

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Agent models for disease propagation

Juliano Genari de Araujo

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-C²MC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Juliano Genari de Araujo

Agent models for disease propagation

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Master in Science. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Tiago Pereira da Silva

USP – São Carlos
March 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G324a Genari de Araújo, Juliano
Agent models for disease propagation / Juliano
Genari de Araújo; orientador Tiago Pereira da
Silva. -- São Carlos, 2024.
127 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2024.

1. agent models. 2. disease simulation. 3.
dynamic populations. 4. stochastic models. I.
Pereira da Silva, Tiago, orient. II. Título.

Juliano Genari de Araujo

Modelos de agentes para propagação de doenças

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Tiago Pereira da Silva

USP – São Carlos
Março de 2024

ACKNOWLEDGEMENTS

Special thanks to my advisor Professor Tiago Pereira (ICMC-USP) for the lessons during this program, and to my co-advisor and life long friend Professor Guilherme Goedert from Escola de Matemática Aplicada (EMAP) at Fundação Getulio Vargas (FGV) who invited me to start this work and guided me throughout, not only this work, but many other moments of my academic life. I'm also gratefully to Professor Claudio Struchiner, from EMAP, for his invaluable insights in epidemiology and modeling.

Additionally, I extend my gratitude to Professors Krerley Oliveira and Sérgio Lira from Universidade Federal de Alagoas (UFAL) and the team at LED-UFAL for their work in data collection and analysis, in partnership with the Municipal Government of Maragogi-AL. Their contribution were crucial to the development of this work.

Finally, I thank to all other members of the ModCovid-19 collaboration for their dedication and efforts in advancing our understanding of the COVID-19 pandemic and epidemiological modeling.

This work was made possible by the ModCovid-19 collaboration, funded in part by the Instituto Serrapilheira (grant No. Serra-1709-16124) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant No. 403679/2020-6). Additionally, this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001.

ABSTRACT

GENARI, J. **Agent models for disease propagation**. 2024. 127 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

As we experienced a major pandemic the necessity of smart interventions became very clear, but the decision for the best interventions to implement are usually based on educated guesses as each disease behaves differently and the macro behavior of the population can be very difficult to predict. Inappropriate interventions usually fail to consider heterogeneities in communities and can put the most susceptible part of the population at risk. To help in the evaluation of interventions, we developed highly modular and configurable software for stochastic agent model simulations: COMORBUSS, a software where the population is constructed in an organic way. Every person in the community is represented in the simulation and has an established routine, some actions are fixed (such as the time when that person goes and comes back from work), and some are randomly taken following probabilities to achieve a mean behavior. COMORBUSS can also be expanded in functionality with modules using some simple interface methods implemented in the main classes. With COMORBUSS and an airborne spread model for inside classrooms we tested different strategies for the return of schools after the first wave of the Covid-19 pandemic for the city of Maragogi-AL, in those simulations we arrived at the conclusion that for the safe opening of schools during a pandemic appropriate NPIs and behavioral protocols must be adopted, the vaccination of school teacher and other school staff is of paramount importance, as those workers are not only more susceptible than students, but they are also the main vectors of transmission. Uncontrolled school opening can be very dangerous as infection rates inside schools can explode leading to a significant increase in cases in the community.

Keywords: agent models, disease simulation, dynamic populations, stochastic models.

LIST OF FIGURES

Figure 1 – Maragogi location in Brazil. Left panel depicts 27 administrative divisions of Brazil, where Alagoas state is highlighted in black. Right panel displays the city of Maragogi (in black) inside Alagoas state.	22
Figure 2 – Maragogi-AL in comparison with Brazilian cities. Cumulative histogram of the total population of Brazilian cities (left panel), GDP <i>per capita</i> (right panel) as a function of the proportion of municipalities (IBGE. . . , 2010). The GDP <i>per capita</i> is conditioned on the group of Brazilian cities between 10000 and 50000 inhabitants.	23
Figure 3 – Age based probabilities of death and hospitalization for COVID-19 calculated for Brazilian cities. Expected death (left panel) and hospitalization (right panel) probabilities for Brazilian cities in the range 10000 and 50000 inhabitants. For hospitalization is assumed any individual developing COVID-19 severe symptoms, see Table 2.	24
Figure 4 – Maragogi-AL in comparison with cities in United States and World. The left panel displays the cumulative histogram of the total population of US cities (UNITED. . . , 2021) and the right panel displays World cities (WORLD. . . , 2021) as a function of the proportion of cities. Dashed lines in black correspond to 10000 and 50000 inhabitants for reference while the orange shows the Maragogi population.	24
Figure 5 – Economic value added from 2002 to 2018. Each economic sector contribution in Maragogi with respect to the total. Administration public includes Defense, Education, Public health and Social security.	25
Figure 6 – Student per class distribution inside schools over cities within the range of 10000 and 50000. The solid black line corresponds to the occupation density distribution over Brazilian cities within the range of 10000 and 50000 inhabitants and Maragogi’s is represented in blue. Left, center and right panels display the distribution for Early Childhood Education, Elementary school and High school, respectively.	26
Figure 7 – Protocol effectiveness. Effectiveness of different mitigation policies measured by cases increase for Maragogi-AL.	38
Figure 8 – Infection placement. Percentile of infections that happen in each service category.	38

Figure 9 – Age distribution. Comparison of real age distribution of Maragogi-AL and for a few randomly generated populations using our algorithm.	42
Figure 10 – Initial household state configurations. Compartment configuration probabilities used for active homes occupied by tree people	44
Figure 11 – Contact probability matrix. Color map representing probabilities that if a person from an age group in the y-axis met someone, that person belongs to the age groups in the x-axis.	48
Figure 12 – Example of a dynamic network for restaurants. Agents (with their ids in circles) in red are workers, agents in blue are visitors. Numbers inside each circle represent the identification number of that agent.	49
Figure 13 – Example of a dynamic network for markets. Agents (circles) in red are workers, agents in blue are visitors. Numbers inside each circle represent the identification number of that agent. A fraction of the workers is designated as cashiers and each client passes through one of them.	50
Figure 14 – Example of a dynamic network for classes in schools. Students are geographically positioned in lines, with one teacher in charge.	51
Figure 15 – Example of a dynamic network for hospitals. Agents (circles) in red are workers, agents in blue are visitors, agents in purple have been admitted to the hospital and are placed in a COVID-19 dedicated ward. Numbers inside each circle represent the id of that agent.	52
Figure 16 – Disease progression. Diagram illustrating how agents can transition between states of the disease.	55
Figure 17 – Group size distribution of the M database.	67
Figure 18 – Database diagram. PSF - Programa Saúde da Família, BOLSA-Programa Bolsa Família and SMC-Sistema de Monitoramento Clínico.	68
Figure 19 – Local picture of the network. One can see the cliques (dense clusters) interconnected by edges. Nodes are colored by age.	69
Figure 20 – Node degree distribution, considering only edges within cliques.	69
Figure 21 – Daily specificity $e(t)$, sensitivity $s(t)$ and constructed probabilities $p_{TP}(t)$ and $p_{TN}(t)$ in Equation (4.7). Note the dashed curves also rely on the sampling incomplete test types, so it changes over each realization of the reconstructed curve. Since the standard deviations are minimal, we chose to plot only the mean curve. Until April 28, only negative results were reported by rapid tests and the moving average has a window of 21 days.	73
Figure 22 – Final estimated curve for exposed, infectious and recovered compartments. The solid lines are the mean over all 400 trials and the dashed ones represent one standard deviation up and below.	76

Figure 23 – **Reference susceptible and infectious compartmental distribution curves for the city of Maragogi-AL in comparison to their calibrated versions calculated from COMORBUSS.** The recovered curves also include the deceased compartment. The solid curves represent the mean over 384 samples, the dotted curves limit a 95% percentile of the distribution, and the colored clear region is bounded by two shifted mean curves. These shifted curves are obtained by summing and subtracting the point-wise standard deviation over the 384 samples. 83

Figure 24 – **Reference exposed and infectious compartmental distribution curves for the city of Maragogi-AL in comparison to their calibrated versions calculated from COMORBUSS.** The solid curves represent the mean over 384 samples, the dotted curves limit a 95% percentile of the distribution, and the colored clear region is bounded by two shifted mean curves. These shifted curves are obtained by summing and subtracting the point-wise standard deviation over the 384 samples. 84

Figure 25 – **Histograms of the final epidemic size for different values of N .** The y -values are normalized so that the histograms represent a distribution. For low values of N the histograms are shifted towards the left side of the vertical dotted line, while for high values of N the tendency flips to the right hand side of the line. The variance decays as N grows, but the shape of the distribution still changes even for high values of N . Low values of N also show evidence of bi-modal behavior. 86

Figure 26 – **Final epidemic size (y -axis in %) vs number of days until the epidemic ends (x -axis in days) for different values of N .**The initial condition is $(S, E, I, R) = (.971, .007, .01, .012)$ for all realizations of the community. The **X** marker inside the clouds is the average over all points. The dotted line is a linear regression on the data, and ρ is the correlation between both variables (epidemic size and its total duration). 87

Figure 27 – Pipeline overview description. Data is collected as patients attend health institutions. Health professionals register patients’ personal, epidemiological, and geolocation data to the Clinical Monitoring System (CMS), which is blended with socio-economical and household data. Using these data, we estimate the number of Exposed (blue), Infectious (red), and Recovered (green) individuals. All the pre-processed data is used to calibrate our stochastic agent-based model, COMORBUSS. From bottom left to right: a schematic representation of the social dynamics of COMORBUSS, producing contacts between individuals in different social contexts. The colored circles represent the state of individuals and the lines represent relevant physical contacts capable of producing contagions. Once calibrated, the model is used to estimate the effectiveness of NPIs.	91
Figure 28 – Most relevant parameters. A non exhaustive classification of parameters used for a COMORBUSS simulation. A detailed description of parameters can be found in Chapter 3, while a complete list of parameters and their values can be found in the Git repository.	93
Figure 29 – Combination of NPIs measures in comparison to the baseline model settings. Left panel: Cases increase under different scenarios with unvaccinated teachers and staff. Right panel: Case increase in different scenarios with vaccinated teachers and staff. The effective teaching hours in hours/week $\frac{h}{w}$ and case increase in school population with respect to baseline are displayed for each NPI combination. In case the active monitoring is also applied, the mean and standard deviation over 60 realizations for the effective teaching hours are shown. The proportional increase in the number of cases is displayed as violin plots (median, lower, and upper quartiles), with kernel density estimates for distributions.	94
Figure 30 – Population fraction infected at the end of the simulation period (77 days) under varying vaccination coverage.	97
Figure 31 – Airborne transmission model inside school environment. The classroom is an enclosed space in which airborne transmission has a high chance of occurrence. Contaminated particles are spread over the classroom, allowing long range infections. The fresh air rate flow Λ quantifies the classroom ventilation. The quanta concentration C varies in the environment depending on the breathing activity.	98
Figure 32 – NPIs description. The icons distinguish the nonpharmaceutical interventions evaluated in this study. In scenarios involving masks, the mask penetration factor p_m is uniform for all individuals, except for teachers wearing PFF2 masks.	100

Figure 33 – Sensitivity analysis across mask penetration factor p_m. Cases increase in school population (solid lines) versus the mask penetration (mean values over 60 realizations for each p_m value).	102
Figure 34 – Sensitivity analysis across ventilation Λ. Cases increase in school population (mean and standard deviation) as a function of classroom ventilation rate. Dashed lines indicate the recommended ventilation rates: $\Lambda_1 = 0.8 h^{-1}$ (unoccupied room), $\Lambda_2 = 3.8 h^{-1}$ (half occupied room), and $\Lambda_3 = 6.6 h^{-1}$ (fully occupied room), following the ASHRAE standard for an average classroom in Maragogi.	103
Figure 35 – Population infected in case of increase in susceptibility. For each intervention scenario, we show the distribution in the percentile of the population infected provided the susceptibility of the population is increased uniformly by a multiplying factor.	103
Figure 36 – Effectiveness comparison for different demographics. Relative increase in cases for different scenarios for Curitiba-PR compared to Maragogi-AL.	126

CONTENTS

1	INTRODUCTION	19
1.1	Motivation	19
1.2	Research questions and the structure of this dissertation	20
1.3	Maragogi: our model city	21
2	AGENT-BASED MODELS: REVIEW	27
2.1	History	27
2.2	EMOD	29
2.3	PanSim	30
2.4	Nosoi	31
2.5	Covasim	32
2.6	COMORBUSS	33
2.7	Model comparison	35
3	COMORBUSS: MODEL DESIGN	37
3.1	Community Model	39
3.1.1	<i>Dynamics: stochastic model for community behavior</i>	39
3.1.1.1	<i>Services as community drivers</i>	40
3.1.1.2	<i>Visitation period</i>	41
3.1.2	<i>Creation: initializing a mimetic community model stochastically</i>	41
3.1.2.1	<i>Creating households while preserving age distribution and average household size</i>	41
3.1.2.2	<i>Household initialization of compartmental data</i>	42
3.1.2.3	<i>Service infrastructure and job allocation</i>	45
3.1.3	<i>Contacts: Service-specific networks</i>	45
3.1.3.1	<i>Standard networks: houses and generic services</i>	46
3.1.3.2	<i>Contact varying with agglomeration</i>	46
3.1.3.3	<i>Networks for environment layer</i>	47
3.1.3.4	<i>Network for restaurants</i>	47
3.1.3.5	<i>Network for markets</i>	48
3.1.3.6	<i>Network for schools</i>	49
3.1.3.7	<i>Network for hospitals</i>	50
3.1.4	<i>Community-defining Parameters</i>	52

3.1.4.1	<i>Services Parameters</i>	52
3.1.5	<i>Transportation layer</i>	53
3.2	Epidemiological Model	54
3.2.1	<i>Progression: stochastic compartmental model for the disease</i>	54
3.2.2	<i>Transmission: disease spread from contacts</i>	55
3.2.2.1	<i>Standard: contact through location-contextualized network</i>	55
3.2.2.2	<i>Specialized: aerosol transmission model in indoor locations</i>	56
3.2.3	<i>Disease-defining Parameters</i>	56
3.2.4	<i>Extra symptoms</i>	57
3.3	Interventions	57
3.3.1	<i>Quarantines</i>	57
3.3.2	<i>Social Isolation</i>	58
3.3.3	<i>Lockdowns and Services Closures</i>	59
3.3.3.1	<i>Decision process</i>	60
3.3.4	<i>Contact tracing</i>	60
3.3.5	<i>Testing</i>	60
3.3.6	<i>Vaccination</i>	61
3.4	Code availability	62
3.4.1	<i>Distribution and Documentation</i>	62
3.4.2	<i>Dependencies</i>	62
4	DATA INTEGRATION	65
4.1	<i>Modeling household networks</i>	65
4.2	SARS-COV2 DATA PROCESSING	70
4.2.1	<i>The reconstruction algorithm</i>	71
4.2.2	<i>Test data correction</i>	72
4.2.2.1	<i>Sampling incomplete test type</i>	72
4.2.2.2	<i>Inference of true positives and true negatives</i>	72
4.2.3	<i>Individual timeline reconstruction</i>	73
4.2.4	<i>Number of Cases estimation</i>	74
4.2.5	<i>The final curve</i>	75
4.3	Parameter estimation	75
4.3.1	<i>Estimation of parameters for household and indoor/outdoor environments</i>	76
4.3.2	<i>Estimation of service's visitation period</i>	77
4.3.3	<i>Estimation of service's contact network parameters</i>	78
4.4	The optimization program	82
4.5	Remarks about the population size	85

5	APPLICATION: PROTOCOL EVALUATION FOR SAFE SCHOOL ACTIVITIES	89
5.1	Materials and methods	90
5.1.1	<i>Data collection</i>	90
5.1.1.1	<i>Services</i>	90
5.1.1.2	<i>Street markets</i>	91
5.1.1.3	<i>Health services</i>	92
5.1.2	<i>Inference of states from data</i>	92
5.1.3	<i>Agent based modeling</i>	92
5.1.3.1	<i>Modeling disease</i>	94
5.1.3.2	<i>Interventions for the schools evaluation</i>	94
5.1.3.3	<i>Intervention in School Dynamics</i>	95
5.1.3.4	<i>Aerosol transmission model: masks and air exchange</i>	95
5.1.3.5	<i>Vaccination model</i>	96
5.1.3.6	<i>Modeling Services</i>	96
5.1.4	<i>Model calibration and closed schools as baseline</i>	97
5.1.5	<i>Poorly ventilated classrooms</i>	98
5.2	Results and Discussion	99
5.2.1	<i>NPIs and vaccination</i>	100
5.2.2	<i>Sensitivity analysis: mask penetration and ventilation</i>	101
5.2.3	<i>Scenarios with more infectious variants</i>	102
6	CONCLUSION	105
6.1	Conclusions on policy evaluation for schools	105
6.2	General conclusions	106
	BIBLIOGRAPHY	109
	APPENDIX A AIRBORNE TRANSMISSION MODEL	115
A.1	Aerosol-based model for infections in a closed environment	115
A.1.1	<i>Relevant length and time scales for aerosol particles</i>	115
A.1.2	<i>Time-evolution of radius-resolved particle concentration</i>	116
A.1.3	<i>Effective airborne transmission</i>	119
A.1.4	<i>Characteristic parameter values</i>	120
A.1.5	<i>Outdoor air exchange rate</i>	120
	APPENDIX B GENERALIZATION FOR CURITIBA-PR	123
B.1	Robustness of results for the capital Curitiba	123
B.1.1	<i>Inference of states from data of Curitiba</i>	123
B.1.2	<i>Baseline scenario</i>	124

B.1.3	<i>Calibration of the model</i>	125
B.1.4	<i>Robustness of results</i>	127

INTRODUCTION

1.1 Motivation

Infectious diseases can spread explosively and be more damaging in communities where they find that their transmission mechanics are compatible with the social structure and dynamics. As an example, the different levels of the social integration of elders in Italian and German families are often used to explain why this group had such different levels of mortality between these two countries ([MORFELD *et al.*, 2021](#)). In order to be effective, public health protocols must identify the most vulnerable groups in a community and the critical infection routes produced by that community structure and behavior, in order to change such elements in ways that suppress the transmission chains.

Finding these optimal changes can also lead to smarter protocols that minimize the social impact of the proposed interventions. However, this is also extremely challenging, and it is an effort that needs to be made case by case, as one cannot assume that the optimal policy for city A will be equally effective in city B. This is due to the extreme heterogeneity found between communities. In many cases, even in the same city can have completely different population densities, service infrastructures, and social behaviors. In a deeply unequal country as Brazil, even two neighbors can live in completely different realities; just take a look at the city of Rio de Janeiro for a classic example.

Although strict containment policies may be necessary (and sometimes even insufficient) in some communities, the same results may be achieved through less restrictive and damaging measures in smaller communities. Moreover, social and economic characteristics may lead to structural vulnerabilities in some communities so that they are disproportionately affected by a pandemic ([COELHO *et al.*, 2022](#)). Clearly, one cannot expect long lockdowns to protect families living in communities with high occupational density in their households, limited access to protective equipment (and in many cases, even clean water and soap) and which have no option

but to carry on with jobs with high exposure due to financial and food insecurity.

In this perspective static social models might not achieve the most realistic picture. Every individual may play multiple roles across different social contexts, and the simple routines prescribed by each role of every person in the community mix together to form a large and complex system. Such a system is also susceptible to change due to internal interaction as a consequence of external interventions. Moreover, this system is reactive to the threats posed by the public health crisis it is facing, and it has memory, with previous infections and interventions interfering with present and future infection chains.

Many agent models for diseases use static contact networks (e.g. [Kerr et al. \(2020\)](#)), usually derived from the contact matrices projected by [Prem, Cook and Jit \(2017\)](#). Such a strategy can be highly efficient (in terms of computational cost and ease of modeling each new population) but can also be limiting when evaluating interventions that can change the behavior of agents. With a static contacts network such changes can only be modeled (in a very limited way) by varying the weight of each edge in the network. It can also fail to capture the heterogeneities of the population, usually disproportionately misrepresenting the most vulnerable part of the population.

As computers grow more powerful, large stochastic agent simulations become possible. In those simulations, the population behavior is modeled in an organic way where each particle has a social role (a defined household, relatives, workplace, shopping habits, etc.) and each action is taken or not depending on a predefined schedule and by a group of probabilities, mimicking as closely as possible the human behavior. The intrinsic stochastic nature of such modeling requires that for each population, the simulation must be run many times with different random seeds to observe the mean behavior (and the possible extremes) for that population.

In this dissertation we will describe COMORBUSS, a bio-social agent model for the study of disease propagation in a community and the evaluation of mitigation measures, and some of its results. COMORBUSS is stochastic in its nature, dynamically generating contact networks in each step of the simulation depending on the location of each particle at each time. COMORBUSS is also highly modular, allowing new behaviors or even new infection mechanisms to be relatively easily programmed and used in simulations ([GENARI et al., 2022](#)).

1.2 Research questions and the structure of this dissertation

Throughout this text, we have taken on many questions regarding the advantages, limitations, and applications of agent-based models in epidemiology. The main questions that are addressed in this work are:

- What are the common features and history behind Agent-based models?

- Can we identify and address limiting modeling choices and features in some of the most interesting models for epidemiological applications?
- Can we tailor intervention policies during a pandemic to a particular community's demography and infrastructure in a systematic and transportable approach?
- How can we integrate epidemiological and social data not only to model heterogeneous behavior that is aware of social context, but also to calibrate these more realistic models?
- Can we build an adaptable interface to our model in order to respond effectively to relevant questions on intervention policy as they emerge in a pandemic (e.g., mix different infection models or types of social contact mechanics)?
- How can thousands of simulations be organized and performed systematically as realizations of counterfactual scenarios in order to quantitatively evaluate the efficacy of interventions in a practical setting?

In the first chapter of this dissertation we motivate the development of agents model and describe Maragogi, and justify why it is our model city. In Chapter 2 we define agents-based models, explore the history of such models, and evaluate four other models comparable to our model and what distinguishes our own model. In Chapter 3 we explore in detail the inner workings of the community model and the epidemiological model inside COMORBUISS and how they interact to generate disease spread and the interventions implemented. In Chapter 4 we describe how all the data necessary to run a community model were collected and processed to be used in COMORBUISS. We show a real application of COMORBUISS in Chapter 5, where we evaluate the impact of schools opening during a pandemic, and evaluate different scenarios to mitigate this impact. And finally in Chapter 6 we discuss the results from the schools evaluation and the potential impact and possible applications of models such as COMORBUISS.

1.3 Maragogi: our model city

The COVID-19 pandemic brought together policy makers, mathematicians, and epidemiologists in order to find effective interventions and minimize the damage of the pandemic. Many of the interventions evaluated in this work were brought to our consideration by the Mayor's Office for the city of Maragogi (AL), with which we developed a very productive partnership, coordinated by Professors Krerley Oliveira and Sérgio Lira (UFAL). It was thanks to this collaboration that we were able to acquire much data on the city's infrastructure, social dynamics, and pandemic response, which allowed for detailed modeling. Moreover, we argue in this section that Mararogi is a good representative of average municipalities in Brazil and that the lessons learned from its modeling can be very valuable for the pandemic response in a large number of cities in our country.

Maragogi is located in the northeast of the state of Alagoas approximately 137 km from the capital Maceió, see Figure 1.

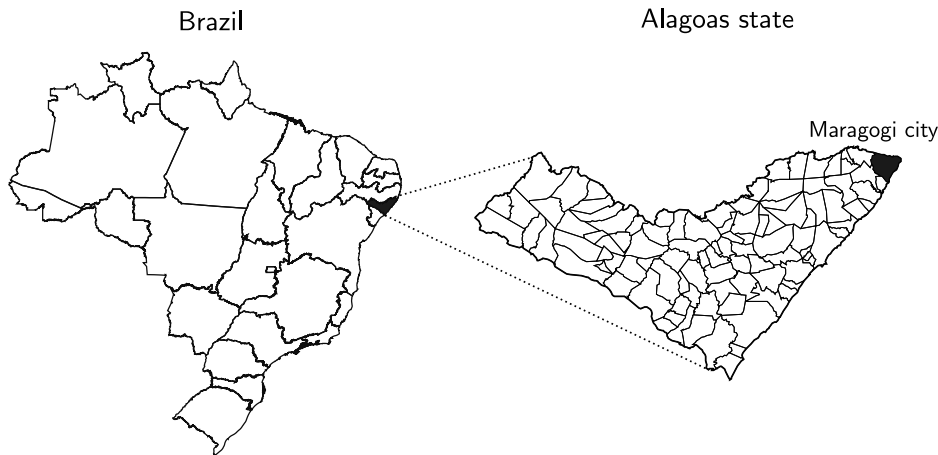


Figure 1 – **Maragogi location in Brazil.** Left panel depicts 27 administrative divisions of Brazil, where Alagoas state is highlighted in black. Right panel displays the city of Maragogi (in black) inside Alagoas state.

Demographics. The national 2010 survey (IBGE..., 2021) estimated that Maragogi had 28749 inhabitants, see Table 1. Note that in 2010 the population was mostly composed of young people (0 - 40 yo) and when compared to the current estimate, we observe a significant shift toward the mid-age (29 - 69 yo).

		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+	Total
2010		6016	6694	5220	4160	2861	1850	1177	539	232	28749
	(%)	20.93	23.28	18.16	14.47	9.95	6.44	4.09	1.87	0.81	100
2019		5542	6276	5967	4704	4102	2954	1933	1005	219	32702
	(%)	16.95	19.20	18.25	14.39	12.54	9.04	5.91	3.07	0.67	100

Table 1 – **Age pyramid of Maragogi.** The age pyramid shown in the first row corresponds to the national 2010 survey (IBGE..., 2021). In the second row, the age pyramid for 2019 is constructed using two databases and corrected due to biases in the data (such as duplicate registers for same individuals).

The national survey for 2019 estimated the population size in 32702 and 33351 in 2021 (IBGE..., 2021). To construct the age pyramid of Maragogi in 2019 we merged two databases. For the interval 0-79 y we used the Programa da Saúde da Família (PSF) — public health assistance program — summing over a total of 34598 inhabitants. For the 80 y - 100 + interval, we imported individuals for each 5-year interval from the Maragogi age pyramid estimated in the national 2019 survey (IBGE..., 2019), and the total number. We constructed the age pyramid of Table 1 multiplying by the factor $32702/34598$ that corresponds to the fraction between the total population size estimated in 2019 and the population size from the PSF data.

Comparing with other Brazilian cities, the left panel of Figure 2 shows the population size range between 10000 and 50000 inhabitants, which is the range corresponding to 44% of

Brazilian cities, and encompasses the city of Maragogi located within it.

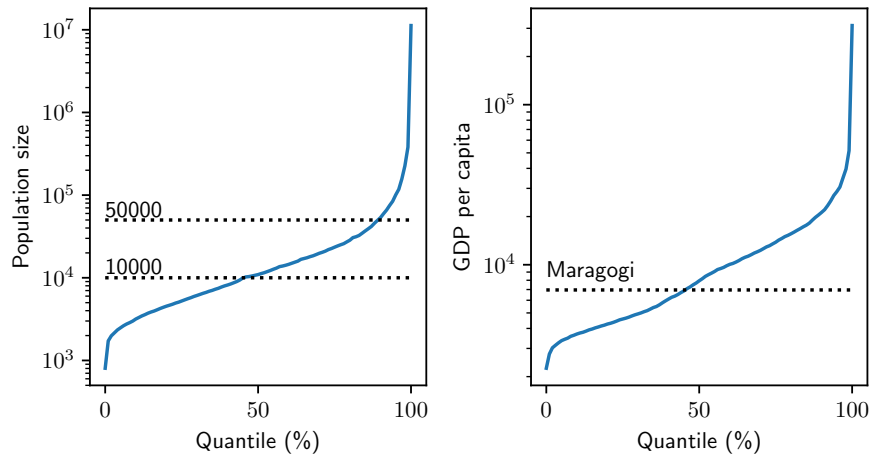


Figure 2 – **Maragogi-AL in comparison with Brazilian cities.** Cumulative histogram of the total population of Brazilian cities (left panel), GDP *per capita* (right panel) as a function of the proportion of municipalities (IBGE. . . , 2010). The GDP *per capita* is conditioned on the group of Brazilian cities between 10000 and 50000 inhabitants.

This range of cities between 10000 and 50000 inhabitants covers mainly cities that share common characteristics in terms of social and epidemiological synergy: small population size, low occupation density, and disease vectors such as public transport are not significant. In addition, there is a small portion of vertical urbanization.

Table 2 contains the probabilities of symptomatic cases, severe cases, and deaths aggregated by age group. Crossing those proportions with Maragogi’s age pyramid (Table 1) we obtain an expected hospitalized/infected ratio of $p_h = 3.304\%$ and a death/infected ratio of $p_d = 0.441\%$ in general. Figure 3 displays the age-based probabilities of death and hospitalization for COVID-19 (computed using the statistics in Table 2) calculated for Brazilian cities.

Age	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
p_{sym}	0.5	0.55	0.6	0.65	0.70	0.75	0.80	0.85	0.9
p_{hosp}	0.0001	0.0001	0.011	0.034	0.043	0.082	0.118	0.166	0.184
p_{death}	0.00002	0.00006	0.0003	0.0008	0.0015	0.006	0.022	0.051	0.093

Table 2 – Age based probabilities for COVID-19.

To put the city of Maragogi into context worldwide, Figure 4 shows the cumulative histogram of the total population from the simplemaps database containing 28372 and 41000 cities corresponding to US and world cities, respectively (UNITED. . . , 2021; WORLD. . . , 2021). We observe that the city of Maragogi is above the center in both cases, suggesting that it is a small urban area with an worldwide average population size (OECD. . . , 2020).

Economic aspects. If we narrow our analysis to this 10000 and 50000 inhabitants range, center panel in Figure 2 shows that Maragogi had GDP *per capita* close to the median in 2010.

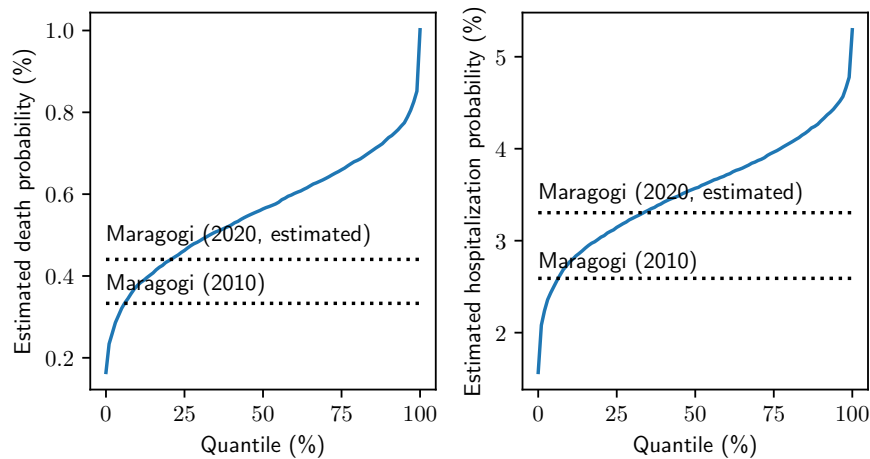


Figure 3 – **Age based probabilities of death and hospitalization for COVID-19 calculated for Brazilian cities.** Expected death (left panel) and hospitalization (right panel) probabilities for Brazilian cities in the range 10000 and 50000 inhabitants. For hospitalization is assumed any individual developing COVID-19 severe symptoms, see Table 2.

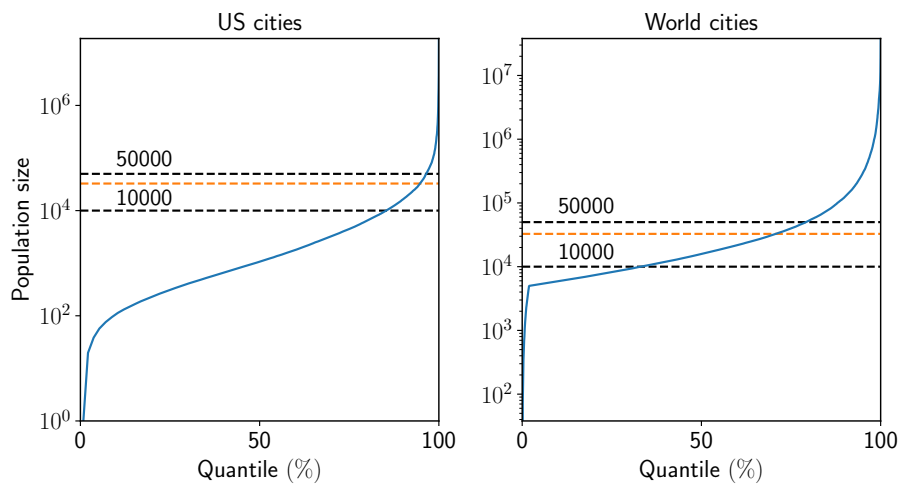


Figure 4 – **Maragogi-AL in comparison with cities in United States and World.** The left panel displays the cumulative histogram of the total population of US cities (UNITED..., 2021) and the right panel displays World cities (WORLD..., 2021) as a function of the proportion of cities. Dashed lines in black correspond to 10000 and 50000 inhabitants for reference while the orange shows the Maragogi population.

To illustrate the distribution of socioeconomic activities in the city, see Figure 5, which shows the economic value added in the last years.

The service sector is represented by a network of hotels and establishments that provide accommodation for travelers. We discarded this hospitality service sector from our analysis because most accommodation establishments were closed during the period of our analysis (Strategic plan May 2020, from City Hall information).

