Disaggregate modeling and policy optimization for the Swiss SARS-CoV-2 outbreak

Cloe Cortes Balcells *   Rico Krueger †   Michel Bierlaire *

November 25, 2022

Report TRANSP-OR 221025
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Federale de Lausanne
transp-or.epfl.ch

*Ecole Polytechnique Federale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland, {cloe.cortesbalcells, michel.bierlaire}@epfl.ch

†Department of Technology, Management and Economics; Technical University of Denmark (DTU), Denmark

Abstract

This work proposes an epidemiological model for the SARS-COV-2 outbreak and an optimization tool to guide policy decision-making. The model accounts for the socioeconomic characteristics of the population and combines mobility information with epidemiological data. We demonstrate the added value of using disaggregate Susceptible-Infected-Recovered-Dead models (SIRD) for analizing SARS-CoV-2 spreading and evaluating the potential of these models to study optimal policy decision-making. We propose a methodology to couple agent-based epidemiological models with Activity-Based Models (ABMs) to capture the heterogeneous activity-travel behaviors in the population. The heterogeneity allows for determining the influence of activity-travel behavior on the pandemic spread and the efficacy of restrictions on specific activities. We compute the probability of infection by modeling it as a function of the most influencing socioeconomic variables per individual. The ABM captures the heterogeneity of activity-travel behavior. Finally, we propose an optimization approach for determining the best policy of activity restrictions to be applied to maximize economic or medical objectives. The policy applied by the Swiss government in the Spring of 2020 is compared with the output of the optimizer, showing how taking actions in the early stage of the pandemic is of fundamental importance.

1 Introduction

In recent years, COVID-19 has reinforced the importance of epidemiological models to study of how diseases spread. Analyses of SARS-COV-2 data, together with projections on the virus's spread, have driven policy decisions all over the world. In particular, the pandemic outbreak of SARS-COV-2 drew attention on the link between transportation and epidemiological research since human mobility is one of the leading causes of the spread of the virus (Tirachini and Cats, 2020, Douglas et al., 2020). This pandemic has revealed the lack of literature that combines the epidemiological field with transportation planning, including activity-travel behavior when studying disease spreading. Coupling mobility and epidemiological data can provide a better spreading model (Tuomisto et al., 2020). Travel-behavior data is of fundamental importance for planning non-pharmaceutical interventions (NPIs), including public transportation scheduling and the partial restriction of people's daily activities (Tirachini and Cats, 2020, Douglas et al., 2020,Zheng et al., 2020 ,Lee and You, 2020).

In the 18th century, the first epidemiological model was used to study the morbidity and mortality of smallpox (Heyde and Seneta, 2001). Thenceforth, multiple models have been developed to explain an epidemic's spread. These approaches range from elementary compartmental models (e.g Kermack et al., 1927) to network models and the latest and more complex approaches: individual or agent-based models (see Eubank et al., 2004 amd Mancastroppa et al., 2020). Deterministic compartmental models (Kelman, 1985), also known as SIR models, depend on differential equations. These equations define the flow dynamics between the different compartments (Susceptible, Infected, Recovered). Deterministic models are unable to represent the social structure relevant to the disease spreading, as they are fully aggregated. Network models can partially cover this gap. Being based on graph theory (Albert and Barabasi, 2002), network models rely of nodes and links representing hosts and their contacts. However, they neglects the characteristics of the contacts and of the individuals that took part in it. This information is captured by agent-based models (Muller et al., 2020), our main focus in developing the SIRD disaggregate model. Agent-based models work by following people over time at an individual level through the different stages of the disease. Agents are the unit of analysis, they act on their own and interact with the environment. Nevertheless, due to the lack of fully disaggregate data, this technique faces the issue of adding aggregated parameters to the model. This results in computationally complex models for the level of final disaggregation reached. For this reason, we study how to couple semi-disaggregate epidemiological and activity-based models to link epidemics and mobility. 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 through disaggregating the model allows for determining the influence of activity-travel behavior on infection, mortality rates, and the efficacy of restrictions on specific activities. We start with a simple SIRD model and disaggregate it to reach the heterogeneity level desired. Multiple authors (see Singu et al., 2020 and Commission, 2020) state the importance of including the influence of individual characteristics like age or income on the spread. Other studies analyze the correlation between positive SARS-CoV-2 tests, mortality rates, and admission to intensive care with socio-economic position (Riou et al., 2021a). Considering socio-economic characteristics such as gender and health care information, or virological characteristics like exposure to the virus, allows a better understanding of the spreading (e.g Commission, 2020). Therefore, it is essential to include variables in the epidemiological model that capture the heterogeneity of the population's behavior. The two main challenges when using activity-based models are, on the one hand, mobility clustering the population to the locations of their activities which leads to contact and contagion, and on the other hand, accounting for heterogeneity (Qian and Ukkusuri, 2021). The final goal is to guide public health authorities in dealing with epidemic situations.

In this work, we propose a model that addresses the heterogeneity issue by disaggregating a traditional epidemiological model to account for the characteristics of the population. 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 for modeling the probability of infection for an individual (Chang et al., 2021), (ii) to obtain computationally efficient modeling compared to agent-based models with the equivalent level of disaggregation on the initial dataset (Tuomisto et al., 2020), (iii) to avoid the use of real–time data-driven interventions and instead rely on model-based solutions (Aleta et al., 2020), and (iv) to make the probability of transmission time-dependent to account for the possibility of implementing policies during a specific moment in time (Mancastroppa et al., 2020). This paper's scope is to demonstrate the added value of using disaggregate models for modeling SARS-CoV-2 spreading and to evaluate the potential of the proposed model-based optimization tool for the implementation of NPIs. We employ a generalized multivariate regression model to compute the probability of infection depending on the socio-economic characteristics of the individuals and a Variable Neighborhood Search (VNS) to solve a multi-objective optimization problem to obtain how and when NPIs should be applied. This method incorporates population heterogeneity, behavior, and contact patterns.

