparing the observed death rates in Table 6 with the estimated RRs in Table 7, we observe how *Model 2* adjusts for significant factors, offering a nuanced perspective on the relationship between sociodemographic clusters and COVID-19 mortality. For instance, although the *Rural with High Human Development* cluster reported a lower death rate than the *Urbanized* cluster in the first year of the pandemic, the model estimates that the *Rural with High Human Development* cluster experienced 30% more deaths after adjusting for covariates. Similarly, the model significantly reduced the association for the *Urbanized with Informal Settlements* cluster, which initially showed the highest death rate.

These findings underscore the complexities of analyzing sociodemographic correlations during the early pandemic phases. For the cumulative period, the observed death rates indicate that the *Urbanized with Informal Settlements*, *Semi-urbanized*, and *Rural with Low Human Development* clusters reported respectively 5%, 41%, and 55% fewer deaths compared to the *Urbanized* cluster after adjusting for covariates. In contrast, *Model 2* estimates these reductions to be 8%, 5%, and 13%, respectively, highlighting the magnitude of the adjustments introduced by the model.

4.2.2 Associations between sociodemographic, vaccination, and political variables and COVID-19 mortality

Table 8 presents the RR estimated by *Model 2*, which assesses the associations between sociodemographic, vaccination, and political variables and COVID-19 mortality across five analysis periods. None variable presented a significant correlation for all periods. However, we highlight that the variables *percentage of the population aged 60 years or more*, *percentage of the urban population*, *percentage of the Indigenous population*, *expected years of schooling at age 18*, and *percentage of votes for Bolsonaro* reported significant correlation for four of the five periods analyzed.

Table 8 – Rate Ratios (RR) (95% Confidence Intervals (CI)) of COVID-19 deaths by sociodemographic, vaccination, and political variables across five analysis periods, as estimated by *Model 2*.

Variable	2020 (first half)	2020	2021	2022	2020-2022
% population 60+ years	0.99 (0.94, 1.04)	**1.11** (1.08, 1.13)	**1.07** (1.05, 1.09)	**1.32** (1.29, 1.35)	**1.08** (1.07, 1.10)
% urban population	1.03 (0.97, 1.10)	**1.08** (1.05, 1.12)	**1.12** (1.09, 1.14)	**1.08** (1.05, 1.12)	**1.12** (1.10, 1.14)
Population density (inhabitants/km²)	**1.03** (1.01, 1.05)	1.00 (0.99, 1.02)	1.00 (0.99, 1.01)	0.99 (0.98, 1.00)	1.00 (0.99, 1.01)
% male population	**1.05** (1.01, 1.09)	1.00 (0.98, 1.02)	**1.03** (1.02, 1.04)	**1.05** (1.03, 1.07)	1.01 (1.00, 1.02)
% Indigenous population	**1.04** (1.02, 1.07)	**1.04** (1.03, 1.06)	**1.02** (1.01, 1.04)	1.02 (1.00, 1.04)	**1.03** (1.02, 1.04)
Gini coefficient	**0.94** (0.90, 0.98)	**0.96** (0.94, 0.98)	1.01 (0.99, 1.02)	0.99 (0.97, 1.01)	1.00 (0.99, 1.02)
% informal settlement households	**1.08** (1.06, 1.10)	**1.04** (1.02, 1.05)	1.00 (0.99, 1.01)	0.99 (0.97, 1.00)	1.01 (1.00, 1.02)
Population density in informal settlement (inhabitants/ha)	0.99 (0.97, 1.02)	1.00 (0.98, 1.01)	1.01 (1.00, 1.02)	1.00 (0.99, 1.01)	1.01 (1.00, 1.02)
% sanitation-related hospitalizations	1.01 (0.97, 1.04)	0.98 (0.97, 1.00)	1.00 (0.99, 1.02)	0.98 (0.96, 1.00)	0.99 (0.98, 1.00)
% self-employed workers	**1.06** (1.01, 1.11)	**1.03** (1.01, 1.06)	0.98 (0.97, 1.00)	**0.94** (0.91, 0.96)	1.00 (0.98, 1.01)
Unemployment rate	**1.22** (1.18, 1.27)	**1.04** (1.02, 1.06)	**0.98** (0.96, 0.99)	0.99 (0.97, 1.01)	1.00 (0.99, 1.01)
% commerce workers	**0.91** (0.87, 0.95)	**0.96** (0.94, 0.98)	0.99 (0.97, 1.00)	**0.95** (0.93, 0.98)	1.01 (0.99, 1.02)
% service workers	0.96 (0.91, 1.01)	**1.06** (1.04, 1.09)	**1.03** (1.02, 1.05)	1.00 (0.97, 1.02)	**1.03** (1.01, 1.05)
% industry workers	0.97 (0.92, 1.01)	1.02 (1.00, 1.05)	0.99 (0.97, 1.00)	**0.96** (0.94, 0.98)	1.00 (0.99, 1.01)
Expected years of schooling at age 18	**1.06** (1.02, 1.11)	1.00 (0.98, 1.02)	**1.06** (1.04, 1.07)	**1.04** (1.02, 1.06)	**1.03** (1.02, 1.04)
% people fully vaccinated	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	**1.08** (1.07, 1.10)	**1.07** (1.05, 1.09)	**1.07** (1.06, 1.08)
% votes for Bolsonaro	1.03 (0.98, 1.08)	**1.08** (1.05, 1.10)	**1.16** (1.14, 1.18)	**1.13** (1.10, 1.16)	**1.15** (1.14, 1.17)

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
Note: The variables were standardized.
Bold: Statistical significance within the 95% CI.

The *percentage of votes for Bolsonaro* exhibited the strongest association, with an RR of 1.15 (95% CI: 1.14–1.17) during the cumulative period (2020–2022). Among the sociodemographic covariates, the *percentage of the urban population* and *percentage of the population aged 60 years or more* presented the highest correlations for the cumulative period, with RR values of 1.12 (95% CI: 1.10–1.14) and 1.08 (95% CI: 1.07–1.10), respectively. Notably, the *percentage of people fully vaccinated* showed a significant positive correlation in the cumulative analysis (RR = 1.07, 95% CI: 1.06–1.08).

In the early pandemic, the first half of 2020, the *unemployment rate* showed the strongest association (RR = 1.22, 95% CI: 1.18–1.27). However, its influence reduced later in 2020 (RR = 1.04, 95% CI: 1.02–1.06) and became negatively correlated in 2021 (RR = 0.98, 95% CI: 0.96–0.99). Similarly, some variables like the *percentage of the male population*, *population density*, and *expected years of schooling at age 18* were positively correlated

in early 2020 but lost significance by the year-end. We also highlight that the *percentage of commerce workers* and the *Gini coefficient* presented negative correlations in 2020.

In 2021, variables that consistently correlated with COVID-19 mortality included the *percentage of votes for Bolsonaro*, the *percentage of the urban population*, the *percentage of the population aged 60 years or more*, and the *percentage of people fully vaccinated*. This pattern aligns closely with those observed for the cumulative period.

By 2022, the *percentage of the population aged 60 years or more* emerged as the most significant factor, with the highest RR in this study (RR = 1.32, 95% CI: 1.29–1.35). The *percentage of votes for Bolsonaro* and the *percentage of the urban population* continued to show strong positive associations.

Conversely, some variables lacked consistent significance. For instance, the *population density in informal settlements* and the *percentage of sanitation-related hospitalizations* showed no significant correlations across periods. Others, such as the *percentage of self-employed workers*, demonstrated period-specific effects, with positive correlations in 2020 and negative correlations in 2022. The *percentage of informal settlement households* presented a significant correlation only in 2020. On the other hand, the *percentage of industry workers* showed a significant negative correlation only in 2022.

4.2.3 Sensitivity analysis

The sensitivity analysis, detailed in Appendix D, supports the robustness of our findings. The analysis of outliers and influential points revealed that their removal enhanced the goodness-of-fit statistics of the *Model 2*. However, this adjustment did not significantly alter the estimated coefficients.

4.3 Discussion

To answer our first research question ("*What profiles of Brazilian municipalities correlate most strongly with COVID-19 mortality?*"), we employed a GMM clustering in which we identified five distinct sociodemographic clusters: *Urbanized*, *Urbanized with Informal Settlements*, *Semi-urbanized*, *Rural with High Human Development*, and *Rural with Low Human Development*. These clusters describe the profiles of the Brazilian municipalities and denote socioeconomic and regional disparities nationwide.

Given this, our findings show that, initially, the *Semi-urbanized* and *Rural with Low Human Development* clusters exhibited the highest positive correlations with COVID-19 mortality, highlighting the differentiated impacts of the disease in the North and Northeast regions in early pandemic. However, since 2021, a notable shift occurred: the *Urbanized*

cluster become significantly more correlated with the mortality than others, surpassing the cumulative mortality rates and RRs of all other clusters. These findings align with the mortality colormap in Figure 6e and the regional dispersion of clusters in Figure 22.

Temporal exposure emerged as a pivotal factor in our analysis, particularly in 2020 and 2022. Our findings denote that this offset in 2020 influenced more *Model 2* than the explanatory variables. Early in 2020, as illustrated in Figure 6a, the initial wave of COVID-19 was restricted to specific regions. As the disease spread more uniformly across the country in 2021, the influence of temporal exposure diminished. Interestingly, the strength of the temporal exposure resurged in 2022. This finding highlights the critical role of accounting for the outbreak onset in a municipality when analyzing COVID-19 mortality within specific timeframes, especially in the early pandemic or when the death events are less frequent.

Additionally, the *Urbanized with Informal Settlements* cluster reported the second-highest COVID-19 mortality rate in the country for the accumulated period. However, our model estimates RRs for this cluster overlapping the CI of *Semi-urbanized* and *Rural with Low Human Development*, the clusters with the lowest death rates. So, our model controlled the correlation for *Urbanized with Informal Settlements* with more intensity, which denotes that the interaction between this cluster and other variables in the model reduced the RR estimated for this cluster. Further, many municipalities from the *Urbanized with Informal Settlements* cluster are in metropolitan areas, which were initial gateways for the coronavirus in Brazil, where temporal exposure plays a significant role.

Discussing our second research question ("*What sociodemographic factors are associated with COVID-19 mortality across Brazilian municipalities?*"), we note that urbanization is a critical factor in this study. Our analysis identifies the *urban population percentage* as a robust indicator correlating with COVID-19 mortality rates within municipalities. Other variables associated with urban living, such as the *percentage of service workers* and *expected years of schooling at age 18*, also showed positive correlations. Our empirical findings align with the insights presented by Keil (2021) that our increasingly urbanized world explains the rapid global dissemination and impact of COVID-19.

Barbosa et al. (2020) and Ziyadidegan et al. (2022) indicate that, early in the pandemic, areas with higher proportions of elderly populations had lower mortality rates. Our study presents a non-significant correlation for this factor in the pandemic onset. So, we analyzed the relationship between the proportion of elderly residents and COVID-19 mortality across the time. Supplementary Figure 46 (Appendix E) shows that the sociodemographic clusters reported either a negative or non-significant correlation between the *percentage of the population aged 60 years or more* and COVID-19 mortality in the early pandemic period. These observations coincide with the period when Brazilian municipalities implemented stringent measures to reduce human mobility in response to COVID-19, see Section 3.3.1. We hypothesize that increased care and protection of

older individuals during the early stages of the pandemic in Brazil could explain this behavior. However, further investigation is needed to determine whether municipalities with larger elderly populations implemented more pronounced mobility reduction measures than others at the pandemic onset.

Another notable finding regarding the elderly population is the elevated RR of 1.32 (95% CI: 1.29–1.35) estimated for 2022, which is substantially higher than the RR values of 1.11 (95% CI: 1.08–1.13) and 1.07 (95% CI: 1.05–1.09) estimated for 2020 and 2021, respectively. While the exact cause of this increase remains uncertain, we hypothesize that the vulnerabilities associated with aging and COVID-19 (Dessie and Zewotir, 2021; Rod et al., 2020) became more pronounced as younger individuals received vaccinations in urbanized areas. This shift may have left elderly populations comparatively more susceptible in 2022.

We confirmed the expectation that cities with a significant presence of *informal settlements* were associated with higher COVID-19 mortality, but only during the first year. Similarly, our findings for *population density* show a significant positive correlation only during the first half of 2020. We believe that a more detailed investigation of *population density* is warranted, potentially requiring analysis at the census tract level to uncover more nuanced associations.

Our study supports concerns about COVID-19 impacts on the Indigenous populations in Brazil (Santos et al., 2020; Croda et al., 2022). We found an RR of 1.03 (95% CI: 1.02 - 1.04) for the *percentage of Indigenous population*. Similarly, the *percentage of the male population* correlates for the early moment in 2020, 2021, and 2022, consistent with existing literature (Gebhard et al., 2020; Jin et al., 2020; Clouston et al., 2021a; Unruh et al., 2022).

In Brazil, several ecological studies have examined the relationship between income inequality and the impact of COVID-19. The *Gini coefficient* has often been identified as a factor associated with COVID-19 outcomes, with findings reported for different time points in the early moment: July 2020 (Martines et al., 2021; Demenech et al., 2020), August 2020 (Figueiredo et al., 2020), and September 2020 (Raymundo et al., 2021). Consistent with these studies, we demonstrate in Appendix E that uncontrolled correlation analyses for the same period indicate a positive association. However, our negative binomial regression reveals a negative correlation between the *Gini coefficient* and COVID-19 mortality during the first pandemic year. This finding highlights the importance of the selected control factors in providing a more nuanced understanding of the complex relationship between income inequality and COVID-19 mortality. Furthermore, from 2021 onward, our *Model 2* estimates no significant correlations for this variable, while uncontrolled analysis in Appendix E suggests a negative relationship.

We note a discrepancy between the perceived vulnerability of regions to COVID-19 at the pandemic onset and the outcomes observed three years later. Models proposed in Brazil to identify the municipalities most vulnerable to COVID-19, such as the Municipal Vulnerability Index - COVID-19 by the Votorantim Institute[1], the Municipal Vulnerability Index to the spread of the Coronavirus by the Perseu Abramo Foundation[2], and Coelho et al. (2020), highlighted the risks in municipalities from the North and Northeast of the country due to their significant socioeconomic vulnerabilities and limited health infrastructure. However, by the end of the first three years of the pandemic, the Southeast, the Midwestern, and the South regions experienced higher mortality rates than the North and Northeast regions.

Despite this, we argue that prioritizing the social and economic challenges faced by poorer populations was justified. In the early stages of the pandemic, variables reflecting these challenges, such as the *unemployment rate*, *percentage of informal settlement households*, and *percentage of self-employed workers*, demonstrated the strongest correlations with COVID-19 deaths. We comprehend that the value of our work is in providing a nuanced understanding that the factors associated with COVID-19 mortality in Brazilian municipalities changed over time, being that from the end of 2020 onward, other factors, including political preference, urban population, and elderly population, consistently showed more pronounced correlations with COVID-19 mortality.

Our findings not only answer our third research question ("*Is the political preference manifested by the municipal populations correlated with COVID-19 mortality?*") as highlight that, among the factors analyzed, the *percentage of votes for Bolsonaro* in the 2022 Presidential Election exhibits the strongest correlation with COVID-19 mortality rates in Brazilian municipalities. This result aligns with previous studies that have identified a link between political preferences and COVID-19 outcomes at the municipal level (Lima et al., 2024a; Xavier et al., 2022b; Hastenreiter Filho and Cavalcante, 2022). However, our study extends these earlier works by incorporating a broader range of explanatory variables and covering a more extended period, whereas prior research only analyzed data up to June 2021 (Lima et al., 2024a; Xavier et al., 2022b; Hastenreiter Filho and Cavalcante, 2022). Additionally, our approach improves upon previous studies by introducing control for temporal exposure.

Another variable that positively correlated with COVID-19 mortality is the *percentage of people fully vaccinated*, replying to our fourth question ("*Did municipalities with higher COVID-19 vaccination coverage experience lower mortality rates?*"). This result challenges the hypothesis that regions with higher vaccination efforts would experience lower death rates. However, it aligns with an alternative hypothesis that areas with higher death rates may have motivated greater engagement in vaccination than regions with milder outbreaks. Several other factors could influence this correlation, including disparities in vaccine

[1] https://institutovotorantim.org.br/ivm/2020/
[2] https://fpabramo.org.br/2020/04/16/estudo-ranqueia-municipios-mais-vulneraveis-ao-coronavirus/

access and distribution across Brazil. Additionally, potential inconsistencies in vaccination data, as discussed in Section 3.3.2, further complicate the interpretation of this relationship.

4.3.1 Related work

Various studies have employed clustering algorithms or regression models to analyze the COVID-19 pandemic, revealing distinct patterns of impact across different countries (Afzal et al., 2021; Erandathi et al., 2021; Meng, 2021; Brida et al., 2021; Zarikas et al., 2020; Gohari et al., 2022; Rizvi et al., 2021; Carrillo-Larco and Castillo-Cara, 2020; Awuah-Mensah and Aidoo, 2024). Additionally, investigations at more localized levels have explored the pandemic effects in specific regions, such as counties in the United States (Vahabi et al., 2021; Ziyadidegan et al., 2022; Maleki et al., 2022; Nicholson et al., 2022; Clouston et al., 2021b), municipalities in Mexico (Pérez-Ortega et al., 2022), municipalities in Italy (De Angelis et al., 2021), districts in Germany (Hoebel et al., 2021), prefectures in Japan (Yoshikawa and Kawachi, 2021), and basic health areas in Catalonia (López-Bazo, 2024).

In the Brazilian context, an ecological study examined data up to October 2020, revealing an association between elevated social vulnerability and COVID-19 mortality until June 2020 (Rocha et al., 2021). This study reported a shift where municipalities with lower social vulnerabilities experienced higher mortality rates in subsequent periods (Rocha et al., 2021), aligning with our findings. Another ecological study (Castro-Alves et al., 2022) focused on the first year of the pandemic in Brazil and found a correlation between well-urbanized municipalities and higher death rates, a finding also observed in our work. Furthermore, Bermudi et al. (2021) analyzed administrative districts in São Paulo, identifying a shift in high-risk areas from those with the best socioeconomic conditions to those with the worst conditions during the first semester of 2020, denoting that enhanced granularity studies reveal more nuance and complex correlations.

To the best of our knowledge, our study is the first to apply regression models to analyze the correlation between sociodemographic clusters and variables, political preferences, vaccination coverage, and temporal exposure to COVID-19 mortality across Brazilian municipalities over a long-term period.

4.3.2 Limitations

We acknowledge limitations in the study presented in this chapter that warrant consideration. First, due to the unavailability of a fully updated 2022 Census, some of the sociodemographic data used in our analysis are based on the 2010 Census (IBGE, 2010). Consequently, municipalities that have experienced substantial changes in their sociodemographic profiles over the past 14 years may introduce biases into the results.

Additionally, the vaccination data employed in this study exhibit apparent inconsistencies, as detailed in Section 3.3.2. Therefore, conclusions regarding the correlation between vaccination and COVID-19 mortality should be approached carefully.

Furthermore, we did not find municipal-level databases on the population's health conditions, meaning these important correlated factors were absent from our analysis (Richardson et al., 2020), which may have influenced our findings. Finally, it is important to emphasize that the ecological associations reported in this chapter do not imply causation (Szklo and Nieto, 2014).

5 A COVID-19 MODEL WITH FUZZY TRANSITIONS BETWEEN EPIDEMIC PERIODS: A NATIONAL CASE STUDY

In this chapter, we aim to comprehensively reproduce the COVID-19 pandemic in Brazil over the initial three years, providing insights into the actual scale, including estimates for underreported cases. We refer to underreported cases as the number of infections estimated to have occurred in a population not captured by the surveillance system over a given period. To achieve this, we implemented a modified Susceptible, Infected, Recovered, Dead, Susceptible (SIRDS) model, allowing us to infer changes in key epidemiological parameters: basic reproduction number (R_0), Infection Fatality Rate (IFR), protected period (also named immunity period), and underreporting of cases. Our model introduces innovation by incorporating time-varying parameters through fuzzy transitions between epidemic periods.

We conducted a sensitivity analysis to validate our model and applied it to data from countries beyond Brazil, ensuring its generalization for Spain, the United Kingdom, and the United States. Regarding data from the first wave, we contextualized our findings with studies that estimated prevalence (Hallal et al., 2020; Pérez-Gómez et al., 2023; Public Health England, 2020c,a,b; Walker et al., 2021; Anand et al., 2020) and IFR (Marra and Quartin, 2021; Picon et al., 2020; Silveira et al., 2020) based in serological research. However, serological surveys have limitations in analyzing epidemics with multiple outbreaks (Gibbons et al., 2014), making epidemiological modeling crucial to estimating the true magnitude of prolonged epidemics.

In this regard, our work relates to other studies that have presented time-varying models to reproduce changes in COVID-19 epidemiological parameters across outbreaks, considering factors like human mobility patterns, emerging variants, and vaccination (Liu et al., 2022; Yang and Shaman, 2022; Ferrante et al., 2022; Ghosh and Ghosh, 2022; Xu and Tang, 2021; Romanescu et al., 2023; Nouvellet et al., 2021; Ribeiro Xavier et al., 2022; Abolpour et al., 2021; Vasconcelos et al., 2021, 2023). For example, Vasconcelos et al. (2023) analyzed COVID-19 mortality trends in Brazil through March 3, 2022, using a multistep logistic-like function to model the waves. While methodologically distinct, our study complements this work by providing a broader examination of the first three years of the pandemic in Brazil and offering additional insights into the temporal dynamics of COVID-19 mortality.

We describe the methodology used in this chapter in Section 5.1. Sections 5.2 and 5.3 present the results and provide a discussion, including a detailed review of related studies. We also published the findings of this chapter in PLOS ONE (Lima et al., 2024c).

5.1 Methodology

5.1.1 Data

Our comprehensive analysis of the COVID-19 pandemic in Brazil spans the first three years, from 2020 to 2022. For estimating the effective reproduction number (R_t), we utilized case data from the SARS database (DATASUS, 2022a), as outlined in Section 3.2.2. Death data from the Mortality Information System (SIM) (DATASUS, 2022b), see Sections 3.1.1 and 3.2, were used to fit the model proposed in this work. Additionally, case data from the Monitoring Panel (DATASUS, 2020a) were employed to estimate underreporting. To determine the population size of Brazil, we used the 2022 Demographic Census (IBGE, 2022) provided by the Brazilian Institute of Geography and Statistics (IBGE).

For broader applicability, we utilized COVID-19 data from Our World in Data (Ritchie et al., 2020) for Spain, the United Kingdom, and the United States. This dataset was employed to illustrate the generalization of our model, as presented in Section 5.1.3.2. Appendix F provides charts with time series for these countries.

5.1.2 SIRDS model with fuzzy epidemic period transitions

In this chapter, we introduce an epidemiological model that incorporates time-varying parameters and leverages fuzzy theory (Zadeh, 1965; Ross, 2005) to enhance the smoothness of transitions between epidemic periods. To propose this model, we regard the critical components of the SIRDS model, as presented in Section 2.2.1. Section 5.1.2.1 outlines the underlying assumptions that guided our decision to propose a model with time-varying parameters and fuzzy transitions. We present the implementation of the model in Section 5.1.2.2. Furthermore, Section 5.1.2.3 delves into the presentation of the objective function and the optimization method employed in our study.

5.1.2.1 Assumptions

In this study, we propose that epidemiological parameters change during a pandemic due to modifications in mobility patterns, mutations in virus properties (such as the emergence of new variants), and fluctuations in vaccination coverage.

We conducted an extensive set of 76,650 epidemic simulations spanning three years to validate these assumptions. These simulations covered a range of values for R_0 (1 to 8). They included parameters such as an infectious period of eight days, IFR of 1%, a protective immunity period of one year, and an initial count of 0.14 infected individuals in a population of 100,000. Figure 23 shows some simulations and illustrates the dynamic patterns of R_t.

In our simulations with a static R_0, we observed a pattern where each outbreak showed a lower R_t peak than its precursor. This sequential drop in R_t peaks continued until reaching an endemic equilibrium, as shown in Fig 23b. While the simulations with static R_0 showed a consistent drop in R_t peaks, aligning with an endemic equilibrium, real-world COVID-19 outbreaks showed a distinct pattern. The dynamics of actual outbreaks did not conform to the consistent drop seen in simulations, suggesting that factors like changes in mobility patterns (Arroyo-Marioli et al., 2021; Nouvellet et al., 2021; Nishiura and Chowell, 2009; Reis et al., 2020) and the emergence of new variants (Liu and Rocklöv, 2021, 2022) contribute to changes in virus transmissibility.

We also noted a pattern in the initial outbreak, where R_t starts at the peak and then declines following (2.9). However, the initial outbreak coincided with a significant reduction in population mobility during the COVID-19 pandemic. Thus, we hypothesize that the initial outbreak began with a transmission pattern that changed for many places even during the first outbreak. We conducted 18,625,950 simulations to investigate, reducing R_0 by 10% to 50% at different points during the first outbreak.

We conducted a comparative analysis of the initial outbreak simulations, comparing those with variable R_0 and their counterparts in simulations with constant R_0. To assess the similarities between these scenarios, we employed the FastDTW algorithm (Salvador and Chan, 2007), a heuristic algorithm designed as an efficient alternative to the Dynamic Time Warping (DTW) algorithm. This approach produces a distance measurement for evaluating the similarity between two-time series (Salvador and Chan, 2007). Figugre 24a shows the assessed similarities, with the lowest DTW similarity recorded at 0.215.

Moving beyond the initial outbreak, we noted that the R_t time series of the subsequent outbreaks present a curve similar to a bell-shaped one as shown in Figure 23. To quantify the similarity between the left and right sides of these curves, we utilized the equation:

$$\text{similarity} = \frac{\sum_{i=p}^{e} R_t(i) - \sum_{i=b}^{p} R_t(i)}{\sum_{i=b}^{e} R_t(i)}, \tag{5.1}$$

where b is the start point, p is the R_t peak, and e is the end point of an outbreak. In this equation, when similarity > 0, the right side is higher than the left side; when similarity < 0, the left side is higher than the right side; and when similarity $= 0$, the left and right sides there is the same sum of R_t. The similarities for simulations without R_0 changes are shown in Figure 24b, with the similarity between sides ranging between -0.39 and 0.22.

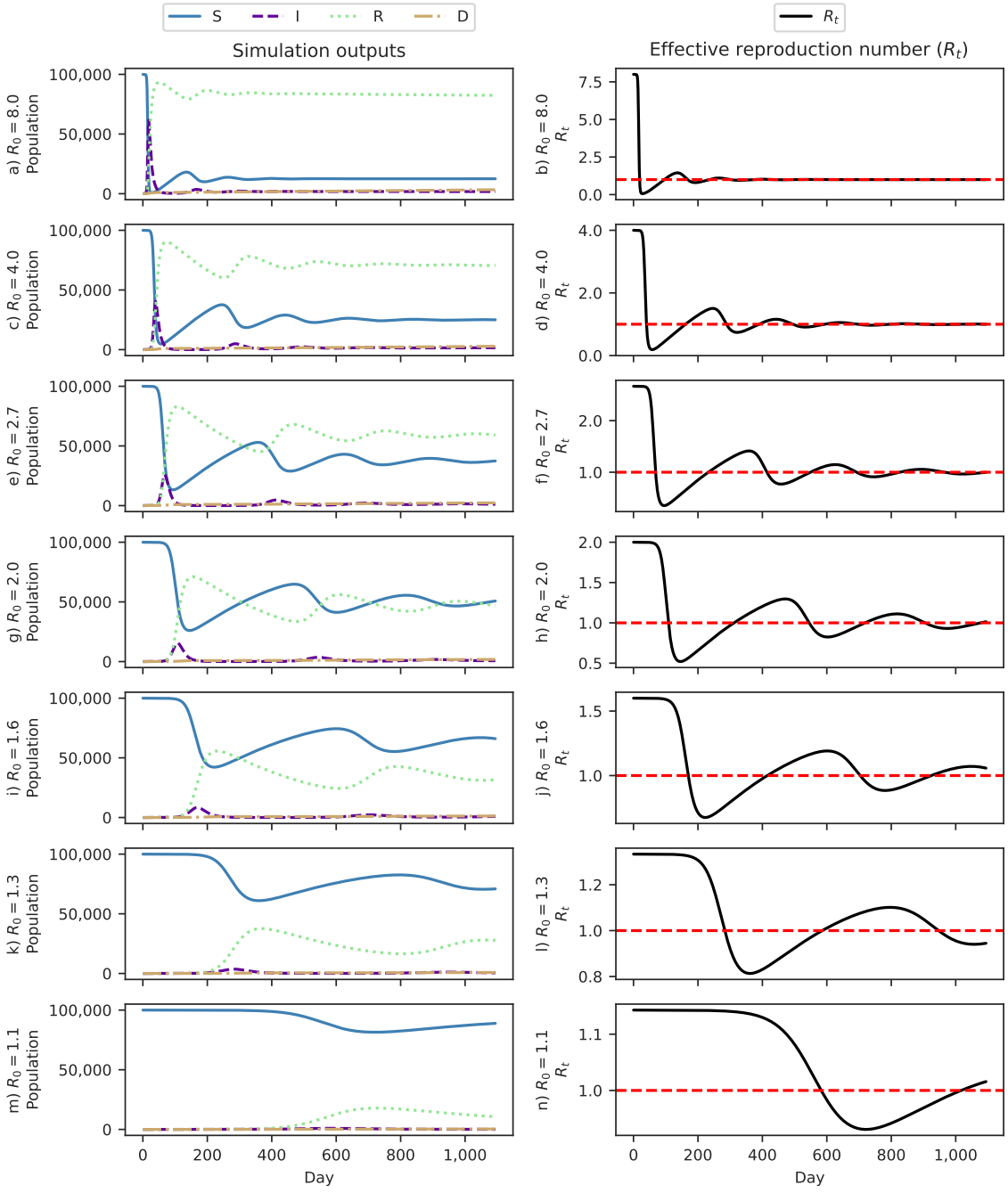


Figure 23 – SIRDS model simulations across three years for different basic reproduction numbers (R_0). Each row corresponds to a specific R_0 value. The charts on the left depict simulation outputs for the Susceptible (S), Infected (I), Recovered (R), and Deceased (D) compartments. On the right side, the charts display the observed effective reproduction number (R_t) over time, with a dashed horizontal line at the reference value ($R_t = 1$) used for epidemic monitoring.

We noted a decline in the lethality across the pandemic. Section 3.2.4 and Appendix G show a CFR drop from the early stages to the end of the study period. We recognize the impact of underreporting on this measurement, mainly during pandemic peaks. However, our observation suggests a drop in the lethality throughout the pandemic.

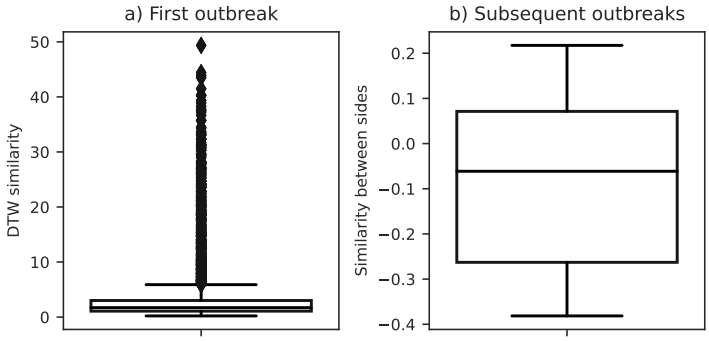


Figure 24 – Comparative boxplots of effective reproduction number (R_t) similarity distributions in synthetic SIRDS outbreaks. (a) Dynamic Time Warping (DTW) similarity for the first outbreak, contrasting synthetic samples with a change in basic reproduction number (R_0) to their counterparts in synthetic samples without changing the R_0. (b) Similarity between left and right sides for subsequent outbreaks in synthetic samples without changing the R_0.

Our study assumes that parameters vary during different epidemic periods. Specifically, the change between periods for the *contact rate* (β) tends to be rapid, while that for the *infection fatality probability* (f) is gradual. Although evidence about the transition between periods for *immunity loss* is lacking, our hypothesis suggests a slow shift similar to f.

5.1.2.2 Model implementation

In this section, we describe the implementation details of our SIRDS model with fuzzy epidemic period transitions, building upon the epidemiological concepts introduced in Section 2.2.1 and the assumptions outlined in Section 5.1.2.1.

Our model incorporates the following parameters:

- Initial infected population ($I(0)$),
- Recovery rate (γ),
- List representing contact rates for different epidemic periods ($\overrightarrow{\beta}$),
- List representing infection fatality probabilities for different epidemic periods ($\overrightarrow{f}$),
- List representing immunity loss rates for different epidemic periods ($\overrightarrow{\omega}$),
- List denoting breakpoints for fast transition between epidemic periods ($\overrightarrow{b_{fast}}$),
- List representing transition days for smoothing fast transitions ($\overrightarrow{\tau_{fast}}$),
- List denoting breakpoints for slow transition between epidemic periods ($\overrightarrow{b_{slow}}$).

It is essential that the minimum size of the parameters $\overrightarrow{\beta}$, $\overrightarrow{f}$, and $\overrightarrow{\omega}$ be one. The parameters $\overrightarrow{b_{fast}}$ and $\overrightarrow{\tau_{fast}}$ are of the same length, both having one item less than $\overrightarrow{\beta}$. Similarly, $\overrightarrow{f}$ and $\overrightarrow{\omega}$ have the same size, both having one item more than $\overrightarrow{b_{slow}}$.

Our model instantiates a fuzzy variable to represent a fast transition (μ_{fast}) between epidemic periods. The universe is the number of days in simulation. A fuzzy partition in this variable represents each epidemic period. These partitions are trapezoidal shapes whose peak, i.e. $\mu_{\text{fast}} = 1$, extends from the correspondent breakpoint to the adjacent one. For each partition, the beginning is advanced by the correspondent transition days considering its peak beginning, and the end is delayed by the adjacent transition days considering its peak end. Thus, the total number of fuzzy partitions is $|\overrightarrow{b_{\text{fast}}}| + 1$.

Our model also instantiates a fuzzy variable to represent a slow transition (μ_{slow}) between epidemic periods. The universe also is the number of days in simulation. A fuzzy partition in this variable represents each epidemic period. These partitions are triangular shapes, beginning at the previous breakpoint, reaching the correspondent breakpoint at the top, i.e. $\mu_{\text{slow}} = 1$, and ending at the next breakpoint. The exception is the last epidemic period, where there is a trapezoidal function with a shape starting at the previous breakpoint and reaching its peak from the correspondent breakpoint to the boundaries of the universe. Thus, the total number of fuzzy partitions is $|\overrightarrow{b_{\text{slow}}}| + 1$.

In our model, the fuzzy variables operate with parameters β, f, and ω to reproduce different epidemic periods. Equation (5.2) defines the inference mechanism for defuzzification at day t for a time-varying epidemic parameter (θ') using a pair of a fuzzy variable (μ) with n partitions and a list of epidemic parameters ($\overrightarrow{\theta}$) also with n items.

$$\theta'(\overrightarrow{\theta}, \mu, t) = \frac{\sum_{i=0}^{n-1}(\mu_i(t) \times \overrightarrow{\theta_i})}{\sum_{i=0}^{n-1}(\mu_i(t))}. \tag{5.2}$$

The algorithm for the SIRDS model with fuzzy epidemic period transitions is summarized in Algorithm 1. The simulation function takes various parameters and computes the rates of change for susceptible (S), infected (I), recovered (R), and deceased (D) populations at a given time step t.

This algorithm provides a comprehensive view of the model dynamics, capturing the influence of fuzzy variables on epidemic parameters and their impact on the SIRDS compartmental model.

Algorithm 1 SIRDS model with fuzzy epidemic period transitions.

procedure SIMULATION(t, S, I, R, D, γ, $\overrightarrow{\beta}$, $\overrightarrow{f}$, $\overrightarrow{\omega}$, μ_{fast}, μ_{slow})
  $N \leftarrow S + I + R + D$
  $\beta \leftarrow \theta'(\overrightarrow{\beta}, \mu_{fast}, t)$
  $f \leftarrow \theta'(\overrightarrow{f}, \mu_{slow}, t)$
  $\omega \leftarrow \theta'(\overrightarrow{\omega}, \mu_{slow}, t)$
  $dS/dt = -\beta IS/N + \omega R$
  $dI/dt = \beta IS/N - \gamma I$
  $dR/dt = (1-f)\gamma I - \omega R$
  $dD/dt = \gamma f I$
  return dS/dt, dI/dt, dR/dt, dD/dt

5.1.2.3 Parameter optimization

We employed the stochastic differential evolution algorithm (Storn and Price, 1997) from the Python *SciPy* library for effective model parameter fitting. The objective was to minimize the discrepancies between original data and simulations for both the *death rate per 100,000 inhabitants in the 7-day moving average* (M) (DATASUS, 2022b) and the *effective reproduction number* (R_t) calculated in Section 3.2.2. To achieve this, we formulated a composite objective function using:

$$\frac{\text{MAE}(M, \hat{M})}{\overline{M}} + \frac{\text{MAE}(R_t, \hat{R}_t)}{\overline{R_t}}, \tag{5.3}$$

where Mean Absolute Error (MAE) is defined by (5.4). Here, $\hat{M}$ and $\hat{R}_t$ are the estimated mortality and effective reproduction numbers from our model, respectively. $\overline{M}$ and $\overline{R_t}$ represent the mean values of M and R_t, respectively.

$$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{5.4}$$

We set the stochastic differential evolution algorithm with a maximum of 10,000 generations, a multiplier factor of five (so, the total population size is five times the quantity of model parameters), an update strategy for the best solution vector once per generation, the mutation strategy 'best1bin', a relative tolerance parameter of 0.01, an absolute tolerance parameter of 0 (a stringent convergence criterion), a mutation parameter range of (0.5, 1), a recombination parameter of 0.7, the Latin hypercube initialization strategy, and an additional local optimization step using the L-BFGS-B method after the global optimization process.

