SURVEILLANCE-BASED ESTIMATES OF THE REPRODUCTIVE NUMBER MAY BE BIASED IN SPATIALLY STRUCTURED POPULATIONS

OCTOBER, 2022

MASTER OF SCIENCE
IN
PHYSICS OF COMPLEX SYSTEMS

Candidate: **PIERO BIRELLO**
Under the supervision of LUCA DALL'ASTA and EUGENIO VALDANO

Université de Paris - Sorbonne Université - Université Paris-Saclay
Politecnico di Torino - SISSA - ICTP

Institute Pierre-Luis d'Epidémiologie et de Santé Publique
Équipe SUMO: surveillance et modélisation des maladies transmissibles

Abstract

The reproductive number R of an epidemic is defined as the expected number of secondary cases produced by a typical infected individual over its entire period of infectiousness. An accurate and timely estimate of the reproductive number is crucial to make projections on the near-future evolution of the epidemic, and to set up the appropriate public health response. Estimates of R often come from surveillance data, as it has been in the case of the SARS-CoV-2 pandemic. This means statistically inferring R from time series of daily reported cases, hospitalizations or deaths. In this study, however, we argue that surveillance-based measures of the reproductive number may not always be accurate measures of the true reproductive number. We focus on structured populations, made up of spatially distinct communities, for which it is known that R corresponds to the dominant eigenvalue of a positive linear operator. We show that the reproductive number measured by surveillance approaches R only after a period of transient behaviour, during which we tipically underestimate R. In some cases, convergence is not ensured at all. Similarly, local (i.e., community-level) estimates of the reproductive number are inaccurate in describing the global epidemic dynamics, reaching R only asymptotically. However, we show that combining surveillance data and mobility data we are able to give reliable estimates of the reproductive number even at the early stages of an epidemic.

Precisely, we consider the SARS-CoV-2 epidemic in France and in Italy. We study the transient period analytically, through SARS-CoV-2 cases data and building a spatial stochastic model with interactions reconstructed from Meta Colocation Maps. We analyse the dependence of the dynamical process leading statistical estimates to the true R on the initial distribution of cases and on the network topology. We study the impact of restrictions to mobility on the accuracy of R estimates. Finally, we propose a new method for the estimate of the reproductive number and test its reliability through simulations of the mentioned stochastic model.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Quantifying the conditions that discriminate between large-scale outbreak and quick disease extinction is at the core of mathematical modeling of epidemic diseases and of public health policy making in response to epidemics [1, 2]. The parameter that is the most used to this aim is the reproduction (or reproductive) number $\mathcal{R}$, representing the average number of secondary cases caused by an infected individual over their entire period of infectiousness. We call basic reproduction number $\mathcal{R}_0$ the reproduction number at the early stages of an epidemic, when the population is completely susceptible [3, 4]. A threshold criterion exists involving $\mathcal{R}_0$: *a disease can invade a population when $\mathcal{R}_0 > 1$, it cannot when $\mathcal{R}_0 < 1$* [5]. Similarly, at later stages, an outbreak occurs if $\mathcal{R} > 1$, it doesn't otherwise. It is therefore crucial to evaluate these parameters to make projections on the near-future evolution of the epidemic and control its spreading.

We will now introduce a well known mathematical model for epidemic spreading and discuss how to analytically obtain the basic reproduction number. We will then consider the case of heterogeneously structured populations, and define the basic reproduction number for such systems.

1.1 Mathematical modeling of epidemic diseases

We now introduce a simple mathematical model for epidemic spreading, the susceptible-infected-recovered (SIR) model [6], and we obtain the basic reproduction number $\mathcal{R}_0$ [7]. The model considers a system with fixed total population N, thus ignoring phenomena such as migrations, births and deaths. Individuals may be in one of three compartments [5, 6, 8, 1] S, I or R. We denote by italic letters S, I, R the number of individuals in each compartment. The transition S $\rightarrow$ I occurs when a susceptible individual interacts with an infectious individual and becomes infected, while the transition I $\rightarrow$ R spontaneously occurs when an individual recovers after some time fighting the disease. In a continuous-time formulation, it is common to assume a Poisson process [9] with rate μ estimated from epidemiological or clinical data, implying that the probability that an individual remains infected for a time τ follows $P(\tau) = \mu e^{-\mu\tau}$, with average infectious time $\langle\tau\rangle = \mu^{-1}$. The S $\rightarrow$ I transition depends on the interaction pattern between individuals in the population instead. However, the most basic approach consists in assuming homogeneous mixing [5], meaning that people randomly interact with each other. This is equivalent to a statistical physics mean field assumption. In this case, we can define the force of infection α, representing the rate for one individual to get infected, simply as:

$$\alpha = \beta I/N, \tag{1.1}$$

with $\beta = \beta' k$, β' being the rate of infection per effective contact, depending on the specific disease, and k the number of contacts. If we assume the infection process to be a Poisson process as well, with rate α, what we obtain is a Markovian description of the epidemic process [10].

To sum up, epidemics can be represented as stochastic reaction-diffusion processes [11], with continuous-time limit equations describing the evolution of the average number of individuals in each compartment. Moreover, it is common to neglect stochastic fluctuations around these mean numbers, according to the assumption that populations are large. In this limit and assuming homogeneous mixing, the system's time evolution can be described by deterministic differential equations obtained applying the law of mass action, with reactions (transitions) determined by specific reaction rates. This law states that the average change in the population density of each compartment due to interactions is given by the product of the force of infection times the average population density [12]. The deterministic differential equations for the SIR model read:

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS/N \\ \frac{dI}{dt} &= \beta IS/N - \mu I \\ \frac{dR}{dt} &= \mu I. \end{aligned} \tag{1.2}$$

Taking the limit $I/N \simeq 0$, which is typically true at the early stages of an epidemic, we can linearize to obtain:

$$\frac{dI}{dt} = (\beta - \mu)I, \tag{1.3}$$

from which:

$$I(t) = I(0)e^{(\beta-\mu)t}. \tag{1.4}$$

The number of infected individuals grows exponentially if $\mathcal{R}_0 = \beta/\mu > 1$. $\mathcal{R}_0$ is here the basic reproduction number, the average number of secondary cases caused by a primary case in a fully susceptible population, and we have recovered the mentioned threshold criterion.

1.2 Epidemics in heterogeneous populations

We should now consider the more complex case of a non homogeneously structured population. Diekmann, Heesterbeek and Metz [13] elaborated a versatile analytical framework to define the basic reproduction number $\mathcal{R}_0$ in a generation process, in the case of a generically heterogeneous population and under linear approximation, i.e., ignoring the fact that the density of susceptibles in the population decreases due to the infection process. Let each individual belong to a heterogeneity state (*h*-state) ξ in the set Ω. This may be the set of age classes, gender, of geographic location and others depending on the epidemic under study and its characteristics. This setting allows us to define $A(\tau, \xi, \eta)$, the specific expected infectivity of an individual which was infected τ time units ago and is in the *h*-state ξ, towards one in the *h*-state η. Let then x be the initial distribution of infectious individuals over the defined *h*-states and $S(\xi)$ the non-normalized density function of susceptibles over classes (its integral gives the total population size). The next-generation operator $Q(S)$ is defined by:

$$(Q(S)x)(\xi) = S(\xi) \int_{\Omega} \int_{0}^{\infty} A(\tau, \xi, \eta) x(\eta) d\tau d\eta \tag{1.5}$$

and tells how many cases are generated from the initial distribution x and how they are distributed over the *h*-states. In the linear approximation, after m generations of the epidemic process, the infected population size over *h*-states is given by $Q(S)^m x$. Under minor conditions on A and S [14] we have:

$$Q(S)^m x \sim c(x) \rho_d^m x_d, \tag{1.6}$$

with $c(x)$ a scalar depending on the initial distribution x, ρ_d the dominant eigenvalue of $Q(S)$ and x_d the corresponding eigenvector. Also, in the long time limit $m \to \infty$, the growth factor per generation is given by the spectral radius of $Q(S)$:

$$r(Q(S)) = \inf_m \|Q(S)^m\|^{1/m} = \lim_{m \to \infty} \|Q(S)^m\|^{1/m}. \tag{1.7}$$

Being $Q(S)$ a positive operator, one can specify conditions under which $r(Q(S)) = \rho_d$ [15]. The per-generation growth factor in the susceptible-only population assumption is exactly what we define as the basic reproduction number. In conclusion, what Diekmann, Heesterbeek and Metz found is that *after a certain period* *of transient behaviour each generation is (in an approximation which improves as time proceeds) $\mathcal{R}_0 = \rho_d$ times as big as the preceding one and distributed over h-state space as described by x_d.*