The farming activity splits into crops (44.7%), pastures (33.6%), woods, and forests (7.9%) in 2017. The Instituto Nacional de Colonização e Reforma Agrária - INCRA has registered

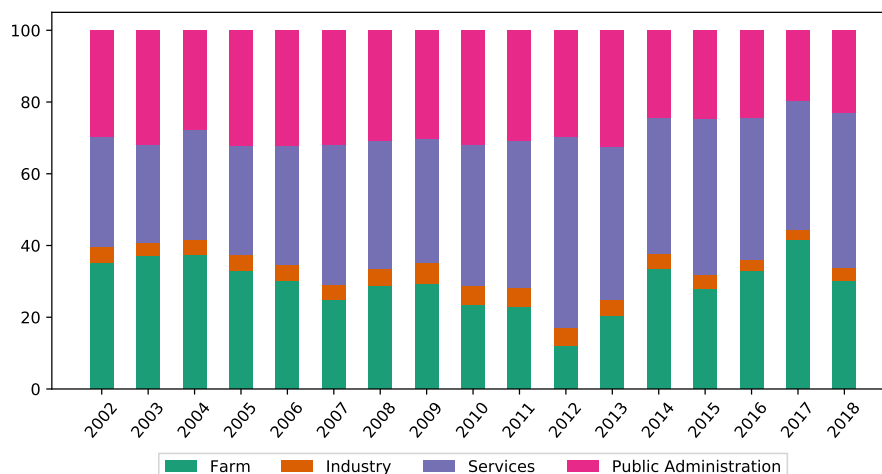


Figure 5 – **Economic value added from 2002 to 2018**. Each economic sector contribution in Maragogi with respect to the total. Administration public includes Defense, Education, Public health and Social security.

in terms of the Cadastro Ambiente Rural (CAR) 363 farming organizations, of which 89% correspond to small-holder farming organizations. Within this category of smallholder farming organizations, approximately 6% consist of rural settlements¹, where 1475 families practice agricultural activities (INCRA..., 2021; SICAR..., 2021). This familiar agricultural activity results in commercialization of products weekly in street market (under initiative of the City Hall).

Education. We filtered the data for schools belonging to municipalities in the range of interest. The data is composed by educational institution and school level (kindergarten, elementary, and high school) of INEP 2020 (INEP..., 2020), see Figure 6. Figure 6 shows the density of school occupation in Maragogi in the context of Brazilian cities within the range of interest. The distributions are similar in all levels of education, in particular Early Childhood Education (ECE), Elementary and High schools in the range of interest, the average of students per class corresponds to 15.25, 17.87 and 25.30, respectively, compared to 19.54, 19.49 and 23.37 of Maragogi schools.

¹ Rural settlement is defined as a portion of land that rural workers undertake to live on the plot and exploit it for their livelihood, using exclusively family labor (ASSENTAMENTO, 2021).

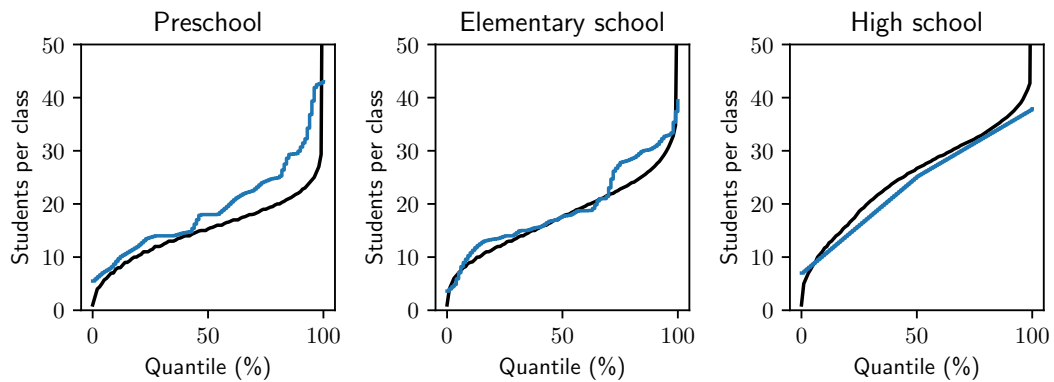


Figure 6 – **Student per class distribution inside schools over cities within the range of 10000 and 50000.** The solid black line corresponds to the occupation density distribution over Brazilian cities within the range of 10000 and 50000 inhabitants and Maragogi's is represented in blue. Left, center and right panels display the distribution for Early Childhood Education, Elementary school and High school, respectively.

AGENT-BASED MODELS: REVIEW

Before discussing the development and application of our agent-based model, COMOR-BUSS, let us take a step back and present a short review of this class of models, as well as some of its most interesting representatives, developed before and during the COVID-19 pandemic. Agent-based models (ABMs) are computational models that simulate and track individual units, which have behaviors defined by sets of rules or probabilities and can interact with each other or their environment. Based on this design, another popular name for this class is Individual-Based Models (IBM). These autonomous agents (which can model human individuals or other entities such as disease vectors or even organizations) can be individually tracked or intervened upon with the objective of assessing their effects on the system as a whole. These models are used in various fields, including economics, ecology, and epidemiology.

Focused on epidemiology, ABMs allow for a better representation of individual heterogeneity of behavior and characteristics, something that is very difficult to represent in differential equation models, for example. ABMs also can allow the behavior of individuals to be constructed in an organic way, where the complete behavior of each individual emerges from a simple set of intuitive rules, which also better represents the heterogeneity of individuals. This detailed characterization, as well as the ability to track each unit, constitutes a natural tool in the modeling and evaluation of public health policies, social interventions, and even the testing and deployment of pharmaceutical interventions (provided that we have a good biological model for the agents).

2.1 History

Before the 2000s, the complexity and applicability of ABMs were notably limited due to several key factors. Firstly, the computational power available at the time was a significant constraint. The intricate calculations required to simulate individual behaviors and interactions in ABMs demand substantial computational resources, which were not as advanced or readily accessible sufficiently before the 2000s. Secondly, the lack of sophisticated data collection and

processing technologies limited the amount and quality of data that could be fed into these models, restricting their accuracy and applicability. Furthermore, the interdisciplinary approach necessary to develop and implement effective ABMs was less prevalent because fields such as computer science, ecology, and social sciences were not as integrated as they are today. Limited software tools and programming languages suitable for creating complex simulations also posed a challenge. All of these factors contributed to the relatively simplistic nature and restricted use of ABMs in various domains before the turn of the millennium.

The review by Willem ([WILLEM *et al.*, 2017](#)) provides a detailed panorama of the development of ABMs between 2006 and 2015, when ABMs experienced significant advances and expansion of applications in the field of infectious disease transmission. One of the key advances was the growing diversification of the diseases modeled using ABMs, extending from endemic to emerging infections. This expansion was accompanied by a greater focus on the incorporation of intervention strategies and the evaluation of economic outcomes within models, indicating a shift from theoretical explorations to more practical policy-oriented applications. This was supported by the increasing ability of the ABMs to simulate complex heterogeneous interactions, both between and within hosts. This reflects a deeper understanding of the stochastic nature of infectious diseases and the importance of individual variability in disease dynamics.

However, the period also highlighted several challenges in the field of ABMs. A major issue was the inconsistency and ambiguity in the terminology used to describe these models, which posed obstacles to effective communication and knowledge sharing between disciplines. The lack of standardized reporting protocols and detailed model descriptions further compounded this problem, making it difficult to replicate studies or build upon previous work. Despite these challenges, [Willem *et al.* \(2017\)](#) underscored the potential of ABMs to improve targeted interventions for endemic infections and the importance of open-source collaboration. This point has been already made, as early as 2006, by [Patlolla *et al.* \(2006\)](#), and reinforced by [Hunter, Namee and Kelleher \(2017\)](#).

In [Patlolla *et al.* \(2006\)](#) the authors emphasize the significance of agent-based models in understanding complex systems, particularly in the realm of public health and epidemiology. It underscores the versatility of ABMs in simulating the spread of infectious diseases, considering the dynamic interactions between individuals and their environment. This work also highlights the potential of ABMs to aid in policy-making by providing insights into the effectiveness of different intervention strategies. It stresses the importance of integrating diverse data sources and the need for interdisciplinary collaboration in developing more comprehensive and accurate models. [Hunter, Namee and Kelleher \(2017\)](#), on the other hand, address a critical aspect of agent-based modeling in epidemiology: the need for a standardized taxonomy to classify and understand the diverse range of ABMs. Their work presents a structured framework to categorize these models based on specific criteria like disease type, societal model, transportation, and environmental factors.

In the next sections, we will explore two ABMs developed before and two developed during the COVID-19 pandemic, and, at the end, a brief introduction of our model. We will explore each model's advantages and disadvantages and discuss the limitations of each model, motivating the development of COMORBUSS.

2.2 EMOD

The Epidemiological MODELing software - EMOD (BERSHTEYN *et al.*, 2018) is a multi-disease agent-based model developed by the Institute for Disease Modeling (IDM), Bellevue, Washington, United States of America. EMOD was originally developed to simulate the spread of malaria but later was extended to other infectious diseases such as Polio, Dengue and HIV, among others. Beyond the pre-programmed disease models, EMOD allows for new custom disease models to be programmed through a class interface either in C++ or Python.

Disease propagation occurs between agents (or vector agents) localized in the same cell of a spatial grid, by default mixing, and infection probabilities are uniform inside each cell. An alternative infection behavior can be configured using the Heterogeneous Intra-Node Transmission option, where the population is divided into configurable groups, and infections between groups are weighted by a transmission matrix. Agents are assigned to a fixed cell, but migration between cells can be configured.

Advantages

- Very efficient to run, as reported by the developers, "typically on the scale of minutes to tens of minutes, depending on how the model is configured" (it is not clear for what size of simulation);
- Multi-disease, it has a set of disease dynamics already configured and an interface for programming custom disease dynamics;
- A large set of self tests already programmed by professional programmers;
- An internal system of messages which can be used to create sophisticated interventions;
- Ability to track all vectors (as agents) or a weighted sub-sample of vectors;
- Modular approach allows base functionality to be extended;

Disadvantages

- Low-resolution spatial grid (minimum cell size $\sim 1km^2$);

- Default infection behavior is uniform inside cells (which are already quite large); Heterogeneous Intra-Node Transmission can be configured to add weights to infections between certain groups of agents, but this is still not as fine-grained as we would like it to be;
- Hard to configure ("configurability comes at the cost of ease-of-use"), configuration is done through large JSON files;
- Static demographic model;
- More than 4 years without active development;

Although EMOD simulations can be very efficient, it does not implement any social dynamics, this limits the kinds of interventions that can be simulated. Heterogeneous Intra-Node Transmission can partially mitigate this issue, but infection is still defined by a fixed matrix of weights between each of the defined population groups. EMOD is highly modular in the disease model, but the population model is rigid and does not allow for any deep customization like COMORBUSS's allows. Both software are complex to configure and require many parameters, but this is a cost of being highly configurable; in this criterion, COMORBUSS mitigates this by offering reasonable default parameters for almost every parameter.

2.3 PanSim

Pandemics Simulator - PanSim ([REGULY et al., 2022](#)) is an agent-based model developed by the PPCU University, Budapest, Hungary, specifically for the COVID-19 pandemic. PanSim was developed to simulate the Hungarian town of Szeged, but other cities can be simulated, given the necessary parameters. It was developed in C++ and is parallelized to run in CPUs using OpenMP or in a GPU using CUDA. This makes the software highly efficient: as reported by the developers, it can run a simulation of one year with 179500 agents in a 10 minute time step in 64 seconds in a single NVidia V100 GPU).

In PanSim agents have different types (infants, elementary school students, full-time workers, etc.), and each type of agent has a few possible schedules to work, go to school, use services, etc. Each of these schedules has a probability associated (it is not clear in the documentation if the definition of the schedule for each agent is done at the start of the simulation and fixed for the whole simulation or if the schedule can change during a single simulation). In each time step, the infection mechanics are based on identifying the location of the agents and, for each location, adding up the infectiousness of the agents in that (weighted by a location-specific factor). Finally, this effective infectiousness value is used to challenge susceptible agents in that location. The disease model in PanSim is a slightly altered SEIRD (Susceptible, Exposed, Infectious, Recovered, Diseased) compartmental model that apparently cannot be easily changed to simulate other kinds of disease. Pansim also implements a vaccine model that modulates the agent's susceptibility and a few NPIs like quarantines, testing policies, services closing, etc.

Advantages

- Highly efficient to run: it can perform simulations of hundreds of thousands of agents for large but fine-grained time intervals in minutes in a powerful enough GPU;
- Reasonably configurable (mortality for different types of agents, quarantines, testing policies, services closing, etc.);
- Agents can have pre-existing conditions that affect mortality and medical services usage;
- Good spatial characterization: available houses, workspaces and schools are passed as parameters with coordinates, other locations are generated at random given the number of each type of location;

Disadvantages

- Not well documented: only the original paper, supplementary information and a few “readme” files available;
- Highly optimized implementation is not easily extended (functionality like contact tracing, public transport, vaccination trials, etc., can not be easily implemented);

PanSim is highly efficient, but this efficiency comes at the cost of extensibility. It has a good set of parameters for the simulation, but it does not have any interface to extend functionality: things we have done and we plan to do with COMORBUSS (mostly by using the modules API), like the airborne transmission model, contact tracing, public transport and vaccination trials, could not be easily done within PanSim. And the lack of public documentation about the usage and internal workings of PanSim only exacerbates those difficulties.

2.4 Nosoi

Nosoi ([LEQUIME *et al.*, 2020](#)) is an agent-based model developed in R focusing on dual host epidemics (such as arboviruses). It is based on the critical assumption that “the number of hosts infected during a simulation is orders of magnitude smaller than the total exposed population”, and on this assumption hosts (both human and vectors) only enter the simulation when they get infected. This allows Nosoi to be used to simulate populations in large geographical regions, since only a fraction of the population needs to be represented in the simulation (as seen in the examples given in the documentation, like on [Lequime and Dellicour \(2021\)](#)).

In a Nosoi simulation, each agent has a probability of moving, exiting the simulation (dying, being cured, leaving the study area, etc.) and of transmitting the pathogen, when a transmission occurs, a new agent enters the simulation. These probabilities are given as functions

by the user; the user also has to provide functions to calculate the standard deviation of the random walk in the space if the agent moves and the number of contacts for each agent.

Advantages

- Very simple and easy to use interface;
- Extra parameters (discrete or continuous) for agents can be defined and used on the core functions;
- Can be used to run large geographical regions (with spatial modeling done by the user);

Disadvantages

- Only applicable to low prevalence epidemics (critical assumption);
- Focused on dual-host type diseases (but it can run single-host diseases);
- Simplistic disease model (no built-in infection mechanics, only simulate infectious agents);
- No built-in social model, making it a bad candidate to evaluate social intervention;

Complex scenarios can be modeled in Nosoi, but all the complexity must be modeled and programmed by the user inside the agents' parameters and the 5 functions to calculate the probabilities, movement, and contacts. This makes Nosoi highly flexible, at the cost of almost every part of the simulation to be modelled by the user. Nosoi itself does not offer any social model and a very simple disease model. We also were unable to find any documentation on Nosoi's efficiency, it probably has reasonable or good efficiency, since it doesn't run the entire population inside the simulation.

2.5 Covasim

COVID-19 Agent-based Simulator - Covasim (KERR *et al.*, 2020) is an agent-based model developed in Python by the Institute for Disease Modeling (IDM) specifically for the COVID-19 pandemic. Covasim uses a modified SEIR model with compartments for multiple levels of symptomatic presentation. It has no discrete spatial model, different contact networks are used depending on where the particle is located.

It has three modes for contacts: (i) static contact networks generated by SynthPops, another software developed by IDM, these networks try to better represent realistic interactions in the different contexts; (ii) random contact networks, these are more dynamic but do not realistically represent most interactions between people; and (iii) a hybrid contact where a mix of

both modes are used. The static, random or hybrid social structure implemented on Covasim fails to capture any heterogeneity in the population, and it also makes difficult to represent certain kinds of interventions that changes the particles behaviour in specific locations.

Covasim also has a few pre-programmed interventions, such as quarantines and a simple contact-tracing (based on a given probability of being traced), as well as a modular interface for interventions, where the user can pass functions that can read and change simulation parameters in each step. It also has a vaccination mechanic, where vaccines uniformly reduce the susceptibility of agents.

Advantages

- Fairly efficient (it can run tens of thousands of particles for 90 days in seconds);
- Simple to configure (at the cost of flexibility);
- Good documentation;
- Powerful set of calibration and analysis tools;

Disadvantages

- High reliance on static contact networks based on inferences from few pre-pandemic observations and low granularity in social settings;
- Very limited interface for extending functionality (only an interface for custom interventions is available);
- Most interventions are based on susceptibility with low or no heterogeneity;

Covasim has a simple but effective way to extend it's functionality for interventions. But it's social model is not organic and heavily relies on synthetic fixed contact networks; this limits the kinds of scenarios that can be simulated with Covasim and in most cases fail to represent heterogeneities in the simulated communities. Covasim also has a fixed SEIR disease model with no interface to extend or replace the disease model.

2.6 COMORBUSS

COmmunitary Malady Observer of Reproduction and Behavior via Universal Stochastic Simulations - COMORBUSS (GENARI *et al.*, 2022), is a bio-social agent model initially developed in Python, by the ModCovid collaboration, for studying the spread of the SARS-CoV-2 within communities and assessing the impact of various mitigation measures. It stands out for

it's dynamic and modular nature, employing a unique approach to model population behavior and interactions.

It implements a modified SEIR model, with different compartments for different symptomatic presentations, but due to it's modular approach this base model can be easily extended. For example we already have implemented an module to extend the symptoms from asymptomatic, light, and severe, to any number of symptoms configured by the user. The social model inside COMORBUSS is built in an organic way, where every agent has a house, place of work, shops it visits, etc, and a defined routine (made of fixed schedules for work, school, etc, and visitation frequency for others like markets, shops, etc) and contacts are generated dynamically on each simulation step (typically 1h), each kind of location has a specific network generator modeled to mimic the typical behavior of agents in each location.

COMORBUSS has implemented mechanics for quarantines, lockdowns, social isolation, vaccinations, public transport, testing policies and others. Beyond all the built in functionality it also offers a robust coding API to be used to modify or extend any part of the simulation, this was extensively used to evaluate the impact of schools reopening using COMORBUSS, where a new airborne infection mechanism was implemented to simulate closed spaces, also a sophisticated school model (class separation, schedule, teacher assignment, etc.) and school specific interventions (reductions in teaching time, classes segregation, class wide and school wide suspension based on testing policies, etc.) where implemented as modules extending the original COMORBUSS functionality for this work.

Advantages

- Organic social modeling, contact networks are dynamically generated at each simulation step based on agents' social roles, locations, and schedules;
- Highly modular, allowing for the integration of new behaviors, interventions, or infection mechanisms (while still providing a robust base model);
- Detailed agent roles, each agent in the simulation has defined social roles, such as household membership, workplace associations, and shopping habits, contributing to the realism of the model;
- Built in tools for running simulations with varying parameters and to visualize and analyse sets of simulations;

Disadvantages

- Computational intensity, due to it's detailed and dynamic nature, COMORBUSS may require significant computational resources (particularly for large populations or extended simulation periods);

- Complex configuration, the complexity and modularity of the model might lead to challenges gathering the necessary data and configure the model;

COMORBUSS offers a unique approach to simulating disease spread and evaluating public health interventions. It's emphasis on dynamic contact networks and the ability to incorporate varying behaviors (capturing population heterogeneities) and modularity make it a powerful tool for understanding complex epidemiological dynamics in diverse community settings. However, these advantages come with the trade-offs of increased computational demands and complexity in model configuration and analysis.

2.7 Model comparison

In this comprehensive review, we have explored a variety of agent-based models (ABMs) designed for epidemiological simulation, each with its unique strengths and limitations. The comparison in Table 3 effectively summarizes their key attributes, offering a clear perspective on their capabilities and suitability for different modeling scenarios. Models like EMOD and PanSim demonstrate high computational efficiency, ideal for large-scale simulations, but they vary in spatial characterization and flexibility. Nosoï, while highly customizable, is tailored for low prevalence, dual-host type diseases, limiting its broader applicability. Covasim, specifically designed for COVID-19, offers a balance of efficiency and functionality but relies heavily on static contact networks and has limited extensibility.

COMORBUSS, our model, stands out for its dynamic and organic social modeling, providing a nuanced simulation of interpersonal interactions and behavior. Its modularity and detailed agent roles allow for a comprehensive exploration of various public health interventions and their impacts on community spread, albeit at the cost of computational intensity and complex configuration requirements. This model's ability to adapt and incorporate a wide range of behaviors and interventions makes it a powerful tool for understanding the intricate dynamics of disease spread in diverse community settings.

The choice of an appropriate ABM for a particular epidemiological study hinges on the specific requirements of the study, including the scale of the population, the level of detail needed in social interactions, the nature of the disease being modeled, and the computational resources available. Each model presents a unique set of features that make it suitable for different aspects of epidemiological modeling and public health policy simulation. The advancement of these models, especially in light of the recent COVID-19 pandemic, underscores the vital role of ABMs in understanding and managing infectious diseases.

	Computational Efficiency	Configurability	Spatial Characterization	Social Model	Interventions	Extensibility	Transmission Model
EMOD	High	High	Large regions at low resolution	None	Flexible interventions with messages system	Limited to disease model and interventions	Contact*, airborne, STI, vector, environmental, can be extended
PanSim	Very High	Medium	Small regions at high resolution	Model based on inputted individual schedules	Limited	No support; opensource, but not well documented	Contact*
Nosoi	Not specified	Limited, most modelling done by the user	Lower density, allows for large regions	Modelling done by the user	Modelling done by the user	Very high (and needed)	Agents generated as infected; Single and dual-host (vectors)
Covasim	High	Limited	None	Static or random	Limited; can be extended	Limited to interventions	Contact*
COMORBUSS	Medium	High	No built-in spatial characterization; but can be extended	Dynamic, based on social roles and contexts	Large set of implemented interventions; can be combined and extended	Very high with API for modules	Contact* and airborne; can be extended

Table 3 – Comparison of the explored Agent-Based Models (* definition of contact is sensitive to the spatial and social models)

COMORBUSS: MODEL DESIGN

Communitary Malady Observer of Reproduction and Behavior via Universal Stochastic Simulations - COMORBUSS, is a bio-social agent model for the study of disease propagation in a community and the evaluation of mitigation measures. Let us clarify each part of this statement.

An agent model is one where we simulate individual agents which represent persons in the modeled community. These agents interact with each other and the environment according to a set of rules and have their own characterization. This allows for the creation of models which capture the heterogeneity of the real community we are studying. Moreover, mitigation measures can be directly modeled by modifying the behavior of the agents (e.g. quarantines, social isolation, reduction of students in classrooms) or the transmission of the pathogen (e.g. masks and vaccination). In this way, the effectiveness of these mitigation policies can be measured and compared directly, see Figure 7.

By bio-social, we mean to emphasize that COMORBUSS at its core is driven by two stochastic models: one for disease progression and propagation based on the individual biology of the agents, and the other for the social dynamics of the agents based on their identities and roles in the community. Connecting these two models is the core modeling assumption that disease transmission rides on social contacts produced by community dynamics. As the social dynamics model drives the individual agents as workers or clients of the services which define the infrastructure of the community (such as hospitals, schools, markets, restaurants, stores, etc.), the agents meet at these locations and possibly infect others. As transmission is contextualized by location and by the roles of the agents involved (e.g., client, worker), we can identify which are the services that contribute the most to the driving force of the infection; see Figure 8.

COMORBUSS as an agent-based model possesses the following remarkable advantages which are derived from **our approach to directly model social dynamics** and the **omniscience the model guarantees to the analyst**:

- individualized and heterogeneous description of the community;

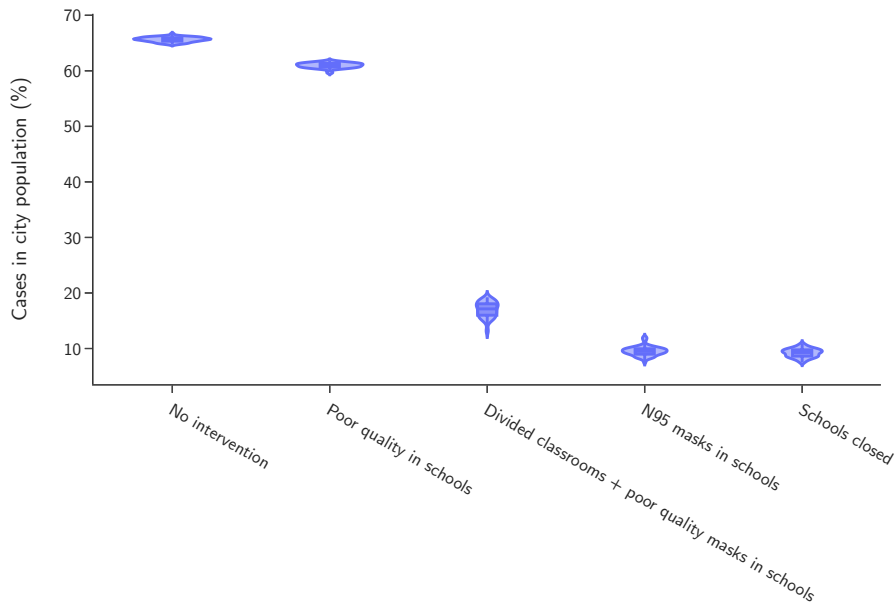


Figure 7 – **Protocol effectiveness.** Effectiveness of different mitigation policies measured by cases increase for Maragogi-AL.

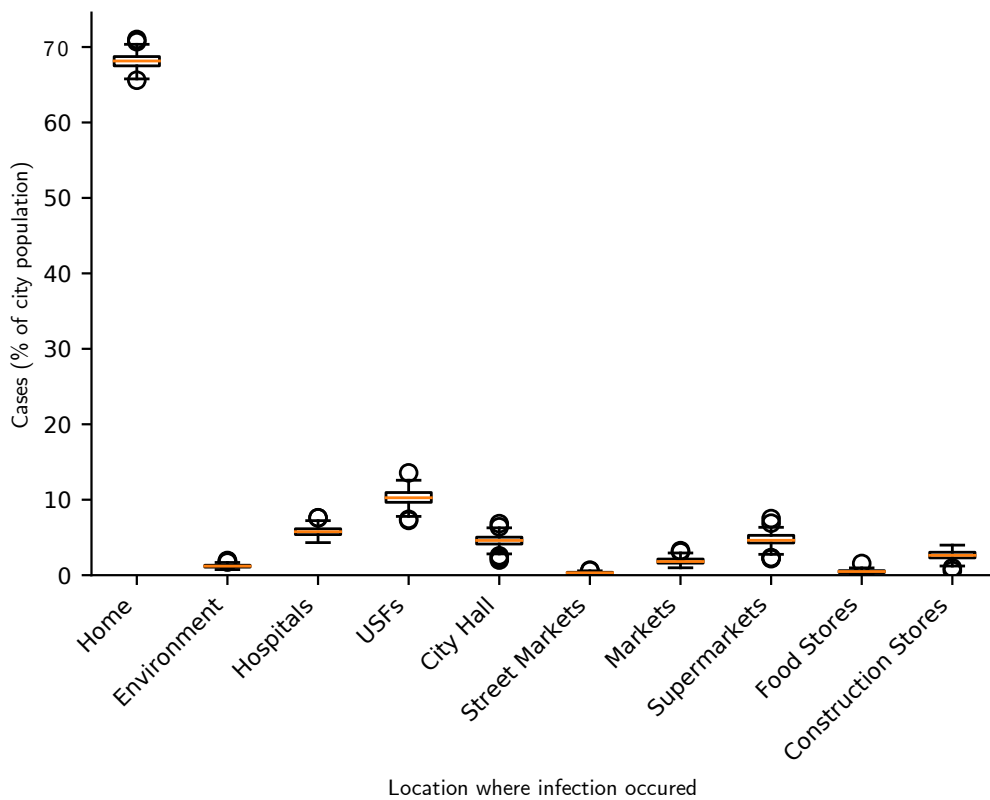


Figure 8 – **Infection placement.** Percentile of infections that happen in each service category.

- behavior models for interventions and their quantitative assessment, even with partial compliance;
- realistic decision making models with dynamic criteria for adoption of interventions;
- ability to produce counterfactual scenarios regardless of the complexity of the scenario, enabling direct comparison in experimentation;

As we saw in the last chapter, all these advantages make COMORBUSS a valuable tool in both the evaluation of policies and the development and testing of new ideas and methods in epidemiology. Now we will explore how the models inside COMORBUSS work and how they interact to create those advantages.

3.1 Community Model

We seek the average epidemiological behavior and the associated variance for a city with a given demography. This is done by simulating multiple realizations of a stochastic model for the disease propagation in this community. In order to eliminate biases introduced by a single societal network, we generate for each random seed a new community representation which captures the following real demographic information of a given city:

- population size;
- age distribution (binned in groups of 5 years);
- household structure (size distribution and age distribution of members);
- service infrastructure;
- job allocation by age group.

3.1.1 Dynamics: stochastic model for community behavior

The core concept of COMORBUSS is the utilization of services to dynamically generate contacts in our model community. As such, each relevant social context is modeled as a service, even "the environment", as is dubbed random meetings on the streets and parks. The services which have been modeled in this work are:

- health facilities: hospitals, public health clinics;
- educational facilities: schools and day-cares;
- essential stores: street markets, markets, supermarkets, food stores, construction stores;
- city hall and environment.

3.1.1.1 Services as community drivers

In the intricate web of community dynamics, services play a pivotal role, acting as the primary drivers that shape the movements and interactions of individuals within a community. These services, encompassing various essential categories such as supermarkets, hospitals, and other public facilities, become focal points where agents converge and interact in different roles.

Agents interact with these services in three distinct roles: as workers, visitors, or guests. Workers are those who are employed by the service, visitors are regular users of the service, and guests are individuals who are temporarily attached to a service due to specific circumstances like hospitalization or quarantine. Each agent is typically linked to a specific instance of a service category, creating a consistent pattern of visitation. However, flexibility is maintained as agents can be reassigned to different instances if circumstances such as service closures arise. This can happen when a service is closed due to a lack of workers (all workers being in quarantine or hospitalized).

Each service has two defining restrictions: its working days and hours, as well as the age groups allowed to use it. Another key parameter is the average period of visitation for that service (e.g. one can say that any person visits the supermarket every week). From this we have the average frequency in days that the service is visited and, using the number of working hours of that service, we compute the hourly probability that an agent will visit that service.

During every hour that a service is open, free visitors are randomly selected and sent to the instance of that service to which they are assigned. If the agent is unable to make a visit (e.g. agent is working or visiting another service), the probability is accumulated to a later hour that the service is open and the agent is available to visit. In this way, we organically produce "rush hours", such as when many workers visit the supermarket after their working hours. After the visit is concluded, the agent is returned to its address until it is selected again for some other activity. Similarly, workers are sent to the instance of the service where they work during their working hours. One can also assign the agents uniformly at different shifts. Guests are so far defined only via hospitalization or quarantines, so their mobility is restricted until the associated measure is completed. They are then returned to their home, where normal social activities are resumed.

Another key component of our model is the interactions between agents within these services, which are facilitated through dynamic service-specific contact networks. The intricacies of these contact networks and their implications on agent interactions will be explored in detail in Section 3.1.3.

This system of service-driven agent dynamics serves as a robust framework for modeling community behavior. It captures the essence of how individual needs and activities are interwoven with community services, creating a realistic and complex portrayal of community life. Any nonpharmaceutical intervention can be modeled as temporary changes in the individual or

collective behavior of the agents, and its consequences can be measured directly.

3.1.1.2 Visitation period

Leaving aside the interaction of agents within the service for a moment, the visitation of agents is what contributes the most to the relevance of services in the disease spreading in the community. COMORBUSS models the visitation of agents to a service by randomly picking them according to a probability p_v . This happens at every time step that the service is opened and that the agents are free to visit, which means that they are not resting at home or visiting any other service. The probability p_v is then given by the inverse of the visitation period v_p : $p_v = 1/v_p$, where v_p is a measure of how many time steps an agent takes to return to a service, given that it is opened¹. To make the visitation period independent of the opening of the service and also of the magnitude of time steps, we assume that it is provided in consecutive days and then we convert it to time steps. The conversion formula is given by

$$v_p = \frac{d_o h_o}{7 \Delta t} v_{pc}, \quad (3.1)$$

where d_o is the total number of days a service is opened on a week, h_o is the total number of hours a service opens for a day, Δt is the time step in hours, and v_{pc} is the visitation period in consecutive days. To calculate v_{pc} , it suffices to know the total number of visitors v_w a service receives during a week, and the total number of agents v_t that can in fact visit the service. With these two values, the visitation period in consecutive days is given by

$$v_{pc} = 7 \frac{v_t}{v_w}. \quad (3.2)$$

3.1.2 Creation: initializing a mimetic community model stochastically

3.1.2.1 Creating households while preserving age distribution and average household size

The agents are created in household groups that are defined sequentially and modified such that real age distribution and average household size are respected. In order to avoid unrealistic household structures (e.g. children living unsupervised) and to consider households with sizes far from the average, we have created and carefully curated an artificial dataset of households mimicking the households in the modeled city.

Although the generated population is smaller than the desired population, a household is sampled from the reference population dataset. We then evaluate the average size of the households created so far: If it is smaller than the desired average household size, a new agent is added to this house; if it is larger, an agent is randomly removed from this house.

¹ A direct implication of this definition is that the visitation period cannot have time length less than that of a time step

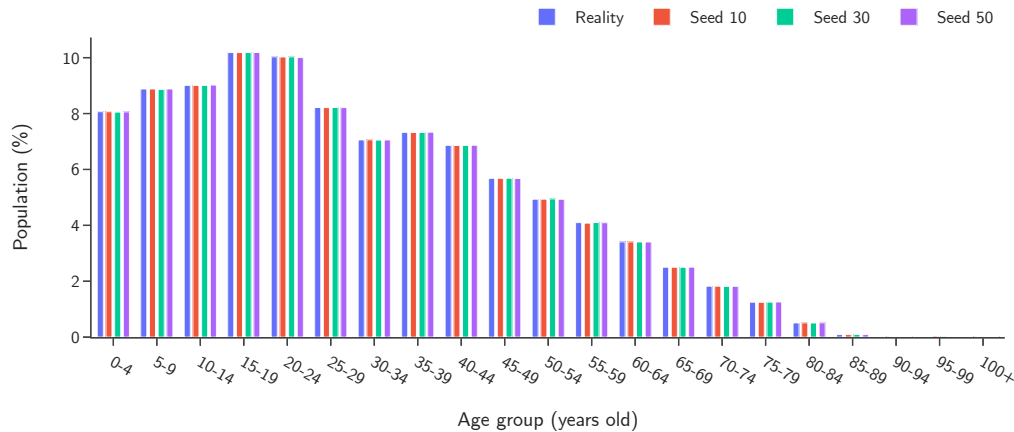


Figure 9 – **Age distribution.** Comparison of real age distribution of Maragogi-AL and for a few randomly generated populations using our algorithm.