2 Methodology

2.1 Dataset specification

To determine the influence of activity-travel behavior on the pandemic spread and the efficacy of restrictions on specific activities, our methodology relies on a dataset containing each individual:

- Daily activity plans;
- Socio-economic and health-related information.

We propose to capture the influence of human mobility and heterogeneity of the population's behavior on the SARS-CoV-2 by linking every human interaction (i.e., contact) with its location and the information on the individuals that took part in it. It is out of the scope of this paper to propose a standardized way to aggregate available data into a suitable dataset. Nevertheless, we discuss in the results section the procedure followed for this study (considering 25% of the Swiss population).

2.2 Activity specification

To capture the influence of human mobility on SARS-CoV-2, we classify contacts between individuals by the type of activity (e.g., leisure, work, education, shopping). We propose to segment of the population into groups characterized by specific features, i.e., age, gender, income, or municipality. The choice of the type and the total number of groups G is twofold. Firstly, the group choice can allow for a better spreading model, capturing correlations between a specific age group or income level with the number of infections. Secondly, once the SIRD model is trained, it is possible to study the implementation of policies for the different population groups. For every activity a, the contacts are represented by the *contact matrix* $\mathbb{C}_a$ (see Equation (1)) such that: A is the total number of considered activities by the study and the term $C_{a_{ij}} = C_a(g_i, g_j)$ is to be interpreted as the number of contacts between individuals of the groups g_i and g_j $\forall i, j \in [1, G]$. The sum of all elements in $\mathbb{C}_a$ is the total number of contacts in a single day, considered representative of the pandemic period:

$$\mathbb{C}_a = \begin{pmatrix} C_a(g_1, g_1) & C_a(g_1, g_2) & \cdots & C_a(g_1, g_G) \\ C_a(g_2, g_1) & C_a(g_2, g_2) & \cdots & C_a(g_2, g_G) \\ \cdots & \cdots & \cdots & \cdots \\ C_a(g_G, g_1) & C_a(g_G, g_2) & \cdots & C_a(g_G, g_G) \end{pmatrix}_a \qquad \forall a. \tag{1}$$

2.3 Probability of infection per individual

Similarly to agent-based models, we model the probability of infection per individual. In addition, we include medical and socio-economic characteristics and evaluate their influence on the SARS-CoV-2 spread through multivariate logistic regression. We are interested in modeling the probability of contracting the virus for an individual. As available datasets refer to testing results, we rely on the assumption that all infected people get tested. For each individual i, we consider his socio-economic characteristics and the characteristics of the environment that surrounds him. We define the probability of infection as:

$$P(\text{infection})_i \sim \boldsymbol{X_i^m} \cdot \boldsymbol{\alpha^m} + \boldsymbol{X_i^p} \cdot \boldsymbol{\alpha^p} = \sum_{j=1}^{J} \alpha_j^m X_{j,i}^m + \sum_{k=1}^{K} \alpha_k^p X_{k,i}^p, \tag{2}$$

where:

- $\boldsymbol{X_i^m} = [X_1^m, X_2^m, \ldots, X_J^m]_i^T$ are the socio-economic characteristics of the individual i,
- $\boldsymbol{X_i^p} = [X_1^p, X_2^p, \ldots, X_K^p]_i^T$ are aggregated indicators of the surroundings of i, i.e. of the municipality where i lives,

- $\boldsymbol{\alpha^m} = [\alpha_1^m, \alpha_2^m, \ldots, \alpha_J^m]$ and $\boldsymbol{\alpha^p} = [\alpha_1^p, \alpha_2^p, \ldots, \alpha_K^p]$ estimate the variable's parameters.

We propose to select the variables $\boldsymbol{X_i^p}, \boldsymbol{X_i^m}, \boldsymbol{\alpha^p}, \boldsymbol{\alpha^m}$ by applying the Generalized Logistic Model Regression (GLM) procedure described by Table 1.

Algorithm 1 Procedure for selecting variables for GLM regression

```
1:  Initialize the Algorithm with an educated guess for the variables X^p_{1,2,...}, X^m_{1,2,...}
    to consider for the analysis
2:  for each variable X^p_{1,2,...,J}, X^m_{1,2,...,K} do
3:      Compute the Variance Inflation Factor (VIF) and the p-value
4:      if p − value < 0.05 & VIF < 5 then
5:          Remove the variable and refit the model (line 1)
6:      else if all p-values and VIF values are accepted with 95% confidence then
7:          Keep the last fitted model (line 5)
8:      end if
9:  end for
10: Output the results)
```

2.4 Probability of infection per group

Once the probability of infection per individual i is defined, according to Equation (2), we aggregate the results to obtain the probability of infection β_g per group g. To obtain β_g, we first calculate the probability of entering in contact with an infected individual $P(\text{contact}|\text{infection})$ and the probability of contact $P(\text{contact})$ from the activity-based model output. From the daily activity schedule of the population we compute statistics on the contacts inside the different facilities (Equations 3 and 4). To reduce computational complexity, $P(\text{contact})$ and $P(\text{contact}|\text{infection})$ are considered to be constant for all individuals and facilities. To obtain the probability of infection given a contact per individual β_i, we apply the Bayes theorem as of Equation (5). Finally, we aggregate the result to obtain the probability of infection per group β_g by averaging the probability of infection given a contact per individual over the population of the group as in Equation (6).

$$P(\text{contact}) = \frac{\text{number of contacts in all facilities}}{\text{number of facilities}} \tag{3}$$

$$P(\text{contact}|\text{infection}) = \frac{\text{total infected people in all facilities} \cdot \text{number of facilities}}{\text{total number of people}} \tag{4}$$

$$\beta_i = P(\text{infection}|\text{contact})_i = \frac{P(\text{contact}|\text{infection}) \cdot P(\text{infection})_i}{P(\text{contact})} \quad \forall \text{ individuals } i \tag{5}$$

$$\beta_g = P(\text{infection}|\text{contact})_g = \frac{\sum_{\forall i \in g} \beta_i}{N_g} \qquad \forall \text{ groups } g \tag{6}$$

N_g is the total number of individuals that belong to the group g. Equation (6) relies on the assumption that β_i for all the individual in group g are mostly similar. This depends on the group choice, which therefore affect the accuracy of the model.