The input parameters for our model vary with the number of outbreaks. We first identify the outbreaks following Section 3.2.3. For each outbreak after the initial one, we define a breakpoint b_{fast} to represent fast transitions between epidemic periods. We estimate R_0 and assess the similarity between the first outbreak R_t and that estimated from its corresponding R_0. If the similarity is higher than 0.22, we add an adjusted b_{fast} to account for changes in disease transmissibility across the outbreak, as explained in Section 5.1.2.1. We also assess the similarity between sides of outbreak curves for the subsequent outbreaks. If not within the range (-0.39, 0.22), we add an adjusted b_{fast} to capture changes in disease transmissibility across the outbreak, again, as presented in Section 5.1.2.1. We define an initial β and one β for each b_{fast}, along with one τ for each b_{fast}.

For representing a slow transition between epidemic periods, we define one breakpoint b_{slow} for each outbreak after the initial one, considering that the interval between outbreaks is at least 180 days. Therefore, in practical terms, is b_{slow} a subset of b_{fast}. We set an initial f and ω, along with one f and ω for each b_{slow}.

We optimize our model considering parameter bounds presented in Table 9, a population of 100,000 individuals, and the simulation period beginning at the start of the case time series. Empirically, we observed that our model performs better when γ is static, so we did not optimize this parameter using the stochastic differential evolution method, as detailed in Section 5.1.3.1. Additionally, it is crucial to highlight that $b_{slow} \subset b_{fast}$, and optimization of these parameters is unnecessary, as b_{fast} has already been optimized.

5.1.3 Experiments

5.1.3.1 Fitting the recovery period

By the insights from Voinsky et al. (2020), indicating a COVID-19 infectious period spanning from 8 to 20 days, we assessed these infectious periods within our model. To estimate the recovery rate (γ), we considered a range of values: $\gamma \in \left\{\frac{1}{8}, \frac{1}{9}, \frac{1}{10}, ..., \frac{1}{20}\right\}$. We conducted 20 model executions for each γ using national data from Brazil. The maximum bound for the parameter initial quantity of infected population ($I(0)$) was set as the sum of case rates per 100,000 inhabitants until the 14^{th} day in the first outbreak (i.e., 0.043).

The outcomes, illustrated in Figure 25a, underscore the optimal performance achieved by our proposed model when the *infectious period* is configured to eight days, corresponding to $\gamma = 1/8$, which we observed error of 0.165 (95% CI: 0.162 - 0.170). Furthermore, Figs 25b and 25c suggest that adopting an infectious period of 8 days incurs a slightly higher optimization cost. Importantly, this incremental cost is well-justified, contributing to refining and enhancing the model outcome. Therefore, all subsequent experiments in this thesis were conducted with $\gamma = 1/8$.

Table 9 – Model parameter bounds for optimization.

Parameter	Description	Minimum value	Maximum value	Reference
$I(0)$	Initial quantity of infected population.	$\frac{1}{\text{population}} \times 100,000$	Empirically defined	Empirical
b^0_{fast}	Adjusted breakpoint in initial outbreak for fast transition between epidemic periods.	Outbreak begin	Outbreak end	Empirical
b_{fast}	Breakpoint for fast transition between epidemic periods.	Outbreak begin	At $R_t^{\uparrow}$	Empirical
b'_{fast}	Adjusted breakpoint in subsequent outbreaks for fast transition between epidemic periods.	At $R_t^{\uparrow}$	Outbreak end	Empirical
τ	Transition days for smoothing fast transitions between epidemic periods.	0	56	Empirical
β_0	Initial contact rate.	$\gamma\overline{R_t}$	γR_0	Empirical
β'_0	Adjusted contact rate in initial outbreak.	$\gamma R_t^{Q_{0.25}}$	γR_0	Empirical
β	Contact rate.	$\max(\inf \beta_{b-1}, \gamma R_t^{\uparrow})$	$\gamma R_t^{\uparrow}/0.3$	Empirical
β'	Adjusted contact rate.	$\gamma R_t^{Q_{0.25}}$	$\gamma R_t^{\uparrow}/0.3$	Empirical
f	IFR in probability.	$\max(\frac{M}{100,000}, 0.0001)$	$\min(\text{CFR}, 0.0133)$	Verity et al. (2020)
ω	Immunity loss rate.	$1/365$	$1/90$	Pulliam et al. (2022)

$R_t^{\uparrow}$: Peak of the effective reproduction number in outbreak.
$\overline{R_t}$: Mean of the effective reproduction number in outbreak.
R_0: Basic reproduction number.
$R_t^{Q_{0.25}}$: First quartile of the effective reproduction number in outbreak.
γ: Recovery rate.
$\inf \beta_{b-1}$: Minimum bound of the previous contact rate.
IFR: Infection Fatality Rate.
CFR: Case Fatality Rate (in probability) in the epidemic period.
M: Death rate per 100,000 inhabitants in the epidemic period.

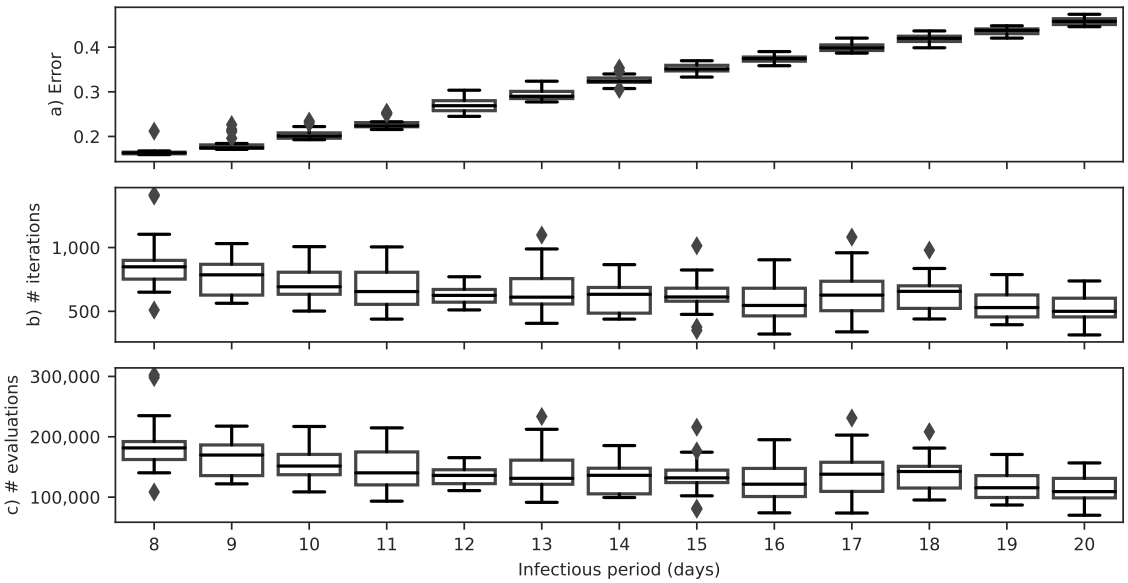


Figure 25 – Boxplots showing key metrics of model optimization for infection periods of 8–20 days: (a) objective function error, (b) iterations, and (c) function evaluations. Boxes represent the first and third quartiles, with the median as a horizontal line and whiskers extending to 1.5 times the interquartile range.

In Section 5.2, we succinctly present the outcomes for the optimal infection period. We evaluate the effectiveness of our model using two key metrics: the Mean Squared Error (MSE) and the coefficient of determination (R^2), defined by (5.5) and (5.6), respectively.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{5.5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{5.6}$$

5.1.3.2 Model application to other countries

To apply our model to data from Spain, the United Kingdom, and the United States, we first estimated the R_t time series using the method outlined in Section 3.2.2. However, adjustments were necessary for the input data of these countries. Specifically, we estimated the onset of symptoms from the new death time series, considering a delay of 19 days between onset and death (Verity et al., 2020). The Appendix H shows charts for R_t series of these places.

We conducted 20 pandemic simulations for each country, assuming an infectious period of eight days ($\gamma = 1/8$). In the cases of Spain and the United States, we set the maximum bound for the parameter $I(0)$ as the sum of case rates per 100,000 inhabitants until the beginning of the first outbreak time series, resulting in 0.0084 and 0.0068, respectively. On the other hand, we set the maximum bound for the United Kingdom for the parameter $I(0)$ to the same value as the minimum bound for this parameter, as presented in Table 9, and set the simulation period began at the peak date of R_t within the initial outbreak.

5.1.3.3 Parameter sensitivity assessment

To evaluate the local sensitivity of our model parameters, we employed the One-Parameter-at-a-Time (OAT) method (Hamby, 1994) using the simulations for Brazil data. The mean values estimated for the parameters in Section 5.1.3.1 served as the baseline. For each parameter, we perturbed its value by 1%, 10%, and 50%, keeping the others fixed.

As defined in (5.7), the absolute elasticity measure was utilized to assess parameter sensitivity, regarding the objective function (5.3) as the output. An absolute elasticity greater than one indicates that the parameter is elastic, meaning higher sensitivity. In other words, a change in the parameter leads to a proportionally more significant change in the output. Conversely, an absolute elasticity of less than one suggests that the parameter is not elastic, indicating lower sensitivity.

$$\text{absolute elasticity} = \left| \frac{\%\text{ change in output}}{\%\text{ change in input}} \right|. \tag{5.7}$$

5.1.3.4 Estimating underreporting factor

The underreporting factor, estimated using (5.8), represents the ratio between estimated infections by a model and reported cases by health authorities. It is essential to clarify that estimated infections encompass not only the reported cases but also the unreported cases and potentially any cases overreported by the health authorities.

$$\text{underreporting factor} = \frac{\text{estimated infections}}{\text{reported cases}}. \tag{5.8}$$

5.2 Results

5.2.1 Retrospective analysis

In this analysis, we focus on the results obtained using our model with an eight-day recovery period ($\gamma = 1/8$) applied to COVID-19 data from Brazil. Fig 26 provides a comprehensive overview of our model outcomes, including comparisons with the original data for various aspects such as the R_t, new cases, new deaths, total cases, and total deaths.

Table 10 presents the evaluation of the model performance, demonstrating a low MSE and high R^2 for both R_t and new deaths, indicating a well-fitted model.

Table 10 – Results for COVID-19 simulation with data from Brazil, Spain, United Kingdom, and United States.

Country	Error	MSE		R^2	
		R_t	New death rate	R_t	New death rate
Brazil	0.165 (0.162-0.170)	0.009 (0.008-0.011)	0.004 (0.004-0.004)	0.872 (0.844-0.888)	0.967 (0.965-0.969)
Spain	0.256 (0.251-0.263)	0.054 (0.044-0.072)	0.004 (0.004-0.004)	0.786 (0.716-0.826)	0.951 (0.949-0.953)
United Kingdom	0.260 (0.244-0.274)	0.044 (0.042-0.045)	0.009 (0.007-0.011)	0.711 (0.699-0.722)	0.943 (0.926-0.957)
United States	0.126 (0.123-0.130)	0.004 (0.004-0.004)	0.001 (0.001-0.002)	0.963 (0.961-0.965)	0.971 (0.970-0.972)

Note: values are presented as the mean with 95% Confidence Interval (CI) bounds in parenthesis.
Error: the objective function error, as defined by (5.3).
MSE: Mean Squared Error. R^2: coefficient of determination. R_t: effective reproduction number. New death rate: per 100,000 inhabitants.

Figure 27a presents the model estimation of the R_0 for Brazil, indicating an initial value of 2.44 (95% CI: 2.42 - 2.46). Notably, the model depicts a subsequent decrease in R_0 to 1.00 (95% CI: 0.99 - 1.01) on May 18, 2020. Following this initial decline, the model suggests a successive increase in R_0 with each outbreak, culminating in its peak value of 5.20 (95% CI: 4.83 - 5.41) observed in the last outbreak.

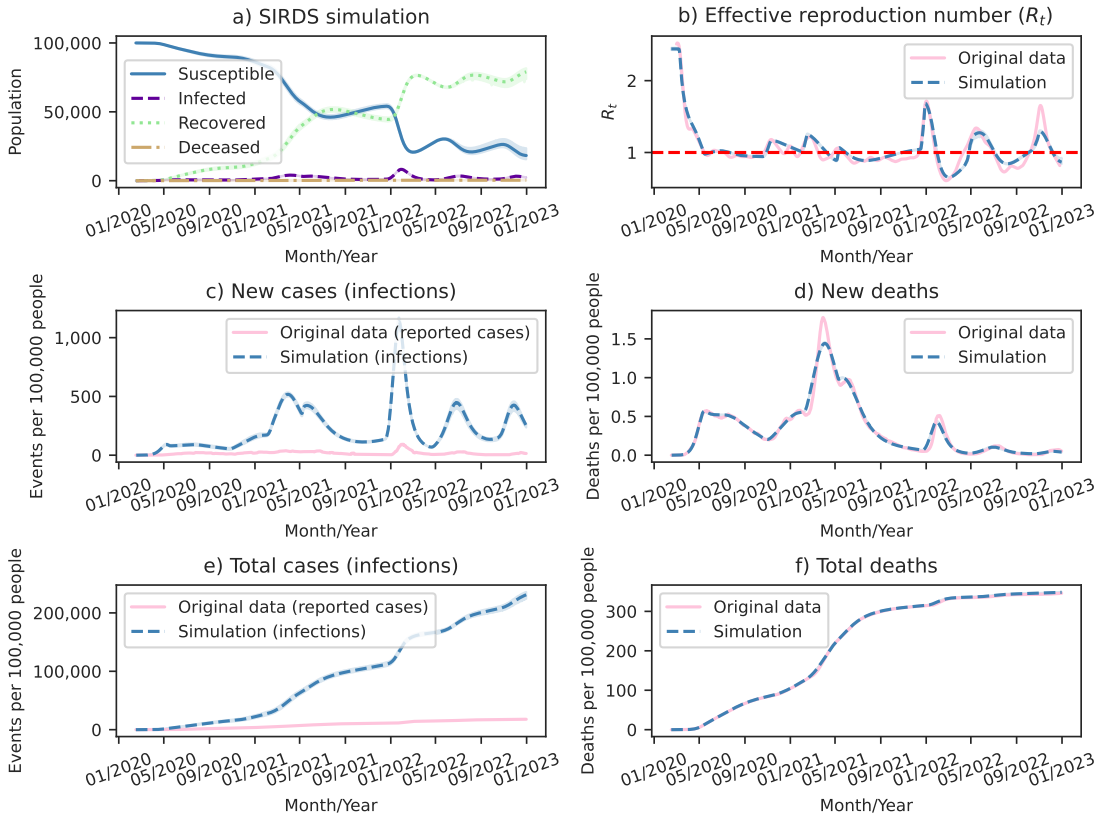


Figure 26 – Comprehensive analysis of simulation results for COVID-19 in Brazil. (a) Model outcomes for an eight-day recovery period detailing the population compartments: Susceptible, Infected, Recovered, and Deceased. (b) Time series comparison between the effective reproduction number (R_t) estimated directly from reported Severe Acute Respiratory Syndrome (SARS) cases and R_t calculated by model simulations. (c) Time series comparison between new cases reported by health authorities and new infections in model simulations. (d) Time series comparison between new deaths reported by health authorities and new deaths in model simulations. (e) Time series comparison between cumulative cases reported by health authorities and cumulative infections in model simulations. (f) Time series comparison between cumulative deaths reported by health authorities and cumulative deaths in model simulations. Shaded regions depict the 95% Confidence Interval (CI).

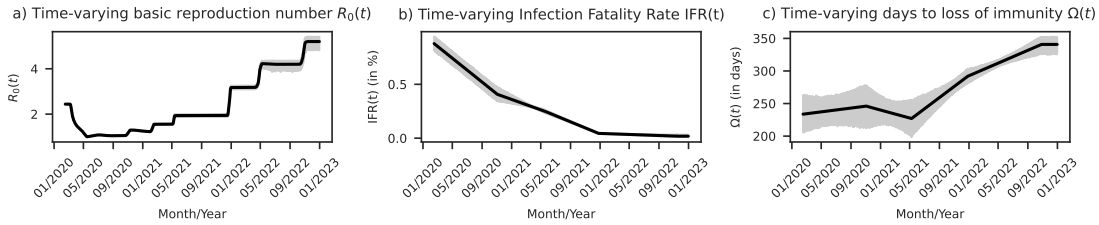


Figure 27 – Time-varying model parameters fitted for COVID-19 in Brazil. (a) Basic reproduction number (R_0) varying with time (t). (b) Infection Fatality Rate (IFR) varying with t. (c) Days to loss of immunity (Ω) varying with t. Shaded regions depict the 95% Confidence Interval (CI).

Figure 27b suggests an initial COVID-19 lethality in Brazil of 0.88% (95% CI: 0.81% - 0.94%). The model further indicates a continuous reduction in IFR throughout the outbreaks, reaching its lowest value of 0.018% (95% CI: 0.011 - 0.033) in the last outbreak.

The parameter about the number of days to loss of immunity, Figure 27c, showed the highest uncertainty. The initial immunity period was estimated to be 234 days (95% CI: 206 - 262), exhibiting an increasing trend, albeit with a reduction observed in mid-2021. The last studied outbreak revealed an immunity period of 341 days (95% CI: 327 - 352).

We also estimate that 63 people (95% CI: 58 - 68) were infected in Brazil when the health authorities reported the first case. For further insights, Appendix I details the fuzzy variables used in this work to facilitate smooth transitions between epidemic periods.

In conclusion, our model reveals a substantial disparity between reported cases in Monitoring Panel (DATASUS, 2020a) and simulated infections, estimating a factor of 12.9 (95% CI: 12.5 - 13.2) more infections than the officially notified cases by Brazilian health authorities until the end of 2022. When analyzing the data for each year, we found that the simulated infections were 5.8 (95% CI: 5.2 - 6.4), 12.9 (95% CI: 12.5 - 13.3), and 16.8 (95% CI: 15.8 - 17.5) times higher than the reported cases in 2020, 2021, and 2022, respectively.

5.2.2 Model generalization

Table 10 illustrates the broad applicability of our model across other countries. We observe reduced errors and well-fitted coefficients of determination for both R_t and new deaths. Notably, the simulations for the United States exhibit the highest level of fitting. Overall, the simulations align more closely with the rate of new deaths than the R_t. An overview of the simulations for Spain, the United Kingdom, and the United States is available in Appendix J.

5.2.3 Comparisons with serological research

Figure 28 illustrates that our simulations for the early pandemic moments align with national serological research for Brazil, the United Kingdom, and the United States. Notably, Spain is the only country where our simulation deviates significantly.

Hallal et al. (2020) conducted two seroprevalence surveys in 133 larger cities in Brazil, using random household and individual selection while excluding children under one year. The first survey, conducted from May 14–21, 2020, had 25,025 individuals, estimating a seroprevalence of 1.9% (95% CI: 1.7 – 2.1), and the second survey, conducted from June 4–7, 2020, had 31,165 individuals, estimating a seroprevalence of 3.1% (95% CI: 2.8 – 3.4). In comparison, our model estimated cumulative infections for Brazil as 2.36% (95% CI: 2.17 - 2.56) on May 14, 2020, and 4.06% (95%CI: 3.72 - 4.44) on June 04, 2020.

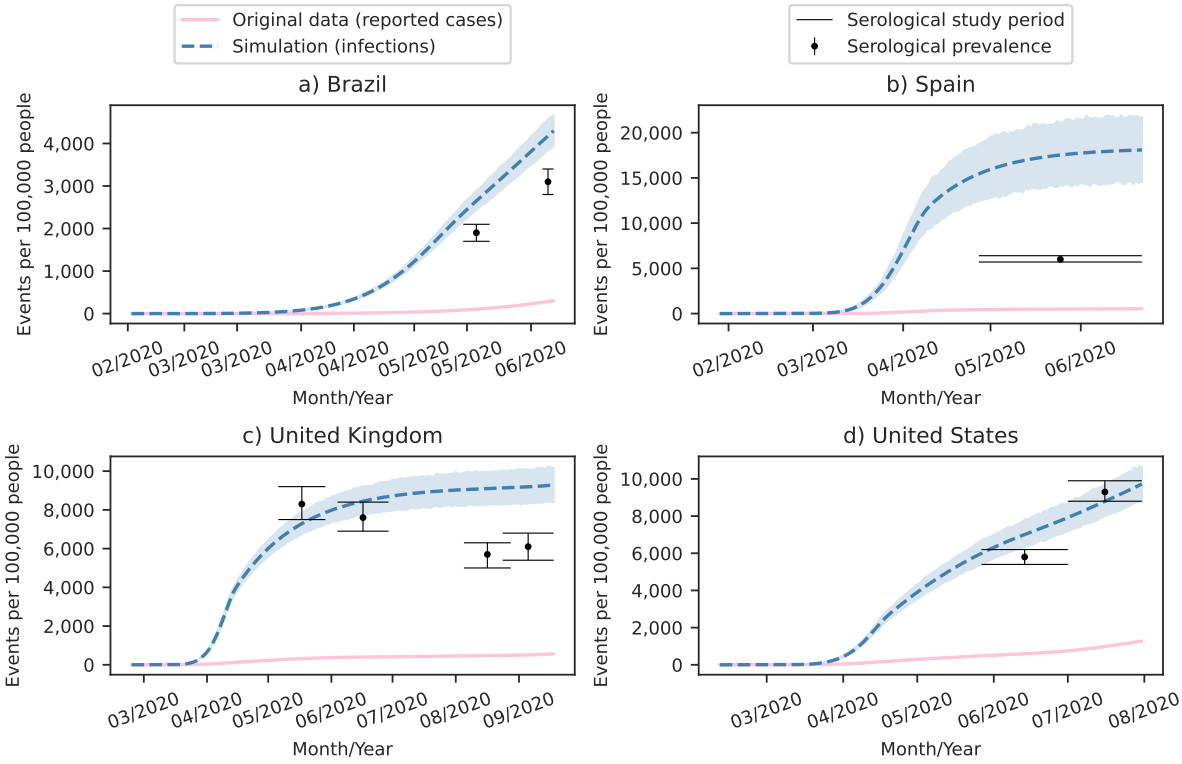


Figure 28 – Comparison of cumulative COVID-19 infections simulated by our model, cumulative reported cases by health authorities, and serological prevalence during the early stages of the pandemic in various countries. (a) Brazil: reported cases from Monitoring Panel (DATASUS, 2020a) and serological prevalence from Hallal et al. (2020). (b) Spain: reported cases from Our World in Data (Ritchie et al., 2020) and serological prevalence from Pérez-Gómez et al. (2023). (c) United Kingdom: reported cases from Our World in Data (Ritchie et al., 2020) and serological prevalence from Public Health England (Public Health England, 2020c,a,b). (d) United States: reported cases from Our World in Data (Ritchie et al., 2020) and serological prevalence from Walker et al. (2021) and Anand et al. (2020). Dashed lines are the simulated cumulative infections, and shaded regions depict the 95% Confidence Interval (CI).

Pérez-Gómez et al. (2023) conducted a representative cohort study of the Spanish population, with 68,287 participants between April 27, 2020, and June 22, 2020. They estimated a seroprevalence in Spain of 6% (95% CI: 5.7 - 6.4) for this period. In contrast, our model estimated a prevalence of 15.5% (95% CI: 12.4 - 18.8) on April 27, 2020.

Public Health England conducted four serological surveys in the first wave in England (Public Health England, 2020c,a,b), which serve as a reference for our simulations in the United Kingdom. The first two surveys (Public Health England, 2020c) were based on healthy blood donors aged 17-69 years, aligning well with our simulations. However, the last two surveys (Public Health England, 2020a,b) included healthy blood donors aged 17 years and older, reducing seroprevalence and causing our model to estimate the prevalence higher than these surveys. Public Health England attributes this decrease in prevalence to

demographic variations in the donor pool, such as the inclusion of donors aged 70 years and older, who were previously restricted during lockdown, and also considers waning antibodies as a potential contributing factor (Public Health England, 2020a,b).

Our prevalence simulations for the United States align with two serological surveys based on dialysis patients. Walker et al. (2021) estimated a prevalence of 5.8% (95% CI: 5.4 - 6.2) among 12,932 dialysis patients from May 27, 2020, to July 1, 2020. Additionally, Anand et al. (2020) estimated a prevalence of 9.3% (95% CI: 8.8 – 9.9) based on a sample of 28,503 randomly selected adult patients receiving dialysis during July 2020.

5.2.4 Sensitivity analysis

In Figure 29, the impact of perturbations in the first breakpoint parameter (b_0) on the COVID-19 simulation for Brazil is evident, showcasing its significant sensitivity. Generally, breakpoint parameters demonstrate considerable sensitivity. Additionally, parameters related to transmissibility, denoted by β, exhibit sensitivity, with the last three showing relatively lower impact. Notably, the initial population infected ($I(0)$) is another parameter demonstrating sensitivity among the model variables.

Conversely, our model exhibits low sensitivity in two crucial epidemiological parameters, namely lethality (f) and loss of immunity (ω). About lethality, we observed some sensitivity for $f_1^{'}$ and f_4 when perturbed by 50%. An additional parameter introduced in our model to facilitate smooth transitions between epidemic periods (τ) demonstrated low sensitivity, with $\tau_0^{'}$ being the only instance where we observed some elasticity.

5.3 Discussion

In this study, we developed a novel mathematical epidemiological model with time-varying parameters driven by fuzzy transitions between epidemic periods. The formulation of this model originated from the considerations outlined in Section 5.1.2.1. We applied and validated our model using COVID-19 data from Brazil, Spain, the United Kingdom, and the United States, demonstrating its robustness and generalization capabilities (Section 5.2.2). Comparative analyses with seroprevalence research in Section 5.2.3 illustrated good model fit, aligning well with the available evidence. Furthermore, the sensitivity analysis in Section 5.2.4 emphasized the crucial role of the time-varying property in capturing the dynamics of the COVID-19 pandemic over three years, particularly concerning changes in transmissibility and breakpoints between epidemic periods. We use this model to present a retrospective analysis of the pandemic in Brazil in Section 5.2.1.

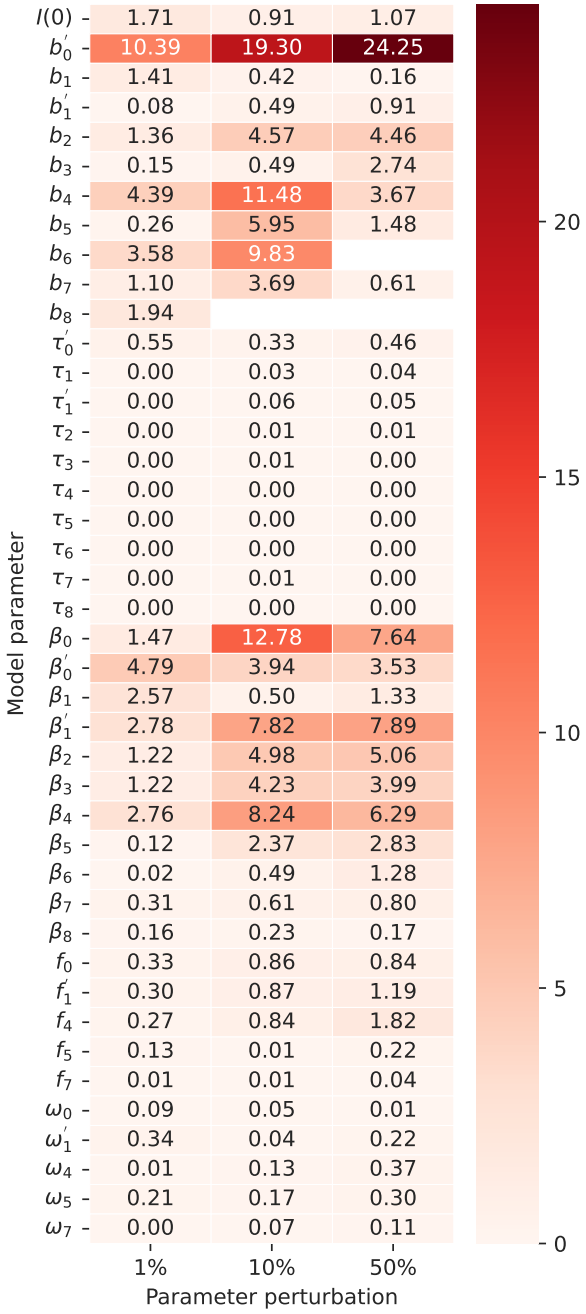


Figure 29 – Sensitivity analysis heatmap for perturbations of 1%, 10%, and 50% in optimized parameters with COVID-19 data in Brazil. Each row corresponds to a specific parameter θ_k, where k denotes the parameter associated with a particular COVID-19 outbreak. When θ_k' is mentioned, it represents an adjustment parameter for atypical outbreak k. The numerical values in each cell represent the elasticity measured for θ under a specific perturbation. The parameters include the initial quantity of infected population $I(0)$, the breakpoint indicating the start of an outbreak (b), transition days between epidemic periods for fast transitions (τ), contact rate (β), infection fatality rate probability (f), and immunity loss rate (ω). Empty cells indicate simulations with errors due to invalid parameter values.

While several studies have leveraged fuzzy theory for COVID-19 pandemic modeling to account for complexity and uncertainties (Melin and Castillo, 2021; Castillo and Melin, 2021; Castillo et al., 2023; Sharma et al., 2021; Abdy et al., 2021), our approach differs from others to employ fuzzy theory in the transitions between epidemic periods, turning the modeling of epidemic period transitions more aligned with our observations of the COVID-19 pandemic. This feature provides a comprehensive view of the epidemiological parameter dynamics, as evidenced in Figure 27, which depicts the time-varying trends for R_0, IFR, and the protection period in the Brazilian context.

We used the proposed model in this work to reproduce the first three COVID-19 pandemic years in Brazil with great reasonability. The estimated R_0 of 2.44 (95% CI: 2.42 - 2.46) aligns with other works (Arroyo-Marioli et al., 2021; Nouvellet et al., 2021). Notably, the time-varying R_0 during the first wave is in line with the adjusted R_0 of 0.91 (95% CI: 0.83 - 1.01) identified by Nouvellet et al. (2021), reflecting the initial reduction in human mobility.

Marra and Quartin (2021), using the same database of Hallal et al. (2020), estimated the IFR for Brazil as 1.03% (95% CI: 0.88% – 1.22%) for serological tests collected between May and June 2020, while our model estimated an IFR of 0.62% (95% CI: 0.57 - 0.68) for the period until June 24, 2020. Our estimate is consistent with other studies that investigated the IFR in the first wave, such as the study conducted by Picon et al. (2020), which estimated an IFR of 0.60% (95% CI, 0.49% – 0.74%) considering antibody prevalence in a Brazilian city, Silveira et al. (2020), which estimated an IFR of 0.38% (95% CI: 0.21% – 0.80%) considering antibody prevalence in a Brazilian state, and Verity et al. (2020), which estimated an IFR of 0.66% (95% CI: 0.39% – 1.33%) using early Polymerase Chain Reaction (PCR) tests in China.

The duration of protection against COVID-19 is still uncertain, as research suggests a fast decay of coronavirus antibodies (Hallal et al., 2020). While some studies indicate that antibody protection could last between five and seven months (Ripperger et al., 2020), others suggest that T and B cells may extend protection (Cohen et al., 2021). Furthermore, there is a risk of reinfection within 90 days of the previous infection (Morris et al., 2022; Pulliam et al., 2022), and also the emergence of new variants that may escape the protection given by a previous infection or vaccine (Pulliam et al., 2022; Nyberg et al., 2022).

Ferrante et al. (2021b) proposed an epidemic model that estimated a protection period of 240 days for COVID-19 based on data from the first and second waves in Manaus/AM. Similar to our model, that estimated a protection period of 234 days (95% CI: 206 – 262) for the early pandemic moments. Morris et al. (2022) used SARS-CoV-2 genomic surveillance data from Johns Hopkins and found a median interval of 377 days between the first infection and reinfection. However, another empirical study showed that reinfection peaks with intervals of six months in South Africa (Pulliam et al., 2022).

The protection period is, without a doubt, the COVID-19 parameter with the most significant uncertainty, as can be seen in our simulation with data from Brazil in Figure 27 and for the other countries in Appendix K. Our sensitivity analysis reinforces this uncertainty, which shows that the protection period is a parameter with low sensitivity.

Estimating the underreporting of cases is a critical factor in comprehending the actual magnitude of an epidemic. In the early stages of the COVID-19 pandemic in Brazil, several studies focused on addressing this issue. For instance, Reis et al. (2020) utilized a mathematical model and estimated that only 10% of COVID-19 infections in Brazil were reported until April 6, 2020. Our model suggests an even more significant underreporting, indicating that Brazilian health authorities reported approximately 4.3% (95% CI: 4.0 - 4.6) of infections during the same period. Bastos et al. (2021), also employing a mathematical model, estimated a range of 8-16 times more infections than reported cases until May 31, 2020, which aligns with our model that estimated a factor of 15 (95% CI: 13 - 16) for the same period. Our findings are consistent with a machine learning model proposed by Noh and Danuser (2021), which estimated that around 20% of infections were reported in Brazil until September 3, 2020.

We recognize that serological surveys are the most reliable method for assessing the underreporting of COVID-19 cases. While the comprehensive COVID-19 serological survey in Brazil conducted by Hallal et al. (2020) did not explicitly calculate the underreporting ratio, we estimated an underreporting factor of 9.1 (95% CI: 8.2 - 10.0), based on their prevalence study on June 4-7, 2020. Our model approximated the factor derived from Hallal et al. (2020), estimating a factor of 12.6 (95% CI: 11.6 - 13.9) until June 7, 2020.

We note that serological surveys also have limitations as they cannot differentiate between historical and current infections or distinguish antibodies resulting from natural exposure and vaccination (Gibbons et al., 2014). Conducting a serological survey after around three years (1,050 days) becomes increasingly challenging, and we have not found recent studies employing this method in the Brazilian context. Therefore, epidemiological modeling is crucial in analyzing a prolonged epidemic such as COVID-19 in Brazil. Our proposed model competes well with solid epidemiological evidence noted in the first wave and successfully captures the mortality rate trends and R_t in other outbreaks. To our knowledge, our model is the first comprehensive investigation of the first three years of COVID-19 in Brazil.

Still, regarding the underreporting of COVID-19 cases in Brazil, our model suggests an increase in underreporting factors after the first pandemic year. We conjecture that the model caught an under-ascertainment phenomenon, where infected individuals do not seek healthcare (Gibbons et al., 2014). Several factors likely contributed to this population behavior, including a sense of reduced risk among the population due to vaccination (Juyal et al., 2021; Dong et al., 2020), the predominance of mild Omicron infections (Nyberg et al., 2022), and the availability of self-tests (Salles, 2023).

The official data about Brazil revealed a 79% reduction in CFR during the 2022 year compared to 2020. Our model estimated an even more substantial 94% (95% CI: 93 - 95) reduction in IFR for the same period, likely associated with increased underreporting during the Omicron phase, as previously discussed. Likewise, for the Omicron phase, other mathematical models by Ribeiro Xavier et al. (2022) estimated a 41% reduction in IFR for Brazil, Liu et al. (2022) reported a 78.7% reduction (95% CI: 66.9% - 85.0%) in South Africa, and Yang and Shaman (2022) observed IFR reductions in nine South African provinces.

5.3.1 Limitations

While our work effectively captures the dynamic changes in epidemiological parameters across outbreaks, we recognize certain limitations inherent in our modeling approach.

Firstly, we maintained a fixed recovery period of 8 days ($\gamma = 1/8$), as discussed in Section 5.1.3.1. Our empirical observation drove this decision that introducing time variability to γ significantly increased computational costs for parameter optimization without substantially improving model outcomes.

Moreover, our model does not explicitly include an *Exposed* (E) compartment, despite the common suggestion in the literature to incorporate an incubation period for modeling COVID-19 (Delli Compagni et al., 2022; Ala'raj et al., 2021; Silva et al., 2020). We addressed this limitation by assuming that our *Infected* (I) compartment implicitly considers aspects of asymptomatic and pre-symptomatic infections. Our observation indicates that the model aligns well with the observed deaths and R_t even with this simplification.

We must note that we relied on notification dates reported by health authorities rather than the actual event dates for data from Spain, the United Kingdom, and the United States. This divergence in data sources introduces uncertainties and may impact the precision of our generalization analyses.

Finally, we acknowledge the lack of direct comparisons with other compartmental models, which would help assess the convergence and divergence of our simulation results relative to alternative approaches.

6 MODELS FOR MEDIUM-TERM FORECASTING COVID-19 MORTALITY IN LARGE BRAZILIAN MUNICIPALITIES

The COVID-19 pandemic (WHO, 2020d) profoundly impacted the world, beginning in 2020 and persisting through subsequent years. In response, governments and health authorities relied on forecasting models to guide critical decisions to interventions, resource allocation, and public health planning (Shinde et al., 2020).

Now, some years after the peak of the COVID-19 crisis, we can retrospectively evaluate the effectiveness of the forecasting models. Although the academic community has contributed with many models, especially for short-term forecasts spanning a few days to weeks (Rahimi et al., 2021; Kamalov et al., 2022), we are interested in medium-term forecasting models that are particularly valuable for public health planning, as they allow for more strategic decisions regarding resource distribution and containment measures.

In the COVID-19 literature, the medium-term horizon is typically defined as four weeks or more (Manley et al., 2024; Drews et al., 2022; Bhatia et al., 2023), sometimes extending to a few months (Dairi et al., 2021; Hasan et al., 2022). Forecasting accurately over this period is particularly challenging due to the pandemic dynamic nature, where different factors can cause sudden shifts in the time series (Drews et al., 2022; Bhatia et al., 2023).

In this chapter, we conducted a retrospective analysis of COVID-19 mortality models for medium-term forecasting across the 41 largest Brazilian municipalities. Our focus was on examining the accuracy of these models over an extended horizon of 84 days, covering nine distinct forecasting windows within a period of nearly 25 months (from April 26, 2020, to May 24, 2022). This timeframe captures various pandemic phases, allowing us to assess model performance under different epidemiological conditions.

We demonstrate the capability of our compartmental model with fuzzy transitions between epidemic periods, introduced in Chapter 5, to provide COVID-19 death forecasts. Additionally, we proposed hybrid models based on our compartmental model. Furthermore, we evaluated the performance of data-driven and ensemble models, comprehensively comparing different forecasting approaches.