The probabilities used in the selection of agents to be added or removed is computed from the difference between the real age distribution and that of the current agent population. We then look at the resulting values for the age groups of the agent candidates.

- If removing an agent: we consider as candidates for removal only the agents whose age groups had negative values in the difference between distributions. We then assign the absolute value of these differences to each agent and normalize them so that they sum to 1. Each value is then used as the probability of removing the corresponding agent.
- If adding an agent: we consider candidates for creation only agents whose age groups had positive values in the difference between distributions. We then filter these positive values and normalize them so that they sum to 1. These are used as the probabilities for selecting an agent of the corresponding age group for creation.

3.1.2.2 Household initialization of compartmental data

In contrast to ODE simulations using compartmental models, which only require the compartment values for initialization, a bio-social agent-based simulation also requires relating compartments with social characteristics in the community. For example, in a community with 250 individuals initialized with 5 infectious ones, having the 5 agents living in the same house or having them living in 5 different houses generates very different results. In the first case, the disease cannot spread more in the same house, while in the second case, it can use the time infected individuals stay at home to spread to others. While a random initialization can still be used to generate a certain tendency in simulations, the high variation in the outcomes demands several realizations in order to reduce standard deviation. Since the preferred environment for spread is always individuals' homes, see (CURMEI *et al.*, 2020), the average results can also be

misleading. The reason is that a random initialization of a few agents will most likely position only one infected agent per home. Another drawback of this approach is that it ties simulation results, and therefore calibrated parameters, to the number of agents in the community, thus making it difficult to export results to other cities of similar but still different attributes.

Ideally, one should be able to relate compartment data to age, social role and household distribution in a time-dependent manner. This level of information would allow a complete disassociation of the probability of infection p , the main parameter calibrated in this work, from the community and its individuals. Unfortunately, it is clear that such data is not available in practice and that collecting it in a meaningful representative way would be nearly impossible. We propose a synthetic half-way solution to this problem, which consists of using data gathered from social behavior and household structure to determine the time-independent probability of a given compartment structure being present at a given home.

The technique we use to synthetically detach the calibrated infection probability p from most population characteristics is to answer the following question: Given a home with n individuals such that it has at least one of its members with an active compartment state, what is the chance that a given compartment configuration is present during the disease life of that home? To answer this question, let us first clarify the meaning of some of these terms:

- Active compartment state: exposed or infectious states.
- Disease life at a home: the period of time in which at least one individual in that home is in an active compartment state.
- Compartment configuration: distributing codes for each compartment, such as S, E, I, R, a configuration is any member of the combinations of n codes out of the 4 possible ones. For example, with $n = 3$, the configuration SEE tells us that the home has one susceptible person and two exposed ones. After some time, the same home can have the following configuration: SEI, which means that one of the exposed persons became infectious.

To determine the probability that a compartment configuration occurs at a home during its disease life, we divide the time a configuration is present at that home by the total time of its disease life. We do that for as many houses as possible, averaging out the probabilities for houses with the same number of people. Figure 10 shows the values we used to initialize active houses with 3 people. From the figure we deduce that if a house with 3 people is active, and all states are possible, then the most likely configuration to occur is SSI, with about 26.6% of chance. The second most likely configuration is SSE, with about 23% chance. The least likely configuration is the one with two exposed individuals and one susceptible one, with about 0.05% of chance, and so on. Notice that a random initialization would most likely have much higher probabilities towards SSE and SSI, with the remaining ones not present in most realizations of the community. It is also important to mention that when the probability of infection changes,

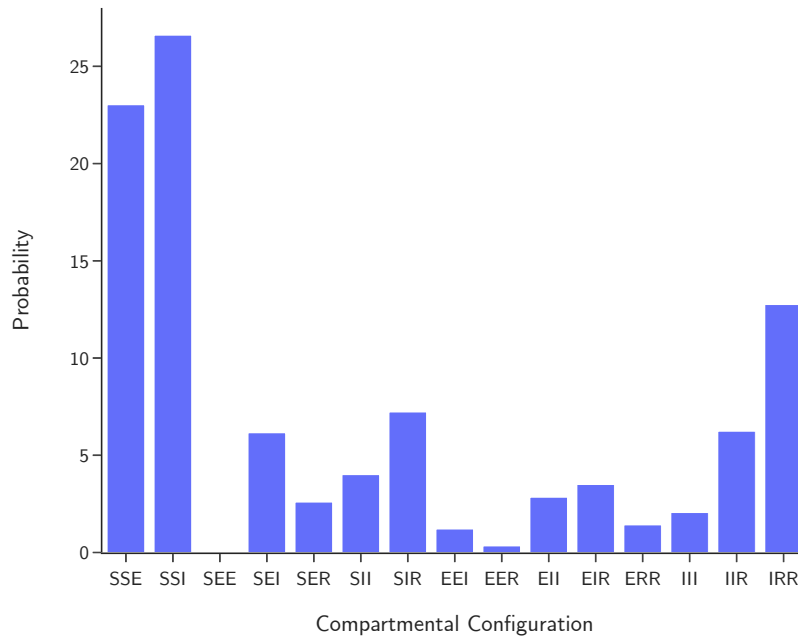


Figure 10 – **Initial household state configurations.** Compartment configuration probabilities used for active homes occupied by tree people

so do the probabilities of compartment configuration. In general, whenever we talk about the probability of infection in this work, we also include the probabilities of inherent compartment configuration that accompany it.

The compartment configuration probabilities should be approximately the same for any point in time, and they also should not depend on the population number, only its household structure and overall social behavior. As a result, we can safely use these probabilities to distribute compartment labels to individuals in a randomly selected active home, as long as there are active compartment labels to distribute. After all active compartment labels are distributed, we randomly select houses to contain a given proportion of susceptible and recovered individuals. This proportion is also estimated from homes that have ended their disease life, and therefore do not contain individuals in the active compartment.

Although the procedure just described allows for a consistent initialization of compartment states along homes in a community, collecting the data necessary to do so is still a hard task. However, if we suppose that the overall social behavior of the community is well captured, and also that the disease modeling inside homes is realistic, then we can conclude that the probability of infection itself determines the probability of each compartment configuration occurring at an active home. This is the argument that allows us to use COMORBUSS to determine the probabilities of compartment configuration. The idea is that, to simulate a community for a given probability of infection p , we first use another simulation to determine the compartment configu-

rations for each possible size of a home. That is, we perform many realizations of a community disease spread using random initialization of states first. Next, we use houses whose disease life is fully captured by the simulations to determine the probabilities for each compartment configuration. Once these probabilities are derived, they are used to initialize compartment states of a second simulation, being that the one that approaches the real life's initial spread the most.

3.1.2.3 Service infrastructure and job allocation

Each service category (e.g. hospitals, supermarkets, schools, etc.) is created as a computer object sharing common defining and operating parameters. Inside each of these objects, we instantiate the same number of these service locations as are known to be had by the modeled city. One of the defining parameters is the average number of workers in the service category and the age groups that are known to function as workers. From this, when the service is created, we randomly select agents in the population of the appropriate age group and assign them as workers for that service. More detailed assignment procedures are, in principle, possible but are unavailable due to lack of required data.

3.1.3 Contacts: Service-specific networks

By collecting the list of visitors, workers, and guests in an instance of a service at any given time, we naturally know the collective history of the community and the sets of agents that can interact. However, how these agents interact is closely associated with the social context at that time. As examples: one does not interact closely between tables in a restaurant while the waiter interacts with the set of tables they are responsible for, as well as coworkers; in a classroom or factory, people are rigidly placed in space for most of the time.

Therefore, we need to consider the social context of the agents in the process of taking the list of agents in a location and producing a contact network. COMORBUSS identifies each particular service having its own network structure, so distinct network models are built when representing restaurants, markets, hospitals, or schools.

All network models share as a common feature the ability to contain in each service tree types of individuals: *workers*, *visitors*, and *guests*. *Workers* of a particular service are individuals who stay in this service for a daily time period during a realization of the stochastic community. In contrast, *visitors* are individuals who visit a single time step that service respecting a frequency of visitation during the simulation. And finally *guests* are individuals that for a temporary duration of time have their default location changed from one's house to that service (e.g. hospitalized or quarantined individuals).

All these types of individuals are specialized for each service to mimic realistic features that one may find in real-world services. For example, waiters in restaurants are modeled as workers who have contacts with visitors. The same idea is applicable to cashiers in markets.

Therefore, these observations must also be taken into account in the modelling of contact networks. Below, we detail the network model for each service.

3.1.3.1 Standard networks: houses and generic services

In houses or generic services, no network configuration can be assumed. As a result, we utilize a standard network model to generate contacts. The contacts in this model are randomly distributed for each agent according to a given average number. This average value may change as the number of agents increases or decreases within the services, as discussed in Section 3.1.3.2. However, the generated contact networks are still dynamic, varying as agents are added or removed.

Contact networks are generated, with few exceptions, using the Erdos-Renyi model, where the probability p_{ER} of an edge being added is given by $p_{ER} = c_{avg}/(N - 1)$. Here, c_{avg} is the average number of contacts, and N is the total number of nodes in the graph. The parameter c_{avg} depends on the definition of contacts, and in this work we assume it to be the following: 'two people two meters or less away from each other for the duration of an hour.'

3.1.3.2 Contact varying with agglomeration

Any contact network needs a fundamental parameter, the average number of contacts (vertices) across the nodes. By default, this input parameter is fixed for each type of network. However, its variation over time may need to be considered in some social contexts due to the high variation of the occupational density of people in that place. For example, in the case of markets, there are rush hours in which the agglomeration is higher. It is also common in this type of service that there are considerably fewer clients at the beginning or end of the work day. To deal with the non-uniformity of the number of agents within each service, we propose a formula to adjust the average number of contacts. The idea comes from supposing that the opportunity for a contact is directly linked to the space available to the agents.

Suppose that two of N agents get in contact with each other whenever they share some specified area A around their position in space. The expression relating $c_{avg}(N)$, A and N is given by

$$E(c_{avg}(N)) = \frac{\frac{N(N-1)}{2}A}{N} = \frac{N-1}{2}A, \quad (3.3)$$

where E is the expectation operator. This formula comes from assuming random walking of N agents within a given service with transit area A . To avoid knowing the transit area, we suppose that a sample N_0 , $c_{avg,0} := c_{avg}(N_0)$ is collectable, and then we approximate A through the formula

$$A \approx \frac{2c_{avg,0}}{N_0 - 1}. \quad (3.4)$$

As a result, the expression for the mean number of contacts $c_{avg}(N)$ varying with the number of agents N is

$$c_{avg}(N) \approx c_{avg,0} \frac{N-1}{N_0-1}. \quad (3.5)$$

Equation (3.5) is used in some of the contact networks introduced in the following. For example, in the case of markets, supermarkets, or street markets, the formula can be used to adjust the average number of contacts between visitors shopping in services. In the case of hospitals, the formula can be used to calculate the average number of contacts among visitors. Another place where such an expression is useful is in the contact network for homes. Assuming equally sized homes, one can infer that the more people at home, the more contacts. Since the number of people at home varies throughout the day, such a formula is well suited to capture the dynamics of movement inside a home.

3.1.3.3 Networks for environment layer

The dynamic in environment layers is very individual-specific and, therefore, we approximate it by random walking. The formula for the average number c_{avg} of contacts between N agents in the environmental layer with transit area A is given by equation (3.3). The transit area A is in this case given by:

$$A = \frac{\pi r^2}{A_u}, \quad (3.6)$$

where r is an *infection radius*, and A_u is the urban area available in the environment layer. The radius of infection is given by half of the largest distance between two agents such that they can be considered in contact. We assume 2 meters as a default value.

Although random, contacts may follow some tendency according to the age of the agents. We used the probabilities exposed in Figure 11, which have been derived from Table 2 of (Del Valle *et al.*, 2007).

3.1.3.4 Network for restaurants

Waiters are restaurant workers who have the greatest potential to become super spreaders of diseases inside their workplace. This happens because they come into contact, as a group, with every visitor who enters the restaurant. As a result, waiters define a special group of workers that must receive special treatment regarding their contact network.

Taking into account the social roles of waiters in restaurants, we model contacts in three categories:

- *visitor-visitor* contacts.
- *waiter-visitor* contacts.
- *worker-worker* contacts.

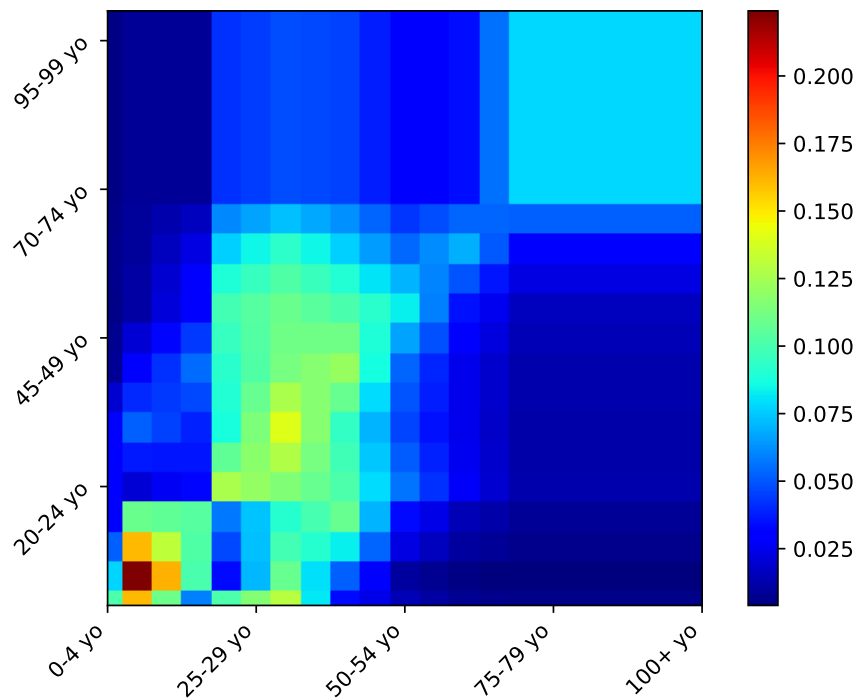


Figure 11 – **Contact probability matrix.** Color map representing probabilities that if a person from an age group in the y-axis met someone, that person belongs to the age groups in the x-axis.

With these categories in mind, the restaurant contact networks are configured by setting the following parameters: the portion of workers who are waiters, the average number of contacts among workers, and the mean number of people sitting around the same table. Because this last parameter is usually difficult to estimate, it can be discarded, in which case the tables are evenly distributed among the waiters in the restaurant.

The contact network for workers is created randomly, always respecting the mean number of contacts provided as input. Among these workers are those composed of waiters, who get in contact with every visitor on the tables they serve. These visitors in turn get in contact with everyone else at the same table.

Figure 12 shows an example of a network for a given restaurant with 5 visitors, 2 waiters, and 3 other workers. Notice that only waiters, identified by ids 1878 and 867, are those who get in contact with visitors. However, it is clear that other workers are in contact with each other. The same thing happens to visitors at the same table.

3.1.3.5 Network for markets

The contact network for markets is similar to that of restaurants in the sense that there exists a class of workers that needs different treatment: cashiers. While other workers usually do not get in contact very frequently with visitors, every visitor is required to make contact with a cashier, either directly or indirectly through shared surfaces, such as shopping belts or credit card machines. Second-order contacts include those among visitors and among workers.

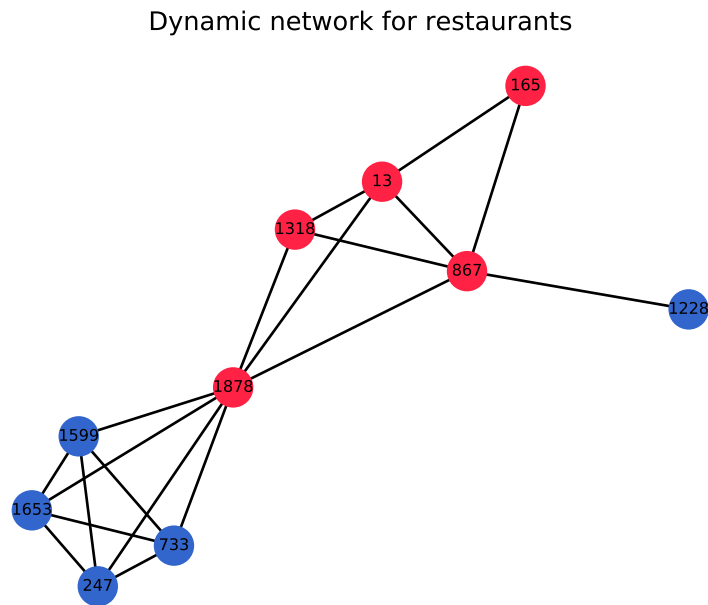


Figure 12 – **Example of a dynamic network for restaurants.** Agents (with their ids in circles) in red are workers, agents in blue are visitors. Numbers inside each circle represent the identification number of that agent.

The contacts between workers and visitors are created randomly, with respect to a given average of contacts provided as input. The contacts between visitors and cashiers are also random, but in this case each visitor is assigned to a cashier. Cashiers are fixed agents that comprise a fixed proportion of all workers in markets.

Figure 13 shows an example of a network for a market. Notice that every visitor (blue agent) gets in contact with at least one worker. Cashiers are workers (red agents) who get in contact with many visitors. Example of cashiers in the figure are those with ids 1479 and 9059. Example of non-cashiers are those with ids 1922 and 2059

3.1.3.6 Network for schools

Schools have two different network models: one for class time and one for break time. During classes, the nature of contacts among students can be very geographical, as students tend to stay seated for long periods of time. During breaks, students are free to walk in public spaces inside the schools. As a result, the distinction between two types of networks is needed.

During class time, we propose a network that connects agents according to nearby neighbors, where students are assumed to follow a geographical disposition of a rectangle. Teachers are treated separately, since they usually move more frequently. The frequency of contact between a student and a teacher may vary according to the age of the student. For

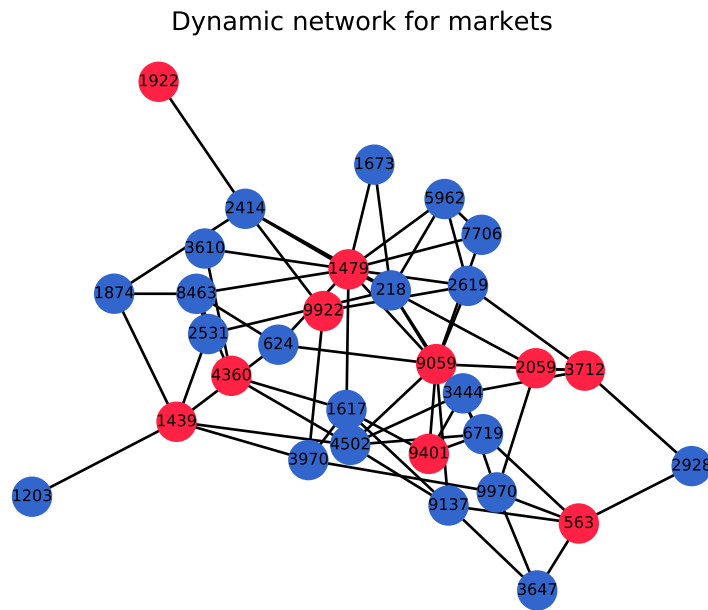


Figure 13 – **Example of a dynamic network for markets.** Agents (circles) in red are workers, agents in blue are visitors. Numbers inside each circle represent the identification number of that agent. A fraction of the workers is designated as cashiers and each client passes through one of them.

example, for students who are toddlers, contact is frequent, but for university students, direct contacts are unusual. The parameters for this type of network are the number of students in a class and the average number of contacts between teachers and students.

During the break time, we propose a simple network in which students get in contact at random. The factor that influences this type of network the most is the number of classes allowed to have a break together, as well as the different ages of the students. The parameters of this network are the number of classes to have a break together and the average number of contacts among students.

Figure 14 shows a network for classes within a school. The teacher is identified by the id 1772.

3.1.3.7 Network for hospitals

Networks for hospitals have, in addition to workers and visitors, admitted persons (hereby labeled guests) who stay in the facility for long periods of time. While these people are admitted, they come into contact with only a few hospital workers. The workers, on the other hand, get in contact with other workers, and some get in contact with visitors as well. Visitors are yet another type of individual which comprises those who seek help in the occasional sickness, as well as those only visiting admitted persons. The need to distinguish between three types

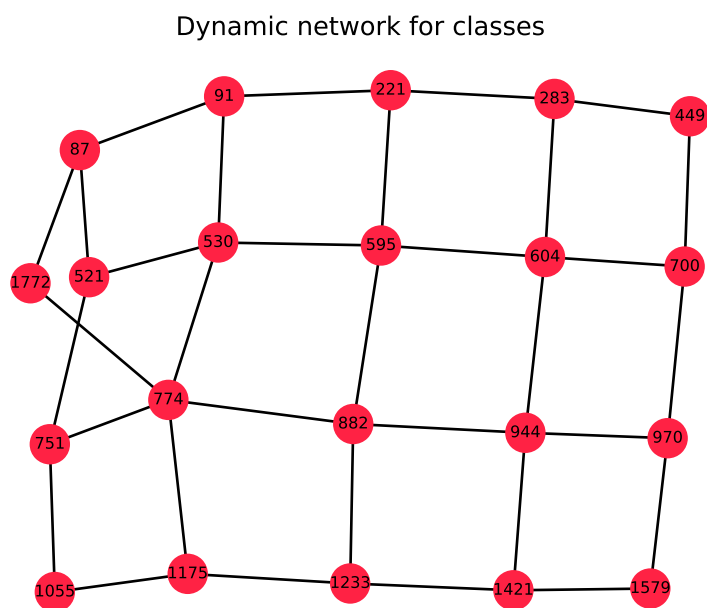


Figure 14 – **Example of a dynamic network for classes in schools.** Students are geographically positioned in lines, with one teacher in charge.

of agent makes this type of network more complex than those introduced previously. Another source of complexity is the fact that some workers are assigned to deal with a specific disease in a pandemic scenario, in an attempt to contain the spread of the disease among workers.

The contact with visitors is adjusted by providing the average number of contacts between themselves, with hospital workers, and with the admitted persons. Contacts among workers take into account the two classes of workers: typical workers and disease workers. The average number of contacts between typical workers, among disease workers, and between typical workers and disease workers must be provided. This last number is typically very small. Finally, the average number of contacts between admitted people and disease workers is a key parameter that can determine the spread of the disease in the hospital.

Figure 15 exemplifies the hospital network. Agents with ids 567, 7372, and 4955, in purple, have been admitted to the hospital. Agents 9943, 7828, 9345 are disease workers, the only workers who get in contact with admitted people. However, they may also get in contact with other workers, in the figure, exemplified by the contact between agent 9345 and agent 435. This last worker may get in contact with another worker, as demonstrated by its connection to the agent 1132. Workers also get in contact with visitors, which can be seen by the connection between agent 435 and agent 8391. Finally, several visitors (in blue) also get in contact with each other, as exemplified by the connection between agent 517 and agent 9300.

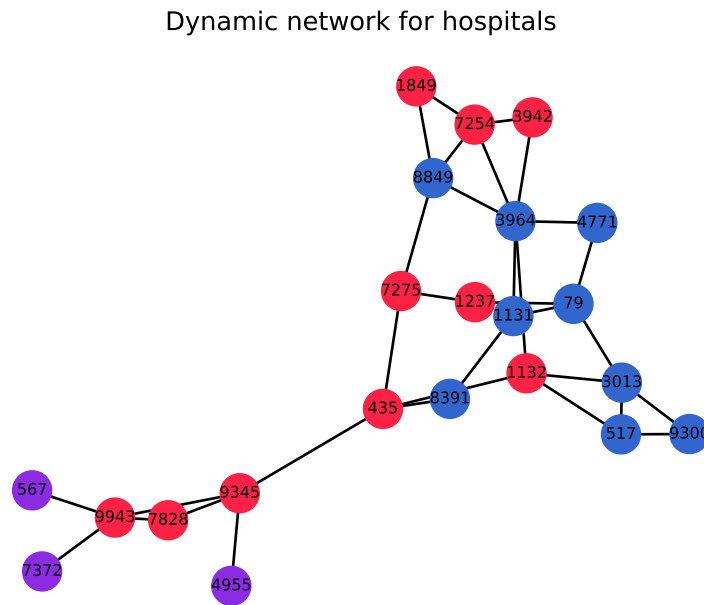


Figure 15 – **Example of a dynamic network for hospitals.** Agents (circles) in red are workers, agents in blue are visitors, agents in purple have been admitted to the hospital and are placed in a COVID-19 dedicated ward. Numbers inside each circle represent the id of that agent.

3.1.4 Community-defining Parameters

The described community model inside a COMORBUSS simulation can be configured with the following parameters:

- `city_name`: Name of the city being simulated;
- `population_ages`: A list for the number of agents in each age group. The age groups are currently separated in intervals of five years from 0 to 100 years, and another age group for 100 years or more. The total number of persons in the city is given from the sum of all of these values;
- `persons_per_home`: Average number of people in a single home;
- `population_graph`: A network containing synthetic or real house structures;
- `services`: A list containing parameters for each service, these parameters are described below.

3.1.4.1 Services Parameters

Every service is defined by the following parameters:

- **name:** Name of the service;
- **number:** Number of instances of this service;
- **days:** Days of the week the service is open to visitation;
- **hours:** Hours of the day the service is open to visitation;
- **visitation_period:** Mean period in days each agent visits this service;
- **age_groups:** List of age groups that visits this service;
- **workers:** Parameters to select workers of different types in this service. Each type can be configured with the following:
 - **name:** Name of the worker type;
 - **number:** Number of this worker type by instance;
 - **shifts:** Shifts available for this type of worker, workers are uniformly distributed between shifts;
 - **age_groups:** List of age groups that can be this type of worker.
- **rooms:** Rooms to distribute workers; each room type is defined with a fixed number for each worker type; rooms of each type are created until there are no more workers of the required type available. At the end all remaining workers are placed in the "public" room that is the same room used by visitors;
- **net_type** and **net_par:** Type of network and its parameters to be used to generate contacts in this service;
- **inf_prob_weight:** Weight applied to the infection probability in this service (used to reduce the infection probability in outdoor services).

3.1.5 *Transportation layer*

A layer for transportation can be optionally activated in COMORBUSS. This layer intercepts all changes in placement during a simulation and places particles in a transport network for a set window of time. In this network, the population is divided into two groups that are randomly assigned at initialization according to the percentage of the population using public transportation services.

- **private transport:** this group is isolated during the time the particle is in the transport layer;

- **public transport:** this group is described by smaller non connected graphs, the size of which is defined by an input distribution with mean corresponding to the average number of users in each vehicle of the public transportation system of the modeled city. The contact networks in each vehicle employ an Erdős-Renyi generator with the mean number of contacts taken as input from the user.

After the desired time in the transportation layer, the particles are placed at their destination. Without the transportation layer enabled, all particle movement is instantaneous.

3.2 Epidemiological Model

3.2.1 Progression: stochastic compartmental model for the disease

At any time, the state of an agent with respect to the modeled disease falls into one of the following compartments:

- (*S*) Susceptible: the susceptible portion of the individuals in the population. This part of the population comprises people who had never had contact with the disease and, therefore, are susceptible to infection.
- (*E*) Exposed: the exposed (or incubating) portion of the individuals in the population. Individuals in this scenario have already had contact with the disease but are still in the incubation stage of the disease. This means that they have been infected but are not infectious.
- (*I*) Infectious: the agent carries the virus and is infectious. The disease itself can manifest itself in different ways, which are subcategorized as
 - (*P_S*) Pre-symptomatic: particles have already become infectious, but they have not yet developed a viral load large enough to show symptoms.
 - (*A_S*) Asymptomatic: this type of particle has passed the activation of the disease, but will never show symptoms. However, they are still infectious.
 - (*S_y*) Symptomatic with mild symptoms: the population in this compartment is those who show mild symptoms.
 - (*S_s*) Symptomatic with severe symptoms: the population in this compartment is those who show severe symptoms.
- (*R*) Recovered: the recovered particles have gone through all the stages of the disease, and that have overcome the disease.
- (*D*) Deceased: the deceased particles have gone through all stages of the disease, developed severe symptoms, and died from it.

When contracting the virus (being exposed), the agents follow the transition diagram shown in Figure 16. The transition between states is stochastic, with transition probabilities being the inverse of the average period in which people remain in that compartment, according to (KERR *et al.*, 2020). The values and references for these periods can be found in the `maragogi_base_conf.py` file². After becoming infectious, an agent remains pre-symptomatic for two days, after which there is the activation event when it is decided whether the disease will manifest as asymptomatic or symptomatic with mild or severe symptoms. Infectious agents recover with a probability estimated from the average duration of the infection; note that the duration of the disease in the case of severe symptoms is longer and such agents can instead convert to the Deceased compartment with a probability dependent on the age group of the agent, see details in Section 4.2.

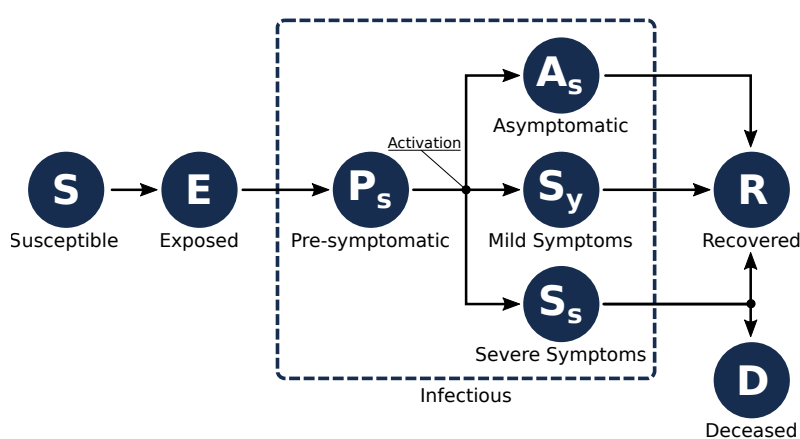


Figure 16 – **Disease progression.** Diagram illustrating how agents can transition between states of the disease.

3.2.2 Transmission: disease spread from contacts

3.2.2.1 Standard: contact through location-contextualized network

The standard transmission model for COMORBUSS is based on contact networks in a location. The first condition for transmission is that a susceptible agent be in contact with an infectious one. Provided such a meeting happens, the susceptible agent converts to the exposed compartment if a random number drawn from a uniform distribution (in unit interval) is less than or equal to the probability of an infection occurring. This probability is the product between the infection probability that is produced by the calibration of the model, the susceptibility of the susceptible agent (which depends on its age group and vaccination status), and a correction parameter which accounts for contacts that do not last the entire time step of an hour.

² <https://gitlab.com/ggoedert/comorbuss/-/blob/paper_school_protocols/schools-paper-scripts/maragogi_base_conf.py>

If this random decision process results in a new infection, the compartment of the previously susceptible agent is rewritten as exposed, and the location, time, and source of the infection are recorded in an infection tree.

3.2.2.2 *Specialized: aerosol transmission model in indoor locations*

In many closed locations where people are present for long periods of time, the main form of infection is not through direct contact with an infectious person, but by inhaling infectious particles that are suspended in the air and accumulate over time. Naturally, modeling this process requires more detailed information on that location since it depends on its volume and rate of air exchange with the outside. These details are not readily available for most services, but for the purposes of this study we acquired the data of the two major schools in the modeled city.

We developed a modified Wells-Riley model which takes into account different parameters for teachers and students. Not only do we consider that these two groups may have different masks, but teachers also release more infectious particles since they speak loudly and continuously.

COMORBUSS naturally tracks all the agents in each classroom and identifies which are infectious. By solving a differential equation for the concentration of infectious particles over time, we compute the balance of absorbed and released particles by students and teachers. We then compute the dose absorbed by each agent in the last time step and from this dose we evaluate the probability that that agent is infected. The modeling details are provided in Section A. Once an infection is produced, we randomly select a source among the infectious individuals in that room and store all the details of this new infection and the usual infection tree.

3.2.3 *Disease-defining Parameters*

The described disease model within a COMORBUSS simulation can be configured with the following parameters:

- `inf_probability`: Probability that an infectious particle will pass the infection in an encounter.
- `susceptibility`: Susceptibility of a particle (defined by age group), the final probability of an infection to occur in an encounter will be given by the `inf_probability` of the source particle multiplied by the `susceptibility` of the susceptible particle.
- `inf_duration`: Mean duration of asymptomatic or mild symptomatic infection (infectious state).
- `inf_severe_duration`: Mean duration of a severe symptomatic infection (infected state).
- `inf_incubation`: Mean duration of the incubation period (exposed state).

- `inf_sympt_timeto`: Time between the transition to the infectious state and the activation of symptoms.
- `inf_prob_sympt`: Probability of an infected particles developing symptoms (defined by age group).
- `inf_severe_sympt_prob`: Probability that an infected particle develops severe symptoms (defined by age group).
- `inf_severe_death_prob`: Probability of an infected particle to die (defined by age group).
- `inf0_perc`: Percentage of particles in each infection compartment at the beginning of the simulation. This is obtained by sampling of the distribution of cases in the initial step inferred in the calibration process.
- `inf0_perc_symp`: Percentage of infected particles in each symptoms compartment at the beginning of the simulation.

3.2.4 Extra symptoms

The extra symptoms module is an easy way to extend the core epidemiological model. With it extra symptoms can be configured with different probabilities (even for each age group); those symptoms can replace, or not, the main symptoms probabilities, if replacing each extra symptoms is given a severity class to determine if the case is considered mild or severe, if an individual has no extra symptom, they are considered asymptomatic.

3.3 Interventions

COMORBUSS implements a few types of interventions; these interventions can be broadly classified into two types: *nonpharmaceutical interventions (NPIs)* are interventions made to simulate public policies like quarantines, lockdowns, closure of services, etc., or changes in behavior such as social isolation, on the other hand *pharmaceutical interventions* are used to simulate vaccines, tests, etc. Most of those interventions can and will use information from other interventions to modulate its parameters, e.g., closure of services depending on the number of positive tests. In this section, we will explore the interventions implemented in our model.