2.5 Force of infection

The force of infection λ in SIRD models describes the transition rate from the compartment of susceptible individuals to the compartment of infectious individuals. We compute it for every group g as follows:

$$\lambda_g(t) = \beta_g \sum_{a=1}^{A} \Theta_{a,g}(t)\, \mathbb{C}_a, \tag{7}$$

where $\lambda_g(t)$ is the force of infection of each individual in group g and $\Theta_{a,g}(t)$ represents the percentage decrease for every activity due to policy application. This term is used to model the effect of policies limiting the mobility of the population. For example, a policy imposing online schooling for a specific age group, is represented by $\Theta_{\text{Education, Age group}} = 0$.

2.6 SIRD specification

We introduce $\lambda_g(t)$ inside the SIRD model for group g, alongside external parameters related to the virus strain, the recovery rate γ_g, and the probabilities of transmission during contact. The SIRD model predicts the number of cases over time for each segment of the population (group g) according to the following Ordinary Differential Equations (ODEs):

$$
\begin{aligned}
\frac{dS_g}{dt}(t) &= -\lambda_g(t)\, I_g(t)^T \frac{S_g(t)}{N}, && (8)\\
\frac{\partial I_g}{\partial t}(t) &= \lambda_g(t)\, I_g(t)^T \frac{S_g(t)}{N} - \gamma_g I_g(t), && (9)\\
\frac{\partial R_g}{\partial t}(t) &= \gamma_g I_g(t)(1-\mu_g), && (10)\\
\frac{\partial D_g}{\partial t}(t) &= \gamma_g \mu_g I_g(t). && (11)
\end{aligned}
$$

Where the terms S, I, R, D, and N stand for the susceptible, infectious, recovered and dead individuals, and the total number of individuals. The ODEs (8)–(11) system can be solved using the *lsoda* (Hindmarsh and Petzold, 2005) 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 $g \in [1, \dots, G]$ where G is the total number of groups in the population P. The process is summarized in Algorithm 2.

Algorithm 2 Summary of the proposed methodology

Require: Dataset with individual information and daily activities schedule

Execute the GLM Algorithm
Compute the contact matrices $\mathbb{C}_a$ for all the possible activities $a \in [1, \dots, A]$
for each population group g **do**
    for each individual $i \in g$ **do**
        Compute the probability of infection per individual β_i with Equation (2)
    end for
    Compute the probability of infection per group β_g and torce of infection $\lambda_g(t)$ with Equation (6)-(7)
    Solve the system of ODEs (8)–(11)
end for
Output the SIRD model

2.7 Optimal policy control

As discussed in the introduction, the final purpose of disaggregating the SIRD models is to assess the policy to be applied to maximize economic or medical objectives. For this reason, we introduced the quantity Θ_a in Equation (7), representing the decrease of contacts per activity due to applying a specific policy. In doing so, we obtained a policy-dependent SIRD model suitable for being used in an optimization problem. The decision variables of the problem are the type of policy to be applied $\mathbf{\Theta}$, the starting day of the application t_{start} , and its duration Δt. We consider the vector $\mathbf{\Theta} = [\Theta_1, \Theta_2, \ldots, \Theta_a, \ldots, \Theta_A]$ as the activity reduction for the different activities. Each element of the latter can assume any value between 0, i.e. the strictest policy possible, and 1, i.e. no policy applied. For example, the complete closure of schools and universities is represented by imposing $\Theta_{education} = 0$. For simplicity, we consider the starting day t_{start} and the duration of the policies Δt to be unique, meaning all the interventions are to be applied at the time t. Nonetheless, this method can be extended to include the possibility of applying different interventions at different times. The decision variables are constrained as follows:

$$0 \le t_{start} \le t_{end} \tag{12}$$

$$0 \le t_{end} - t_{start} \tag{13}$$

$$0 \le \ \Theta_a \ \le 1 \quad \forall a \in [1, \ldots, A]. \tag{14}$$

where t_{end} represents the end of the simulated period. Each policy has a twofold effect: it helps contain the spread of the pandemic by reducing the number of contacts but on the other hand, it negatively affects the country's economy. For this reason, we target an optimal policy that minimizes two costs:

1. A health cost H, computed as the death toll of the pandemic,
2. An economic cost E resulting from people unemployed due to the lockdown policy or from people unable to work because infected.

The health cost H can be computed as:

$$H = \text{sum}\{\mathbf{D}\} = \sum_{\forall g} D_g, \tag{15}$$

where D_g is the death toll of group g, the output of the SIRD model when subject to the policy vector $\mathbf{\Theta}$, as described in Equation (11). The economic cost E is modeled as the decrease of the GDP due to diseased and dead workers together with people unemployed due to the lockdown policy (Colas et al., 2021) by means of the Cobb-Douglas function:

$$E = E(\mathbf{I}, \mathbf{D}) = \frac{1}{T_{end}} \sum_{t=1}^{T_{end}} \left[Y_0 - A K_0^{\gamma_k} \left(\mathbf{L}(\mathbf{t})^{1-\gamma_k} \right) \right] \frac{1}{Y_0}, \tag{16}$$

where Y_0 is the GDP before the pandemic, the Cobb-Douglas function $AK_0^{\gamma_k} \left(L(t)^{1-\gamma_k} \right)$ estimates the GDP at any time t during the pandemic and $L(t)$ is the active population employed. The constants A, K_0, γ_k, are the exogenous technical progress, the initial capital stock and the capital elasticity, respectively. The values for the case of Switzerland are shown in the result section. To capture the effect of lockdown policies on the active employed population $L(t)$, we propose to express the latter as:

$$L(t) = [1 - u(t)] \cdot ER \cdot (N - G(t)), \tag{17}$$

where $u(t)$ is the level of partial unemployment due to the lockdown policies, $G(t) = I(t) + D(t)$ is the size of the ill or dead population output by the SIRD model and ER is the employment rate . We relate the variable $u(t)$ to the reduction of the percentage due to the lockdown policy:

$$u(t) = (\mathbf{1} - \mathbf{\Theta}) \cdot \mathbf{c}\frac{\Delta t}{365}, \tag{18}$$

where $\mathbf{c}$ is the GDP contribution vector representing the contribution of each activity to the national GDP. In particular, for the days t when no policy is applied, then $\mathbf{1} = \mathbf{\Theta}$ and therefore $u(t) = 0$ since there is no activity reduction. Finally, we propose to solve the following multi-objective optimization problem:

$$\begin{aligned} &\min_{t_{start}, \Delta t, \mathbf{\Theta}} \begin{bmatrix} H \\ E \end{bmatrix} \\ &\text{subject to:} \\ &\text{Cost functions (15)-(18),} \\ &\text{SIRD model (8)-(11),} \\ &\text{Decision variable constrains (12)-(14).} \end{aligned} \tag{19}$$

