SARS-CoV-2 epidemiological model based on socio-economic variables in Switzerland

Cloe Cortes Balcells

Michel Bierlaire

Rico Krueger

STRC conference paper 2022 **May 12, 2022**

STRC | **22nd Swiss Transport Research Conference**
Monte Verità / Ascona, May 18-20, 2022

SARS-CoV-2 epidemiological model based on socio-economic variables in Switzerland

Cloe Cortes Balcells
Transport and Mobility Laboratory
Ecole Polytechnique Federale de Lausanne (EPFL)
Lausanne
cloe.cortesbalcells@epfl.ch

Michel Bierlaire
Transport and Mobility Laboratory
Ecole Polytechnique Federale de Lausanne (EPFL)
michel.bierlaire@epfl.ch

Rico Krueger
Department of Technology, Management and Economics
Technical University of Denmark (DTU), Denmark
rickr@dtu.dk

May 12, 2022

Abstract

Existing epidemiological models analyze the transmission of infectious diseases considering a perfect homogenous population. However, the COVID-19 emergency has shown the importance of considering activity-travel behavior when studying the spreading of the virus. With increasing epidemiological data available, and the outburst on agent-based activity models, we can move beyond aggregation and start including individual features. To the best of the authors knowledge, this is the most in-depth study of how socio-economic and virological features impact the spreading of COVID-19 to date. We use chronological data from the Federal Office of Public Health (FOPH) from mid-February 2020 to mid-September 2021. We derive the influence of socio-economic characteristics with a novel semi-disaggregated SIRD model, obtaining the total number of infections per specific group of the population sharing pre-determined features. Finally, we validate our model with Google data and compute the reinfection rate by applying non-pharmaceutical interventions. Five features, including information about the individual and the municipality, have a $\geq 95\%$ probability of being correlated with the endogenous variable of positive testing for COVID-19. In addition, we find that certain variables, including age or the population density per square meter, remain representative for all waves, whereas others, like household income, are dependent on the epidemiological wave studied. Our results suggest a strong dependency on individual and municipality characteristics and the force of infection of COVID-19.

Keywords

Agent-based model, SIR model, policy decision making, COVID-19.

Contents

List of Tables

List of Figures

1 Introduction

Since the 18$^{\text{th}}$ century, epidemiological models are used to study the spread of infectious diseases (see Heyde and Seneta, 2001). However, models that include activity-travel behavior and disease spreading are much more new to the scientific community. COVID-19 represents a turning point in the history of transportation and epidemiological research for two main reasons: the magnitude of the pandemic and the amount of data collected during its evolution. This situation has revealed the lack of literature that combines the epidemiological field with public transportation planning. According to Tirachini and Cats (2020); Douglas *et al.* (2020), human mobility is one of the main causes of the spread of COVID-19. Therefore, coupling mobility and epidemiological data can provide a better spreading model and flexibility when addressing an epidemiological crisis. Mobility information is of fundamental importance for planning Non-Pharmaceutical Interventions (NPIs), including public transportation logistics and the partial restriction of people's daily activities (see Tirachini and Cats, 2020; Douglas *et al.*, 2020; Zheng *et al.*, 2020; Lee and You, 2020) during and after an epidemic crisis.

For this reason, we propose a model that addresses the subject at a disaggregate level. To understand human mobility, we need to capture the heterogeneity of behavior in the population, not only in the mobility model but also inside the epidemiological model. Capturing heterogeneity allows for determining the influence of activity-travel behavior on mortality rates and the efficacy of restrictions on specific activities. Concerning the influence of individual characteristics on the spread, while the impact of the vaccination status is rather clear, multiple authors (see Singu *et al.*, 2020; Oertelt-Prigione, 2020) state the importance of other socio-economic characteristics like age or income. Also, Riou *et al.* (2021c) presents the correlation between positive SARS-CoV-2 tests, mortality rates, and admission to intensive care with the SocioEconomic Position (SEP). In European Institute for Gender Equality (2021) they state that 76% of health workers in Europe are women. For this reason, studies like Oertelt-Prigione (2020) consider factors such as gender, exposure to the virus, symptoms, and health care information. Also, in Klein and Flanagan (2016) we find the potential different immune responses according to biological characteristics. It is important to include variables in the epidemiological model that capture the heterogeneity of the population's behavior. In Qian and Ukkusuri (2021a) they point out the two main challenges when using activity-based models: (i) mobility clusters the population to the locations of their activities which leads to contact and contagion (ii) to account for heterogeneity, we need to take into account additional assumptions. For this reason, we aim to address the following limitations found in the existing literature: (i) to define a clear methodology that establishes which variables are meaningful inside an

epidemiological model, for example income or residence place (Chang *et al.*, 2021a) (ii) to overcome the issue of adding aggregated parameters inside agent-based models due to the lack of data (Tuomisto *et al.*, 2020), (iii) to avoid the use of real-time data-driven analysis in order to define more targeted and less disruptive interventions (Aleta *et al.*, 2020), and (iv) to make the probabilities of transmission time dependant, since an early adoption can potentially allow to contain the epidemics (Mancastroppa *et al.*, 2020a)

Therefore, this paper's scope is firstly to demonstrate the added value of using disaggregate models for modelling SARS-CoV-2 spreading. Secondly, to describe the preliminary considerations and define a model that accounts for virological and socio-economic variables. And finally, to evaluate the potential of this model to study SARS-CoV-2 policy decision making. For this task, we employ score matching causal inference combined with generalised multivariate regression model to compute the probability of infection depending on the socio-economic characteristics of the individuals. This method is especially fitting since it incorporates the population heterogeneity and their behavior and contact patterns.

2 Literature review

2.1 Epidemiological models

In epidemiology, SIR models (Susceptible-Infectious-Recovered) have been widely used to study the spread of multiple diseases (see Choisy *et al.*, 2007). These models divide the population into three groups: *Susceptible*, *Infected* and *Recovered* individuals. *Susceptible* individuals are not infected and do not present any immunity or resistance to the disease. *Infected* individuals remain in this state for a certain amount of time and may contaminate others. *Recovered* individuals are usually considered immune to the disease. This group usually also includes deceased people. In the literature, most of the works that study the COVID-19 outbreak from the perspective of mobility use an adapted version of the SIR, mainly the SEIR model (see Chang *et al.*, 2021b; Aleta *et al.*, 2020; Müller *et al.*, 2020; Qian and Ukkusuri, 2021b; Tuomisto *et al.*, 2020). The SEIR model integrates the SIR model with an 'Exposed' state to consider individuals already infected but cannot contaminate others yet. A further step is taken in Aleta *et al.* (2020), where pre-symptomatic, symptomatic, or asymptomatic states are considered. Müller *et al.*

(2020) refines this model by splitting the Infectious state into four sub-states representing four levels of increasing symptoms severity (pre/asymptomatic, symptomatic, seriously sick, critical). Moreover, Lemaitre *et al.* (2020b); Tuomisto *et al.* (2020) use a single Infectious state but instead includes states indicating if the patient is hospitalized or in the Intensive Care Unit (ICU). Both Lemaitre *et al.* (2020b) and Tuomisto *et al.* (2020) also separate recovered and deceased patients, as the average duration of the hospitalization differs between both cases.

2.2 Aggregated versus disaggregated models

2.2.1 Aggregated models

We define the *aggregation* of an epidemiological model as the level of detail with which we compute the progression of the disease within the population. Fully aggregated SIR models result in Equations (1)-(3), describing the evolution of the size of the Susceptible $S(t)$, Infected $I(t)$ and Recovered $R(t)$ compartments of the population as a function of the time t (Choisy *et al.*, 2007; Lemaitre *et al.*, 2020b).

$$\frac{dS}{dt}(t) = -\lambda(t)\frac{I(t)}{N}S(t) \tag{1}$$

$$\frac{dI}{dt}(t) = \lambda(t)\frac{I(t)}{N}S(t) - \gamma I(t) \tag{2}$$

$$\frac{dR}{dt}(t) = \gamma I(t). \tag{3}$$

$$S(t)+I(t) + R(t) = N \quad \forall t \tag{4}$$

where N is the constant population size, $\lambda(t)$ is the number of contacts per person per unit of time and γ is the recovery rate. The dynamics of the spread depending on the ratio:

$$R_0 = \lambda/\gamma. \tag{5}$$

We define R_0 as the average number of new infections that individuals who carry the disease cause in an early stage of the epidemic inside a susceptible population (see Diekmann and Heesterbeek, 2000; Anderson and Mary, 1992). If R_0 is smaller than 1, the number of infected people decreases over time. This system of equations can be solved analytically to

obtain the curve of the number of cases or prevalence of the disease over time (Lemaitre *et al.*, 2020b). The analysis of the equation system's equilibrium yields the result stating that the value $R_0 = 1$ is the threshold defining whether an epidemic declines or grows exponentially (Choisy *et al.*, 2007). This parameter depends on the number of contacts within the population and the probability of contagion during those contacts.

The main disadvantage of compartmental models lies in their assumption that compartments are fully mixed. For this reason, they risk oversimplifying the problem by neglecting 'imperfect mixture' such as heterogeneity in the population, contact patterns, and complex behavior of individuals (see Soper, 1929).