3.3.1 Quarantines

Quarantine policies are a group of NPIs implemented in COMORBUSS intended to isolate individuals inside a simulation based on an individual's parameters. Quarantines in COMORBUSS can:

- Isolate symptomatic patients in their house, where they will only leave the house if their symptomatic state changes to severe;
- Isolate suspected cases in a hotel or other service, where the individuals don't go home or outside during the quarantine;
- Isolate hospital workers from their families, where they go to a hotel or other service after their work shift instead of their house;

Among other types of quarantine. When setting up a quarantine policy in COMORBUSS, the user must specify a filter to select individuals to enter and to exit quarantine, a new default placement (service or home) for the quarantined individuals, and whether the quarantined individuals should work or visit other services. Each quarantine policy can be configured with the following parameters:

- `name`: Name of the quarantine policy;
- `delay`: Delay between the particle is selected to quarantine and the start of the quarantine in days;
- `filter_in`: A sequence of nested tuples of strings and values to be evaluated to select agents to enter quarantine, strings can be population attributes (see particles states attributes), comparative operators, binary operators or markers;
- `filter_out`: Filters to select agents to end quarantine;
- `placement`: Placement marker or name of the service for the place where the agent should quarantine;
- `confine_particles`: If true agent will not work or visit services;
- `allow_requarantine`: If true, allow agents to be quarantined more than once in this quarantine.

3.3.2 Social Isolation

Social isolation in COMORBUSS models a change in behavior of the population. When social isolation is active, a percentage of the population reduces their social activities by reducing or stopping the visitation of certain services and reducing the time they spend outside their homes. Social isolation can be configured with the following parameters:

- `social_isolation`: Enables social isolation mechanics;
- `isol_pct_time_series`: Day-dependent array with the fraction of population that follows social isolation;

- `isol_stay_prob`: Probability p that particles already in social isolation in a day remain in social isolation in the next day. Therefore, the probability that particles not in social isolation follow social isolation is $1 - p$. Notice that even if $p = 0$, the particles in the first age group are still isolated at home. In fact, it may happen that the actual percentage of the population being isolated is larger than the one provided because of this fact. However, if $p < 0$, then no social isolation measure is applied. Quarantined particles are still isolated at the respective quarantine places.

3.3.3 Lockdowns and Services Closures

Lockdowns are NPIs implemented in COMORBUSS intended to restrict the movement of the entire population or a fraction of the population. When setting up a lockdown policy, the user can pass an array with the days inside the simulation where the lockdown must be followed or set limits on population parameters (number of cases, etc.) to dynamically start and end lockdowns. The percentage of the population that follows the lockdowns can also be set. During lockdowns, the visitation to services is reduced or stopped depending on the settings; individuals will not leave the house to go to the environment. Services can also be completely closed depending on population parameters, or in configured days, each service can be configured with its own set of closure policies. Lockdowns and services closures can be configured with the following parameters:

- `lockdown`: A boolean value that informs if the lockdown measure is to be applied to the community or not;
- `lockdown_adhere_percent`: the percentage of the population adhering to the lockdown measure. During lockdown, the percentage of the population effectively being isolated home is given by the maximum value between `lockdown_adhere_percent` keyword's value and `isol_pct_time_series` keyword's value for that day;
- `lockdown_decision_offset`: A time period in days to delay the decision on whether to start or end the lockdown measure.
- `decision`: The decision process to be used on lockdowns and closure of services, see sub-section below (specific decisions can be set individually for each service with the `decision` parameter in the service's parameters dictionary);
- `decision_par_lockdown`: Parameters for the decision process to start and stop lockdowns;
- `decision_par_services_closing`: Global parameters for the decision process to open or close services (specific parameters can be set individually for each service with the `decision_par` parameter in the service's parameters dictionary);

3.3.3.1 Decision process

The decision process used on closure of services and lockdown are protocols already implemented in the software, to use those mechanics the user must select one of the decision process available and also set the parameters for the decision process for lockdowns and services closures. The decision process for closure of services can be set globally, or it can be set individually in the service's parameters dictionary. The available decision processes are as follows:

- **BY_DIAGNOSTICS**: Decisions are taken by the percentage of diagnosed particles (decision parameters: `start_frac` and `stop_frac`);
- **BY_INFECTIOUS**: Decisions are taken by the percentage of infectious particles (decision parameters: `start_frac` and `stop_frac`);
- **BY_SYMPTOMATICS**: Decisions are taken by the percentage of symptomatic particles (decision parameters: `start_frac` and `stop_frac`);
- **FIXED_PERIOD**: Decisions are taken by a fixed period (decision parameters: `start_day` and `stop_day`).

3.3.4 Contact tracing

Contacts between particles can be traced inside COMORBUSS, by default this data is not stored due to the computational cost, but it can be optionally stored. However, the power of contact tracing is where it is setup to work with other interventions, for example, a quarantine policy can be configured to use tracing data to isolate individuals who had contact with confirmed cases. A tracing percentage can also be configured to better represent imperfect tracing, with the following parameter:

- `tracing_percent`: Fraction of the population that has tracing capability.

3.3.5 Testing

Our model also has a robust testing/diagnostics module, allowing multiple testing policies to be configured in each simulation. For each testing policy, the user can configure who takes the tests with an individuals filter, how many tests are applied per day, the sensitivity and specificity of the test, the time between the test is applied and the result, the period in the simulation where the policy is active, and if individuals can be retested. After an individual is tested, they are marked as positive or negative, and this information can then be used to guide other interventions such as when to start/stop lockdowns, close services, etc. The information of whether an individual's test was true or not is also stored in the simulation and can be used in the simulation analysis. The parameters for each testing policy are:

- `name`: Name of the testing policy;
- `start_day` and `end_day`: Dates to start and stop the testing policy inside the simulation;
- `filter_particles`: A sequence of nested tuples of strings and values to be evaluated to select agents to apply testing, strings can be population attributes (see particles states attributes), comparative operators, binary operators or makers;
- `number_per_day`: Number of tests to be applied in the selected population, particles selected with `filter_particles` will be selected randomly to test if there are not enough tests. If set to -1 will test all particles selected with `filter_particles`;
- `sensitify`: Sensitivity value for the test in the $[0., 1.]$ interval;
- `specificity`: Specificity value for the test in the $[0., 1.]$ interval;
- `test_delay`: Time in days between an agent is selected to be tested and the test is made;
- `result_delay`: Time in days between an agent is tested and the result is available;
- `allow_retest`: Allow particles to retest the test ("yes", "no" or "negative");
- `retest_delay`: Time since the last result for a particle to be allowed retest.

3.3.6 Vaccination

Finally, COMORBUSS has a highly flexible vaccination module. This module allows different vaccine models and different vaccination campaigns to be integrated into simulations. Each vaccination campaign can be configured with:

- `name`: Name of the vaccination policy;
- `start_day` and `stop_day`: Dates to start and stop the vaccination inside the simulation;
- `filter`: Filters to select individuals to receive vaccine (can be passed as logical expression or function);
- `filter_parameters`: Parameters of the filter functions (if given as functions);
- `priority_function`: Function to prioritize individuals to receive vaccines;
- `effectiveness`: Dictionary of biological effectiveness for the vaccine, the effectiveness can affect any individual parameter such as susceptibility, infectiousness, probability of symptoms or death, etc. Each effectiveness can be passed as a single value, an array with an immunological curve, or as a function that generates the immunological curve depending on the individual parameters.

- `effectiveness_parameters`: Parameters of the effectiveness functions (if given as functions);
- `adverse_effects`: Names and probabilities of adverse effects (can be configured per age group);
- `doses_per_day`: Number of doses to be applied per day during the campaign;
- `keep_last_value`: Determines if the last value on the immunological curve should be kept or the individual should return to the initial state;
- `only_not_vaccinated`: Determines if the vaccination should be applied in already vaccinated individuals;
- `vaccinate_at_start`: Determines whether vaccination should be applied at the start of the simulation, for example, to simulate a previously vaccinated population;

This set of parameters allows COMORBUSS to model from simple 0-1 vaccines where the individual is granted full or no immunity, to leaky vaccines or more sophisticated vaccine models with dynamic immunological curves or effectiveness to other individual parameters such as probability of symptoms. Each vaccine also has configurable adverse effects, this can be used, for example, with the extra symptoms module to evaluate cost-benefit scenarios of vaccinations.

3.4 Code availability

3.4.1 *Distribution and Documentation*

COMORBUSS has a project webpage under the link <https://comorbuss.org>, where all the developments, results, and links are assembled.

The source code for COMORBUSS is available in the repository <https://gitlab.com/ggoedert/comorbuss> under license [AGPLv3](#). The version of the code together with all required input files and simulation scripts is available under the tag `Paper_Maragogi_Schools`.

The complete documentation of the COMORBUSS library is available on <https://docs.comorbuss.org/> under license [CC BY-SA 4.0](#).

3.4.2 *Dependencies*

Here we specify all the versions of the computer libraries used for the present work. COMORBUSS is written in Python (version 3.7.7) and requires the following modules:

- `numpy v.1.18`

-
- matplotlib v.3.1.3
 - pandas v.1.0.5
 - seaborn v.0.10.1
 - h5py v.3.10
 - h5dict v.0.2.2
 - scipy v.1.5.0
 - portion v.2.0.2
 - networkx v.2.5.1
 - tqdm v.4.46.0
 - numba v.0.53.1

DATA INTEGRATION

In the design of COMORBUSS flexibility was a paramount feature. However, this flexibility comes at the cost of requiring a substantial number of parameters to accurately represent the myriad of dynamics in community interactions and epidemic progression. This chapter focuses on the crucial aspects of data collection and processing which are indispensable for the operation of the COMORBUSS model. We will explore the various sources from which data was gathered, encompassing a wide range of demographic, socioeconomic, and behavioral patterns. In addition, the chapter will detail the methodologies employed in processing these data, ensuring their compatibility with our model.

4.1 Modeling household networks

We used three databases to reconstruct a social network of household contacts: *Programa Saúde da Família (PSF)*, *Programa Bolsa Família (BOLSA)* and *Sistema de Monitoramento da COVID-19 (SMC)*. These city-owned databases correspond, respectively, to a public health assistance program, a social assistance program, and a software to register covid-19 health attendance. The data from the first two databases was previously collected from non-structured sources such as PDF files and processed. By combining data from the other two sources, we managed to capture the distribution of household family size for at least $2/3$ of the city population.

Each of the following tables were constructed containing one column that specifies for each person (row in the table) which household group it belongs to, hence by grouping rows in the table by this *group-column* value we can obtain the network structure.

The *Programa da Saúde da Família* is the largest database containing 26721 rows, but it is the poorest in detail with only 4 columns, namely:

- *nome*: Token representing the name of each person;

- *cns*: Token representing the health program id, (*cadastro nacional de saúde - cns*) of each person (will be used later for merging the tables);
- *idade*: The age of each person;
- *codigo_familiar_psf*: A token representing the group (family code) of each person;

Since the data was available in PDF format, we made a JavaScript script to download files from each family, extracting the text using the Python library called *pdfplumber*. We noticed that the fields were well defined between some specific sets of words, so we used string matching techniques to filter the information from each field and structure the text.

The *Programa Bolsa Família* database is the second largest database, containing 18682 rows and 11 columns concerning rich data about the beneficiary of the social program, with 99% of the data collected in 2016 or later. For this reason, we chose this table to be the fundamental source of data for the construction of the network, as will be detailed in Section II. The columns contain data related to:

- *Personal information*: Tokens representing the name of each person, the name of their parents, *cns* id and *cpf* id which stands for *cadastro de pessoa física*, an individual id used in Brazil that can uniquely represent each person throughout the databases. The age of each person.
- *Work information*: Various columns detailing work information.
- *Family information*: A token representing the group (family code) of each person in the database, as well as a field (column) describing the family-role of the beneficiary. The address of the house where each person lives (is the same for each person in the same group).

The *Sistema de Monitoramento COVID-19* database is the one containing detailed information about the health status of people in the city, concerning the actual pandemic. It is by far the smallest database with only 1602 rows and 14 columns with data related to:

- *Personal information*: Tokens representing the name of each person, the name of their mother, *CNS* id and *CPF*. The age of each person.
- *Medical information*: Various columns detailing medical information about the evolution of the patient's status throughout the year (date of first symptoms, date that the patient tested positive or recovered), etc.
- *Family information*: A token representing the group (family code) of each person in the database. The address (neighborhood) of the house where each person lives.

The idea of integration of the three databases assumes the following propositions:

- BOLSA database offers us a reasonable idea of the distribution of families across the population.
- Medical information from all patients should be used, whenever possible.

An important fact to note is that the three databases do not share a common key, e.g., *CPF* or *CNS* in order to merge the data without repetition. Hence, we propose the following approach to merge the databases (let B , S , P denote BOLSA, SMC and PSF databases respectively):

1. Let $I_1 = B \cap S$. The intersection is found using the key *CPF*.
2. If $C = S - I_1$, let $I_2 = C \cap P$. The intersection is found using the key *CNS*.
3. The merged database M is obtained by the disjoint union of the following tables: $I_1 \cup I_2 \cup (B - I_1) \cup (S - I_2 \cup I_1)$.

We finish this process with a final table M with 19973 rows and several columns, containing roughly $2/3$ of the city's population, regarding household contact, economic and health data of its citizens. This table is used to assign each person to a group, based on the *grouping column* (family code) of each source database.

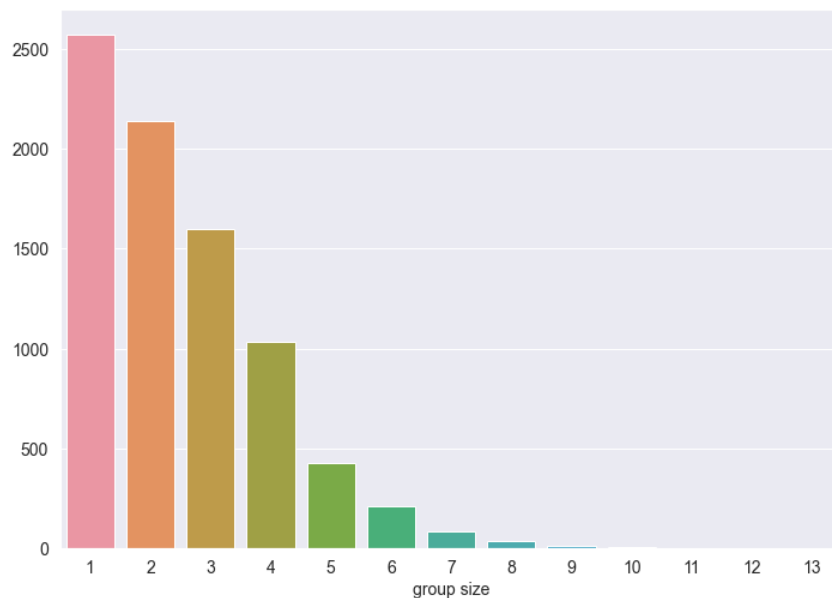


Figure 17 – Group size distribution of the M database.

We further expand the database M by incorporating persons registered only in the table P that have a relative (the same person in the group) in the table I_2 , as described above, resulting in a final table with 20350 rows.

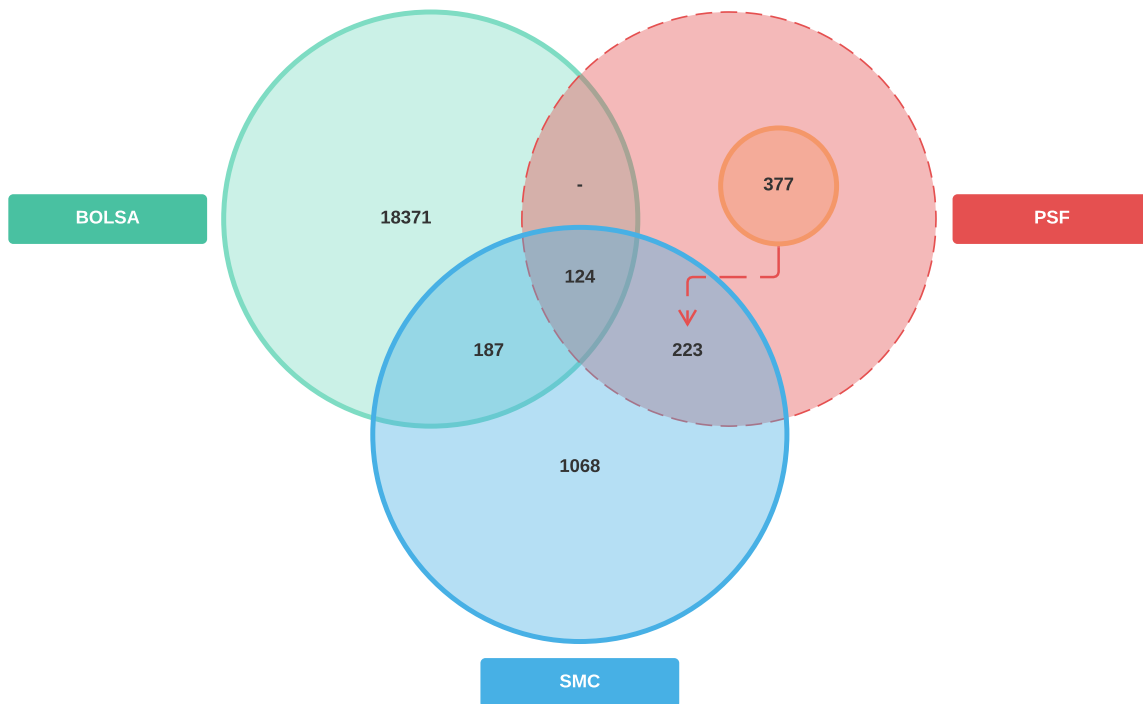


Figure 18 – **Database diagram.** PSF - Programa Saúde da Família, BOLSA-Programa Bolsa Família and SMC-Sistema de Monitoramento Clínico.

The merged database M is used to construct the network of contacts G through the following steps:

1. Create, for each row in M a node in G with its respective attributes (columns of M).
2. Group each node in G by its *grouping column* (family code). If a node has more than 1 valid family code we choose to group it by the following priority: firstly family code from B , secondly from P , and thirdly from S .
3. If vertices u and v are in the same group as constructed in the previous step, create an edge (u, v) in the network. That process ultimately yields a network composed of only fully connected components (cliques), which represents the household contacts.
4. We try to further connect cliques in the network by checking parenthood relationships based on the mother / father names in the B section of the database (those nodes who have a valid B family code attribute). We first select nodes that have unique names and then map each name to its node label. For each node u in the network, we find nodes f and m that have the name attribute equal to the father name and the mother name attribute of node u , respectively. If f and m both have the same family code, we create edges (u, m) and (u, f) . Notice that it is possible to find only node f or only node m (exclusively). In such a case, we just create the edge between the child and parent node.

The caution taken in the last step of checking the uniqueness of the names and if nodes m and f belong to the same family is due to the fact that common names might pose a problem on creating edges in such a way, as they would form clusters that are little related to the real parenthood relationships.

The edges created by step 3 are labeled *INTRAFAMILIAR edges* (relating to contacts within families), while edges created by step 4 are labeled *INTERFAMILIAR edges* (relating to contacts between distinct but related families).

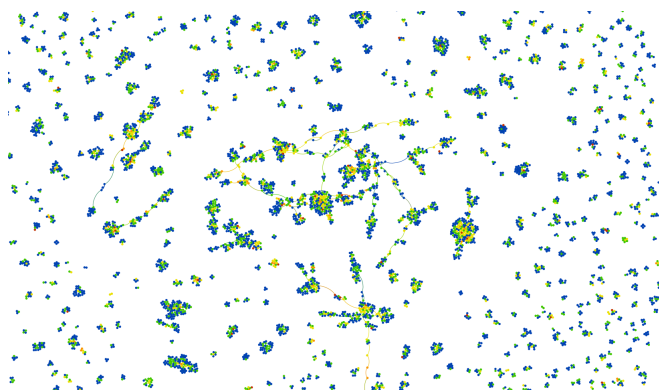


Figure 19 – **Local picture of the network.** One can see the cliques (dense clusters) interconnected by edges. Nodes are colored by age.

The resulting network has 27235 edges, 24596 being *INTRAFAMILIAR*. The average degree of the network considering only such edges is about 2.4 as we can see in the histogram below:

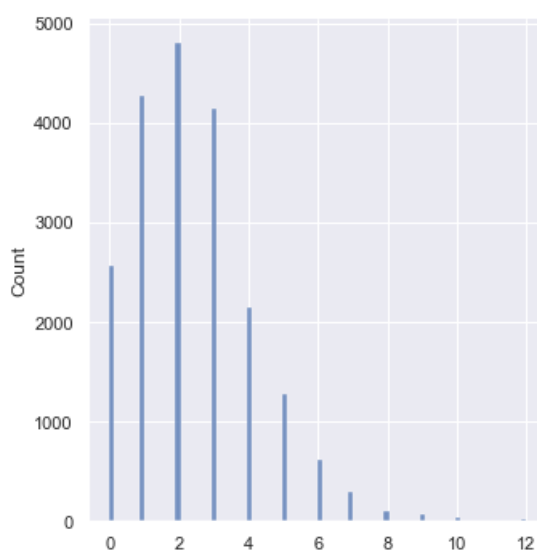


Figure 20 – Node degree distribution, considering only edges within cliques.

In the end to ensure complete anonymity of the data, we ran the algorithm described in

Section 3.1.2.1 to generate a new synthetic database to be used and distributed with COMOR-BUSS, while still preserving the same macro characteristics as the real data.

4.2 SARS-COV2 DATA PROCESSING

Our reconstruction is based on data collected in Maragogi-AL using the *Sistema de Monitoramento COVID-19*. From the anonymous database, we accessed the attendances of each tested patient, in case of hospitalization, the hospitalization date, and in case of death, the death date. Each attendance entry is composed of attendance date, symptom onset date, test type (rapid or RT-PCR) and test result (positive or negative), see Listing 1.

```

1: "5": { //anonymized patient id
2:     "attendances": {
3:         "1174": { //attendance unique id
4:             "result": "negative",
5:             "test_type": "rapid",
6:             "attendance_date": "2020-05-11",
7:             "symptom_onset_date": "2020-05-07"
8:         },
9:         "1375": {
10:            "result": "positive",
11:            "test_type": "TR-PCR",
12:            "attendance_date": "2020-06-17",
13:            "symptom_onset_date": null //unfilled attendance
14:            date
15:        }
16:    },
17:    "hospitalization_date": "2020-06-18",
18:    "death_date": null,
19: }

```

Source code 1 – Patient data example. This patient had two appointments, the first with a negative result and the last, one month later, with a positive result. The patient was hospitalized one day after the second appointment, but did not die.

Most quantities required for the reconstruction, such as number of hospitalizations, deaths, and attendances, evolve over time. We chose to reconstruct the curve until July 25.

To account for false negatives and false positives, we also needed information about the specificity and sensitivity of the tests. In general, only 52 of the 1722 tests performed until July 25 were RT-PCR tests.

Different brands of rapid tests were used throughout the year. The utilization dates in Table 4 were informed by Maragogi's health professionals, and the accuracies were taken from

(ANVISA, 2020). The RT-PCR test was assumed to have 100% specificity and sensitivity.

Test brand	Utilization	Specificity	Sensitivity
Wondfo One-Step COVID-2019 Test	Apr 11 - Jun 25	99,57%	86,43%
MedTeste MedTeste Coronavirus (COVID-19) IgG/IgM	May 01 - Jun 22	99,3%	97,4%
Advagen COVID-19 IgG/IgM LF	Jun 23 - Aug 31	96%	85%
Lungene COVID-19 IgG/IgM Rapid Test. Cassette	Sep 01 - Oct 28	96,48%	91,06%

Table 4 – Usage and accuracy of rapid tests.

Using data from Table 4 and assuming that when more than one rapid test is available, they are used equally, we obtain the overall daily specificity and sensitivity of the rapid tests (see Figure 21). From Table 2 in Section 1.3, we use the expected probabilities resulting from the hospitalized / infected ratio $p_h = 3.304\%$ and the death/infected ratio of $p_d = 0.441\%$ in general.

Finally, the distributions of infection times were given by (KERR *et al.*, 2020), namely:

- Incubation period (length of time between exposition and viral shedding): log-normal with mean 4.6 days and deviation 4.8;
- Symptom onset period (length of time after viral shedding has begun and before an individual has symptoms, when one has symptoms): log-normal with mean 1 day and deviation 1;
- Recovery period (length of time after incubation while the individual is infectious): log-normal with mean 8 days and deviation 2 for non hospitalized patients or with mean 14 days and deviation 2.4 for hospitalized patients.

4.2.1 The reconstruction algorithm

Reconstruction of susceptible, exposed, infectious, and recovered curves was performed by taking the mean over 400 curves generated stochastically. Each generated curve is also saved for later use in calibration.

To build these curves, we need to know, for each infected person, when one enters and leaves each compartment. For example, we know the date of attendance of the patient in Listing 1 for both attendances. We also know the hospitalization date and the symptom onset date for the first attendance. But the date on which the patient was exposed to the virus, when it became infectious, or recovered is unknown. This missing information will be reconstructed using previously known distributions, as listed in the last section, or resampling from the data.

The reconstruction has three main steps: test data correction, individual timeline reconstruction, and case estimate, each will be described on the next sections.

4.2.2 Test data correction

The test data correction step relies on two minor steps: sampling incomplete test type and inference of true positives (TP) and true negatives (TN).

4.2.2.1 Sampling incomplete test type

First of all, we treated incomplete data. For each incomplete test type field (104 out of 1722) with date t , we sampled its type (either rapid or RT-PCR) using all tests with known test type from the same date t .

4.2.2.2 Inference of true positives and true negatives

The next step is to arbitrate whether the test results are correct or not (for rapid tests, since RT-PCR tests are always assumed to be correct).

Let TP be the percentage of true positives, TN of true negatives, FP of false positives and FN of false negatives. By definition, specificity (e) and sensitivity (s) are given by

$$e = \frac{TN}{TN + FP} \text{ and } s = \frac{TP}{TP + FN}, \quad (4.1)$$

but we want to evaluate the probability of true positives (p_{TP}) and true negatives (p_{TN}), i.e.,

$$p_{TP} = \frac{TP}{TP + FP} \text{ and } p_{TN} = \frac{TN}{TN + FN}. \quad (4.2)$$

We aim to write both equations above in terms of known quantities: the specificity and sensitivity are known from the technical notes (ANVISA, 2020), and $p = TP + FP$ comes from the total number of positive tests throughout the period. From Equation (4.1) we have

$$TN \left(1 - \frac{1}{e}\right) + FP = 0 \text{ and } TP \left(1 - \frac{1}{s}\right) + FN = 0 \quad (4.3)$$

and from $TP + FP + TN + FN = 1$,

$$TP + FP = p \text{ and } TN + FN = 1 - p. \quad (4.4)$$

Thus,

$$TP - TN \left(1 - \frac{1}{e}\right) = p \text{ and } TN - TP \left(1 - \frac{1}{s}\right) = 1 - p. \quad (4.5)$$

Solving (4.5), we have

$$TP = \frac{p + (1 - p) \left(1 - \frac{1}{e}\right)}{1 - \left(1 - \frac{1}{s}\right) \left(1 - \frac{1}{e}\right)} \text{ and } TN = \frac{(1 - p) + p \left(1 - \frac{1}{s}\right)}{1 - \left(1 - \frac{1}{s}\right) \left(1 - \frac{1}{e}\right)}. \quad (4.6)$$

Therefore,

$$p_{TP} = \frac{1 + \frac{1-p}{p} \left(1 - \frac{1}{e}\right)}{1 - \left(1 - \frac{1}{s}\right) \left(1 - \frac{1}{e}\right)} \text{ and } p_{TN} = \frac{1 + \frac{p}{1-p} \left(1 - \frac{1}{s}\right)}{1 - \left(1 - \frac{1}{s}\right) \left(1 - \frac{1}{e}\right)}. \quad (4.7)$$

These quantities evolve with time, as the proportion of positive tests varies over time. So, for a given day t we let $p(t)$ be the ratio p calculated using a 21-day window centered on t (which matches the disease cycle used in the calibration). Also, let $e(t)$ and $s(t)$ be the mean specificity and sensitivity of the rapid tests available on day t .

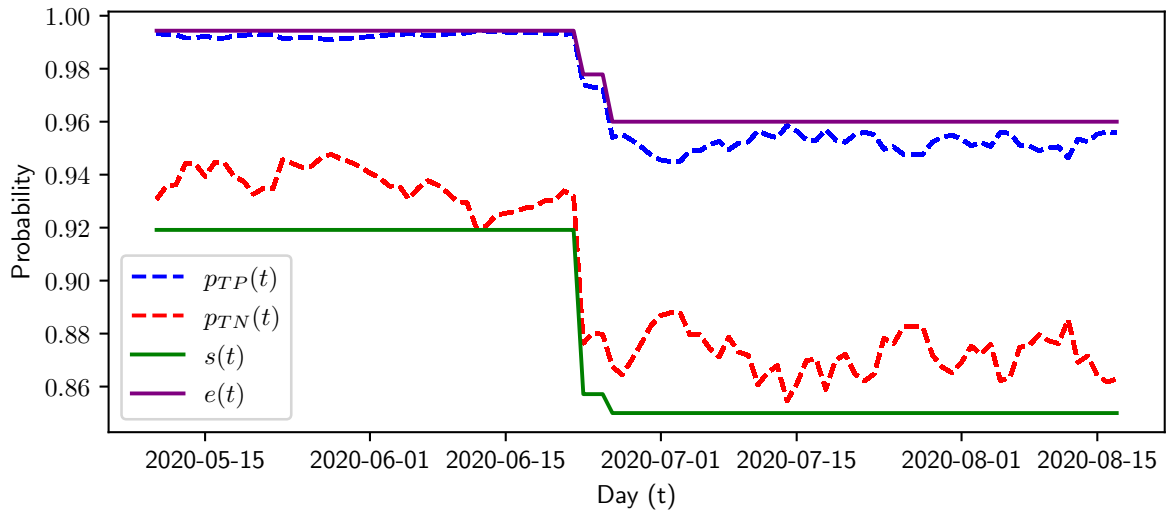


Figure 21 – **Daily specificity $e(t)$, sensitivity $s(t)$ and constructed probabilities $p_{TP}(t)$ and $p_{TN}(t)$ in Equation (4.7).** Note the dashed curves also rely on the sampling incomplete test types, so it changes over each realization of the reconstructed curve. Since the standard deviations are minimal, we chose to plot only the mean curve. Until April 28, only negative results were reported by rapid tests and the moving average has a window of 21 days.

Using the curves p_{TP} and p_{TN} , one can determine whether a given rapid test was positive or negative. We run that decision stochastically for each attendance with a rapid test. From now on, when we refer to positive tests, we are talking about the tests we judged as positive.

4.2.3 Individual timeline reconstruction

Each individual has one or more attendances. From the first attendance with a positive test result, if it exists, we took the date of onset of the symptom and the date of attendance.

Let i be an agent, τ_E^i its exposition date, τ_I^i the day it becomes infectious, τ_{sym}^i the symptom onset day and τ_R^i the recovery (or death) date, the individual disease timeline is the tuple $(\tau_E^i, \tau_I^i, \tau_{sym}^i, \tau_R^i)$. The value τ_{sym}^i is the only one we know and any other value can be

stochastically constructed using the distributions in (KERR *et al.*, 2020), namely:

$$\begin{aligned}\tau_I^i - \tau_E^i &\sim \text{lognormal}(4.6, 4.8) \\ \tau_{sym}^i - \tau_I^i &\sim \text{lognormal}(1, 1)\end{aligned}$$

and for non hospitalized patients

$$\tau_R^i - \tau_I^i \sim \text{lognormal}(8, 2)$$

or hospitalized patients

$$\tau_R^i - \tau_I^i \sim \text{lognormal}(14, 2.4).$$

Some attendances do not have information on the date of onset of symptoms (around 23% of positive cases). Again, from the filled data we derived the distribution of the time between symptom onset and medical attendance, only over positive cases, and then sampled the onset date of unfilled entries.

4.2.4 Number of Cases estimation

The ratios

$$\frac{\text{number of hospitalizations}}{\text{number of cases}} \text{ and } \frac{\text{number of deaths}}{\text{number of cases}} \quad (4.8)$$

should approximate the inferred ratios $p_h = 0.03304$ and $p_d = 0.00441$, respectively. Let $NB(q, n)$ be the negative binomial distribution with success probability q , which counts the number of Bernoulli failures that should occur until n success. In the period until July 25, a total of 18 individuals died and 119 were hospitalized. So, we can model the number of cases as

$$T_h = NB(p_h, 119) + 119 \text{ or as } T_d = NB(p_d, 18) + 18. \quad (4.9)$$

Using the number of hospitalizations, we have $\mathbb{E}(T_h) = \frac{119}{p_h} \approx 3601$ with a 90% confidence interval of [2966, 4033]. Using the number of deaths, we have $\mathbb{E}(T_d) = \frac{18}{p_d} \approx 4086$ with a 90% confidence interval of [2623, 5760]. Both confidence intervals agree, although the confidence interval estimated using deaths is larger. Since it has a narrower confidence interval, we use $T = T_h$ to estimate the total number of cases.

It is also interesting to note that the data appear consistent, the ratio between recorded deaths and recorded hospitalizations is $\frac{18}{119} \approx 15.1\%$ and the ratio $\frac{p_d}{p_h}$ is approximately 13.3%, a small difference.

4.2.5 The final curve

Let H be the set of all hospitalized individuals and N the set of all non-hospitalized infected individuals, define

$$E_H(t) = \sum_{i \in H} \mathbf{1}_{[\tau_E^i, \tau_I^i)}(t), I_H(t) = \sum_{i \in H} \mathbf{1}_{[\tau_I^i, \tau_R^i)}(t), R_H(t) = \sum_{i \in H} \mathbf{1}_{[\tau_R^i, \infty)}(t). \quad (4.10)$$

Let $E_N(t)$, $I_N(t)$, and $R_N(t)$ be defined analogously. Also, let

$$C_H = \sum_{i \in H} \mathbf{1}_{[\tau_A^i, \infty)}(t) \text{ and } C_N = \sum_{i \in N} \mathbf{1}_{[\tau_A^i, \infty)}(t), \quad (4.11)$$

where τ_A^i is the first attendance date with a positive result.

