



HAL
open science

Probabilistic and mean-field model of COVID-19 epidemics with user mobility and contact tracing

Marianne Akian, Luca Ganassali, Stéphane Gaubert, Laurent Massoulié

► **To cite this version:**

Marianne Akian, Luca Ganassali, Stéphane Gaubert, Laurent Massoulié. Probabilistic and mean-field model of COVID-19 epidemics with user mobility and contact tracing. 2020. hal-02941123

HAL Id: hal-02941123

<https://hal.science/hal-02941123>

Preprint submitted on 16 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Probabilistic and mean-field model of COVID-19 epidemics with user mobility and contact tracing

M. Akian^{*}, L. Ganassali[†], S. Gaubert[‡], L. Massoulié[§]

September 16, 2020

Abstract

We propose a detailed discrete-time model of COVID-19 epidemics coming in two flavours, mean-field and probabilistic. The main contribution lies in several extensions of the basic model that capture i) user mobility – distinguishing *routing*, i.e. change of residence, from *commuting*, i.e. daily mobility – and ii) contact tracing procedures. We confront this model to public data on daily hospitalizations, and discuss its application as well as underlying estimation procedures.

Contents

1	Single Population Model	3
1.1	Deterministic version: mean-field model	5
1.2	Probabilistic version: multi-type branching process	6
1.3	Extension with contact tracing	7
2	Inference methods	9
2.1	Perron’s theory for fitting and inference in the mean-field setting	9
2.2	Kalman filtering for the branching process	10
3	Numerical illustration on Paris hospitalization data	12
3.1	Simulation settings	12
3.2	Results	13
4	Variable contact rates	14
5	Model with Routing Mobility	18
5.1	Model parameters for subpopulations	18
5.2	Routing step in the mean field version	19
5.3	Estimation of routing fractions and numbers of contacts	20

Introduction

A profusion of mathematical models of the COVID-19 epidemic propagation has been developed in a very short time lapse. Specific models were designed to address particular objectives, such as

^{*}INRIA and CMAP, École polytechnique, IP Paris, CNRS. Email: marianne.akian@inria.fr.

[†]INRIA, DI/ENS, PSL Research University, Paris, France. Email: luca.ganassali@inria.fr.

[‡]INRIA and CMAP, École polytechnique, IP Paris, CNRS. Email: stephane.gaubert@inria.fr.

[§]MSR-Inria Joint Centre, INRIA, DI/ENS, PSL Research University, Paris, France. Email: laurent.massoulie@inria.fr.

forecasting the epidemics, understanding the impact of various interventions such as restraining mobility and contacts between users, or deploying contact tracing and case isolation.

The goal of this study is to propose a single model of COVID-19 epidemics based on a well understood family of probabilistic models – that of multi-type branching processes – and to detail simple extensions of this model that capture the impact of contact tracing procedures as well as user mobility.

Our model can serve several purposes: It can be used to assess from available time series so-called contact rates, and in turn the impact of user mobility on epidemic propagation. It can also be used in surveillance platforms to infer the status of specific sub-populations (in particular fraction of infectious or immunized persons), a prerequisite for the implementation of targeted testing campaigns.

Paper organization We introduce in Section 1 a discrete time epidemiological model of COVID-19 for a homogeneous population. We detail a probabilistic version of the model, that is a multi-type branching process. It represents in a precise manner the durations of the principal phases of a patient’s condition and the transition rates between such phases. A deterministic version of this model is then described. We then introduce a simple extension of the model to represent the impact of contact tracing and case isolation.

In Section 2, we describe two methods that both use model structure for inference of parameters and prediction. For the deterministic, or mean-field version of the model, we propose an approach to perform inference that relies on Perron’s theory. For the probabilistic version of the model, by exploiting the fact that we are dealing with a hidden Markov model, we propose a Kalman filtering approach to perform inference.

Next, we present a numerical illustration in Section 3, where the model is fitted to daily counts of hospitalizations in the Paris area. There we estimate piecewise constant contact rates corresponding to distinct time periods, namely pre-confinement, intermediate, and confinement periods. We discuss the application of this fitting procedure to forecasting the future number of hospitalizations.

We then describe in Section 4 an extension for the contact rates to take daily mobility – or other covariates – into account, making these parameters depend on time.

We finally discuss in Section 5 the extension of the model to multiple sub-populations, corresponding e.g. to geographical regions and age ranges. Key parameters in this extension are contact matrices, describing the frequency of contacts between sub-populations. The routing dimension of this extension is detailed. We revisit the question of estimating model parameters, specifically contact matrices in this multi sub-populations model. We assume potentially time-varying contact matrices and consider estimation of such contact matrices based on user mobility traces.

Related work SIR or SEIR-type models of COVID-19 featuring asymptomatic or undocumented infectious individuals have been considered in [11], [14] and [2]. Studies [14, 2] consider same contact rates for symptomatic and asymptomatic cases. Some works [13] also consider a split of the infectious phase to explicit the occurrence of aggravated symptoms.

General distributions of times spent in a given state have been studied. Prodromic phase, during which individuals are infectious, yet have not displayed symptoms, is considered in [3]. The model of [3] comprises a detailed breakdown in various phases, yet relies on geometric/exponential durations for time spent in each phase, while the model of [15] describes general discrete time distributions for the time spent in states.

A broad range of studies have addressed the issue of estimating the time spent in a given state. In [12], the incubation period is modeled with a log-normal distribution with mean 5.8 days. In [9], the median incubation period was estimated to be 5.1 days, and 97.5% of those who develop symptoms will do so within 11.5 days of infection.

[1] reviews studies of the infectious period's duration, reporting several estimates with substantial variations, giving for asymptomatic cases an estimate of 6.5-9.5 days. Pre-symptomatic infectious period is estimated to 0-4 days across studies. [16] also investigates the time between infection and remission/death.

The model of Figure 1 is a simplified version of the model considered in [3], which features additional states, including a breakdown between hospitalized cases entering Intensive Care Units or not, and a distinction between several severity levels of the infection. It is similar to the model considered in [13], with a key difference in our model that is that we take the discrete time t into account.

1 Single Population Model

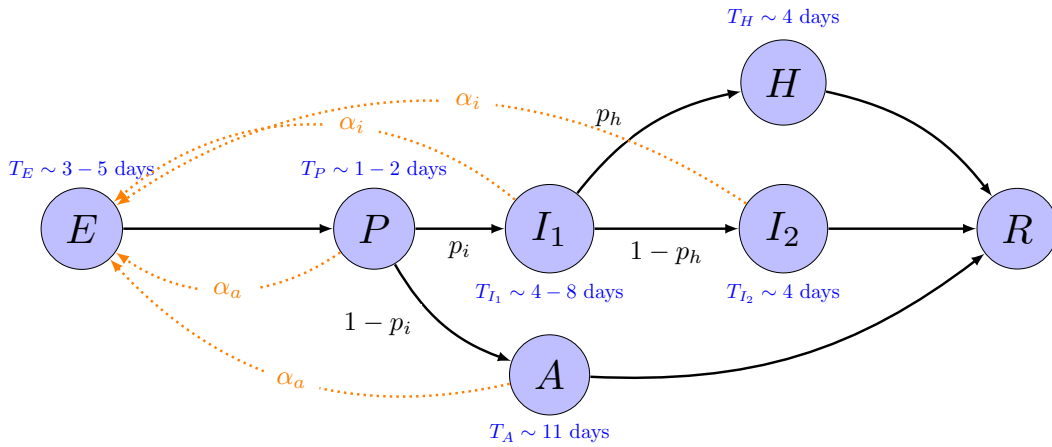


Figure 1: State diagram for individual's condition in model 1. Dotted arrows show the contamination of new cases: the mean number of new E cases per day per A/P (resp. per I_1/I_2) individual is α_a (resp. α_i).

We consider the discrete time SEIR-type model depicted in Figure 1. We classify the states of an individual as E for *exposed*; P for *prodromic*; A for *infectious asymptomatic*; I_1 for *phase 1 infectious symptomatic*; H for *hospitalized*; I_2 for (non-hospitalized) *phase 2 infectious symptomatic*; and R for *removed* from the contamination chain. After having moved to prodromic phase P , an individual becomes either asymptomatic (state A) or symptomatic, entering state I_1 . After some time in the I_1 state, two situations arise: either the individual is hospitalized, entering state H , either there is no aggravation, leading to state I_2 . Then, the last state is always R , where the individual is either dead or recovered.

We shall in fact adopt a more detailed representation for the state of an individual: its detailed state will be represented by a couple (a, t) where $a \in \{E, A, I_1, H, I_2, R\}$ is the stage of the disease, and t the number of days spent in this stage. For instance, an individual in state $(E, 3)$ just completes their third day in the exposed phase.

Remark 1.1. (i) In epidemiological models one often tracks individuals in so-called *susceptible* state S : these are those individuals who may get exposed to infection. In the case of a negligible fraction of infected individuals, in particular in the early stages of the epidemics, on which we focus, it is not necessary to track them, hence our choice not to include state S in the diagram of Figure 1.