We are interested in obtaining the Pareto frontier for the multi-objective optimization problem (19) as we aim to provide the authorities with a set of optimal solutions, to guide the policy choice. We propose to solve (19) using the multi-objective Variable Neighborhood Search algorithm introduced by Ortelli et al., 2021, and available in the PandasBiogeme package for Python (Bierlaire, n.d.). The focus is to use a metaheuristic approach to solve the multi-objective problem. VNS's main advantage is the flexibility to define the list of neighborhood structures. This algorithm starts from a candidate solution and then iteratively moves to a neighbor solution. A neighborhood is the set of all potential solutions that differ from the current solution to the minimal possible extent. Therefore the algorithm requires a neighborhood relation to be defined on the search space. For each element of $\mathbf{\Theta}$, we assign an increasing and decreasing operator. Policy's duration operators work by increasing or decreasing the length of the policy, and by shifting the policy in time with fixed length. This algorithm searches for a local optimum for a given neighborhood structure, and when it finds it, it continues with another structure. The result of the algorithm returns a set of optimal policies that, according to the authors, can help guide the authorities responsible for managing a pandemic crisis or adapting to a post-pandemic situation. In particular, by choosing different points on the Pareto frontier the authorities can prioritize or not the economical stability of the country over the national health. Nevertheless, while we want to provide a tool for the authorities to choose the right policy, it is out of the scope of this contribution to take part in such a choice.

3 Results

The proposed approach is validated on data concerning 25% of the Swiss population. We segment the population 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. It allows to:(i) study the correlation of the probability that an individual gets infected given its socio-economic characteristics, (ii) evaluate NPIs policies. The dataset requirements as input for the SIRD disaggregate model include activity and medical information about the individuals. Since no synthetic population in the literature includes all the needed features, we compute a matching algorithm to combine different datasets. The underlying reason is that we need to account for each individual's daily activitiy plan, socio-economic characteristics, and SARS-COV-2 medical-related information. For this reason, we manipulate data from the Federal Office of Public Health (FOPH) from mid-February 2020 to mid-September 2021 (Riou et al., 2021b). The dataset contains the positive tests in Switzerland and information about the tested individuals. It includes age, gender, municipality, vaccination doses, hospitalization, and causalities. We include open-source data (of Public Health, 2020) 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. Finally, we match the FOPH and the municipality data with a calibrated MATSim simulation output from ETH Zurich (Horl and Balac, 2021). The final dataset is 2M individuals and contains three individual socio-economic characteristics (sex, age, and municipality) and forty-one variables at the municipality level (Office, n.d.). We merge the FOPH Data with the municipality data to obtain the distribution of positive tests in Switzerland. This analysis gives an overview of the disparities in infection in the different municipalities, which makes these variables significant to model the positive tests of the individuals. To provide an example of such variables we show in Figure 1 the population density and the percentage of young inhabitants per municipality, together with the number of cases per capita.

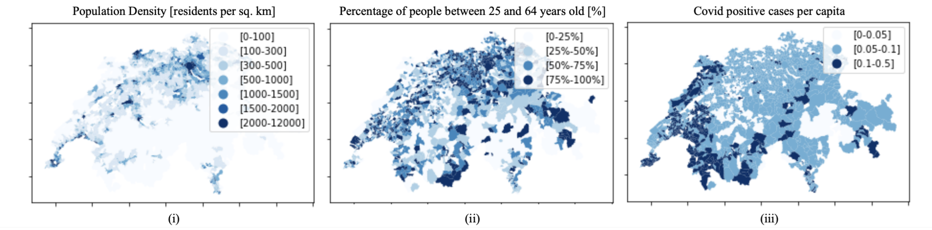

Figure 1: Distribution of positive tests against their location in Switzerland, in addition to some data concerning the municipalities to analyze: *(i)* the population density, expressed in number of residents per km square, *(ii)* the percentage of people between 20 and 64 years old, *(iii)* the number of positive tests per capita.

Activity contact matrix

As seen in Equation (7), the force of infection is a vector whose dimension depends on the segmentation of the population. Since *Age* is the explanatory variable with the highest correlation to an individual's infection probability, we stratify the model into four age groups. P_C, which contains 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 ($G = P_C, P_{A1}, P_{A2}, P_E$). We consider seven activity categories: (i) home, (ii) work, (iii) leisure, (iv) service, (v) education, (vi) shop and (vii) car. The policy vector $\mathbf{\Theta}$, introduced in Equation (7) will assume the following form:

$$\mathbf{\Theta} = [\Theta_{home} \quad \Theta_{work} \quad \Theta_{leisure} \quad \Theta_{service} \quad \Theta_{education} \quad \Theta_{shop} \quad \Theta_{car}]. \tag{20}$$

The contact matrix $\mathbb{C}_a$ defined in Equation (1) encodes the number of contacts between and among each of the four age groups g and 7 activities a per time step. By analyzing the mean of contacts per activity between each age group, we observe that most contacts occur inside *Education*. The number of people inside this facility is very high for a long time (approximately eight hours), explaining the outlier in this activity. Also, the highest number of contacts among adults are during work time, for people under 18 during leisure activities, and for elders during grocery shopping or inside a car.