Assuming no subnotification among hospitalizations and deaths. Also, using T cases on July 25th, we define

$$\alpha = \frac{T - C_H(t^*)}{C_N(t^*)}, \quad (4.12)$$

where t^* is July 25th, $C_H(t^*) = 119$ and $C_N(t^*)$ varies depending on the missing data reconstruction, the inference of test results and the individual timeline reconstruction. Then, the quantity α captures the ratio between the overall mild cases and the mild cases followed. On average, only 16.75% of patients with mild or no symptoms sought medical help.

Finally, we reconstruct the curves:

$$E(t) = E_H(t) + \alpha E_N(t), I(t) = I_H(t) + \alpha I_N(t) \text{ and } R(t) = R_H(t) + \alpha R_N(t). \quad (4.13)$$

The procedure has four stochastic steps: test type resampling, test result correction, symptom onset date resampling, and individual timeline reconstruction. The final curve is given by the mean of 400 trials; see Figure 22. Of course, the curve of susceptible individuals ($S(t)$) is given by the total population minus the sum $E(t) + I(t) + R(t)$.

4.3 Parameter estimation

In this section, we describe how to use data collected from Maragogi-AL city to estimate some main parameters of the model, such as those in the definition of services. We also specify the calibration procedure used to approximate the poorly estimated or unknown parameters whose variance influences the SEIR curves the most.

In the sections to come, we first focus on the estimation of some key parameters in the definition of services. Section 4.3.1 gives reasoning to the choice of the average number of contacts within homes, and also to the relation between the probability values of indoor and outdoor infection. Sections 4.3.2 and 4.3.3 are intended to explain the estimation of the most relevant service parameters with respect to disease transmission: the visitation period and the network parameters. Finally, the next two sections detail the calibration process and the sensitivity of the results with respect to the population size, respectively.

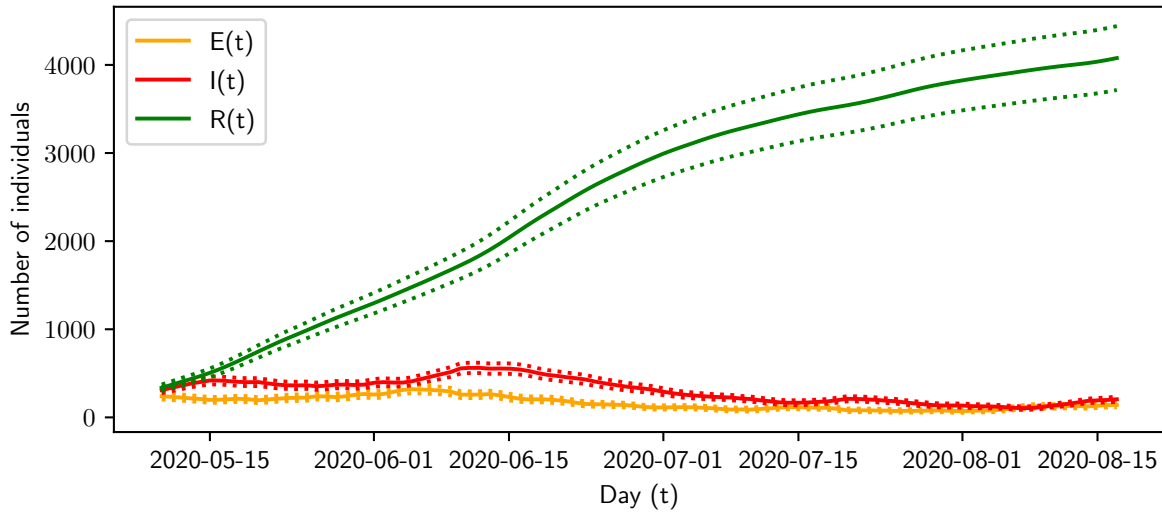


Figure 22 – **Final estimated curve for exposed, infectious and recovered compartments.** The solid lines are the mean over all 400 trials and the dashed ones represent one standard deviation up and below.

4.3.1 Estimation of parameters for household and indoor/outdoor environments

Regarding the specific epidemic history of Maragogi-AL city, one notices that a large portion of the population stayed home during the period we consider in this study. This fact is confirmed from the *Inloco* geolocation data (currently under the name of *Incognia* (INCOGNIA... , 2021))¹. From the high level of social isolation in this period, we assume that the transmission rate in homes was higher compared to the rate in other environments, such as essential services. The transmission rate of the household, denoted R_h , is introduced in Curmei *et al.* (2020), and is defined by the average number of new infections caused by an infected individual within its household. Given the intense social isolation in Maragogi-AL, we have used the highest value of R_h estimated by Curmei *et al.* (2020) as our reference. This leads to choosing the average number of 1 hour contacts c_{homes} inside a home so that the total number of new infections in houses during the period considered is about 70% of the total. In our simulations, we have used $c_{homes} = 0.7$.

Transmission rates also vary considerably for indoor and outdoor environments². From the meta-analysis of Nishiura *et al.* (2020), it is inferred that indoor environments increase 18.7 times the probability of disease contagion compared to outdoor environments. We consider this aspect in our simulations, multiplying the infection probability within outdoor environments by

¹ The company uses high resolution smartphone geolocation data to generate the social isolation index time series, see (INCOGNIA... , 2021). However, we must point out that due to geographic limitations, the regional cellphone signal is not captured with high quality, causing an underestimation of the social isolation index

² Outdoor environments in Maragogi-AL, for the sake of our study, include the environmental layer (see Section 3.1.3.3) and street markets

a weight equal to $1/20$.

4.3.2 Estimation of service's visitation period

The services visitation period is one of the parameters that most influences the spread of the disease in a given city, since it controls the influx of agents within services. In Section 3.1.1.2 we define the visitation period and also how to estimate it. In this section, we describe how we collected the data for the actual values of this parameter for each service in the city of Maragogi-AL.

Street market: street market opens at Saturdays, from 6:00 to 12:00. During the pandemics, an average of 3000 people visited the street fair every day it opened; see Section 1.3. Assuming that all people in age groups 5 and above can visit the fair, the calculation for v_{pc} is

$$v_{pc} = 7 \frac{v_t}{v_w} = 7 \frac{20884}{3000} \approx 48.73 \text{ days.}$$

Hospitals: the total number of hospital visits in each hospital unit (UPA and SAMU) from April 29 to June 28, 2020 was 3579 and 304, respectively. To estimate the actual number of people who visited the hospitals during this period of time, we must take into account the other people accompanying these attendees. In order to do that, let us suppose that at least children from the first 3 age groups are accompanied by an adult, and that the same happens to elderly from the 13th age group and above.

Assuming that everyone in Maragogi had the same number of contacts with the disease, we estimate from susceptibility p_{sus} and the probability of developing severe symptoms p_{sev} , what the portion po of attendees who brought another person with them to the hospital:

$$po = \frac{\langle p_{sus} * p_{sev}, pop_{ce} \rangle}{\langle p_{sus} * p_{sev}, pop_t \rangle} \approx 0.2496,$$

where $*$ is the point-wise multiplication of vectors, $\langle \cdot, \cdot \rangle$ is the inner product, pop_t is the vector of all people from all age groups, and pop_{ce} agrees with pop_t for children or elderly, but has null entries otherwise.

The UPA attendees do not all come from Maragogi, but the hospital estimates that at least half of them do. Taking all this information into account, we estimate the visitation to the hospitals to be

$$v_{pc} = (180 - 120) \frac{32702}{(0.5 * 3579 + 304) * 1.2496} \approx 750 \text{ days.}$$

USFs: the USFs open during the week only, but they receive much more people than hospitals. From day 130 to day 210 of 2020, they have attended a total of 7334 people. Taking into account people who come accompanied, the visitation period for USFs is

$$v_{pc} = (210 - 130) \frac{32702}{7334 * 1.2496} \approx 285.5$$

Supermarkets: to account for visitation routines in supermarkets, 5 of the largest of its kind have been interviewed. The supermarkets “Preço bom” have reported 3000 attendances every week, while supermarkets “Supermar”, “Mercado Nacional” and “Mercadinho Durare” have reported an average of 550 attendances weekly. As a result, the visitation period for this category of services in days is

$$v_{pc} = 7 \frac{20884}{3550} \approx 41.18,$$

as long as all age groups above 20 years old are considered consumers.

Markets: given the big difference in the contact network for supermarkets and other types of markets, we decided to separate them into two distinct types of service. For markets, which are more local and smaller in size, we gathered information from two representatives, namely markets “Mini Carrefour” and “Mercadinho do Beto”. These two markets reported an average of 50 visitors per week. We then assumed a similar visitation for all other 37 instances, which allowed us to estimate the following visitation period, in days:

$$v_{pc} = 7 \frac{20884}{50 * 39} \approx 74.97.$$

Food Stores and Construction stores: the other types of services that received people and that remained opened during the period considered were grouped into two categories:

- Food stores: all other types of services that sell specialized food, such as fruits and vegetables and beverages. This category also includes pharmacies.
- Construction stores: all types of stores that sell maintenance equipment, such as those for civil engineering, household equipment, vehicle parts, etc.

These two types of stores are small, and we assumed that their visitation periods were twice and four times longer than the visitation period of markets, respectively.

4.3.3 Estimation of service's contact network parameters

In this section, we describe the data used in the contact network parameters for Maragoggi-AL services, according to their definition given in Section 3.1.3.

Contacts are a way to quantify the opportunity for disease spreading if agents entering contact have the proper compartmental state. In this work, we have assumed that a contact has the following definition: “two people two meters or less from each other for the duration of one hour”.

A general procedure to quantify the network parameters, as described in the following sections, is

1. estimate the average amount of time t_{cont} people at two meters or less away from each other;

2. quantify the number of contacts by taking into account t_{cont} instead of one hour;
3. derive a weight w_{cont} to be multiplied by the parameter values, scaling contacts to the duration of one hour.

As an example, w_{cont} would be given by $w_{cont} = 1/12$ if t_{cont} is 5 minutes. That would mean that a person having 12 contacts of five minutes would equal having one of one hour.

Street Markets: for street markets, we have used the model described in section 3.1.3.5 with some few modifications:

- All workers are cashiers, hereby denominated sellers;
- Visitors can get in contact with more than one cashier/seller.

Since there are only two categories of agents inside street markets, cashiers and visitors, we need to approximate three parameters: the average number of contacts among sellers, the average number of contacts among visitors, and the average number of contacts between the two of them.

Before obtaining values for the average number of contacts, we need to estimate its average length of time t_{cont} . We have done so using recordings of individuals collected by drones on one of the days the street market opened. Following the routine of anonymous people on the street market, we calculated t_{cont} to be 5 minutes. Now we are in a position to estimate the average number of contacts.

Contacts among sellers: There were 185 sellers in the street markets during the time considered, distributed along 120 stands. 55 of these stands were owned by a seller and 65 of them had two sellers as owners. We assume that the stands with two sellers were constantly in contact and that an average of 3 contacts of 5 minutes happened between sellers of different stands each hour. As a result, the average number of contacts $\bar{c}_{sellers}$ between sellers per hour is

$$\bar{c}_{sellers} \approx \frac{65 * 12 + 3 * 120}{120} \approx 6. \quad (4.14)$$

Contacts between sellers and visitors: from the frames collected by drones during the opening hours of the street market, we estimate that about 300 people stayed around the stands every hour. We have also estimated that visits took 50 to 60 minutes on average. As a result, an average of 2 visitors were found around 60 stands, while 3 visitors stayed constantly close to 60 stands. In the worst case scenario, we have all groups of three people getting in contact with 2 sellers, 5 groups of two people getting in contact with 2 sellers, and the remaining 55 groups of two people getting in contact with 1 seller. As a result, the maximum number of contacts of 5 minutes that visitors have with sellers $c_{vis \rightarrow sell}^{max}$ is

$$c_{vis \rightarrow sell}^{max} \approx 12 \frac{60 * 3 * 2 + 5 * 2 * 2 + 55 * 2 * 1}{458} \approx 12.8, \quad (4.15)$$

where 458 is the average number of visitors present in the street market per hour, considering visits of 55 minutes. Analogously, in the best case scenario, 55 groups of three visitors are found around the stands of one seller, 5 groups of three visitors stay close to the stands with two sellers, and the remaining 60 groups of two visitors get in contact with two sellers per hour. In this case, we see that the minimum number of contacts of 5 minutes that visitors have with sellers $c_{vis \rightarrow sell}^{min}$ is

$$c_{vis \rightarrow sell}^{min} \approx 12 \frac{55 * 3 * 1 + 5 * 3 * 2 + 60 * 2 * 1}{458} \approx 8.3. \quad (4.16)$$

Taking the average between the worst and best case scenarios, we see that the average number of contacts $\bar{c}_{vis \rightarrow sell}$ that visitors have with sellers is

$$\bar{c}_{vis \rightarrow sell} \approx \frac{c_{vis \rightarrow sell}^{max} + c_{vis \rightarrow sell}^{min}}{2} \approx 10.6. \quad (4.17)$$

Contacts among visitors: From the data acquired through drone observations, we know that about 3000 people attend the street market when it opens. In addition, since visitors take about 55 minutes to shop, we also know that about 458 people visit the street market per hour. Of these people, some are shopping, and some are assumed to be randomly walking in the transit area of the fair. The remaining few formed clusters of people socializing. From our observations, the average number and amount of people in each cluster is

- 1 cluster of 5 people: $5 \times 4/2 = 10$ 1-hour contacts;
- 3 clusters of 4 people: $3(4 \times 3)/2 = 18$ 1-hour contacts;
- 11 clusters of 3 people: $11(3 \times 2)/2 = 33$ 1-hour contacts;
- 23 clusters of 2 people: $23(2 \times 1)/2 = 23$ 1-hour contacts.

The number of 1-hour contacts happening in stands where 2 visitors could be found is $60(2 \times 1/2) = 60$, and the total number of contacts that occur in the stands where 3 visitors could be found is $60(3 \times 2/2) = 180$. Finally, for the random walking of the remaining 62 people, we assume an infectious radius of 2 meters. Whenever agents are found at less than this distance from each other, we count a contact. However, since the transit area of the fair is approximately $1607 m^3$, it makes sense to consider this type of contact only for a number of people larger than $1607/4\pi \approx 128$. As a result, only the above two types of contacts are considered and, therefore,

$$\bar{c}_{visitors} \approx 12 \frac{324}{458} \approx 8.5, \quad (4.18)$$

where $\bar{c}_{visitors}$ is the average number of 5-minute contacts happening among visitors per hour.

Hospitals and other health facilities: for hospitals and other health facilities networks (that do not treat diagnosed individuals), data have been acquired from the city hall. For this type of service, the contact network employed is that introduced in Section 3.1.3.7, for which we have the following parameters:

- $p_{dis.w.}$: percentage of hospital workers that deal specifically with the pandemic disease in question;
- $\bar{c}_{workers}$: average number of 1-hour contacts among non-disease workers;
- $\bar{c}_{dis.w.}$: average number of 1-hour contacts among disease workers;
- $\bar{c}_{dis.w. \rightarrow w.}$: average number of 1-hour contacts from disease workers to non-disease workers;
- $\bar{c}_{visitors}$: average number of 1-hour contacts among visitors;
- $\bar{c}_{vis. \rightarrow w.}$: average number of 1-hour contacts from visitors to non-disease workers.
- $\bar{c}_{guests \rightarrow dis.w.}$: average number of 1-hour contacts from guests (admitted persons) to visitors.

According to city hall data, the values of the above parameters for campaign hospitals are: $p_{dis.w.} = 0.19$, $\bar{c}_{workers} = 2$, $\bar{c}_{dis.w.} = 2.9$, $\bar{c}_{dis.w. \rightarrow w.} = 0.2$, $\bar{c}_{visitors} = 2$, $\bar{c}_{vis. \rightarrow w.} = 1$, and $\bar{c}_{guests \rightarrow dis.w.} = 0.15$. For other types of health facility, the difference is that there are no disease workers who specifically deal with admitted persons. Therefore, the non-zero values for the above parameters are: $\bar{c}_{workers} = 2$, $\bar{c}_{visitors} = 2$, and $\bar{c}_{vis. \rightarrow w.} = 1$.

Markets, supermarkets, food stores and construction stores: the data used in the network parameters of these services have also been collected from city hall estimates. The type of network used here is that presented in Section 3.1.3.5, whose main parameters are

- $\bar{c}_{workers}$: average number of 5-minute contacts among workers;
- $\bar{c}_{visitors}$: average number of 5-minute contacts among visitors;
- $\bar{c}_{vis. \rightarrow w.}$: average number of 5-minute contacts from visitors to workers;
- $p_{cashier}$: percentage of workers that are cashiers.

For supermarkets, the above parameters have been estimated to be equal to: $\bar{c}_{workers} = 3$, $\bar{c}_{visitors} = 3$, $\bar{c}_{vis. \rightarrow w.} = 0.25$, and $\bar{c}_{vis. \rightarrow w.} = 0.22$. For the remaining services, these parameters are: $\bar{c}_{workers} = 3$, $\bar{c}_{visitors} = 3$, $\bar{c}_{vis. \rightarrow w.} = 0.25$, and $\bar{c}_{vis. \rightarrow w.} = 0.29$.

City hall: for the city hall we have used the standard Erdos-Renyi model, where the probability p_{ER} of an edge being added is given by $p_{ER} = c_{avrg} / (N - 1)$. Here, c_{avrg} is the average number of contacts, and N is the total number of nodes on the graph. However, the value of c_{avrg} has been calibrated along with the probability of infection due to lack of information, and also due to the high number of workers compared to the remaining services (355 versus 916). The calibrated value ended up being $c_{avrg} = 0.61$ 1-hour contacts per hour. See Section 4.4.

Environmental layer: for the environmental layer, which comprises agents out of home who are not in either of the other services, we have used the network model explained in Section

3.1.3.3. The only customizable parameter in this network is the urban area, which in the case of the Maragogi-AL city is: 7.654 km^2 .

Schools: for schools we have used an entirely different transmission model which is not based on physical contacts but rather on the aerosol transmission of infectious particles. See Section A for details.

4.4 The optimization program

After directly calculating or estimating all parameters that we consider relevant for simulating a community, we are left with the task of approximating the infection probability $p \in [0, 1]$ and the mean number of 1-hour contacts $c \in \mathbb{R}_+$ between workers at the City Hall. Since infection probability is a very behavior-dependent parameter, it is difficult to approximate it directly. Similarly, the contact network inside the City Hall could not be assumed from a priori information. To find parameter values that best fit the disease data, we use an optimization program to estimate these parameters for the period considered. In this section, we describe the methodology used in this optimization step, and we also provide numerical evidence that it is, in fact, well suited for the task.

Let \hat{x} be a candidate for approximating $x = (p, c)$. We evaluate how close \hat{x} is to x using the Wasserstein distance as a goodness of fit as follows:

- Let \mathcal{D} be a set of time markers (in our case, days), and let \mathcal{E} be defined by

$$\mathcal{E} = \{(s, e, i, r) \in [0, 1]^4 : s + e + i + r = 1\}. \quad (4.19)$$

Then $\mathcal{X} = \mathcal{D} \times \mathcal{E}$ contains any SEIR curve evaluated at times in \mathcal{D} . We define $\{X_i\}_{i=1}^n$ as the possible SEIR trajectories generated by our SEIR curve reconstruction (see Appendix 4.2) and $\{X_j^y\}_{j=1}^m$ be m i.i.d. trajectories generated by our model when we use the parameters in $(y_1, y_2) = y \in [0, 1] \times \mathbb{R}_+$ as the infection probability $p = y_1$ and the mean number of 1-hour contacts $c = y_2$. We also set $\hat{\nu}$ as the empirical measure given by $\{X_i\}_{i=1}^n$ and $\hat{\mu}^y$ as the empirical measure obtained from $\{X_j^y\}_{j=1}^m$.

- The L_1 -Wasserstein distance between $\hat{\nu}$ and $\hat{\mu}^y$ is given by

$$W_1(\hat{\nu}, \hat{\mu}^y) = \inf_{\gamma \in \Gamma(\hat{\nu}, \hat{\mu}^y)} \int_{\mathcal{X} \times \mathcal{X}} \|X - Y\|_1 d\gamma(X, Y), \quad (4.20)$$

where $\Gamma(\hat{\nu}, \hat{\mu}^y)$ is the set of all couplings of $\hat{\nu}$ and $\hat{\mu}^y$. In our case, with empirical measures having finite support, one can evaluate Equation 4.20 using linear programming, so we employ the solution implemented on the *Python Optimal Transport package* (FLAMARY *et al.*, 2021).

- We evaluate

$$\hat{x} = \operatorname{argmin}_y W_1(\hat{\nu}, \hat{\mu}^y) \quad (4.21)$$

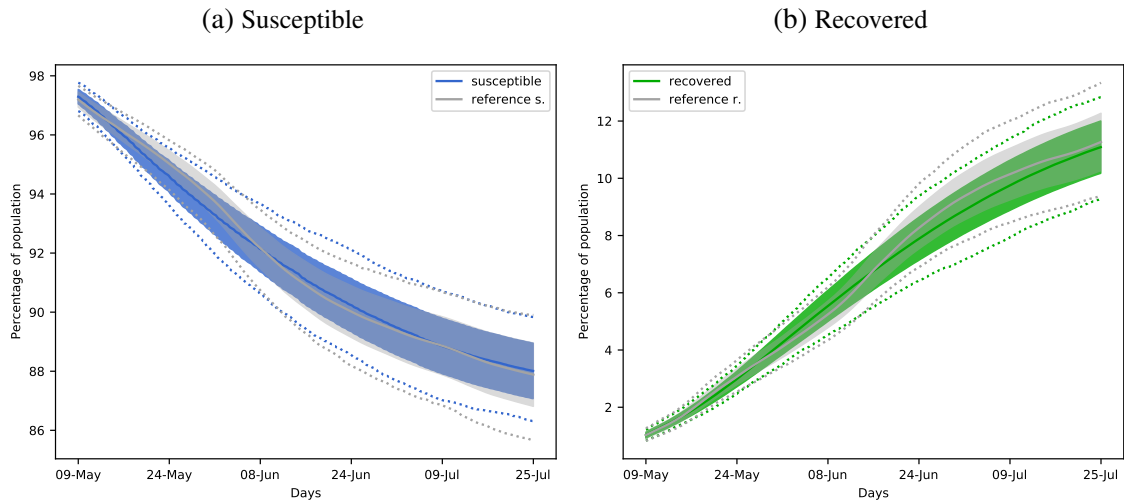


Figure 23 – **Reference susceptible and infectious compartmental distribution curves for the city of Maragogi-AL in comparison to their calibrated versions calculated from COMORBUSS.** The recovered curves also include the deceased compartment. The solid curves represent the mean over 384 samples, the dotted curves limit a 95% percentile of the distribution, and the colored clear region is bounded by two shifted mean curves. These shifted curves are obtained by summing and subtracting the point-wise standard deviation over the 384 samples.

in three steps: 1) using a population size of $N = 10000$, we perform a grid search to narrow down the search space; 2) still using $N = 10000$, the search for an optima in the narrowed space is performed by applying the Nelder-Mead algorithm; 3) we apply Nelder-Mead with $N = 32702$ (the real population size). The first two steps using a small population size reduce the computational cost of the process, and the last one corrects any artifact produced by rescaling to $N = 10000$.

In practice, we use $n = m = 384$ and \mathcal{D} as the days between May 9 and July 25 2020. The calibration procedure just described generates the following approximations for (p, c) : $p \approx 0.1356$ and $c \approx 0.6116$. The L_1 -Wasserstein distance between the approximated optima and the reference curves is then given by $W_1(\hat{\nu}, \hat{\mu}^{(0.1356, 0.6116)}) = 9.3 \times 10^{-3}$. The resulting SEIR curves are compared to the reference curves in Figures 23,24. We notice a very good fit, especially for the susceptible and recovered compartments. These compartments are, in fact, usually the ones obtained with the highest accuracy for the reference curves.

The quality of \hat{x} . The Wasserstein distance is a widely used goodness-of-fit measure (SOMMERFELD; MUNK, 2016; ARJOVSKY; CHINTALA; BOTTOU, 2017) for determining how close are two distributions. It has well-known concentration bounds when the measures are empirically approximated, which is exactly our case (see (DEDECKER; MERLEVÈDE, 2019; ARJOVSKY; CHINTALA; BOTTOU, 2017)). A good indicator of quality for the estimate \hat{x} is how closely one can recover a calibration parameter when COMORBUSS generates the input SEIR data using a given value for this parameter. We check this property experimentally by using the infection probability as the calibration parameter aforementioned. The experimental protocol is as follows:

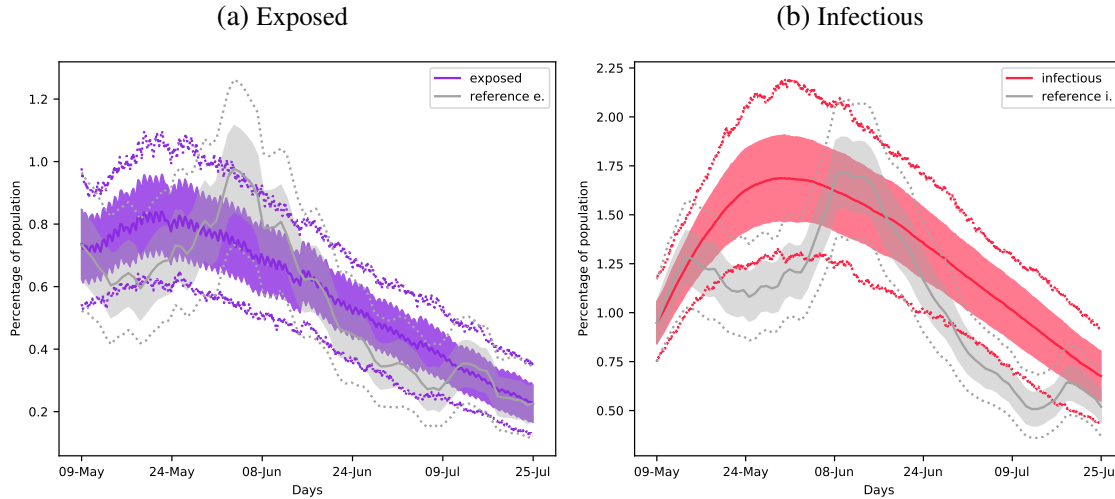


Figure 24 – **Reference exposed and infectious compartmental distribution curves for the city of Maragogi-AL in comparison to their calibrated versions calculated from COMORBUSS.** The solid curves represent the mean over 384 samples, the dotted curves limit a 95% percentile of the distribution, and the colored clear region is bounded by two shifted mean curves. These shifted curves are obtained by summing and subtracting the point-wise standard deviation over the 384 samples.

- Let $p \in [0, 1]$ be fixed and S_1, \dots, S_{50} be 50 disjoint sets of 384 seeds each (we took $S_1 = \{1001, \dots, 1384\}$, $S_2 = \{1384 + 1, \dots, 1000 + 2 \cdot 384\}$, etc.);
- We set in COMORBUSS the infection probability as $p = 0.15$, the mean number of 1-hour contacts in the City Hall as $c = 0.3$, and the population size as the full value $N = 32702$, and we run simulations using the seeds from S_i , $i = 1, \dots, 50$. This procedure generates 50 empirical measures $\hat{v}_i^{(0.15, 0.3)}$, $i = 1, \dots, 50$;
- For each $\hat{v}_i^{(0.15, 0.3)}$, $i = 1, \dots, 50$, we solve

$$\hat{x}_i = \operatorname{argmin}_{y \in [0, 1]} W_1(\hat{v}_i^{(0.15, 0.3)}, \hat{\mu}^y). \quad (4.22)$$

To simplify the procedure and reduce computational cost, we fix y_2 , the second coordinate of y as $y_2 = 0.3$. That is, we effectively only calibrate for the infection probability in this test. Nevertheless, this showcases the effectiveness of the proposed calibration procedure.

After trying to recover $p = 0.15$ as the infection probability using the procedure described above, we obtain the following approximation \hat{p} for p : $\hat{p} = 0.147 \pm 0.0008$. We notice that the approximation for p is very close to the original value we attempted to recover. This simulation asserts not only that the optimization program is effective for approaching the real observed value for (p, c) , according to the input data, but also that the scaling made in COMORBUSS for the population size is effective (see the section below).

4.5 Remarks about the population size

The most critical parameter for controlling computational time is the population size N . As a result, understanding the impact of this parameter on changes in results is essential.

Sensitivity analysis on population size N generally focuses on how the distribution of the final epidemic size (i.e., the distribution of the total number of cases after the epidemic ends) evolves with N . The dependence of a classical stochastic compartmental SEIR model with respect to N has been analyzed in (GREENWOOD; GORDILLO, 2009; BIBBONA; SIROVICH, 2017). In (GREENWOOD; GORDILLO, 2009), the authors provide experimental evidence that although the aforementioned distributions converge as N grows, their convergence is slow. This fact is verified by noticing that even with N on the order of 10^4 , significant differences can still be observed as N grows.

We designed a similar experiment for our model. Unlike the compartmental model, our stochastic agent-based model constructs an entire city and assigns individuals to networks (e.g. family structures, school networks, services networks). Approximating real populations using values of N distinct from the real population size may incur rescaling errors. Looking at the total number of individuals assigned to each relevant social activity modeled in the city of Maragogi-AL, we determined that the minimum population size necessary to keep at least one individual in each social role is $N = 1000$.

To test how the final epidemic size changes with respect to N , we evaluate the results obtained from COMORBUSS by setting $N \in \{1000, 2000, 3000, 4000, 5000, 10000, 15000, 20000, 30000\}$. We make 384 simulations for each value of N , and each simulation is run until the sum of exposed and infectious individuals becomes zero. Subsequently, we evaluated the percentage of the population that was infected, calling it the final size of the epidemic. The results are shown in Figure 25.

The outcome of our tests, shown in Figure 25, agrees with the results exposed in (GREENWOOD; GORDILLO, 2009). We understand these results from a probabilistic perception. For small population sizes, statistical fluctuations are more significant, since probabilistic events such as spreading the disease or recovering from it occur less frequently. This can lead to rapid decay in epidemic measures in more realizations of the community, leading even to bimodal distributions for the final epidemic size (see Figure 25). On the other hand, for large population sizes, the number of agents is prone to sustain the epidemic for a longer period of time. This is because we have a larger number of probabilistic events, which smooth out probabilistic fluctuations. This behavior helps to shift the distribution of the final epidemic size towards its right-sided mode (the process is clearly seen in Figure 25, where the histograms tend to the right hand side of the vertical dotted line as N increases).

In (GREENWOOD; GORDILLO, 2009) the authors point out that the final epidemic size distributions display a bimodal behavior with two peaks. Our simulations also give evidence

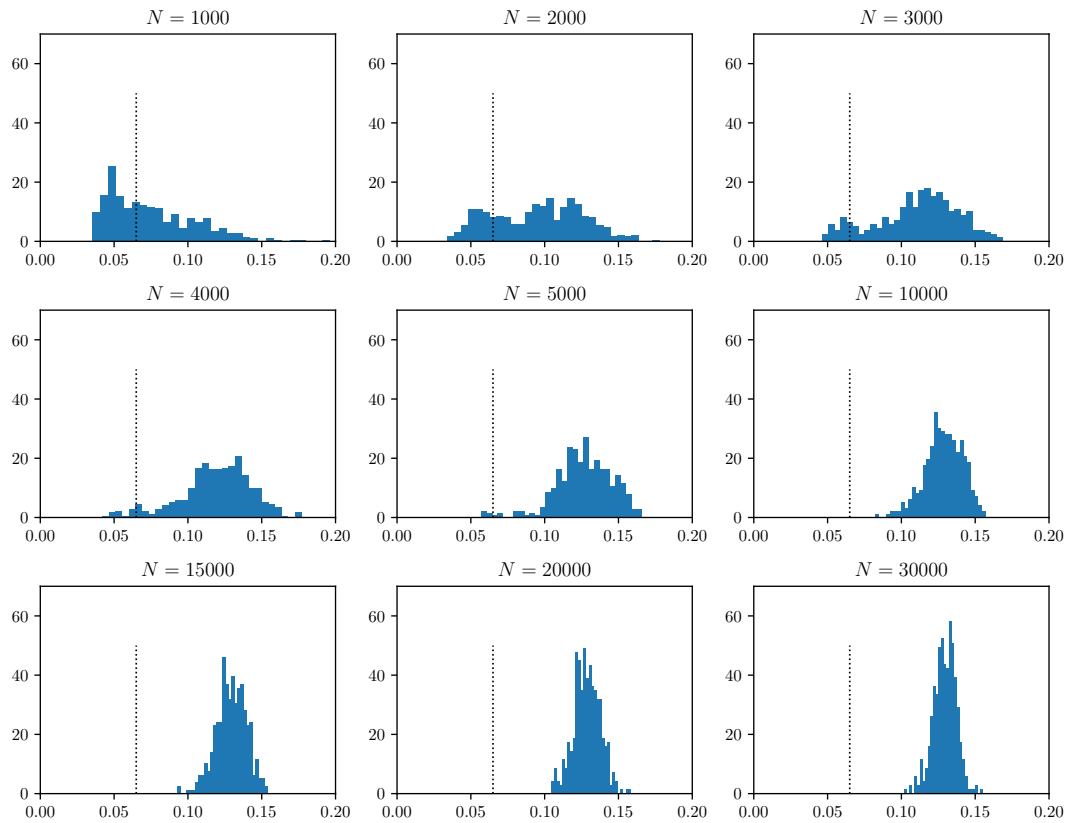


Figure 25 – **Histograms of the final epidemic size for different values of N .** The y-values are normalized so that the histograms represent a distribution. For low values of N the histograms are shifted towards the left side of the vertical dotted line, while for high values of N the tendency flips to the right hand side of the line. The variance decays as N grows, but the shape of the distribution still changes even for high values of N . Low values of N also show evidence of bi-modal behavior.

of the bimodal structure, especially for small populations (see Figure 25). This shows that COMORBUISS is capable of incorporating the classical properties of stochastic compartmental models.

Figure 26 helps to summarize how N affects the model's behavior, which we can outline as two regimes:

- *Low values (N of order 10^3).* Here, the epidemic has a more unpredictable behavior (the clouds are less concentrated), and it finishes sooner without infecting a large number of people (the clouds in the figure are shifted southwest). In fact, the average final epidemic size and the average final day for $N = 1000$ were 7.5% and 80.3 days, respectively. For $N = 30000$ they were 12.9% and 199.0 days, respectively;
- *High values (N of order 10^4).* Here, the epidemic has a more predictable behavior (the clouds in the figure are concentrated) and also a longer duration. Although for low values of N a longer duration is associated with larger epidemic sizes, this behavior is softened by high values of N : the correlation ρ between both variables decreases as N increases.