2.2.2 Semi-aggregation: metapopulation models

A possible level of disaggregation consists in the formulation of *metapopulation* models, in which the population is represented as a clustered network. Each cluster is a *subpopulation* (groups) characterized by its number of individuals within each disease progression state. While the subpopulations may exchange individuals and interact with one another, the model does not consider interactions at the individual level within each group (Hackl and Dubernet, 2019). The subpopulations are usually built on geographic and demographic criteria: in Qian and Ukkusuri (2021b) New York City is divided into 15 zones that interact via the various transportation systems. Interestingly, this study also considers separate SEIR states for the populations *commuting between two zones* to precisely assess the proportion of contagions that take place while traveling within the city. The subpopulations can also depend on data availability: in Chang *et al.* (2021b) the individuals are aggregated in Census Block Groups (CBGs), which makes it coherent with the demographic data from the US Census as well as the mobility data. This type of semi-aggregated model already allows for revealing inequalities in the outbreak prevalence across the subpopulations.

In conclusion, these models can capture the disease dynamics and compute the infection evolution in disease outbreaks. However, they cannot analyze the contact between two individuals nor the characteristics of the individuals interact. Moreover, they do not leverage activity-based information, therefore neglecting the possible correlation between the location of an individual and its potential interactions. Finally, precise network-based modeling requires a high number of input variables, increasing the dimension of the problem and resulting in computationally expensive calculations.

2.2.3 Disaggregated / Agent-based models

The most advanced level of disaggregation is considering the population at the levels of individuals. As each individual has its characteristics, actions, and interactions with the others, they are often referred to as *agents*, and the resulting model is called *Agent-based* (Hackl and Dubernet, 2019; Tuomisto *et al.*, 2020; Aleta *et al.*, 2020). Each agent is characterized by a disease state at each time step and may contaminate or be contaminated by the other agents with whom there is a contact. Just as the number of contacts and the transmission rate need to be chosen or calibrated in aggregated SIR models (section 2.2.1), an agent-based model heavily relies on the *contact policy* as well as the probability of transmission. Smieszek (2009) demonstrates that redundant contacts (constant contacts between the agents) tend to reduce the rate at which the disease spreads (lower R_0). It also supports the intuitive idea that the average number of contacts per time unit is essential in determining the magnitude of the outbreak peak.

Agent-based models present many advantages as they represent a more realistic approach to the problem of simulating an epidemic. Metapopulation models have the issue of generalizing the behavior and faith of the individuals that cohexist within the same subpopulation (Hackl and Dubernet, 2019). For example, a metapopulation model by nature cannot simulate Super-spreading events where a single individual contaminates many others. Agent-based models can reproduce those phenomena, and it is notably among the main objectives in Aleta *et al.* (2020).They also allow studying the effect of precise NPIs such as contact tracing, and individual quarantine (Tuomisto *et al.*, 2020).

On the other hand, agent-based models are technically more challenging for two main reasons: (i) their need for fine-grained data at the individual level in order to realistically simulate contacts between the agents (this is linked to the mobility models which are discussed in Subsection 2.3); (ii) the computational cost, as the interactions between potentially hundreds of thousands of unique and varying agents need to be computed. In addition, those models are usually stochastic (contrary to the basic aggregated SIR model), which requires the model to be run for a significant amount of time (Chang *et al.*, 2021b; Hackl and Dubernet, 2019). This can be solved by optimizing and parallelizing the calculations as in Chang *et al.* (2021b). Still, most studies limit the size to a specific city, or a relatively limited sample of the real population (Müller *et al.*, 2020; Hackl and Dubernet, 2019; Tuomisto *et al.*, 2020). All in all, these models are suited for describing the heterogeneity of the population in terms of mobility. However, there is a lack of contributions to defining heterogeneous populations in mobility and epidemiology

2.3 Contact and mobility models

This subsection describes the methods used to simulate the contacts within a population, which is a crucial aspect of any epidemiological model, as mentioned in Section 2.1.

2.3.1 Contact policies

Similar to the overall epidemiological model, the contacts policy used in a simulation can be more or less aggregated. In the basic SIR model described in Choisy *et al.* (2007) leading to equations 1-3, the contact policy is reduced to a single parameter β which is the average number of contacts of a single person per unit of time. While simplifying the simulation to allow for the system to be solved analytically, it ignores all differences in behavior and activity between the individuals. An improvement of the scalar β is the contact matrix: for a population that can be divided into m groups, we define a matrix of size mxm such that $M_{ij} = \beta_{ij}$ where β_{ij} is the number of contacts per unit of time between individuals from a group i and those from group j. In the case of a metapopulation model, the contact matrix may indicate the contacts between the subpopulations.

Nonetheless, metapopulation or agent-based models are often based on a disaggregated contact policy: at each timestep, the model computes which agent / which subpopulation has been in contact with which others. A first method is to use a synthetic population and arbitrarily decide on the contacts: in Smieszek (2009) each individual is put in contact with n other random individuals, which may or not change throughout the simulation. A similar policy is used in Mancastroppa *et al.* (2020b), with a crucial difference: each node (individual) has an *activity potential* that influences the distribution of its number of contacts and an *attractiveness* that calibrates its likelihood of being chosen as a contact. This allows the author to study the impact of individual quarantine policies by adjusting the activity potentials and attractivenesses of infectious agents.

What seems most realistic yet technically challenging are data-driven activity-based contact models. Those rely on datasets that list the mobility of all individuals within a population. This includes home locations and successive locations visited over a day, including commuting. The type of activity performed is also indicated as it matters in the number of contacts. Such datasets are used in Hackl and Dubernet (2019); Chang *et al.* (2021b); Tuomisto *et al.* (2020); Müller *et al.* (2020); Aleta *et al.* (2020); Qian

and Ukkusuri (2021b). In Qian and Ukkusuri (2021b); Hackl and Dubernet (2019) the mobility data comes from a survey about the transportation system. In Aleta *et al.* (2020); Chang *et al.* (2021b) it is obtained via Cuebiq and SafeGraph, respectively, two private companies specialized in gathering mobility data, including via mobile phones. In Müller *et al.* (2020) the authors indicate that their base mobility dataset stems from mobile phones data too.

2.4 Relationship between infections and socio-economic and demographic variables

This paper aims to show the advantages of adapting an aggregate model adapting a disaggregate socio-economic and demographic data. Among the variables that may influence the probability of transmission and contact policy are age, population density, and income. Thus, a key component will be identifying which variables play a role by correlating to the probability and force of infection. DuPre *et al.* (2021) studies the correlations between socio-economic variables and COVID-19 cases and deaths trajectories across 3,141 counties in the U.S.A. between January and June 2020. Their results show that older median age, a 1% higher proportion of females, and a higher proportion of black or Hispanic residents increased the probability of a surge in the number of cases in June (when no lockdown was in place) as well as being in the worst-case scenario regarding the number of cases. The same conclusions could be made regarding the correlation with the deaths trajectory. According to the authors, these results reflect that black and Hispanic workers are more likely to be employed in the industry as essential workers. Consequently, those occupations offer fewer alternatives to reduce mobility and are more at risk as they imply more close-up contacts (such as healthcare or transportation systems employees).

On the other hand, DuPre *et al.* (2021) shows that counties with a higher proportion of individuals below the poverty level or without health insurance were less likely to be among the counties experiencing the worst death trajectories. Besides, Sannigrahi *et al.* (2020) studied the correlation between socio-demographic variables and COVID-19 cases and fatalities across Europe. Their work identifies income, poverty, and total population as critical variables in explaining the variance of the COVID-19 incidence. As mentioned in Sannigrahi *et al.* (2020), the Ordinary Least Square regression does not capture the spatial auto-correlation: two areas (e.g., European countries) that are close to each other have a statistical tendency to have similar values. Thus the authors use spatial regression models: Geographically Weighted Regression (GWR), Spatial Error Model (SEM), and

Spatial Lag Model (SLM). The results show that the GWR captures better the relationship between the variables and the number of cases and deaths than the OLS regression (R^2_{max} of 0.88 compared to 0.76).

In summary, epidemiological models are widely used to study the spread of infectious diseases. Nonetheless, potential contributions can enrich the scientific community, bridging the epidemiological and transportation fields. Specifically regarding the implementation of a methodology that allows considering the fitting variables of the individual. After including them in the infection probability as interpretable explanatory variables. This level of disaggregation enables the study of activity-based information. Activity-based information allows us to assess features like the correlation between the socio-economic characteristics of the population and their transportation mode choice or provide insight on how to plan for transportation of a city in a post-pandemic world. In other words, these models are attractive in describing the heterogeneity of the population in terms of mobility. However, there is a lack of contributions to defining heterogeneous populations in mobility and epidemiology.