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 ($P(\text{infection})_i$), (ii) select those variables and proceed with parameter elimination, (iii) estimate the parameters to compute $P(\text{infection})_i$. The GLM algorithm computes the logistic regression where the outcome variable is a binary variable indicating infection status. The result reveals that out of the 41 considered variables, six are meaningful for modeling the probability of infection per individual:

$$P(\text{infection})_i \sim \alpha_\Lambda \Lambda + \alpha_\chi \chi + \alpha_\Upsilon \Upsilon + \alpha_\kappa \log(\kappa) + \alpha_\phi \log(\phi) + \alpha_\epsilon \epsilon, \qquad (21)$$

where the variables Λ, χ, Υ, κ, ϕ and ϵ are described in Table 1, together with their coefficients α_Λ, α_χ, α_Υ, α_κ, α_ϕ and α_ϵ, their standard error and their p-value. Specifically, $\boldsymbol{X^m} = [\Lambda \quad \kappa \quad \epsilon]$ are the socio-economic characteristics of the individual, and $\boldsymbol{X^p} = [\chi \quad \Upsilon \quad \phi]$ are aggregate indicators of the surroundings (in our case, the municipality).

Variable	**Description**	**Coeff.** (α)	**Std err**	**P > |z|**
Λ	Age	-0.0056	0.000	0.000
χ	Urban area	0.0501	0.010	0.000
Υ	Population density per km^2	0.0143	0.008	0.050
κ	Household income	0.9325	0.082	0.000
ϕ	Percentage of population between 20 and 55 years old	0.0224	0.004	0.000
ϵ	Employed	0.3470	0.008	0.000

Table 1: Coefficients using an Iteratively Reweighted Least Squares GLM

Model validation

We structure the validation of our methodology in three blocks to assess the capability of the proposed model in capturing: (i) the trend of age-dependency in the probability of infection given contact, (ii) the aggregated daily infections of the first pandemic wave, and (iii) the dependency of the pandemic on activity-based behavior. For the first part of the validation, we apply Equations 3-7 and compute β_g and the reproduction rate per group R_{0g}, defined as:

$$R_{0g} = \lambda_g / \gamma_g. \qquad (22)$$

By observing the probability of infection given contact per group, also known as susceptibility in the literature, we can show how our methodology can capture this parameter's age dependence. Many studies like Davies et al., 2020, and Goldstein et al., 2021 present a positive correlation between the age of the individuals and their probability of infection given contact. In particular, the literature suggests (see Davies et al., 2020) that the susceptibility to infection in individuals under 20 years of age is approximately half that of adults aged over 20 years, in full accordance with the results obtained with our methodology. Moreover, it is possible to observe how the values for the parameters R_0 have a similar trend, as also discussed by many literature contributions (see Goldstein et al., 2021). The values of

Age Group	Probability of infection given contact per group (β_g)	Basic reproduction number per group (R_{0g})
Children (C)	0.172	0.754
Adults1 (A1)	0.341	1.467
Adults1 (A2)	0.440	3.634
Eldery (E)	0.562	4.981

Table 2: Probability of infection and basic reproduction number per group

R_0 from Table 2 are in agreement with some authors (Goldstein et al., 2021) and comparable to the R_0 of neighboring countries at the beginning of the pandemic.

To further validate the model, we compare the aggregated number of cases generated from its output with the registered cases recorded by an external data service. We initialize the discrete integration process (see Algorithm 2, lines 3 to 10) to obtain the total number of cases per time-step (i.e., day) per population group. Then, 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, considering data from $2020-02-24$to $2020-04-30$. The period object of study is chosen because it has regular social contacts, and the population is not yet vaccinated. Note that we set some activity restrictions in the contact matrices $\mathbb{C}_a$ from mid-march. In particular, the policy vector $\boldsymbol{\Theta}$ is applied with t_{start} = 16th of March following the activation of the restrictions by the Swiss government (Molloy et al., 2021,of Public Health, 2020). The values of $\boldsymbol{\Theta}$ are chosen according to the implemented measures. For example, the complete closure of schools and universities is represented by imposing $\Theta_{education} = 0$. Finally, we plot the cumulative cases over the same period from the Epidemiology table from Google COVID-19 Open Data (*Google Covid data*, n.d.) in Figure 2. The numbers obtained by our epidemiological model are initialized accordingly with the initial values of infected individuals from official public data. By comparing the curves, we can state that the developed model can capture the evolution of the positive cases with reasonable accuracy.

While Table 2 proves the methodology to accurately represent the dependency of the pandemic spread on the age groups and the socio-economic variables, Figure 2 validates the overall SIRD approach by comparing the aggregated number of cases. Nevertheless, the model is not yet proven to capture the correlation between activity-travel behavior and infections. To prove our methodology's ability to capture such correlation, we use the reduction of the reinfection rate as the parameter to compare to other studies (Muller et al., 2020). In their study, the reinfection rate is defined as:

$$R = \frac{\text{Reinfection cases with restriction/Reinfection cases with no restriction}}{\text{Total number of individuals}}. \tag{23}$$

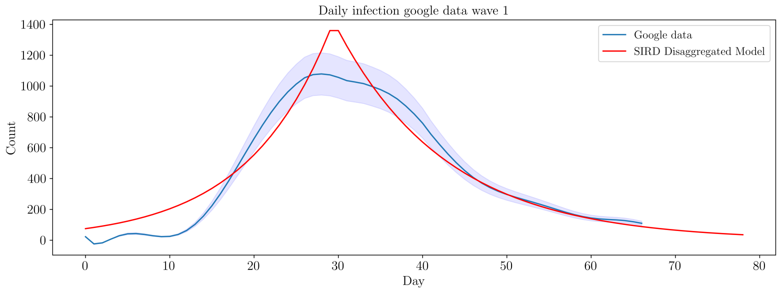

Figure 2: Daily infection from google data for the first wave (blue line) against the aggregated output of our model (red line).