- (ii) We distinguish between asymptomatic and symptomatic individuals, as new symptomatic cases can be observed when the corresponding individuals consult their physician or enter hospital after experiencing serious symptoms.
- (iii) We distinguish between hospitalized and non-hospitalized cases. The number of new hospital admissions can be measured, for instance, from medical emergency services records.
- (iv) We distinguish according to the number of days an individual has spent in its current state for the following reason. Traditional SEIR-type models implicitly assume that the time that an individual spends in a particular state admits an exponential distribution (for continuous time ODE-based models) or a geometric distribution (for discrete time models). However the durations spent in the states of interest for COVID-19 appear to be very far from exponentially distributed. Thus, for an accurate short-term prediction of the epidemics evolution, one needs to take this number of days into account.

State variables We represent the dynamics by means of the following state variables. On each day t , we represent the population by a vector $x(t)$ with entries:

- $x_{E,d}(t)$: number of individuals in phase E , having spent d days in this phase;
- $x_{P,d}(t)$: number of individuals in phase P , having spent d days in this phase;
- $x_{I_1,d}(t)$: phase 1 infectious individuals, having spent d days in this phase;
- $x_{I_2,d}(t)$: phase 2 infectious individuals having spent d days in this phase;
- $x_{A,d}(t)$: asymptomatic individuals having spent d days in this phase;
- $x_H(t)$: number of individuals newly hospitalized on day t .

By considering the state variable $x_H(t)$, instead of $x_{H,d}(t)$, we ignore the number of days an individual spends at hospital. This stems from the assumption that such cases are efficiently isolated and do not contribute to propagating the epidemics, an assumption that could be revised.

Distribution of phase durations For each $\tau \in \{E, P, I_1, I_2, A\}$, define $r_\tau(d)$ the probability that corresponding phase will end on the following day, given that it has lasted d days. Denoting by $p_\tau(d)$ the probability that phase τ lasts d days, $r_\tau(d)$ is the associated failure rate, defined by

$$r_\tau(d) = \begin{cases} \frac{p_\tau(d)}{\sum_{\delta \geq d} p_\tau(\delta)} & \text{if } \sum_{\delta=1}^d p_\tau(\delta) > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Available statistics suggest that typical durations lie in the ranges given on Figure 1. A reasonable baseline assumption could consist in taking

- $p_E(d) = 1/3$ for $d \in \{3, 4, 5\}$ and $p_E(d) = 0$ otherwise,
- $p_P(1) = p_P(2) = 1/2$ and $p_P(d) = 0$ otherwise,
- $p_{I_1}(d) = 1/3$ for $d \in \{5, 6, 7\}$ and $p_{I_1}(d) = 0$ otherwise,
- $p_{I_2}(d) = 1$ for $d = 4$ and $p_{I_2}(d) = 0$ otherwise,
- $p_A(d) = 1$ for $d = 11$ and $p_A(d) = 0$ otherwise.

Note however that far longer durations have been observed, e.g. incubation lasting 10 days or more; I_1 phase between first symptoms and aggravation lasting as long as 20 days. This baseline is therefore by no means an accepted characterization of the dynamics of an individual's condition.

Transition probabilities

- p_i : probability that previously exposed individual becomes symptomatic at end of incubation. A baseline assumption consists in taking $p_i = 0.7$. Again, no consensus has emerged yet on the fraction of asymptomatic infected users and competing assumptions have been proposed (for baselines, data from APHP <https://www.aphp.fr> can be used, as done e.g. in [3]).
- p_h : probability that previously phase 1 infectious individual develops aggravated form at end of phase 1. A baseline assumption consists in taking $p_h = 0.05$, again with the same cautionary note.

Contact rates

- α_i : average number of new exposed individuals generated by single infectious individual (I_1, I_2) in a single day. Here we do not distinguish between phases 1 and 2, an assumption which could be revisited.
- α_a : average number of new exposed individuals generated by single asymptomatic or prodromic individual in a single day.

In the sequel we assume that $\alpha_a = \alpha_i$. This assumption is often made, for instance in [14, 2]. In contrast, [11] makes an explicit distinction in contact rates α for documented cases and $\mu\alpha$ for undocumented ones. Especially, an estimate for μ in China between January 10th and January 23th is given: $\mu \sim 0.55$ (95% confidence interval: (0.46, 0.62)).

Remark 1.2. This model can be complemented by equations giving the number of deaths and immunizations per day. This, together with a tracking of the volume of compartment S , could be useful if one wishes to estimate the proportion of immunized individuals in the population. Denoting by p_d the probability of death conditional on having been hospitalized, and assuming that only hospitalized individuals die, this leads to two other state variables:

$$\begin{aligned} x_{\text{death}}(t+1) &= p_d x_H(t), \\ x_{\text{immun}}(t+1) &= (1 - p_d) x_H(t) + \sum_{\delta > 0} x_{A,\delta}(t) r_A(\delta) + x_{I_2,\delta}(t) r_{I_2}(\delta). \end{aligned}$$

The number of deaths as written neglects time between hospitalization and death.

Observables We assume that coordinates $x_H(t)$ are observed by the end of day t . This would correspond to the number of new hospitalized patients with COVID-19 pathology. In contrast, the states variables $x_E(t), x_A(t), x_P(t)$ are never observed. They will later be referred to as *hidden trajectories*. It would be straightforward to adjust the model to incorporate other observables, such as the number of patients consulting their physician on each day.

1.1 Deterministic version: mean-field model

This is a deterministic, linear model, based on the assumption that the fraction of susceptible individuals in the total population is close to 1. In other words it is meant to represent stages of the epidemics where collective immunization is yet negligible. Simple adjustments can be made to transform this model to account for a non-negligible, time-evolving, fraction of immunized population.

Assume there is a maximal number of days, h , that each of the phases E, P, I_1, I_2, A can last. Let $x_E = (x_{E,1}, \dots, x_{E,h})^\top$, $x_P = (x_{P,1}, \dots, x_{P,h})^\top$, and similarly for h -dimensional vectors x_A, x_{I_1}, x_{I_2} . The state space on day t , $x(t)$, is given by

$$x(t) := \begin{pmatrix} x_E(t) \\ x_P(t) \\ x_{I_1}(t) \\ x_A(t) \\ x_{I_2}(t) \\ x_H(t) \end{pmatrix} \quad (2)$$

The dynamics are then given by

$$x(t+1) = Mx(t), \quad (3)$$

where matrix M is specified by the detailed equations, for $d \in [h-1]$:

$$\begin{aligned} \forall \tau \in \{E, P, I_1, I_2, A\}, \quad x_{\tau, d+1}(t+1) &= x_{\tau, d}(t)(1 - r_\tau(d)), \\ x_{E,1}(t+1) &= \sum_{\delta > 0} [\alpha_i(x_{I_1, \delta} + x_{I_2, \delta})(t) + \alpha_a(x_{A, \delta} + x_{P, \delta})(t)], \\ x_{P,1}(t+1) &= \sum_{\delta > 0} x_{E, \delta}(t)r_E(\delta), \\ x_{I_1,1}(t+1) &= p_i \sum_{\delta > 0} x_{P, \delta}(t)r_P(\delta), \\ x_{A,1}(t+1) &= (1 - p_i) \sum_{\delta > 0} x_{P, \delta}(t)r_P(\delta), \\ x_{I_2,1}(t+1) &= (1 - p_h) \sum_{\delta > 0} x_{I_1, \delta}(t)r_{I_1}(\delta), \\ x_H(t+1) &= p_h \sum_{\delta > 0} x_{I_1, \delta}(t)r_{I_1}(\delta). \end{aligned}$$

1.2 Probabilistic version: multi-type branching process

We now transform the previous mean field model into a probabilistic counterpart: a multi-type (or multi-dimensional) branching process. Consideration of probabilistic behaviour is particularly relevant when we deal with not-so-large infected populations, or in sub-critical cases where the largest eigenvalue $\lambda_1(M)$ of matrix M defined in (3) verifies $\lambda_1(M) \leq 1$. Indeed, when $\lambda_1(M) > 1$, by results of Kesten and Stigum [7], there is a random variable Z such that for large t , the state $X(t)$ renormalized by $\lambda_1(M)^{-t}$ converges to Zu , where u is the eigenvector of M associated with eigenvalue $\lambda_1(M)$. This provides a theoretical justification of the mean field model in the case of large infected populations. Figure 2 gives a numerical illustration of this result.

The number of newly exposed individuals now follows a Poisson distribution with mean as specified in the mean field model. Determination of durations in each state, and choice of next state at branches in the state diagram of Figure 1, are made with independent random coins with the prescribed probabilities.

For clarity, to distinguish between the two versions, random variables will be denoted in capital letters through the rest of this article. Binomial (resp. Poisson) distributions are denoted by \mathcal{B} (resp. \mathcal{P}).