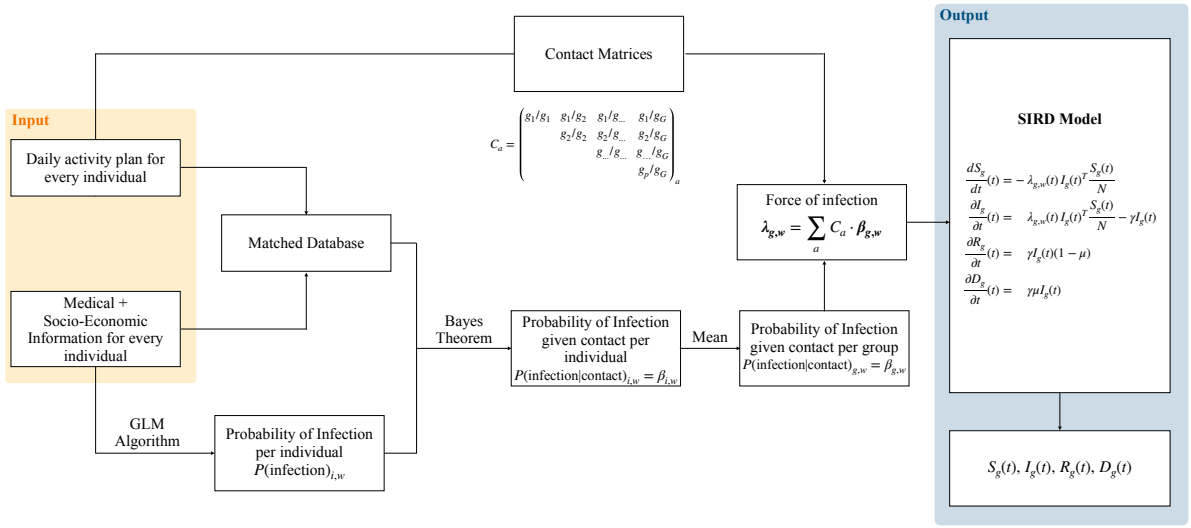
3 Methodology

Compartmental models, also known as SIR models, assume a uniformly distributed population. We understand a uniformly distributed population as a population in which all individuals are subject to the same infection risk per time step (Kelman, 1985). However, in Riou *et al.* (2021a), it is established that population structure is critical when accounting for COVID-19.

For this reason, we implement a semi-disaggregated SIRD model with an activity-based approach. We establish the epidemiological disaggregation through the force of infection $\lambda(t)$ (see Equations (1)-(3)). Indeed, the latter is defined conditional to the belonging portion g of the total population P and the pandemic spreading wave (w).

Figure 1 displays an overview of the workflow of our model:

Figure 1: Model formulation diagram



3.1 Input of the model

The two primary inputs required to run the model exposed in Figure 1 are the daily activity plan for every individual and the medical and socio-economic characteristics of every individual. To obtain the dataset, we propose to match the activities schedule of the individuals and their medical information through some socio-economic characteristics.

To obtain the daily activity plan for every individual, we use agent-based modeling. ABM is a system modeled as a collection of autonomous decision-making entities called agents. Each agent individually assesses their situation and makes decisions based on rules. Agents may execute various behaviors appropriate for the system they represent—for example, producing, consuming, moving, and most importantly, interacting with other agents. We use an activity model to produce the event files of each individual in our simulation. Its use of dependency injection and agent-based modeling provides a fine-grained modular framework. This information is completed by defining the different activities, such as the transportation mode, leisure, home, errands, and work. Also, another file defines the characteristics of the population, which is used for matching. Finally, we obtain the schedule of each individual for every day.

The other input for the model is the medical information of the individual. Specifically,

the requirements are information about the date the individual tested positive, vaccination and hospitalization date and its socio-economic characteristics.

3.2 Activity specification

We define the contact matrix C_a in Equation (6), in order to encode the number of contacts per time step between and among each group for every given activity a

$$C_a = \begin{pmatrix} g_1/g_1 & g_1/g_2 & g_1/g_{...} & g_1/g_G \\ & g_2/g_2 & g_2/g_{...} & g_2/g_G \\ & & g_{...}/g_{...} & g_{...}/g_G \\ & & & g_p/g_G \end{pmatrix}_a \quad \forall a \in [1, A] \tag{6}$$

where A is the total number of considered activities by the study and the term g_i/g_j is to be interpreted as the number of contact between individuals of the groups g_i and g_j $\forall i, j \in [1, G]$.

The contact matrix allows for accounting for information related to the contacts by type of activity and by group g, which may have very different social behavior. Indeed, as discussed in Section 2, different types of activities yield probabilities of infection that differ in their likelihood and in how we prevent them. For example, places of recreation are responsible for a higher proportion of infections than places of business (see Chang *et al.*, 2021c). They are also subject to different types of regulation (e.g., closure of restaurants and bars or non-essential shops). Differentiating the contact matrix allows for studying different public policies by directly modifying the number of contacts under a particular policy, i.e., a compulsory "telework" measure would directly result in a work-related contact matrix (C_a, where a = telework) close to zero, at least for teleworking businesses.

3.3 Probability of infection $\beta_{g,w}$

To define the probability of infection per group, we need to determine the explanatory variable that will define the probability of infection for each individual. Therefore, we estimate a regression function that will define the probability of infection for each individual.

For this reason, we define and estimate the parameters that explain the $P(\text{infection})_{i,w}$ for each wave w and individual i. The two methods used are: (1) we find correlation through multivariate logistic regression between the explanatory variables and the binary response for testing positive for an individual i, and (2) we check for causality by implementing a propensity score causal inference algorithm.

3.3.1 Probability of infection per individual

The probability of infection per individual is computed using a multivariate logistic regression with a binary dependent variable. The variable we are interested in modeling is an individual's positive test, an indicator for whether an individual has been SARS-CoV-2 tested positive (infected = yes) or negative (infected = no). The regressors that ought to have power in explaining whether an individual has been tested positive are their socio-economic characteristics and the characteristics of the environment that surrounds them.

This process involves the implementation of multivariate logistic regression and a matching score causal inference method. The causal inference method aims to balance the distribution of covariates in the treated and control groups. The regression has the objective of finding the correlation between the probability of infection and the selected explanatory variables. It is worth mentioning that matching methods should not be seen in conflict with regression adjustment. The two methods are complementary and best used in combination. We define the probability of infection per individual as:

$$P(\text{infection})_{i,w} \sim \sum_{m=1}^{M} (\beta_m \log X_{m,i} + \beta_p \log X_{P,i}) \tag{7}$$

Where X_m are the socio-economic characteristics of the individual, X_p are aggregate indicators of the surroundings, and β_m and β_p estimate the variable's parameters. The procedure to select the variables for the Generalized Linear Model Regression (GLM) is detailed in Algorithm 1.

Algorithm 1 Procedure for selecting variables for GLM regression

1: Split the dataset into 70% train and 30% test
2: Select all the variables from the list that we want to consider for the analysis
3: Compute the GLM algorithm
4: Compute the Variance Inflation Factor (VIF) for all the variables
5: **if** High p-value and high VIF **then**
6: Remove the variables with high p-value and high VIF
7: Refit the model (line 3)
8: **else if** all p-values and VIF values are accepted with 95% confidence **then**
9: Output the the result
10: **end if**

3.3.2 Probability of infection per group

Computing the probability of infection per individual $\beta_{i,w}$ allows us to define the probability of infection $\beta_{g,w}$ per group g and per wave w. To obtain $\beta_{g,i}$, we first calculate $P(\text{contact}|\text{infection})$ and $P(\text{contact})$ from an activity-based model output:

$$P(\text{contact}) = \frac{\text{total number of interactions in all facilites}}{\text{total number of facilities}} \tag{8}$$

$$P(\text{contact}|\text{infection}) = \frac{\text{number of infected people in facilities} * \text{total number of facilities}}{\text{total number of people}} \tag{9}$$

This output contains the daily activity schedule of the individuals in the population. For this reason, it is possible to compute statistics on the contacts inside the different facilities and the characteristics of these encounters. To obtain the probability of infection, given that there is contact between two individuals $\beta_{i,w}$ we apply the Bayes theorem (10):

$$\beta_{i,w} = P(\text{infection}|\text{contact}) = \frac{P(\text{contact}|\text{infection})P(\text{infection})}{P(\text{contact})} \tag{10}$$

Aggregating the result for each group g we obtain:

$$\beta_{g,w} = P(\text{infection}|\text{contact})_{g,w} = \frac{\sum\limits_{\forall i \in g} \beta_{i,w}}{N_g} \tag{11}$$

where the term $\sum_{\forall i \in g}$ indicates the sum for all the individuals i that belongs to the group g and N_g is the total number of individuals that belong to the group g.

3.4 Force of infection ($\lambda_{g,w}(t)$)

This paper aims to define the probability of infection of an individual given its socio-economic characteristics and activity-travel behavior. Consequently, we compute the activity contact matrices C_a together with the mean of the probability of infection given contact per group ($\beta_{g,w}$). As already mentioned, to obtain this parameter, we first model the probability of infection given contact depending on the socio-economic characteristics of the individual ($\beta_{i,w}$)

$$\lambda_{g,w}(t) = \beta_{g,w} * \sum_{a=1}^{A} C_a \tag{12}$$

where $\lambda_{g,w}(t)$ is the mean force of infection of all the individuals that belong to the same group g for wave w. The entire process is summarized into Algorithm 2.