We will consider the case where Ω is the set of geographic sub-populations of a country, focusing in particular on French departments and Italian provinces. We will reconstruct the expected between-departments (or provinces) infectivity from Meta Colocation Maps describing citizens mobility from the start of the COVID-19 pandemic [16]. We will study how the mobility network of the two chosen countries determines the equilibrium growth factor in France and Italy as well as the equilibrium *h*-state space distribution. Most importantly, we will focus on the time and spatial scales of the convergence process. We will provide an analytical, simulations based and data based description of the out-of-equilibrium phase, in order to get some insight on this complex dynamics and determine the error we commit when estimating $\mathcal{R}$ out of equilibrium. Eventually, we will propose a new non biased method for the estimate of the reproductive number and compare its performance to the one of canonical estimates.

Chapter 2

Analytical treatment

Our aim is to expand Diekmann *et al.* work with an analytical treatment of the transient phase of an epidemic, as well as to include different space scales of observation in our analysis. We want to describe how the measured reproduction number -both at a local and a global scale, as we will later see- behaves out-of-equilibrium. With this purpose, we will set our analytical framework and describe the dynamics for the involved quantities. We will be able to determine the laws for the time evolution of the defined reproduction numbers, and to identify the variables that are relevant in this process.

2.1 The analytical framework

We here expand Diekmann *et al.* work with an analytical treatment of the transient phase of an epidemic in a generic discrete Ω case and, consequently, a discrete between classes expected infectivity operator. Consider M nodes, possibly representing spatial patches in a geographic territory. Let A_{ij} be the expected infectivity of an individual of the *h*-class j towards one in the *h*-class i, and S the susceptible population vector, with S_i the number of susceptibles in *h*-class i. x is now the distribution vector of infected individuals over the M states. Take a discrete Ω, time integrated version of Eq. (1.5):

$$(Q(S)x)_i = S_i \sum_j A_{ij} x_j, \tag{2.1}$$

from which

$$Q_{ij}(S) = S_i A_{ij}. \tag{2.2}$$

The element $Q_{ij}(S)$ of this discrete, time independent next-generation operator $Q(S) \in \mathbb{R}^{M,M}$, which we will call *reproduction operator*, gives the expected number

of infectious cases that a single case in j generates in i, given S [[19]]. We use the notation Q for $Q(S)$, dropping the S dependence, and the notation Q_0 in the case the whole population is susceptible. According to Diekmann and more recent studies [[17], [18]], the basic reproduction number is $\mathcal{R}_0 = r(Q_0)$ and the reproduction number is $\mathcal{R} = \mathcal{R}(S) = r(Q)$.
Some further definitions are needed for later use. We define the vector I as I_i being the number of actively infectious in i, with $I_{tot} = \sum_i I_i$, $I = xI_{tot}$. We will refer to the potential vector q^T as the vector such that $q_i = \sum_j Q_{ji}$. q_i is hence the number of secondary cases that a case in i generates, regardless of where it generates them. A key quantity in our analysis is the *observed* reproductive number $\mathcal{S}$, describing how many secondary cases currently derive from a primary one. It is given by the product between the potential vector and the actively infectious distribution vector:

$$\mathcal{S} = q^T x = \sum_{ij} Q_{ij} x_j. \tag{2.3}$$

Finally, the *observed local* reproductive number is:

$$s_i = \frac{(QI)_i}{I_i} = \frac{(Qx)_i}{x_i}, \tag{2.4}$$

measuring the number of cases that are currently produced in i due to interactions with the whole system, per each infected in i.

When investigating an epidemic spread, the role of all the introduced numbers should be clear. The true reproduction number $\mathcal{R}$ is the one that determines the dynamics of the system, as we will see in Eq. (2.5), and it identifies whether there is a mode - associated to the dominant eigenvalue of the reproduction operator - that grows exponentially and ultimately leads to an epidemic wave. The observed reproductive number is instead what can be measured from surveillance data. People typically infer it from time lines of cases, hospitalisations or deaths using statistical models (e.g., `EpiEstim` [[20]]). In our framework, $\mathcal{S}$ is simply the average number of cases generated in generation $t+1$, by a case in generation t. It can be computed on the whole system ($\mathcal{S}$), on patches (s_i) or even on subsets of patches.