Thus, conditioning to state vector $X(t)$ at time t , the state $X(t+1)$ at time $t+1$ verifies:

$$\forall \tau \in \{E, P, I_1, I_2, A\}, X_{\tau, d+1}(t+1) \sim \mathcal{B}(X_{\tau, d}(t), 1 - r_{\tau}(d)),$$

$$X_{E,1}(t+1) \sim \mathcal{P}\left(\sum_{\delta>0} [\alpha_i(X_{I_1, \delta} + X_{I_2, \delta})(t) + \alpha_a(X_{A, \delta} + X_{P, \delta})(t)]\right),$$

$$X_{P,1}(t+1) = \sum_{\delta>0} [X_{E, \delta}(t) - X_{E, \delta+1}(t+1)],$$

$$X_{I_1,1}(t+1) \sim \mathcal{B}\left(\sum_{\delta>0} [X_{P, \delta}(t) - X_{P, \delta+1}(t+1)], p_i\right),$$

$$X_{A,1}(t+1) = \sum_{\delta>0} [X_{P, \delta}(t) - X_{P, \delta+1}(t+1)] - X_{I_1,1}(t+1),$$

$$X_{I_2,1}(t+1) \sim \mathcal{B}\left(\sum_{\delta>0} [X_{I_1, \delta}(t) - X_{I_1, \delta+1}(t+1)], 1 - p_h\right),$$

$$X_H(t+1) = \sum_{\delta>0} [X_{I_1, \delta}(t) - X_{I_1, \delta+1}(t+1)] - X_{I_2,1}(t+1).$$

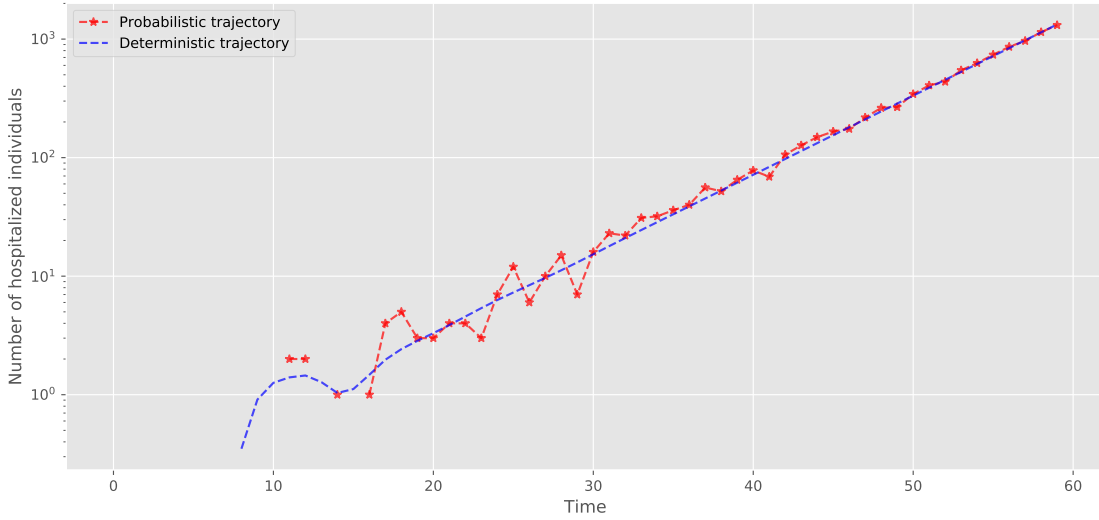


Figure 2: Trajectories of newly hospitalized individuals of length 60 days from the probabilistic and deterministic versions, with $p_a = 0.3$, $p_h = 0.05$, $\alpha_i = 0.4$, $\alpha_a = 0.3$, initialized with 200 individuals entering in E .

1.3 Extension with contact tracing

One motivation of the present study is the control of the epidemics via case isolation, that is quarantining of identified cases so that they do not contribute to further virus propagation. The main approach to deploy case isolation relies on identification of infectious cases via tests. Thus, when testing capacity is limited, efficient case isolation crucially depends on targeted test campaigns. Contact tracing is of paramount importance for targeting tests: it allows to determine who has been in contact with infectious cases, and hence focus tests on such contacts.

We now describe an extension of our previously introduced model to include contact tracing and case isolation. In a nutshell, the model assumes that those individuals subsequently infected by a specific person (its children in the branching tree associated to the branching process)

get tested with some probability once that specific person is positively tested (and subsequently isolated). This formalization of contact tracing has already been considered in [8] for a simple branching process, analyzing the efficiency of contact tracing plus case isolation to control epidemics. We give here a more general version, providing a generic method to expand the state space of a multi-type branching process in order to capture the introduction of contact tracing plus case isolation.

Consider then a multi-type branching process where user types are indexed by $j \in [J]$. For the sake of generality, we shall assume that types j encode the succession of future phases ϕ that an individual will visit over the coming days. Thus an individual with type $j = \{E, E, P, I_1, H\}$ is in state E for the present day and the next, and will then spend a single day in the states P , then I_1 , and then H subsequently. Also, an individual with type $j = \{\phi_1, \dots, \phi_\ell\}$ on a given day will become an individual with type $j' = \{\phi_2, \dots, \phi_\ell\}$ on the next day.

Note that the previous branching process gives rise to an equivalent branching process with this set of types, as can be seen by sampling, upon birth of an individual, the sequence of states ϕ it will visit after its birth.

We now assume that there is a set Φ_0 of distinguished states $\phi \in \Phi_0$ such that an individual entering a state ϕ in Φ_0 automatically incurs a *positive test*. Concretely, the state $\phi = H$, namely hospitalization of the individual, would automatically trigger a positive test, and we may think of Φ_0 as the singleton $\{H\}$.

We then assume that, for some fixed tracing probability $p_t > 0$, the contact between an infectious individual and another individual that became infected is being recorded. The original branching process dynamics are then modified as follows: once an individual is tested positive, it is isolated (removed) and its traced contacts are subsequently tested positive on the following day, and subsequently isolated, and so on.

The key observation is that this modification preserves the branching structure, provided we perform the following extension of the original type space $[J]$. The new type space becomes

$$\mathcal{J} = \{(j, d) : j \in [J], d \in [D_{\max}] \cup \{+\infty\}\},$$

where the second coordinate d represents the number of days until the corresponding individual will be positively tested.

Upon birth of an individual whose parent has type (j, d) , the newly born individual will have type (j', d') . We now specify the distribution of the pair (j', d') . First sample a sequence $j'' = \{\phi''_1, \dots, \phi''_{d'}\} \in [J]$ conditionally on j with the same conditional distribution as in the original branching process. Let $d_1 = \inf\{k \geq 1 : \phi''_k \in \Phi_0\}$ be the number of days before the newly born individual enters some state in Φ_0 .

We then let

$$d' = \begin{cases} d_1 & \text{with probability } 1 - p_t, \\ \min(d_1, d + 1) & \text{with probability } p_t. \end{cases} \quad (4)$$

this reflects the fact that if traced, the new individual will become tested either when it enters some state in Φ_0 , or on the day after its parent gets tested, whichever happens first. If untraced, it only becomes tested on the day it enters some state in Φ_0 . Finally, we let $j' = \{\phi''_1, \dots, \phi''_{d'}\}$, i.e. we truncate the sequence j'' to length d' , thereby encoding the impact of case isolation, by which a positively tested individual can no longer infect anyone else.

Note that further extensions can be considered that preserve the branching process structure. Here is an example of such an extension. Assume that on each day, each infected individual may be submitted to a randomly administered test, and trigger a positive test reply, this happening independently with probability ε on any given day. To reflect such random tests, the above-described dynamics would then need to be amended in the following manner. For an individual newly born from a (j, d) -parent, sample a random variable X distributed as

$$X = G \cdot \mathbf{1}_{G \leq D_{\max}} + \infty \cdot \mathbf{1}_{G > D_{\max}}, \quad (5)$$

where G follows a geometric distribution of parameter ε .

With j'' and d_1 as above, let then

$$d' = \begin{cases} \min(d_1, X) & \text{with probability } 1 - p_t, \\ \min(d_1, X, d + 1) & \text{with probability } p_t. \end{cases} \quad (6)$$

The construction then goes through unchanged, letting $j' = \{\phi'_1, \dots, \phi''_{d'}\}$ and the newly born individual's type being (j', d') .

The above extensions to the original branching process can be used to assess how the tracing probability p_t , or the fraction ε of random tests, affects the potency of the epidemics. In particular, one may consider the mean progeny matrix M , indexed by extended types $j \in [J']$ and associated with the branching process that captures contact tracing.

A crude criterion for success of contact tracing consists in requiring that the spectral radius $\rho(M)$ be strictly less than 1. Under such conditions, given a vector $x(0)$ capturing the initial number of individuals of each type, then the expectation of X_{infected} , the total number individuals ultimately infected, reads

$$\mathbb{E}(X_{\text{infected}}) = \sum_{t \geq 0} \langle x(0), M^t e \rangle = \langle x(0), (I - M)^{-1} e \rangle. \quad (7)$$

In the above, e denotes the all-ones vector and the second equality is valid whenever $\rho(M) < 1$.

This expression for $\mathbb{E}(X_{\text{infected}})$ can be further processed to obtain simple upper bounds. Particularly simple upper bounds have been obtained in [4] in the case where M is symmetric. However symmetry does not hold for the matrices M considered here, so that we cannot rely on the bounds of [4].

2 Inference methods

In this section, we propose inference methods for our model. We consider a general deterministic (respectively, probabilistic) setting where the state vector at time t is denoted by $x(t)$ (resp. $X(t)$).