3.5 SIRD Model

The final step is to introduce $\lambda_{g,w}(t)$ inside a SIRD model. For each segment of the population (group g), characterized by a set of specific features, i.e., age, gender, income, or municipality, we define the following Ordinary Differential Equations (ODEs):

$$\frac{dS_g}{dt}(t) = -\lambda_{g,w}(t)\, I_g(t)^T \frac{S_g(t)}{N} \tag{13}$$

$$\frac{\partial I_g}{\partial t}(t) = \lambda_{g,w}(t)\, I_g(t)^T \frac{S_g(t)}{N} - \gamma I_g(t) \tag{14}$$

$$\frac{\partial R_g}{\partial t}(t) = \gamma I_g(t)(1-\mu) \tag{15}$$

$$\frac{\partial D_g}{\partial t}(t) = \gamma \mu I_g(t) \tag{16}$$

Where the terms S, I, R, D, N stand for the total susceptible, infectious, recovered, and dead individuals, and the total number of individuals, respectively. The recovery rate (γ) and the fatality rate (μ) are defined from the literature (see Lemaitre *et al.*, 2020a). Note that the periods considered for the waves are based on Roelens *et al.* (2021). The system of ODEs (13)–(16) can be solved using the *lsoda* solver for ordinary differential equations thanks to its ability to switch between stiff and non-stiff integration methods automatically. This method allows for computing S, I, R, and D for each group $g \in [1, \ldots, G]$ where G is the total number of groups that compose the population P. We compute it t times at which explicit estimates for the output are desired.

Algorithm 2 Summary of the proposed methodology

Require: Dataset with individual information and smtg else
1: Execute Algorithm 1
2: Compute the contact matrices C_a for all the possible activities $a \in [1, \ldots, A]$
3: **for** each pandemic wave w **do**
4: **for** each population group g **do**
5: **for** each individual i **do**
6: Compute the probability of infection per individual $\beta_{i,w}$ with Equation (7)
7: **end for**
8: Compute the probability of infection per group $\beta_{g,w}$ with Equation (11)
9: Compute the force of infection $\lambda_{g,w}(t)$ with Equation (12)
10: Solve the system of ODEs (13)–(16)
11: **end for**
12: Output the model for each pandemic wave w
13: **end for**

It comes without saying that the above specification depends on the dataset available to define $P(\text{infection})_{i,w}$. The dataset should include the characteristics of the individual and virological attributes. Individual attributes can include age, gender, or income, and virological attributes are the parameters like viral load, contact intensity, or ventilation characteristics. However, segmenting the population into smaller groups results in a higher dimension of the matrix C_a that can lead to higher computational cost. To the best of

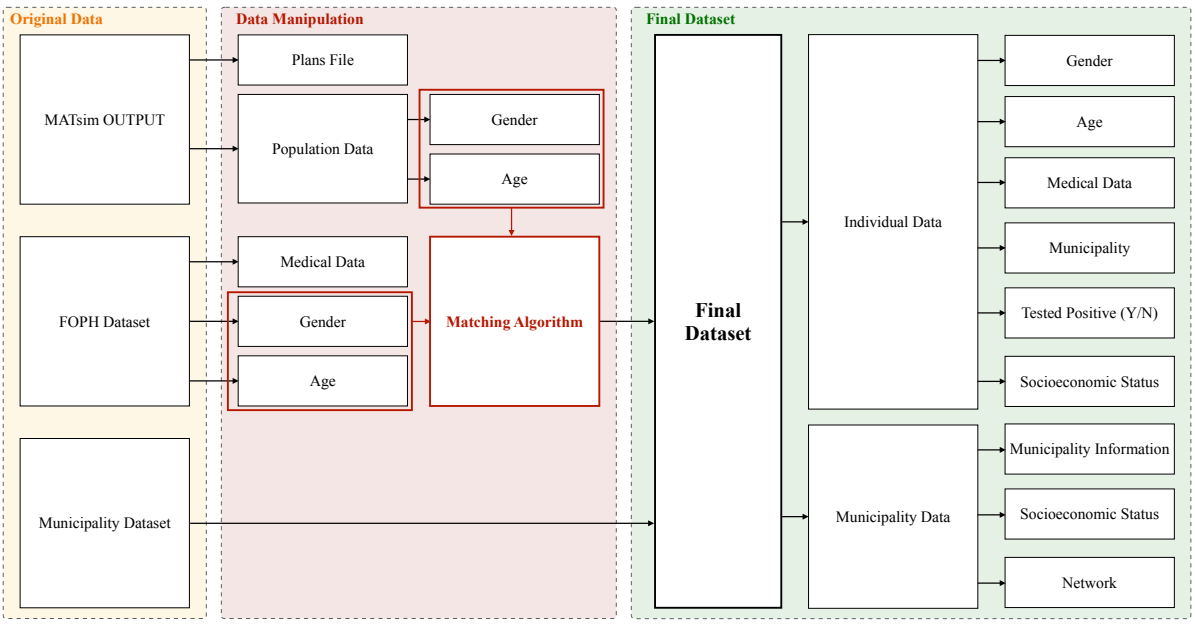
the authors' knowledge, high-performance computing is required if more than 15 groups of the population need to be taken into account.

4 Results

The study case used to validate the proposed approach relies on data concerning 5% of the Swiss population. The segmentation of the population per group is based on the individual's age. We estimate the force of infection by including socio-economic variables of the individuals and their daily activities. By capturing these two phenomena, we can get information about the activity-travel behavior of the Swiss population. Specifically, it allows to: (i) study the causality and the correlation of the probability that an individual gets infected given its socio-economic characteristics, (ii) evaluate targeted policies by including associated variables.

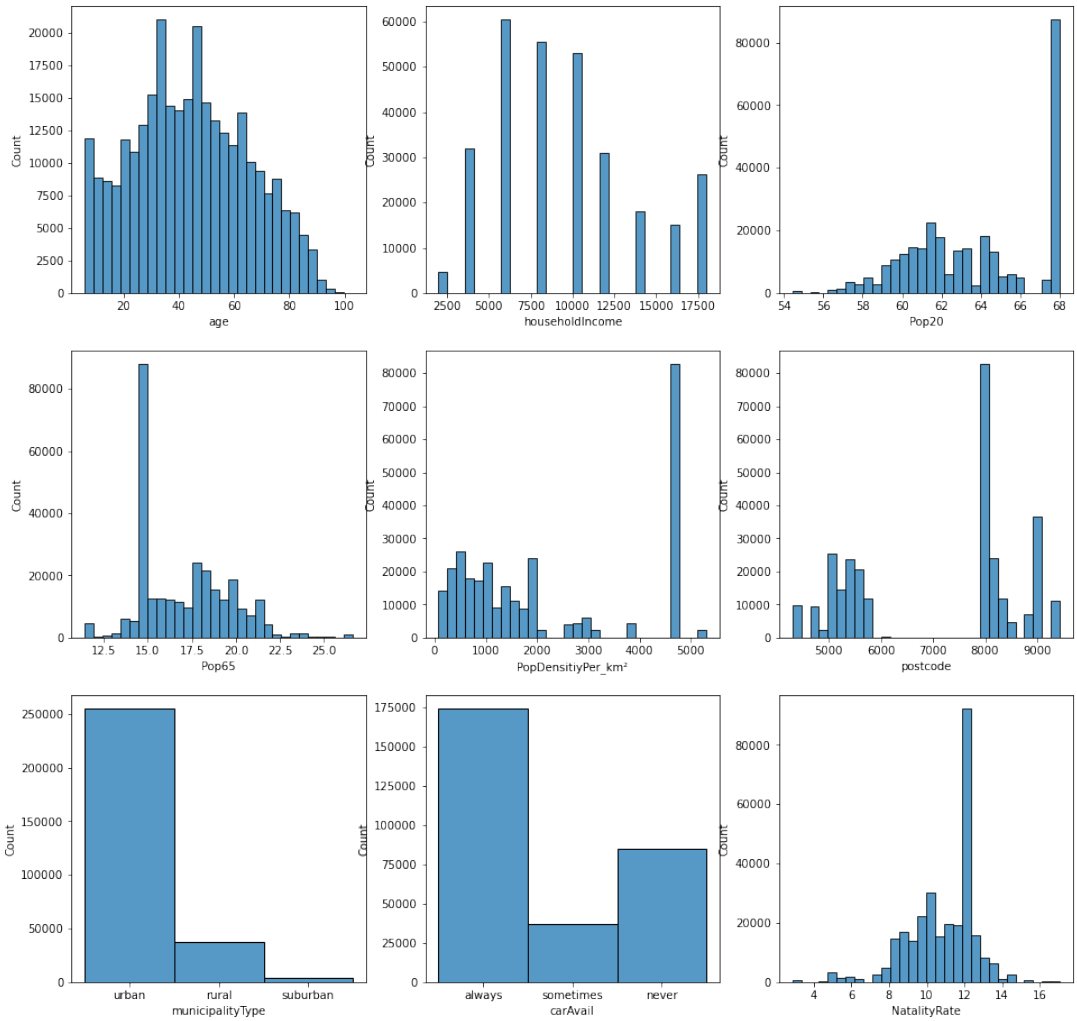
4.1 Data

Figure 2: Pre-process of the dataset



The dataset requirements as an input for this model include activity and medical information about the individuals as visible from Figure 1. Since there is no synthetic population in the literature that includes all the needed features, we compute a matching algorithm to combine different datasets (see Figure 2). The underlying reason is that we need to account for each individual's daily activities and socio-economic characteristics together with their COVID-19 medical-related information.