6.1 Methodology

6.1.1 Data

This study utilized daily COVID-19 death data from Brazilian municipalities, extracted from the Mortality Information System (SIM) (DATASUS, 2022b). The data collection covered the period from the early stages of the pandemic in 2020 until May 21, 2022. We concentrated our analysis on a sample of the 41 largest Brazilian municipalities with populations exceeding 500,000, as detailed in Section 3.2.1.4.

For tracking cases in cities, we accessed data from the Monitoring Panel (DATASUS, 2020a) and the SARS database (DATASUS, 2022a), both detailed in Section 3.1.1.

6.1.2 Forecasting horizon and analysis windows

In this study, we consider a horizon of 84 days. We assessed nine windows, each covering one horizon, with maximum fitting dates as follows: (i) April 25, 2020, (ii) July 18, 2020, (iii) October 10, 2020, (iv) January 2, 2021, (v) March 27, 2021, (vi) June 19, 2021, (vii) September 11, 2021, (viii) December 4, 2021, and (ix) February 26, 2022. Figure 30 highlights these analysis moments across our target variable, the COVID-19 death rate.

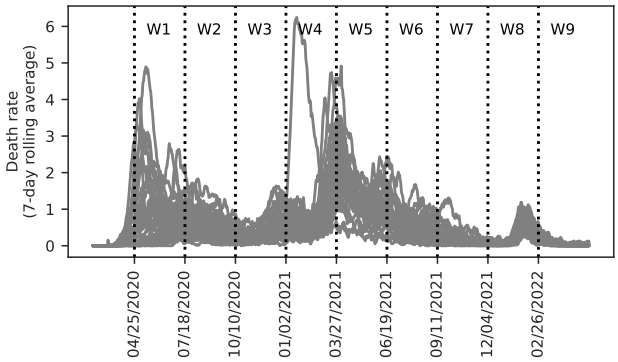


Figure 30 – COVID-19 death rate per 100,000 inhabitants in a 7-day rolling average for the 41 largest Brazilian municipalities. The dotted vertical lines indicate the forecast dates for the analysis windows, with each 84-day window labeled as Wi where i ranges from the first window ($W1$) to the ninth window ($W9$).

Source: Mortality Information System (SIM) (DATASUS, 2022b).

6.1.3 Forecasting models

We employed five models to predict the rate of new COVID-19 deaths per 100,000 population on a 7-day moving average for the municipalities in our dataset. We fine-tuned each model for each forecast window. In this section, we present each model.

6.1.3.1 Fuzzy SIRDS model

We employed the Fuzzy Susceptible-Infected-Recovered-Deceased-Susceptible (SIRDS) model, introduced in Chapter 5, to forecast COVID-19 deaths across municipalities in multiple windows. While Chapter 5 focused on using this mathematical model to estimate time-varying epidemiological parameters for national-level data, this chapter examines the model forecasting performance at the municipal level.

Since our model uses epidemic periods to guide its input parameters, we utilized the outbreaks identified for the 41 largest Brazilian municipalities, as detailed in Section 3.2.3. We followed the procedures outlined in Section 5.1.2.3 to optimize our municipal dataset. The only exception was the maximum bound for the parameter that denotes the initial number of infected individuals ($I(0)$), which we refined, setting it to the case rate per 100,000 inhabitants until the 56^{th} day in the first outbreak. For each municipality in our sample and each forecasting window, we performed 20 simulations, assuming an infectious period of eight days ($\gamma = 1/8$). Figure 31 illustrates the delimitation of data used to fit the model for São Paulo city in each analysis window. We employed the same approach for all other municipalities in our dataset.

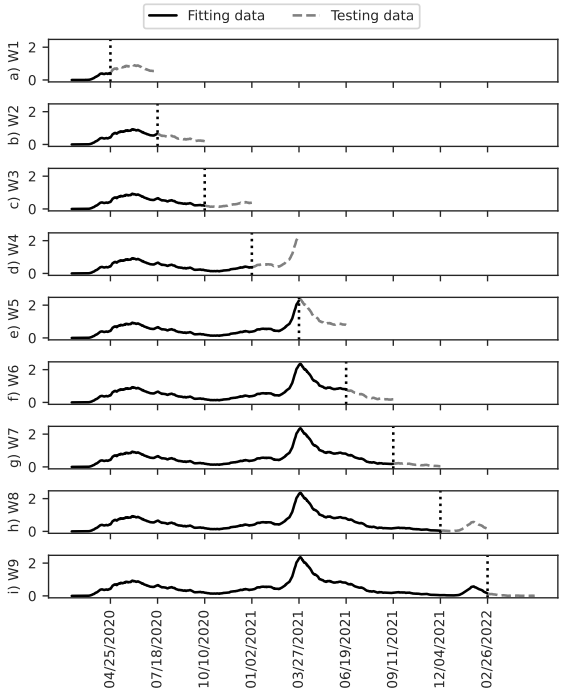


Figure 31 – Death data for fitting and evaluating the Fuzzy SIRDS model across nine 84-day forecast windows (a–i: $W1$–$W9$). Charts show COVID-19 death rates per 100,000 inhabitants (7-day rolling average) for São Paulo, with vertical lines marking observed and test data boundaries. Source: DATASUS (2022b).

6.1.3.2 LSTM model

We employed an univariate Long Short-Term Memory (LSTM) neural network model. We trained this model with global data, using data from all municipalities to train the model and then making forecasts individually for each municipality. The forecasting horizon was 84 days; however, for the first three forecasting windows, we used 28 days due to data limitations in the early pandemic. For the model input, we used an interval lagged by 84 days, but for the first three windows, we used intervals lagged by 7, 14, and 21 days, respectively, again due to early pandemic data limitations. We tuned the model for each forecasting window, using one forecasting horizon as validation data. Figure 32 illustrates this process.

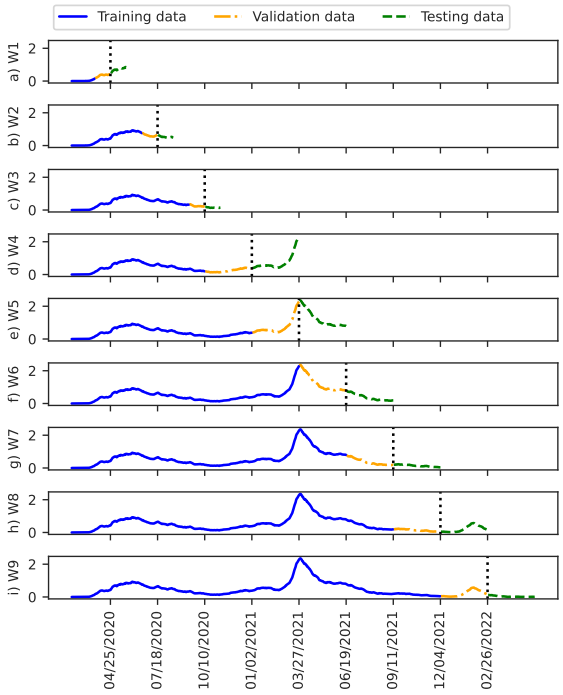


Figure 32 – Death data for training and evaluating the LSTM model across nine forecast windows ($W1$–$W9$). Panels (a–i) show COVID-19 death rates per 100,000 inhabitants (7-day rolling average) for São Paulo. Forecast windows are 84 days, except $W1$–$W3$ (28 days). Testing and validation sequences match the forecast window size. The model was trained for each window using data from all other municipalities in the dataset. Source: DATASUS (2022b).

Additionally, we randomly defined configurations with one to three hidden layers, 32 to 128 neurons per hidden layer, dropout layers with rates of 0.1 to 0.5, and the activation function of the output layer as either ReLU or sigmoid. We conducted 64 trials for each analysis moment. We trained with 128 epochs, using a patience of ten epochs. So, we selected the 20 best models per analysis moment to provide confidence intervals for our results. Last, we refitted the selected models, integrating the validation and training data. We executed this pipeline using *TensorFlow* and *Keras* in Python.

6.1.3.3 Hybrid LSTM model

Our Hybrid LSTM model integrates the outputs of the Fuzzy SIRDS model as inputs for a multivariate LSTM model. Specifically, the compartmental time series (Susceptible (S), Infected (I), Recovered (R), and Deceased (D)) were used as features for the model. In addition, the rate of new COVID-19 deaths per 100,000 inhabitants in a 7-day rolling average was also a feature of this model. The pipeline for training this model followed the same approach outlined for the univariate LSTM model in Section 6.1.3.2. Figure 33 illustrates the structure of the Hybrid LSTM model.

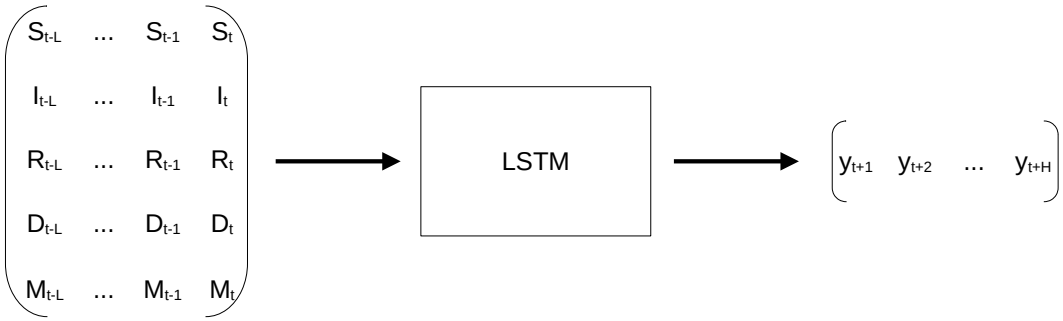


Figure 33 – Structure of the Hybrid LSTM model, highlighting the input and output windows. The input features include the outputs of the Fuzzy SIRDS model: Susceptible (S), Infected (I), Recovered (R), and Deceased (D). Additionally, the rate of new COVID-19 deaths per 100,000 inhabitants in a 7-day rolling average, denoted by M, is incorporated as an input feature. Each input feature is a time series of L samples, representing days lagged from time t. The output, y_{t+h}, represents the predicted death rate for a municipality over the subsequent H days in the forecast horizon.

Source: Mortality Information System (SIM) (DATASUS, 2022b).

6.1.3.4 Hybrid SIRDS model

The Hybrid SIRDS model customizes the Fuzzy SIRDS model by adjusting the IFR and R_0 parameters for the forecast period based on the LSTM model output. We defined the adjusted IFR (IFR$'$) using the deaths estimated by the LSTM model in the first forecasting week and the mean infected population estimated by the Fuzzy SIRDS model in the last week before the forecasting period. So, (6.1) defines IFR$'$:

$$\text{IFR}' = \frac{7\sum_{i=1}^{7}\hat{y}_{t+i}}{\sum_{i=0}^{6}\hat{I}_{t-i}}, \tag{6.1}$$

where $\hat{y}$ is the estimated death rate by the LSTM model, $\hat{I}$ is the compartment of infected individuals estimated by the Fuzzy SIRDS model, and t is the maximum date to fit the model in an analysis window.

To estimate the adjusted R_0 (R'_0), we first estimated the effective reproduction number ($\hat{R}_t$) using the death time series from the LSTM model. We then used the $\hat{R}_t$ mean in the first forecasting week and the susceptible ($\hat{S}$) population average estimated by the Fuzzy SIRDS model in the last week before the forecasting period to estimate:

$$R'_0 = \sum_{i=1}^{7} (\hat{R}_t)_{t+i} \frac{N}{\sum_{i=0}^{6} \hat{S}_{t-i}}, \tag{6.2}$$

where N is the total population. We combined pairs of models from the Fuzzy SIRDS model and the LSTM model, ultimately producing 20 Hybrid SIRDS models.

6.1.3.5 Ensemble model

We developed an Ensemble model by averaging the predictions of four base models: Fuzzy SIRDS, LSTM, Hybrid LSTM, and Hybrid SIRDS. We conducted 20 simulations for each base model, municipality, and analysis window to ensure robust results. To produce CI for our ensemble estimation, we calculated the COVID-19 death rate predictions by:

$$\hat{y}^{\text{Ensemble}}_{w,m,s,t} = \frac{\hat{y}^{\text{Fuzzy SIRDS}}_{w,m,s,t} + \hat{y}^{\text{LSTM}}_{w,m,s,t} + \hat{y}^{\text{Hybrid LSTM}}_{w,m,s,t} + \hat{y}^{\text{Hybrid SIRDS}}_{w,m,s,t}}{4}, \tag{6.3}$$

where $\hat{y}$ is the estimation by a model for day t in window w, city m, and simulation s.

6.1.4 Evaluation of model performances

We performed 20 forecasting simulations per municipality for each model in every window to generate a 95% Confidence Intervals (CI). We considered the average time series forecasted in our analysis to compare the performance of the models.

We evaluate the forecast accuracy using the Root Mean Squared Error (RMSE). For a relative evaluation, we used the Symmetric Mean Absolute Percentage Error (SMAPE) as an alternative to the Mean Absolute Percentage Error (MAPE). Since our target is the COVID-19 death rate, which often includes values close to zero, SMAPE can provide a more balanced evaluation than MAPE. The RMSE and SMAPE are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{6.4}$$

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{\frac{|\hat{y}_i|+|y_i|}{2}} \times 100, \tag{6.5}$$

where n is the number of days in the analyzed period, y_i represents the actual values, and $\hat{y}_i$ represents the values forecasted by a model.

We evaluated the RMSE and SMAPE across all forecasted days within each forecasting window and for specific timeframes. These included short-term periods (the first week, first two weeks, and first three weeks) and medium-term periods (the first four weeks, weeks five through eight, and weeks nine through twelve). Furthermore, we identified the best-performing model for each municipality by selecting the one with the lowest SMAPE.

6.2 Results

Figure 34 illustrates the forecasts provided by our models for the four most populous Brazilian municipalities: São Paulo/SP, Rio de Janeiro/RJ, Brasília/DF, and Fortaleza/CE. We provide outcome plots for all municipalities in our dataset on the online supplementary material: `https://github.com/helderseixas/thesis`.

To provide an overview of model performance across all municipalities in our dataset, we created heatmaps that visualize SMAPE metrics for the entire forecasting window and specific periods: the first four weeks, weeks 5 to 8, and the last four weeks. Figure 35 presents the performance of the Fuzzy SIRDS model in forecasting COVID-19 deaths across the 41 largest Brazilian municipalities. The results indicate that this model performed significantly better during the first four weeks than longer-term forecasts. Additional heatmaps for the other models analyzed in this study are available in Appendix L.

Figure 36 illustrates the magnitude of errors observed for the models, with the lowest RMSEs occurring in the last two windows and the highest errors reported in the first window. In contrast, Figure 37 displays the relative errors, showing the lowest errors in Windows 4 and 5 and the highest in the first and last windows. A consistent pattern emerges across all models, with errors increasing over time within each window, i.e., shorter-term predictions generally exhibit lower errors than longer-term predictions. Furthermore, all models demonstrate performance within similar error ranges. Notably, the LSTM model frequently achieves a lower median error than other models from Window 4 to Window 8. However, in the last window, the LSTM model performed the worst, whereas the Fuzzy SIRDS model delivered the best performance.

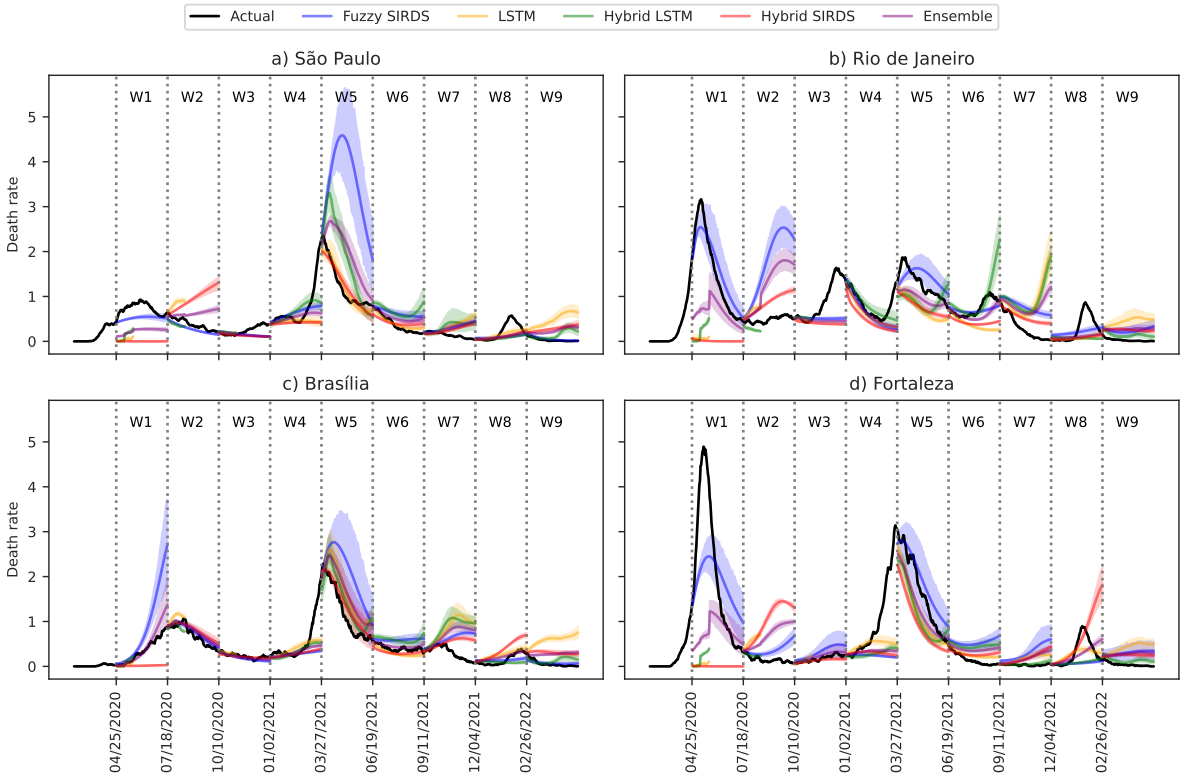


Figure 34 – COVID-19 death rate forecasts per 100,000 inhabitants across nine forecasting windows for the (a) São Paulo/SP, (b) Rio de Janeiro/RJ, (c) Brasília/DF, and (d) Fortaleza/CE municipalities. The black line represents the observed death rate displayed as a 7-day rolling average, while the colored lines illustrate the predictions from the various models. Shaded areas indicate the 95% Confidence Intervals (CI). The dotted vertical lines indicate the forecast dates for the analysis windows, with each 84-day window labeled as Wi where i ranges from the first window ($W1$) to the ninth window ($W9$).

Figure 37 shows that the models present median SMAPE lower than 50% for the first four weeks from the second to the seventh window, highlighting the period from Window 4 to 6 when the models in its majority reported SMAPE third quartile lower than 50%. The relative error is significantly reduced for short-term forecasting, as shown in Figure 38.

Figure 39 illustrates the frequency that each model achieved the lowest SMAPE for a municipality across the nine forecasting windows. We observe a notable variation in the models that best predicted the COVID-19 death rate over time. The Fuzzy SIRDS model was the best predictor for several municipalities during the first three windows, the Window 7 and the final window. On the other hand, between the fourth and sixth windows, the LSTM model provided the most accurate predictions for the higher proportion of municipalities. The Hybrid SIRDS also stood out, delivering the best forecasts in Windows 3, 7, and 8. The Hybrid LSTM and Ensemble models also achieved the best predictions for many municipalities at various moments.

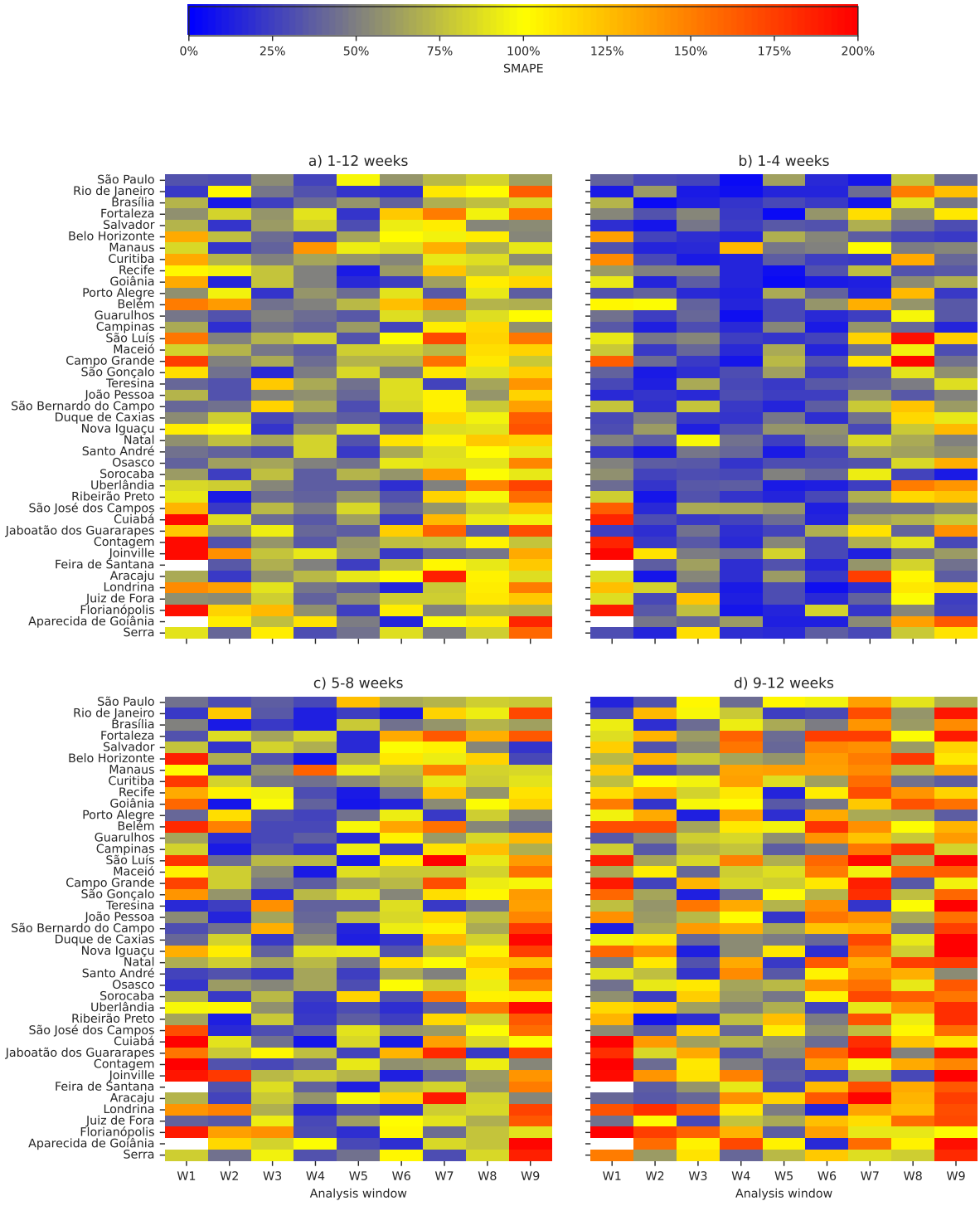


Figure 35 – Heatmaps illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across the 41 largest Brazilian municipalities, based on the Fuzzy SIRDS model predictions over nine forecasting windows. The x-axis represents the forecast windows, with each 84-day window labeled as Wi where i ranges from the first window ($W1$) to the ninth window ($W9$). SMAPE values are shown for (a) the entire forecast period, (b) the first four weeks, (c) weeks five to eight, and (d) weeks nine to twelve.

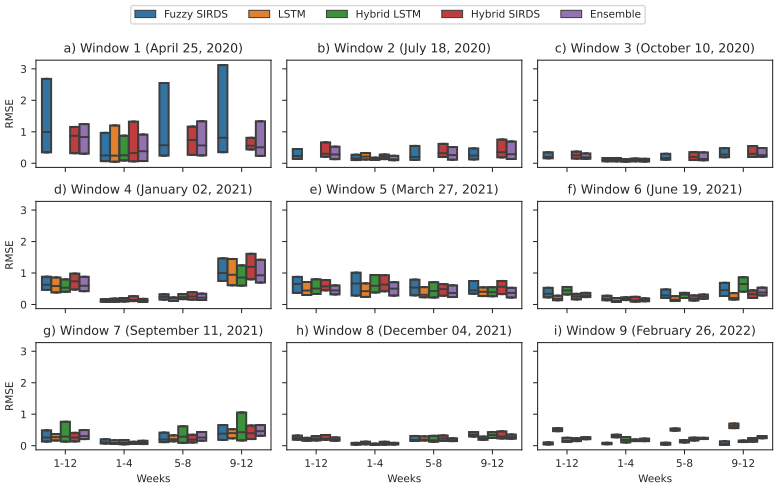


Figure 36 – Boxplots illustrating the Root Mean Squared Error (RMSE) for the COVID-19 death forecasting models across the 41 largest Brazilian municipalities in the medium term: the entire twelve-week forecast period, as well as for the first four weeks, weeks five to eight, and weeks nine to twelve. Each chart plots this metric across nine forecasting windows, ranging from (a) the first window (predicted on April 25, 2020) to (i) the final window (predicted on February 26, 2022). Each boxplot represents the interquartile range (IQR), with the box spanning the first and third quartiles and the median indicated by a horizontal line within the box. Whiskers and outliers are omitted for clarity.

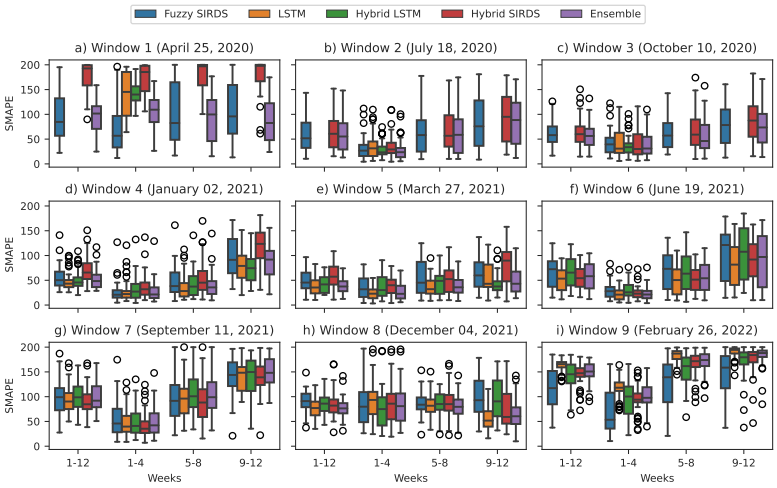


Figure 37 – Boxplots illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for the COVID-19 death forecasting models across the 41 largest Brazilian municipalities in the medium term: the entire twelve-week forecast period, as well as for the first four weeks, weeks five to eight, and weeks nine to twelve. Each chart plots this metric across nine forecasting windows, ranging from (a) the first window (predicted on April 25, 2020) to (i) the final window (predicted on February 26, 2022). Each boxplot represents the interquartile range (IQR), with the box spanning the first and third quartiles and the median indicated by a horizontal line within the box. The whiskers extend to the minimum and maximum values within 1.5 times the IQR, and individual points represent the outliers.

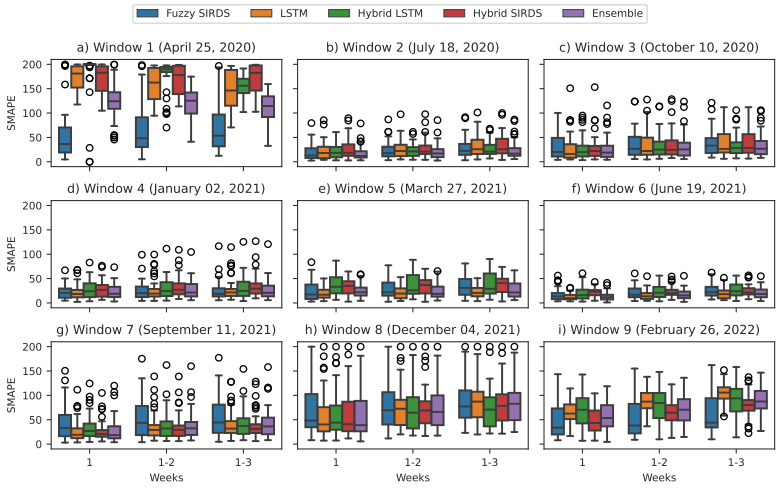


Figure 38 – Boxplots illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for the COVID-19 death forecasting models across the 41 largest Brazilian municipalities in the short term: the first week, the two first weeks, and three first weeks. Each chart plots this metric across nine forecasting windows, ranging from (a) the first window (predicted on April 25, 2020) to (i) the final window (predicted on February 26, 2022). Each boxplot represents the interquartile range (IQR), with the box spanning the first and third quartiles and the median indicated by a horizontal line within the box. The whiskers extend to the minimum and maximum values within 1.5 times the IQR, and individual points represent the outliers.

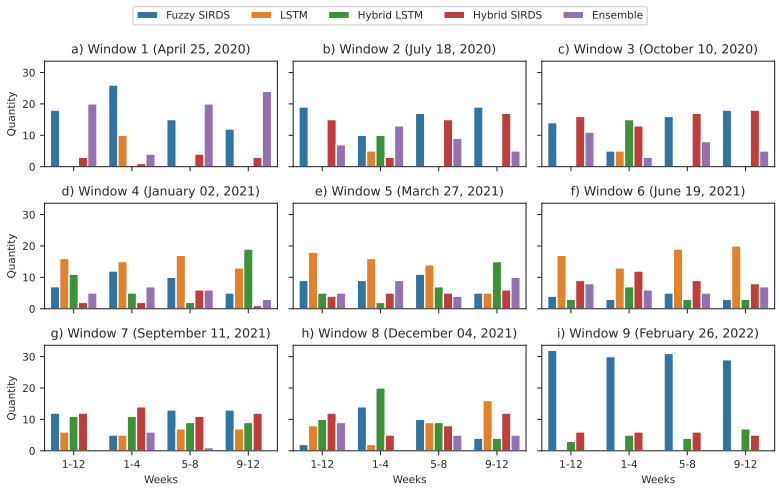


Figure 39 – Bar plots illustrating the number of times each COVID-19 death forecast model achieved the lowest Symmetric Mean Absolute Percentage Error (SMAPE) for a municipality across nine forecasting windows, corresponding to the forecast dates: ranging from (a) the first window (predicted on April 25, 2020) to (i) the final window (predicted on February 26, 2022). SMAPE values were calculated for the 41 largest Brazilian municipalities over the entire forecast period (twelve weeks), as well as for the first four weeks, weeks five to eight, and weeks nine to twelve.

6.3 Discussion

In this study, we empirically evaluated various models for medium-term forecasting of COVID-19 deaths, defined as a forecasting horizon of 84 days (12 weeks). Our analysis focused on the 41 largest Brazilian municipalities across nine distinct forecasting windows, starting from April 26, 2020, shortly after Brazil reported its first case in late February 2020 (DATASUS, 2020a), and continuing through May 24, 2022. Our contribution lies in providing empirical insights into the effectiveness of different models for medium-term COVID-19 death forecasting, which is crucial for epidemiological surveillance, as it enables public health authorities to anticipate and respond proactively to emerging trends.

Our findings indicate no dominant or infallible model for medium-term COVID-19 death forecasting. The performance of the models varied across different forecasting windows and for different municipal time series.

Forecasting COVID-19 deaths in the medium term, particularly at the onset of an epidemic, poses significant challenges for all model types, whether data-driven or compartmental. As expected, Figure 37a shows that the LSTM model, lacking sufficient data, failed to provide accurate forecasts during the first forecasting window. The Fuzzy SIRDS model only offered reasonable forecasts for the first four weeks. Therefore, during the early stages of an epidemic, compartmental models serve as useful tools for providing reasonable forecasts for up to four weeks. These findings align with the literature, which highlights the strength of compartmental models in simulating the spread dynamics of infectious diseases and the limitations of data-driven models that require substantial retrospective data (Rahimi et al., 2021; Shankar et al., 2021).

The Fuzzy SIRDS and LSTM models showed improved forecasting accuracy as more data became available, as illustrated in Figure 37. However, the SMAPE values indicate a decline in accuracy during the last three windows. We attribute this decline to the metric sensitivity to values near zero, which were reported by many municipalities during that period, resulting in inflated relative errors. Notably, the RMSE values were comparatively low for these same windows, reflecting a minor absolute error. On the other hand, the error magnitude increased when municipalities reported higher death rates, as observed in Windows 1, 4, and 5. These observations emphasize the importance of a nuanced and comprehensive analysis when interpreting model performance metrics.

We observed that the LSTM model slightly outperformed the Fuzzy SIRDS model from Window 4 to Window 8. However, the performance of the LSTM model declined in the last window (Window 9). Forecasting during Window 9 posed challenges for the LSTM model due to an unseen pattern in the training data. Specifically, the input time series (Window 8) presented initially stable trends followed by two shifts, first a positive trend and then a negative one, reflecting the onset of the Omicron wave in Brazil. In contrast, the time series

in the forecast window (Window 9) exhibited a decreasing death rate at the start, followed by an extended flat period, as illustrated in Figure 30. Despite leveraging two years of training data, such patterns were infrequent, complicating the data-driven model predictions.

Window 9 highlights the robustness of compartmental models in capturing epidemic dynamics. The Fuzzy SIRDS model showed this strength by frequently providing reliable predictions during this period, resulting in reduced RMSE values (Figure 36). While the LSTM model performed slightly better during periods where the training data allowed fine-tuned fitting, the Fuzzy SIRDS model demonstrated to be a more stable predictor in scenarios marked by heightened uncertainty.

The extended models (Hybrid LSTM, Hybrid SIRDS, and Ensemble) significantly contributed to forecasting COVID-19 death rates in the municipalities. These models frequently emerged as the best predictors for a considerable proportion of municipalities across the analysis windows, as illustrated in Figure 39. However, their performance depends heavily on the accuracy of the base models. This dependency was evident in Window 9, where the LSTM model failed to provide reliable forecasts. Consequently, the performance of the extended models was also affected, leading to a reduced proportion of municipalities for which they provided the best predictions (Figure 39i).

So, we recommend that epidemiological surveillance teams and policymakers account for the diversity of model types to gain a more comprehensive understanding of epidemic trends. Accurate COVID-19 forecasting beyond four weeks is challenging, mainly due to abrupt trend changes within forecast windows, such as the abrupt shift observed at the end of Window 4 (Figure 30), which led to higher error magnitudes (Figure 36d). Despite the increased uncertainty in medium-term predictions, our presented models have demonstrated valuable tools for monitoring pandemics over a long period.

6.3.1 Related work

Various studies have proposed models to forecast COVID-19 data in the medium term. These studies tested with different forecast horizons, ranging from 28 days (de Araújo Morais and da Silva Gomes, 2022; Liao et al., 2021), 30 days (Farooq and Bazaz, 2021), and 31 days (Barnard et al., 2022), to five weeks (Wang et al., 2022), nine weeks (Al-Rashedi and Al-Hagery, 2023; Babashova, 2022), ten weeks (Ramazi et al., 2021; Congdon, 2021), 12 weeks (Appadu et al., 2021), and 13 weeks (Dairi et al., 2021; Samrin et al., 2022; Hasan et al., 2022). These works explored various types of models, including compartmental models (Farooq and Bazaz, 2021; Wang et al., 2022; Barnard et al., 2022; Liao et al., 2021), deterministic models (Appadu et al., 2021; Babashova, 2022), statistical models (Samrin et al., 2022; Hasan et al., 2022; Al-Rashedi and Al-Hagery, 2023; de Araújo Morais and da Silva Gomes, 2022; Congdon, 2021), machine learning models (Farooq and Bazaz, 2021; Ramazi et al.,

2021; Dairi et al., 2021; Al-Rashedi and Al-Hagery, 2023; de Araújo Morais and da Silva Gomes, 2022; Wang et al., 2022; Liao et al., 2021), and hybrid models (Farooq and Bazaz, 2021; de Araújo Morais and da Silva Gomes, 2022; Wang et al., 2022; Liao et al., 2021).

While some studies focused on forecasts for a single country (Ramazi et al., 2021; Samrin et al., 2022; Al-Rashedi and Al-Hagery, 2023; Wang et al., 2022; Congdon, 2021; Barnard et al., 2022), others extended their analysis to a small set of countries (Dairi et al., 2021; Appadu et al., 2021; Hasan et al., 2022; Babashova, 2022; Liao et al., 2021) or regions (Farooq and Bazaz, 2021). Notably, none of these studies conducted medium-term forecasts for more than ten locations. Although these studies produced significant findings, they often lacked a comprehensive and robust approach to assess their forecasting. Some conducted only future forecasts (Samrin et al., 2022; Hasan et al., 2022; Congdon, 2021; Barnard et al., 2022), while others only assessed predictions for a single moment in the pandemic (Farooq and Bazaz, 2021; Ramazi et al., 2021; Dairi et al., 2021; Appadu et al., 2021; de Araújo Morais and da Silva Gomes, 2022; Wang et al., 2022; Babashova, 2022; Liao et al., 2021). Additionally, these studies analyzed a limited number of locations.

On the other hand, four comprehensive studies have conducted retrospective analyses of medium-term COVID-19 forecasts (Adiga et al., 2021; Manley et al., 2024; Drews et al., 2022; Bhatia et al., 2023). These studies primarily employed ensemble models, with a forecasting horizon of 28 days (Adiga et al., 2021; Manley et al., 2024; Bhatia et al., 2023), except for one that extended to a 60-day horizon (Drews et al., 2022). These models were used to forecast COVID-19 confirmed cases (Adiga et al., 2021; Drews et al., 2022), deaths (Manley et al., 2024; Bhatia et al., 2023), hospital admissions (Manley et al., 2024), and hospital bed occupancy (Manley et al., 2024). The least comprehensive study focused only on national data from England (Manley et al., 2024), while others expanded their scope to include data from ten countries (Drews et al., 2022), with the most extensive analyses covering 81 countries (Bhatia et al., 2023) and the U.S. counties (Adiga et al., 2021).