2.2 Dynamics in time

Let's now derive the dynamical equations for I and x. Here one time step is one generation of the epidemic process. We denote I_t the vector of cases in the t-th generation. These will generate new cases in the next generation, and recover. We will now assume constant Q. This means keeping Diekmann assumption that the susceptible fraction is constant and considering A to be non varying with time.

The I-dynamics is then linear:

$$I_{t+1} = QI_t, \tag{2.5}$$

which means

$$I_t = Q^t I_0. \tag{2.6}$$

The above equation gives the expected number of cases in each patch, in each generation.

The dynamics of the distribution vector is nonlinear. Eq. (2.5) can be rewritten as:

$$x_{t+1} I_{t+1,tot} = Qx_t I_{t,tot} \tag{2.7}$$

from which

$$x_{t+1} = \frac{Qx_t}{q^T x_t}, \tag{2.8}$$

$$x_t = \frac{Q^t x_0}{q^T Q^{t-1} x_0}. \tag{2.9}$$

There is one case when the x-dynamics is linear, precisely if Q is proportional to a left-stochastic matrix, i.e., $Q = q_0 C$, with q_0 scalar, $C \in \mathbb{R}^{M,M}$, and $\sum_j C_{ji} = 1$. Physically, this means that each case, independently of its *h*-class, generates always the same number of secondary cases (q_0), and C just tells where they generate them. In this case:

$$x_{t+1} = Cx_t. \tag{2.10}$$

The dynamics is at *equilibrium* when $x_{t+1} = x_t$. This means that I_{tot} may still change (increase or decrease), but the shape of the distribution is no longer changing, so that the dynamics is effectively scalar: $I_{tot,t+1} = \mathcal{R} I_{tot,t}$. This happens when x is proportional to the dominant (*Perron*) eigenvector of the positively definite operator Q. We call this eigenvector v, and normalize it so that $\sum_i v_i = 1$. The equilibrium is then $x = v$. Note that v is the only non-negative eigenvector due to *Perron Frobenius theorem* [21], so it must be the only physical equilibrium, since each entry of x needs to be between zero and one.

2.3 $\mathcal{R}$ and $\mathcal{S}$ evolution

We define $\Delta = \mathcal{S} - \mathcal{R}$ as the difference between the reproductive number measured from surveillance ($\mathcal{S}$), and the true one ($\mathcal{R}$). This quantity is representative of the error we commit when inferring the reproduction number from recorded cases, which is the main focus of this work.

We use the following set of right eigenvectors for Q: v is the principal one, with eigenvalue $\mathcal{R}$, then I have w_μ with eigenvalues Λ_μ, $\mu \in [1, M-1]$. These eigenvalues may be repeated. Impose $\sum_i v_i = 1$, and

$$\sum_i w_{\mu,i} = \begin{cases} 1 \text{ if possible} \\ 0 \text{ if it is the case.} \end{cases} \tag{2.11}$$

Define then $\sum'_\mu$ as the sum on those μ for which $\sum_i w_{\mu,i} \neq 0$. We can decompose $x = gv + \sum_\mu h_\mu w_\mu$. From the normalization of x, I have $g = 1 - \sum'_\mu h_\mu$. We get:

$$\begin{aligned} \Delta &= \mathcal{S} - \mathcal{R} \\ &= \sum_{ij} Q_{ij} x_j - \mathcal{R} \\ &= \sum_{ij} Q_{ij} [(1 - \sum_\mu{}' h_\mu) v_j + \sum_\mu h_\mu w_{\mu,j}] - \mathcal{R} \\ &= \mathcal{R}(1 - \sum_\mu{}' h_\mu) + \sum_\mu{}' h_\mu \Lambda_\mu - \mathcal{R} \\ &= - \sum_\mu{}' \left(\mathcal{R} - \Lambda_\mu\right) h_\mu. \end{aligned} \tag{2.12}$$

There are two cases in which it is clear that $\Delta = 0$ identically. This happens when the dynamics is at equilibrium ($x = v$) or when Q is proportional to a stochastic matrix ($Q = q_0 C$).