Figure 3: Distribution of positives test a function of (from top left to bottom right): *(i)* age, *(ii)* aggregated income per household, *(iii)* percentage of people per municipality with an age between 20 and 65 years old, *(iv)* percentage of people per municipality with an age over 65 years old, *(v)* population density per km^2, *(vi)* location based on postcode, *(vii)* type of municipality, *(viii)* availability of private means of transportation (i.e., car) and *(ix)*natality rate.



We manipulate data from the Federal Office of Public Health (FOPH) from mid-February 2020 to mid-September 2021 (see Riou *et al.*, 2021b). The dataset contains the positive

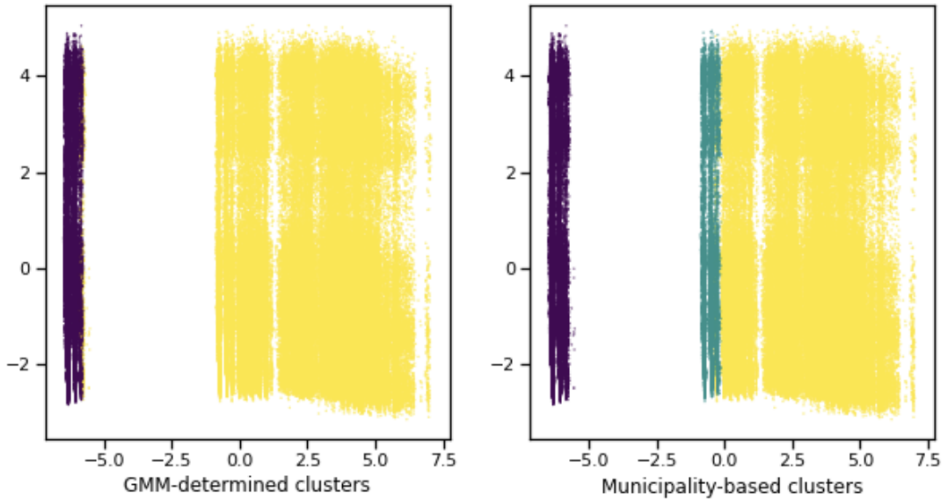
tests in Switzerland, together with information about the tested individuals. It includes information about the age, gender, municipality, vaccination doses, hospitalization, and if they died. Moreover, we add open-source data (see Admin suisse, 2022) from the Swiss municipalities. These variables per municipality include the median income, the social security rate, the percentage of people working in the tourism sector, or population density per square meter. We match the FOPH and the municipality data with a calibrated MATSim simulation output from ETH Zurich (see Hörl and Balac, 2021). The final dataset size is 400k individuals and is formed by three individual socio-economic characteristics (sex, age, and municipality) and forty-one variables at the municipality level (see BFS, 2022).

Figure 3 plots the distribution of positive tests against different explanatory variables. We observe that the distributions are not horizontal, reinforcing the assumption that these variables are fitting to describe the correlation with the positive testing variable.

To further analyze the underlying reason for the outliers in the municipality aggregated indicators, we compute a Principal Component Analysis (PCA) to reduce the explanatory variables to 2 dimensions and a Gaussian Mixture Model (GMM) clustering to visualize the clusters better and use them for further analysis. Figure 4 displays the two clusters: Zurich (violet) and St. Gallen (green). The outliers of the municipality indicators are reasonable since more individuals of the same municipality explain more counts in the histogram. On the other hand, the individual-based indicators are almost uniformly distributed. For this reason, we decide to test these variables as explanatory to compute the probability of infection per individual.

In Figure 5, we merge the FOPH Data with the municipality data to plot the positive tests in Switzerland. This analysis gives an overview of the disparities in infection in the different municipalities, which makes these variables suitable to define the positive tests of the individuals.

Figure 4: Visualization of the clusters using the GMM algorithm. The figure from the right adds a column in the analysis indicating if individuals live in Zurich, St-Gallens, or some other municipality.



4.2 Activity contact matrix

As seen in subsection 3.4, the force of infection is a vector whose dimension depends on the segmentation of the population. For our case of study, we stratify the model into 4 age groups: P_C which contains the individuals younger than 18 years old, P_{A1} individuals between 19 and 35 years-old, P_{A2} individuals between 36 and 55 years-old and P_E individuals over 56 years-old ($G = P_C, P_{A1}, P_{A2}, P_E$). The contact matrix C_a defined in Equation 6 assumes the form visible in Table 1. This matrix encodes the number of contacts between and among each age group per time step. In Figure 6, we can observe the mean of contacts per activity between each age group.

Table 1: Contact matrix structure for each activitity

	child (C)	*adult1 (A1)*	*adult2 (A2)*	*elderly (E)*
child	child / child	child / adult1	child / adult2	child / elderly
adult1	-	adult1 / adult1	adult1 / adult2	adult1 / elderly
adult2	-	-	adult2 / adult2	adult2 / elderly
elderly	-	-	-	elderly / elderly

For instance, $C - C$ defines the contacts for each activity between the same age group C. It is fascinating to observe that most contacts take place inside *Education*, where the

Figure 5: Distribution of positive tests against their location in Switzerland, in addition to some data concerning the municipalities to analyze(from top left to bottom right): *(i)* the population density, *(ii)*the average size of the households per person,*(iii)* the rate of persons receiving social financial aid, *(iv)* the percentage of people between 0 and 19 years old, *(v)* the percentage of people between 20 and 64 years old, *(vi)* the percentage of people over 64 years old, *(vii)* the proportion of private housing, *(viii)* the number of positive tests, and *(ix)* the number of positive tests per capita.

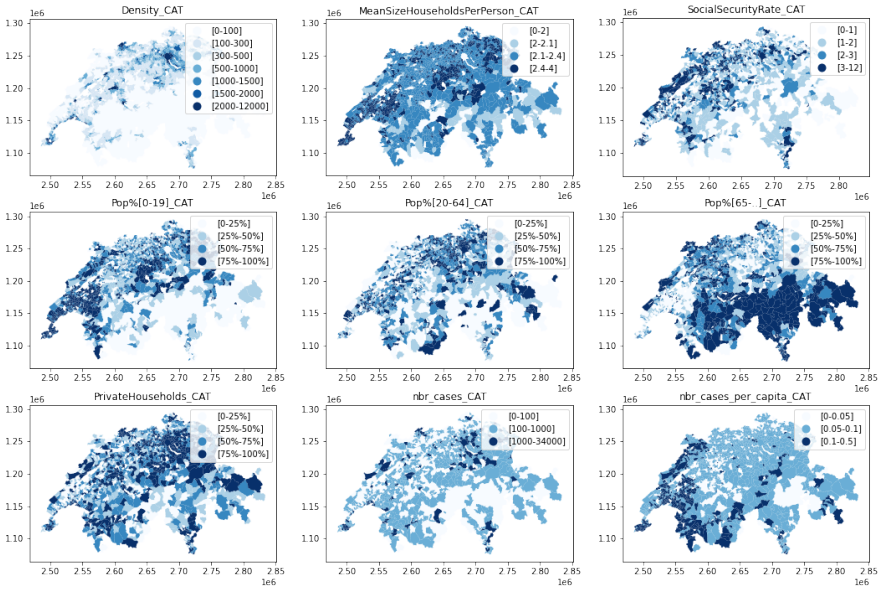
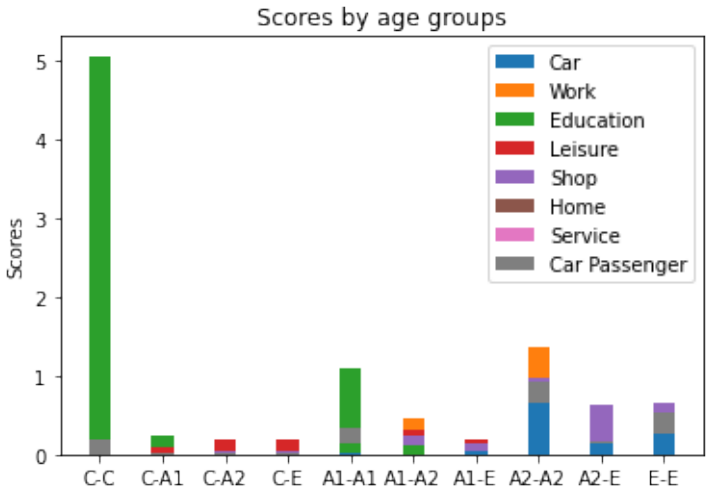
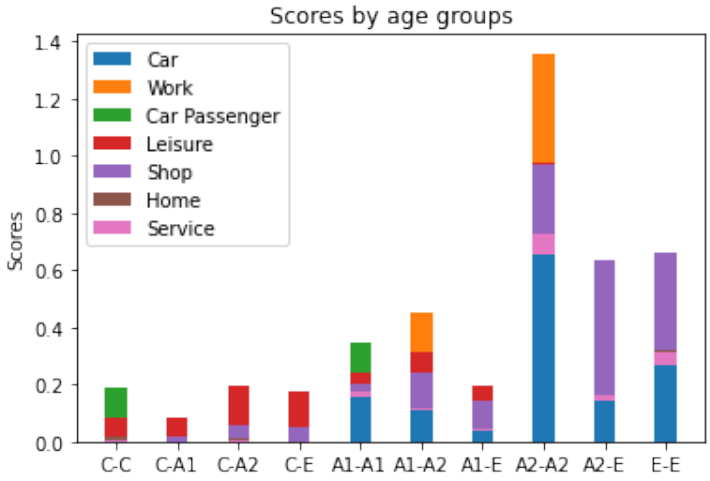


Figure 6: Mean contacts per activities and groups



number of people inside an education facility is very high for a long time (≈ eight hours).