Adiga et al. (2021) compared several models, including ARIMA, LSTM, compartmental, and Kalman filter models, alongside an ensemble model. Adiga et al. (2021) conducted weekly forecasts over nearly six months, from August 2020 to January 2021. The study found that the ensemble model was more stable than the base models, though the authors stressed the importance of the base methods. They observed that different models contributed significantly to forecasting accuracy depending on the spatial region and period.

Manley et al. (2024) evaluated four compartmental models, two data-driven models, one agent-based model, and an ensemble model that integrated these approaches. Forecasting spanned nearly 14 months, from November 2021 to December 2022. The study concluded that the ensemble model enhanced robustness and reduced biases associated with single-model predictions. It performed exceptionally well during exponential growth or decline in the epidemic, although its accuracy diminished around epidemic peaks and valleys.

Drews et al. (2022) utilized a compartmental model and a Holt-Winters statistical model, along with an ensemble model combining these techniques. Drews et al. (2022) conduct forecasts across five windows over nearly seven months, from May to November 2020. The study found that forecasts were most accurate when the epidemic did not change abruptly and noted that the ensemble model consistently outperformed base models. However, they observed that reliable forecasts were generally limited to short-term horizons of a few weeks, with substantial variation across locations and periods.

Bhatia et al. (2023) employed an ensemble model combining three statistical models, conducting 39 analyses over nearly nine months, from March to November 2020. The study found that simple statistical models could reliably capture epidemic trajectories in multiple countries for up to four weeks when fitted to routine disease surveillance data. However, forecasting accuracy declined significantly beyond this time horizon, emphasizing the challenges of medium-term predictions.

The findings presented in this chapter align closely with these state-of-the-art medium-term retrospective forecast studies (Adiga et al., 2021; Manley et al., 2024; Drews et al., 2022; Bhatia et al., 2023). We also observed that different model types contribute to more careful epidemiological surveillance, showing variation in performance across different moments and municipalities. Moreover, forecasting COVID-19 beyond four weeks remains highly challenging. Our study introduces novel contributions to the literature on medium-term retrospective forecasting of COVID-19. We conducted forecasts over a more extended period of nearly 25 months (from April 26, 2020, to May 24, 2022), covering various pandemic phases. Additionally, we are the first to medium-term forecasts for the municipal-level data in the Brazilian context.

6.3.2 Limitations

The primary limitations of this chapter derive from the computational demands associated with fitting both the LSTM and Fuzzy SIRDS models to long-time series data. In particular, the LSTM model requires substantial computational resources to fine-tune its parameters, mainly when deeper architectures are employed to enhance accuracy. Similarly, the complexity of the Fuzzy SIRDS model increases with the length of the time series, as the number of parameters to be optimized grows with the number of outbreaks recorded.

These computational challenges limit our capacity to apply these models to larger datasets and more refined versions of the models, which could produce more accurate forecasts. Despite these constraints, we conducted 20 simulations for each model, window, and municipality to account for uncertainty in our predictions. Increased computational power would enable more extensive analyses and enhance the robustness of the models.

Another limitation is that the data used in this chapter is suitable only for retrospective forecasting, as COVID-19 death records from SIM (DATASUS, 2022b) are consolidated annually. For nowcasting, the only available dataset is the Monitoring Panel (DATASUS, 2020a), which suffers from reporting delays by health authorities. Consequently, applying the models in this chapter for nowcasting would likely reduce accuracy due to data quality issues.

Although we recommend that epidemiological surveillance professionals consider the diversity of forecast model types to support their decision-making, this thesis does not provide a clear guideline for identifying the most reliable forecast when models produce divergent predictions. Further efforts are needed to develop more practical tools that can assist professionals in evaluating and selecting among different model outputs.

7 ANALYZING THE COVID-19 PARAMETERS FOR LARGE BRAZILIAN MUNICIPALITIES

In this chapter, we delve into the dynamics of COVID-19 across the initial three years (2020-2022), focusing on the 41 largest cities in Brazil. Leveraging a mathematical model with fuzzy transitions between epidemic periods as shown in Chapter 5, we estimate time-varying parameters such as the *basic reproduction number* (R_0) and the Infection Fatality Rate (IFR). Via an ample analysis, we correlate these model outputs with data about social isolation measures, vaccination indices, and the emergence of new variants. We aim to offer insights into the factors that influenced the pandemic at the municipal level.

Different studies utilized models to forecast the first COVID-19 wave in the Brazilian context (Bastos and Cajueiro, 2020; Melo et al., 2020; Oliveira et al., 2021). Notably, a work analyzed epidemiological parameters across 29 inner municipalities during the initial wave, emphasizing potential differences in control effectiveness across regions (Almeida et al., 2021). Other work examined the impact of mobility on the variation of R_0 in Brazil and other countries (Nouvellet et al., 2021). Ferrante et al. (2022) correlated epidemiological parameters with various factors to elucidate the dynamics of the first two waves in the Brazilian city of Manaus/AM. We explored in Chapter 5 the variation of epidemic parameters across time using national data from Brazil. In this chapter, we extend the Chapter 5 exploring the dynamic of epidemiological parameters over time for the 41 largest municipalities in Brazil.

We note a need for more work that conducts comprehensive pandemic analysis across a more significant period at the municipal level. So, our study contributes by analyzing the pandemic in larger Brazilian cities across three years. Our study provides an understanding of the factors shaping the pandemic via the correlation of the model outcomes with data on social isolation, vaccination indices, and the emergence of variants. We also examine the cities exhibiting outlier death rates (Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO), contrasted with bottom outliers (Feira de Santana/BA, São Luís/MA, and Florianópolis/SC), which illuminates the diverse drivers behind divergent pandemic outcomes.

We organized the remainder of this chapter as follows: Section 7.1 outlines the methodology, while Section 7.2 presents the results and discussion. The findings discussed in this chapter were also presented at the XXIV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2024) (Lima and Guimarães, 2024).

7.1 Methodology

7.1.1 Data

In this chapter, we utilized the same dataset of daily COVID-19 cases and deaths for the 41 largest Brazilian municipalities as presented in Chapter 6. However, for this analysis, we extended the study period from the early pandemic stages in 2020 to December 31, 2022.

Additionally, we examined our results with correlated data on human mobility, vaccination, and COVID-19 variants. To monitor human mobility in Brazilian municipalities, we utilized data from the COVID-19 Community Mobility Report (Google, 2020) produced by Google. We used vaccination data for Brazilian municipalities from the Brazilian Health Ministry (Brasil, 2022). Additionally, state-level data on COVID-19 genomic surveillance in Brazil, reported monthly by the Fundação Oswaldo Cruz (Fiocruz, 2020), were incorporated. Chapter 3 provides a detailed presentation of data used in this chapter.

7.1.2 Simulations

In this chapter, we utilized our mathematical model proposed in Chapter 5 to simulate the progression of the COVID-19 pandemic in Brazilian municipalities during the early stages of the pandemic through December 31, 2022. Following the processes described in Section 6.1.3.1, we fitted the model parameters to the municipal data. We conducted 25 simulations for each municipality in our sample.

7.1.3 Data analysis

This section outlines the methods employed for analyzing the obtained results. Initially, we evaluated the performance of our model using data from our municipality sample. This assessment relied on two key metrics: the Mean Squared Error (MSE) (5.5) and the coefficient of determination (R^2) (5.6).

For each city, we conducted a cross-correlation analysis (Shumway et al., 2000) between stay-at-home index (Δ_H) and R_0. The cross-correlation coefficient $C_k(x, y)$, given by (7.1), was computed, where x_t and y_t are the series, and k is the lag (Shumway et al., 2000).

$$C_k(x,y) = \frac{\sum(x_{t+k} - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2}\sqrt{\sum(y_t - \bar{y})^2}}. \tag{7.1}$$

To establish the significance limits for cross-correlation, we introduced adjusted factors to mitigate spurious correlations resulting from autocorrelated time series, as suggested by Dean and Dunsmuir (2016). We calculated the weighted cross-correlation

significance limits using (7.2), where a and b denote the autocorrelation coefficients at lag 1 for x_t and y_t , respectively.

$$\pm\frac{1.96}{\sqrt{n}}\sqrt{\frac{1+ab}{1-ab}}. \tag{7.2}$$

Before the cross-correlation analysis, we conducted stationarity transformations for the time series Δ_H and R_0 of the municipalities. We applied a 7-day differencing to generate a new time series denoted as $\Delta_H^{'}$ and $R_0^{'}$. Additionally, $\Delta_H^{'}$ was prewhitened using an ARIMA model of order (28, 0, 7), yielding $\Delta_H^{''}$.

We assessed the stationarity of the transformed series using the Augmented Dickey Fuller (ADF) test (Dickey and Fuller, 1979). We also applied the Ljung–Box test (Ljung and Box, 1978) at lag one that showed that despite prewhitening, $\Delta_H^{''}$ remained non-white noise.

Subsequently, for each municipality and specific year, we computed cross-correlation analyses $C_k(\Delta_H^{''}, R_0)$. Only significant $C_k(x, y)$ values were considered. The *xcorr* method from the Python library *matplotlib.pyplot* was utilized for estimating sample cross-correlation coefficients, while the ADF test and Ljung–Box test were performed using the Python library *statsmodels*.

Lastly, we explored the relationship between virus variants and R_0 and the percentage of vaccinated individuals and IFR. We present these analyses in Section 7.2, which presents our findings and discusses observations regarding outlier municipalities.

7.2 Results and discussion

Table 11 shows that the model closely matches the observed data about mortality in larger Brazilian municipalities. On the other hand, the R^2 assessed for R_t indicates that the model captures moderately the variance of the actual observations.

Table 11 – Results of the COVID-19 simulations for the 41 largest Brazilian municipalities.

Error	MSE		R^2	
	R_t	New death rate	R_t	New death rate
0.278 (0.255-0.320)	0.018 (0.014-0.023)	0.014 (0.010-0.019)	0.600 (0.430-0.678)	0.942 (0.912-0.966)

Note: values are presented as the median with the first and third quartiles in parentheses.
Error: the objective function error, as defined by (5.3). MSE: Mean Squared Error. R^2: coefficient of determination. R_t: effective reproduction number. New death rate: per 100,000 inhabitants.

While the model may not precisely capture the variability of R_t, the overall outcomes suggest that it effectively captures the trend of the original data. As an illustration, Figure 40 depicts the simulation results for Cuiabá/MT, the city with the highest death rate in our sample. The median R^2 for R_t in Cuiabá/MT is 0.55, but still, Figure 40

indicates a reasonable reproduction of the pandemic dynamics. We produced outcome plots for all municipalities in our sample and have them available in the supplementary material of our published paper[1] (Lima and Guimarães, 2024).

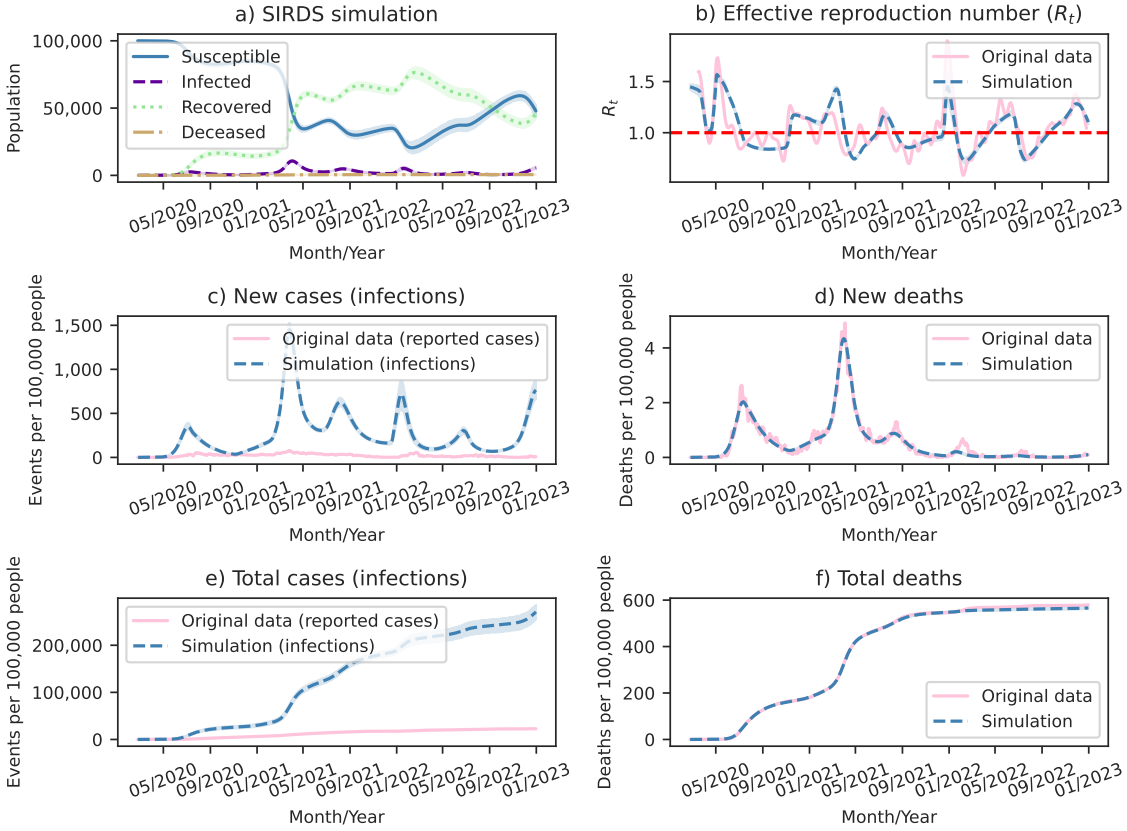


Figure 40 – Comprehensive analysis of simulation results for Cuiabá/MT, the municipality with the highest COVID-19 mortality in our sample. (a) Model outcomes for an eight-day recovery period detailing the population compartments: Susceptible, Infected, Recovered, and Deceased. (b) Time series comparison between the effective reproduction number (R_t) estimated directly from reported Severe Acute Respiratory Syndrome (SARS) cases and R_t calculated by model simulations. (c) Time series comparison between new cases reported by health authorities and new infections in model simulations. (d) Time series comparison between new deaths reported by health authorities and new deaths in model simulations. (e) Time series comparison between cumulative cases reported by health authorities and cumulative infections in model simulations. (f) Time series comparison between cumulative deaths reported by health authorities and cumulative deaths in model simulations. Shaded regions depict the 95% Confidence Interval (CI).

Figure 41 shows the time-varying estimates of R_0, IFR, and days to loss of immunity (Ω) for the municipalities in our sample. In particular, Ω exhibits the highest uncertainty, aligning with the findings observed in Chapter 5, which reported a low sensi-

[1] Supplementary material of Lima and Guimarães (2024): https://github.com/helderseixas/covid-brazilian-municipalities.

tivity for this parameter. We note a trend in cities to reduce R_0 to near the baseline value of 1 during the early stages of the pandemic. However, this trend reverses after mid-2020, with the model estimating an increase in R_0, which persists until 2021. 2022 is marked by an increase in uncertainty in R_0 estimates. Regarding IFR, we note that the highest confusion was in the early stages of the pandemic. In 2021, the model suggests a declining trend in IFR. Notably, the model estimates significantly lower IFR for municipalities in 2022 compared to the preceding years.

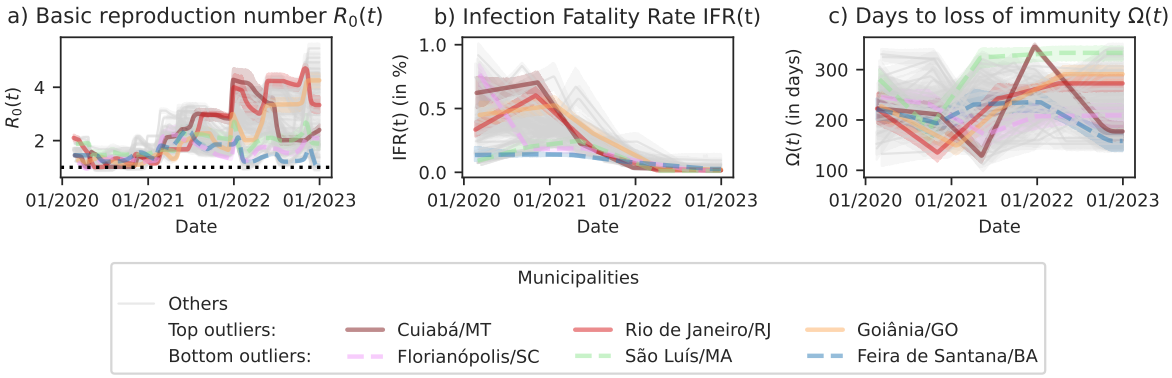
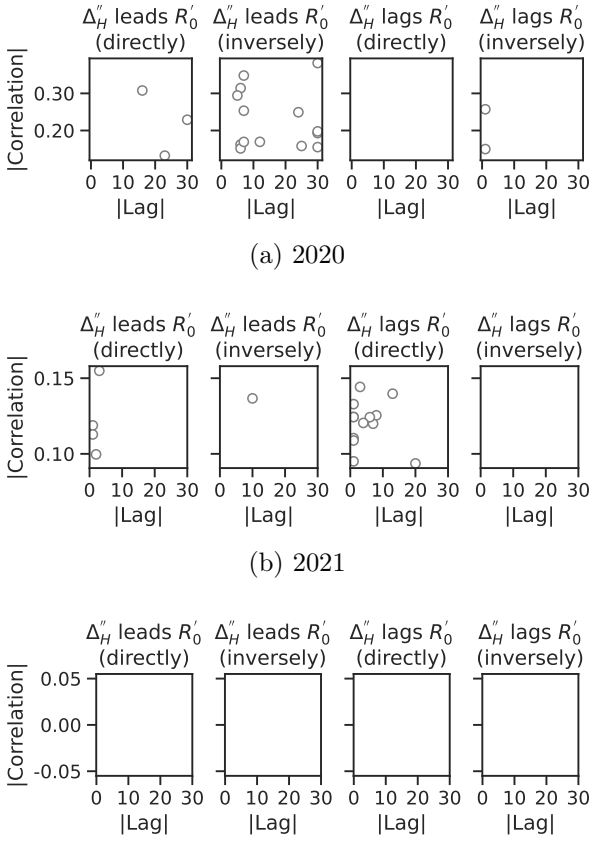


Figure 41 – Model parameters varying by time (t) estimated for COVID-19 in the 41 largest Brazilian municipalities from 2020 to 2022. (a) Basic reproduction number (R_0), with a dotted horizontal line representing the reference value ($R_0 = 1$). (b) Infection Fatality Rate (IFR). (c) Days to loss of immunity (Ω). The plot highlights cities with notable deviations from the average mortality rate, categorized as top outliers (Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO) and bottom outliers (Feira de Santana/BA, São Luís/MA, and Florianópolis/SC). Shaded regions depict the 95% Confidence Interval (CI).

Our analysis delved into the cross-correlation between the time series of Δ_H and the time-varying R_0 for municipalities in our sample, as depicted in Figure 42. Across the years, distinct patterns emerged: in 2020, we observed two distinct groups of municipalities where Δ_H inversely led to R_0, which we noted eight municipalities with a lag of around seven days and six municipalities with a lag of around 28 days. Regarding 2021, we note that Δ_H is lagged by R_0 for 13 municipalities, in which 75% of the correlations were with the lag between one day and seven days. However, in 2022, no notable correlation was observed. These findings shed light on the effectiveness of social isolation measures over time: stringent early stages reduced R_0, while reactive measures in 2021 failed to replicate the same impact. The absence of significant correlations in 2022 indicates that social isolation was no longer the primary measure to combat COVID-19 in municipalities since a large portion of the population was already vaccinated.



(a) 2020

(b) 2021

(c) 2022

Figure 42 – Scatter plots for the correlation absolute coefficients and lag values resulting from cross-correlation analysis between $\Delta_H^{''}$ (stay-at-home index) and COVID-19 $R_0^{'}$ (time-varying basic reproductive number) estimated for the 41 largest Brazilian municipalities in the years (a) 2020, (b) 2021, and (c) 2022. Each point represents the highest significant correlation coefficient observed for a municipality in the respective analysis.

We calculated the median R_0 for each coronavirus variant during the months in which it predominated in the states of the 41 largest Brazilian municipalities. As illustrated in Table 12, the initial dominant variants, categorized as *Others*, exhibited a median R_0 of 1.30. Subsequently, variants such as Gamma and Delta emerged as dominant in 2021, showcasing an R_0 hovering around 2. The Omicron phase marked the period with the highest median R_0, with values exceeding 2.30, except for the BA.4.* subvariant, which reported a lower R_0. These observations underscore the capability of the model to capture the rising trend of R_0 in Brazilian municipalities in response to the emergence of new variants.

Table 13 illustrates a consistent reduction in the IFR as the percentage of the population fully vaccinated against COVID-19 increases. Our analysis reveals a robust negative correlation between the proportion of fully vaccinated individuals and IFR, with a Spearman correlation coefficient of -0.81.

Table 12 – Basic reproduction number (R_0) for each coronavirus variant during months when it was dominant in states from the 41 largest Brazilian municipalities.

Variant	R_0
Others	1.30 (1.16 - 1.45)
P.1.* (Gamma)	1.95 (1.62 - 2.41)
B.1.617.2+AY.* (Delta)	1.98 (1.66 - 2.36)
BA.1.* (Omicron)	2.53 (1.81 - 3.22)
BA.2.* (Omicron)	2.94 (1.99 - 3.54)
BA.4.* (Omicron)	1.53 (1.24 - 1.81)
BA.5.* (Omicron)	2.32 (1.69 - 3.66)

Note: values are presented as the median with the first and third quartiles in parentheses.

Table 13 – Infection Fatality Rate (IFR) for different ranges of the fully vaccinated population against COVID-19 in the 41 largest Brazilian municipalities.

Population fully vaccinated	Infection Fatality Rate (IFR)
0%	0.32% (0.21 - 0.49)
$> 0\%$ and $\leq 10\%$	0.29% (0.23 - 0.37)
$> 10\%$ and $\leq 20\%$	0.23% (0.17 - 0.29)
$> 20\%$ and $\leq 30\%$	0.18% (0.13 - 0.24)
$> 30\%$ and $\leq 40\%$	0.15% (0.11 - 0.21)
$> 40\%$ and $\leq 50\%$	0.13% (0.09 - 0.19)
$> 50\%$ and $\leq 60\%$	0.11% (0.07 - 0.15)
$> 60\%$ and $\leq 70\%$	0.06% (0.04 - 0.10)
$> 70\%$ and $\leq 80\%$	0.04% (0.02 - 0.07)
$> 80\%$	0.03% (0.01 - 0.05)

Note: values are presented as the median with the first and third quartiles in parentheses.

We analyzed the correlations among the epidemic parameters and the usual interventions against COVID-19, such as social isolation and vaccination. We pay special attention to outlier cities that exhibited notably elevated mortality rates during the study period, exemplified by Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO, as well as those demonstrating reduced mortality rates, such as Feira de Santana/BA, São Luís/MG, and Florianópolis/SC. Despite our efforts, we did not uncover a definitive explanation for why top outlier cities experienced 2.5 times more deaths than their bottom outlier counterparts.

If Δ_H and vaccination could not clearly explain the pandemic outcome in outlier cities, we note that epidemiological parameters estimated are much more explicit in suggesting the differences between outlier groups. Notably, we noted a contrast between these two groups of cities during the initial two years, with death rates in top outliers being 2.65 times higher than those in bottom outliers. By 2022, however, the death rate in top outliers had reduced to 65% higher than in bottom outliers. Also, we note that all outliers reduced R_0 during the early stages of 2020, as depicted in Figure 41a. Subsequently, in early 2021, R_0 increased for all outliers; however, we note that after the initial rise in R_0,

the municipalities from the bottom outlier group were able to maintain a stable R_0, while municipalities such as Cuiabá and Rio de Janeiro experienced R_0 values exceeding 2.5. Finally, the model estimated that the IFR for top outlier municipalities was higher than that observed in bottom outlier municipalities during the initial two years, as illustrated in Figure 41b. However, by 2022, the model indicated a convergence of the IFR in top outliers towards the same pattern observed in bottom outliers, suggesting a mitigating effect of mass vaccination efforts on lethality disparities among cities.

7.2.1 Limitations

We identified twelve cities with vaccination rates exceeding 100% of their population due to noise in the database (Brasil, 2022), so, as discussed in Section 3.3.2, we should analyze these findings cautiously. Also, Google advises caution when using the COVID-19 Community Mobility Report (Google, 2020) for comparisons between locations or periods. Despite this, we assessed that the data remained reasonably reliable and did not damage the analysis conducted in this chapter.

8 CONCLUSIONS

This thesis explores the key factors associated with COVID-19 mortality rates across Brazilian municipalities during the first three years of the pandemic. We examined the relationships between demographic, social, economic, and political factors and mortality through comprehensive analysis. We also examined the impacts of interventions such as social isolation measures and vaccination efforts and the emergence of new variants. Our findings provide critical insights into the epidemiological dynamics of COVID-19 in Brazil and offer recommendations for future health crises. In summary, this study presents mathematical evidence supporting widely held perceptions of COVID-19 in Brazil.

We demonstrated that Brazilian municipalities effectively mitigated the initial COVID-19 wave by reducing the time-varying basic reproduction number (R_0) through robust social isolation measures. Cross-correlation analysis showed that early preventive actions in 2020 played a crucial role in controlling the spread of the virus. However, a shift to reactive social isolation in 2021, combined with the emergence of Gamma and Delta variants, correlates with higher R_0 values and increased mortality. Notably, the vaccination effort, beginning in 2021 and consolidating in 2022, is significantly associated with a reduced IFR, culminating in a convergent reduction in mortality across municipalities. These findings emphasize the effectiveness of proactive measures in mitigating the impact of the pandemic.

We also demonstrated that municipalities with lower human development experienced a higher correlation with COVID-19 mortality in the early pandemic stages. However, we note a significant change in the associations among sociodemographic factors and mortality across the pandemic. Since 2021, the profile of municipalities more correlated with increased COVID-19 mortality has been that with better social development, which are urbanized municipalities. This finding denotes that urbanization plays a significant and robust role in COVID-19 spreading.

Our insight is that the urban lifestyle contributes to the propagation of the virus and, consequently, elevated COVID-19 mortality in urbanized places. However, our two experiments, the regression analysis and epidemic simulations, denote that the factors impacting the Brazilian municipalities during the COVID-19 pandemic were not static. We argue that measures such as social isolation and vaccination controlled the impact of urban lifestyle in the municipalities.

For instance, our findings show that in the early moment, COVID-19 mortality correlates to variables that denote population poverty, such as unemployment and households in informal settlements. At that moment, the coronavirus outbreak mainly impacted municipalities in the North region, Northeast coast, and metropolitan areas. Our findings also show that the municipalities administrated the outbreak in the early pandemic, reducing the R_0. This reduction correlates with the preventive social isolation noted at that moment. So, this measure can have controlled the outbreak in urbanized municipalities, mitigating the magnitude of deaths in the first wave. Consequently, relevant risk factors related to poverty can be more pronounced in the early moments.

The year 2021 reported 60% of deaths in Brazil in the study period, denoting that the pandemic was out of control. We noted disproportional impact in Southeast, Midwestern, and South, the more urbanized regions. That year, the percentage of urban population variable consolidated as the most robust sociodemographic factor correlated with COVID-19 mortality in the Brazilian municipalities. At that moment, the municipalities left preventive social isolation and employed reactive social isolation. Our findings show that the municipalities failed to reduce the R_0 quickly, which was different from the observed in the first wave. So, the urban factor was uncontrolled, leading to increased deaths because the urbanized municipalities represent the highest proportion of the national population.

On the other hand, in 2022, the vaccination collaborated to mitigate the urban effect of the pandemic. That year, the municipalities already reported a significant proportion of the fully vaccinated population against COVID-19. Additionally, this year had the Omicron variant as dominant, characterized by low lethality (Beppu et al., 2024; Rana et al., 2022; Nyberg et al., 2022; Ward et al., 2022; Lorenzo-Redondo et al., 2022). We note a significant reduction of the Infection Fatality Rate (IFR) at that moment. With the urban factor controlled, we note that the percentage of the elderly population, a notable risk factor (Richardson et al., 2020), emerged as the strongest correlation with COVID-19 mortality.

Our study also showed that political preference is associated with COVID-19 mortality, being the percentage of votes for Jair Bolsonaro in the 2022 Presidential Election, the variable investigated that reported the strongest correlation. Our findings also identified correlations for other sociodemographic attributes: percentage of Indigenous people, percentage of service workers, and expected years of schooling at age 18. Denoting that an interplay of factors influences COVID-19 mortality, levering different vulnerabilities for different regions.

The implications of this thesis for health authorities and policymakers is to provide methods and insights to help analyze and monitor future epidemics. For the COVID-19 pandemic, we found that proactive measures such as social isolation and vaccination are related to mitigating the factor played by urbanization in the increasing mortality. So, for epidemics similar to COVID-19, taking measures to account for urbanization vulnerabilities is essential. Additionally, measures to control vulnerabilities related to poverty in the

municipalities are significant since our findings show a positive correlation for variables such as households in informal settlements in the early pandemic. Specific places also demand increased care, as our findings suggest that municipalities with higher Indigenous populations correlate with more mortality. Finally, combating misinformation in health crises also plays an important role since information propagated by political leaders can influence people to increase their exposure to a virus, aggravating the risk of vulnerable groups, such as the vulnerability of elderly individuals to COVID-19.

This thesis makes several contributions. First, we highlight the critical importance of incorporating temporal exposure into regression analyses of COVID-19 mortality, enabling a more nuanced understanding of the evolving factors driving mortality rates. Second, we underscore the need for a dynamic, time-sensitive approach to assessing COVID-19 mortality risk in Brazilian municipalities, reflecting the shifting correlations between sociodemographic factors and mortality throughout the pandemic.

Additionally, we developed a modified epidemiological model prepared for analyzing multi-outbreak epidemics, which incorporates fuzzy transitions between epidemic phases to account for temporal variations. We demonstrated the robustness of the model through its application to national-level data from Brazil, Spain, the United Kingdom, and the United States, as well as municipal-level data from the 41 largest Brazilian municipalities. We validated the model accuracy by comparing its outcomes with serological survey data (Hallal et al., 2020; Pérez-Gómez et al., 2023; Public Health England, 2020c,a,b; Walker et al., 2021; Anand et al., 2020), and we demonstrated application of the model for generating COVID-19 death forecasts for Brazilian municipalities.

In forecast application, we recommend employing a diverse set of models in epidemiological surveillance to enhance the reliability and comprehensiveness of medium-term predictions. Our model also contributes to a deeper analysis of the pandemic in Brazil, shedding light on the potential underreporting of COVID-19 cases in official datasets. Finally, we analyzed the relationship between the model outcomes and critical factors such as social isolation, vaccination rates, and the emergence of new variants, offering a holistic perspective on the drivers shaping the pandemic trajectory.

We hope this thesis contributes to advancing the computational epidemiology field. We aspire that our findings contribute to a comprehensive understanding of the COVID-19 pandemic in Brazilian municipalities and that our methods can help health authorities and policymakers improve their decision-making for future epidemics. Furthermore, we desire this thesis to motivate researchers to produce deeper investigations and methods to understand the local factors aggravating the impact of epidemics.

8.1 Limitations

The main limitation of this thesis lies in its dependence on ecological data. So, the associations and conclusions based on ecological data do not allow causal inferences (Szklo and Nieto, 2014). Despite this limitation, our findings can provide valuable contributions to hypothesis formulation and support the development of public policies (Bonita et al., 2006).

8.2 Future work

This thesis opens multiple perspectives for future research. A more detailed exploration of the urbanization factor can enhance our regression analysis of variables correlated with COVID-19 mortality. We plan to conduct a higher granularity analysis of the urban population percentage using census tract-level data to uncover more nuanced associations. Refined data on population density may reveal stronger correlations.

Another promising work involves adapting our Fuzzy SIRDS model to other diseases, such as dengue, chikungunya, Mpox, and avian influenza. In other work related to our model, we aim to analyze the optimal objective function for calibrating the model to COVID-19 data, using as ground truth the serological research data from Brazil, Spain, the United Kingdom, and the United States. Within this line of work, we also intend to explore alternative stochastic metaheuristics, such as Cross-Entropy Optimization (CEopt) and PSO, to improve parameter optimization. Also, we aim to compare our approach with other methods for smoothing shifts between epidemic periods, including multistep logistic-like functions, spline-based models, Gaussian processes, and Kalman filters.

For the forecasting application, we propose integrating our epidemiological model with spatial virus spreading models to enhance predictions for municipalities. Additionally, a comparison of model performance in both nowcasting and retrospective analysis could provide valuable insights. Another important direction is the development of a practical tool to assist epidemiological surveillance professionals in evaluating and selecting among different model outputs, making model-based decision support more accessible and actionable.

This thesis identifies large municipalities with atypical COVID-19 mortality rates, categorizing them as top outliers (Cuiabá/MT, Rio de Janeiro/RJ, and Goiânia/GO) and bottom outliers (Feira de Santana/BA, São Luís/MA, and Florianópolis/SC). While our analysis primarily attributes these outlier patterns to variations in R_0 and IFR, future studies should conduct a more in-depth investigation into correlated factors, accounting for population-specific variables in these municipalities to better explain these discrepancies. One potential strategy could involve adopting an augmented level of granularity, examining COVID-19 data at the neighborhood level within these cities.

By following these directions, we aim to deepen the understanding of the local factors driving epidemic dynamics and refine predictive and explanatory models for public health applications.

8.3 Publications

1. Lima, H. S., Tupinambás, U., and Guimarães, F. G. (2024c). Estimating time-varying epidemiological parameters and underreporting of Covid-19 cases in Brazil using a mathematical model with fuzzy transitions between epidemic periods. *PLOS ONE*, 19(6):1–35.

2. Lima, H. and Guimarães, F. (2024). Analyzing the COVID-19 parameters for large Brazilian municipalities using a model with fuzzy transitions between epidemic periods. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 70–81, Porto Alegre, RS, Brasil. SBC.

3. Lima, H. S., Silva, P. C. d. L. e., Meira Jr., W., Tupinambás, U., Costa, M. A., and Guimarães, F. G. (2024b). Sociodemographic Factors Associated with Covid-19 Mortality in Brazilian Municipalities Across Three Years: an Approach Supported by Gaussian Mixture Clustering. *Spatial and Spatio-Temporal Epidemiology*. Under review.

REFERENCES

Abdy, M., Side, S., Annas, S., Nur, W., and Sanusi, W. (2021). An SIR epidemic model for COVID-19 spread with fuzzy parameter: the case of Indonesia. *Advances in difference equations*, 2021:1–17.

Abolpour, R., Siamak, S., Mohammadi, M., Moradi, P., and Dehghani, M. (2021). Linear parameter varying model of COVID-19 pandemic exploiting basis functions. *Biomedical Signal Processing and Control*, 70:102999.

Acuña-Zegarra, M. A., Santana-Cibrian, M., and Velasco-Hernandez, J. X. (2020). Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance. *Mathematical Biosciences*, 325:108370.

Adiga, A., Wang, L., Hurt, B., Peddireddy, A., Porebski, P., Venkatramanan, S., Lewis, B. L., and Marathe, M. (2021). All Models Are Useful: Bayesian Ensembling for Robust High Resolution COVID-19 Forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 2505–2513, New York, NY, USA. Association for Computing Machinery.

Afzal, A., Ansari, Z., Alshahrani, S., Raj, A. K., Kuruniyan, M. S., Saleel, C. A., and Nisar, K. S. (2021). Clustering of COVID-19 data for knowledge discovery using c-means and fuzzy c-means. *Results in Physics*, 29:104639.

Ajzenman, N., Cavalcanti, T., and Da Mata, D. (2023). More than Words: Leaders' Speech and Risky Behavior during a Pandemic. *American Economic Journal: Economic Policy*, 15(3):351–371.

Al-Rashedi, A. and Al-Hagery, M. A. (2023). Deep Learning Algorithms for Forecasting COVID-19 Cases in Saudi Arabia. *Applied Sciences*, 13(3).

Ala'raj, M., Majdalawieh, M., and Nizamuddin, N. (2021). Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections. *Infectious Disease Modelling*, 6:98–111.

Albani, V. V., Albani, R. A., Bobko, N., Massad, E., and Zubelli, J. P. (2022). On the role of financial support programs in mitigating the SARS-CoV-2 spread in Brazil. *BMC Public Health*, 22(1):1781.

Almeida, G., Vilches, T., Ferreira, C., and Fortaleza, C. (2021). Addressing the COVID-19 transmission in inner Brazil by a mathematical model. *Scientific Reports*, 11(1):10760.

Alzahrani, S. I., Aljamaan, I. A., and Al-Fakih, E. A. (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. *Journal of Infection and Public Health*, 13(7):914–919.

Anand, S., Montez-Rath, M., Han, J., Bozeman, J., Kerschmann, R., Beyer, P., Parsonnet, J., and Chertow, G. M. (2020). Prevalence of SARS-CoV-2 antibodies in a large nationwide sample of patients on dialysis in the USA: a cross-sectional study. *The Lancet*, 396(10259):1335–1344.

Appadu, A., Kelil, A., and Tijani, Y. (2021). Comparison of some forecasting methods for COVID-19. *Alexandria Engineering Journal*, 60(1):1565–1589.

Arora, P., Kumar, H., and Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals*, 139:110017.

Arroyo-Marioli, F., Bullano, F., Kucinskas, S., and Rondón-Moreno, C. (2021). Tracking R of COVID-19: A new real-time estimation using the Kalman filter. *PloS one*, 16(1):e0244474. https://doi.org/10.1371/journal.pone.0244474.

Asano, C. L., Ventura, D. d. F. L., Aith, F. M. A., Reis, R. R., and Ribeiro, T. B. (2021). Direito e pandemia: ordem jurídica e sistema judiciário não foram suficientes para evitar graves violações. *DIREITOS NA PANDEMIA-MAPEAMENTO E ANÁLISE DAS NORMAS JURÍDICAS DE RESPOSTA À COVID-19 NO BRASIL*.