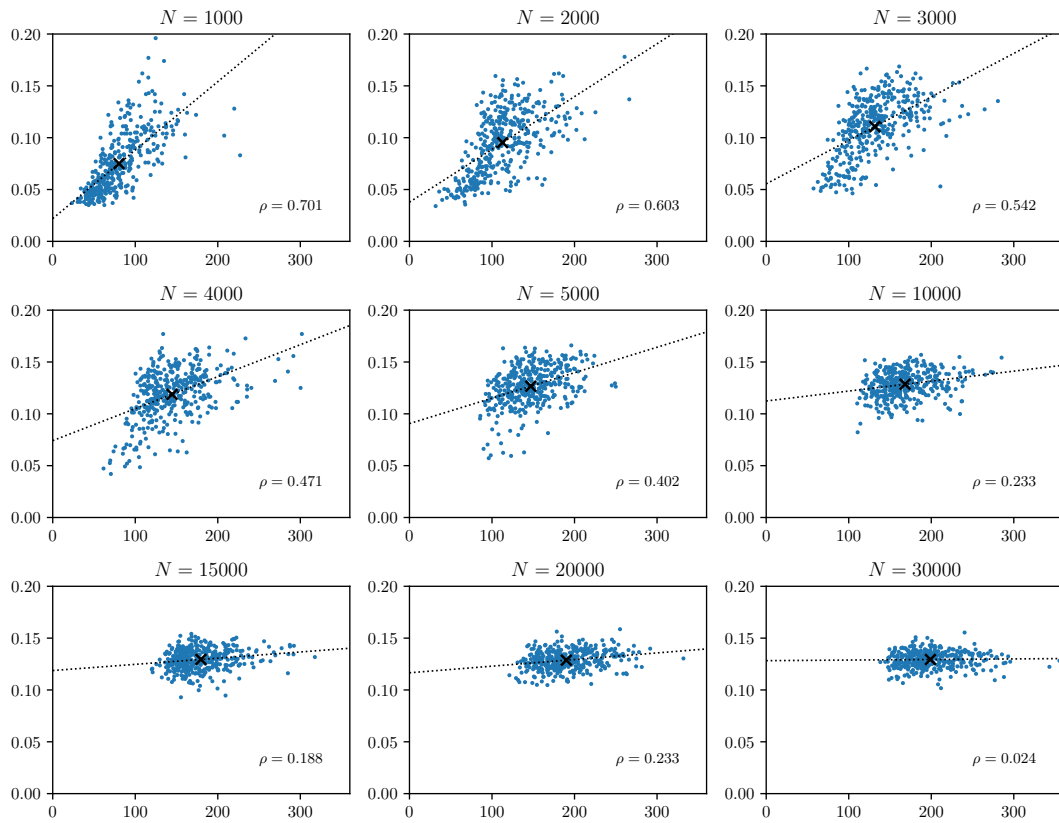


Figure 26 – **Final epidemic size (y-axis in %) vs number of days until the epidemic ends (x-axis in days) for different values of N .** The initial condition is $(S, E, I, R) = (.971, .007, .01, .012)$ for all realizations of the community. The X marker inside the clouds is the average over all points. The dotted line is a linear regression on the data, and ρ is the correlation between both variables (epidemic size and its total duration).

These results show that one must avoid approximating population size of order 10^4 using population sizes of order 10^3 whenever possible. Approximations between the same magnitudes are possible since the population sizes of 10000, 20000 and 30000 display average final epidemic sizes of 12.85%, 12.87% and 12.94%, respectively. Other variables are not as robust with respect to changes in population size. For instance, the average total duration of the pandemic for a population size equal to 10000, 20000, and 30000 was 168.2, 190.1 and 199.0 days, respectively.

As a rule of thumb, we choose to approximate the population of Maragogi-AL (32702 individuals) using $N = 10000$ on the most computationally expensive and repetitive routines, such as the calibration process described above. For less expensive routines, such as those comparing different opening scenarios for schools, we make no approximation ($N = 32702$).

APPLICATION: PROTOCOL EVALUATION FOR SAFE SCHOOL ACTIVITIES

The educational system plays a fundamental role in the socio-intellectual development and mental health of children and adolescents. During the COVID-19 pandemic, the impact of school closures on society has been enormous. UNESCO reported that, as of April 8, 2020, up to 188 countries closed schools nationwide. In developing countries, such as Brazil, the nutritional well-being of children was put in jeopardy, as families rely on school meals. And yet, in Brazil alone, schools remained closed full-time for 191 days in 2020 affecting 44.3 million children. However, given the frequent contact during a school day, the prevalence of mild symptoms in children and the role of school as a source of contacts that bridge family nuclei, there is understandable concern that face-to-face classes could drive uncontrolled spreading of the virus.

In view of the negative physical and mental consequences for students, together with the educational deficit imposed by school closures, the ECDC agency points out that transmission mitigation measures are necessary for students to have a safe socialization and learning environment (COVID-19..., 2021). Therefore, a major concern is the evaluation of mitigation protocols (THOMPSON; AL., 2020) to understand the impact of each measure within the school community.

Living in a household with a child who goes to school physically increases the risk of being infected by up to 38%. Similarly, school teachers are 1.8 times more likely to be infected than those working from home (LESSLER *et al.*, 2021) and the return of face-to-face classes has been directly related to outbreaks (COVID-19..., 2021). Mitigation measures such as separating student groups, quarantining exposed students and professionals, wearing masks, maintaining adequate air ventilation, vaccinating risk groups, and monitoring the emergence of cases can all decrease the number of new cases (GURDASANI *et al.*, 2021; MUNDAY *et al.*, 2021; LESSLER *et al.*, 2021).

Often, mitigation measures are put in place simultaneously, making it difficult to disen-

tangle their individual impact on transmission from temporal case report datasets. The lack of infrastructure, personnel, and laboratory equipment may also limit the use of these measures in developing countries, especially when they are based on resource intensive practices such as testing and subsequent contact tracing of cases. Thus, it becomes crucial to identify effective mitigation practices *a priori*.

Our aim in this chapter is to quantitatively assess the effects of vaccination (SILVA *et al.*, 2021) and NPIs protocols and find effective protocols for school activities. Our study shows that classes can be kept open safely, provided that the correct combination of measures is adopted. Relying on a single measure is mostly not effective or stable, but simple measures can go a long way when properly combined and implemented.

5.1 Materials and methods

5.1.1 Data collection

The city of Maragogi in Northeast Brazil has 33,000 inhabitants (IBGE... , 2021) and is a representative of at least 40% of Brazilian cities in terms of income and demographics. Moreover, its demography is also typical worldwide, being located above the 50% quantile in a sample of 28,372 North American cities and 41,000 global cities, using the `simplemaps` database (UNITED... , 2021; WORLD... , 2021), see Section 1.3 for further details.

Through a partnership with the city of Maragogi, established since March 2020, we developed a Clinical Monitoring System to track and trial all severe acute respiratory syndrome patients. We also geolocalized the patients and integrated this information with public data to obtain household socio-economic data and family clusters Section 4.1. The data integration is illustrated in the upper left panel of Fig. 27. For our study, we used data from May 9, 2020, to July 25, 2020, consisting of 18 confirmed deaths and 119 hospitalizations. In this period 1722 tests were performed, namely 52 RT-PCR tests and 1670 antibody tests (in majority COVID-19 IgG/IgM, see Section 4.2 for further details).

This study was approved by UFAL institutional Ethics Committee (CAAE: 43058821.9.0000.5013).

5.1.1.1 Services

We mapped the services that were allowed to be open during the period under government regulations and interviewed a sample of businesses to estimate daily occupation. The bulk of such services are food stores, building supply stores, restaurants, and other minor retail services as described in 3.1.1.

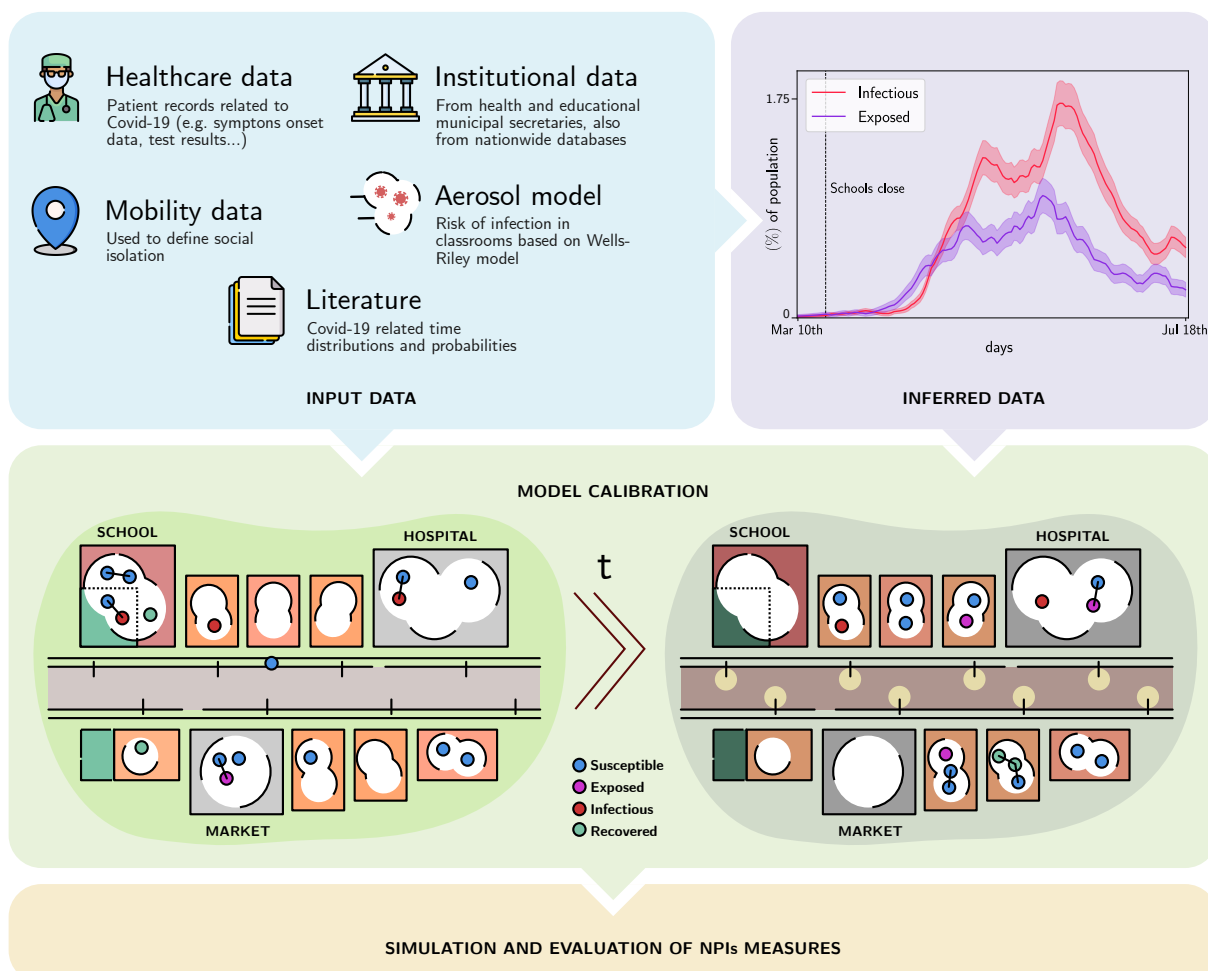


Figure 27 – **Pipeline overview description.** Data is collected as patients attend health institutions. Health professionals register patients’ personal, epidemiological, and geolocation data to the Clinical Monitoring System (CMS), which is blended with socio-economical and household data. Using these data, we estimate the number of Exposed (blue), Infectious (red), and Recovered (green) individuals. All the pre-processed data is used to calibrate our stochastic agent-based model, COMORBUSS. From bottom left to right: a schematic representation of the social dynamics of COMORBUSS, producing contacts between individuals in different social contexts. The colored circles represent the state of individuals and the lines represent relevant physical contacts capable of producing contagions. Once calibrated, the model is used to estimate the effectiveness of NPIs.

5.1.1.2 Street markets

We estimated the usage of important open air services such as street markets by images collected by drones. We processed the images using the Drone Deploy mapping software marking tool (DRONE... , 2020) to evaluate the mean size and duration of the cluster of people less than 2 meters apart during opening hours, as well as the average time spent by individuals in the street market. In Section 1.3, we also show that cities with demographics similar to Marogogi have analogous street market behavior.

5.1.1.3 Health services

During the period considered, the triage of all COVID-19 related cases was performed in a field hospital. We interviewed the health secretary's staff to obtain data on the mean appointment time and the mean number of contacts a patient has with doctors and other patients. This also provided data on the mean number of contacts among staff, see 3.1.3.7.

5.1.2 Inference of states from data

We estimate the epidemiological SEIR curve from attendance data from our Clinical Monitoring System. The SEIR curve corresponds to the trajectory of the population over the period of observation in the states: susceptible, exposed, infectious, and recovered. The challenge is to transform the information of an individual reported in the attendance data into these states of the entire city population over time, correcting for subnotification.

Under the hypothesis that all severe cases (hospitalization and death) are reported in our Clinical Monitoring System, for each reported individual, we estimated the number of unreported infected individuals using a negative binomial (NB) distribution and consequently the total number of cases in the city over the period of observation. We modeled the total number of cases by $T = NB(p_h, 119) + 119$, where $p_h \approx 3.304\%$ is the estimated probability of hospitalization for the city. We assume that these unreported individuals present their first symptoms at the same time as reported individuals.

Having all individuals carrying the virus, we estimated how they progress across the SEIR compartments based on the severity of the case and the distribution of the permanence of each state (WÖLFEL *et al.*, 2020). We rerun the statistical model 400 times to obtain SEIR curve samples for the city, see Section 4.2 for further details. We denote by $\hat{\nu}$ the (empirical) distribution induced by these samples, for example, the measure given by the uniform distribution over the 400 obtained samples.

5.1.3 Agent based modeling

Agent based models are a class of computational models that track individual units (agents) of objects of interest. In the case of communal disease transmission, the natural choice for agents is the people who form that community and on whose contact the disease transmission is based. The two most important advantages of these models are: i) we can directly incorporate the biological and social heterogeneity of that community and investigate how it influences transmission patterns; ii) we are omniscient regarding the simulated histories of the agents and can reliably evaluate the effects of specific public health protocols via counterfactual analysis of these histories.

Our agent-based model, called COMORBUISS (COMMunitary Malady Observer of REproduction and Behavior via Universal Stochastic Simulations), takes all these advantages a

few steps further: we built a full model for the social dynamics of general communities in order to produce the contacts that drive disease propagation. We achieve this via a general modeling procedure of a city's infrastructure, which can be systematically applied to any city via data integration. Moreover, our model is aware of the different roles the agents play in the various services that compose the infrastructure and produces contacts accordingly. This allows us to pinpoint the impact of a specific service and related mitigation protocols on disease spreading, as well as track the resultant infection tree. To avoid over-specialized simulations of a single city, COMORBUSS stochastically produces for every simulation a realization of the transmission trajectory for the city in the class defined by the desired demographic and infrastructure data. For instance, each simulation has its own household network while satisfying the same distributions that describe the household structure in that community. In the following, we describe the main parts of the model and elaborate on its many details in Chapter 3. The most important parameters are classified and explained in Figure 28.

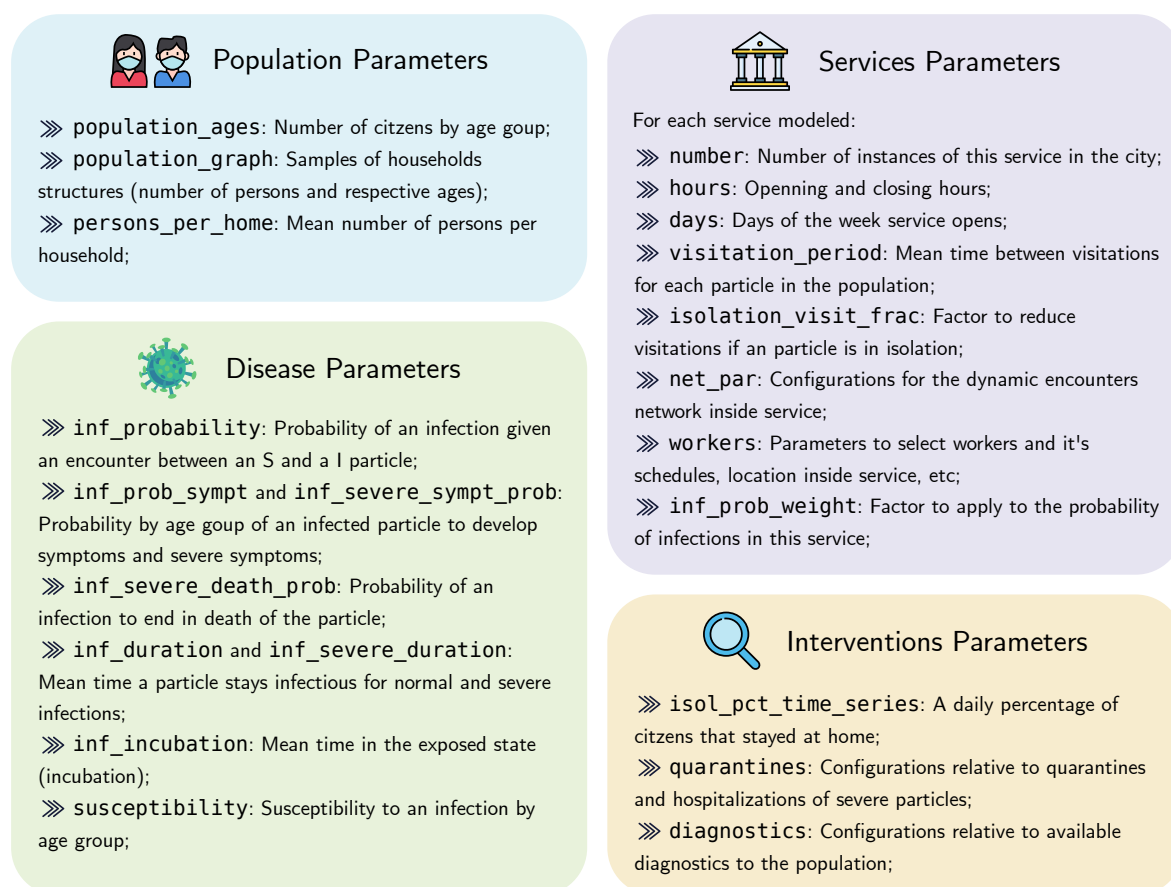


Figure 28 – **Most relevant parameters.** A non exhaustive classification of parameters used for a COMORBUSS simulation. A detailed description of parameters can be found in Chapter 3, while a complete list of parameters and their values can be found in the [Git repository](#).

5.1.3.1 Modeling disease

Each agent is characterized by its age, which determines the susceptibility of the agent, the probability of developing symptoms and the probability of dying from the disease. When a susceptible agent encounters an infectious one (pre-symptomatic, asymptomatic, mildly, or severely symptomatic), it has a probability of becoming exposed. After an incubation period, this agent becomes pre-symptomatic, and after an activation period, its state is converted to either asymptomatic, mildly, or severely symptomatic. The distribution of these states is empirically estimated from actual statistics (LINTON *et al.*, 2020; VERITY *et al.*, 2020). After a random period, the agents are converted to recovered (or deceased); see 3.2.

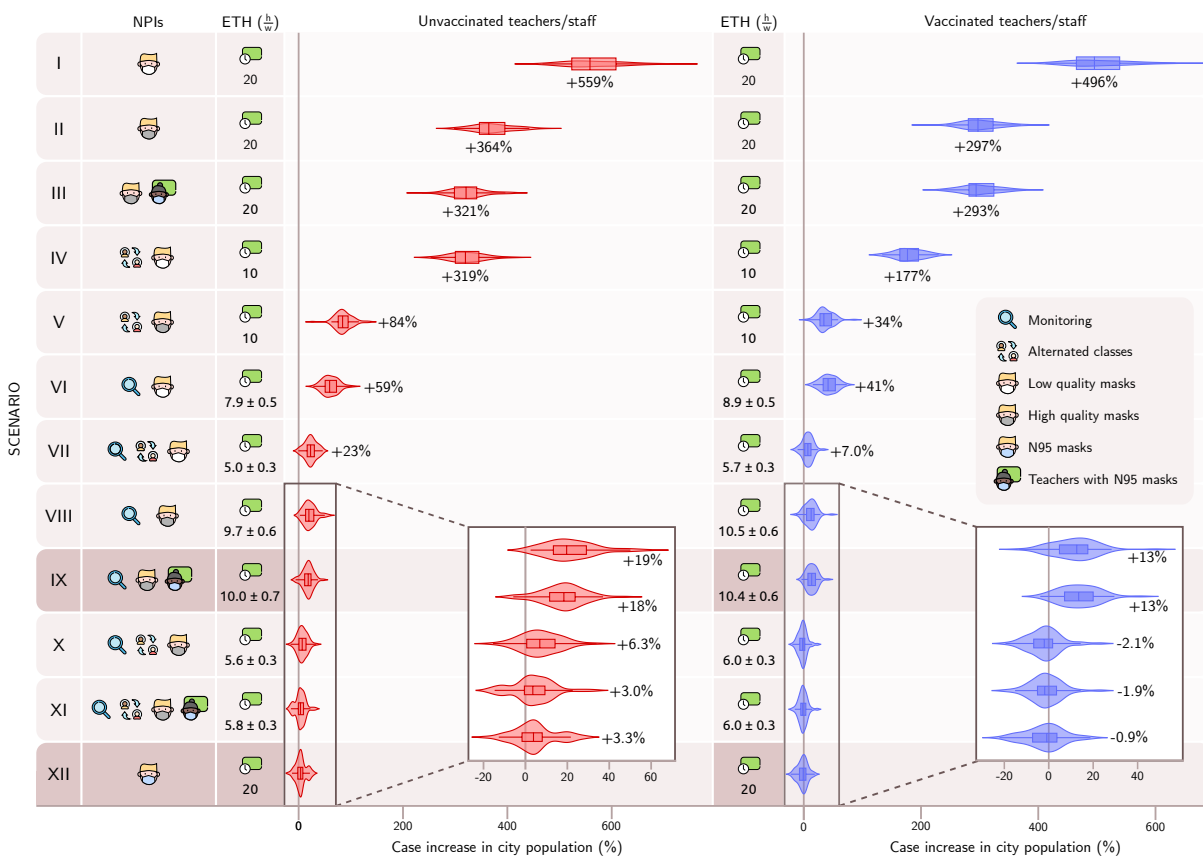


Figure 29 – **Combination of NPIs measures in comparison to the baseline model settings.** Left panel: Cases increase under different scenarios with unvaccinated teachers and staff. Right panel: Case increase in different scenarios with vaccinated teachers and staff. The effective teaching hours in hours/week $\frac{h}{w}$ and case increase in school population with respect to baseline are displayed for each NPI combination. In case the active monitoring is also applied, the mean and standard deviation over 60 realizations for the effective teaching hours are shown. The proportional increase in the number of cases is displayed as violin plots (median, lower, and upper quartiles), with kernel density estimates for distributions.

5.1.3.2 Interventions for the schools evaluation

The scenarios simulated in this study are based on the first wave of infection in Maragogi, so services related to tourism are closed. The other standard NPI adopted in the base scenario is

social isolation based on telephonic triangulation data processed by (INCOGNIA... , 2021) to provide the daily percentage of people who stayed home. This is modeled by randomly selecting at the beginning of each simulation day the desired number of agents and confining them to their homes for that day.

Standard testing policy is the serological testing of symptomatic agents. Diagnosed agents are quarantined at home if they present mild symptoms or are hospitalized if their symptoms are severe. Quarantines and hospitalizations are lifted when agents leave the infectious compartment after recovering or dying.

5.1.3.3 Intervention in School Dynamics

We implemented and combined the following NPIs in the context of schools:

- Reduced workload: daily teaching hours are reduced from 4 to 2 hours;
- Alternating groups: students are separated into two groups which attend the classroom in alternating days;
- Use of masks: students and professors are supplied masks with given penetration factors;
- Active monitoring: suspicious cases are monitored and intermittent closing is declared upon discovery of cases
 - suspicious cases are students, professionals or their relatives which present symptoms;
 - suspicious cases are tested and if the diagnose is positive the student is quarantined;
 - the classroom associated to the quarantined person is closed for 14 days;
 - if using alternating groups, only the group associated to the quarantined person is suspended;
 - if more than one classroom is closed in the span of a week, the whole school is closed for a week.

The effects of these NPIs and their combinations are the main results of this work.

5.1.3.4 Aerosol transmission model: masks and air exchange

Interventions in the aerosol model are made by parameterizing Equation (A.14) in Appendix A. We introduce values for the penetration factor of masks p_m^i used by students and professionals and test the efficacy of different scenarios with various values of the volume flow rate Λ of air with the exterior. For reference, we highlight documented or recommended values of these parameters.

5.1.3.5 Vaccination model

The vaccination model used in this study is a simple binary infection model. Vaccinated agents can become immune (susceptibility 0) with a probability given by the effectiveness of the vaccine after a given period. If vaccination of an agent does not lead to immunity, its susceptibility remains unchanged. In the present work, we assumed a worst-case scenario where vaccines were not widely available and were prioritized for teachers and staff. Simulations with vaccinated teachers and staff, they are initialized assuming that they have protective neutralizing antibodies against COVID-19. This blocks any possible infection chain that begins with these individuals.

Secondly, we investigate the effects of NPI adoption under different scenarios of partial vaccination for the general population (see Fig. 30). Our main interest in this analysis is to evaluate the viability of the proposed measures for countries with different vaccination coverage, both in the well-covered European continent and in the undervaccinated African continent. We observe that the correct choice of NPIs can effectively protect the community even for low vaccination coverage, while poor adoption of NPIs can lead to high infection rates even for high vaccination coverage. Since we are dealing with larger segments of the population rather than just the school sub-population, these simulations were performed with a more realistic vaccination model that only partially protects each agent with a biological efficacy of 98% for infection, resulting in an effective vaccine efficacy of 90% for the scenario where no NPIs are adopted. Although it tends to be more realistic, this model is highly complex to adjust and interpret because the measured vaccine efficacy is closely related to the running epidemiological scenario which responds to the adopted NPIs (KASLOW, 2021; STRUCHINER; HALLORAN, 2007; MADEWELL *et al.*, 2021).

5.1.3.6 Modeling Services

The city infrastructure is modeled by creating individual instances for each service (schools, hospitals, markets, restaurants, shops, etc.) and by assigning agents to work/visit that location if they belong to an appropriate age group (a child may not work at a shop, and an adult may not attend class). Worker agents are relocated to that service location during their shifts, whereas the visits of client agents are simulated stochastically. An hourly visitation rate of a service by an agent is empirically estimated, taking into account the service's opening hours and average visitation frequency of real clients; for details, see 3.1.1.2. Additionally, agents may be assigned as guests to special services, which implies that their standard location is changed from their homes to that service instance. In this way, we distinguish between hospitalizations, hotel quarantines, and nursing home patients.

A novel point of our model is the creation of contact networks contextualized by social activity. The ratio of encounters between workers and clients, as well as the clustering properties of a contact network, naturally depend on the observed social context. For example, in restaurants

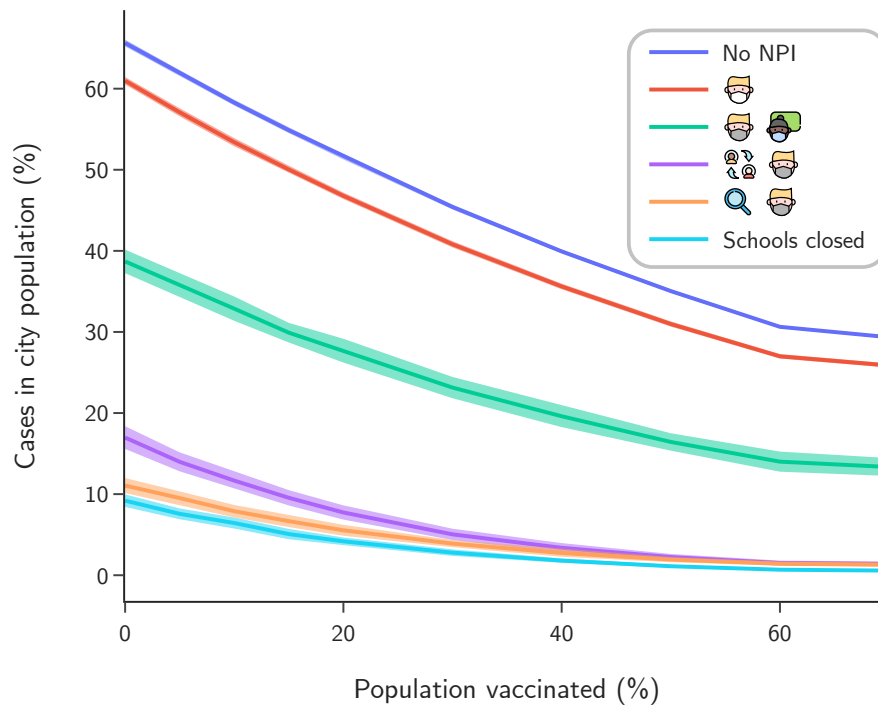


Figure 30 – **Population fraction infected at the end of the simulation period (77 days) under varying vaccination coverage.**

there is a clustering of clients belonging to the same table, and contact between different tables is mediated by the contact of a shared waiter. Contact networks in schools, hospitals, stores, etc., are all considerably different from each other. COMORBUSS updates random contact networks every hour for all agents in service instances, while respecting the characteristic architecture of the contact network of that type of service and distinguishing between the social roles of agents. Details and examples may be found in 3.1.3.

5.1.4 Model calibration and closed schools as baseline

We aggregate socio-geographical data, as well as epidemiological data to COMORBUSS from May 9th 2020 to July 25th 2020, and leave the infection probability p and the mean number of contacts c in the City Hall to be calibrated using the empirical measure $\hat{\nu}$ obtained from the inference of states from data (see Section C). For a given $y = (q, d) \in [0, 1] \times \mathbb{R}_+$ we denote by $\hat{\mu}^y$ the empirical measure given by 400 independent realizations of COMORBUSS with p and c chosen as $(p, c) = y$. We construct an estimate \hat{x} for $x = (p, c)$ by minimizing over all y the L_1 -Wasserstein distance between $\hat{\nu}$ and $\hat{\mu}^y$, see Section 4.3.

We initialize the community according to its demographics and household distribution, see Section 4.3. The disease state of agents is proportional to the average inferred epidemiological data for day May 9th 2020. The calibrated model is in excellent agreement with the estimated data and we use it as a baseline. This scenario resulted in an average of 3007 ± 249 new infections

in the population, in which 25% of those infections occurred in the school population, a measure that will serve as a baseline for keeping schools open in study cases.

5.1.5 Poorly ventilated classrooms

In poorly ventilated classrooms, the main transmission mechanism is aerosols emitted by an infected agent. Aerosols can remain suspended in the air, thereby reaching agents far from the original emitters (MORAWSKA; CAO, 2020; POYDENOT *et al.*, 2021). To model this exhaled air without reference to the microscopic pathogen concentration, we follow the exposition in (MILLER *et al.*, 2021; BAZANT; BUSH, 2021), which describes the evolution of the quanta concentration in a closed space. *Quanta*, introduced by Wells, measure the expected rate of disease transmission, interpreted as the transference of the quanta of infection between pairs of infected and susceptible agents (RILEY; MURPHY; RILEY, 1978).

In our model, we denote by C ($quanta/m^3$) the total concentration of quanta inside a classroom of volume V . Classrooms contain a total of N agents, with S susceptible individuals, I_s infected students and I_t infected teachers. All breathe uniformly at a rate $B = 0.5 m^3/h$. Since mask wearing can decrease the amount of aerosols emitted into the air, we denote for each agent the penetration mask factor $p_m^i \in (0, 1)$, with $i = s, t$: see Fig. 31.

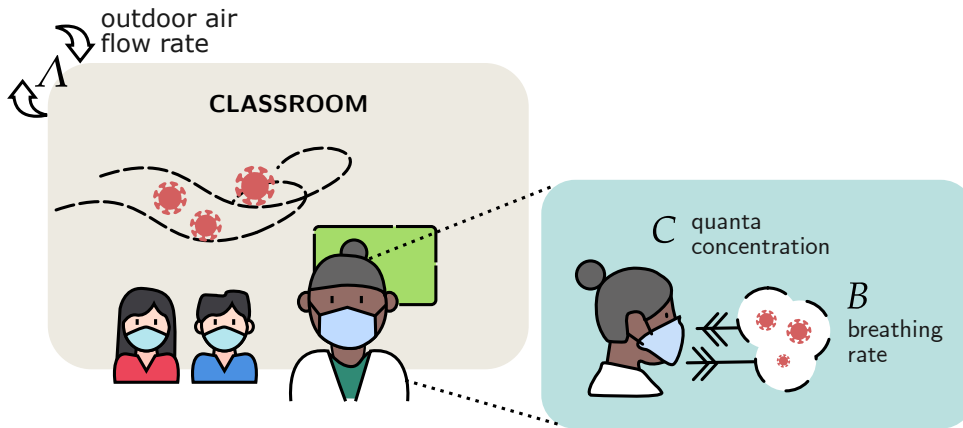


Figure 31 – **Airborne transmission model inside school environment.** The classroom is an enclosed space in which airborne transmission has a high chance of occurrence. Contaminated particles are spread over the classroom, allowing long range infections. The fresh air rate flow Λ quantifies the classroom ventilation. The quanta concentration C varies in the environment depending on the breathing activity.

Each person exchanges quanta with the air depending on breathing activity. We introduce the concentration of quanta expelled by students $C_s = 40$ ($quanta/m^3$) and teachers $C_t = 72$ ($quanta/m^3$) (BAZANT; BUSH, 2021) (corresponding to voice counting (MORAWSKA *et al.*, 2009)). Under a well-mixed room assumption, the total concentration of quanta C ($quanta/m^3$) inside the classroom satisfies the mass equation:

$$V \frac{dC}{dt} = -(\Lambda V + NB)C + B(C_s I_s p_m^s + C_t I_t p_m^t). \quad (5.1)$$

Note that our setting is based on the fact that airborne particles remain airborne before being extracted by outdoor air flow Λ (typically reported as air changes per hour or ACH) or inhaled by an agent. We investigate the poor ventilation limit $\Lambda = 0$ and the fresh air flow ($\Lambda > 0$), see Appendix A.