Figure 7: Mean contacts per activities and groups (Education excluded from the list of activities)



To better visualize the contacts by activity and group, we delete the education activity from Figure 6 and re-scale the y-axis, obtaining Figure 7. Here, we observe that the most contacts among adults are during work time, for people under 18 during leisure activities, and for elders during grocery shopping or inside a car. The activity per groups modularity shapes our model to implement policies with high flexibility and dynamism for each age group.

4.3 Parameter estimates

As previously mentioned, we want to: (i) determine the impact that the socio-economic variables have on the probability of infection of an individual ($P(\text{infection})_{i,w}$), (ii) select those variables and proceed with parameter elimination, (iii) estimate the parameters to compute $P(\text{infection})_{i,w}$. To do so, Algorithm 1 is executed for both the first and the second pandemic waves (2020-02-24 to 2020-04-30 and 2020-10-01 to 2021-02-14, respectively). We compute the propensity score for both estimations by running a probit model where the outcome variable is a binary variable indicating infection status, and we verify the results with multiple logistic regression. The results reveal that the two waves share a high correlation with the same set of variables, namely: the age of the individual Λ, percentage of the population above 65 years old for a specific municipality χ, percentage of the population between 20 and 65 years old for a specific municipality Υ and population density per km κ, with the only exception for the income Φ, considerable

as variable only for the second wave. Specifically, the analytic forms obtained as the output of Algorithm 1 for the two waves are:

$$P(\text{infection})_{i,1} \sim \beta_\Lambda \log(\Lambda) + \beta_\chi \log(\chi) + \beta_\Upsilon \log(\Upsilon) + \beta_\kappa \log(\kappa). \tag{17}$$

$$P(\text{infection})_{i,2} \sim \beta_\Lambda \log(\Lambda) + \beta_\chi \log(\chi) + \beta_\Upsilon \log(\Upsilon) + \beta_\kappa \log(\kappa) + \beta_\Phi \Phi. \tag{18}$$

Table 2: Summary statistics of the list of covariates

	1^{st} WAVE			2^{nd} WAVE		
Stratified by infection	0	1	SMD	0	1	SMD
n	269642	414		281576	14499	
$\Lambda(mean(SD))$	44.42 (20.99)	50.91 (20.24)	0.315	44.43(21.05)	43.33(19.68)	0.054
$\Upsilon(mean(SD))$	3.89 (3.27)	65.17 (3.07)	0.0400	63.63(3.30)	63.91(3.25)	0.086
$\kappa(mean(SD))$	2399.74 (1760.49)	3123.01 (1733.89)	0.414	2222.49(1771.15)	2401.70(1751.82)	0.102
$\chi(mean(SD))$	16.89 (2.50)	16.26 (2.29)	0.0264	17.04(2.56)	16.84(2.48)	0.077
$\Phi(mean(SD))$				0.02(0.13)	0.02(0.12)	0.002

In Table 2, we see in the control group 281576 subjects and the treated group 14499 subjects. First, we analyze the means and standard deviations of the different variables. Secondly, we observe that the Standardized Mean Differences (SDM) are lower than 0.1, which means that none of them are showing imbalance. Therefore, we can apply to match. Finally, we compute a propensity score matching analysis and test for non-linearity for the non-binary attributes. We find that the log-transformation reduces skewness and allows us to fulfill the condition $SMD \lesssim 0.1$ (see Table 3).

Table 3: Standardized Mean Difference for the list of variables

	1^{st} WAVE			2^{nd} WAVE		
Stratified by infection	Means Treated	Means Control	SMD	Means Treated	Means Control	SMD
distance	0.00200	0.00150	0.0523	0.0499	0.0499	-0.000
$\log \Lambda(mean(SD))$	3.835	3.646	0.0405	3.654	3.649	0.0104
$\log \Upsilon(mean(SD))$	2.779	2.8160	0.0272	4.156	4.156	0.00210
$\log \kappa(mean(SD))$	4.176	4.156	0.0413	7.427	7.426	0.000700
$\chi(mean(SD))$	7.802	7.427	0.0467	2.813	2.814	-0.00240
$\log \Phi(mean(SD))$				0.0159	0.0104	0.0436

In Table 3, we find the same list of variables as in Table 2. If we look at the SMD, we observe very low values, so we can accept our matching results (see Table 5). Overall, the variables have an outstanding balance as the standardized mean difference is never even close to 0.1.

Moreover, if we compute the algorithm to obtain the estimates of the explanatory variables from Equations (17) and (18), we can observe that β_Λ takes a positive value. This can be explained by the fact that older people are more likely to test positive than children. Nevertheless, it does not imply that mortality is lower. Also, it is interesting to see that a more significant percentage of adults has a higher impact on infection than a more significant percentage of the elderly population for the first wave. On the other hand, we can observe a negative impact on the percentage of elderly and adults during the second wave. The vaccination measures can explain this. Note that by the end of the first wave (end of April 2020), 20% of the population was vaccinated. Lastly, as we obtain high values for the SMD in multiple variables when including the *income* related variable Φ, we can state that the latter is not representative of the first wave.

Table 4: Coefficients using Ordinary Least Squares (OLS) method using Least Squares

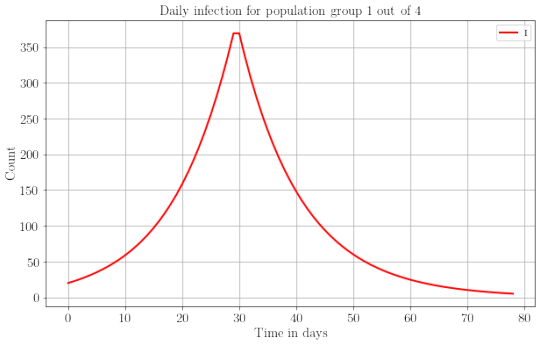
	1^{st} WAVE				2^{nd} WAVE			
Variable	Est.	SE	z-val.	p-val.	Est.	SE	z-val.	p-val.
$\log(\Lambda)$	0.000800	0.000	6.264	0.000	0.00160	0.00100	2.359	0.0180
$\log(\chi)$	0.00170	0.00100	1.661	0.0970	-0.0197	0.00500	-3.918	0.000
$\log(\Upsilon)$	0.0105	0.00400	2.841	0.00500	-0.0800	0.0180	-4.424	0.000
$\log(\kappa)$	0.000400	0.000	2.658	0.00800	0.00890	0.00100	13.988	0.000
Φ					-0.00250	0.00100	-3.124	0.00200

Table 5: Coefficients using Matching score algorithm

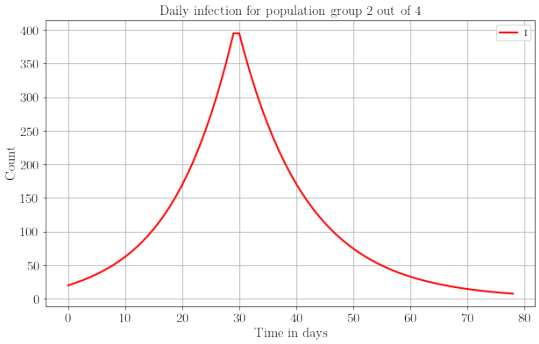
	1^{st} WAVE				2^{nd} WAVE			
Variable	Est.	SE	z-val.	p-val.	Est.	SE	z-val.	p-val.
$\log(\Lambda)$	0.641	0.101	6.321	$2.590\text{e}-10$	0.0301	0.0143	2.100	0.0358
$\log(\chi)$	1.395	0.773	1.806	0.0701	-0.386	0.107	-3.608	0.000309
$\log(\Upsilon)$	7.388	2.893	2.554	0.0106	-2.0381	0.391	-5.211	$1.880e-07$
$\log(\kappa)$	0.281	0.109	2.594	0.00950	0.187	0.0141	13.246	$< 2\text{e}-16$
Φ					-0.0556	0.0261	-2.131	0.0331