Asher, M., Lomax, N., Morrissey, K., Spooner, F., and Malleson, N. (2023). Dynamic calibration with approximate bayesian computation for a microsimulation of disease spread. *Scientific Reports*, 13(1):8637.

Awuah-Mensah, Y. K. and Aidoo, E. N. (2024). Modelling the spatial varying relationships between socioeconomic inequalities and COVID-19 mortality in the African subregion. *Earth Science Informatics*, pages 1–13.

Azevedo, W. O. d. D., Santa Cruz, F., Dias, J. C., Davidovich, L., and Moreira, I. d. C. (2020). Pacto pela vida e pelo Brasil.

Babashova, S. (2022). Predicting the Dynamics of Covid-19 Propagation in Azerbaijan based on Time Series Models. *WSEAS Transactions on Environment and Development*, 18:1036–1048. Publisher: WSEAS.

Barberia, L. G. and Gómez, E. J. (2020). Political and institutional perils of Brazil's COVID-19 crisis. *Lancet (London, England)*, 396(10248):367. https://doi.org/10.1016/S0140-6736(20)31681-0.

Barbosa, I. R., Galvão, M. H. R., Souza, T. A. d., Gomes, S. M., Medeiros, A. d. A., and Lima, K. C. d. (2020). Incidence of and mortality from COVID-19 in the older Brazilian population and its relationship with contextual indicators: an ecological study. *Revista Brasileira de Geriatria e Gerontologia*, 23:e200171.

Barnard, R. C., Davies, N. G., Jit, M., and Edmunds, W. J. (2022). Modelling the medium-term dynamics of SARS-CoV-2 transmission in England in the Omicron era. *Nature Communications*, 13(1):4879. Publisher: Nature Publishing Group.

Barreto, I. C. d. H. C., Costa Filho, R. V., Ramos, R. F., Oliveira, L. G. d., Martins, N. R. A. V., Cavalcante, F. V., Andrade, L. O. M. d., and Santos, L. M. P. (2021). Health collapse in Manaus: the burden of not adhering to non-pharmacological measures to reduce the transmission of Covid-19. *Saúde em Debate*, 45:1126–1139.

Bartholo, T. L., Koslinski, M. C., Tymms, P., and Castro, D. L. (2022). Learning loss and learning inequality during the Covid-19 pandemic. *Ensaio: Avaliação e Políticas Públicas em Educação*.

Bastos, S. B. and Cajueiro, D. O. (2020). Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil. *Scientific Reports*, 10(1):19457.

Bastos, S. B., Morato, M. M., Cajueiro, D. O., and Normey-Rico, J. E. (2021). The COVID-19 (SARS-CoV-2) uncertainty tripod in Brazil: Assessments on model-based predictions with large under-reporting. *Alexandria Engineering Journal*, 60(5):4363–4380. https://doi.org/10.1016/j.aej.2021.03.004.

Batistela, C. M., Correa, D. P., Bueno, Á. M., and Piqueira, J. R. C. (2021). SIRSi compartmental model for COVID-19 pandemic with immunity loss. *Chaos, Solitons & Fractals*, 142:110388.

Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., and Favre, G. (2020). Real estimates of mortality following COVID-19 infection. *The Lancet infectious diseases*, 20(7):773.

Bender, R. (2009). Introduction to the Use of Regression Models in Epidemiology. In Walker, J. M. and Verma, M., editors, *Cancer Epidemiology*, volume 471, pages 179–195. Humana Press, Totowa, NJ. Series Title: Methods in Molecular Biology.

Beppu, H., Fukuda, T., Otsubo, N., Akihisa, T., Kawanishi, T., Ogawa, T., Abe, Y., Endo, M., Hanawa, T., Sugita, C., Kikkawa, Y., Yamada, T., and Wakai, S. (2024). Comparative

outcomes of hemodialysis patients facing pre-Omicron and Omicron COVID-19 epidemics. *Therapeutic Apheresis and Dialysis*, 28(1):51–60.

Bermudi, P. M. M., Lorenz, C., Aguiar, B. S. d., Failla, M. A., Barrozo, L. V., and Chiaravalloti-Neto, F. (2021). Spatiotemporal ecological study of COVID-19 mortality in the city of São Paulo, Brazil: Shifting of the high mortality risk from areas with the best to those with the worst socio-economic conditions. *Travel Medicine and Infectious Disease*, 39:101945.

Bhatia, S., Parag, K. V., Wardle, J., Nash, R. K., Imai, N., Elsland, S. L. V., Lassmann, B., Brownstein, J. S., Desai, A., Herringer, M., Sewalk, K., Loeb, S. C., Ramatowski, J., Cuomo-Dannenburg, G., Jauneikaite, E., Unwin, H. J. T., Riley, S., Ferguson, N., Donnelly, C. A., Cori, A., and Nouvellet, P. (2023). Retrospective evaluation of real-time estimates of global COVID-19 transmission trends and mortality forecasts. *PLOS ONE*, 18(10):1–17.

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S. A., Zhang, T., et al. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet infectious diseases*, 20(8):911–919. https://doi.org/10.1016/S1473-3099(20)30287-5.

Birello, P., Re Fiorentin, M., Wang, B., Colizza, V., and Valdano, E. (2024). Estimates of the reproduction ratio from epidemic surveillance may be biased in spatially structured populations. *Nature Physics*, 20(7):1204–1210. Publisher: Nature Publishing Group.

Bjørnstad, O. N., Shea, K., Krzywinski, M., and Altman, N. (2020a). Modeling infectious epidemics. *Nat. Methods*, 17(5):455–456. https://doi.org/10.1038/s41592-020-0822-z.

Bjørnstad, O. N., Shea, K., Krzywinski, M., and Altman, N. (2020b). The seirs model for infectious disease dynamics. *Nature methods*, 17(6):557–559.

Bonita, R., Beaglehole, R., and Kjellström, T. (2006). *Basic epidemiology*. World Health Organization.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer.

Brasil (2022). Campanha Nacional de Vacinação contra Covid-19. https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao . Accessed: Oct 25, 2024.

Brida, J. G., Alvarez, E., Limas, E., et al. (2021). Clustering of time series for the analysis of the COVID-19 pandemic evolution. *Economics Bulletin*, 41(3):1082–1096.

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., and Zhilin, O. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, 1(3).

Cantó, B., Coll, C., and Sánchez, E. (2017). Estimation of parameters in a structured SIR model. *Advances in Difference Equations*, 2017(1):33.

Carr, S., Unwin, N., and Pless-Mulloli, T. (2007). *An introduction to public health and epidemiology*. McGraw-Hill Education (UK).

Carrillo-Larco, R. M. and Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. *Wellcome Open Research*, 5.

Castillo, O., Castro, J. R., and Melin, P. (2023). Forecasting the covid-19 with interval type-3 fuzzy logic and the fractal dimension. *International Journal of Fuzzy Systems*, 25(1):182–197.

Castillo, O. and Melin, P. (2021). A new fuzzy fractal control approach of non-linear dynamic systems: The case of controlling the COVID-19 pandemics. *Chaos, Solitons & Fractals*, 151:111250.

Castro-Alves, J., Silva, L. S., Lima, J. P., and Ribeiro-Alves, M. (2022). Were the socio-economic determinants of municipalities relevant to the increment of COVID-19 related deaths in Brazil in 2020? *PLoS ONE*, 17(4):e0266109.

Cavalini, L. T. and de Leon, A. C. M. P. (2008). Morbidity and mortality in Brazilian municipalities: a multilevel study of the association between socioeconomic and healthcare indicators. *International Journal of Epidemiology*, 37(4):775–783.

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of The Total Environment*, 729:138817.

Chimmula, V. K. R. and Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, 135:109864.

Choi, S. and Ki, M. (2020). Estimating the reproductive number and the outbreak size of COVID-19 in Korea. *Epidemiology and health*, 42:e2020011.

Cintrón-Arias, A., Castillo-Chávez, C., Betencourt, L., Lloyd, A. L., and Banks, H. T. (2008). The estimation of the effective reproductive number from disease outbreak data. In: North Carolina State University [Internet]. Available from: https://repository.lib.ncsu.edu/bitstream/handle/1840.4/4036/crsc-tr08-08.pdf?sequence=1. Accessed: Jan 12, 2024.

Clouston, S. A., Natale, G., and Link, B. G. (2021a). Socioeconomic inequalities in the spread of coronavirus-19 in the united states: A examination of the emergence of social inequalities. *Social Science & Medicine*, 268:113554.

Clouston, S. A. P., Natale, G., and Link, B. G. (2021b). Socioeconomic inequalities in the spread of coronavirus-19 in the United States: A examination of the emergence of social inequalities. *Social Science & Medicine*, 268:113554.

Coelho, F. C., Lana, R. M., Cruz, O. G., Villela, D. A. M., Bastos, L. S., Piontti, A. P. y., Davis, J. T., Vespignani, A., Codeço, C. T., and Gomes, M. F. C. (2020). Assessing the spread of COVID-19 in Brazil: Mobility, morbidity and social vulnerability. *PLOS ONE*, 15(9):e0238214. Publisher: Public Library of Science.

Cohen, K. W., Linderman, S. L., Moodie, Z., Czartoski, J., Lai, L., Mantus, G., Norwood, C., Nyhoff, L. E., Edara, V. V., Floyd, K., et al. (2021). Longitudinal analysis shows durable and broad immune memory after SARS-CoV-2 infection with persisting antibody responses and memory B and T cells. *Cell Reports Medicine*, 2(7):100354. https://doi.org/10.1016/j.xcrm.2021.100354.

Congdon, P. (2021). Mid-Epidemic Forecasts of COVID-19 Cases and Deaths: A Bivariate Model Applied to the UK. *Interdisciplinary Perspectives on Infectious Diseases*, 2021(1):8847116.

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9):1505–1512. https://doi.org/10.1093/aje/kwt133.

Croda, M. G., Barbosa, M. d. S., Marchioro, S. B., Nascimento, D. D. G. d., Melo, E. C. P., Cruz, O. G., Torres, A. J. L., Oliveira, L. A. d., Ganem, F., and Simionatto, S. (2022). The first year of the COVID-19 pandemic in an indigenous population in Brazil: an epidemiological study. *Revista do Instituto de Medicina Tropical de São Paulo*, 64:e69.

Cunha Jr, A., Barton, D. A., and Ritto, T. G. (2023). Uncertainty quantification in mechanistic epidemic models via cross-entropy approximate Bayesian computation. *Nonlinear dynamics*, 111(10):9649–9679.

Dairi, A., Harrou, F., Zeroual, A., Hittawe, M. M., and Sun, Y. (2021). Comparative study of machine learning methods for COVID-19 transmission forecasting. *Journal of Biomedical Informatics*, 118:103791.

Dansana, D., Kumar, R., Bhattacharjee, A., and Mahanty, C. (2022). COVID-19 Outbreak Prediction and Analysis of E-Healthcare Data Using Random Forest Algorithms. *International Journal of Reliable and Quality E-Healthcare (IJRQEH)*, 11(1):1–13. Publisher: IGI Global.

Das, A., Ghosh, S., Das, K., Basu, T., Das, M., and Dutta, I. (2020). Modeling the effect of area deprivation on COVID-19 incidences: a study of Chennai megacity, India. *Public Health*, 185:266–269.

DATASUS (2020a). Painel de casos de doença pelo coronavírus 2019 (COVID-19) no Brasil pelo Ministério da Saúde. Database: Gov.BR [Internet]. Available from: https://covid.saude.gov.br/. Accessed: Nov 23, 2023.

DATASUS (2020b). Tabnet. Database: Gov.BR [Internet]. Available from: https://datasus.saude.gov.br/informacoes-de-saude-tabnet/. Accessed: Nov 23, 2023.

DATASUS (2022a). Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19. Database: Gov.BR [Internet]. Available from: https://opendatasus.saude.gov.br. Accessed: Nov 23, 2023.

DATASUS (2022b). Sistema de Informação sobre Mortalidade – SIM. Database: Gov.BR [Internet]. Available from: https://opendatasus.saude.gov.br/fa_IR/dataset/sim. Accessed: Nov 23, 2023.

Davarci, O. O., Yang, E. Y., Viguerie, A., Yankeelov, T. E., and Lorenzo, G. (2023). Dynamic parameterization of a modified SEIRD model to analyze and forecast the dynamics of COVID-19 outbreaks in the United States. *Engineering with Computers*.

De Angelis, E., Renzetti, S., Volta, M., Donato, F., Calza, S., Placidi, D., Lucchini, R. G., and Rota, M. (2021). COVID-19 incidence and mortality in Lombardy, Italy: An ecological study on the role of air pollution, meteorological factors, demographic and socioeconomic variables. *Environmental Research*, 195:110777.

De Angelis, E., Renzetti, S., Volta, M., Donato, F., Calza, S., Placidi, D., Lucchini, R. G., and Rota, M. (2021). Covid-19 incidence and mortality in lombardy, italy: An ecological study on the role of air pollution, meteorological factors, demographic and socioeconomic variables. *Environmental Research*, 195:110777.

de Araújo Morais, L. R. and da Silva Gomes, G. S. (2022). Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model. *Applied Soft Computing*, 126:109315.

de Leon, F. L. L., Malde, B., and McQuillin, B. (2023). The effects of emergency government cash transfers on beliefs and behaviours during the COVID pandemic: Evidence from Brazil. *Journal of Economic Behavior & Organization*, 208:140–155.

de Souza, W. M., Buss, L. F., Candido, D. d. S., Carrera, J.-P., Li, S., Zarebski, A. E., Pereira, R. H. M., Prete Jr, C. A., de Souza-Santos, A. A., Parag, K. V., et al. (2020). Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nature human behaviour*, 4(8):856–865. https://doi.org/10.1038/s41562-020-0928-4.

Dean, R. T. and Dunsmuir, W. T. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior research methods*, 48:783–802.

Delli Compagni, R., Cheng, Z., Russo, S., and Van Boeckel, T. P. (2022). A hybrid Neural Network-SEIR model for forecasting intensive care occupancy in Switzerland during COVID-19 epidemics. *Plos one*, 17(3):e0263789.

Demenech, L. M., Dumith, S. d. C., Vieira, M. E. C. D., and Neiva-Silva, L. (2020). Income inequality and risk of infection and death by COVID-19 in Brazil. *Revista Brasileira de Epidemiologia*, 23:e200095.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dessie, Z. G. and Zewotir, T. (2021). Mortality-related risk factors of COVID-19: a systematic review and meta-analysis of 42 studies and 423,117 patients. *BMC Infectious Diseases*, 21(1):855.

Devaraj, J., Madurai Elavarasan, R., Pugazhendhi, R., Shafiullah, G. M., Ganesan, S., Jeysree, A. K., Khan, I. A., and Hossain, E. (2021). Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant? *Results in Physics*, 21:103817.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431.

Dobson, A. J. and Barnett, A. G. (2018). *An introduction to generalized linear models*. Chapman and Hall/CRC.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.

Drews, M., Kumar, P., Singh, R. K., De La Sen, M., Singh, S. S., Pandey, A. K., Kumar, M., Rani, M., and Srivastava, P. K. (2022). Model-based ensembles: Lessons learned from retrospective analysis of COVID-19 infection forecasts across 10 countries. *Science of The Total Environment*, 806:150639.

Dunn, P. K., Smyth, G. K., et al. (2018). *Generalized linear models with examples in R*, volume 53. Springer.

Dutta, A. (2022). Covid-19 waves: Variant dynamics and control. *Scientific Reports*, 12(1):1–9.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

El-Gilany, A.-H. (2021). Covid-19 caseness: An epidemiologic perspective. *Journal of Infection and Public Health*, 14(1):61–65.

El-Shabasy, R. M., Nayel, M. A., Taher, M. M., Abdelmonem, R., Shoueir, K. R., et al. (2022). Three wave changes, new variant strains, and vaccination effect against COVID-19 pandemic. *International Journal of Biological Macromolecules*.

Erandathi, M., Chung Wang, W. Y., and Hsieh, C.-C. (2021). Clustering the countries for quantifying the status of Covid-19 through time series analysis. *Information Discovery and Delivery*, ahead-of-print(ahead-of-print).

Fan, C., Meng, Y., Sun, X., Wu, F., Zhang, T., and Li, J. (2021). Parameter estimation for the SEIR model using recurrent nets. *arXiv preprint arXiv:2105.14524*.

Fanelli, D. and Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals*, 134:109761.

Fantin, R., Barboza-Solís, C., Hildesheim, A., and Herrero, R. (2023). Excess mortality from COVID 19 in Costa Rica: a registry based study using Poisson regression. *The Lancet Regional Health – Americas*, 20. Publisher: Elsevier.

Farooq, J. and Bazaz, M. A. (2021). A deep learning algorithm for modeling and forecasting of COVID-19 in five worst affected states of India. *Alexandria Engineering Journal*, 60(1):587–596.

Ferrante, L., Duczmal, L., Steinmetz, W. A., Almeida, A. C. L., Leão, J., Vassão, R. C., Tupinambás, U., and Fearnside, P. M. (2021a). How Brazil's President turned the country into a global epicenter of COVID-19. *Journal of Public Health Policy*, 42(3):439–451. https://doi.org/10.1057/s41271-021-00302-0.

Ferrante, L., Duczmal, L. H., Capanema, E., Steinmetz, W. A. C., Almeida, A. C. L., Leão, J., Vassao, R. C., Fearnside, P. M., and Tupinambás, U. (2022). Dynamics of COVID-19 in Amazonia: a history of government denialism and the risk of a third wave. *Preventive medicine reports*, 26:101752. https://doi.org/10.1016/j.pmedr.2022.101752.

Ferrante, L., Duczmal, L. H., Steinmetz, W. A., Almeida, A. C. L., Leão, J., Vassão, R. C., Tupinambás, U., and Fearnside, P. M. (2021b). Brazil's COVID-19 epicenter in manaus: how much of the population has already been exposed and are vulnerable to SARS-COV-2? *Journal of racial and ethnic health disparities*, pages 1–7. https://doi.org/10.1007/s40615-021-01148-8.

Figueiredo, A. M. d., Figueiredo, D. C. M. M. d., Gomes, L. B., Massuda, A., Gil-García, E., Vianna, R. P. d. T., and Daponte, A. (2020). Social determinants of health and COVID-19 infection in Brazil: an analysis of the pandemic. *Revista brasileira de enfermagem*, 73.

Fiocruz (2020). Fiocruz's genomics network. Database: Fiocruz [Internet]. Available from: https://www.genomahcov.fiocruz.br/en/. Accessed: Feb 20, 2024.

Friston, K. J., Parr, T., Zeidman, P., Razi, A., Flandin, G., Daunizeau, J., Hulme, O. J., Billig, A. J., Litvak, V., Price, C. J., et al. (2020). Second waves, social distancing, and the spread of COVID-19 across the USA. *Wellcome Open Research*, 5.

Gallos, L. K. and Fefferman, N. H. (2015). The effect of disease-induced mortality on structural network properties. *PloS one*, 10(8):e0136704. https://doi.org/10.1371/journal.pone.0136704.

Gao, S., Shen, M., Wang, X., Wang, J., Martcheva, M., and Rong, L. (2023). A multi-strain model with asymptomatic transmission: Application to COVID-19 in the US. *Journal of Theoretical Biology*, 565:111468.

Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R., and Klein, S. L. (2020). Impact of sex and gender on COVID-19 outcomes in Europe. *Biology of sex differences*, 11:1–13.

Ghosh, K. and Ghosh, A. K. (2022). Study of COVID-19 epidemiological evolution in India with a multi-wave SIR model. *Nonlinear Dynamics*, 109(1):47–55. https://doi.org/10.1007/s11071-022-07471-x.

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., et al. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC public health*, 14(1):1–17. https://doi.org/10.1186/1471-2458-14-147.

Giesecke, J. (2014). Primary and index cases. *The Lancet*, 384(9959):2024.

Gohari, K., Kazemnejad, A., Sheidaei, A., and Hajari, S. (2022). Clustering of countries according to the COVID-19 incidence and mortality rates. *BMC Public Health*, 22(1):1–12.

Google (2020). Google COVID-19 community mobility reports. Available from: https://www.google.com/covid19/mobility/. Aaccessed: May 5, 2023].

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., et al. (2020). Practical considerations for measuring the effective reproductive number, R_t. *PLoS computational biology*, 16(12):e1008409.

Hale, T., Angrist, N., Hale, A. J., Kira, B., Majumdar, S., Petherick, A., Phillips, T., Sridhar, D., Thompson, R. N., Webster, S., et al. (2021). Government responses and covid-19 deaths: Global evidence across multiple pandemic waves. *PLoS One*, 16(7):e0253116.

Hallal, P. C., Hartwig, F. P., Horta, B. L., Silveira, M. F., Struchiner, C. J., Vidaletti, L. P., Neumann, N. A., Pellanda, L. C., Dellagostin, O. A., Burattini, M. N., et al. (2020). SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys. *The Lancet Global Health*, 8(11):e1390–e1398. https://doi.org/10.1016/S2214-109X(20)30387-9.

Halpern, C. (2021). Distant learning: The experiences of Brazilian schoolteachers during the COVID-19 school closures. *Journal of Ethnic and Cultural Studies*, 8(1):206–225.

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental monitoring and assessment*, 32:135–154.

Hasan, A., Putri, E., Susanto, H., and Nuraini, N. (2022). Data-driven modeling and forecasting of COVID-19 outbreak for public policy making. *ISA Transactions*, 124:135–143.

Hastenreiter Filho, H. N. and Cavalcante, L. R. (2022). Variáveis associadas à mortalidade por covid-19 nos municípios brasileiros: um estudo exploratório. *RPER*, page 57–70.

Hernandez-Matamoros, A., Fujita, H., Hayashi, T., and Perez-Meana, H. (2020). Forecasting of COVID19 per regions using ARIMA models and polynomial functions. *Applied Soft Computing*, 96:106610.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hoebel, J., Michalski, N., Diercke, M., Hamouda, O., Wahrendorf, M., Dragano, N., and Nowossadeck, E. (2021). Emerging socio-economic disparities in COVID-19-related deaths during the second pandemic wave in Germany. *International Journal of Infectious Diseases*, 113:344–346.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, 395(10223):497–506.

IBGE (2010). Censo 2010. Data retrieved from Censo 2010, https://censo2010.ibge.gov.br/.

IBGE (2020). Aglomerados subnormais 2019 : classificação preliminar e informações de saúde para o enfrentamento à COVID-19 : notas técnicas. Technical report, Instituto Brasileiro de Geografia e Estatística (IBGE).

IBGE (2022). Censo 2022. Available from: https://censo2022.ibge.gov.br/. Accessed: Jul 29, 2023.

Iyengar, K. P., Ish, P., Botchu, R., Jain, V. K., and Vaishya, R. (2022). Influence of the Peltzman effect on the recurrent COVID-19 waves in Europe. *Postgraduate medical journal*, 98(e2):e110–e111.

Jin, J.-M., Bai, P., He, W., Wu, F., Liu, X.-F., Han, D.-M., Liu, S., and Yang, J.-K. (2020). Gender differences in patients with COVID-19: focus on severity and mortality. *Frontiers in public health*, 8:545030.

Johnson, D. P. and Owusu, C. (2024). Examining associations between social vulnerability indices and COVID-19 incidence and mortality with spatial-temporal Bayesian modeling. *Spatial and Spatio-temporal Epidemiology*, 48:100623.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Jung, S., Kim, J.-H., Hwang, S.-S., Choi, J., and Lee, W. (2023). Modified susceptible–exposed–infectious–recovered model for assessing the effectiveness of non-pharmaceutical interventions during the COVID-19 pandemic in Seoul. *Journal of Theoretical Biology*, 557:111329.

Juyal, D., Pal, S., Thaledi, S., and Pandey, H. C. (2021). COVID-19: The vaccination drive in India and the Peltzman effect. *Journal of Family Medicine and Primary Care*, 10(11):3945. https://doi.org/10.4103/jfmpc.jfmpc_739_21.

Kamalov, F., Rajab, K., Cherukuri, A. K., Elnagar, A., and Safaraliev, M. (2022). Deep learning for Covid-19 forecasting: State-of-the-art review. *Neurocomputing*, 511:142–154.

Kamrujjaman, M., Saha, P., Islam, M. S., and Ghosh, U. (2022). Dynamics of SEIR model: A case study of COVID-19 in Italy. *Results in Control and Optimization*, 7:100119.

Keeling, M. J. and Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton university press.

Keil, R. (2021). Covid-19: pandemic on an urban planet. In *COVID-19 and Similar Futures: Pandemic Geographies*, pages 259–265. Springer.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Lalwani, P., Salgado, B. B., Pereira Filho, I. V., da Silva, D. S. S., de Morais, T. B. d. N., Jordão, M. F., Barbosa, A. R. C., Cordeiro, I. B., de Souza Neto, J. N., de Assuncao, E. N., et al. (2021). SARS-CoV-2 seroprevalence and associated factors in Manaus, Brazil: baseline results from the DETECTCoV-19 cohort study. *International Journal of Infectious Diseases*, 110:141–150.

Lancet, T. (2020). COVID-19 in Brazil:"So what?". *Lancet (London, England)*, 395(10235):1461.

Lasco, G. (2020). Medical populism and the COVID-19 pandemic. *Global Public Health*, 15(10):1417–1429. https://doi.org/10.1080/17441692.2020.1807581.

Liao, Z., Lan, P., Fan, X., Kelly, B., Innes, A., and Liao, Z. (2021). SIRVD-DL: A COVID-19 deep learning prediction model based on time-dependent SIRVD. *Computers in Biology and Medicine*, 138:104868.

Lima, E. E. C. d., Costa, L. C. C. d., Souza, R. F., Rocha, C. O. d. E., and Ichihara, M. Y. T. (2024a). Presidential election results in 2018-2022 and its association with excess mortality during the 2020-2021 COVID-19 pandemic in Brazilian municipalities. *Cadernos de Saúde Pública*, 40:e00194723. Publisher: Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz.

Lima, H. and Guimarães, F. (2024). Analyzing the COVID-19 parameters for large Brazilian municipalities using a model with fuzzy transitions between epidemic periods. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 70–81, Porto Alegre, RS, Brasil. SBC.

Lima, H. S., Silva, P. C. d. L. e., Meira Jr., W., Tupinambás, U., Costa, M. A., and Guimarães, F. G. (2024b). Sociodemographic Factors Associated with Covid-19 Mortality in Brazilian Municipalities Across Three Years: an Approach Supported by Gaussian Mixture Clustering. *Spatial and Spatio-Temporal Epidemiology*. Under review.

Lima, H. S., Tupinambás, U., and Guimarães, F. G. (2024c). Estimating time-varying epidemiological parameters and underreporting of Covid-19 cases in Brazil using a mathematical model with fuzzy transitions between epidemic periods. *PLOS ONE*, 19(6):1–35.

Liu, Y. and Rocklöv, J. (2021). The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *Journal of travel medicine*, 28(7):taab124.

Liu, Y. and Rocklöv, J. (2022). The effective reproductive number of the Omicron variant of SARS-CoV-2 is several times relative to Delta. *Journal of Travel Medicine*, 29(3):taac037.

Liu, Y., Yu, Y., Zhao, Y., and He, D. (2022). Reduction in the infection fatality rate of Omicron variant compared with previous variants in South Africa. *International Journal of Infectious Diseases*, 120:146–149. https://doi.org/10.1016/j.ijid.2022.04.029.

Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Loomba, S., De Figueiredo, A., Piatek, S. J., De Graaf, K., and Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*, 5(3):337–348.

López, L. and Rodó, X. (2020). The end of social confinement and COVID-19 re-emergence risk. *Nature Human Behaviour*, 4(7):746–755.

Lorenz, C., Bermudi, P. M. M., de Aguiar, B. S., Failla, M. A., Toporcov, T. N., Chiaravalloti-Neto, F., and Barrozo, L. V. (2021). Examining socio-economic factors to understand the hospital case fatality rates of COVID-19 in the city of São Paulo, Brazil. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 115(11):1282–1287.

Lorenzo-Redondo, R., Ozer, E. A., and Hultquist, J. F. (2022). Covid-19: is omicron less lethal than delta? *bmj*, 378. https://doi.org/10.1136/bmj.o1806.

Lotta, G., Wenham, C., Nunes, J., and Pimenta, D. N. (2020). Community health workers reveal COVID-19 disaster in Brazil. *The Lancet*, 396(10248):365–366. https://doi.org/10.1016/S0140-6736(20)31521-X.

López-Bazo, E. (2024). The complex link between socioeconomic deprivation and COVID-19. Evidence from small areas of Catalonia. *Spatial and Spatio-temporal Epidemiology*, 49:100648.

Mahase, E. (2020). Covid-19: death rate is 0.66% and increases with age, study estimates. *BMJ: British Medical Journal (Online)*, 369.

Maleki, M., Bidram, H., and Wraith, D. (2022). Robust clustering of COVID-19 cases across US counties using mixtures of asymmetric time series models with time varying and freely indexed covariates. *Journal of Applied Statistics*, pages 1–15.

Manley, H., Bayley, T., Danelian, G., Burton, L., Finnie, T., Charlett, A., Watkins, N. A., Birrell, P., De Angelis, D., Keeling, M., Funk, S., Medley, G., Pellis, L., Baguelin, M., Ackland, G. J., Hutchinson, J., Riley, S., and Panovska-Griffiths, J. (2024). Combining models to generate consensus medium-term projections of hospital admissions, occupancy and deaths relating to COVID-19 in England. *Royal Society Open Science*, 11(5):231832. Publisher: Royal Society.

Marra, V. and Quartin, M. (2021). A Bayesian estimate of the early COVID-19 infection fatality ratio in Brazil based on a random seroprevalence survey. *International Journal of Infectious Diseases*, 111:190–195. https://doi.org/10.1016/j.ijid.2021.08.016.

Martines, M. R., Ferreira, R. V., Toppa, R. H., Assunção, L., Desjardins, M. R., and Delmelle, E. M. (2021). Detecting space–time clusters of COVID-19 in Brazil: mortality, inequality, socioeconomic vulnerability, and the relative risk of the disease in Brazilian municipalities. *Journal of Geographical Systems*, 23(1):7–36.

Martins, C. M., Gomes, R. Z., Muller, E. V., Borges, P. K. d. O., Coradassi, C. E., and Montiel, E. M. d. S. (2020). PREDICTIVE MODEL FOR COVID-19 INCIDENCE IN A MEDIUM-SIZED MUNICIPALITY IN BRAZIL (PONTA GROSSA, PARANÁ). *Texto & Contexto - Enfermagem*, 29:e20200154. Publisher: Universidade Federal de Santa Catarina, Programa de Pós Graduação em Enfermagem.

Massard, M., Eftimie, R., Perasso, A., and Saussereau, B. (2022). A multi-strain epidemic model for COVID-19 with infected and asymptomatic cases: Application to French data. *Journal of Theoretical Biology*, 545:111117.

Mbuvha, R. and Marwala, T. (2020). Bayesian inference of COVID-19 spreading rates in South Africa. *PLOS ONE*, 15(8):e0237126. Publisher: Public Library of Science.

Melin, P. and Castillo, O. (2021). Spatial and Temporal Spread of the COVID-19 Pandemic Using Self Organizing Neural Networks and a Fuzzy Fractal Approach. *Sustainability*, 13(15):8295.

Melo, G. C. d., Duprat, I. P., Araújo, K. C. G. M. d., Fischer, F. M., and Araújo Neto, R. A. d. (2020). Prediction of cumulative rate of COVID-19 deaths in Brazil: a modeling study. *Revista Brasileira de Epidemiologia*, 23.

Meng, T. (2021). Clusters in the Spread of the COVID-19 Pandemic: Evidence From the G20 Countries. *Frontiers in Public Health*, 8:948.

Millevoi, C., Pasetto, D., and Ferronato, M. (2024). A Physics-Informed Neural Network approach for compartmental epidemiological models. *PLOS Computational Biology*, 20(9):e1012387. Publisher: Public Library of Science.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Morgenstern, H. (1995). Ecologic studies in epidemiology: concepts, principles, and methods. *Annual Review of Public Health*, 16(1):61–81.

Morris, C. P., Eldesouki, R. E., Fall, A., Gaston, D. C., Norton, J. M., Gallagher, N. D., Luo, C. H., Abdullah, O., Klein, E. Y., and Mostafa, H. H. (2022). Sars-cov-2 reinfections during the delta and omicron waves. *JCI insight*, 7(20). https://doi.org/10.1172/jci.insight.162007.

Morrissey, K., Spooner, F., Salter, J., and Shaddick, G. (2021). Area level deprivation and monthly COVID-19 cases: The impact of government policy in England. *Social Science & Medicine*, 289:114413.

Mourao, P. and Junqueira, A. (2021). Through the irregular paths of inequality: An analysis of the evolution of socioeconomic inequality in brazilian states since 1976. *Sustainability*, 13(4).

Mpinganzima, L., Ntaganda, J. M., Banzi, W., Muhirwa, J. P., Nannyonga, B. K., Niyobuhungiro, J., Rutaganda, E., Ngaruye, I., Ndanguza, D., Nzabanita, J., Masabo, E., Gahamanyi, M., Dushimirimana, J., Nyandwi, B., Kurujyibwami, C., Ruganzu, L. F. U., Nyagahakwa, V., Mukeshimana, S., Ngendahayo, J. P., Nsabimana, J. P., Niyigena, J. D. D., Uwonkunda, J., and Mbalawata, I. S. (2023). Compartmental mathematical modelling of dynamic transmission of COVID-19 in Rwanda. *IJID Regions*, 6:99–107.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Nicholson, C., Beattie, L., Beattie, M., Razzaghi, T., and Chen, S. (2022). A machine learning and clustering-based approach for county-level COVID-19 analysis. *PLoS ONE*, 17(4):e0267558.

Nishiura, H. and Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. *Mathematical and statistical estimation approaches in epidemiology*, pages 103–121.

Noh, J. and Danuser, G. (2021). Estimation of the fraction of COVID-19 infected people in US states and countries worldwide. *PloS one*, 16(2):e0246772. https://doi.org/10.1371/journal.pone.0246772.

Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N. F., Cattarino, L., Cooper, L. V., et al. (2021). Reduction in mobility and COVID-19 transmission. *Nature communications*, 12(1):1090. https://doi.org/10.1038/s41467-021-21358-2.

Nyberg, T., Ferguson, N. M., Nash, S. G., Webster, H. H., Flaxman, S., Andrews, N., Hinsley, W., Bernal, J. L., Kall, M., Bhatt, S., et al. (2022). Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B. 1.1. 529) and delta (B. 1.617. 2) variants in England: a cohort study. *The Lancet*, 399(10332):1303–1312.

Oliveira, J. F., Jorge, D. C., Veiga, R. V., Rodrigues, M. S., Torquato, M. F., da Silva, N. B., Fiaccone, R. L., Cardim, L. L., Pereira, F. A., de Castro, C. P., et al. (2021). Mathematical modeling of COVID-19 in 14.8 million individuals in Bahia, Brazil. *Nature communications*, 12(1):333.

Pereira, I. G., Guerin, J. M., Silva Júnior, A. G., Garcia, G. S., Piscitelli, P., Miani, A., Distante, C., and Gonçalves, L. M. G. (2020). Forecasting Covid-19 Dynamics in Brazil: A Data Driven Approach. *International Journal of Environmental Research and Public Health*, 17(14):5115. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

Pérez-Gómez, B., Pastor-Barriuso, R., Fernández-de Larrea, N., Hernán, M. A., Pérez-Olmeda, M., Oteo-Iglesias, J., Fernández-Navarro, P., Fernández-García, A., Martín, M., Cruz, I., et al. (2023). SARS-CoV-2 Infection During the First and Second Pandemic Waves in Spain: the ENE–COVID Study. *American Journal of Public Health*, 113(5):533–544.

Pérez-Ortega, J., Almanza-Ortega, N. N., Torres-Poveda, K., Martínez-González, G., Zavala-Díaz, J. C., and Pazos-Rangel, R. (2022). Application of data science for cluster analysis of COVID-19 mortality according to sociodemographic factors at municipal level in Mexico. *Mathematics*, 10(13):2167.

Petherick, A., Kira, B., Barberia, L., Furst, R., Goldszmidt, R., Luciano, M., and Majumdar, S. (2020). Brazil's fight against COVID-19: risk, policies, and behaviours. *BSG Working Paper Series*, 36:1–49.

Picon, R. V., Carreno, I., da Silva, A. A., Mossmann, M., Laste, G., de Campos Domingues, G., Heringer, L. F. F., Gheno, B. R., Alvarenga, L. L., and Conte, M. (2020). Coronavirus disease 2019 population-based prevalence, risk factors, hospitalization, and fatality rates in southern brazil. *International Journal of Infectious Diseases*, 100:402–410. https://doi.org/10.1016/j.ijid.2020.09.028.

Pinto Neto, O., Kennedy, D. M., Reis, J. C., Wang, Y., Brizzi, A. C. B., Zambrano, G. J., de Souza, J. M., Pedroso, W., de Mello Pedreiro, R. C., de Matos Brizzi, B., Abinader, E. O., and Zângaro, R. A. (2021). Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil. *Nature Communications*, 12(1):418. Number: 1 Publisher: Nature Publishing Group.

PNUD, IPEA, and FJP (2013). Índice de desenvolvimento humano municipal brasileiro. http://www.atlasbrasil.org.br/.

PNUD, IPEA, and FJP (2017). Registros administrativos (2012 até 2017). Data retrieved from Registros Administrativos (2012 até 2017) database, http://atlasbrasil.org.br/acervo/biblioteca.

Public Health England (2020a). National COVID-19 surveillance report: 11 September 2020 (week 37). In: Gov.UK [Internet]. Available from: https://www.gov.uk/government/publications/national-covid-19-surveillance-reports. Accessed: Dec 6, 2023.

Public Health England (2020b). National COVID-19 surveillance report: 2 October 2020 (week 40). In: Gov.UK [Internet]. Available from: https://www.gov.uk/government/publications/national-covid-19-surveillance-reports. Accessed: Dec 6, 2023.