The amount of quanta inhaled by the i th agent within the class over a period of time t is the inhaled dose $D_i(t) = Bp_m^i \int_0^t C(t)dt$. We evaluate this integral over the solution to Equation (5.1). Using the inhaled dose of each agent, we plug it into the Wells-Riley model to calculate the probability that a susceptible individual is infected (MILLER *et al.*, 2021; POYDENOT *et al.*, 2021), which consists of estimating the risk of infection in indoor environments via the inhaled dose.

$$P_{indoor}^i(t) = 1 - e^{-r_i D_i(t)},$$

where r_i is the relative susceptibility (an age-based measure (ZHANG *et al.*, 2020)) for the agent i . We set the relative susceptibility of children (age 0 to 14 years), adults (age 15 to 64 years), and the elderly (over 65 years) to $r_i = 0.23, 0.68, 1$, respectively. To determine the source of infection of a particular exposed individual, we pick a random individual uniformly from all the infectious individuals in the enclosed space; see Appendix A.

5.2 Results and Discussion

We present three classes of results, each with their own implications for the design of health protocols: i) effectiveness analysis of a large set of protocols; ii) analysis of how the most relevant protocols depend on good mask practices and ventilation; iii) predictions on the effectiveness of the protocol when challenged by more infectious viral strains.

It must be noted that, while our model can be easily applied to other communities via our systematic data integration procedure, acquiring good quality datasets and ensuring their compatibility is the most limiting challenge in our methodology. For example, we have found that in many cities the census data and the database describing the available services are offset by a few years. We had the experience of modeling cities which had explosive growth during those years and these two datasets became so incompatible that there were not enough agents from the demographic data to work on the most recent infrastructure. We naturally need to rely on interpolation and extrapolation of historical datasets in such cases. Regardless, we find that a close collaboration with city managers, as we had in Maragogi, is ideal for ensuring the quality of the data, as well as in identifying trends and supporting modeling choices. This is critical in order to evolve the model as we learn more about the disease, and the social behavior also changes in response to it. Our experience in this process was documented on Chapter 4.

5.2.1 NPIs and vaccination

In 27 schools, the total school population is 8,528, with 7,557 students. We quantified the effects of five NPIs on the school population, which consists of teachers, school staff, and students. Each NPI is described in Fig. 32. Although there is still controversy in the literature on the efficiency of surgical masks for filtering particles (CHENG *et al.*, 2021) and side effects (KISIELINSKI *et al.*, 2021), we assign mask quality via their permeability factors p_m , as indicated in Fig. 32.









NPI	Description
 Reduced Workload	Schools function with shifts of two hours instead of four hours.
 Alternating Groups	Schools function with reduced class sizes, and in particular classes are separated into 2 groups having in-person activities on alternate days.
 Use of Mask	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;">  Low quality: $p_m = 0.5$ </div> <div style="text-align: center;">  N95 or PFF2: $p_m = 0.05$ </div> </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <div style="text-align: center;">  Good quality: $p_m = 0.3$ </div> <div style="text-align: center;">  Teachers and staff with N95. </div> </div>
 Active Monitoring	Schools function under the following measures: <ul style="list-style-type: none"> • Symptomatic people are tested; • If a case is found in a classroom, their activities are suspended for 14 days; • Students are tested and isolated (14 days) when they are symptomatic or a family member is confirmed positive; • Teachers which had contact with a classroom in which there were confirmed cases are tested and suspended for 14 days in the case of positive result; • School is closed for one week if there are two cases in distinct classes within a week.

Figure 32 – **NPIs description.** The icons distinguish the nonpharmaceutical interventions evaluated in this study. In scenarios involving masks, the mask penetration factor p_m is uniform for all individuals, except for teachers wearing PFF2 masks.

We simulate school activities with different NPI and compute the percentage increase in cases with respect to the baseline. The results are presented in Fig. 29 along with the effective teaching hours. Conducting classes in full shift and wearing only poor quality masks leads to a 559% increase in infections. We note that the wearing of N95 masks by teachers and staff is particularly effective in reducing the number of cases compared to other scenarios, and we

highlight this NPI in Fig. 29 (darker color). Active monitoring prevents spread, at the expense of the effective number of teaching hours.

In the simulation, we assume that the vaccinated teachers and staff are initialized with protective neutralizing antibodies against COVID-19. This blocks any possible infection chain starting from these individuals. The right panel of Fig. 29 displays the effectiveness of NPI combinations with vaccinated employees. If employees are not vaccinated, case rates increase in all scenarios. The case increases in the highlighted (darker color) scenarios are reduced for both unvaccinated and vaccinated employees, indicating that they are a potential source of infection for the school population.

We also analyze the robustness of our results when considering a larger city, using, as an example, the regional capital of Curitiba with almost 2 million inhabitants. We observe how bad protocols lead to a sharp increase in infections while good ones successfully avoid this phenomenon. Most notably, the relative effectiveness rank between intervention is preserved, even if the case increase relative to the baseline is less pronounced, see more details in the Appendix B. This not only shows the stability of the protocols but also indicates that smaller cities are more vulnerable and need appropriate protocols.

We also consider the effectiveness of NPI scenarios under different levels of vaccination coverage, see Fig. 30. Our motivation is to assess the viability and safety of public health decisions even in countries with low coverage, such as African countries. In fact, even with a low vaccination coverage, we find that a good choice of NPIs in schools also protects the larger community better. At the same time, poorly chosen or nonexistent NPIs may leave communities highly exposed, regardless of vaccination coverage. We therefore stress the importance of appropriate NPIs and protocols, whether or not the underlying country enjoys good vaccine coverage. We recall that cities are modeled with only essential services operating, including schools. The lessons learned here extend to other services and social contexts to avoid the worsening of outbreaks.

5.2.2 Sensitivity analysis: mask penetration and ventilation

We quantify the relevance of mask penetration factor p_m and ventilation air flow rate Λ for the increase in COVID-19 cases in cities. Assuming that all pupils wear masks with the same p_m , Fig. 33 shows the impact of the penetration factor on the number of cases if schools are kept open. The results are sensitive to the penetration factor of the masks, as seen by comparing the first (poor quality or practices, $p_m = 0.5$) and second (high quality masks, $p_m = 0.3$) simulation scenarios, showing a decrease of almost 200% in cases regardless of the vaccination status of employees. We also observe that the use of N95 masks by employees increases the effective teaching hours in scenarios with active monitoring.

Fig 34 shows the sensitivity analysis when the ventilation rate is varied inside classrooms.

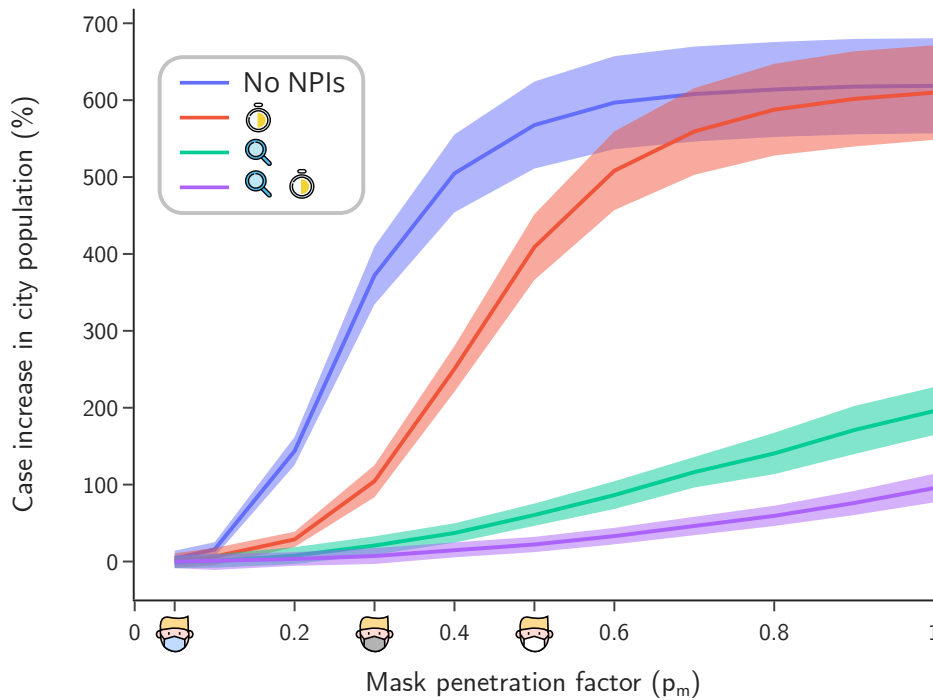


Figure 33 – **Sensitivity analysis across mask penetration factor p_m .** Cases increase in school population (solid lines) versus the mask penetration (mean values over 60 realizations for each p_m value).

Based on the recommendations of the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) (ASHRAE . . . , 2021), we calculated the minimal ventilation rate of $\Lambda_1 = 0.8 h^{-1}$ for unoccupied classrooms using their average dimensions in Maragogi. The ventilation rates for the half full and full classrooms are $\Lambda_2 = 3.8 h^{-1}$ and $\Lambda_3 = 6.6 h^{-1}$, respectively; for more details, see Appendix A.

5.2.3 Scenarios with more infectious variants

When investigating the effectiveness of school safety protocols during infection waves caused by new, more infectious variants, we are drawn to the limiting worst-case scenarios. As such, we assume that the new variant completely avoids acquired immunity from vaccination or previous infections. New variants are modeled by an increase in population susceptibility, therefore encompassing both our contact and aerosol transmission models. Susceptibility is increased by the multiplying factor over all age groups as a limiting case.

The results are depicted in Fig. 35. As expected, the total infected population increases monotonically with the increase in susceptibility, with poor protocols for school activities leading to extreme infection rates in the community. Most importantly, not only do good protocols still lead to a remarkable decrease in infection rates, but the relative rank of effectiveness between protocols is preserved regardless of how much susceptibility is increased. This shows the stability of good protocols and makes the point that their adoption should always be a top priority, even when faced with potentially new variants.

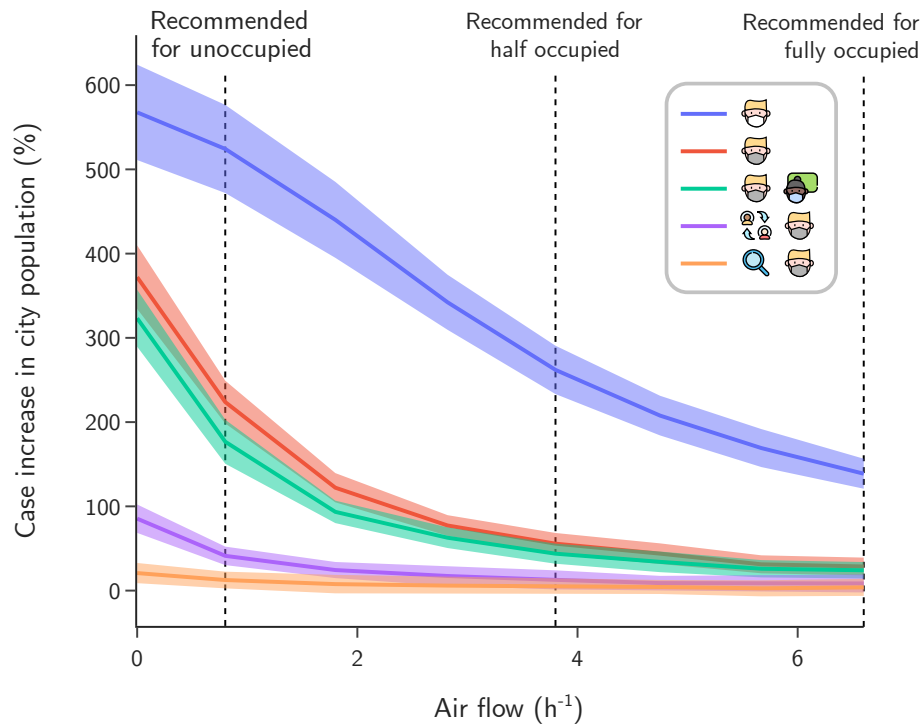


Figure 34 – **Sensitivity analysis across ventilation Λ .** Cases increase in school population (mean and standard deviation) as a function of classroom ventilation rate. Dashed lines indicate the recommended ventilation rates: $\Lambda_1 = 0.8 h^{-1}$ (unoccupied room), $\Lambda_2 = 3.8 h^{-1}$ (half occupied room), and $\Lambda_3 = 6.6 h^{-1}$ (fully occupied room), following the ASHRAE standard for an average classroom in Maragogi.

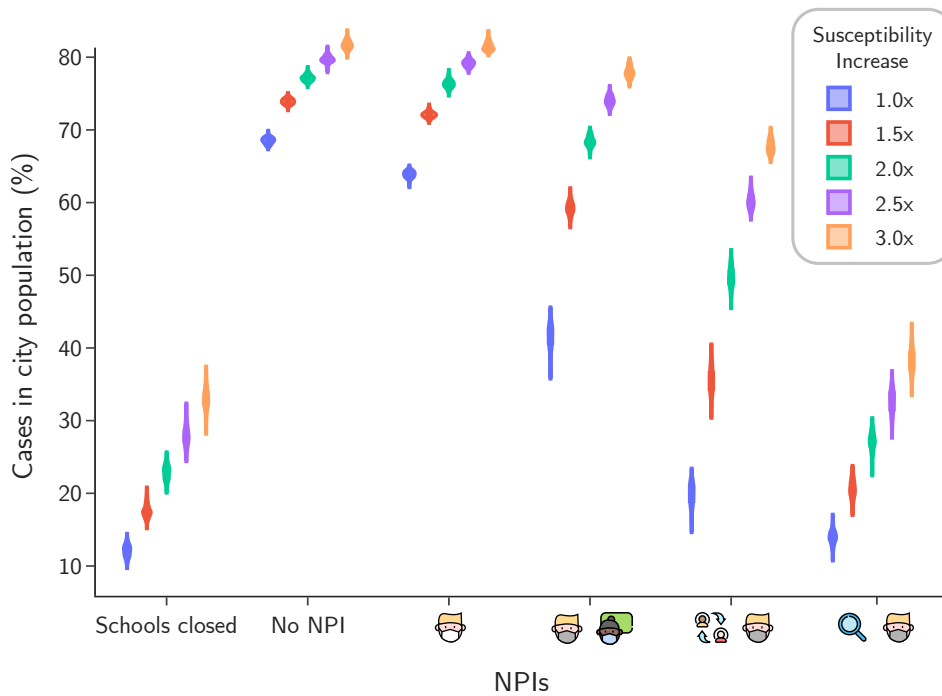


Figure 35 – **Population infected in case of increase in susceptibility.** For each intervention scenario, we show the distribution in the percentile of the population infected provided the susceptibility of the population is increased uniformly by a multiplying factor.

CONCLUSION

6.1 Conclusions on policy evaluation for schools

The airborne transmission mechanism of COVID-19 is the main cause of infections in school environments in classrooms with poor air circulation. Since many classrooms are equipped with air conditioning or heating, most have poor air circulation. Therefore, reducing the size of the class does not necessarily curb spread because an infected person can emit aerosols that stay in the air and infect students far away from the same classroom.

Vaccination of employees is an essential measure. However, in the absence of other measures, such as monitoring and quarantines, the number of cases in cities is likely to increase by 177% if only the use of low-quality masks and alternating classes is implemented.

The penetration factors provided by manufacturers and used in our simulations are idealized. In practice, the fit of a mask and the practices of users result in lower filtration efficacy. In fact, after testing a contagion model based on a study of Canadian classrooms ([HOU; KATAL; WANG, 2021](#)), we compared the ensuing results with our own aerosol model under the same class conditions but varying penetration factors, in order to estimate its value in these classrooms. We were alarmed to find that the effective penetration factor for Canadian classrooms in that study was only 0.5, despite the assumption of high-quality masks. It would therefore be of great benefit to educate the general population on proper mask use. Otherwise, the potential effectiveness of the sanitary protocols will be compromised as the penetration factor achieving increases (Fig. 33).

All these findings can be explained by three facts: teachers are more susceptible than children, they expel more virulent particles since they are constantly speaking loudly, and they are the most effective bridges of transmission between isolated classes. Therefore, high quality masks not only protect the individual teacher, but also suppress community infection.

Our most striking result is that one must adopt the appropriate NPIs and behavioral

protocols to safely continue school activities during a pandemic, regardless of vaccination coverage. Good protocols can protect countries even with poor vaccination coverage. In contrast, bad protocols can seriously aggravate the underlying public health crises even in countries with very high vaccination coverage. This is in great part due to the long duration of social contacts in schools, which easily leads to breakthrough infections without proper protocols. This is particularly relevant given that in many countries children are not routinely vaccinated for COVID-19, or when preparing for the emergence of new variants with potentially low cross immunity.

There is no single solution to a pandemic, but we draw hope in showing that the proper combination of NPIs, vaccination, and behaviors allows the safe continuation of activities as fundamental and important as teaching.

6.2 General conclusions

COMORBUSS, as an advanced bio-social agent model, has a large potential in the realm of epidemiological simulations. Its intricate design and the ability to simulate complex community dynamics and disease propagation provide an invaluable tool for public health research and policy making. The versatility of the model is demonstrated through its application to various scenarios, including the impact of COVID-19 in school environments, demonstrating its potential to inform and guide effective public health strategies.

A key strength of COMORBUSS lies in its dynamic nature, allowing for the organic representation of community behavior. This is achieved through a bottom-up approach, where complex interactions emerge from simple rules determining behavior. The flexibility of the model, with its ability to adapt to different community structures and behavioral patterns, makes it universally applicable, transcending geographical and demographic boundaries.

Incorporating non-pharmaceutical and pharmaceutical interventions in the model offers a comprehensive view of the potential outcomes of various public health strategies. This includes detailed simulations of quarantine protocols, social isolation measures, lockdowns, contact tracing, testing policies, and vaccination campaigns. Each intervention can be fine-tuned, providing a realistic representation of its implementation and effectiveness in different community settings.

We have verified the extensive and powerful applicability of agent models throughout this master's program. In particular in our own work to assess the effects of public health protocols during the COVID-19 pandemic, we anticipate the significant social impact of designing public health policies tailored to specific communities. At the same time, we recognize limitations and difficulties in the development and use of these types of model, many of which were addressed in our development of COMORBUSS. However, we still see the need for a better theoretical foundation for analytical methods and standardization of experimental frameworks, especially employing the language of causal inference (MARSHALL; GALEA, 2015). We shall dedicate

future works in this direction, along with other practical applications, such as optimizing vaccine efficacy tests or the evaluation of risk-benefit scenarios.

In conclusion, COMORBUSS stands as a testament to the power of interdisciplinary collaboration in tackling complex public health challenges. By bridging the gap between epidemiology, social sciences, and computational modeling, it offers a nuanced understanding of disease dynamics within communities. As we continue to face global health challenges, tools like COMORBUSS will be instrumental in shaping informed, effective, and context-sensitive public health policies.

BIBLIOGRAPHY

ANVISA. **Acurácia dos testes diagnósticos registrados na ANVISA para a COVID-19**. [S.l.], 2020. Citations on pages 71 and 72.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. Wasserstein generative adversarial networks. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 214–223. Citation on page 83.

ASHRAE 62.1-2019 - Ventilation for Acceptable Indoor Air Quality (ANSI Approved). 2021. <https://www.techstreet.com/ashrae/standards/ashrae-62-1-2019?product_id=2088533>. Accessed: 2022-09-29. Citation on page 102.

ASSENTAMENTO. 2021. <<https://antigo.incra.gov.br/pt/assentamentos.html>>. Accessed: 2021-08-24. Citation on page 25.

BAZANT, M. Z.; BUSH, J. W. M. A guideline to limit indoor airborne transmission of covid-19. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 118, n. 17, 2021. ISSN 0027-8424. Available: <<https://www.pnas.org/content/118/17/e2018995118>>. Citations on pages 98, 115, 116, 118, 119, and 120.

BAZANT, M. Z.; KODIO, O.; COHEN, A. E.; KHAN, K.; GU, Z.; BUSH, J. W. M. Monitoring carbon dioxide to quantify the risk of indoor airborne transmission of covid-19. **medRxiv**, Cold Spring Harbor Laboratory Press, 2021. Available: <<https://www.medrxiv.org/content/early/2021/04/16/2021.04.04.21254903>>. Citations on pages 115, 116, and 118.

BERSHTEYN, A.; GERARDIN, J.; BRIDENBECKER, D.; LORTON, C. W.; BLOEDOW, J.; BAKER, R. S.; CHABOT-COUTURE, G.; CHEN, Y.; FISCHLE, T.; FREY, K. *et al.* Implementation and applications of emod, an individual-based multi-disease modeling platform. **Pathogens and disease**, Oxford University Press, v. 76, n. 5, p. fty059, 2018. Citation on page 29.

BIBBONA, E.; SIROVICH, R. Strong approximation of density dependent markov chains on bounded domains. **arXiv preprint arXiv:1704.07481**, 2017. Citation on page 85.

BUONANNO, G.; MORAWSKA, L.; STABILE, L. Quantitative assessment of the risk of airborne transmission of sars-cov-2 infection: Prospective and retrospective applications. **Environment International**, v. 145, p. 106112, 2020. ISSN 0160-4120. Available: <<https://www.sciencedirect.com/science/article/pii/S0160412020320675>>. Citation on page 119.

CHEN, C.; WILLEKE, K. Characteristics of face seal leakage in filtering facepieces. **American Industrial Hygiene Association Journal**, Taylor & Francis, v. 53, n. 9, p. 533–539, 1992. Pmid: 1524028. Available: <<https://doi.org/10.1080/15298669291360120>>. Citation on page 117.

CHENG, Y.; MA, N.; WITT, C.; RAPP, S.; WILD, P. S.; ANDREAE, M. O.; PÖSCHL, U.; SU, H. Face masks effectively limit the probability of sars-cov-2 transmission. **Science**, American Association for the Advancement of Science, v. 372, n. 6549, p. 1439–1443, 2021. ISSN 0036-8075. Available: <<https://science.sciencemag.org/content/372/6549/1439>>. Citation on page 100.

CNES-HEALTH Professionals Data. 2020. <<http://cnes.datasus.gov.br/pages/profissionais/extracao.jsp>>. Accessed: 2020-11-12. Citation on page 125.

COELHO, L. E.; LUZ, P. M.; PIRES, D. C.; JALIL, E. M.; PERAZZO, H.; TORRES, T. S.; CARDOSO, S. W.; PEIXOTO, E. M.; NAZER, S.; MASSAD, E.; SILVEIRA, M. F.; BARROS, F. C.; VASCONCELOS, A. T.; COSTA, C. A.; AMANCIO, R. T.; VILLELA, D. A.; PEREIRA, T.; GOEDERT, G. T.; SANTOS, C. V.; RODRIGUES, N. C.; GRINSZTEJN, B.; VELOSO, V. G.; STRUCHINER, C. J. Prevalence and predictors of anti-sars-cov-2 serology in a highly vulnerable population of rio de janeiro: A population-based serosurvey. **The Lancet Regional Health - Americas**, n. 15, p. 100338, 2022. Citation on page 19.

COVID-19 in children and the role of school settings in transmission - second update. 2021. <<https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-in-children-and-the-role-of-school-settings-in-transmission-second-update.pdf>>. Accessed: 2021-08-17. Citation on page 89.

CURMEI, M.; ILYAS, A.; EVANS, O.; STEINHARDT, J. Estimating household transmission of sars-cov-2. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Available: <<https://www.medrxiv.org/content/early/2020/06/27/2020.05.23.20111559>>. Citations on pages 42 and 76.

DEDECKER, J.; MERLEVÈDE, F. Behavior of the empirical wasserstein distance in \mathbb{R}^d under moment conditions. **Electronic Journal of Probability**, Institute of Mathematical Statistics and Bernoulli Society, v. 24, p. 1–32, 2019. Citation on page 83.

Del Valle, S.; HYMAN, J.; HETHCOTE, H.; EUBANK, S. Mixing patterns between age groups in social networks. **Social Networks**, v. 29, n. 4, p. 539–554, 2007. ISSN 0378-8733. Available: <<https://www.sciencedirect.com/science/article/pii/S0378873307000330>>. Citation on page 47.

DRONE Deploy mapping software. 2020. <<https://www.dronedeploy.com/>>. Accessed: 2021-09-22. Citation on page 91.

FLAMARY, R.; COURTY, N.; GRAMFORT, A.; ALAYA, M. Z.; BOISBUNON, A.; CHAMBON, S.; CHAPEL, L.; CORENFLOS, A.; FATRAS, K.; FOURNIER, N.; GAUTHERON, L.; GAYRAUD, N. T.; JANATI, H.; RAKOTOMAMONJY, A.; REDKO, I.; ROLET, A.; SCHUTZ, A.; SEGUY, V.; SUTHERLAND, D. J.; TAVENARD, R.; TONG, A.; VAYER, T. Pot: Python optimal transport. **Journal of Machine Learning Research**, v. 22, n. 78, p. 1–8, 2021. Available: <<http://jmlr.org/papers/v22/20-451.html>>. Citation on page 82.

GENARI, J.; GOEDERT, G. T.; LIRA, S. H.; OLIVEIRA, K.; BARBOSA, A.; LIMA, A.; SILVA, J. A.; OLIVEIRA, H.; MACIEL, M.; LEDOINO, I. *et al.* Quantifying protocols for safe school activities. **Plos one**, Public Library of Science San Francisco, CA USA, v. 17, n. 9, p. e0273425, 2022. Citations on pages 20 and 33.

GREENWOOD, P. E.; GORDILLO, L. F. Stochastic epidemic modeling. In: **Mathematical and statistical estimation approaches in epidemiology**. [S.l.]: Springer, 2009. p. 31–52. Citation on page 85.

GURDASANI, D.; ALWAN, N. A.; GREENHALGH, T.; HYDE, Z.; JOHNSON, L.; MCKEE, M.; MICHIE, S.; PRATHER, K. A.; RASMUSSEN, S. D.; REICHER, S.; RODERICK, P.; ZIAUDDEEN, H. School reopening without robust covid-19 mitigation risks accelerating the pandemic. **The Lancet**, Elsevier, v. 397, n. 10280, p. 1177–1178, 2021. ISSN 0140-6736. Available: <[https://doi.org/10.1016/S0140-6736\(21\)00622-X](https://doi.org/10.1016/S0140-6736(21)00622-X)>. Citation on page 89.

HOU, D.; KATAL, A.; WANG, L. L. Bayesian calibration of using co2 sensors to assess ventilation conditions and associated covid-19 airborne aerosol transmission risk in schools. **medRxiv**, Cold Spring Harbor Laboratory Press, , 2021. Available: <<https://www.medrxiv.org/content/early/2021/02/03/2021.01.29.21250791>>. Citation on page 105.

HUNTER, E.; NAMEE, B. M.; KELLEHER, J. D. A taxonomy for agent-based models in human infectious disease epidemiology. **Journal of Artificial Societies and Social Simulation**, Jasss, v. 20, n. 3, 2017. Citation on page 28.

IBGE- SIDRA tabela 6450. 2019. <<https://sidra.ibge.gov.br/Tabela/6450>>. Accessed: 2020-11-12. Citations on pages 124 and 125.

IBGE 2019 population size estimate. 2019. <https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2019/estimativa_dou_2019.pdf>. Accessed: 2021-12-15. Citation on page 22.

IBGE census 2010. 2010. <<https://censo2010.ibge.gov.br/>>. Accessed: 2021-05-12. Citations on pages 9 and 23.

IBGE panorama - Maragogi-AL. 2021. <<https://cidades.ibge.gov.br/brasil/al/maragogi/panorama>>. Accessed: 2021-12-15. Citations on pages 22 and 90.

INCOGNIA company. 2021. <<https://www.incognia.com/pt/a-marca-inloco-agora-e-incognia>>. Accessed: 2020-11-12. Citations on pages 76 and 95.

INCRA - Beneficiários. 2021. <<https://saladacidadania.incra.gov.br/Beneficiario/ConsultaPublica>>. Accessed: 2021-08-24. Citation on page 25.

INEP - dados abertos. 2020. <<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/inep-data>>. Accessed: 2020-11-12. Citation on page 125.

INEP, Students per class. 2020. <<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais/media-de-alunos-por-turma>>. Accessed: 2021-05-12. Citations on pages 25 and 125.

KASLOW, D. C. Force of infection: a determinant of vaccine efficacy? **.npj Vaccines**, v. 6, 2021. Available: <<https://www.nature.com/articles/s41541-021-00316-5>>. Citation on page 96.

KERR, C. C.; STUART, R. M.; MISTRY, D.; ABEYSURIYA, R. G.; HART, G.; ROSENFELD, K.; SELVARAJ, P.; NUNEZ, R. C.; HAGEDORN, B.; GEORGE, L. *et al.* Covasim: an agent-based model of covid-19 dynamics and interventions. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Citations on pages 20, 32, 55, 71, 74, and 124.

KHOVALYG, D.; KAZANCI, O. B.; HALVORSEN, H.; GUNDLACH, I.; BAHNFLETH, W. P.; TOFTUM, J.; OLESEN, B. W. Critical review of standards for indoor thermal environment and air quality. **Energy and Buildings**, v. 213, p. 109819, 2020. ISSN 0378-7788. Available: <<https://www.sciencedirect.com/science/article/pii/S0378778819314719>>. Citations on pages 115 and 120.

KISIELINSKI, K.; GIBONI, P.; PRESCHER, A.; KLOSTERHALFEN, B.; GRAESSEL, D.; FUNKEN, S.; KEMPSKI, O.; HIRSCH, O. Is a mask that covers the mouth and nose free from undesirable side effects in everyday use and free of potential hazards? **International Journal of Environmental Research and Public Health**, v. 18, n. 8, 2021. ISSN 1660-4601. Available: <<https://www.mdpi.com/1660-4601/18/8/4344>>. Citation on page 100.

LEQUIME, S.; BASTIDE, P.; DELLICOUR, S.; LEMEY, P.; BAELE, G. nosoi: A stochastic agent-based transmission chain simulation framework in r. **Methods in ecology and evolution**, Wiley Online Library, v. 11, n. 8, p. 1002–1007, 2020. Citation on page 31.

LEQUIME, S.; DELLICOUR, S. **Spread of an Ebola-like virus in continuous space**. 2021. <<https://slequime.github.io/nosoi/articles/examples/ebola.html>>. Accessed: 2023-12-26. Citation on page 31.

LESSLER, J.; GRABOWSKI, M. K.; GRANTZ, K. H.; BADILLO-GOICOECHEA, E.; METCALF, C. J. E.; LUPTON-SMITH, C.; AZMAN, A. S.; STUART, E. A. Household covid-19 risk and in-person schooling. **Science**, American Association for the Advancement of Science, v. 372, n. 6546, p. 1092–1097, 2021. ISSN 0036-8075. Available: <<https://science.sciencemag.org/content/372/6546/1092>>. Citation on page 89.

LINTON, N. M.; KOBAYASHI, T.; YANG, Y.; HAYASHI, K.; AKHMETZHANOV, A. R.; JUNG, S.-m.; YUAN, B.; KINOSHITA, R.; NISHIURA, H. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. **Journal of Clinical Medicine**, v. 9, n. 2, 2020. ISSN 2077-0383. Available: <<https://www.mdpi.com/2077-0383/9/2/538>>. Citation on page 94.

MADEWELL, Z. J.; DEAN, N. E.; BERLIN, J. A.; COPLAN, P. M.; DAVIS, K. J.; STRUCHINER, C. J.; HALLORAN, M. E. Challenges of evaluating and modelling vaccination in emerging infectious diseases. **Epidemics**, v. 37, p. 100506, 2021. ISSN 1755-4365. Available: <<https://www.sciencedirect.com/science/article/pii/S1755436521000554>>. Citation on page 96.

MARSHALL, B. D.; GALEA, S. Formalizing the role of agent-based modeling in causal inference and epidemiology. **American journal of epidemiology**, Oxford University Press, v. 181, n. 2, p. 92–99, 2015. Citation on page 106.

MELLAN, T. A.; HOELTGEBAUM, H. H.; MISHRA, S.; WHITTAKER, C.; SCHNEKENBERG, R. P.; GANDY, A.; UNWIN, H. J. T.; VOLLMER, M. A.; COUPLAND, H.; HAWRYLUK, I. *et al.* Report 21: Estimating covid-19 cases and reproduction number in brazil. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Citation on page 123.

MILLER, S. L.; NAZAROFF, W. W.; JIMENEZ, J. L.; BOERSTRA, A.; BUONANNO, G.; DANCER, S. J.; KURNITSKI, J.; MARR, L. C.; MORAWSKA, L.; NOAKES, C. Transmission of sars-cov-2 by inhalation of respiratory aerosol in the skagit valley chorale superspreading event. **Indoor Air**, v. 31, n. 2, p. 314–323, 2021. Available: <<https://onlinelibrary.wiley.com/doi/abs/10.1111/ina.12751>>. Citations on pages 98 and 99.

MORAWSKA, L.; CAO, J. Airborne transmission of sars-cov-2: The world should face the reality. **Environment International**, v. 139, p. 105730, 2020. ISSN 0160-4120. Available: <<https://www.sciencedirect.com/science/article/pii/S016041202031254X>>. Citation on page 98.

MORAWSKA, L.; JOHNSON, G.; RISTOVSKI, Z.; HARGREAVES, M.; MENGERSEN, K.; CORBETT, S.; CHAO, C.; LI, Y.; KATOSHEVSKI, D. Size distribution and sites of origin of droplets expelled from the human respiratory tract during expiratory activities. **Journal of Aerosol Science**, v. 40, n. 3, p. 256–269, 2009. ISSN 0021-8502. Available: <<https://www.sciencedirect.com/science/article/pii/S0021850208002036>>. Citations on pages 98, 115, and 120.

MORFELD, P.; TIMMERMANN, B.; GROSS, J. V.; LEWIS, P.; COCCO, P.; ERREN, T. C. Covid-19: Heterogeneous excess mortality and “burden of disease” in germany and italy and their states and regions, january–june 2020. **Frontiers in Public Health**, Frontiers Media SA, v. 9, p. 663259, 2021. Citation on page 19.