Individual types are now denoted by $j \in [J]$, and $M_{i,j}$ is the mean number of type i -children from a type j -parent. M^\top is known as the *mean-progeny matrix* in branching process theory. It depends on a parameter of interest $\theta \in \Theta$, that encapsulates for instance $p_i, p_h, \alpha_a, \alpha_i$ and time distributions in Section 1, as well as p_t and c_{\max} if we consider contact tracing as in Section 1.3.

We assume that the observed trajectories are the first h coordinates of $x(t)$ (resp. $X(t)$), denoted by $x_H(t)$ (resp. $X_H(t)$), all other coordinates being hidden trajectories.

2.1 Perron's theory for fitting and inference in the mean-field setting

For simplicity, assume that the remaining unknown parameters of the model are the contact rates α_a and α_i , all other parameters having correctly been estimated, e.g. from patient statistics compiled at hospitals.

Constant contact rates In the case where contact rates are constant at $\theta_0 = (\alpha_{i,0}, \alpha_{a,0})$, the state at time t is given by $x(t) = M(\theta_0)^t x(0)$. Its behavior is determined at first order by the largest eigenvalue of matrix $M = M(\theta_0)$, $\lambda_1(M)$, $x(t)$ being close to $c \lambda_1(M)^t u$, where u is the eigenvector of M associated with eigenvalue $\lambda_1(M)$ and c some constant. The exponential rate of growth $\lambda_1(M)$ of observable $x_H(t)$ can be estimated, e.g. by fitting a regression line to the semilog plot up to time t of the corresponding time series, thus producing the estimate $\hat{\lambda}(t)$.

For a single population, in the mean field model, estimation of the exponent $\lambda_1(M)$ gives one relation constraining the unknown parameters, namely:

$$\hat{\lambda}(t) = \lambda_1(M(\theta_0)) =: F(\theta_0). \quad (8)$$

Remark 2.1. Let p be a polynomial such that $p(M) = 0$ (p could be the characteristic polynomial of M for instance). Writing $p(z) = z^d - \sum_{i=1}^d a_{d-i} z^i$, from the expression $x_H(t) = P_H M^t x(0)$ valid in the mean-field model, it readily follows that

$$x_H(t) = \sum_{i=1}^d a_i x_H(t-i). \quad (9)$$

In other words, the observables $x_H(t)$ follow an auto-regressive dynamics. Instead of estimating unknown coefficients of matrix M , one could therefore estimate directly the auto-regressive coefficients a_i . The latter approach is potentially simpler, and suffices if one is only interested in forecasting $x_H(t)$. The coefficients a_i however do not have a direct physical interpretation, in contrast to the coefficients of M . Estimation of M is therefore better suited to assess the impact of specific measures.

Reaction to shocks If on given day t_0 (start of confinement say), the parameters α_i and α_a abruptly change, the dynamics from t_0 onwards would again be linear, with a modified matrix $M(\theta_1)$ with $\theta_1 = (\alpha_{i,1}, \alpha_{a,1})$ corresponding to the new contact rates. This suggests the following approach for inferring these parameters.

Estimate the state vector $x(t_0)$ as Cu , for some constant $C = C(\theta_0)$ such that $Cu_H = x_H(t_0)$, where u_H is the projection of $u = u(\theta_0)$ on the first h vectors of the canonical basis.

Leverage then the transient behaviour of the post-shock dynamics as follows. Making the dependency of the post-shock matrix $M(\theta_1)$ on the unknown θ_1 explicit, estimate the unknown (θ_0, θ_1) as the minimizers of the optimization problem:

$$\min_{\substack{\theta_0, \theta_1 \\ F(\theta_0) = \hat{\lambda}(t_0)}} \sum_{t > T_0} \left([M(\theta_1)^{t-t_0} C(\theta_0) u(\theta_0)]_H - x_H(t) \right)^2. \quad (10)$$

Remark 2.2. The above optimization criterion is essentially motivated by the assumption of constant contact parameters before and after the shock. This is a strong assumption, verified at best in an approximate sense. See [10] which considers multiple phases in an SIR dynamics, to be fitted to observations. In Section 3 we will consider two shocks. Time-varying contact rates will be discussed below in Section 4.

We just saw how to leverage Perron's theory to perform inference in the mean-field version of the model. Let us now describe an inference method in the probabilistic model which – as already explained – is more relevant when dealing with small populations, e.g. when studying sub-critical evolutions in a decay phase, or spread relaunch phases.

2.2 Kalman filtering for the branching process

In the following paragraph we describe an inference method in the probabilistic setting that enables to estimate the state $X(t)$ (including hidden trajectories) by updating the estimation at every new time step. The vector of average numbers per type at generation t , conditionally on the initial state to be $X(0) \in \mathbb{R}^J$, is $M^t X(0)$.

Measurement errors First, we describe a model for measures, that is the observed data X_H^* , as follows:

$$X_H^*(t) = P_H X(t) + v(t), \quad (11)$$

where P_H is the $h \times J$ matrix of the projection on the h last components of the canonical basis, and $v(t)$ is a centered noise, with covariance matrix $R(t)$, representing the confidence in measurements.

Model errors Now, we introduce the filtration $\mathcal{F}_t := \sigma(X_H^*(0), \dots, X_H^*(t))$, that is the σ -algebra generated by all measurements up to time t . We provide estimates of states $X(t)$, as well as the estimated errors. We will use the following notations:

- $\hat{X}(t)$: an estimate of state $X(t)$ given \mathcal{F}_t , and $P(t)$ its covariance matrix,
- $\hat{X}(t|t-1)$: an estimate of $X(t)$, given \mathcal{F}_{t-1} , and $P(t|t-1)$ its covariance matrix.

At time t , we write

$$X(t) = MX(t-1) + u(t),$$

with $u(t) = X(t) - MX(t-1)$ a centered vector. For $j \in [J]$, let S_j denote the covariance matrix of vector of children from type j -parent. The covariance matrix $Q(t)$ of $u(t)$ can be computed as follows:

$$\begin{aligned} Q(t) &:= \text{Var}(u(t)) \\ &= \mathbb{E}[\text{Var}(X(t) - MX(t-1)|X(t-1))] + \text{Var}(\mathbb{E}[X(t) - MX(t-1)|X(t-1)]) \\ &= \mathbb{E}\left[\sum_{j \in [J]} X(t-1)_j S_j\right] + 0 \\ &= \sum_{j \in [J]} (M^{t-1}X(0))_j S_j. \end{aligned}$$

The fact that closed-form expressions for second moments are available is a consequence of the Markovian structure of our model. The same Markovian structure also implies that $u(t)$ is decorrelated from any \mathcal{F}_{t-1} -measurable variable.

Kalman filtering The strategy is then to compute $\hat{X}(t|t-1)$ at each step, and to adjust the prediction with the measurement $X_H^*(t)$ to build $\hat{X}(t)$, assuming that $\hat{X}(t)$ is of the form

$$\hat{X}(t) := \hat{X}(t|t-1) + K(t) \left(X_H^*(t) - P_H \hat{X}(t|t-1) \right).$$

Optimizing $K(t)$ in the quadratic error sense leads to *Kalman filtering* (see e.g. [6]), which is the best linear recursive estimator in our setting, namely a hidden Markov model. With this method, at each time t , we perform the following predictions:

$$\begin{aligned} \hat{X}(t|t-1) &= M\hat{X}(t-1), \\ P(t|t-1) &= M \cdot P(t-1) \cdot M^\top + Q(t), \end{aligned} \tag{12}$$

then we adjust our estimates, updating with the new measure $X_H^*(t)$:

$$\begin{aligned} K(t) &:= P(t|t-1) \cdot P_H^\top \cdot (P_H \cdot P(t|t-1) \cdot P_H^\top + R(t))^{-1}, \\ \hat{X}(t) &= \hat{X}(t|t-1) + K(t) \cdot (X_H^*(t) - P_H \hat{X}(t|t-1)), \\ P(t) &= (\text{Id}_J - K(t) \cdot P_H) P(t|t-1). \end{aligned} \tag{13}$$

The intermediate matrix $K(t)$ of size $J \times h$ is usually called the *Kalman gain*. A possible initial condition consists for instance in setting $X(0) = e_i$, that is one initial individual in a given state $i \in [J]$, with $P(0) = 0_J$, the $J \times J$ null matrix.

The above discussion shows how to leverage the model structure to provide an alternative approach to maximum likelihood estimation approaches.

Such techniques can be used in a single population scenario, as we illustrate in the next section. They could however become prohibitive for the multi-population scenarios we shall consider later, and for which the Kalman filtering approach may be an appealing alternative.

3 Numerical illustration on Paris hospitalization data

In this section, we illustrate the probabilistic version of our model by fitting it to daily hospitalization data, and show how to use it to forecast future numbers of hospitalizations.

The population we will consider is that of department 75 (Paris, France). We still assume that the only observed trajectory is $X_H^*(t)$, the number of new hospitalizations, all other trajectories being hidden. A noisy version of this data, namely the number of new COVID-19 hospitalizations every day in Paris, can be found in public data provided by SurSaUD syndromic surveillance system.