Public Health England (2020c). National COVID-19 surveillance report: 9 July 2020 (week 28). In: Gov.UK [Internet]. Available from: https://www.gov.uk/government/publications/national-covid-19-surveillance-reports. Accessed: Dec 6, 2023.

Pulliam, J. R., van Schalkwyk, C., Govender, N., von Gottberg, A., Cohen, C., Groome, M. J., Dushoff, J., Mlisana, K., and Moultrie, H. (2022). Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science*, 376(6593):eabn4947. https://doi.org/10.1126/science.abn4947.

Putra, S., Mu'tamar, K., and Zulkarnain (2019). Estimation of Parameters in the SIR Epidemic Model Using Particle Swarm Optimization. *American Journal of Mathematical and Computer Modelling*, 4(4):83. Number: 4 Publisher: Science Publishing Group.

Rahimi, I., Chen, F., and Gandomi, A. H. (2021). A review on COVID-19 forecasting models. *Neural Computing and Applications*, pages 1–11.

Rahman, M. S. and Chowdhury, A. H. (2022). A data-driven eXtreme gradient boosting machine learning model to predict COVID-19 transmission with meteorological drivers. *PLOS ONE*, 17(9):e0273319. Publisher: Public Library of Science.

Ramazi, P., Haratian, A., Meghdadi, M., Mari Oriyad, A., Lewis, M. A., Maleki, Z., Vega, R., Wang, H., Wishart, D. S., and Greiner, R. (2021). Accurate long-range forecasting of COVID-19 mortality in the USA. *Scientific Reports*, 11(1):13822. Publisher: Nature Publishing Group.

Ramos, A. M., Ferrández, M. R., Vela-Pérez, M., Kubik, A. B., and Ivorra, B. (2021). A simple but complex enough θ-SIR type model to be used with COVID-19 real data. Application to the case of Italy. *Physica D: Nonlinear Phenomena*, 421:132839.

Rana, R., Kant, R., Huirem, R. S., Bohra, D., and Ganguly, N. K. (2022). Omicron variant: Current insights and future directions. *Microbiological Research*, 265:127204.

Raymundo, C. E., Oliveira, M. C., Eleuterio, T. d. A., André, S. R., da Silva, M. G., Queiroz, E. R. d. S., and Medronho, R. d. A. (2021). Spatial analysis of COVID-19 incidence and the sociodemographic context in Brazil. *Plos one*, 16(3):e0247794.

Reis, R. F., de Melo Quintela, B., de Oliveira Campos, J., Gomes, J. M., Rocha, B. M., Lobosco, M., and Dos Santos, R. W. (2020). Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil. *Chaos, Solitons & Fractals*, 136:109888. https://doi.org/10.1016/j.chaos.2020.109888.

Reis, R. F., Oliveira, R. S., Quintela, B. d. M., Campos, J. d. O., Gomes, J. M., Rocha, B. M., Lobosco, M., and Dos Santos, R. W. (2021). The quixotic task of forecasting peaks of covid-19: Rather focus on forward and backward projections. *Frontiers in Public Health*, 9:623521. Publisher: Frontiers Media SA.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., and dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fractals*, 135:109853.

Ribeiro Xavier, C., Sachetto Oliveira, R., da Fonseca Vieira, V., Lobosco, M., and Weber dos Santos, R. (2022). Characterisation of omicron variant during COVID-19 pandemic and the impact of vaccination, transmission rate, mortality, and reinfection in South Africa, Germany, and Brazil. *BioTech*, 11(2):12.

Ricard, J. and Medeiros, J. (2020). Using misinformation as a political weapon: Covid-19 and bolsonaro in brazil. *Harvard Kennedy School Misinformation Review*, 1(3). https://doi.org/10.37016/mr-2020-013.

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., Barnaby, D. P., Becker, L. B., Chelico, J. D., Cohen, S. L., et al. (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *Jama*, 323(20):2052–2059.

Ripperger, T. J., Uhrlaub, J. L., Watanabe, M., Wong, R., Castaneda, Y., Pizzato, H. A., Thompson, M. R., Bradshaw, C., Weinkauf, C. C., Bime, C., et al. (2020). Orthogonal sars-cov-2 serological assays enable surveillance of low-prevalence communities and reveal durable humoral immunity. *Immunity*, 53(5):925–933. https://doi.org/10.1016/j.immuni.2020.10.004.

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., and Roser, M. (2020). Coronavirus Pandemic (COVID-19). https://ourworldindata.org/coronavirus. [accessed 14 July 2023].

Ritto, T. G., Cunha Jr, A., and Barton, D. A. (2021). Parameter calibration and uncertainty quantification in an SEIR-type COVID-19 model using approximate Bayesian computation. In *3rd Pan American congress on computational mechanics (PANACM 2021)*.

Rizvi, S. A., Umair, M., and Cheema, M. A. (2021). Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons & Fractals*, 151:111240.

Rocha, R., Atun, R., Massuda, A., Rache, B., Spinola, P., Nunes, L., Lago, M., and Castro, M. C. (2021). Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. *The Lancet Global Health*.

Rod, J., Oviedo-Trespalacios, O., and Cortes-Ramirez, J. (2020). A brief-review of the risk factors for covid-19 severity. *Revista de saude publica*, 54.

Romanescu, R., Hu, S., Nanton, D., Torabi, M., Tremblay-Savard, O., and Haque, M. A. (2023). The effective reproductive number: modeling and prediction with application to the multi-wave Covid-19 pandemic. *Epidemics*, page 100708. https://doi.org/10.1016/j.epidem.2023.100708.

Ross, T. J. (2005). *Fuzzy logic with engineering applications*. John Wiley & Sons.

Salata, A. (2020). Race, Class and Income Inequality in Brazil: A Social Trajectory Analysis. *Dados*, 63(3):e20190063.

Salles, S. (2023). Falta de comunicação de resultados de autotestes gera subnotificação, apontam especialistas. Available from: https://www.cnnbrasil.com.br/saude/falta-de-comunicacao-de-resultados-de-autotestes-gera-subnotificacao-apontam-especialistas. Accessed: Jul 11, 2023.

Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Samrin, N. A., Suzan, M. M. H., Hossain, M. S., Mollah, M. S. H., and Haque, M. D. (2022). Analysis of COVID-19 Trends in Bangladesh: A Machine Learning Analysis. In Boulouard, Z., Ouaissa, M., Ouaissa, M., and El Himer, S., editors, *AI and IoT for Sustainable Development in Emerging Countries: Challenges and Opportunities*, Lecture Notes on Data Engineering and Communications Technologies, pages 611–625. Springer International Publishing, Cham.

Santos, V. S., Souza Araújo, A. A., de Oliveira, J. R., Quintans-Júnior, L. J., and Martins-Filho, P. R. (2020). COVID-19 mortality among Indigenous people in Brazil: a nationwide register-based study. *Journal of Public Health*, 43(2):e250–e251.

Sarkar, K., Khajanchi, S., and Nieto, J. J. (2020). Modeling and forecasting the COVID-19 pandemic in India. *Chaos, Solitons & Fractals*, 139:110049.

Shah, N. H., Sheoran, N., Jayswal, E., Shukla, D., Shukla, N., Shukla, J., and Shah, Y. (2022). Modelling covid-19 transmission in the united states through interstate and foreign travels and evaluating impact of governmental public health interventions. *Journal of Mathematical Analysis and Applications*, 514(2):124896.

Shankar, S., Mohakuda, S. S., Kumar, A., Nazneen, P., Yadav, A. K., Chatterjee, K., and Chatterjee, K. (2021). Systematic review of predictive mathematical models of COVID-19 epidemic. *Medical journal armed forces India*, 77:S385–S392.

Sharma, M. K., Dhiman, N., Mishra, V. N., et al. (2021). Mediative fuzzy logic mathematical model: A contradictory management prediction in COVID-19 pandemic. *Applied Soft Computing*, 105:107285.

Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., and Hassanien, A. E. (2020). Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art. *SN Computer Science*, 1(4):197.

Shrivastav, L. K. and Jha, S. K. (2021). A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. *Applied Intelligence*, 51(5):2727–2739.

Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). *Time series analysis and its applications*, volume 3. Springer.

Silva, P. C., Batista, P. V., Lima, H. S., Alves, M. A., Guimarães, F. G., and Silva, R. C. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals*, 139:110088.

Silveira, M. F., Barros, A. J., Horta, B. L., Pellanda, L. C., Victora, G. D., Dellagostin, O. A., Struchiner, C. J., Burattini, M. N., Valim, A. R., Berlezi, E. M., et al. (2020). Population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil. *Nature Medicine*, 26(8):1196–1199. https://doi.org/10.1038/s41591-020-0992-3.

Smith, T. J. and McKenna, C. M. (2013). A comparison of logistic regression pseudo r2 indices. *Multiple Linear Regression Viewpoints*, 39(2):17–26.

Storn, R. and Price, K. (1997). Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341. https://doi.org/10.1023/A:1008202821328.

Suchoski, B., Stage, S., Gurung, H., and Baccam, P. (2022). GPU Accelerated Parallel Processing for Large-Scale Monte Carlo Analysis: COVID-19 Parameter Estimation and New Case Forecasting. *Frontiers in Applied Mathematics and Statistics*, 8. Publisher: Frontiers.

Szklo, M. and Nieto, F. J. (2014). *Epidemiology: beyond the basics*. Jones & Bartlett Publishers, Burlington, MA.

Tang, Y., Serdan, T. D. A., Masi, L. N., Tang, S., Gorjao, R., and Hirabara, S. M. (2020). Epidemiology of COVID-19 in Brazil: using a mathematical model to estimate the outbreak peak and temporal evolution. *Emerging Microbes & Infections*, 9(1):1453–1456.

Thakur, V., Bhola, S., Thakur, P., Patel, S. K. S., Kulshrestha, S., Ratho, R. K., and Kumar, P. (2021). Waves and variants of SARS-CoV-2: understanding the causes and effect of the COVID-19 catastrophe. *Infection*, pages 1–16.

The Open University (2016). *Epidemiology: An introduction*. The Open University.

Tian, D., Sun, Y., Zhou, J., and Ye, Q. (2021). The global epidemic of the SARS-CoV-2 delta variant, key spike mutations and immune escape. *Frontiers in immunology*, page 5001.

TSE (2022). Divulgação dos resultados das eleições 2022. Database: TSE [Internet]. Available from: https://www.tse.jus.br/eleicoes/eleicoes-2022/divulgacao-dos-resultados-das-eleicoes-2022. Accessed: Nov 4, 2024.

Unruh, L. H., Dharmapuri, S., Xia, Y., and Soyemi, K. (2022). Health disparities and COVID-19: A retrospective study examining individual and community factors causing disproportionate COVID-19 outcomes in Cook County, Illinois. *PLOS ONE*, 17(5):e0268317. Publisher: Public Library of Science.

Vahabi, N., Salehi, M., Duarte, J. D., Mollalo, A., and Michailidis, G. (2021). County-level longitudinal clustering of COVID-19 mortality to incidence ratio in the United States. *Scientific reports*, 11(1):1–22.

Vasconcelos, G. L., Brum, A. A., Almeida, F. A. G., Macêdo, A. M. S., Duarte-Filho, G. C., and Ospina, R. (2021). Standard and Anomalous Waves of COVID-19: A Multiple-Wave Growth Model for Epidemics. *Brazilian Journal of Physics*, 51(6):1867–1883.

Vasconcelos, G. L., Pessoa, N. L., Silva, N. B., Macêdo, A. M. S., Brum, A. A., Ospina, R., and Tirnakli, U. (2023). Multiple waves of COVID-19: a pathway model approach. *Nonlinear Dynamics*, 111(7):6855–6872.

Velavan, T. P. and Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical medicine & international health*, 25(3):278.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G., Fu, H., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 20(6):669–677. https://doi.org/10.1016/S1473-3099(20)30243-7.

Vinceti, M., Filippini, T., Rothman, K. J., Di Federico, S., and Orsini, N. (2021). SARS-CoV-2 infection incidence during the first and second COVID-19 waves in Italy. *Environmental research*, 197:111097.

Voinsky, I., Baristaite, G., and Gurwitz, D. (2020). Effects of age and sex on recovery from COVID-19: Analysis of 5769 Israeli patients. *Journal of Infection*, 81(2):e102–e103. https://doi.org/10.1016/j.jinf.2020.05.026.

Volpatto, D. T., Resende, A. C. M., dos Anjos, L., Silva, J., Dias, C. M., Almeida, R., and Malta, S. (2023). A generalised SEIRD model with implicit social distancing mechanism: A Bayesian approach for the identification of the spread of COVID-19 with applications in Brazil and Rio de Janeiro state. *Journal of Simulation*, 17(2):178–192. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17477778.2021.1977731.

Walker, A. G., Sibbel, S., Wade, C., Moulton, N., Marlowe, G., Young, A., Fadem, S. Z., and Brunelli, S. M. (2021). SARS-CoV-2 antibody seroprevalence among maintenance dialysis patients in the United States. *Kidney Medicine*, 3(2):216–222.

Wang, X., Wang, H., Ramazi, P., Nah, K., and Lewis, M. (2022). From Policy to Prediction: Forecasting COVID-19 Dynamics Under Imperfect Vaccination. *Bulletin of Mathematical Biology*, 84(9):90.

Ward, I. L., Bermingham, C., Ayoubkhani, D., Gethings, O. J., Pouwels, K. B., Yates, T., Khunti, K., Hippisley-Cox, J., Banerjee, A., Walker, A. S., et al. (2022). Risk of covid-19 related deaths for SARS-CoV-2 omicron (B. 1.1. 529) compared with delta (B. 1.617. 2): retrospective cohort study. *Bmj*, 378. https://doi.org/10.1136/bmj-2022-070695.

Watson, G. L., Xiong, D., Zhang, L., Zoller, J. A., Shamshoian, J., Sundin, P., Bufford, T., Rimoin, A. W., Suchard, M. A., and Ramirez, C. M. (2021). Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. *PLOS Computational Biology*, 17(3):e1008837. Publisher: Public Library of Science.

WHO (2020a). Covid-19 strategy update, 14 april 2020. Technical report, World Health Organization. https://www.who.int/publications/m/item/covid-19-strategy-update. Accessed: Jul 14, 2023.

WHO (2020b). Getting your workplace ready for COVID-19: how COVID-19 spreads, 19 March 2020. Technical report, World Health Organization.

WHO (2020c). Mask use in the context of COVID-19: interim guidance, 1 December 2020. Technical report, World Health Organization.

WHO (2020d). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.

WHO (2023). Covid-19 vaccine tracker and landscape. Technical report, World Health Organization. Accessed: Feb 27, 2024.

Wilson, S. L. and Wiysonge, C. (2020). Social media and vaccine hesitancy. *BMJ Global Health*, 5(10).

Xavier, C. R., Oliveira, R. S., Vieira, V. d. F., Rocha, B. M., Reis, R. F., Quintela, B. d. M., Lobosco, M., and Santos, R. W. d. (2022a). Timing the race of vaccination, new variants, and relaxing restrictions during COVID-19 pandemic. *Journal of Computational Science*, 61:101660.

Xavier, D. R., Silva, E. L. e., Lara, F. A., Silva, G. R. R. e., Oliveira, M. F., Gurgel, H., and Barcellos, C. (2022b). Involvement of political and socio-economic factors in the spatial and temporal dynamics of COVID-19 outcomes in Brazil: A population-based study. *The Lancet Regional Health – Americas*, 10. Publisher: Elsevier.

Xiang, Y., Jia, Y., Chen, L., Guo, L., Shu, B., and Long, E. (2021). COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models. *Infectious Disease Modelling*, 6:324–342.

Xu, J. and Tang, Y. (2021). Bayesian framework for multi-wave COVID-19 epidemic analysis using empirical vaccination data. *Mathematics*, 10(1):21. https://doi.org/10.3390/math10010021.

Yang, W. and Shaman, J. L. (2022). COVID-19 pandemic dynamics in South Africa and epidemiological characteristics of three variants of concern (Beta, Delta, and Omicron). *Elife*, 11. https://doi.org/10.7554/eLife.78933.

Yoo, D.-s., Hwang, M., Chun, B. C., Kim, S. J., Son, M., Seo, N.-K., and Ki, M. (2022). Socioeconomic Inequalities in COVID-19 Incidence During Different Epidemic Phases in South Korea. *Frontiers in Medicine*, 9. Publisher: Frontiers.

Yoshikawa, Y. and Kawachi, I. (2021). Association of Socioeconomic Characteristics With Disparities in COVID-19 Outcomes in Japan. *JAMA Network Open*, 4(7):e2117060.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*.

Zaki, M. J. and Meira Jr, W. (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press, Cambridge, UK.

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., and Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*, 31:105787.

Zimerman, R. A., Ferrareze, P. A. G., Cadegiani, F. A., Wambier, C. G., Fonseca, D. d. N., De Souza, A. R., Goren, A., Rotta, L. N., Ren, Z., and Thompson, C. E. (2022). Comparative genomics and characterization of SARS-CoV-2 P. 1 (gamma) variant of concern from Amazonas, Brazil. *Frontiers in medicine*, 9:141. https://doi.org/10.3389/fmed.2022.806611.

Ziyadidegan, S., Razavi, M., Pesarakli, H., Javid, A. H., and Erraguntla, M. (2022). Factors affecting the COVID-19 risk in the US counties: an innovative approach by combining unsupervised and supervised learning. *Stochastic Environmental Research and Risk Assessment*, pages 1–16.

APPENDIX A - Descriptive statistics of Brazilian municipalities

Table 14 – Descriptive statistics of variables for 5,560 Brazilian municipalities, categorized by COVID-19 mortality, sociodemographic indices, vaccination coverage, and political attributes.

	Variable	Mean (SD)	Median (Q1, Q3)	Source
Deaths	COVID-19 deaths (2020/1)	13.98 (179.25)	1.00 (0.00, 3.00)	(DATASUS, 2022b)
	COVID-19 deaths (2020)	38.21 (336.20)	6.00 (2.00, 17.00)	(DATASUS, 2022b)
	COVID-19 deaths (2021)	76.25 (476.90)	17.00 (7.00, 41.00)	(DATASUS, 2022b)
	COVID-19 deaths (2022)	11.75 (69.33)	3.00 (1.00, 7.00)	(DATASUS, 2022b)
	COVID-19 deaths (2020-2022)	126.21 (871.14)	27.00 (12.00, 66.00)	(DATASUS, 2022b)
Sociodemographic	% population 0-19 years	27.55 (4.85)	26.88 (24.27, 30.12)	(IBGE, 2022)
	% population 20-39 years	28.76 (2.96)	28.87 (26.96, 30.57)	(IBGE, 2022)
	% population 40-59 years	26.28 (2.66)	26.72 (24.88, 28.07)	(IBGE, 2022)
	% population 60+ years	17.43 (4.59)	17.26 (14.33, 20.27)	(IBGE, 2022)
	Life expectancy (years)	73.09 (2.68)	73.47 (71.15, 75.16)	(PNUD et al., 2013)
	% urban population	63.81 (22.04)	64.64 (47.07, 82.16)	(IBGE, 2022)
	Population density (inhabitants/km^2)	116.04 (596.27)	24.27 (11.34, 53.50)	(IBGE, 2022)
	Average household size	2.84 (0.32)	2.78 (2.65, 2.95)	(IBGE, 2022)
	% crowded households	25.13 (12.99)	23.07 (15.41, 32.58)	(PNUD et al., 2013)
	% male population	50.00 (1.58)	49.89 (49.07, 50.77)	(IBGE, 2022)
	% Indigenous population	1.20 (6.14)	0.07 (0.03, 0.19)	(IBGE, 2022)
	% black and brown population	56.01 (22.66)	60.79 (37.46, 75.30)	(IBGE, 2022)
	Per capita income (BRL)	493.34 (242.96)	467.38 (281.03, 650.37)	(PNUD et al., 2013)
	Gini coefficient	0.49 (0.07)	0.49 (0.45, 0.54)	(PNUD et al., 2013)
	Social transfer per capita (BRL)	137.84 (110.38)	99.48 (47.28, 214.78)	(PNUD et al., 2017)
	% informal settlement households	1.05 (4.51)	0.00 (0.00, 0.00)	(IBGE, 2020)
	% population in informal settlements	0.52 (3.28)	0.00 (0.00, 0.00)	(IBGE, 2010)
	Population density in informal settlement (inhabitants/ha)	4.65 (26.60)	0.00 (0.00, 0.00)	(IBGE, 2010)
	% households without bathroom	3.87 (7.97)	0.43 (0.06, 3.80)	(IBGE, 2022)
	% sanitation-related hospitalizations	3.16 (4.65)	1.47 (0.59, 3.71)	(PNUD et al., 2017)
	Activity rate	55.37 (9.24)	55.63 (49.24, 61.22)	(PNUD et al., 2013)
	% self-employed workers	24.69 (9.85)	22.66 (18.08, 28.69)	(PNUD et al., 2013)
	Unemployment rate	6.74 (3.83)	6.27 (4.16, 8.62)	(PNUD et al., 2013)
	% informal workers	56.51 (19.27)	57.17 (40.26, 73.68)	(PNUD et al., 2013)
	% poor population spending 1+ hour to work	1.39 (1.56)	0.90 (0.33, 1.91)	(PNUD et al., 2013)
	% agriculture workers	35.57 (18.25)	36.47 (21.84, 49.30)	(PNUD et al., 2013)
	% commerce workers	10.57 (4.41)	10.04 (7.25, 13.45)	(PNUD et al., 2013)
	% service workers	32.46 (8.89)	31.89 (26.13, 38.00)	(PNUD et al., 2013)
	% industry workers	9.61 (8.92)	6.53 (3.32, 13.29)	(PNUD et al., 2013)
	Illiteracy rate	11.81 (7.56)	9.38 (5.57, 17.97)	(IBGE, 2022)
	Expected years of schooling at age 18	9.46 (1.10)	9.47 (8.75, 10.21)	(PNUD et al., 2013)
Vac.	% people fully vaccinated (2021)	73.32 (14.26)	74.43 (65.28, 82.66)	(Brasil, 2022)
	% people fully vaccinated (2021-2022)	85.05 (15.02)	86.08 (76.73, 94.80)	(Brasil, 2022)
Pol.	% votes for Bolsonaro (2022 first round)	37.84 (16.66)	39.34 (22.74, 51.21)	(TSE, 2022)

SD: Standard deviation.
Q1: First quartile.
Q3: Third quartile.
Vac.: Vaccination data.
Pol.: Political preference data.

APPENDIX B - Sociodemographic clustering details

We determined the optimal number of clusters using the validation metrics: Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index, as presented in Figure 43. Based on these metrics, we selected the model with five sociodemographic clusters for the statistical analysis. Figure 44 illustrates the dispersion of municipalities in the first two principal components. We labeled the clusters as Urbanized, Urbanized with Informal Settlements, Semi-urbanized, Rural with High Human Development, and Rural with Low Human Development, based on the descriptive statistics presented in Table 15.

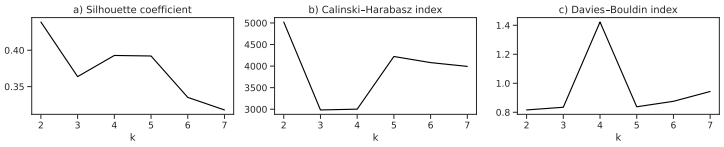


Figure 43 – Validation metrics for the sociodemographic clustering with different numbers of clusters (k). Metrics include: (a) Silhouette Coefficient, (b) Calinski-Harabasz Index, and (c) Davies-Bouldin Index.

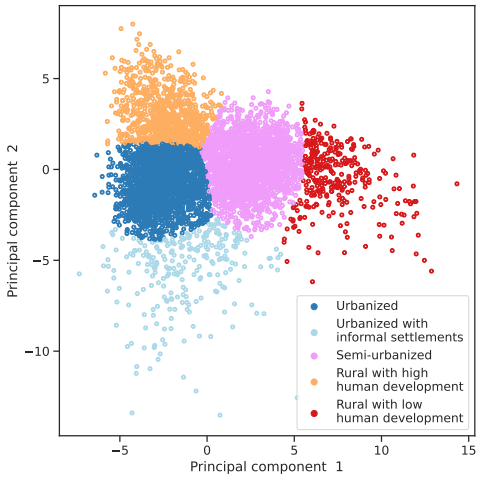


Figure 44 – Scatter plot of the sociodemographic clusters considering the principal component 1 and principal component 2. The point colors represent the sociodemographic cluster of the municipalities.

Table 15 – Descriptive statistics of variables across the different sociodemographic clusters.

Variable	Median (Quartile 1, Quartile 3)				
	Urbanized	Urbanized with informal settlements	Semi-urbanized	Rural with high human development	Rural with low human development
% population 0-19 years	25.12 (23.46, 26.87)	27.80 (25.82, 30.43)	29.63 (27.61, 31.83)	23.10 (21.13, 24.97)	38.23 (35.62, 41.96)
% population 20-39 years	28.66 (27.06, 30.35)	31.74 (30.17, 33.37)	29.31 (27.95, 30.70)	25.55 (23.86, 27.18)	30.78 (29.43, 31.95)
% population 40-59 years	27.56 (26.68, 28.48)	26.82 (25.09, 27.79)	25.15 (23.91, 26.36)	28.37 (27.38, 29.44)	20.15 (18.38, 21.85)
% population 60+ years	18.46 (16.19, 20.76)	13.29 (11.09, 15.90)	15.72 (13.74, 17.83)	22.51 (20.26, 25.38)	10.25 (8.21, 12.52)
Life expectancy (years)	75.06 (74.01, 76.00)	74.39 (73.23, 75.70)	70.98 (69.68, 72.29)	74.66 (73.57, 75.80)	70.09 (68.47, 71.41)
% urban population	82.90 (74.07, 90.82)	96.07 (89.29, 99.72)	54.68 (41.15, 66.28)	47.14 (34.06, 58.35)	43.33 (31.62, 53.58)
Population density (inhabitants/km²)	32.92 (16.82, 74.94)	463.72 (158.88, 1819.34)	20.59 (8.81, 46.51)	18.77 (11.27, 28.25)	10.62 (2.39, 20.67)
Average household size	2.70 (2.61, 2.78)	2.83 (2.71, 2.97)	2.92 (2.80, 3.05)	2.62 (2.54, 2.71)	3.55 (3.34, 4.02)
% crowded households	17.59 (13.56, 22.22)	31.84 (26.42, 38.90)	30.91 (25.67, 36.69)	11.81 (8.39, 16.03)	53.43 (45.55, 65.29)
% male population	49.52 (48.86, 50.19)	48.22 (47.57, 48.89)	50.00 (49.20, 50.85)	50.65 (50.00, 51.32)	51.18 (50.38, 51.90)
% indigenous population	0.07 (0.03, 0.13)	0.16 (0.09, 0.29)	0.09 (0.03, 0.31)	0.04 (0.00, 0.10)	0.16 (0.05, 8.04)
% black and brown population	44.19 (32.39, 57.27)	68.47 (54.58, 74.26)	74.65 (66.63, 80.73)	26.58 (15.50, 42.79)	81.43 (74.59, 86.01)
Per capita income (BRL)	646.90 (549.92, 770.12)	620.89 (466.69, 820.97)	277.95 (238.25, 334.07)	579.40 (479.55, 720.21)	195.93 (169.71, 229.36)
Gini coefficient	0.46 (0.42, 0.50)	0.51 (0.46, 0.56)	0.52 (0.49, 0.55)	0.46 (0.42, 0.49)	0.59 (0.55, 0.62)
Social transfer per capita (BRL)	49.89 (30.34, 76.40)	72.54 (43.97, 119.79)	219.56 (164.17, 279.66)	52.94 (27.94, 82.26)	307.56 (248.57, 383.70)
% informal settlement households	0.00 (0.00, 0.00)	9.54 (3.46, 18.15)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
% population in informal settlements	0.00 (0.00, 0.00)	4.81 (0.51, 12.29)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Population density in informal settlement (inhabitants/ha)	0.00 (0.00, 0.00)	49.01 (8.01, 109.21)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
% households without bathroom	0.06 (0.02, 0.17)	0.22 (0.08, 0.78)	3.69 (1.60, 7.50)	0.12 (0.03, 0.37)	25.66 (17.47, 36.38)
% sanitation-related hospitalizations	0.91 (0.40, 2.00)	0.74 (0.43, 1.47)	2.42 (1.03, 5.47)	1.25 (0.40, 2.74)	7.23 (3.01, 14.67)
Activity rate	59.30 (55.68, 62.66)	57.66 (53.78, 61.45)	49.41 (44.77, 53.94)	63.81 (57.88, 70.85)	47.17 (41.42, 51.34)
% self-employed workers	20.11 (16.98, 24.03)	19.57 (17.21, 22.22)	22.62 (18.05, 27.80)	35.04 (27.34, 44.68)	28.24 (22.06, 36.31)
Unemployment rate	5.94 (4.35, 7.57)	9.96 (8.06, 11.96)	7.56 (5.54, 10.12)	2.53 (1.38, 4.04)	6.91 (4.82, 9.67)
% informal workers	38.27 (29.97, 47.55)	37.65 (31.89, 44.28)	52.12 (45.24, 58.62)	29.57 (22.06, 39.04)	65.52 (53.72, 75.47)
% poor population spending 1+ hour to work	0.43 (0.19, 0.90)	1.45 (0.69, 3.12)	1.62 (0.94, 2.53)	0.29 (0.07, 0.73)	3.67 (2.37, 5.49)
% agriculture workers	20.95 (12.39, 29.96)	3.00 (1.13, 8.83)	43.84 (34.65, 52.21)	50.67 (43.23, 58.98)	54.08 (44.38, 61.48)
% commerce workers	12.38 (9.80, 15.02)	16.43 (14.64, 19.01)	9.24 (6.85, 11.78)	6.99 (5.40, 8.75)	7.38 (5.11, 9.67)
% service workers	36.03 (31.27, 41.58)	47.94 (42.32, 52.12)	30.41 (25.98, 35.12)	24.60 (20.21, 29.29)	25.80 (21.95, 30.43)
% industry workers	13.86 (7.72, 21.44)	10.01 (6.56, 16.07)	3.64 (2.21, 6.03)	6.37 (3.56, 10.40)	2.71 (1.66, 4.34)
Illiteracy rate	5.75 (4.12, 7.78)	4.61 (3.26, 8.24)	18.84 (14.31, 22.63)	6.87 (4.53, 9.08)	20.02 (13.98, 24.63)
Expected years of schooling at age 18	9.96 (9.34, 10.59)	9.53 (9.05, 10.05)	8.94 (8.41, 9.48)	10.21 (9.50, 10.89)	8.14 (7.37, 8.83)

APPENDIX C - Detailed results of the regression models

Table 16 – Estimated coefficients (95% Confidence Intervals (CI)) and goodness-of-fit statistics for *Model 1*.

Variable	2020 (first half)	2020	2021	2022	2020-2022
Intercept	-9.53 (-9.61, -9.44)	-7.53 (-7.57, -7.49)	-6.29 (-6.32, -6.27)	-8.07 (-8.11, -8.03)	-5.91 (-5.93, -5.89)
Urbanized with informal settlements	0.38 (0.20, 0.57)	0.10 (0.01, 0.19)	-0.16 (-0.23, -0.09)	-0.04 (-0.12, 0.03)	-0.09 (-0.15, -0.03)
Semi-urbanized	0.88 (0.74, 1.01)	0.34 (0.28, 0.41)	-0.17 (-0.21, -0.13)	-0.08 (-0.14, -0.02)	-0.03 (-0.07, 0.00)
Rural with high human development	-0.01 (-0.21, 0.19)	0.06 (-0.02, 0.14)	-0.07 (-0.12, -0.02)	-0.08 (-0.16, 0.00)	-0.02 (-0.06, 0.02)
Rural with low human development	1.28 (1.05, 1.51)	0.43 (0.32, 0.54)	-0.32 (-0.39, -0.24)	-0.30 (-0.42, -0.17)	-0.10 (-0.16, -0.03)
% population 60+ years	-0.28 (-0.34, -0.23)	0.01 (-0.01, 0.03)	0.05 (0.04, 0.07)	0.24 (0.22, 0.27)	0.06 (0.05, 0.08)
% urban population	-0.02 (-0.08, 0.05)	0.08 (0.05, 0.11)	0.12 (0.10, 0.14)	0.11 (0.08, 0.14)	0.11 (0.10, 0.13)
Population density (inhabitants/km²)	0.04 (0.01, 0.07)	0.01 (-0.00, 0.03)	-0.00 (-0.01, 0.01)	-0.01 (-0.02, -0.00)	-0.00 (-0.01, 0.01)
% male population	-0.17 (-0.22, -0.12)	-0.06 (-0.08, -0.03)	0.02 (0.00, 0.03)	0.03 (0.01, 0.05)	0.00 (-0.01, 0.01)
% Indigenous population	0.02 (-0.01, 0.05)	0.03 (0.02, 0.05)	0.03 (0.01, 0.04)	0.01 (-0.01, 0.03)	0.03 (0.02, 0.04)
Gini coefficient	-0.01 (-0.06, 0.04)	0.00 (-0.02, 0.02)	0.02 (0.00, 0.03)	0.02 (-0.00, 0.04)	0.01 (0.00, 0.03)
% informal settlement households	0.14 (0.11, 0.17)	0.05 (0.04, 0.07)	0.00 (-0.01, 0.01)	-0.01 (-0.03, -0.00)	0.02 (0.01, 0.03)
Population density in informal settlement (inhabitants/ha)	0.02 (-0.02, 0.05)	0.01 (-0.01, 0.03)	0.01 (0.00, 0.02)	0.00 (-0.01, 0.02)	0.01 (-0.00, 0.02)
% sanitation-related hospitalizations	-0.04 (-0.08, -0.01)	-0.02 (-0.04, -0.00)	-0.00 (-0.01, 0.01)	-0.02 (-0.05, -0.00)	-0.01 (-0.02, -0.00)
% self-employed workers	0.23 (0.17, 0.28)	0.07 (0.04, 0.09)	-0.01 (-0.03, 0.00)	-0.07 (-0.10, -0.04)	0.00 (-0.01, 0.02)
Unemployment rate	0.19 (0.15, 0.23)	0.05 (0.03, 0.07)	-0.02 (-0.04, -0.01)	-0.01 (-0.03, 0.01)	0.00 (-0.01, 0.02)
% commerce workers	0.06 (0.01, 0.11)	0.05 (0.02, 0.07)	0.02 (-0.00, 0.03)	0.02 (-0.01, 0.04)	0.03 (0.01, 0.04)
% service workers	0.17 (0.11, 0.23)	0.10 (0.07, 0.13)	0.03 (0.01, 0.04)	-0.02 (-0.05, -0.01)	0.03 (0.02, 0.05)
% industry workers	0.15 (0.10, 0.21)	0.08 (0.06, 0.11)	-0.01 (-0.02, 0.01)	-0.03 (-0.05, -0.01)	0.01 (0.00, 0.03)
Expected years of schooling at age 18	0.08 (0.03, 0.13)	-0.01 (-0.03, 0.01)	0.05 (0.03, 0.06)	0.01 (-0.01, 0.03)	0.03 (0.01, 0.04)
% people fully vaccinated	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.08 (0.06, 0.09)	0.06 (0.04, 0.09)	0.06 (0.05, 0.08)
% votes for Bolsonaro	0.02 (-0.04, 0.07)	0.10 (0.07, 0.12)	0.16 (0.14, 0.18)	0.14 (0.11, 0.16)	0.15 (0.13, 0.16)
Statistic					
Degrees of freedom	20	20	21	21	21
Residual degrees of freedom	5,539	5,539	5,538	5,538	5,538
Deviance	5,323.38	6,339.26	6,148.99	6,167.09	5,751.35
Pearson's chi-squared statistic (χ^2)	6,833.98	5,973.34	6,049.98	5,586.53	5,657.04
Cox & Snell pseudo-R^2 (R^2_{CS})	0.32	0.21	0.61	0.37	0.56
McFadden pseudo-R^2 (R^2_{McF})	0.09	0.04	0.12	0.09	0.10
Log-likelihood	-10,213	-16,423	-19,829	-12,378	-21,455
Akaike Information Criterion (AIC)	20,468	32,888	39,702	24,799	42,953
Bayesian Information Criterion (BIC)	20,607	33,027	39,848	24,945	43,099

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
Note: The variables were standardized.

Table 17 – Estimated coefficients (95% Confidence Intervals (CI)) and goodness-of-fit statistics for *Model 2*.