We compute the reinfection rate R indicator by setting to zero the elements of the policy vector $\mathbf{\Theta}$ and running the model. The obtained results are shown in Table 3 and compared to previous authors' results (see Brauner et al., 2020, Haug et al., 2020, and Muller et al., 2020), where it is possible to observe that the results obtained with our disaggregate SIRD Model are in line with the ones proposed by Brauner et al., 2020.

Measure	**Brauner Brauner et al., 2020**	**Haug Haug et al., 2020**	**Disaggregated SIRD Model**
Schools closed	50	16	38
Most businesses suspended	26		27
Work ban	34		36
Gatherings limited to $\leq$ 1000	16		19
Gatherings limited to $\leq$ 100	17		21
Gatherings limited to $\leq$ 10	28		32
Mass gathering cancellation $\geq$ 50		27	31
Small gathering cancellation $\leq$ 50		17	22
Stay-at-home order with exemptions	14		12

Table 3: Percentage point reduction of Muller et al., 2020

Optimal policy control

The optimal policy control problem is computed based on the validated SIRD model. The following assumptions are made on the policy vector $\mathbf{\Theta}$: (i) t_{start} is bounded between the beginning to the end of modeled period, i.e., 80 days (ii) the policy duration is implemented at least for 7 days (iii) the policy vector can be modified with a precision of 0.05, meaning the authorities can modify the contacts between the individual with steps of 5% on the total number of the contact. To compute the total GDP loss and the economic cost of diseased workers and unemployed people due to the lockdown policy we choose the values of the GDP contribution vector $\boldsymbol{c}$ from the literature. The contributions of each activity to the general economy (GDP) c_a are: 1% for home, 1% (*gdp loss work*, 2020) for work, 4.4% (*gdp loss supermarket*, n.d.) for leisure, 1.7% (OECD, n.d.) for services, 1.5% Hanushek and Woessmann, n.d. for education, 5% (*gdp loss supermarket*, n.d.) for supermarkets, and 1.5% for cars. We run the optimization problem for the two objectives, and we show in Figure 3 the health cost H (Equation (15)) on the y-axis and the loss in GDP in percentage (Equation (16)) on the x-axis. Together with the Pareto frontier in yellow, the figure shows in green the set of solutions that have been considered at some point by the algorithm and in blue the ones that have been in the Pareto, but have been removed because dominated by another solution.

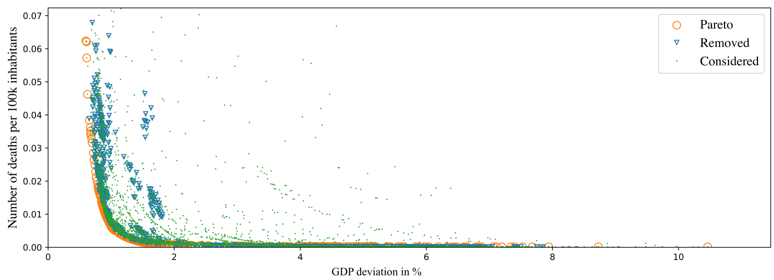

Figure 3: Pareto solution from the VNS algorithm (368 points) .

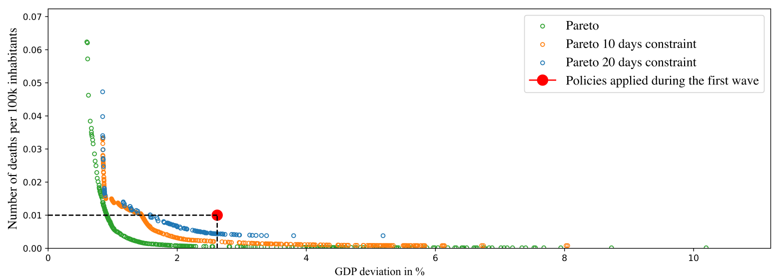

Figure 4: Analysis of different policies simulation outcomes dependant on the t_{start}

In Figure 4, we display different Pareto frontiers obtained from the VNS algorithm and analyze different policies. In particular, the green curve represents the Pareto frontier considering that the policy vector can be applied at any time between the start of the pandemic and the end. Nevertheless, it is possible to prove that most of the points composing this frontier are characterized by an application of the policy vector in the early stages of the pandemic, i.e. in the first week, making this Pareto unrealistic. In fact, it is improbable for a country to apply

NPIs against a pandemic a few days after the appearance of the first positive tests. Therefore, we examine how much the policy's efficiency depends on early implementation. We compute two additional Pareto frontiers using the VNS algorithm by adding a new constraint on $t_{start} \leq 10, 20$ days respectively. As t_{start} increases the Pareto frontier worsens since the optimal points imply higher losses in GDP (E) and the number of deaths (H).

To assess the policy decision-making process, we start by plotting the point that reflects the policies applied during the first wave (red dot). The applied measures consist in the closing of all shops, restaurants, bars and leisure facilities, together with mandatory online schooling and remote working when possible of Public Health, 2020. These activity restrictions were implemented the 16^{th} of March 2020 ($t_{start} = 21$) and modified the 11^{th} of May ($T = 61$). The values for the policy vector can be found in Table 4.

As further validation of the proposed methodology, the GDP loss computed by means of Equation (16) and visible as ΔE in Table 4, is perfectly in line with the real GDP loss registered in the first quarter of 2022, accounting for -2.7 % of the Swiss GDP. We project the policy applied by the Swiss government (red dot) on the Pareto frontiers, to check for policies characterized by the same economical or health cost, obtaining 6 alternative scenarios. In particular, Scenarios 1-3 are the projection of the red dot on the different Pareto frontiers, being equal the economical cost. Scenario 1 shows how an early application of policies would have allowed to reduce the number of causalities, being similar the amount of restriction on the population (i.e. being equal Θ_{mean} and the economical losses E). Acting in later stages, as in Scenario 3, calls for stricter measures. Nonetheless, the latter scenario safeguards the service activity, proposing to keep facilities such as restaurants open. Scenarios 4-6 are the projection of the red dot on the three Pareto frontiers, being equal to the health cost, i.e. the number of deaths. Scenario 4 shows how an early policy only imposing online schooling would have reduced by a factor of three the GDP losses. For all scenarios, we can observe that the later the restriction is implemented, the stricter they have to be to remain on the Pareto frontier.