MUNDAY, J. D.; JARVIS, C. I.; GIMMA, A.; WONG, K. L.; ZANDVOORT, K. v.; GROUP, C. C.-. W.; FUNK, S.; EDMUNDS, W. J. Estimating the impact of reopening schools on the reproduction number of sars-cov-2 in england, using weekly contact survey data. **medRxiv**, Cold Spring Harbor Laboratory Press, , 2021. Available: <<https://www.medrxiv.org/content/early/2021/03/08/2021.03.06.21252964>>. Citation on page 89.

NISHIURA, H.; OSHITANI, H.; KOBAYASHI, T.; SAITO, T.; SUNAGAWA, T.; MATSUI, T.; WAKITA, T.; SUZUKI, M. Closed environments facilitate secondary transmission of coronavirus disease 2019 (covid-19). **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Available: <<https://www.medrxiv.org/content/early/2020/04/16/2020.02.28.20029272>>. Citation on page 76.

OECD urban population size. 2020. <<https://data.oecd.org/popregion/urban-population-by-city-size.htm>>. Accessed: 2021-11-15. Citation on page 23.

OPENDATASUS - SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19. 2020. <<https://opendatasus.saude.gov.br/dataset/srag-2020>>. Accessed: 2021-08-12. Citation on page 123.

PATLOLLA, P.; GUNUPUDI, V.; MIKLER, A. R.; JACOB, R. T. Agent-based simulation tools in computational epidemiology. In: SPRINGER. **Innovative Internet Community Systems: 4th International Workshop, IICS 2004, Guadalajara, Mexico, June 21-23, 2004. Revised Papers 4**. [S.l.], 2006. p. 212–223. Citation on page 28.

POYDENOT, F.; ABDOURAHAMANE, I.; CAPLAIN, E.; DER, S.; HAIECH, J.; JALLON, A.; KHOUTAMI, I.; LOUCIF, A.; MARINOV, E.; ANDREOTTI, B. Risk assessment for long and short range airborne transmission of sars-cov-2, indoors and outdoors, using carbon dioxide measurements. **medRxiv**, Cold Spring Harbor Laboratory Press, 2021. Available: <<https://www.medrxiv.org/content/early/2021/05/07/2021.05.04.21256352>>. Citations on pages 98, 99, and 119.

PREM, K.; COOK, A. R.; JIT, M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 13, n. 9, p. e1005697, 2017. Citation on page 20.

REGULY, I. Z.; CSERCSIK, D.; JUHÁSZ, J.; TORNAI, K.; BUJTÁR, Z.; HORVÁTH, G.; KEÖMLEY-HORVÁTH, B.; KÓS, T.; CSEREY, G.; IVÁN, K. *et al.* Microsimulation based quantitative analysis of covid-19 management strategies. **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 18, n. 1, p. e1009693, 2022. Citation on page 30.

RILEY, E. C.; MURPHY, G.; RILEY, R. L. Airborne spread of measles in a suburban elementary school. **American Journal of Epidemiology**, v. 107, n. 5, p. 421–432, 05 1978. ISSN 0002-9262. Available: <<https://doi.org/10.1093/oxfordjournals.aje.a112560>>. Citation on page 98.

SICAR - Cadastro Ambiental Rural (CAR). 2021. <<https://www.car.gov.br/publico/imoveis/index>>. Accessed: 2021-08-24. Citation on page 25.

SILVA, P. J. S.; SAGASTIZÁBAL, C.; NONATO, L. G.; STRUCHINER, C. J.; PEREIRA, T. Optimized delay of the second covid-19 vaccine dose reduces icu admissions. **Proceedings of the National Academy of Sciences of the United States of America**, v. 118, n. 35, 2021. Citation on page 90.

SOMMERFELD, M.; MUNK, A. Inference for empirical wasserstein distances on finite spaces. **arXiv preprint arXiv:1610.03287**, 2016. Citation on page 83.

STRUCHINER, C. J.; HALLORAN, M. E. Randomization and baseline transmission in vaccine field trials. **Epidemiology and Infection**, Cambridge University Press, v. 135, n. 2, p. 181–194, 2007. Citation on page 96.

THOMPSON, R. N.; AL., C. J. S. et. Key questions for modelling covid-19 exit strategies. **Proceedings of the Royal Society B**, v. 287, 2020. Citation on page 89.

UNITED States Cities Database. 2021. <<https://simplemaps.com/data/us-cities>>. Accessed: 2021-09-06. Citations on pages 9, 23, 24, and 90.

VERITY, R.; OKELL, L. C.; DORIGATTI, I.; WINSKILL, P.; WHITTAKER, C.; IMAI, N.; CUOMO-DANNENBURG, G.; THOMPSON, H.; WALKER, P. G. T.; FU, H.; DIGHE, A.; GRIFFIN, J. T.; BAGUELIN, M.; BHATIA, S.; BOONYASIRI, A.; CORI, A.; CUCUNUBÁ, Z.; FITZJOHN, R.; GAYTHORPE, K.; GREEN, W.; HAMLET, A.; HINSLEY, W.; LAYDON, D.; NEDJATI-GILANI, G.; RILEY, S.; ELSLAND, S. van; VOLZ, E.; WANG, H.; WANG, Y.; XI, X.; DONNELLY, C. A.; GHANI, A. C.; FERGUSON, N. M. Estimates of the severity of coronavirus disease 2019: a model-based analysis. **The Lancet Infectious Diseases**, Elsevier, v. 20, n. 6, p. 669–677, Jun 2020. ISSN 1473-3099. Available: <[https://doi.org/10.1016/S1473-3099\(20\)30243-7](https://doi.org/10.1016/S1473-3099(20)30243-7)>. Citation on page 94.

WILLEM, L.; VERELST, F.; BILCKE, J.; HENS, N.; BEUTELS, P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). **BMC infectious diseases**, Springer, v. 17, p. 1–16, 2017. Citation on page 28.

WÖLFEL, R.; CORMAN, V. M.; GUGGEMOS, W.; SEILMAIER, M.; ZANGE, S.; MÜLLER, M. A.; NIEMEYER, D.; JONES, T. C.; VOLLMAR, P.; ROTHE, C.; HOELSCHER, M.; BLEICKER, T.; BRÜNINK, S.; SCHNEIDER, J.; EHMANN, R.; ZWIRGLMAIER, K.; DROSTEN, C.; WENDTNER, C. Virological assessment of hospitalized patients with covid-2019. **Nature**, v. 581, n. 7809, p. 465–469, May 2020. ISSN 1476-4687. Available: <<https://doi.org/10.1038/s41586-020-2196-x>>. Citation on page 92.

WORLD Cities Database. 2021. <<https://simplemaps.com/data/world-cities>>. Accessed: 2021-09-06. Citations on pages 9, 23, 24, and 90.

ZHANG, J.; LITVINOVA, M.; LIANG, Y.; WANG, Y.; WANG, W.; ZHAO, S.; WU, Q.; MERLER, S.; VIBOUD, C.; VESPIGNANI, A.; AJELLI, M.; YU, H. Changes in contact patterns shape the dynamics of the covid-19 outbreak in china. **Science**, American Association for the Advancement of Science, v. 368, n. 6498, p. 1481–1486, 2020. ISSN 0036-8075. Available: <<https://science.sciencemag.org/content/368/6498/1481>>. Citations on pages 99, 119, and 120.

AIRBORNE TRANSMISSION MODEL

A.1 Aerosol-based model for infections in a closed environment

A.1.1 *Relevant length and time scales for aerosol particles*

Pathogen-carrying aerosol particles are expelled by infected individuals in a range of radii ranging from $0.1 \mu\text{m}$ to 1mm . Most of these particles are in the sub-micrometer scale, and the size distribution of the droplets depends on the breathing activity, varying from $0.1 \mu\text{m}$ to $5.0 \mu\text{m}$ with a peak around $0.5 \mu\text{m}$ (MORAWSKA *et al.*, 2009).

Pathogens carried by airborne droplets have a typical lifetime inside the enclosed space, so we consider the damping rate of the pathogen concentration λ_c . This rate depends on the radius r of airborne droplets (BAZANT; BUSH, 2021; BAZANT *et al.*, 2021), and it encompasses four distinct mechanisms

$$\lambda_c(r) = \lambda_a + \lambda_f(r) + \lambda_s(r) + \lambda_v(r), \quad (\text{A.1})$$

where λ_a accounts for outdoor air exchange rate, λ_f is the room filtration rate (filtration due to mechanical ventilation or people breathing in the room and absorbing infectious airborne particles), λ_s is the net sedimentation rate, and λ_v stands for the deactivation rate of the aerosolized pathogen (which depends on humidity and droplet size).

Although the definition of air quality inside enclosed space varies over international standards (KHOVALYG *et al.*, 2020), ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) described in the technical notes¹ that the minimum recommended outdoor air exchange rate depends on the environment. Namely, for American homes $\lambda_a = 0.35 \text{h}^{-1}$, while for classrooms of children aged 5 to 9 years $\lambda_a = 0.8 \text{h}^{-1}$. Those are the minimal

¹ ASHRAE 62.1 — Ventilation for Acceptable Indoor Air Quality.

recommended values and as we will see correspond to the largest order of magnitude among all other terms in Equation (A.1).

For most air conditioning systems in Brazil, a filtration system is absent and is not coupled to the mechanical ventilation. However, in our model we assume that aerosol consumption arises from people breathing in the classroom and filtering air in their respiratory system. Therefore, we consider that the filtration rate can be estimated by $\lambda_f = NB/V$, where N is the number of people in the room, B is the average breathing rate, and V is the volume of the classroom. We consider the values of $B = 0.5 \text{ m}^3/h$, $V = 150 \text{ m}^3$ (average volume of Maragogi classrooms) and $N = 20$, which yields $\lambda_f = 0.07 \text{ h}^{-1}$.

The size of the droplets determines the sedimentation rate λ_s . For droplets larger than a critical radius $r > r_c$, the sedimentation rate due to gravity is high and contributes significantly to λ_c . Hereafter, we consider airborne transmission as that associated with droplets with radius $r < r_c$, since those droplets remain suspended in the air for long periods of time (typically a few hours in a closed classroom) and contain viral loads capable of producing long-range airborne transmission. The realistic values for r_c range from $1.3 \text{ }\mu\text{m}$ to $5.5 \text{ }\mu\text{m}$ (BAZANT; BUSH, 2021).

The sedimentation rate (drop settling rate) is given by $\lambda_s = \bar{v}_s(\bar{r})/H$, where H is the height of the enclosed space. Fixing the sedimentation velocity $\bar{v}_s = 0.108 \text{ m/h}$ (BAZANT *et al.*, 2021) (the effective radius of respiratory drop is $\bar{r} = 0.5 \text{ }\mu\text{m}$), and the height H of the Maragogi classrooms being in the range of $2.57 - 2.85 \text{ m}$, we estimate that λ_s lies in the interval $0.038 - 0.042 \text{ h}^{-1}$. Therefore, for biologically relevant droplets of submicrometer radius, settling can be safely neglected (BAZANT; BUSH, 2021).

In the following section, we will closely follow (BAZANT; BUSH, 2021; BAZANT *et al.*, 2021), and assume a size-dependent sedimentation rate $\lambda_s(r) = v_s(r)/H = \lambda_a(r/r_c)^2$ as the inverse of the time taken for a drop in radius r to sediment from ceiling to floor in a quiescent room. Hence, Bazant and co-authors propose that for the relevant droplet size range in consideration, one may write

$$\lambda_c(r) = \lambda_a \left[1 + \left(\frac{r}{r_c} \right)^2 \right] + \lambda_v(r) + \lambda_f(r). \quad (\text{A.2})$$

The viral deactivation rate (non-infectious) $\lambda_v(r)$ depends on the droplet radius and other quantities, such as temperature and humidity. Therefore, by aggregating the data of influenza viruses, we can extrapolate a linear relationship between relative humidity in the environment RH for SARS-CoV-2 (BAZANT; BUSH, 2021). We adopted $\lambda_v = 0.6RH \text{ h}^{-1}$ (since Maragogi is a coastal tropical city, RH can be a significantly high factor).

A.1.2 Time-evolution of radius-resolved particle concentration

We assume that the air is well mixed in the room to evaluate the concentration of infectious airborne pathogens dependent on time suspended in a classroom of volume V occupied

by N individuals, I infected and $N - I$ susceptible individuals. Following Bazant and Bush (2021), we assume that the radius-resolved concentration of infectious aerosol-borne pathogen in a classroom with well-mixed air conditions evolves according to

$$V \frac{\partial c(r,t)}{\partial t} = \sum_{j=1}^I P_j(r,t) - V \lambda_c(r,t) c(r,t), \quad (\text{A.3})$$

where $c(r,t)$ is the number-density of virion particles in the room carried by aerosol droplets with radius r (given in virions per volume per radius), $P_j(r,t)$ is the pathogen production rate due to respiratory activity of a given infectious individual j in the room, and $\lambda_c(r,t)$ is the pathogen concentration relaxation rate.

The production term of a single infectious individual is given by

$$P_j(r,t) = B_j(t) p_m^j(r) q_j(r,t), \quad (\text{A.4})$$

where $B_j(t)$ is the individual breathing rate, $p_m^j(r)$ is the mask penetration factor of droplets of radius r , and $q_j(r,t)$ is an activity dependent concentration of exhaled virions in droplets of radius r (number of virions per volume of air per radius of droplet). Moreover, we may specify that for each infectious individual $q_j(r,t) = n_d^j(r,t) V_d(r) c_v(r)$, where $n_d^j(r,t)$ is the size distribution of the emitted droplets (number density of the expelled droplets of radius r), $V_d(r) = 4\pi r^3/3$ is the volume of the droplet, and $c_v(r)$ is a microscopic viral concentration (concentration of virions per volume of the droplet).

We point out that infected individuals emit virions in droplets with a given size distribution that quickly evolves (in a time scale shorter than one second) to a stationary profile $q(r)$ that can be suspended in the air for longer time (for minutes or hours). Therefore, for the relevant contagion time scale in a closed room (from minutes to hours), the production term P in Equation (A.4) is time independent under a constant breathing rate B . Moreover, we also assume $\lambda_c(r,t) = \lambda_c(r)$ for steady ventilation conditions.

For simplicity, we assume that the average breathing rate for students and teachers is a constant value B regardless of their activity. The mask penetration factor $p_m(r)$ lies in the unit interval $[0, 1]$ - so it might be associated with a probability that a particle will penetrate the mask tissue - and depends on the droplet size distribution. Based on experimental observations (CHEN; WILLEKE, 1992), from now on we assume that the mask penetration factor is approximately constant in this submicrometer size range and evaluate $\overline{p_m} = p_m(\bar{r})$ at an effective aerosol radius \bar{r} to be defined below in Equation (A.11).

Consider that at $t = 0$, N individuals enter a room of volume V and zero initial concentration of airborne viral particles, $c_0(r) = c(r, t = 0) = 0$. These individuals wear masks with equal penetration factor $\overline{p_m}$ and only one individual is infectious among them. They remain in the room for a given period of time τ , maintaining constant respiratory activity (breathing and

talking). The time evolution of the radius-resolved concentration is given by

$$\frac{1}{\lambda_c(r)} \frac{\partial c(r,t)}{\partial t} = \frac{P(r)}{V\lambda_c(r)} - c(r,t), \quad (\text{A.5})$$

which can be integrated to

$$c(r,t) = c_0(r)e^{-\lambda_c(r)t} + \frac{P(r)}{V\lambda_c(r)} [1 - e^{-\lambda_c(r)t}], \quad (\text{A.6})$$

where $P(r) = B \bar{p}_m q(r)$ and $\lambda_c(r) > 0$ for the relevant range of droplet size.

The probability of a susceptible person to be infected in the room depends not only on the total number of virions inhaled, but also on the power of a virion to cause an infection when it carries a droplet of a given radius r . Therefore, we define the infectious dose inhaled by an individual exposed to the room from $t = 0$ to $t = \tau$ as

$$D(\tau) = \int_0^\tau dt \int_0^\infty dr B p_m(r) c(r,t) i(r), \quad (\text{A.7})$$

where $i(r)$ is the infectivity of the aerosolized pathogen in a droplet of radius r . $i(r)$ can be interpreted as proportional to the probability that a single virion causes an infection in a susceptible person when it is inhaled in a droplet of radius r (in Refs. (BAZANT; BUSH, 2021; BAZANT *et al.*, 2021), $i(r)$ is equivalent to $c_i(r)$).

The transient term in Equation (A.6) vanishes after long exposition times $\tau \gg \lambda_c^{-1}$. In this condition we have the following linear dependence of the inhaled dose with τ ,

$$D(\tau) \approx \frac{B^2}{V \bar{p}_m^2} \tau \int_0^\infty dr \frac{q(r)i(r)}{\lambda_c(r)} = \frac{B^2}{V \bar{p}_m^2} \tau \frac{C_q}{\bar{\lambda}_c}, \quad (\text{A.8})$$

where as in (BAZANT; BUSH, 2021) we have defined

$$C_q \equiv \int_0^\infty dr q(r)i(r), \quad (\text{A.9})$$

$$\bar{\lambda}_c^{-1} \equiv \frac{\int_0^\infty dr q(r)i(r)\lambda_c(r)^{-1}}{\int_0^\infty dr q(r)i(r)}. \quad (\text{A.10})$$

Moreover, the effective infectious drop radius \bar{r} can now be chosen such that

$$\lambda_c(r = \bar{r}) = \bar{\lambda}_c. \quad (\text{A.11})$$

The realistic physical parameters give us a range of $\bar{r} = 0.3 - 5 \mu m$. Bazant and co-authors (BAZANT; BUSH, 2021) have used $\bar{r} = 2 \mu m$ to fit data from super-spreading events and the Wuhan outbreak; to monitor air quality indoors Ref. (BAZANT *et al.*, 2021) uses $\bar{r} = 0.5 \mu m$ for a closed space.

We consider that the probability $p(\tau)$ of a susceptible individual to be infected when inhaling a given aerosolized pathogen dose $D(\tau)$ is given by the exponential distribution (Wells-Riley model)

$$p(\tau) = 1 - e^{-s_r D(\tau)}, \quad (\text{A.12})$$

where s_r is the age-dependent relative susceptibility of infection (an age-based measure (ZHANG *et al.*, 2020)) for a person. This expression follows from the simplest assumption that any infectious viral particle can trigger an infection by independent action of all inhaled viral particles, leading to a Poisson process (POYDENOT *et al.*, 2021). For low-dose inhalation, $D \ll 1$, the probability can be approximated by $p(\tau) \approx s_r D(\tau)$. This result is equivalent to the probability calculated for the school safety guidelines in (BAZANT; BUSH, 2021).

A.1.3 Effective airborne transmission

For our epidemiological model, it suffices to estimate the mean infectious viral load concentration of exhaled air C_q defined in Eq. (A.9). We will consider $q(r,t) = q(r)$ for any infectious individual in a room, so C_q is a time independent constant that represents its average concentration of exhaled “quanta”, depending on its respiratory activity. C_q is typically expressed in units of quanta per volume of air and represents the important epidemiological parameter that can be numerically estimated based on real outbreak data.

Infectivity $i(r)$ (quanta RNA copies⁻¹) represents the probability that a pathogen surviving inside the host will initiate infection, or we can interpret taking the inverse of infectivity i^{-1} , which corresponds to the “infectious dose” of pathogens from aerosol droplets inhaled that cause infection with probability $1 - (1/e) = 63\%$.

To convert the infectious dose quantified in terms of RNA copies to infectious quanta (which is the measure we use in our model), two parameters must be known a priori: i) the number of infectious particles (RNA copies) needed to initiate infection (c_{RNA} , RNA copies PFU⁻¹), and (ii) the conversion parameter quanta to plate-forming unit (PFU) (c_{PFU} , PFU quanta⁻¹). Hence, the expression for determining $i(r)$ is

$$i(r) = \frac{1}{c_{RNA}(r) c_{PFU}(r)}.$$

Currently there are no c_{PFU} values available for SARS-CoV-2 in the scientific literature for this value (BUONANNO; MORAWSKA; STABILE, 2020), or characterization of size-dependent distributions $q(r)$, $n_d(r)$ and $c_v(r)$. Therefore, we estimate the adopting values for SARS-CoV-1. On the other hand, the parameter c_{RNA} has been estimated to be 1.3×10^2 RNA copies PFU⁻¹.

Equation (A.9) implies that we should be able to characterize the concentration of virions suspended in the air on droplets of all sizes that are capable of causing an infection. Therefore, we define the total concentration of infectious aerosolized virions per volume of air as

$$C(t) = \int_0^\infty c(r,t) i(r) dr, \quad (\text{A.13})$$

where C is given in units of quanta per volume of air. Multiplying Equation (A.3) by $i(r)$ and integrating for all r one derives

$$V \frac{dC}{dt} = -(\Lambda V + NB)C + B(C_s N_s p_m^s + C_t N_t p_m^t), \quad (\text{A.14})$$

where C is the quanta per unit of volume of air in the room, $\Lambda + NB/V = \bar{\lambda}_c$ is the effective rate of relaxation of quanta concentration, $p_m = \bar{p}_m = p_m(\bar{r})$ is the effective mask penetration factor. We consider the masks of the teachers to have p_m^t , and the masks of the students to be present p_m^s . The effective radius \bar{r} for relevant infectious aerosol droplets is given by Eq. (A.11), where we make the following approximation

$$\int_0^\infty \lambda_c(r) c(r, t) i(r) dr \approx \bar{\lambda}_c C(t).$$

We consider in Equation (A.14) that the volume V classroom is occupied by N individuals, in which S are susceptible, N_s are infected students and N_t are infected teachers. Each person exchanges air masses with the environment at an average breathing rate B , inhales a $C(t)$ quanta concentration, and exhales a different concentration. We introduce heterogeneity in the concentration of quanta expelled by students and teachers, assuming that they perform different breathing activities (BAZANT; BUSH, 2021): $C_s = 40$ (quanta/m³) is the concentration of quanta expelled from students such that $C_q^{students} = C_s$, and $C_t = 72$ (quanta/m³) denotes the concentration expelled by teachers (corresponding to voiced counting (MORAWSKA *et al.*, 2009)), such that $C_q^{teachers} = C_t$.

The amount of quanta inhaled by a person inside the class over an exposition time τ is the inhaled dose in Eq. (A.7), which can be written as

$$D(\tau) = B p_m \int_0^\tau C(t) dt,$$

where $t = 0$ stands for the time the person enters the room and the total concentration of quanta C (quanta/m³) inside the classroom evolves according to Eq. (A.14). Finally, the probability $p(\tau)$ of a susceptible individual being infected when inhaling a given aerosolized pathogen dose $D(\tau)$ is given by Eq. (A.12).

Infectivity is known to differ between different age groups and pathogen strains, a variability captured by the relative susceptibility s_r in Eq. (A.12). For example, based on the study of transmission in quarantined households in China (ZHANG *et al.*, 2020), Bazant and Bush (BAZANT; BUSH, 2021) suggest assigning $s_r = 1$ to the elderly (over 65 years old), $s_r = 0.68$ to adults (aged 15-64) and $s_r = 0.23$ to children (aged 0-14) for the original Wuhan strain of SARS-CoV-2, which we adopt here as well.

A.1.4 Characteristic parameter values

A.1.5 Outdoor air exchange rate

Although the definition of air quality within an enclosed space varies according to international standards (KHOVALYG *et al.*, 2020), we selected ASHRAE. As described in the technical notes of ASHRAE 62.1 (Ventilation for acceptable indoor air quality), additional requirements

to take into account airborne transmission are not covered by the minimum ventilation rates used here. For ASHRAE 62.1 the minimum ventilation rate is calculated as

$$\lambda_a = \Lambda_p N + \Lambda_a A \quad (\text{A.15})$$

where Λ_p is the outdoor airflow rate required per person, N is the number of people in the ventilation zone during use, Λ_a is the outdoor air flow rate required per unit area and A is the net occupiable floor area of the ventilation zone. Both Λ_p and Λ_a are reference ventilation rates determined by the ASHRAE standard and depend on the type of enclosed space (we adopted values of Educational Facilities - Classrooms of ages 5 to 8 and age 9 plus). As mentioned in the main text, we adopted three reference values regarding distinct situations rather than any arbitrary values:

- *Unoccupied*: it consists of the minimum ventilation rate letting $N = 0$. Take the mean area of the Maragogi classroom group in our database, we obtained the ventilation rate as $\Lambda_1 = 0.8 \text{ h}^{-1}$.
- *Half occupied density*: assumes half the occupation density for classrooms. So, the ventilation rate accounts for both factors N and A . Using the same mean area value as previously, we obtain $\Lambda_2 = 3.8 \text{ h}^{-1}$.
- *Full occupied density* : it consists of full occupation density, and by repeating a similar calculation we obtain $\Lambda_3 = 6.6 \text{ h}^{-1}$.

Note that all reference outdoor exchange air flows above are larger than sedimentation and inactivation rate in the model.

GENERALIZATION FOR CURITIBA-PR

B.1 Robustness of results for the capital Curitiba

We show some results of our investigation on the effects of mitigation protocols in schools for the city Curitiba, the largest state capital in the south of Brazil with nearly 2 million inhabitants. This is a very well developed city, among the highest ranked in the country with respect to HDI, which is in the *very high* range.

The results presented consider potential interventions during the infection wave that occurred between June 14th 2020 and October 12th 2020.

We look at the main scenarios of Figure 4, namely scenarios I, III, V, and VIII, as well as the scenarios where schools remain open with no NPIs and the baseline where schools are closed. We observe that, while the city of Curitiba is less susceptible to the measures, with an increase in cases showing a lower magnitude, **the results are structurally robust and present the same relative hierarchy of effectiveness as the one shown in our main study.**

B.1.1 Inference of states from data of Curitiba

The inference of states in the case of Curitiba is similar to the inference of states made for Maragogi in Section 4.2. The data used for this inference are available at OPENDATASUS (OPENDATASUS..., 2020). The structure of these data differs from the structure of the data collected for Maragogi mainly because we have no information about the brands of the tests used, meaning that we cannot take into consideration false positives or false negatives.

The population of Curitiba is approximately 60 times bigger than the population of Maragogi. This enables us to avoid dealing with attendance or hospitalization data, which are prone to higher bias, and use the more robust death data to infer the states on a daily basis (as done in (MELLAN *et al.*, 2020)).

To infer the states, we use a negative binomial with the daily number of deaths and the overall probability of death (computed using the Table 2), then we infer, using the distributions in (KERR *et al.*, 2020), the time each reconstructed individual spent in each state. As in the main study, this process is repeated 400 times to generate a distribution.

B.1.2 Baseline scenario

The baseline scenario we consider in this section is the one obtained from the modeling of COVID-19 disease in the city during the period of June 14th 2020 to October 12th 2020. During the period considered, the city of CURITIBA-PR was also in lockdown, though interventions were softer compared to those applied to the city of MARAGOGI-AL during the first wave of the disease. From the city's official instructions regarding the opening/closure of services during the first wave, we grouped the services allowed to open during the period into the following categories:

- Hospitals: it comprises all type of health facilities in which possible COVID-19 infected patients were received, including campaign hospitals or not;
- Health Facilities: it includes all other type of health facilities not contained in the category above;
- Supermarkets: the set of all market facilities commercializing mainly food, of medium to large size according to (IBGE-..., 2019) (code 47.11 – 3);
- Markets: the set of all market facilities commercializing mainly food, of small size according to (IBGE-..., 2019) (code 47.12 – 1);
- Food stores: the set of small food stores commercializing essential products (meat, dairy, etc., codes 47.21 – 1, 47.22 – 9, 47.23 – 7, 47.24 – 5);
- Construction stores: the set of store facilities which sell construction equipment, sell vehicle fuel and provide maintenance to vehicle engines (codes 47.3, 47.4 and 45);
- Drug stores: the set of pharmacies and similar stores (code 47.7);
- Industry: the set of industries which depend on production lines to deliver its products (codes 10 to 17, and 19 to 33);
- Construction: the set of companies specialized in construction, which demand physical presence of many workers on site (codes 41 to 43);
- Non-essential: all type of services not included above, except schools.

Data on the total number of facilities and the total number of employees have been gathered, for most services, from (IBGE-..., 2019). Data for the total number of facilities and the total number of hospital and health facility employees were taken from (CNES-HEALTH..., 2020). Schools were not opened during the period considered, but we have taken them into account in comparison scenarios (see Section B.1.4). Data on students and teachers, as well as classes and schools, have all been taken from (INEP..., 2020; INEP..., 2020).

During the period considered, according to the city's official instructions, almost all of the services mentioned above were opened, most with restrictions on the opening time and total number of people per square meter. In our simulations, we have considered that from July 1st 2020 to July 21st 2020, construction stores and non-essential services remained closed. These services were opened during the rest of the period considered in normal opening time. Other services were also opened at normal times during the period considered. The impact of restrictions on opening time and people capacity for services has been taken into account in the calibration of the average number of contacts in these services. See Section B.1.3 for details.

The visitation period and contact network parameters for hospitals, health facilities, markets, supermarkets, food stores, and construction stores have been assumed to be the same as those collected for Maragogi-AL (see Section 4.3). The visitation period for drug stores was used 4 times that of the markets, and the network parameters for this service were chosen equal to those of the markets as well. Services named construction, construction stores, industry and non-essential services did not receive clients, therefore, their visitation period was conceptually infinite. However, the contact network parameters assumed for these services have been calibrated from the SEIR data (see Section B.1.3 below).

We have also considered that modeling the public transportation system was relevant for the spread of COVID-19 in Curitiba-PR (as opposed to what was assumed for Maragogi-AL). The contact network and general behavior of the transportation system are described in Section 3.1.5.

B.1.3 Calibration of the model

The calibration process used in the city of Curitiba-AL is identical to that exposed in Section 4.4, except that more parameters were optimized in this case. We have calibrated 4 parameters in total, which are listed in the following:

- p : the infection probability parameter, the same type calibrated for the city of Maragogi-AL;
- f_{viol} : the fraction of non-essential services that violated city hall instructions regarding their opening during the period considered. This parameter was not assumed necessary at first, but it proved needed eventually during the calibration process;

- c_{ne} : the average number of 1-hour contacts between workers of industry, construction and non essential services. Notice that we have assumed the same parameter for the three types of service;
- c_{transp} : the average number of 1-hour contacts between users of the public transportation system.

We have observed from a simple sensibility analysis that the first two of these parameters caused a much higher impact on the SEIR curves generated from COMORBUSS as an output. Since calibrating the four parameters simultaneously has been proved to be an intense and nearly impractical computational task, we have chosen to calibrate them in two steps. First, we optimize c_{ne} and c_{transp} , keeping p as in Section 4.4 and $f_{viol} = 0$. This first calibration procedure gave us the following approximate values for these parameters: $c_{ne} = 0.2$ and $c_{transp} = 0.1$. Fixing c_{ne} and c_{transp} by these calibrated values, we optimized for p and f_{viol} in a second step. The final values for these last parameters were found to be $p = 0.0434$ and $f_{viol} = 0.879$, with an L^1 -Wasserstein distance of 1.35×10^{-2} between the SEIR curve distributions of the target and reference.

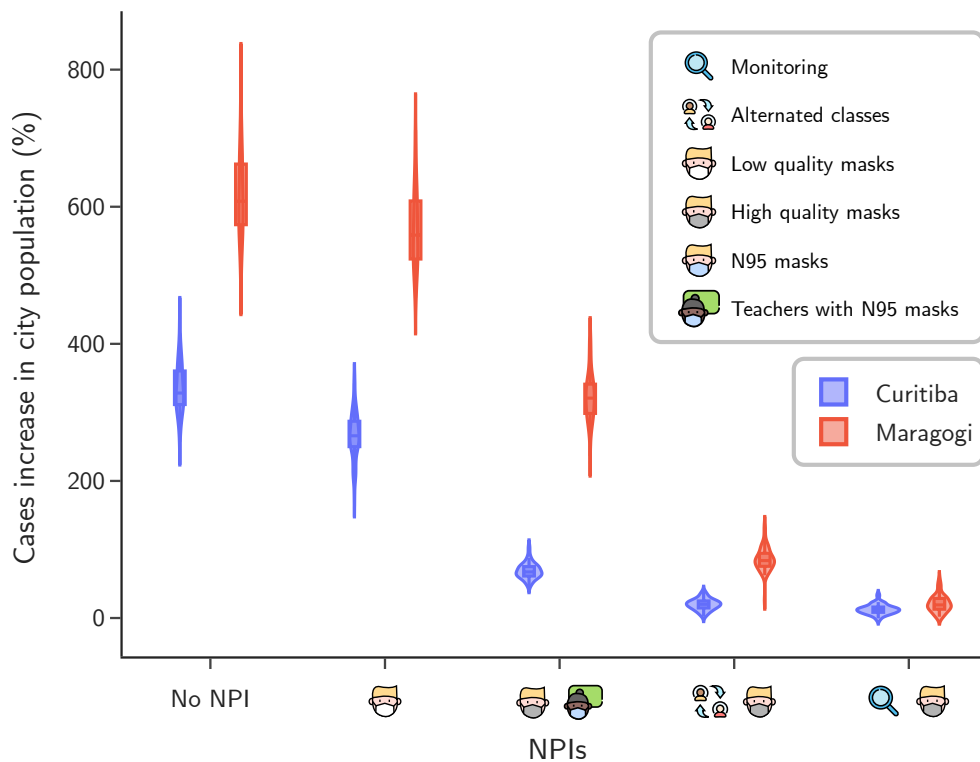


Figure 36 – **Effectiveness comparison for different demographics.** Relative increase in cases for different scenarios for Curitiba-PR compared to Maragogi-AL.

B.1.4 Robustness of results

After the modeling and calibration for the city of Curitiba, we perform simulations with 60 seeds using five different policy scenarios which are compared again to the baseline where schools are kept closed. The increase in cases relative to this baseline is depicted for each scenario in Figure 36.

Most remarkably, the relative rank of protocol effectiveness is the same as observed for a city of small demography, such as Maragogi. This highlights the robustness of the protocols across different demographics. Second, we note that cities of smaller demographics are susceptible to greater case increase due to bad choices of protocols. This highlights their greater vulnerability and, coupled with their larger representation in national and international demographic distributions, justifies our choice of focus for this study.