Though Kalman filtering approach – as described in Section 2 – would be natural here, we take a more direct, brute-force method, using Monte-Carlo simulation to infer the parameters.

3.1 Simulation settings

For simulations of our probabilistic model, we take $p_i = 0.7$, $p_h = 0.05$ (see Transition probabilities in Section 1). We initialize the time range on February 8th, which is 18 days before the first local SARS-CoV2 death case in France. We focus on the pre-confinement and confinement phases, so the fitting is made up to May 5th, on a total of 72 days for the train set. The period from May 6th to May 12th is kept as the test set for prediction (see below).

Recall that for contact parameters, we assume $\alpha_i = \alpha_a = \alpha$. Furthermore we assume that it is piecewise constant on the time range, as follows:

- between day 0 (February 8th) and day t_2 , $\alpha = \alpha_1$.
- between day t_2 and day t_3 , $\alpha = \alpha_2$.
- between day t_3 and day T (May 5th), $\alpha = \alpha_3$.

The three phases can be identified to the *pre-confinement phase*, the *confinement phase I* and *confinement phase II*, each phase being the result of several changes: mainly political decisions, but also work habits, observance of the rules, etc. In order to get a fitting as realistic as possible, the boundaries of these three phases are also inferred.

Inferred parameters, loss function In total, six parameters $(X_E(0), \alpha_1, t_2, \alpha_2, t_3, \alpha_3)$ are tuned. $X_E(0)$ is the initial condition, that is the initial number of exposed individuals. This quantity is then divided and allocated to all variables $X_{E,d}(0)$ according to distribution p_E described in Section 1.

Given $(X_E(0), \alpha_1, t_2, \alpha_2, t_3, \alpha_3)$, we sample several independent trajectories of $X(t)$. Now we consider two choices for the loss function.

We first define the L^1 loss function as follows:

$$\mathcal{L}(X_E(0), \alpha_1, t_2, \alpha_2, t_3, \alpha_3) := \mathbb{E}_{X_E(0)} \left[\frac{1}{T} \sum_{t=0}^T |X_H(t) - X_H^*(t)| \right], \quad (14)$$

where X_H is the random trajectory under model 1, and X_H^* is the observed data.

Second, since we expect exponential growth for the variables of interest (see e.g. 1.2 or 2.1), we shall also consider the L^1 -log loss function:

$$\tilde{\mathcal{L}}(X_E(0), \alpha_1, t_2, \alpha_2, t_3, \alpha_3) := \mathbb{E}_{X_E(0)} \left[\frac{1}{T} \sum_{t=0}^T |\log X_H(t) - \log X_H^*(t)| \right]. \quad (15)$$

For both choices, we aim to minimize the loss function on our train set. To do so, we perform a naive grid search in six dimensions, computing at each step an empirical expectation of the loss for the chosen parameters over sampled trajectories, in order to compute approximate solutions $(\hat{X}_E(0), \hat{\alpha}_1, \hat{t}_2, \hat{\alpha}_2, \hat{t}_3, \hat{\alpha}_3)$ of

$$\arg \min_{\substack{X_E(0), \alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}_+ \\ t_2, t_3 \in [T]}} \text{Loss}(X_E(0), \alpha_1, t_2, \alpha_2, t_3, \alpha_3), \quad (16)$$

for $\text{Loss} \in \{\mathcal{L}, \tilde{\mathcal{L}}\}$.

Prediction After having estimated all the parameters, we use our model to predict the evolution of trajectories in a future time range. In our example, this time range runs from $T + 1$ (May 6th) to T_{pred} (May 12th), that is a period of one week.

The performance of our prediction (over our test set) is evaluated in terms of the mean L^1 norm

$$\mathcal{L}_{\text{pred}} := \mathbb{E}_{\text{fit}} \left[\frac{1}{T_{\text{pred}} - T} \sum_{t=T+1}^{T_{\text{pred}}} |X_H(t) - X_H^*(t)| \right], \quad (17)$$

where the probability \mathbb{P}_{fit} is defined for previously inferred parameters $(\hat{X}_E(0), \hat{\alpha}_1, \hat{t}_2, \hat{\alpha}_2, \hat{t}_3, \hat{\alpha}_3)$ in equation (16).

We chose as our test criterion in (17) absolute deviation rather than deviation between logarithms because this corresponds to an error in the absolute numbers of hospital entrances, which has arguably a clearer operational meaning.

3.2 Results

The inferred parameters with SurSaUD data, together with the training and test errors, are summed up in Table 1. Corresponding curves of inferred parameters, predicted hidden variables and hospital entrances are given in Figure 3 and Figure 4.

Loss function	$\hat{X}_E(0)$	$\hat{\alpha}_1$	\hat{t}_2	$\hat{\alpha}_2$	\hat{t}_3	$\hat{\alpha}_3$	Fitting error	Prediction error $\mathcal{L}_{\text{pred}}$
$L^1(\mathcal{L})$	59	0.360	03-15	0.215	03-20	0.042	11.1195 ± 0.0232	4.0093 ± 0.0185
$L^1\text{-log}(\tilde{\mathcal{L}})$	18	0.455	03-15	0.265	03-20	0.040	0.29904 ± 0.00062	3.7201 ± 0.0125

Table 1: Approximate values of inferred parameters, for department 75, approximated with the probabilistic model, with SurSaUD data. All errors are given with their 95% confidence interval obtained by Monte Carlo simulations.

Several remarks can be made from these results:

- The $\mathcal{L}_{\text{pred}}$ prediction error is of order 4, which is to be compared to the typical diameter of a 95% confidence interval in the prediction time range in Figure 3 and Figure 4, which is about 10. We measure a prediction error which is less than half a typical uncertainty on a sampled stochastic trajectory $X_H(t)$ with fitted parameters. This supports the claim that a prediction error of order 4 is the best possible with our model.

- For both loss functions, the optimal transition points t_1 and t_2 are identified on March 15th and March 20th. Some other strategies could have been used to identify phases, e.g. by fitting a piecewise linear regression model on the observations $\log X_H^*(t)$, and reporting them earlier on the contact rates, according to the distributions of phase durations, as done in recent work by Gaubert et al. [5]. For the Paris population, the major political decisions and events occurred on March 12th (closure of all schools and universities), March 14th (closure of all non essential public places), March 15th (municipal elections) and March 17th (lock-down). These phases do not necessarily exactly coincide with the transition points, this can be caused e.g. by assumptions made for time parameters (see Section 1), and delays due to observance of the rules. In addition, we fixed three phases, but this number is arbitrary and could be discussed (see e.g. [10]).
- From the figures, the L^1 -log Loss function $\tilde{\mathcal{L}}$ seems to give better results for the fitting part (which is not surprising since we are in log scale), but this is also true for prediction, as presented in Table 1. This can justify the use of the L^1 -log Loss function $\tilde{\mathcal{L}}$ in future work.
- These results, though based on noisy data and perfectible with other datasets or broader inference, validate our model: with few parameters, the evolution of epidemics can be described in a satisfactory way: with the L^1 -log Loss function for instance, the mean relative error is $\sim 12.1\%$ (in logarithmic scale) for the trajectory of $X_H(t)$.

As shown in this section, the model of Figure 1 can be useful to describe the evolution of crucial variables such as the number of daily hospitalized individuals, but also show hidden trajectories, and predict hospital load on future days.

We considered piecewise constant contact rates in Section 2 for inference purposes, and performed corresponding numerical illustration in the present Section. We shall now consider the case of contact rates that are no longer piecewise constant, being potentially dependent not just on “shocks” such as the start of confinement, but also on other covariates. For instance, regular day-to-day mobility or commuting flows of population could also impact contact rates, as we now investigate.

4 Variable contact rates

Contact rates are now assumed to depend on piece-wise constant rates depending on specific temporal phases, as described before, and daily variables that are responsible for additional variability. Among these daily variables, some are related to user mobility, and in particular on the daily outflow, that is the number of day-return trips outside the region of interest.

The model we advocate is then exactly as in Section 1, except that contact rates α_i, α_a are now time-dependent.