Policy	Δ**t**	$\mathbf{t_{start}}$	$\mathbf{\Theta}$	$\mathbf{\Theta_{mean}}$	**H**	**E**
Policies applied during the first wave	61	21	[1.00 0.75 0.00 0.00 0.00 1.00 1.00]	0.54	251	2.7
Scenario 1 (green)	64	1	[1.00 0.00 0.85 1.00 0.00 1.00 0.00]	0.55	13	2.55
Scenario 2 (orange)	51	11	[1.00 0.00 0.20 1.00 0.00 1.00 0.00]	0.46	39	2.61
Scenario 3 (blue)	44	21	[1.00 0.00 0.00 1.00 0.00 0.05 0.00]	0.29	100	2.62
Scenario 4 (green)	61	6	[1.00 1.00 1.00 1.00 0.00 1.00 1.00]	0.86	249	0.91
Scenario 5 (orange)	49	14	[1.00 0.00 1.00 1.00 0.00 1.00 1.00]	0.71	251	0.95
Scenario 6 (blue)	34	26	[1.00 0.05 0.00 1.00 0.00 0.00 0.10]	0.31	250	1.09

Table 4: Policy parameters with related economical losses and deaths

4 Conclusions

This paper describes the design and evaluation of a disaggregate SIRD 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, (iii) we calibrate the epidemiological model on real data, (iv) we validate multiple aspects of the disaggregate model by comparison to real Swiss data and existing literature, (v) we propose a method to compute optimal policies to control the spread of the virus. 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. Nevertheless, we propose alternative ways to validate the model against aggregated data and literature. According to the authors, the optimal policy algorithm returns essential information that can help guide the authorities responsible for managing a pandemic crisis or adapting to a post-pandemic situation. Future works might include adding different groups to the model based on other socio-economic variables (subject to data availability). Adding more groups will allow us to explore different policy strategies and their efficiency. Also, the probability of infection is correlated with age, and different age groups have different activity-travel behavior. Endogeneity issues should be analyzed carefully by modeling the probability of infection using a latent model. Finally, to extend the model to the different COVID-19 variants to evaluate its performance and consider the methodology for other non-vector-borne diseases.

Data availability

The data used in this study are available from the Swiss Federal Office of Public Health (FOPH) and the Swiss Federal Institute of Technology in ZÃŒrich (ETHZ). Restrictions apply to the availability of the data. The data were used under license for the current study and are therefore not publicly available. The data are however available from the authors upon reasonable request and with permission of FOPH and ETHZ. The data have undergone an ethical-legal check and have been evaluated for their compliance with the regulations set out by the EPFL Ethics Affairs office. The study was reviewed and approved by the EPFL Research Ethics Compliance officer. The data fall under the Federal Act on the Control of Communicable Diseases in Humans (Epidemics Act) and the Federal Act on Data Protection (Data Protection Act), together with the associated ordinances and provisions. These Federal Acts legislate the timely detection, monitoring, prevention, and control of crises, a model applicable to normal, particular, and extraordinary health situations. The collection and reporting of SARS-CoV-2 data are managed by the FOPH via a notification form for the sake of infectious diseases notification. The data considered in this study are coded, hence preventing the authors from tracing back the identity of human subjects. The authors had access to anonymized demographic and geographic data that does not represent any risk for patients' identification.

References

Albert, R. and Barabasi, A.-L. (2002). Statistical mechanics of complex networks, *Reviews of Modern Physics* **74**(1): 47–97.
URL: *https://link.aps.org/doi/10.1103/RevModPhys.74.47*

Aleta, A., MartÃn-Corral, D., Bakker, M. A., y Piontti, A. P., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N. E., Halloran, M. E., Longini, I. M., Pentland, A., Vespignani, A., Moreno, Y. and Moro, E. (2020). Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas, *preprint*, Epidemiology.
URL: *http://medrxiv.org/lookup/doi/10.1101/2020.12.15.20248273*

Bierlaire, M. (n.d.). A short introduction to PandasBiogeme.

Brauner, J. M., Mindermann, S., Sharma, M., Stephenson, A. B., GavenÄiak, T., Johnston, D., Leech, G., Salvatier, J., Altman, G., Norman, A. J., Monrad, J. T., Besiroglu, T., Ge, H., Mikulik, V., Hartwick, M. A., Teh, Y. W., Chindelevitch, L., Gal, Y. and Kulveit, J. (2020). The effectiveness of eight non-pharmaceutical interventions against COVID-19 in 41 countries, *preprint*, Epidemiology.
URL: *http://medrxiv.org/lookup/doi/10.1101/2020.05.28.20116129*

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D. and Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening, *Nature* **589**(7840): 82–87.
URL: *http://www.nature.com/articles/s41586-020-2923-3*

Colas, C., Hejblum, B., Rouillon, S., Thiebaut, R., Oudeyer, P.-Y., Moulin-Frier, C. and Prague, M. (2021). EpidemiOptim: A Toolbox for the Optimization of Control Policies in Epidemiological Models, *Journal of Artificial Intelligence Research* **71**: 479–519.
URL: *https://jair.org/index.php/jair/article/view/12588*

Commission, E. (2020). The impact of sex and gender in the COVID-19 pandemic.
URL: *https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/impact-sex-and-gender-covid-19-pandemic*$_e$*n*

Davies, N. G., Klepac, P., Liu, Y., Prem, K., Jit, M. and Eggo, R. M. (2020). Age-dependent effects in the transmission and control of COVID-19 epidemics, *Nature Medicine* **26**(8): 1205–1211. Number: 8 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41591-020-0962-9*

Douglas, M., Katikireddi, S. V., Taulbut, M., McKee, M. and McCartney, G. (2020). Mitigating the wider health effects of covid-19 pandemic response, *BMJ* p. m1557.
URL: *https://www.bmj.com/lookup/doi/10.1136/bmj.m1557*

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z. and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks, *Nature* **429**(6988): 180–184. Number: 6988 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nature02541*

gdp loss supermarket (n.d.).
URL: *https://www.seco.admin.ch/seco/en/home/seco/nsb-news.msg-id-79316.html*

gdp loss work (2020).
URL: *https://www.pharmaceutical-technology.com/analysis/remote-working-gdp-covid/*

Goldstein, E., Lipsitch, M. and Cevik, M. (2021). On the Effect of Age on the Transmission of SARS-CoV-2 in Households, Schools, and the Community, *The Journal of Infectious Diseases* **223**(3): 362–369.
URL: *https://academic.oup.com/jid/article/223/3/362/5943164*

Google Covid data (n.d.).
URL: *https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-epidemiology.md*

Hanushek, E. A. and Woessmann, L. (n.d.). gdp loss education.