4.4 Model fit

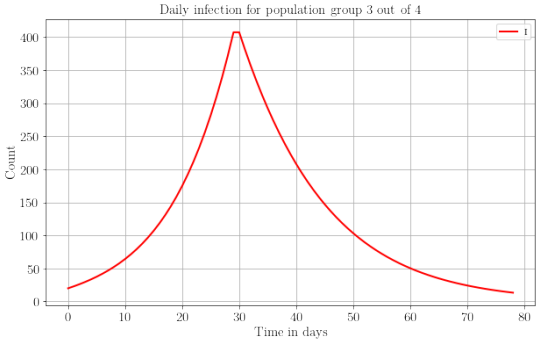
We initialize the discrete integration process (see Algorithm 2, lines 3 to 13) to obtain the total number of cases per timestep per group of population. We test our model to represent the infection dynamics observed during the COVID-19 pandemic. In particular, the case study focuses on the first pandemic wave, therefore considering data from 2020-02-24 to 2020-04-30. The choice of the time period object of study is supported by two



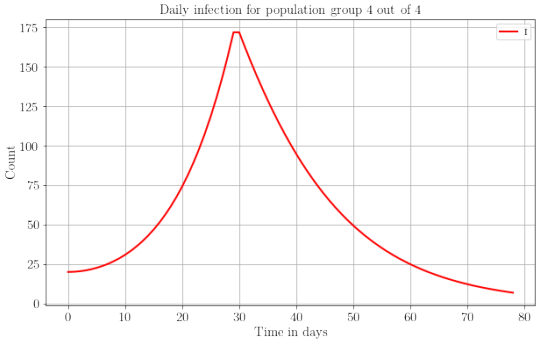
(a) Results SIRD for G_C



(b) Results SIRD for G_{A1}



(c) Results SIRD for G_{A2}



(d) Results SIRD for G_E

Figure 8: Infected population by age group for the first wave

main features that make it suitable to be studied by the proposed model: it has regular social contacts, and the population is not yet vaccinated. Note that we set some activity restrictions in the contact matrix C_a from mid-march to: (i) reduce the force of infection (ii) fit the total number of cases from Google data (see Google, 2022). The evolution of the *Infected* population are shown in Figure 8. In particular, Figures 8(a) to 8(d) show the evolution for the 4 different age groups. Finally, we plot the number of cumulative COVID-19 cases over the same period from the Google data in Figure 9, with the numbers obtained by our epidemiological model initialized accordingly with the initial values of infected individuals from official public data. By comparing the curves, we can state that the developed model is able to capture the evolution of the positive cases with good accuracy.

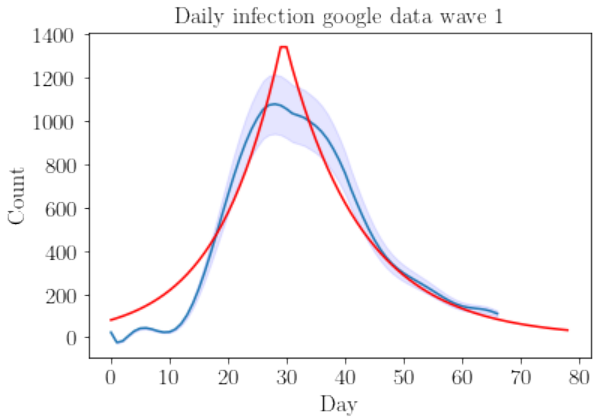


Figure 9: Daily infection from google data for the first wave (blue line) against the aggregated output of our model (red line).

4.5 Comparison with the state-of-the-art

For the sake of comparing the obtained results and assessing the quality of the proposed disaggregation approach, we refer to Müller *et al.* (2020), where an age-dependent progression model from Episim is used. The probability for individual i become infected given its contacts by this process in a time step t is described in Equation 19 as:

$$P_{i,t} = \beta_{i,w}^{\text{Episim}} = 1 - \exp\left[-\theta \sum_{m} q_{m,t}.ci_{im,t} \cdot in_{i,t} \cdot \tau_{im,t}\right] \tag{19}$$

where m indicates an agent other than the studied agent i, θ is a calibration parameter, q the shedding rate (microbial load), ci_{im} the contact intensity between agent i and agent m, in the intake (reduced, e.g., by a mask), and τ_{im} the duration of interaction between the two individuals.

However, in Müller *et al.* (2020) they assume $q = ci = in = 1$ since none of these values are known for COVID-19. We compute the force of infection per group using their methodology and our dataset to compute the mean duration of interaction between the two individuals τ, and the weights per age group to compute $\beta_{g,w}^{\text{Episim}}$. The results are visible in Table 6. The first two columns show the parameters used by Episim in terms of the probability of transitioning to contagious and developing symptoms. Since the Episim model assumes $q = ci = in = 1$, the two columns are populated by the same values. The third column indicates the probability of transitioning to contagious computed with our dataset. This table shows that the Episim algorithm might not provide promising results, especially for lower age groups, as their estimate differs from the real data. It

is important to mention that the Episim model is run for a sample of individuals based in Berlin, Germany, while the dataset refers to Switzerland. This might introduce some deviation in the results but not justify the difference between Episim estimated data and real measurements. As a direct consequence, the values of $\beta_{g,w}^{\text{Episim}}$ and $\beta_{g,w}$ computed by our model are very different, except for the older population group, where Episim uses a more accurate P(contagious) values.

Table 6: Age-dependant progression model from Episim and comparison for the values of the probabilities of infection given contact per age group

Age group	P^{Episim}contagious %	P^{Episim}symptoms %	Weights	$\beta_{g,w}^{\text{Episim}}$	$\beta_{g,w}$
0 to 9	0.1	0.1	22.3	0.034	0.32
10 to 19	0.3	0.3			
20 to 29	1.2	1.2	19.1	0.029	0.19
30 to 39	3.2	3.2			
40 to 49	4.9	4.9	30.7	0.045	0.14
50 to 59	10.2	10.2			
60 to 69	16.6	16.6			
70 to 79	24.3	24.3	27.5	0.042	0.05
80 +	27.3	27.3			

Since the MATSim scenario of Switzerland is not open source, we cannot compute the scenario to compare the infection outputs. For this reason, we decide to use the reduction of the reinfection rate as the parameter to compare to other studies as in Müller *et al.* (2020). In their study, the reinfection rate is defined as:

$$R = \frac{\frac{\text{Reinfection cases with restriction}}{\text{Reinfection cases with no restriction}}}{\text{Total number of individuals}} \tag{20}$$

To compute R, as discussed in Section 3.2, it is sufficient to set to zero elements of the contact matrix C_a and run the model. The obtained results are shown in Table 7 and compared to Brauner *et al.* (2020) Haug *et al.* (2020) and Müller *et al.* (2020), where it is possible to observe that the results obtained with our disaggregated SIRD Model are in line with the ones proposed by Brauner *et al.* (2020).

Table 7: Percent reduction of R in Müller *et al.* (2020)

Measure	Brauner *et al.* (2020)	Haug *et al.* (2020)	Disaggregated SIRD Model
Schools closed	50	16	38
Most businesses suspended	26		27
Work ban	34		36
Gatherings limited to ≤ 1000	16		19
Gatherings limited to ≤ 100	17		21
Gatherings limited to ≤ 10	28		32
Mass gathering cancellation ≥ 50		27	31
Small gathering cancellation ≤ 50		17	22
Event ban			
Gathering ban			
Venue closure			
Stay-at-home order with exemptions	14		12

5 Conclusion

This paper describes the design and evaluation of a semi-disaggregated activity-based model. We aim to create an interdisciplinary bridge between transportation and the epidemiological community. The most significant contributions are: (i) we capture how the socio-economic characteristics of an individual define the force of infection, (ii) we obtain a self-explanatory model, defined by the estimates of the variables that characterize the spreading event, and (iii) high goodness of fit of our model with Google data. Moreover, the effortlessness with which the activities can be modified per population group makes this tool fitting for testing activity restrictions policies. Concerning the performance assessment of the model, the lack of individual data makes it very challenging to have a rigorous analysis of how disaggregation performs in this kind of model. The main reason is that we have to validate our model with aggregated data. Also, we cannot examine the correlation between adding levels of disaggregation and the goodness of fit with the real data. Furthermore, we have defined the population by age and classified them into four groups. We believe it would be interesting to design policies by combined socio-economic and/or activity parameters and allocate the population to them. This multiple categorization will allow for dynamically testing policies. Moreover, this model is suited to work together with an optimization algorithm. However, this extension is out of the scope of this paper.

Future works might include: (i) developing and adding an optimization tool extension to the model. This will allow us to explore different policy strategies and their efficiency. Secondly, (ii) extent the model to the different COVID-19 variants to evaluate its performance and consider the methodology for other non-vector-borne diseases.

6 References

Admin suisse (2022) Admin: Population, `https://www.bfs.admin.ch/bfs/fr/home/statistiques/population.html`. Accessed: 2022-01-05.

Aleta, A., D. Martín-Corral, M. A. Bakker, A. P. y. Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini, A. Pentland, A. Vespignani, Y. Moreno and E. Moro (2020) Quantifying the importance and location of sars-cov-2 transmission events in large metropolitan areas, *medRxiv*.