Mobility-dependent contact rates Assuming that the daily outflows $f(t)$ (assumed to be normalized so as to have zero mean and standard deviation 1) are known, the contact rates α_a, α_i are now mobility-dependent in the following way:

$$\begin{aligned}\alpha_i(t) &= F_i(\phi(t), f(t)), \\ \alpha_a(t) &= F_a(\phi(t), f(t)).\end{aligned}\tag{18}$$

In the above, $\phi(t)$ is a discrete phase (e.g. $\phi(t) \in \{1, 2, 3\}$ as in Section 3), and F_i, F_a are fixed functions. These could be taken as logistic functions in the variable $f(t)$; alternatively, one might take

$$\begin{aligned}\alpha_i(t) &= \alpha_{i,\phi(t)}(1 + \gamma_i f(t)), \\ \alpha_a(t) &= \alpha_{a,\phi(t)}(1 + \gamma_a f(t)).\end{aligned}\tag{19}$$

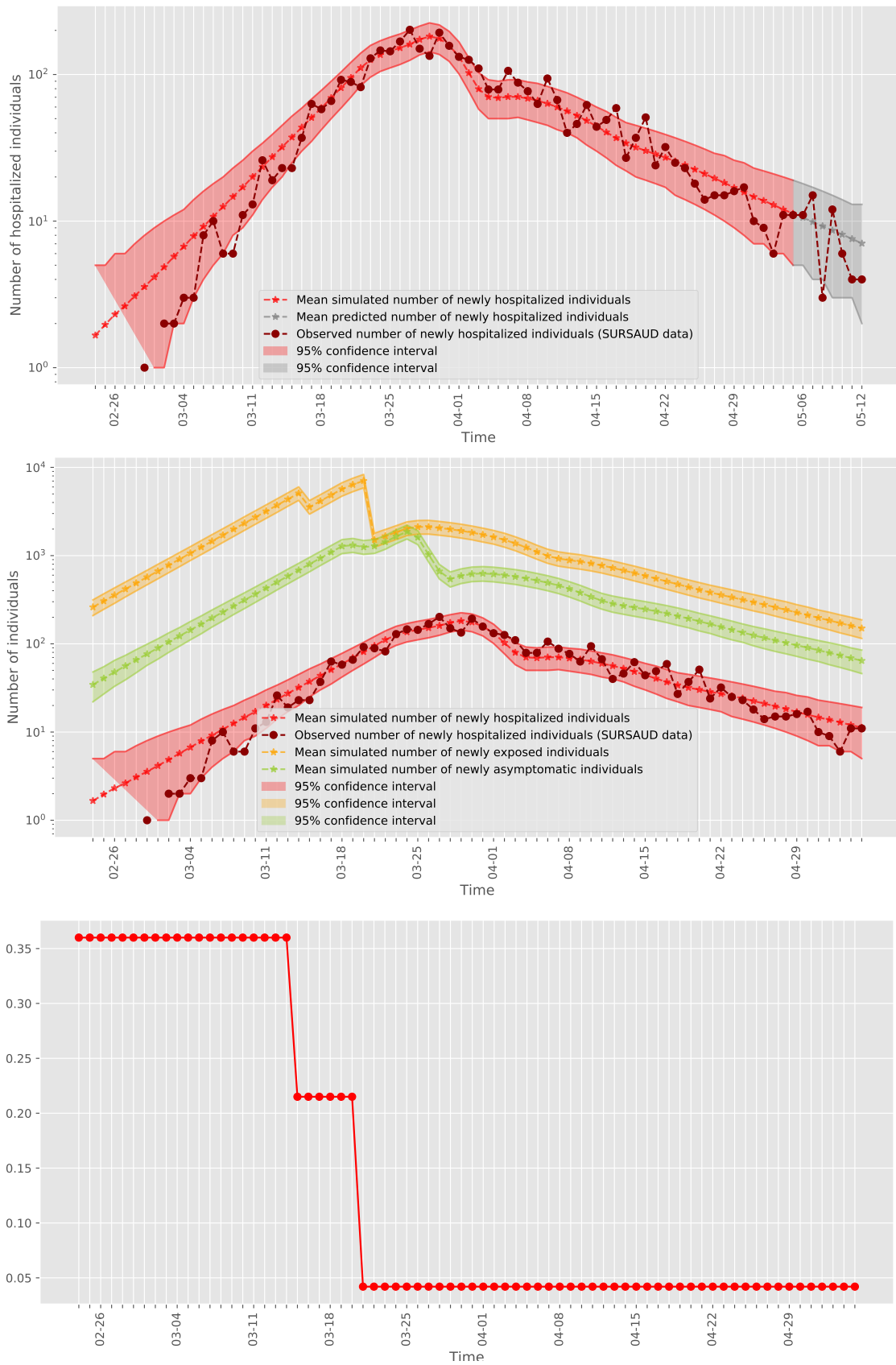


Figure 3: Best fitting of our probabilistic model with SurSaUD data for Paris with L^1 cost function, with prediction, some hidden trajectories, and variation of contact rates $\alpha_a = \alpha_i$.

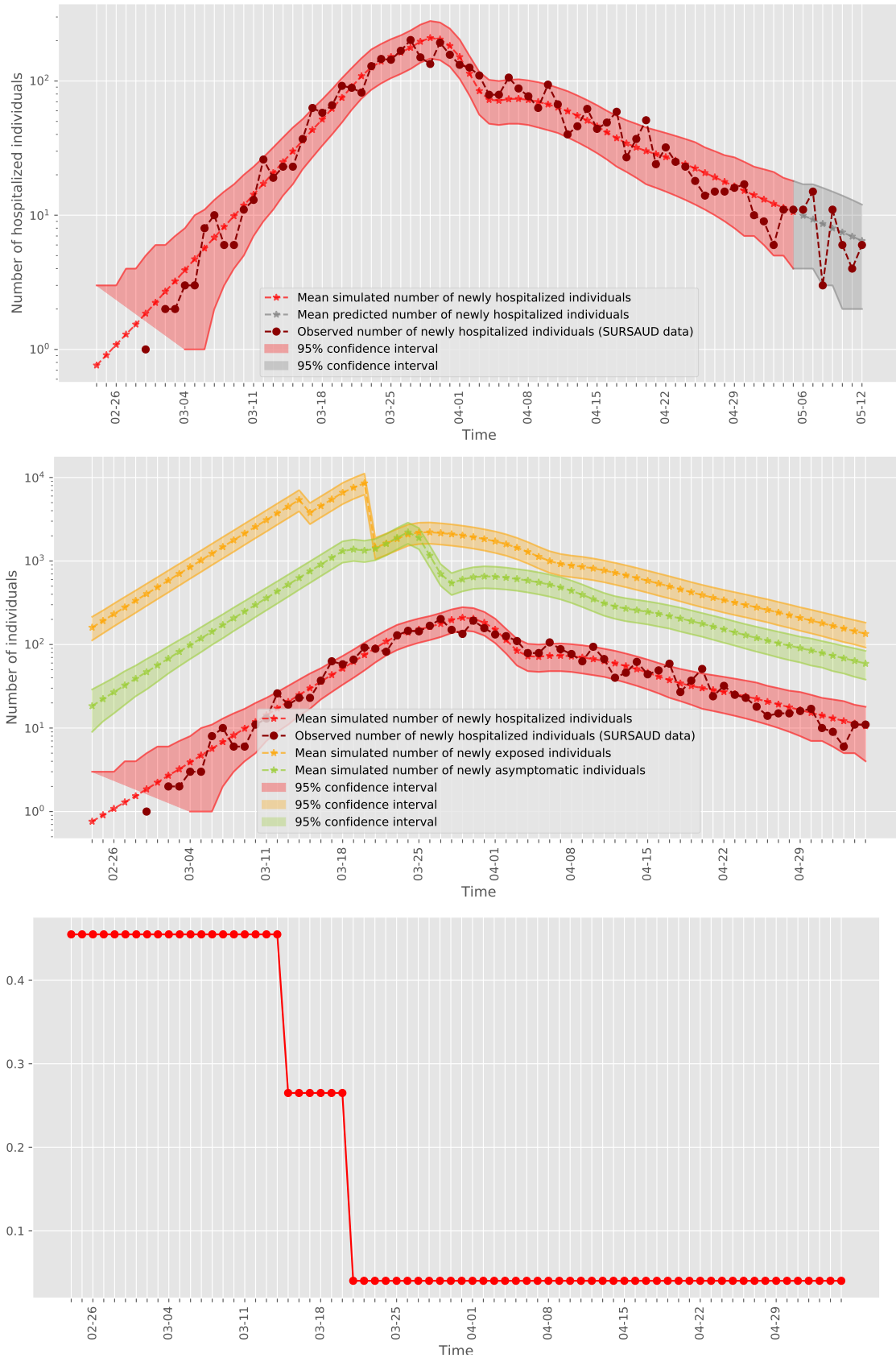


Figure 4: Best fitting of our probabilistic model with SurSaUD data for Paris with $\log-L^1$ cost function, with prediction, some hidden trajectories, and variation of contact rates $\alpha_a = \alpha_i$.

With this formulation, variables $\alpha_{i,\phi(t)}$ (resp. $\alpha_{a,\phi(t)}$) is the value of α_i (resp. α_a) when the commuting flow is at its equilibrium, and γ_i, γ_a are non-negative parameters.

Figure 5 below gives illustrates the evolution of the contact rates for Paris based on Equation (19), where the observables $f(t)$ are obtained from SFR mobile operator data.