Let's define the auxiliary variable $\lambda_\mu = \Lambda_\mu / \mathcal{R}$. Due to the Perron-Frobenius theorem $|\lambda_\mu| < 1$. We can derive the evolution in time of g, h_μ from Eq. (2.9):

$$\begin{aligned} g_t v + \sum_\mu h_{t,\mu} w_\mu &= \frac{Q^t (g_0 v + \sum_\mu h_{0,\mu} w_\mu)}{q^T Q^{t-1} (g_0 v + \sum_\mu h_{0,\mu} w_\mu)} \\ &= \frac{\mathcal{R}^t g_0 v + \sum_\mu \Lambda_\mu^t h_{0,\mu} w_\mu}{q^T (\mathcal{R}^{t-1} g_0 v + \sum_\mu \Lambda_\mu^{t-1} h_{0,\mu} w_\mu)} \\ &= \frac{\mathcal{R}^t g_0 v + \sum_\mu \Lambda_\mu^t h_{0,\mu} w_\mu}{\mathcal{R}^t g_0 + \sum_\mu{}' \Lambda_\mu^t h_{0,\mu}} \\ &= \frac{g_0 v + \sum_\mu \lambda_\mu^t h_{0,\mu} w_\mu}{g_0 + \sum_\mu{}' \lambda_\mu^t h_{0,\mu}}. \end{aligned} \tag{2.13}$$

Then:

$$g_t = \frac{g_0}{g_0 + \sum_\nu{}' \lambda_\nu^t h_{0,\nu}}, \tag{2.14}$$

$$h_{t,\mu} = \frac{\lambda_\mu^t h_{0,\mu}}{1 - \sum_\nu{}' (1 - \lambda_\nu^t) h_{0,\nu}}. \tag{2.15}$$

We stress that the denominators in Eq. (2.14) and Eq. (2.15) are the same, they are different ways of writing the same thing, using $g = 1 - \sum_\mu' h_\mu$. Also, this denominator is positive, since it is the observed growth ratio after t generations. As known, since $-1 < \lambda_\mu < 1$ the dynamics always brings towards $x = v$:

$$\left| \frac{h_{t,\mu}}{g_t} \right| = \left| \frac{h_{0,\mu}}{g_0} \right| \left| \lambda_\mu \right|^t . \tag{2.16}$$

The equation for the evolution in time of Δ can be obtained inserting Eq. (2.15) in Eq. (2.12):

$$\Delta_t = -\mathcal{R} \frac{\sum_\mu{}' \lambda_\mu^t (1 - \lambda_\mu) h_{0,\mu}}{1 - \sum_\nu{}' (1 - \lambda_\nu^t) h_{0,\nu}} . \tag{2.17}$$

We observe that the higher are the Λ_μ, the longer the time needed for $|h_{t,\mu}/g_t|$ to converge to zero. In particular, the second largest eigenvalue of Q, let it be Λ_1, is the one that is the most influential in determining the process convergence time. Also notice that convergence of $\mathcal{S}$ to $\mathcal{R}$ may present oscillations. If $\lambda_\mu < 0$, $h_{t,\mu}$ can be alternating. Then, *if negative eigenvalues exist, the measurement error can oscillate between positive and negative values while damping towards zero*. Roughly speaking, if Q is weakly coupled (almost diagonal), probably all eigenvalues are strictly positive, because they are close to the diagonal entries, which are of course positive. As a consequence, the error has always the same sign. If the system is strongly coupled, meaning that Q is highly non-diagonal, negative eigenvalues emerge and can lead to oscillations.

The concept of weak or strong coupling can be formalized as follows. The system is weakly coupled if Q is *strictly column diagonally dominant*, meaning $Q_{ii} > \sum_{j \neq i} Q_{ji}$, $\forall i$ [22]. Basically Q is weakly coupled if, for each h-class i, most of the secondary cases are generated locally (Q_{ii} is large), than in other h-classes ($\sum_{j \neq i} Q_{ji}$). The system is strongly coupled otherwise.

We can prove that oscillations can occur only if Q is strongly coupled. By *Gershgorin's circle theorem* [23],

$$\lambda_\mu \geq \min_i \left\{ Q_{ii} - \sum_{j \neq i} Q_{ji} \right\} . \tag{2.18}$$

If Q is weakly coupled then, by definition,

$$\lambda_\mu \geq \min_i \left\{ Q_{ii} - \sum_{j \neq i} Q_{ji} \right\} > 0. \tag{2.19}$$

This proves that weak coupling is a necessary (possibly not sufficient) condition for oscillations of $\mathcal{S}$ around $\mathcal{R}$ to occur.