Variable	2020 (first half)	2020	2021	2022	2020-2022
Intercept	-12.56 (-12.64, -12.48)	-12.68 (-12.72, -12.65)	-12.12 (-12.15, -12.10)	-13.79 (-13.82, -13.75)	-12.71 (-12.73, -12.69)
Urbanized with informal settlements	0.29 (0.15, 0.43)	0.14 (0.06, 0.22)	-0.14 (-0.20, -0.08)	0.02 (-0.05, 0.08)	-0.09 (-0.14, -0.03)
Semi-urbanized	0.63 (0.51, 0.75)	0.27 (0.21, 0.33)	-0.15 (-0.19, -0.11)	-0.02 (-0.08, 0.04)	-0.05 (-0.08, -0.01)
Rural with high human development	0.51 (0.30, 0.72)	0.26 (0.18, 0.34)	-0.05 (-0.09, 0.00)	0.02 (-0.06, 0.09)	0.01 (-0.04, 0.05)
Rural with low human development	0.61 (0.42, 0.81)	0.25 (0.14, 0.35)	-0.29 (-0.36, -0.21)	-0.13 (-0.25, -0.02)	-0.14 (-0.20, -0.07)
% population 60+ years	-0.01 (-0.06, 0.04)	0.10 (0.08, 0.13)	0.06 (0.05, 0.08)	0.28 (0.25, 0.30)	0.08 (0.06, 0.09)
% urban population	0.03 (-0.03, 0.09)	0.08 (0.05, 0.11)	0.11 (0.09, 0.13)	0.08 (0.05, 0.11)	0.11 (0.10, 0.13)
Population density (inhabitants/km^2)	0.03 (0.01, 0.05)	0.00 (-0.01, 0.02)	-0.00 (-0.01, 0.01)	-0.01 (-0.02, -0.00)	-0.00 (-0.01, 0.01)
% male population	0.05 (0.01, 0.09)	0.00 (-0.02, 0.02)	0.03 (0.02, 0.04)	0.05 (0.03, 0.07)	0.01 (0.00, 0.02)
% Indigenous population	0.04 (0.02, 0.07)	0.04 (0.03, 0.06)	0.02 (0.01, 0.03)	0.02 (-0.00, 0.04)	0.03 (0.02, 0.04)
Gini coefficient	-0.07 (-0.11, -0.03)	-0.04 (-0.06, -0.02)	0.01 (-0.01, 0.02)	-0.01 (-0.03, 0.01)	0.00 (-0.01, 0.02)
% informal settlement households	0.08 (0.05, 0.10)	0.04 (0.02, 0.05)	-0.00 (-0.01, 0.01)	-0.01 (-0.03, -0.00)	0.01 (0.00, 0.02)
Population density in informal settlement (inhabitants/ha)	-0.01 (-0.03, 0.02)	-0.00 (-0.02, 0.01)	0.01 (-0.00, 0.02)	-0.00 (-0.01, 0.01)	0.01 (-0.00, 0.02)
% sanitation-related hospitalizations	0.01 (-0.03, 0.04)	-0.02 (-0.04, -0.00)	0.00 (-0.01, 0.02)	-0.02 (-0.04, -0.00)	-0.01 (-0.02, -0.00)
% self-employed workers	0.06 (0.01, 0.11)	0.03 (0.01, 0.06)	-0.02 (-0.03, -0.00)	-0.06 (-0.09, -0.04)	-0.00 (-0.02, 0.01)
Unemployment rate	0.20 (0.16, 0.24)	0.04 (0.02, 0.06)	-0.02 (-0.04, -0.01)	-0.01 (-0.03, 0.01)	0.00 (-0.01, 0.01)
% commerce workers	-0.10 (-0.14, -0.05)	-0.04 (-0.06, -0.02)	-0.01 (-0.03, 0.00)	-0.05 (-0.07, -0.02)	0.01 (-0.01, 0.02)
% service workers	-0.04 (-0.10, 0.01)	0.06 (0.04, 0.09)	0.03 (0.02, 0.05)	-0.00 (-0.03, 0.02)	0.03 (0.01, 0.04)
% industry workers	-0.04 (-0.08, 0.01)	0.02 (0.00, 0.05)	-0.01 (-0.03, 0.00)	-0.04 (-0.06, -0.02)	0.00 (-0.01, 0.01)
Expected years of schooling at age 18	0.06 (0.02, 0.10)	-0.00 (-0.02, 0.02)	0.05 (0.04, 0.07)	0.04 (0.02, 0.06)	0.03 (0.02, 0.04)
% people fully vaccinated	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.08 (0.07, 0.10)	0.07 (0.05, 0.09)	0.07 (0.06, 0.08)
% votes for Bolsonaro	0.03 (-0.02, 0.07)	0.08 (0.05, 0.10)	0.15 (0.13, 0.16)	0.12 (0.10, 0.15)	0.14 (0.13, 0.16)
Statistic					
Degrees of freedom	20	20	21	21	21
Residual degrees of freedom	5,539	5,539	5,538	5,538	5,538
Deviance	3,468.36	5,403.91	5,732.97	4,461.53	5,685.81
Pearson's chi-squared statistic (χ^2)	5,274.50	7,263.28	5,826.01	6,715.25	5,712.67
Cox & Snell pseudo-R^2 (R^2_{CS})	0.12	0.09	0.59	0.40	0.58
McFadden pseudo-R^2 (R^2_{McF})	0.04	0.02	0.11	0.11	0.10
Log-likelihood	-8,457	-15,534	-19,443	-11,162	-21,179
Akaike Information Criterion (AIC)	16,956	31,109	38,930	22,368	42,402
Bayesian Information Criterion (BIC)	17,095	31,248	39,076	22,514	42,548

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
Note: The variables were standardized.

APPENDIX D - Sensitivity analysis for regression model

We performed a sensitivity analysis for *Model 2* to evaluate its robustness. Specifically, we generated 30 bootstrap resamples for each period to assess variability in model performance. The results, summarized in Table 18, show minimal variation in goodness-of-fit statistics across resamples. Similarly, Table 19 highlights that the estimated coefficients for the model variables remained consistent, indicating no significant changes.

Table 18 – Comparison of goodness-of-fit statistics (95% Confidence Intervals (CI)) between *Model 2* using 30 bootstrap resamples and the reference dataset.

Statistic	2020 (first half)	2020	2021	2022	2020-2022
R^2_{CS}	0.004 (0.0, 0.007)	0.005 (0.002, 0.007)	0.004 (-0.0, 0.008)	-0.004 (-0.011, 0.003)	-0.0 (-0.004, 0.004)
R^2_{McF}	0.001 (0.0, 0.002)	0.001 (0.0, 0.001)	0.001 (0.0, 0.003)	-0.001 (-0.003, 0.002)	-0.0 (-0.001, 0.001)
LL	-14.544 (-62.831, 33.742)	-17.624 (-60.656, 25.408)	55.747 (20.92, 90.575)	37.834 (6.808, 68.86)	-0.619 (-50.579, 49.341)
AIC	29.089 (-67.483, 125.661)	35.248 (-50.817, 121.313)	-111.495 (-181.15, -41.84)	-75.668 (-137.72, -13.616)	1.239 (-98.681, 101.159)
BIC	29.089 (-67.483, 125.661)	35.248 (-50.817, 121.313)	-111.495 (-181.15, -41.84)	-75.668 (-137.72, -13.616)	1.239 (-98.681, 101.159)

Note: The values represent differences between the statistic calculated from the bootstrap resamples and the corresponding statistic from the reference dataset, with 95% CI provided.
Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
R^2_{CS}: Cox & Snell pseudo-R^2.
R^2_{McF}: McFadden pseudo-R^2.
LL: Log-likelihood.
AIC: Akaike Information Criterion.
BIC: Bayesian Information Criterion.

We also assessed model robustness by conducting experiments excluding outliers and influential points. Points with Pearson residuals greater than 3 were classified as outliers, while influential points were identified by a Cook's Distance greater than $\frac{4}{observation\ quantity}$. Table 20 shows that removing outliers and influential points improved model performance for *Model 2*. However, this removal did not significantly change the coefficients.

Table 19 – Comparison of the coefficients (95% Confidence Intervals (CI)) between *Model 2* with 30 bootstrap resamples and *Model 2* using the reference dataset.

Variable	2020 (first half)	2020	2021	2022	2020-2022
Intercept	-0.008 (-0.023, 0.007)	0.003 (-0.005, 0.010)	0.003 (-0.002, 0.008)	0.003 (-0.004, 0.010)	0.001 (-0.004, 0.005)
Urbanized with informal settlements	0.006 (-0.024, 0.036)	0.005 (-0.015, 0.024)	-0.002 (-0.016, 0.012)	0.004 (-0.011, 0.019)	0.004 (-0.004, 0.013)
Semi-urbanized	0.010 (-0.014, 0.034)	-0.001 (-0.012, 0.010)	-0.006 (-0.015, 0.002)	0.002 (-0.010, 0.014)	-0.004 (-0.012, 0.004)
Rural with high human development	0.017 (-0.022, 0.057)	0.001 (-0.013, 0.015)	-0.006 (-0.015, 0.004)	-0.006 (-0.018, 0.006)	0.003 (-0.006, 0.012)
Rural with low human development	0.011 (-0.023, 0.045)	-0.008 (-0.027, 0.011)	-0.006 (-0.024, 0.013)	0.003 (-0.022, 0.029)	-0.007 (-0.018, 0.004)
% population 60+ years	0.001 (-0.009, 0.011)	-0.001 (-0.005, 0.003)	-0.000 (-0.004, 0.003)	0.000 (-0.004, 0.005)	-0.003 (-0.006, -0.000)
% urban population	0.002 (-0.011, 0.014)	-0.003 (-0.009, 0.002)	-0.000 (-0.004, 0.004)	-0.005 (-0.012, 0.001)	-0.002 (-0.005, 0.001)
Population density (inhabitants/km²)	-0.001 (-0.004, 0.003)	0.000 (-0.001, 0.002)	0.000 (-0.001, 0.001)	-0.002 (-0.005, 0.000)	-0.001 (-0.002, 0.001)
% male population	0.004 (-0.005, 0.012)	0.002 (-0.002, 0.005)	0.001 (-0.003, 0.005)	0.004 (0.001, 0.008)	-0.001 (-0.003, 0.001)
% indigenous population	0.003 (-0.003, 0.009)	0.001 (-0.002, 0.003)	-0.001 (-0.004, 0.002)	0.003 (-0.001, 0.007)	-0.001 (-0.003, 0.001)
Gini coefficient	0.005 (-0.002, 0.013)	-0.004 (-0.009, 0.000)	0.002 (-0.001, 0.004)	-0.004 (-0.008, -0.001)	0.000 (-0.002, 0.003)
% informal settlement households	0.002 (-0.001, 0.006)	0.001 (-0.001, 0.003)	-0.001 (-0.003, 0.002)	0.000 (-0.002, 0.003)	-0.001 (-0.002, 0.000)
Population density in informal settlement (inhabitants/ha)	0.000 (-0.004, 0.005)	-0.001 (-0.003, 0.001)	0.001 (-0.001, 0.003)	0.001 (-0.002, 0.005)	0.000 (-0.001, 0.002)
% sanitation-related hospitalizations	0.001 (-0.004, 0.007)	0.000 (-0.002, 0.003)	-0.001 (-0.003, 0.002)	-0.003 (-0.006, 0.000)	-0.001 (-0.003, 0.001)
% self-employed workers	-0.010 (-0.018, -0.001)	-0.000 (-0.005, 0.004)	-0.002 (-0.006, 0.002)	0.001 (-0.004, 0.006)	0.000 (-0.003, 0.003)
Unemployment rate	0.000 (-0.007, 0.008)	0.001 (-0.003, 0.006)	-0.000 (-0.003, 0.003)	0.002 (-0.001, 0.006)	0.001 (-0.001, 0.003)
% commerce workers	0.003 (-0.006, 0.011)	0.001 (-0.003, 0.006)	0.001 (-0.002, 0.005)	0.000 (-0.004, 0.004)	0.001 (-0.002, 0.004)
% service workers	-0.009 (-0.021, 0.003)	-0.001 (-0.006, 0.004)	-0.002 (-0.006, 0.002)	0.005 (0.000, 0.010)	-0.002 (-0.005, 0.002)
% industry workers	-0.004 (-0.014, 0.006)	-0.001 (-0.006, 0.004)	-0.000 (-0.003, 0.002)	0.000 (-0.004, 0.004)	-0.000 (-0.003, 0.002)
Expected years of schooling at age 18	0.012 (0.003, 0.021)	-0.000 (-0.004, 0.003)	0.000 (-0.003, 0.004)	0.000 (-0.003, 0.004)	-0.001 (-0.003, 0.002)
% people fully vaccinated	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	-0.000 (-0.003, 0.002)	0.002 (-0.001, 0.005)	0.001 (-0.002, 0.004)
% votes for Bolsonaro	0.003 (-0.005, 0.011)	0.003 (-0.002, 0.008)	0.001 (-0.003, 0.005)	0.000 (-0.005, 0.005)	0.000 (-0.003, 0.004)

Note: The values represent the differences between the coefficients estimated from *Model 2* with bootstrap resamples and the corresponding coefficients estimated from *Model 2* using the reference dataset, along with their 95% CI.
Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).

Table 20 – Estimated coefficients (95% Confidence Intervals (CI)) and goodness-of-fit statistics for *Model 2*, calculated using the dataset excluding outliers and influential points.

Variable	2020 (first half)	2020	2021	2022	2020, 2022
Intercept	-12.76 (-12.84, -12.68)	-12.77 (-12.81, -12.74)	-12.16 (-12.18, -12.13)	-13.84 (-13.87, -13.80)	-12.74 (-12.76, -12.72)
Urbanized with informal settlements	0.14 (0.01, 0.28)	0.11 (0.03, 0.18)	-0.19 (-0.25, -0.13)	-0.01 (-0.06, 0.05)	-0.12 (-0.17, -0.07)
Semi, urbanized	0.70 (0.58, 0.82)	0.30 (0.24, 0.36)	-0.17 (-0.21, -0.13)	-0.02 (-0.08, 0.03)	-0.06 (-0.09, -0.03)
Rural with high human development	0.46 (0.25, 0.66)	0.24 (0.16, 0.31)	-0.06 (-0.11, -0.02)	0.02 (-0.05, 0.10)	-0.02 (-0.06, 0.02)
Rural with low human development	0.70 (0.51, 0.90)	0.22 (0.12, 0.32)	-0.38 (-0.45, -0.31)	-0.13 (-0.25, -0.02)	-0.19 (-0.25, -0.12)
% population 60+ years	0.03 (-0.02, 0.08)	0.10 (0.08, 0.12)	0.07 (0.06, 0.09)	0.27 (0.25, 0.29)	0.08 (0.07, 0.09)
% urban population	0.03 (-0.03, 0.08)	0.08 (0.05, 0.11)	0.11 (0.10, 0.13)	0.09 (0.06, 0.12)	0.12 (0.10, 0.14)
Population density (inhabitants/km²)	0.03 (0.01, 0.06)	0.00 (-0.01, 0.02)	-0.00 (-0.01, 0.01)	-0.01 (-0.02, -0.01)	-0.00 (-0.01, 0.01)
% male population	0.07 (0.03, 0.11)	0.00 (-0.02, 0.02)	0.04 (0.03, 0.05)	0.06 (0.04, 0.08)	0.02 (0.00, 0.03)
% indigenous population	0.04 (0.01, 0.07)	0.04 (0.03, 0.06)	0.03 (0.02, 0.05)	0.02 (-0.01, 0.04)	0.03 (0.02, 0.04)
Gini coefficient	-0.06 (-0.10, -0.03)	-0.05 (-0.07, -0.03)	0.00 (-0.01, 0.02)	-0.01 (-0.03, 0.01)	-0.00 (-0.01, 0.01)
% informal settlement households	0.10 (0.08, 0.12)	0.05 (0.04, 0.06)	-0.00 (-0.01, 0.01)	-0.02 (-0.03, -0.01)	0.01 (0.00, 0.02)
Population density in informal settlement (inhabitants/ha)	0.01 (-0.02, 0.03)	-0.00 (-0.02, 0.01)	0.01 (0.00, 0.03)	-0.00 (-0.01, 0.01)	0.01 (0.00, 0.02)
% sanitation-related hospitalizations	-0.01 (-0.04, 0.02)	-0.03 (-0.05, -0.01)	0.00 (-0.01, 0.01)	-0.03 (-0.05, -0.01)	-0.01 (-0.02, -0.00)
% self-employed workers	0.08 (0.04, 0.13)	0.03 (0.01, 0.05)	-0.02 (-0.04, -0.01)	-0.07 (-0.10, -0.05)	-0.00 (-0.01, 0.01)
Unemployment rate	0.24 (0.20, 0.27)	0.05 (0.03, 0.06)	-0.02 (-0.03, -0.00)	-0.01 (-0.03, 0.01)	0.01 (-0.00, 0.02)
% commerce workers	-0.14 (-0.19, -0.10)	-0.04 (-0.06, -0.01)	-0.00 (-0.02, 0.01)	-0.05 (-0.07, -0.02)	0.01 (-0.00, 0.02)
% service workers	-0.04 (-0.09, 0.01)	0.06 (0.03, 0.08)	0.04 (0.02, 0.06)	0.00 (-0.02, 0.03)	0.03 (0.01, 0.04)
% industry workers	-0.05 (-0.09, 0.00)	0.02 (-0.00, 0.04)	-0.02 (-0.03, -0.00)	-0.05 (-0.07, -0.03)	-0.01 (-0.02, 0.00)
Expected years of schooling at age 18	0.08 (0.04, 0.12)	-0.00 (-0.02, 0.02)	0.06 (0.05, 0.07)	0.04 (0.02, 0.06)	0.03 (0.02, 0.04)
% people fully vaccinated	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.08 (0.07, 0.09)	0.09 (0.07, 0.11)	0.07 (0.06, 0.08)
% votes for Bolsonaro	0.04 (-0.00, 0.09)	0.08 (0.05, 0.10)	0.14 (0.12, 0.15)	0.12 (0.10, 0.14)	0.14 (0.13, 0.15)
Statistic	**2020 (first half)**	**2020**	**2021**	**2022**	**2020, 2022**
Degrees of freedom	20	20	21	21	21
Residual degrees of freedom	5,210	5,205	5,220	5,201	5,245
Deviance	2,788.97	4,678.02	5,368.03	3,831.77	5,341.52
Pearson's chi, squared statistic (χ^2)	3,121.01	4,787.65	5,135.19	3,941.4	5,148.05
Cox & Snell pseudo-R^2 (R^2_{CS})	0.18	0.11	0.71	0.49	0.69
McFadden pseudo-R^2 (R^2_{McF})	0.07	0.02	0.16	0.15	0.14
Log, likelihood	-6,809	-13,736	-17,518	-9,572	-19,299
Akaike Information Criterion (AIC)	13,660	27,514	35,080	19,187	38,643
Bayesian Information Criterion (BIC)	13,797	27,651	35,224	19,331	38,787

Analysis periods: (i) the first half of 2020, (ii) the year 2020, (iii) the year 2021, (iv) the year 2022, and (v) the cumulative period (2020-2022).
Note: The variables were standardized.

APPENDIX E - Time series of correlations between mortality rates and sociodemographic variables

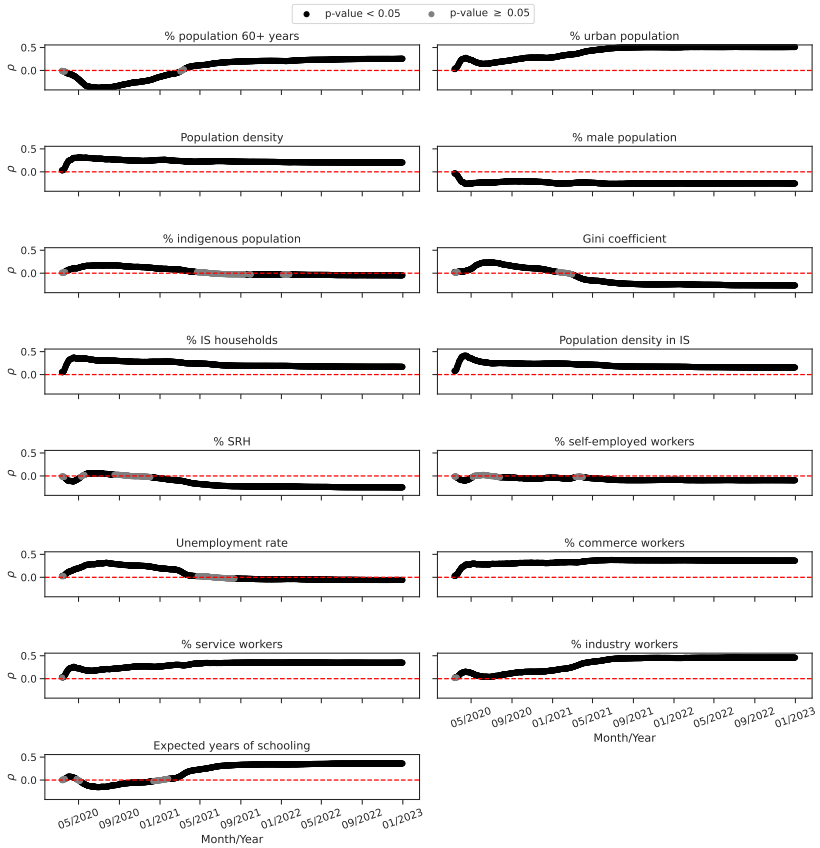


Figure 45 – Time series plot of the Spearman correlation coefficient ρ between the municipal accumulated mortality rate per 100,000 inhabitants and the sociodemographic variables. The dashed horizontal line highlights the threshold that distinguishes positive and negative correlations.

IS: Informal settlements.
SRH: sanitation-related hospitalizations.

Figure 46 – Small multiples with time series plot of the Spearman correlation coefficient ρ between the municipal accumulated mortality rate per 100,000 inhabitants and the sociodemographic variables (rows) for the sociodemographic clusters (columns): (a) Semi-urbanized, (b) Urbanized, (c) Rural with high human development, (d) Urbanized with informal settlements, and (e) Rural with low human development. The dashed horizontal line highlights the threshold that distinguishes positive and negative correlations.

IS: Informal settlements.
SRH: sanitation-related hospitalizations.

APPENDIX F - Epidemiological Time Series of COVID-19 for Spain, the United Kingdom, and the United States

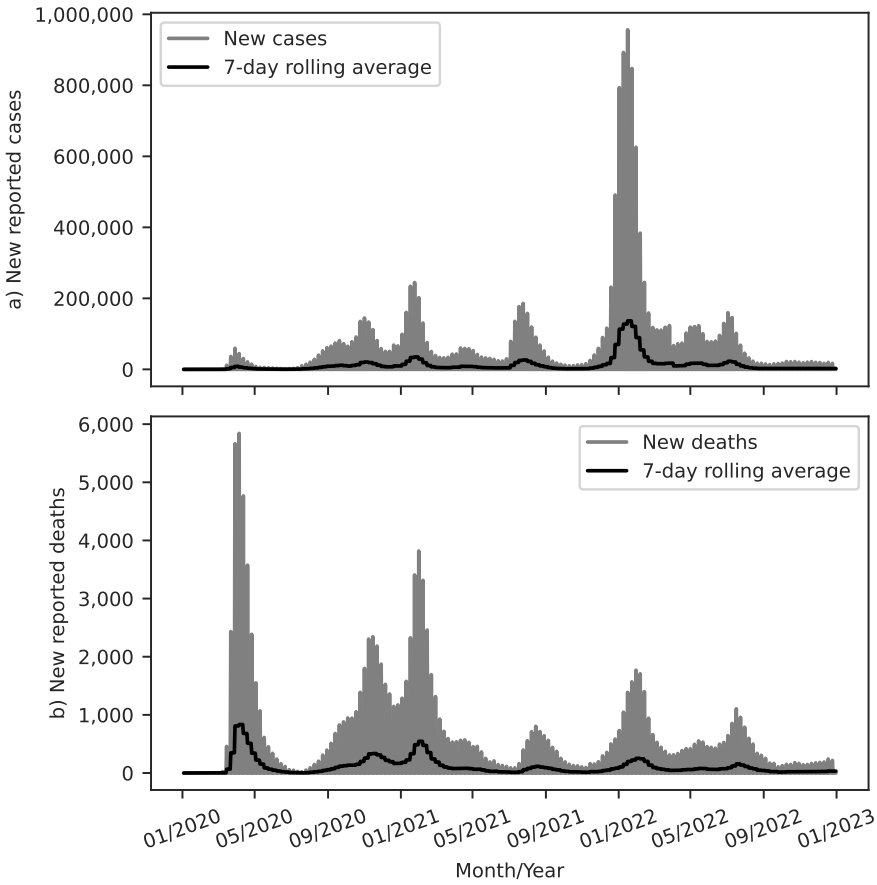


Figure 47 – Time series of COVID-19 in Spain at the reported date by the health authorities. (a) new reported cases and (b) new reported deaths.

Source: Our World in Data (Ritchie et al., 2020)

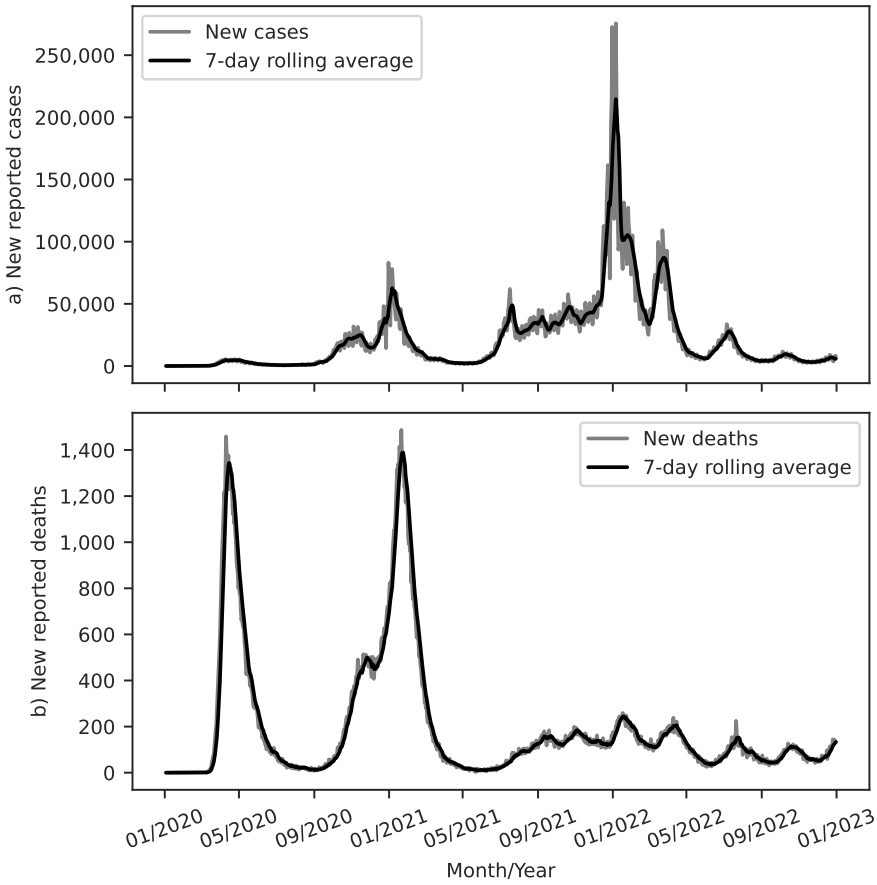


Figure 48 – Time series of COVID-19 in the United Kingdom at the reported date by the health authorities. (a) new reported cases and (b) new reported deaths.

Source: Our World in Data (Ritchie et al., 2020)

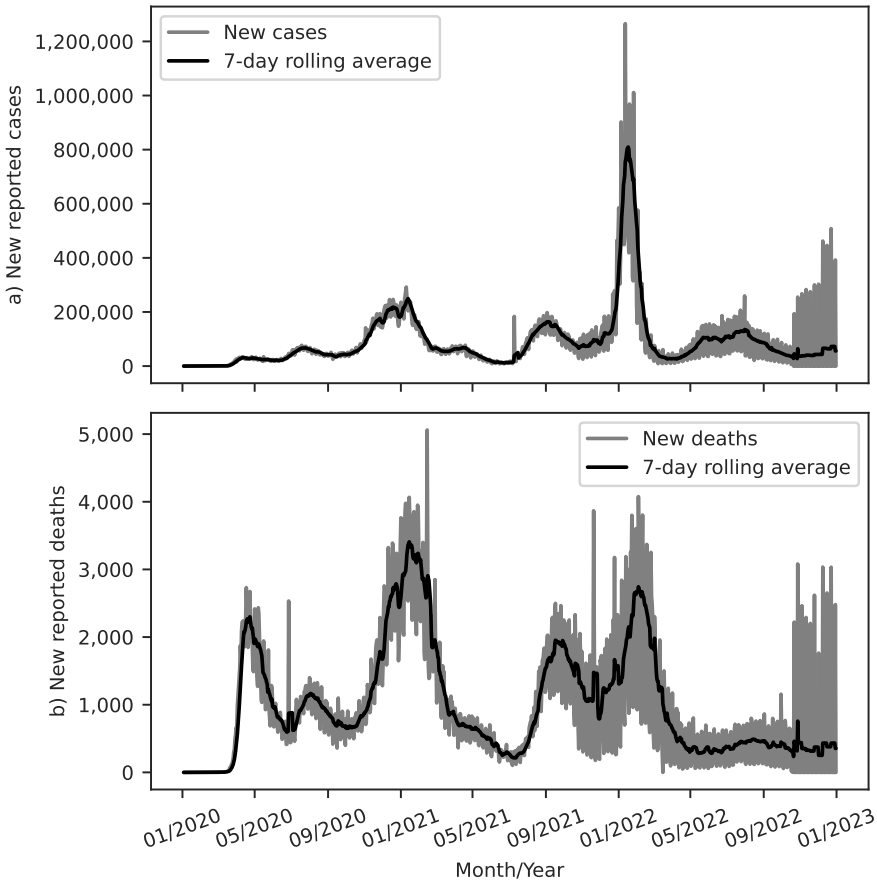


Figure 49 – Time series of COVID-19 in the United States at the reported date by the health authorities. (a) new reported cases and (b) new reported deaths.

Source: Our World in Data (Ritchie et al., 2020)

APPENDIX G - Time series of COVID-19 Case Fatality Rate (CFR) for Spain, the United Kingdom, and the United States

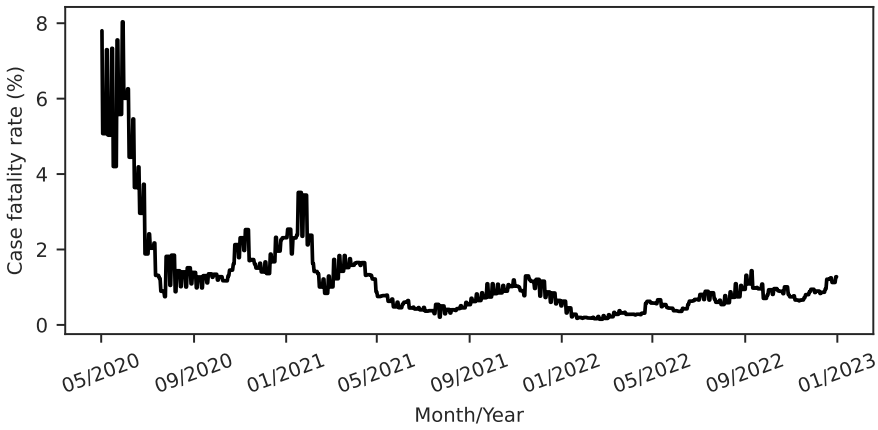


Figure 50 – Time series of COVID-19 Case Fatality Rate (CFR) in Spain.

Source: Our World in Data (Ritchie et al., 2020)

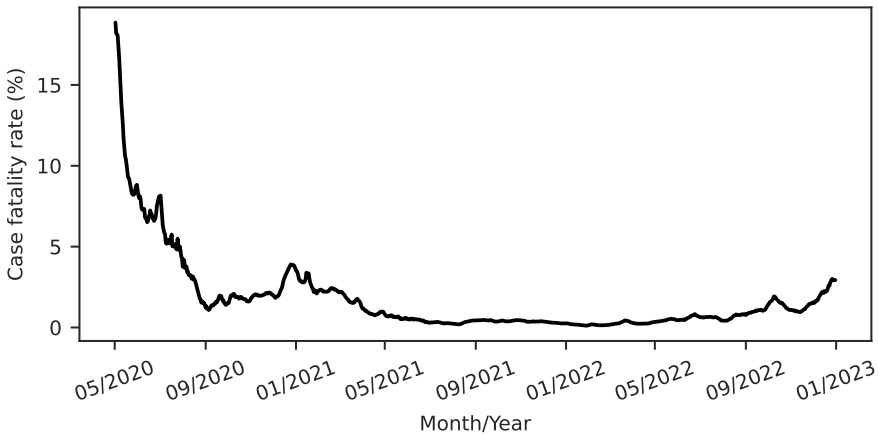


Figure 51 – Time series of COVID-19 Case Fatality Rate (CFR) in the United Kingdom.

Source: Our World in Data (Ritchie et al., 2020)

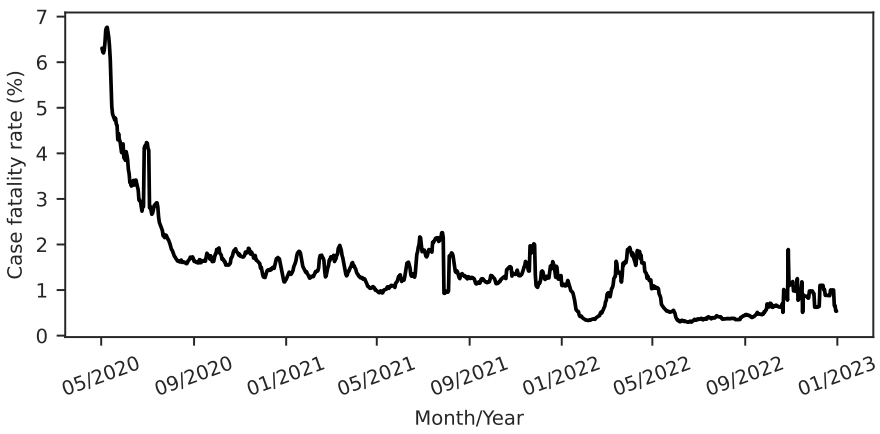


Figure 52 – Time series of COVID-19 Case Fatality Rate (CFR) in the United States. Source: Our World in Data (Ritchie et al., 2020)

APPENDIX H - Effective reproduction number for Spain, the United Kingdom, and the United States

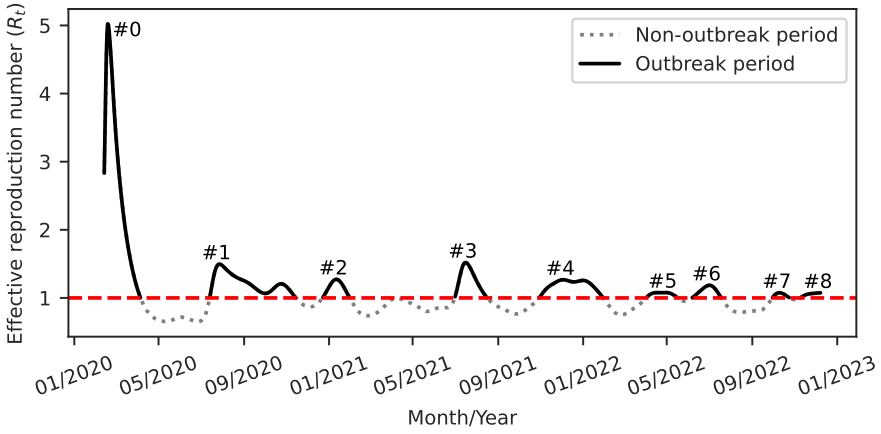


Figure 53 – Effective reproduction number (R_t) for COVID-19 in Spain. The dashed horizontal line represents the reference value ($R_t = 1$) used to monitor epidemics. The R_t time series alternates between outbreak periods (solid line) and non-outbreak periods (dotted line). Nine outbreak periods, labeled from #0 to #8, were identified.

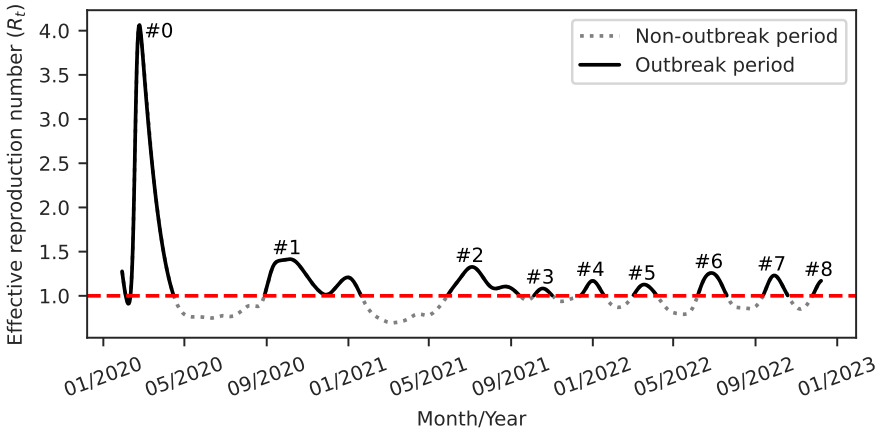


Figure 54 – Effective reproduction number (R_t) for COVID-19 in the United Kingdom. The dashed horizontal line represents the reference value ($R_t = 1$) used to monitor epidemics. The R_t time series alternates between outbreak periods (solid line) and non-outbreak periods (dotted line). Nine outbreak periods, labeled from #0 to #8, were identified.

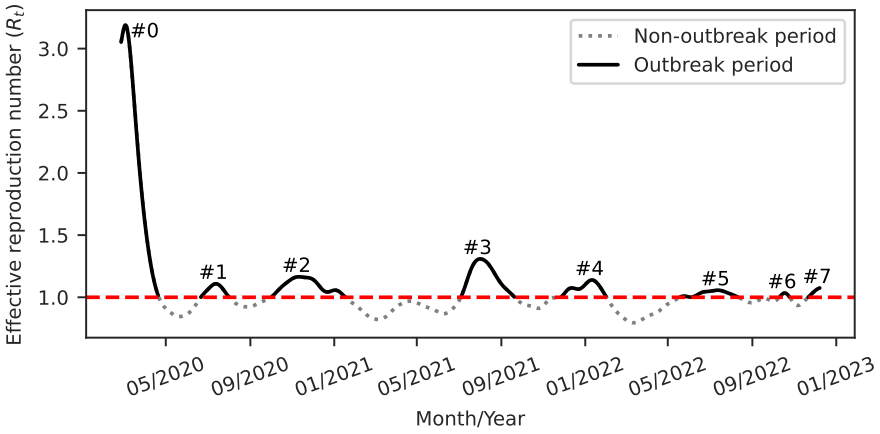


Figure 55 – Effective reproduction number (R_t) for COVID-19 in the United States. The dashed horizontal line represents the reference value ($R_t = 1$) used to monitor epidemics. The R_t time series alternates between outbreak periods (solid line) and non-outbreak periods (dotted line). Eight outbreak periods, labeled from #0 to #7, were identified.