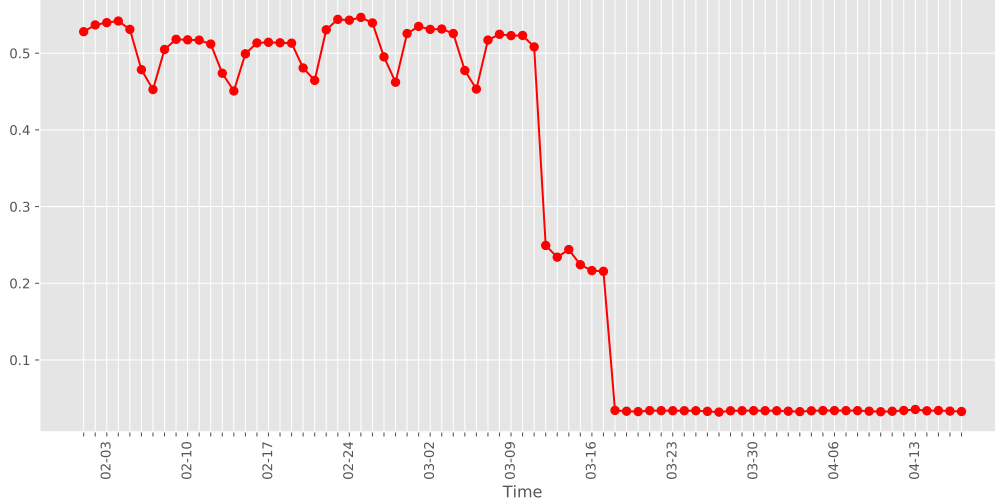


Figure 5: Example of mobility-dependent contact rates with equation (19) for Paris, based on SFR data. Parameters: $\alpha_{i,1} = \alpha_{a,1} = 0.45$, $\alpha_{i,2} = \alpha_{a,2} = 0.3$, $\alpha_{i,3} = \alpha_{a,3} = 0.05$, $\gamma_i = \gamma_a = 0.15$.

The shape of the resulting contact rates reflects the discrete phases as before, together with weekly periodicity patterns. Contact rates are lower on Saturdays and Sundays, when daily commuting trips are less numerous, as expected.

Remark 4.1. In its mean-field version, the model with variable contact rates is

$$x(t+1) = \left[M_0 + \sum_{k=1}^K \alpha_k(t) M_k \right] x(t),$$

for fixed matrices M_0, \dots, M_K and scalars $\alpha_k(t)$ that depend on mobility covariates relative to time t . Assume for tractability that the covariates $\alpha_k(t)$ are small, so that heuristically we have the first order approximation, for $\tau \in [d]$,

$$x(t) \approx M_0^\tau x(t-\tau) + \sum_{i=1}^{\tau} \sum_{k=1}^K \alpha_k(t-i) M_0^{i-1} M_k x(t-i).$$

Following on Remark 2.1, let $p(z) = z^d - \sum_{i=1}^d a_i z^{d-i}$ be a degree- d polynomial such that $p(M_0) = 0$. We then have the first-order expansion in the $\alpha(t)$:

$$\begin{aligned} x(t) &\approx \sum_{i=1}^d a_i M_0^{d-i} x(t-d) + \sum_{i=1}^d \sum_{k=1}^K \alpha_k(t-i) M_0^{i-1} M_k x(t-i) \\ &\approx \sum_{i=1}^d a_i \left[x(t-i) - \sum_{j=1}^{d-i} \sum_{k=1}^K \alpha_k(t-i-j) M_0^{j-1} M_k x(t-i-j) \right] + \sum_{i=1}^d \sum_{k=1}^K \alpha_k(t-i) M_0^{i-1} M_k x(t-i). \end{aligned}$$

Provided that for all $i = 0, \dots, d-1$ and all $k \in [K]$, there exist matrices $N_{k,i}$ and polynomials $p_{k,i}(z) = \sum_{r=0}^{d(k,i)} b_r(k,i)z^r$ of degree $d(k,i) \leq i$ such that $P_H M_0^i M_k = N_{k,i} P_H p_{k,i}(M_0)$, then up to terms of first order one has

$$P_H M_0^i M_k x(s) = N_{k,i} \sum_{r=0}^{d(k,i)} b_r(k,i) P_H M_0^r x(s) \approx N_{k,i} \sum_{r=0}^{d(k,i)} b_r(k,i) x_H(s+r).$$

Plugged into the previous display, this entails that the observations $x_H(t)$ satisfy an auto-regressive relation with coefficients involving the covariates α , of general form:

$$x_H(t) = \sum_{i,j=1}^d \alpha(t-i) P_{i,j} x_H(t-j), \quad (20)$$

where $P_{i,j}$ are constant matrices. A possible approach could then consist, for prediction purposes, in fitting such an auto-regressive relation while learning suitable functions $\alpha(t)$ of available covariates.

Besides commuting outflows mentioned above, another aspect of mobility that plays an important role is that of *routing*, that is change of residence location: this happens during holiday periods, but has also happened at the start of lockdown, where a significant exodus from dense urban areas towards countryside has been observed.

We now describe an extension of our model that captures such routing.

5 Model with Routing Mobility

To introduce routing, we must first consider an extension of our basic model to cover multiple cohorts, or sub-populations. For the sake of readability we only describe the generalized model in its mean-field version. However a probabilistic version could easily be described along the same lines as in the single population case.

We thus introduce a subdivision of the global population into cohorts $c \in \mathcal{C}$. These could for instance be given by a couple (r, a) characterizing geographical regions r and age ranges a .

A potential alternative to the above definition of cohort c is to define sub-population according to a given age range a , the region r in which the corresponding persons usually live, and the region r' in which they slept the previous night. In that case, we have $c = (r, r', a)$. This notion of cohort is particularly important to describe our routing model, and account for viral dissemination through regions.

We denote by $\mathcal{N}_c(t)$ the size of cohort c at time t .

5.1 Model parameters for subpopulations

The model is as follows. Each cohort $c \in \mathcal{C}$ has its own biological parameters θ_c (encapsulating e.g. contact rates, distributions of phase durations, etc.). A natural assumption is to take these to depend only on age range. Per age parameters could be obtained from hospital-collected statistics.

Routing dynamics of $\mathcal{N}_c(t)$ For a given cohort $c \in \mathcal{C}$, the time evolution of $\mathcal{N}_c(t)$ is the result of population flows. We represent such flows in the following manner: we have

$$\mathcal{N}_c(t+1) = \sum_{c' \in \mathcal{C}} R_{c',c}(t) \mathcal{N}_{c'}(t) + \mathcal{E}_c(t+1), \quad (21)$$

where $R_{c',c}(t)$ is the fraction of individuals counted in $\mathcal{N}_{c'}(t)$ on day t , that have migrated to be counted in $\mathcal{N}_c(t+1)$ on day $t+1$. The case $c' = c$ of sedentary individuals (no migration) is also

taken into account. New external arrivals in c (say, from a foreign country) are accounted for by $\mathcal{E}_c(t+1)$. We also assume that for all cohort c' ,

$$\sum_{c \in \mathcal{C}} R_{c',c}(t) \leq 1, \quad (22)$$

allowing for departures outside the considered population (say, abroad) when the sum is strictly less than 1.

Daily activity, contact between sub-populations On day t , a contact intensity factor $n_{c,c'}(t)$ gives the average number of individuals from cohort c' that a typical individual from cohort c will encounter. It is natural to assume that

$$\mathcal{N}_c(t)n_{c,c'}(t) = \mathcal{N}_{c'}(t)n_{c',c}(t) \quad (23)$$

Parameter $q_{c,c'}$ represents the probability that upon a contact between individuals from population c and population c' , infection gets propagated from that in c to that in c' . It is natural to assume that $q_{c,c'}$ only depends on the age ranges of c and c' . We then let

$$\alpha_{c,c'}(t) := q_{c,c'}n_{c,c'}(t), \quad (24)$$

the infection rate from c to c' on day t . For simplicity we do not distinguish between asymptomatic/prodromic (previously α_a) and symptomatic individuals (previously α_i) here.

Note that the above dynamics, and in particular Equations (23), (24) generalize the mobility-dependent contact rates of Section 4 to the multi-subpopulation case.

5.2 Routing step in the mean field version

Each sub-population c on day t is characterized by state vector $x^c(t)$, given by concatenation of vectors $x_E^c, x_P^c, x_{I_1}^c, x_A^c, x_{I_2}^c, x_H^c$. As before this is a vector of size $5h+1$, h being the maximal number of days each phase lasts.

To specify the transition between day t and day $t+1$ we consider the following two-stage mechanism: a first stage consists in an operation similar to the previously considered dynamics, extended to account for multiple cohorts c , giving vector $y(t)$ as a linear transform of vector $x(t)$. A second stage representing the routing of individuals between cohorts gives $x(t+1)$ as a linear transform of $y(t)$. Specifically we have

$$y^c(t) = N_{c,c}(t)x^c(t) + \sum_{c'} \alpha_{c',c}(t)M_{c,c'}x^{c'}(t), \quad (25)$$

where matrices $N = (N_{c,c})_{c,c' \in \mathcal{C}}$ and $M = (M_{c,c'})_{c,c' \in \mathcal{C}}$ are characterized by the equations

$$\begin{aligned} \forall \tau \in \{E, P, I_1, I_2, A\}, \quad y_{\tau,d+1}^c(t) &= x_{\tau,d}^c(t)(1 - r_\tau(d)), \\ y_{E,1}^c(t) &= \sum_{c' \in \mathcal{C}} \alpha_{c',c}(t) \sum_{\delta > 0} \left[(x_{I_1,\delta}^{c'} + x_{I_2,\delta}^{c'})(t) + (x_{A,\delta}^{c'} + x_{P,\delta}^{c'})(t) \right], \\ y_{P,1}^c(t) &= \sum_{\delta > 0} x_{E,\delta}^c(t)r_E(\delta), \\ y_{I_1,1}^c(t) &= p_i \sum_{\delta > 0} x_{P,\delta}^c(t)r_P(\delta), \\ y_{A,1}^c(t) &= (1 - p_i) \sum_{\delta > 0} x_{P,\delta}^c(t)r_P(\delta), \\ y_{I_2,1}^c(t) &= (1 - p_h) \sum_{\delta > 0} x_{I_1,\delta}^c(t)r_{I_1}(\delta), \\ y_H^c(t) &= p_h \sum_{\delta > 0} x_{I_1,\delta}^c(t)r_{I_1}(\delta). \end{aligned}$$

The routing stage is then provided by:

$$\begin{pmatrix} x_E^c \\ x_P^c \\ x_{I_1}^c \\ x_A^c \\ x_{I_2}^c \end{pmatrix} (t+1) = \sum_{c'} R_{c',c}(t) \begin{pmatrix} y_E^c \\ y_P^c \\ y_{I_1}^c \\ y_A^c \\ y_{I_2}^c \end{pmatrix} (t).$$

5.3 Estimation of routing fractions and numbers of contacts

Observation/estimation of routing fractions An ideal scenario is when routing fractions $R_{c,c'}(t)$ are directly available. However, this is typically not the case. We therefore describe a more plausible situation, for which we also propose a potential approach for estimating these fractions.