Some interesting properties can be observed concerning local observed reproductive numbers too. At equilibrium,

$$s_i = \frac{(Qgv)_i}{gv_i} = \mathcal{R} = \mathcal{S} \qquad \forall i. \tag{2.20}$$

Every patch measures the same reproductive number, which is also the global reproductive number of the system (both the true $\mathcal{R}$ and the observed $\mathcal{S}$, which are the same at equilibrium). Note that the patch is not measuring its own true reproductive number q_i, representing secondary cases generated anywhere. It is measuring the global reproductive number of the system. In general, however, q_i is observable only if the system is fully decoupled, meaning Q is diagonal.

The equation for the evolution in time of the s_i can be obtained as well. Let's first expand x:

$$\begin{aligned} s_i &= \frac{(Q(gv + \sum_\mu h_\mu w_\mu))_i}{gv_i + \sum_\mu h_\mu w_{\mu,i}} \\ &= \frac{(1 - \sum_\mu{}' h_\mu)\mathcal{R} v_i + \sum_\mu h_\mu \Lambda_\mu w_{\mu,i}}{(1 - \sum_\mu{}' h_\mu) v_i + \sum_\mu h_\mu w_{\mu,i}}. \end{aligned} \tag{2.21}$$

Inserting Eq. (2.15):

$$\begin{aligned} s_{t,i} &= \mathcal{R}\frac{(1 - \sum_\mu{}' h_{0,\mu}) v_i + \sum_\mu \lambda_\mu^{t+1} h_{0,\mu} w_{\mu,i}}{(1 - \sum_\mu{}' h_{0,\mu}) v_i + \sum_\mu \lambda_\mu^{t} h_{0,\mu} w_{\mu,i}} \\ &= \mathcal{R}\frac{g_0 v_i + \sum_\mu \lambda_\mu^{t+1} h_{0,\mu} w_{\mu,i}}{g_0 v_i + \sum_\mu \lambda_\mu^{t} h_{0,\mu} w_{\mu,i}}. \end{aligned} \tag{2.22}$$

We now see that:

$$s_{t,i} > R \iff \sum_\mu \lambda_\mu^{t+1} h_{0,\mu} w_{\mu,i} > \sum_\mu \lambda_\mu^{t} h_{0,\mu} w_{\mu,i}.$$

This condition may be verified depending on the λ_μ, $h_{1,\mu}$ and $w_{\mu,i}$ which can all in principle assume negative values. Note in particular that we know all w_μ have at least one negative component due to Perron Frobenius theorem. Interestingly, we see that even in the weakly coupled case, where we showed that $\mathcal{S}$ cannot oscillate around $\mathcal{R}$, oscillations of the s_i around $\mathcal{R}$ can occur due to the negative signs of some of the $h_{0,\mu}$ and $w_{\mu,i}$.

Lastly, we point out that $\mathcal{S}$ can be written as a linear combination of the s_i:

$$\mathcal{S} = \sum_i s_i x_i. \tag{2.23}$$

This is relevant in terms of the threshold criterion. Since $x_i < 1 \; \forall i$, in order to observe a reproductive number that is larger than one we need at least one of the s_i to be larger than one.

2.4 Notes on Perron-Frobenius

Some notes can be made regarding the reproduction operator Q's properties. If we assume that $Q_{ii} > 0$, because each h-class always generates a nonzero amount of cases in itself, then Q has maximal rank, because it is the sum of a diagonal matrix with maximal rank, and a perturbation with some rank. This assumption is clearly true when h-classes are spatial patches.