Anderson, R. and R. Mary (1992) Infectious diseases of humans : dynamics and control / roy m. anderson and robert m. may, *SERBIULA (sistema Librum 2.0)*, 01 1992.

BFS (2022) Bfs municipality open data, `https://www.bfs.admin.ch/bfs/de/home/statistiken/regionalstatistik/regionale-portraets-kennzahlen/gemeinden.html`.

Brauner, J. M., S. Mindermann, M. Sharma, A. B. Stephenson, T. Gavenčiak, D. Johnston, J. Salvatier, G. Leech, T. Besiroglu, G. Altman, H. Ge, V. Mikulik, M. Hartwick, Y. W. Teh, L. Chindelevitch, Y. Gal and J. Kulveit (2020) The effectiveness and perceived burden of nonpharmaceutical interventions against covid-19 transmission: a modelling study with 41 countries, *medRxiv*.

Chang, S., E. Pierson, P. Koh, J. Gerardin, B. Redbird, D. Grusky and J. Leskovec (2021a) Mobility network models of covid-19 explain inequities and inform reopening, *Nature*, **589** (7840) 82–87, ISSN 0028-0836.

Chang, S., E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky and J. Leskovec (2021b) Mobility network models of covid-19 explain inequities and inform reopening, *Nature*, **589** (7840) 82–87, Jan 2021, ISSN 1476-4687.

Chang, S., E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky and J. Leskovec (2021c) Mobility network models of covid-19 explain inequities and inform reopening, *Nature*, **589** (7840) 82–87.

Choisy, M., J.-F. Guégan and P. Rohani (2007) *Mathematical Modeling of Infectious Diseases Dynamics*, chap. 22, 379–404, John Wiley & Sons, Ltd, ISBN 9780470114209.

Diekmann, O. and J. Heesterbeek (2000) Mathematical epidemiology of infectious dis-

eases: Model building, analysis and interpretation, *Wiley Series in Mathematical and Computational Biology, Chichester, Wiley*, 01 2000.

Douglas, M., S. V. Katikireddi, M. Taulbut, M. McKee and G. McCartney (2020) Mitigating the wider health effects of covid-19 pandemic response, *BMJ*, **369**.

DuPre, N. C., S. Karimi, C. H. Zhang, L. Blair, A. Gupta, L. M. A. Alharbi, M. Alluhibi, R. Mitra, W. P. McKinney and B. Little (2021) County-level demographic, social, economic, and lifestyle correlates of covid-19 infection and death trajectories during the first wave of the pandemic in the united states, *Science of The Total Environment*, **786**, 147495, ISSN 0048-9697.

European Institute for Gender Equality (2021) European Institute for Gender Equality, `https://eige.europa.eu/covid-19-and-gender-equality/essential-workers`. Accessed: 2022-01-22.

Google (2022) Open data repository on COVID-19, `https://health.google.com/covid-19/open-data/`. Accessed: 2022-01-05.

Hackl, J. and T. Dubernet (2019) Epidemic spreading in urban areas using agent-based transportation models, *Future Internet*, **11**, 92, 04 2019.

Haug, N., L. Geyrhofer, A. Londei, E. Dervic, A. Desvars-Larrive, V. Loreto, B. Pinior, S. Thurner and P. Klimek (2020) Ranking the effectiveness of worldwide covid-19 government interventions, *medRxiv*.

Heyde, C. C. and E. Seneta (2001) *Statisticians of the Centuries*, Springer.

Hörl, S. and M. Balac (2021) Synthetic population and travel demand for paris and Île-de-france based on open and publicly available data, *Transportation Research Part C: Emerging Technologies*, **130**, 103291, ISSN 0968-090X.

Kelman, A. (1985) Compartmental models and their application, *International Journal of Bio-medical Computing - INT J BIO MED COMPUT*, **16**, 294–295, 05 1985.

Klein, S. L. and K. L. Flanagan (2016) Sex differences in immune responses, *Nature Reviews Immunology*, **16** (10) 626–638.

Lee, M. and M. You (2020) Psychological and behavioral responses in south korea

during the early stages of coronavirus disease 2019 (covid-19), *International Journal of Environmental Research and Public Health*, **17** (9), ISSN 1660-4601.

Lemaitre, J., J. Perez-Saez, A. Azman, A. Rinaldo and J. Fellay (2020a) Assessing the impact of non-pharmaceutical interventions on sars-cov-2 transmission in switzerland, *Swiss Medical Weekly*, **150**, 05 2020.

Lemaitre, J. C., J. Perez-Saez, A. S. Azman, A. Rinaldo and J. Fellay (2020b) Assessing the impact of non-pharmaceutical interventions on SARS-CoV-2 transmission in switzerland, *Swiss Med. Wkly*, **150**, w20295.

Mancastroppa, M., R. Burioni, V. Colizza and A. Vezzani (2020a) Active and inactive quarantine in epidemic spreading on adaptive activity-driven networks, *Physical Review E*, **102** (2), aug 2020.

Mancastroppa, M., R. Burioni, V. Colizza and A. Vezzani (2020b) Active and inactive quarantine in epidemic spreading on adaptive activity-driven networks, *Physical Review E*, **102** (2), Aug 2020, ISSN 2470-0053.

Müller, S. A., M. Balmer, B. Charlton, R. Ewert, A. Neumann, C. Rakow, T. Schlenther and K. Nagel (2020) Using mobile phone data for epidemiological simulations of lockdowns: government interventions, behavioral changes, and resulting changes of reinfections, *medRxiv*.

Oertelt-Prigione, S. (2020) *The impact of sex and gender in the COVID-19 pandemic*, European Union.

Qian, X. and S. V. Ukkusuri (2021a) Connecting urban transportation systems with the spread of infectious diseases: A trans-seir modeling approach, *Transportation Research Part B: Methodological*, **145**, 185–211, ISSN 0191-2615.

Qian, X. and S. V. Ukkusuri (2021b) Connecting urban transportation systems with the spread of infectious diseases: A trans-seir modeling approach, *Transportation Research Part B: Methodological*, **145**, 185–211, ISSN 0191-2615.

Riou, J., R. Panczak, C. Althaus, C. Junker, D. Perisa, K. Schneider, N. Criscuolo, N. Low and M. Egger (2021a) Socioeconomic position and the cascade from sars-cov-2 testing to covid-19 mortality: Population-based analysis of swiss surveillance data, *SSRN Electronic Journal*, 01 2021.

Riou, J., R. Panczak, C. Althaus, C. Junker, D. Perisa, K. Schneider, N. Criscuolo, N. Low and M. Egger (2021b) Socioeconomic position and the cascade from sars-cov-2 testing to covid-19 mortality: Population-based analysis of swiss surveillance data, *SSRN Electronic Journal*, 01 2021.

Riou, J., R. Panczak, C. L. Althaus, C. Junker, D. Perisa, K. Schneider, N. G. Criscuolo, N. Low and M. Egger (2021c) Socioeconomic position and the covid-19 care cascade from testing to mortality in switzerland: a population-based analysis, *The Lancet Public Health*, **6** (9) e683–e691.

Roelens, M., A. Martin, B. Friker, F. M. Sousa, A. Thiabaud, B. Vidondo, V. Buchter, C. Gardiol, J. Vonlanthen, C. Balmelli, M. Battegay, C. Berger, M. Buettcher, A. Cusini, D. Flury, U. Heininger, A. Niederer-Loher, T. Riedel, P. Schreiber and O. Keiser (2021) Evolution of covid-19 mortality over time: results from the swiss hospital surveillance system (ch-sur), *Swiss Medical Weekly*, 09 2021.

Sannigrahi, S., F. Pilla, B. Basu, A. S. Basu and A. Molter (2020) Examining the association between socio-demographic composition and covid-19 fatalities in the european region using spatial regression approach, *Sustainable Cities and Society*, **62**, 102418, ISSN 2210-6707.

Singu, S., A. Acharya, K. Challagundla and S. N. Byrareddy (2020) Impact of social determinants of health on the emerging covid-19 pandemic in the united states, *Frontiers in public health*, **8**, 406.

Smieszek, T. (2009) A mechanistic model of infection: Why duration and intensity of contacts should be included in models of disease spread, *Theoretical biology and medical modelling*, **6**, 25, 11 2009.

Soper, H. (1929) The interpretation of periodicity in disease prevalence, *J. R. Statist. Soc*, **A 92**, 34–73, 01 1929.

Tirachini, A. and O. Cats (2020) Covid-19 and public transportation: Current assessment, prospects, and research needs, *Journal of Public Transportation*, **22**, 07 2020.

Tuomisto, J. T., J. Yrjölä, M. Kolehmainen, J. Bonsdorff, J. Pekkanen and T. Tikkanen (2020) An agent-based epidemic model reina for covid-19 to identify destructive policies, *medRxiv*.

Zheng, R., Y. Xu, W. Wang, G. Ning and Y. Bi (2020) Spatial transmission of covid-19 via public and private transportation in china, *Travel Medicine and Infectious Disease*, **34**, 101626, ISSN 1477-8939.