Assume that $c = (r, r', a)$, r representing the usual address of individuals and r' the place where they spent the previous night. Assume that we observe the quantities $\mathcal{N}_c(t)$, as well as, for all pair of regions (r_1, r_2) , the quantities:

$$\Delta_{r_1, r_2, a}(t) := \sum_{r, r'} \mathcal{N}_{(r, r_1, a)}(t) R_{(r, r_1, a), (r', r_2, a)}(t), \quad (26)$$

that is the flow of people of age range a who slept on night before t in r_1 and slept on the following night in r_2 . The usual address of these individuals is however assumed unknown.

In such a situation, we propose the following approach, popular in the literature on traffic matrices for management of communication networks. Estimate the unobserved quantities $R_{c,c'}(t)$ as the solution $\hat{R}_{c,c'}(t)$ of

$$\max_{R_{c,c'} \geq 0} \sum_{c, c'} R_{c,c'} \ln \left(\frac{1}{R_{c,c'}} \right) \quad (27a)$$

$$\text{subject to } \forall c', \sum_c R_{c',c} \leq 1 \quad (27b)$$

$$\forall (r_1, r_2), \sum_{r, r'} \mathcal{N}_{r, r_1, a}(t) R_{(r, r_1, a), (r', r_2, a)} = \Delta_{r_1, r_2, a}(t). \quad (27c)$$

This is a maximum entropy criterion. As a concave maximization program, it can be solved with efficient numerical methods.

Estimation of mean numbers of contacts $n_{c,c'}(t)$ Contacts counted in the contact intensity factor $n_{c,c'}(t)$ may occur in many circumstances. Among these we distinguish contacts at home, at work, at school, in transit (public transportation) as parameters on which preventive measures can be brought to bear. We may split such contacts accordingly, writing

$$n_{c,c'}(t) = n_{c,c'}^H(t) + n_{c,c'}^W(t) + n_{c,c'}^S(t) + n_{c,c'}^T(t). \quad (28)$$

Let us consider estimation of $n_{c,c'}^T(t)$. Assume we have access to additional observations about the mobility of individuals of all sub-populations. For instance, we may have statistics showing that on day t , approximately $n_{c,z}(t)$ individuals of population c have visited a location z . We may represent by z a time-space location, e.g. being in site s during a given hour h of the day.

Remark 5.1. The variables $n_{c,z}(t)$ could be unprocessed counts observed on day t . It may however be more appropriate to apply preliminary filtering on such unprocessed counts (e.g. applying a weighted averaging of raw counts over a past time window) to improve estimation quality of contact rates.

We may then let

$$\mathcal{N}_c(t)n_{c,c'}^T(t) = \sum_z \beta_z n_{c,z}(t)n_{c',z}(t), \quad (29)$$

where the parameters β_z are to be inferred, and capture the density of contacts to be expected in (time-space) location z . In this formulation, the parameters to be inferred are: parameters β_z , and infection probabilities $q_{c,c'}$. This model is over-parameterized, since multiplying the β_z by η and dividing the $q_{c,c'}$ by η leaves the model unchanged for all $\eta \neq 0$. This can be easily solved, forcing for instance the parameters β_z to have mean 1. While this over-parameterization is easily circumvented, it remains a challenge to effectively fit all the free parameters of the above model, as well as to define the proper spatio-temporal granularity associated with the time-space locations z .

Conclusion

In this study we laid the foundations of a general model of COVID-19 epidemics which captures user mobility and contact tracing. We described it as a multi-type branching process, proposed inference methods to perform short-term prediction, that we illustrated on the Paris hospitalization data. We also introduced extensions to capture contract tracing and case isolation, and finally proposed extensions to capture user mobility, distinguishing between commuting and routing mobility.

Future work will further exploit this model together with hospital incidence and mobility data to forecast epidemic progress and assess impact of mobility on infectious contacts.

Acknowledgements

The authors are grateful to SFR for making available to them time series of human mobility across departments in France. This data, partially illustrated by Figure 5, has inspired the design of the models described in the last two sections.

This work was partially supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

References

- [1] Andrew W. Byrne, David McEvoy, Aine Collins, Kevin Hunt, Miriam Casey, Ann Barber, Francis Butler, John Griffin, Elizabeth Lane, Conor McAloon, Kirsty O'Brien, Patrick Wall, Kieran Walsh, and Simon More. Inferred duration of infectious period of sars-cov-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic covid-19 cases. *medRxiv*, 2020.
- [2] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. A time-dependent sir model for covid-19 with undetectable infected persons. 2020. http://gibbs1.ee.nthu.edu.tw/A_TIME_DEPENDENT_SIR_MODEL_FOR_COVID_19.PDF.
- [3] L. Di Domenico, G. Pullano, C. Sabbatini, P.-Y. Boëlle, and V. Colizza. Expected impact of lockdown in Île-de-france and possible exit strategies. <https://www.medrxiv.org/content/10.1101/2020.04.13.20063933v1>, 2020.
- [4] Moez Draief, Ayalvadi Ganesh, and Laurent Massoulié. Thresholds for virus spread on networks. *Ann. Appl. Probab.*, 18(2):359–378, 04 2008.
- [5] Stéphane Gaubert, Marianne Akian, Xavier Allamigeon, Marin BOYET, Baptiste Colin, Théotime Grohens, Laurent Massoulié, David P. Parsons, Frederic Adnet, Érick Chanzy, Laurent Goix, Frédéric Lapostolle, Éric Lecarpentier, Christophe Leroy, Thomas Loeb, Jean-Sébastien Marx, Caroline Télion, Laurent Treluyer, and Pierre Carli. Understanding and monitoring the evolution of the Covid-19 epidemic from medical emergency calls: the example of the Paris area. arXiv:2005.14186, June 2020.
- [6] Mohinder Grewal and Angus Andrews. Kalman filtering: theory and practice using matlab. *New York: John Wiley and Sons*, 14, 01 2001.
- [7] H. Kesten and B. P. Stigum. A limit theorem for multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1211–1223, 1966.
- [8] Amaury Lambert. A mathematical assessment of the efficiency of quarantining and contact tracing in curbing the COVID-19 epidemic. *medRxiv*, 2020.
- [9] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020. PMID: 32150748.
- [10] Marc Lavielle. Modelling some covid-19 data. <http://webpopix.org/covidix19.html>.
- [11] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, 2020.
- [12] Conor G McAloon, Aine Collins, Kevin Hunt, Ann Barber, Andrew Byrne, Francis Butler, Miriam Casey, John M Griffin, Elizabeth Lane, David McEvoy, Patrick Wall, Martin J Green, Luke O'Grady, and Simon J More. The incubation period of covid-19: A rapid systematic review and meta-analysis of observational research. *medRxiv*, 2020.
- [13] Jonathan Roux, Clément Massonnaud, and Pascal Crépey. Covid-19: One-month impact of the french lockdown on the epidemic burden. *medRxiv*, 2020.
- [14] Rajesh Singh and R. Adhikari. Age-structured impact of social distancing on the covid-19 epidemic in india. <https://arxiv.org/pdf/2003.12055.pdf>, 2020.

- [15] Mircea T. Sofonea, Bastien Reyné, Baptiste Elie, Ramsès Djidjou-Demasse, Christian Selinger, Yannis Michalakis, and Samuel Alizon. Epidemiological monitoring and control perspectives: application of a parsimonious modelling framework to the covid-19 dynamics in france. *medRxiv*, 2020.
- [16] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick Walker, Han Fu, Amy Dighe, Jamie Griffin, Anne Cori, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma M Cucunuba, Rich Fitzjohn, Katy A M Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Daniel Laydon, Gemma Nedjati-Gilani, Steven Riley, Sabine van Elsland, Erik Volz, Haowei Wang, Yuanrong Wang, Xiayoue Xi, Christl Donnelly, Azra Ghani, and Neil Ferguson. Estimates of the severity of covid-19 disease. *medRxiv*, 2020.