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

worldwide COVID-19 government interventions, *preprint*, Epidemiology.
URL: *http://medrxiv.org/lookup/doi/10.1101/2020.07.06.20147199*

Heyde, C. C. and Seneta, E. (2001). *Statisticians of the Centuries*, Springer Science & Business Media. Google-Books-ID: uS3dq_grwr0C.

Hindmarsh, A. and Petzold, L. (2005). LSODA, Ordinary Differential Equation Solver for Stiff or Non-Stiff System. Place: Nuclear Energy Agency of the OECD (NEA) INIS Reference Number: 41086668.
URL: *https://inis.iaea.org/search/search.aspx?orig$_q$ = RN* : 41086668

Horl, S. and Balac, M. (2021). Synthetic population and travel demand for Paris and Ãle-de-France based on open and publicly available data, *Transportation Research Part C: Emerging Technologies* **130**: 103291.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0968090X21003016*

Kelman, A. (1985). Compartmental models and their application, *International Journal of Bio-Medical Computing* **16**(3-4): 294–295.
URL: *https://linkinghub.elsevier.com/retrieve/pii/0020710185900637*

Kermack, W. O., McKendrick, A. G. and Walker, G. T. (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **115**(772): 700–721. Publisher: Royal Society.
URL: *https://royalsocietypublishing.org/doi/10.1098/rspa.1927.0118*

Lee, M. and You, M. (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): E2977.

Mancastroppa, M., Burioni, R., Colizza, V. and Vezzani, A. (2020). Active and inactive quarantine in epidemic spreading on adaptive activity-driven networks, *Physical Review E* **102**(2): 020301.
URL: *https://link.aps.org/doi/10.1103/PhysRevE.102.020301*

Molloy, J., Schatzmann, T., Schoeman, B., Tchervenkov, C., Hintermann, B. and Axhausen, K. W. (2021). Observed impacts of the Covid-19 first wave on travel behaviour in Switzerland based on a large GPS panel, *Transport Policy* **104**: 43–51.
URL: *https://www.sciencedirect.com/science/article/pii/S0967070X21000159*

Muller, S. A., Balmer, M., Charlton, B., Ewert, R., Neumann, A., Rakow, C., Schlenther, T. and Nagel, K. (2020). Using mobile phone data for epidemiological simulations of lockdowns: government interventions, behavioral changes, and resulting changes of reinfections, *preprint*, Epidemiology.
URL: *http://medrxiv.org/lookup/doi/10.1101/2020.07.22.20160093*

OECD (n.d.). gdp loss services.
URL: *https://www.oecd.org/coronavirus/policy-responses/covid-19-and-the-retail-sector-impact-and-policy-responses-371d7599/*

of Public Health, F. O. (2020). Coronavirus: Federal Council declares âextraordinary situationâ and introduces more stringent measures.
URL: *https://www.bag.admin.ch/bag/en/home/das-bag/aktuell/medienmitteilungen.msg-id-78454.html*

Office, F. S. (n.d.). Portraits of the communes.
URL: *https://www.bfs.admin.ch/bfs/en/home/statistiken/regionalstatistik/regionale-portraets-kennzahlen/gemeinden.html*

Ortelli, N., Hillel, T., Pereira, F. C., de Lapparent, M. and Bierlaire, M. (2021). Assisted specification of discrete choice models, *Journal of Choice Modelling* **39**: 100285.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S175553452100018X*

Qian, X. and Ukkusuri, S. V. (2021). Connecting urban transportation systems with the spread of infectious diseases: A Trans-SEIR modeling approach, *Transportation Research Part B: Methodological* **145**: 185–211.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0191261521000175*

Riou, J., Panczak, R., Althaus, C. L., Junker, C., Perisa, D., Schneider, K., Criscuolo, N. G., Low, N. and Egger, M. (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* .
URL: *https://www.ssrn.com/abstract=3845990*

Riou, J., Panczak, R., Althaus, C. L., Junker, C., Perisa, D., Schneider, K., Criscuolo, N. G., Low, N. and Egger, M. (2021b). 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. Publisher: Elsevier.
URL: *https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(21)00160-2/fulltext*

Singu, S., Acharya, A., Challagundla, K. and Byrareddy, S. N. (2020). Impact of Social Determinants of Health on the Emerging COVID-19 Pandemic in the United States, *Frontiers in Public Health* **8**: 406.
URL: *https://www.frontiersin.org/article/10.3389/fpubh.2020.00406/full*

Tirachini, A. and Cats, O. (2020). COVID-19 and Public Transportation: Current Assessment, Prospects, and Research Needs, *Journal of Public Transportation* **22**(1).
URL: *https://scholarcommons.usf.edu/jpt/vol22/iss1/1*

Tuomisto, J. T., Yrjola, J., Kolehmainen, M., Bonsdorff, J., Pekkanen, J. and Tikkanen, T. (2020). An agent-based epidemic model REINA for COVID-19 to identify destructive policies, *preprint*, Infectious Diseases (except HIV/AIDS).
URL: *http://medrxiv.org/lookup/doi/10.1101/2020.04.09.20047498*

Zheng, R., Xu, Y., Wang, W., Ning, G. and Bi, Y. (2020). Spatial transmission of COVID-19 via public and private transportation in China, *Travel Medicine and Infectious Disease* **34**: 101626.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S1477893920300946*