APPENDIX I - Fuzzy variables fitted for smoothing transitions between epidemic periods

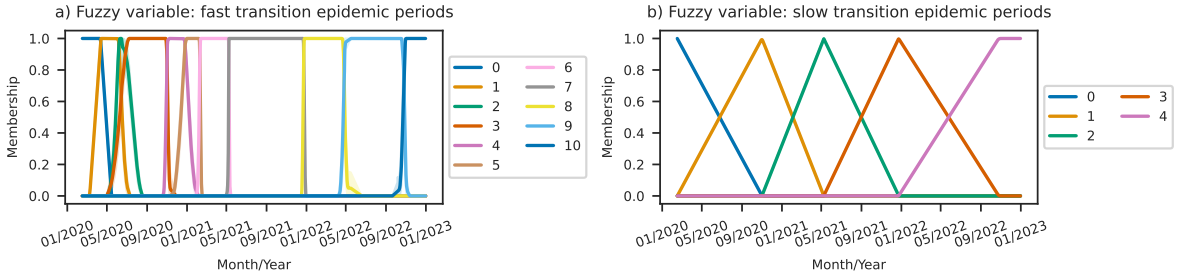


Figure 56 – Fuzzy variables fitted for smoothing transitions between epidemic periods in Brazil. (a) Fuzzy variable with partitions representing epidemic periods with fast transitions. (b) Fuzzy variable with partitions representing epidemic periods with slow transitions. The colors represent partitions, and shaded regions depict the 95% Confidence Interval (CI).

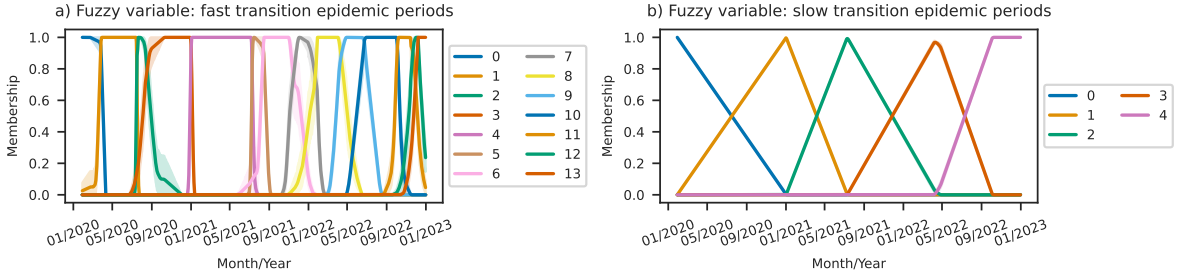


Figure 57 – Fuzzy variables fitted for smoothing transitions between epidemic periods in Spain. (a) Fuzzy variable with partitions representing epidemic periods with fast transitions. (b) Fuzzy variable with partitions representing epidemic periods with slow transitions. The colors represent partitions, and shaded regions depict the 95% Confidence Interval (CI).

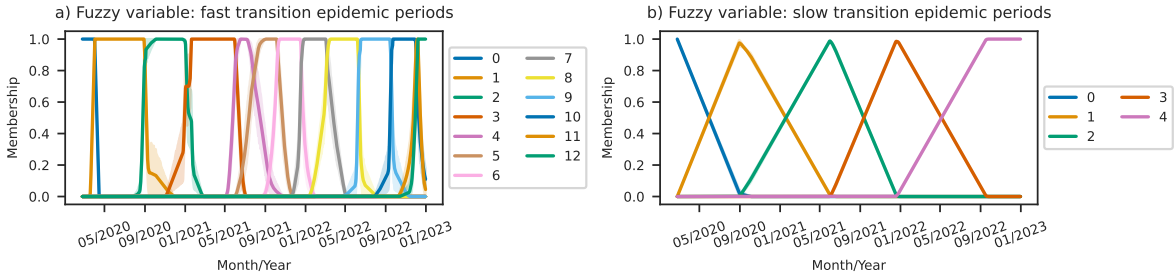


Figure 58 – Fuzzy variables fitted for smoothing transitions between epidemic periods in the United Kingdom. (a) Fuzzy variable with partitions representing epidemic periods with fast transitions. (b) Fuzzy variable with partitions representing epidemic periods with slow transitions. The colors represent partitions, and shaded regions depict the 95% Confidence Interval (CI).

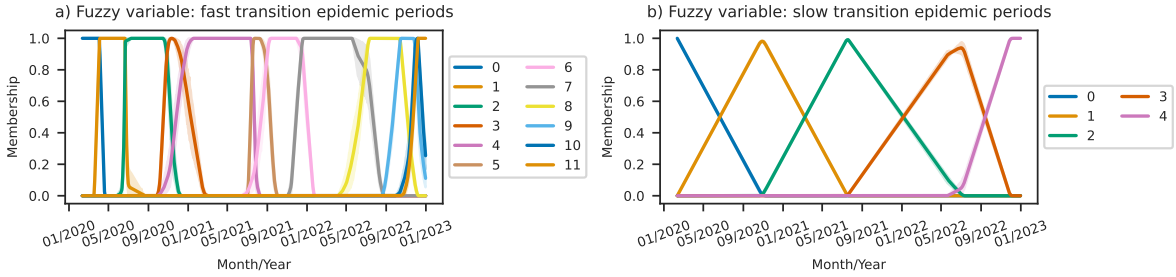


Figure 59 – Fuzzy variables fitted for smoothing transitions between epidemic periods in the United States. (a) Fuzzy variable with partitions representing epidemic periods with fast transitions. (b) Fuzzy variable with partitions representing epidemic periods with slow transitions. The colors represent partitions, and shaded regions depict the 95% Confidence Interval (CI).

APPENDIX J - Comprehensive analysis of simulation results for COVID-19 in Spain, the United Kingdom, and the United States

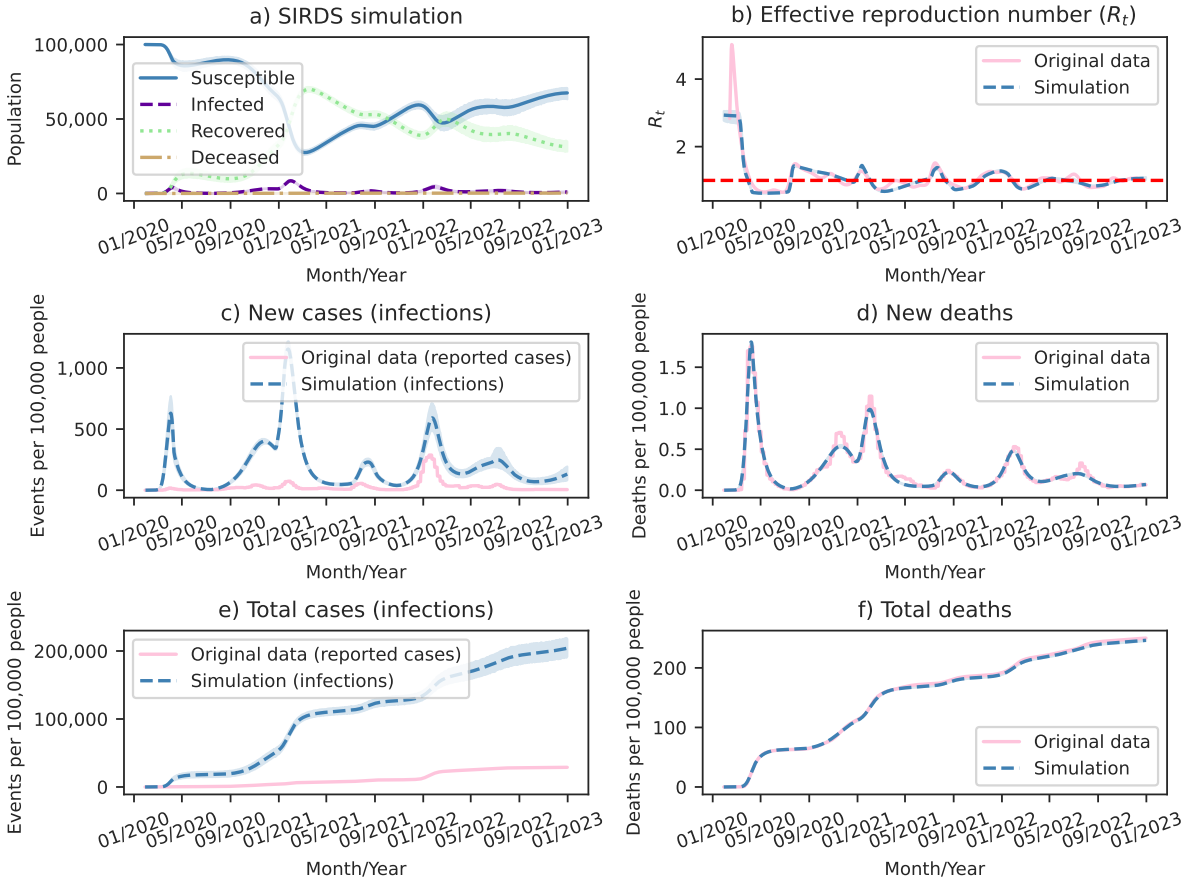


Figure 60 – Comprehensive analysis of simulation results for COVID-19 in Spain. (a) Model outcomes for an eight-day recovery period detailing the population compartments: Susceptible, Infected, Recovered, and Deceased. (b) Time series comparison between the effective reproduction number (R_t) estimated directly from reported deaths and R_t calculated by model simulations. (c) Time series comparison between new cases reported by health authorities and new infections in model simulations. (d) Time series comparison between new deaths reported by health authorities and new deaths in model simulations. (e) Time series comparison between cumulative cases reported by health authorities and cumulative infections in model simulations. (f) Time series comparison between cumulative deaths reported by health authorities and cumulative deaths in model simulations. Shaded regions depict the 95% Confidence Interval (CI).

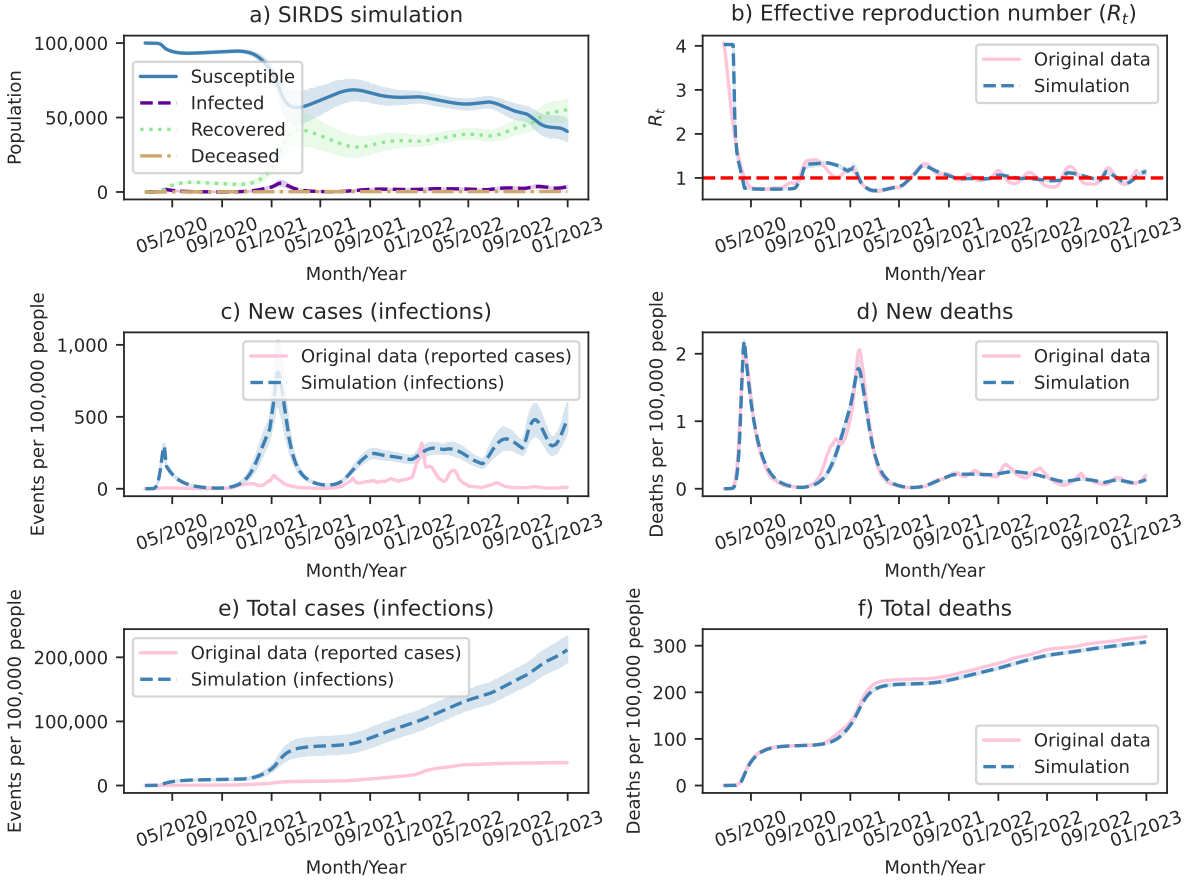


Figure 61 – Comprehensive analysis of simulation results for COVID-19 in the United Kingdom. (a) Model outcomes for an eight-day recovery period detailing the population compartments: Susceptible, Infected, Recovered, and Deceased. (b) Time series comparison between the effective reproduction number (R_t) estimated directly from reported deaths and R_t calculated by model simulations. (c) Time series comparison between new cases reported by health authorities and new infections in model simulations. (d) Time series comparison between new deaths reported by health authorities and new deaths in model simulations. (e) Time series comparison between cumulative cases reported by health authorities and cumulative infections in model simulations. (f) Time series comparison between cumulative deaths reported by health authorities and cumulative deaths in model simulations. Shaded regions depict the 95% Confidence Interval (CI).

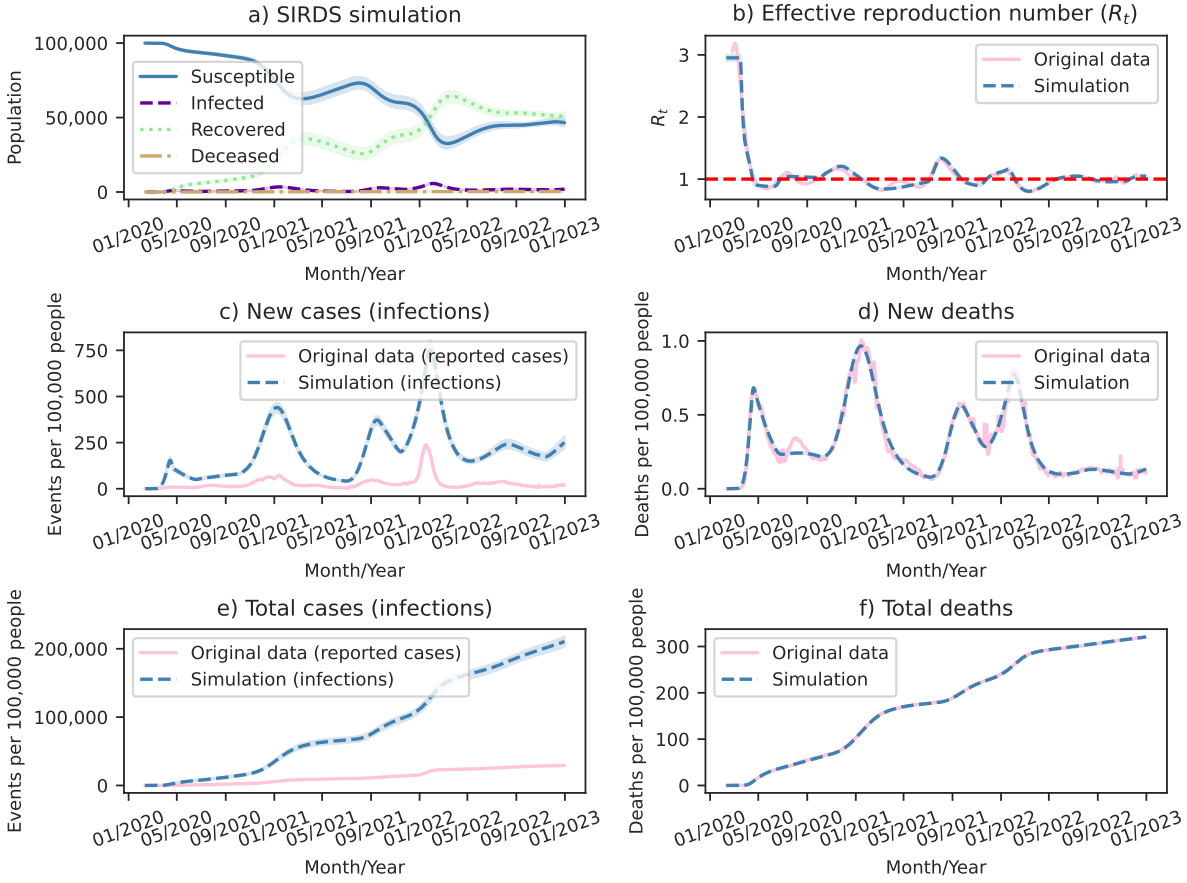


Figure 62 – Comprehensive analysis of simulation results for COVID-19 in the United States. (a) Model outcomes for an eight-day recovery period detailing the population compartments: Susceptible, Infected, Recovered, and Deceased. (b) Time series comparison between the effective reproduction number (R_t) estimated directly from reported deaths and R_t calculated by model simulations. (c) Time series comparison between new cases reported by health authorities and new infections in model simulations. (d) Time series comparison between new deaths reported by health authorities and new deaths in model simulations. (e) Time series comparison between cumulative cases reported by health authorities and cumulative infections in model simulations. (f) Time series comparison between cumulative deaths reported by health authorities and cumulative deaths in model simulations. Shaded regions depict the 95% Confidence Interval (CI).

APPENDIX K - Time-varying model parameters fitted for COVID-19 in Spain, the United Kingdom, and the United States

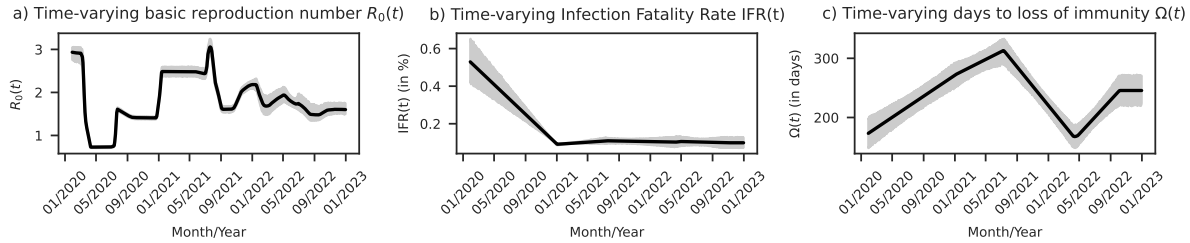


Figure 63 – Time-varying model parameters fitted for COVID-19 in Spain. (a) Basic reproduction number (R_0) varying with time (t). (b) Infection Fatality Rate (IFR) varying with t. (c) Days to loss of immunity (Ω) varying with t. Shaded regions depict the 95% Confidence Interval (CI).

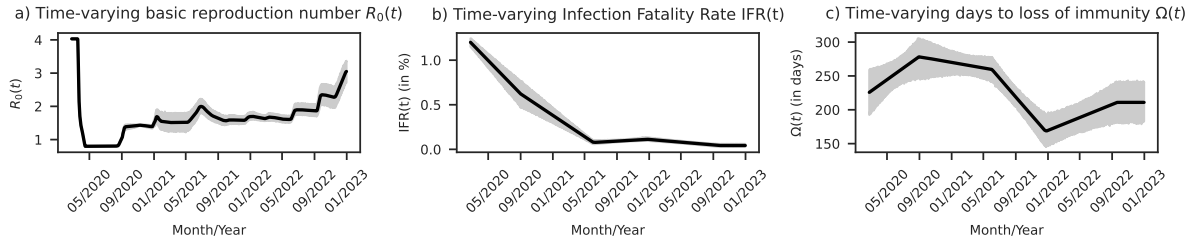


Figure 64 – Time-varying model parameters fitted for COVID-19 in the United Kingdom. (a) Basic reproduction number (R_0) varying with time (t). (b) Infection Fatality Rate (IFR) varying with t. (c) Days to loss of immunity (Ω) varying with t. Shaded regions depict the 95% Confidence Interval (CI).

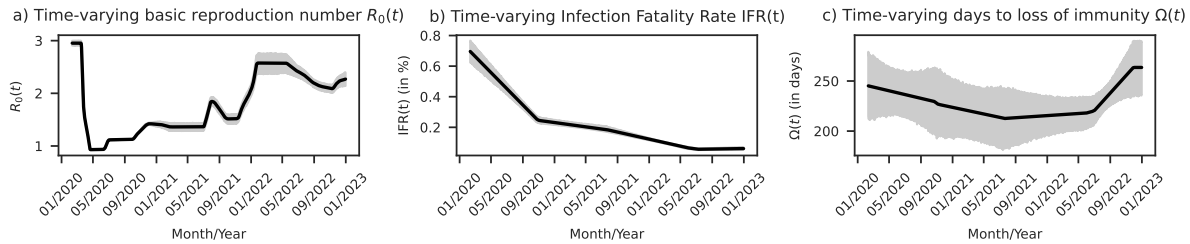


Figure 65 – Time-varying model parameters fitted for COVID-19 in the United States. (a) Basic reproduction number (R_0) varying with time (t). (b) Infection Fatality Rate (IFR) varying with t. (c) Days to loss of immunity (Ω) varying with t. Shaded regions depict the 95% Confidence Interval (CI).

APPENDIX L - Heatmaps illustrating the model performance for COVID-19 death forecast across the 41 largest Brazilian municipalities

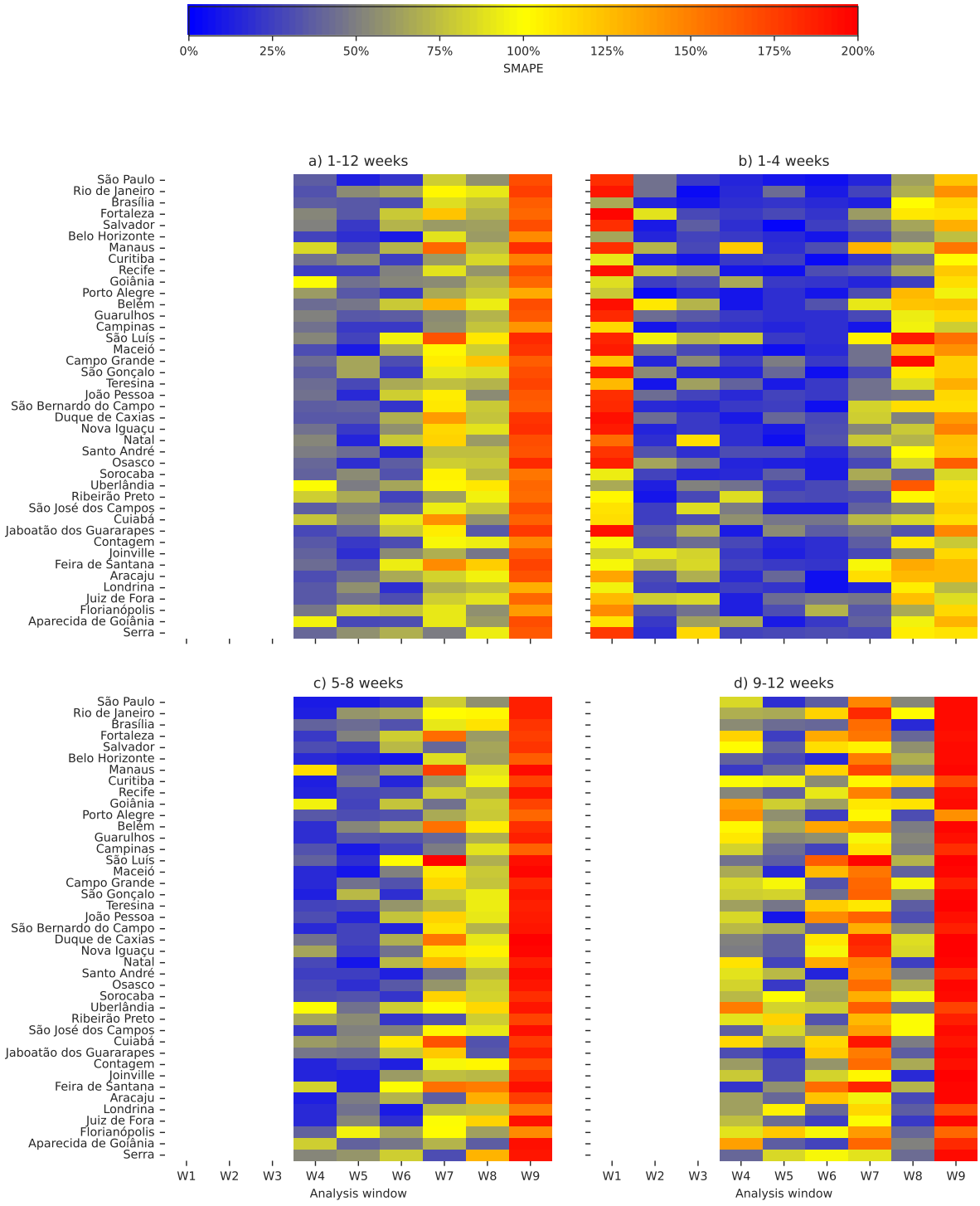


Figure 66 – Heatmaps illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across the 41 largest Brazilian municipalities, based on the LSTM model predictions over nine forecasting windows. The x-axis represents the forecast windows, with each 84-day window labeled as Wi where i ranges from the first window ($W1$) to the ninth window ($W9$). SMAPE values are shown for (a) the entire forecast period, (b) the first four weeks, (c) weeks five to eight, and (d) weeks nine to twelve.

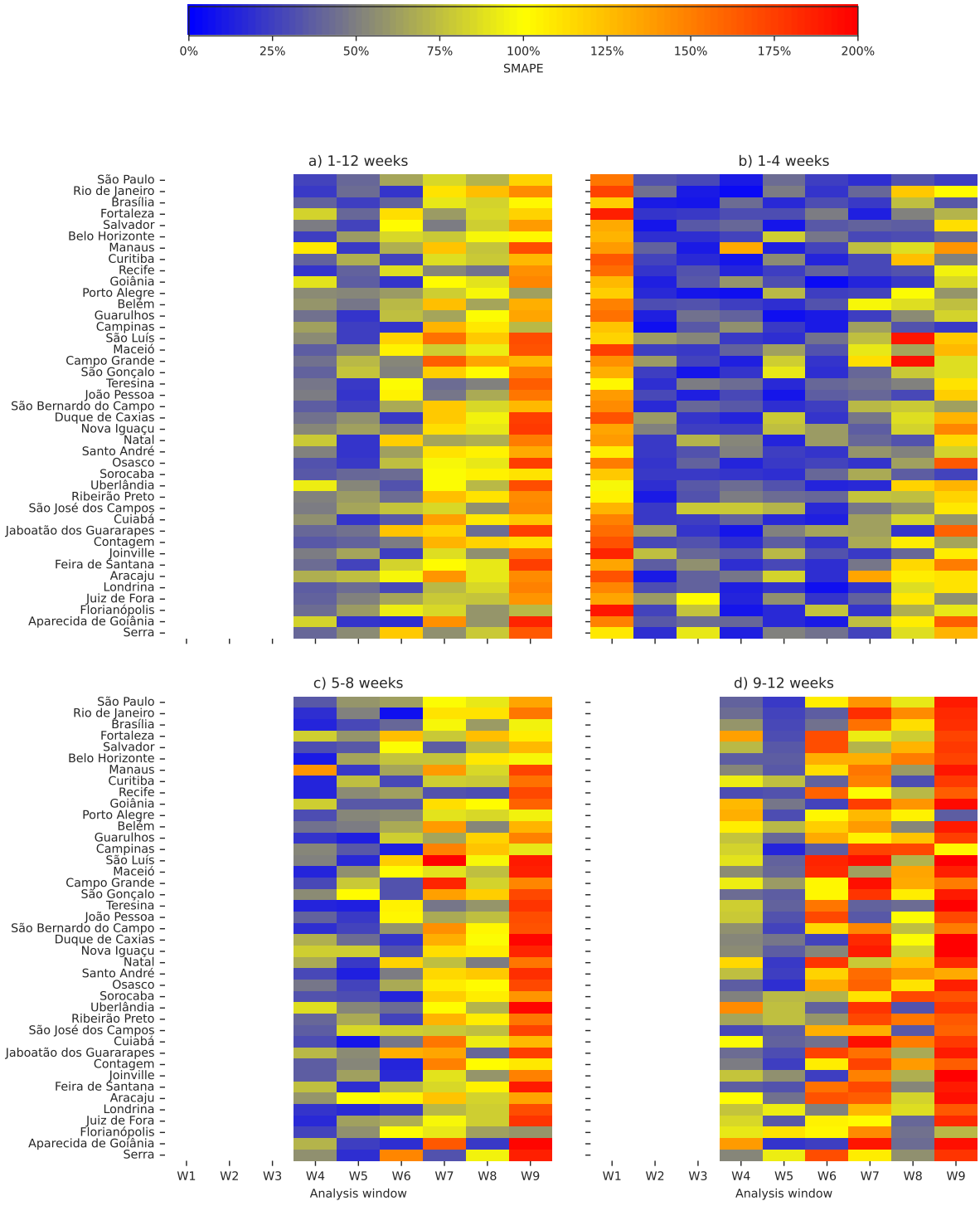


Figure 67 – Heatmaps illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across the 41 largest Brazilian municipalities, based on the Hybrid LSTM model predictions over nine forecasting windows. The x-axis represents the forecast windows, with each 84-day window labeled as Wi where i ranges from the first window ($W1$) to the ninth window ($W9$). SMAPE values are shown for (a) the entire forecast period, (b) the first four weeks, (c) weeks five to eight, and (d) weeks nine to twelve.

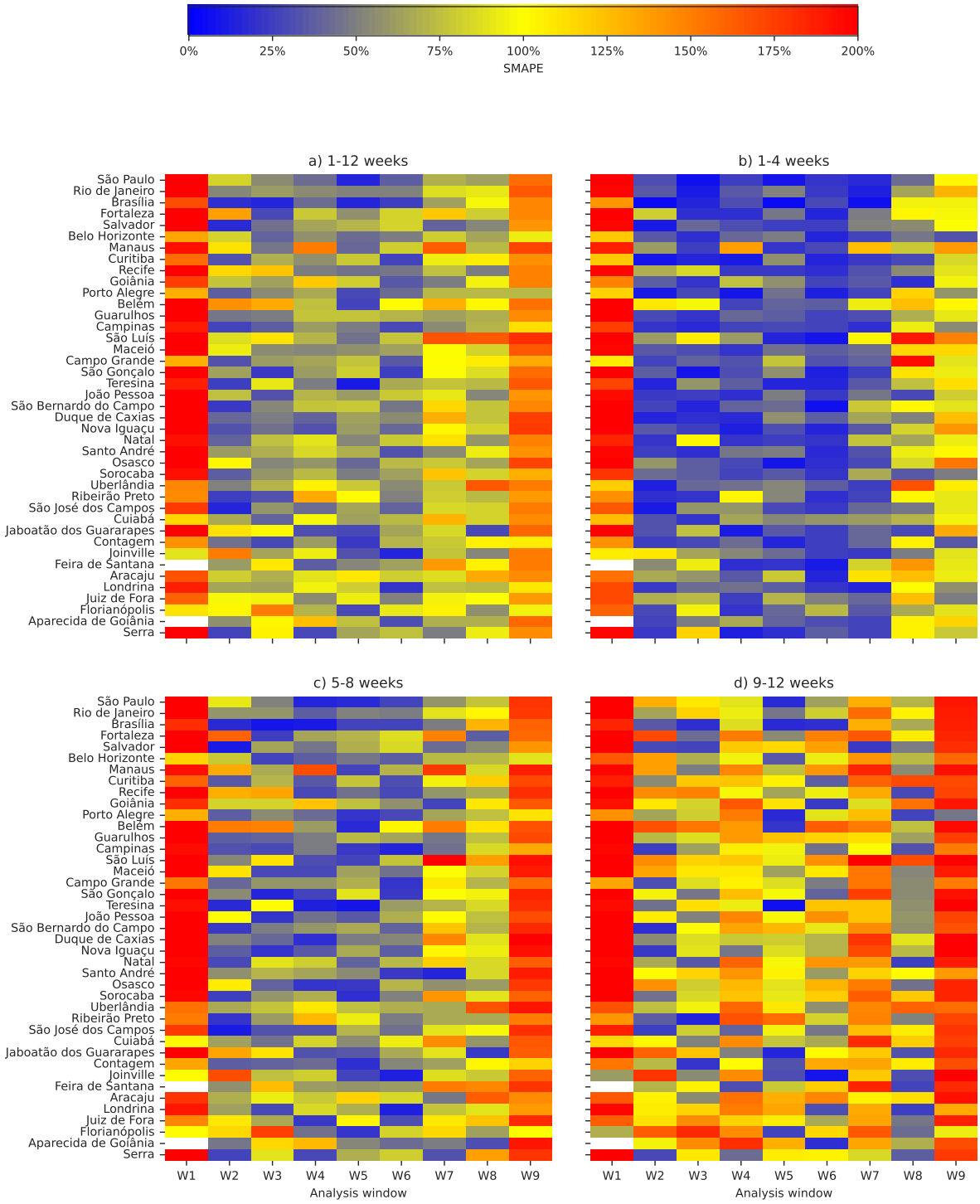


Figure 68 – Heatmaps illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across the 41 largest Brazilian municipalities, based on the Hybrid SIRDS model predictions over nine forecasting windows. The x-axis represents the forecast windows, with each 84-day window labeled as Wi where i ranges from the first window ($W1$) to the ninth window ($W9$). SMAPE values are shown for (a) the entire forecast period, (b) the first four weeks, (c) weeks five to eight, and (d) weeks nine to twelve.

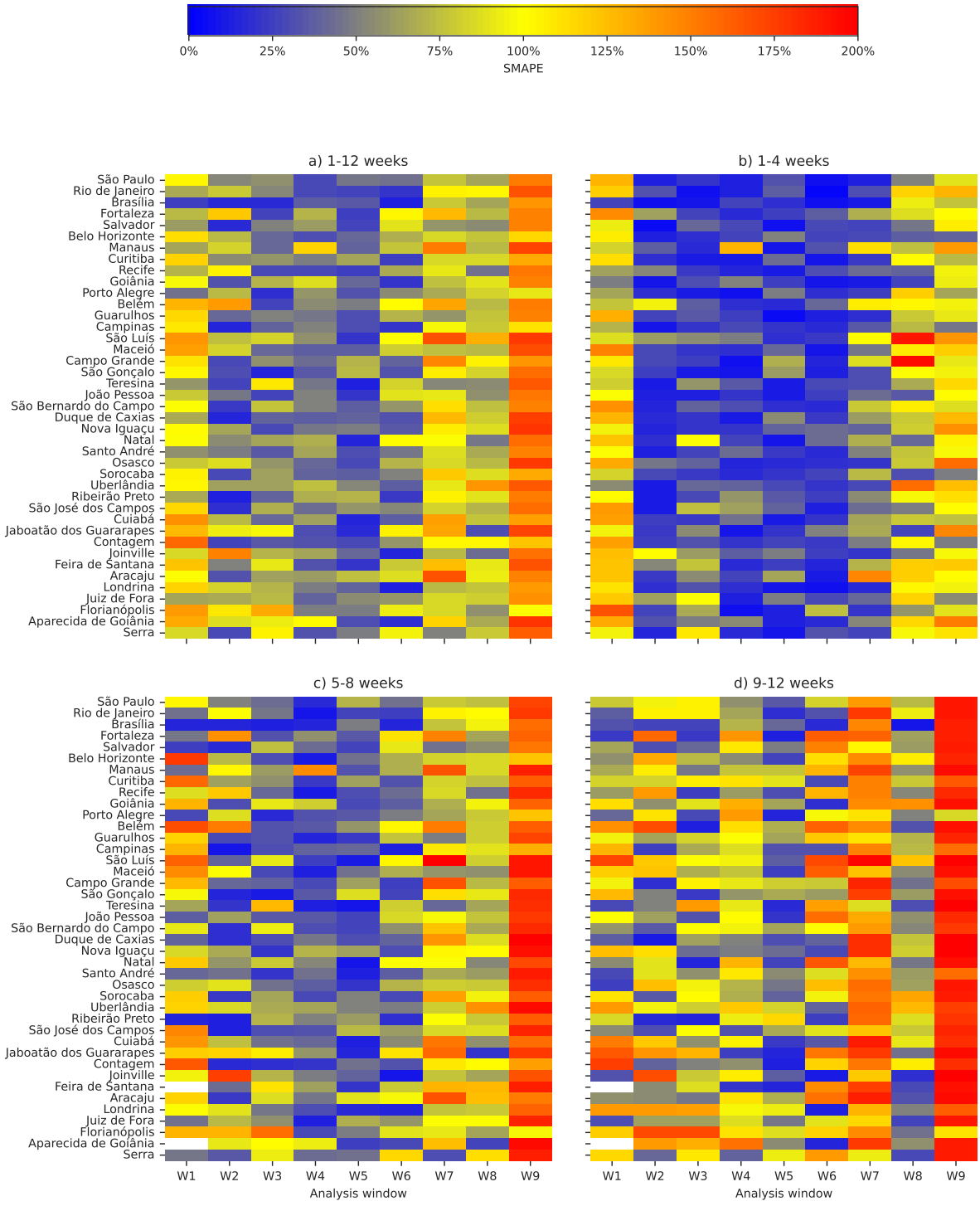


Figure 69 – Heatmaps illustrating the Symmetric Mean Absolute Percentage Error (SMAPE) for COVID-19 death forecasts across the 41 largest Brazilian municipalities, based on the Ensemble model predictions over nine forecasting windows. The x-axis represents the forecast windows, with each 84-day window labeled as Wi where i ranges from the first window ($W1$) to the ninth window ($W9$). SMAPE values are shown for (a) the entire forecast period, (b) the first four weeks, (c) weeks five to eight, and (d) weeks nine to twelve.