As a second comment, Perron-Frobenius theorem requires that either Q is strictly positive or Q is non negative and irreducible. Q is irreducible if the associated directed graph is strongly connected. Under this conditions $\mathcal{R}$ is strictly positive and non degenerate and v has strictly positive components. Typically this is true, particularly if we think of h-classes as spatial patches. However, Perron-Frobenius theorem can be recovered even in the case requirements are not satisfied. Consider:

$$Q = \left(\begin{array}{c|c|c} T_u & 0 & 0 \\ \hline B_1 & Q_{scc} & 0 \\ \hline B_2 & B_3 & T_d \end{array} \right) . \tag{2.24}$$

where T_u, T_d are lower diagonal and Q_{scc} is the strongly connected component (SCC). The spectrum of Q is the union of the diagonal entries of T_u, T_d and the spectrum of Q_{scc}. If $\mathcal{R}$ is one diagonal element of T_u, T_d this is trivial: it means that some peripheral communities have the largest reproductive number. Assume instead that $\mathcal{R}$ belongs to the spectrum of Q_{scc}. Then we can write the Perron eigenvector as follows:

$$v = \left(\begin{array}{c} v_u \\ \hline v_{scc} \\ \hline v_d \end{array} \right) . \tag{2.25}$$

If we write by blocks the eigenvector equation $Qv = \mathcal{R}v$, on the top block we have $T_u v_u = \mathcal{R} v_u$, whose only solution is $v_u = 0$ as $\mathcal{R}$ is not an eigenvalue of T_u. This means $Q_{scc} v_{scc} = \mathcal{R} v_{scc}$, and $v_d = (\mathcal{R} - T_d)^{-1} B_3 v_{scc}$. The dynamics is thus completely determined by the SCC, and we can study the SCC isolated, recovering the full force of the Perron-Frobenius theorem.

2.5 Summary of analytical results

The analytical analysis carried out expands Diekmann results with a description of the out-of-equilibrium phase and of locally observed reproductive numbers behaviour. Out-of-equilibrium the measured reproductive number ($\mathcal{S}$) is different from the true reproductive number ($\mathcal{R}$) and each patch may measure a different reproductive number ($s_i \neq \mathcal{S} \neq \mathcal{R}$). It was found that during the convergence process oscillations of $\mathcal{S}$ around $\mathcal{R}$ may occur if Q is strongly coupled, meaning that $\mathcal{S} - \mathcal{R}$ may change sign. The exact formulas for the time evolution of weights g and h_μ in the linear approximation were found, see Eq. (2.14) and Eq. (2.15). The expression for the time evolution of Δ was given, see Eq. (2.12), as well as the one for the time evolution of the s_i, see Eq. (2.22). The role of the second largest eigenvalue Λ_1 in determining the convergence time was highlighted.

At equilibrium, instead, every *h*-class measures the same reproductive number, which also coincides with the true dynamical one.

Chapter 3

Data reconstruction of *Q*

We proceed reconstructing the reproduction operator Q as in Eq. (2.2), in the case where h-classes are spatial patches. In particular, we choose to focus on France and Italy, and we take departments and provinces respectively as spatial patches. We choose Colocation Maps [16], which are spatial network datasets that have been developed within Facebook's Data For Good program, as the main data to obtain the expected infectivity A_{rs} of an infected in patch s towards one in patch r. We then correct these data using Movement Range Maps [32] and discuss the obtained reproduction operators in France and in Italy.

3.1 Meta colocation Maps

Colocation Maps are available as weekly data starting from March 2020 for several countries in the world. Hence, information concerning the major changes that took place in mobility patterns during the last two years due to COVID-19 response measures can be obtained from these data. The weekly resolution was chosen by developers because it is believed that such a time period is well representative of current time human mobility. Colocation Maps are elaborated from localization data of people who use Facebook on a mobile device and who opt in to Location History (LH) and Background Location collection (BC). As a consequence, the tracked population does not correspond with the actual one. In some cases, we can imagine that this reduced population could be not representative of the actual one. The most influential source of bias could be spatial heterogeneity in income in developing countries, where mobile phones are not yet spread in the whole population and the fraction of Facebook users in the population is small. On the other hand, we believe this is not the case for France and Italy. Also note that Colocation Maps have already been used for epidemiological studies in these countries [28].

These Maps estimate how often people from different regions are colocated, i.e., simultaneously located in the same place. These data were appositely designed for epidemiological *metapopulation models*, which are models in which groups of spatially separated individuals interact with each other [24]. The level of coupling between these sub-populations clearly depends on how much they come into contact that is sufficient to transmit the considered disease. Epidemiologists have been parametrizing these couplings indirectly, using for example counts of individuals moving from one region to another [25, 26, 27]. Colocation Maps are instead non-localized data (we do not know where people from two regions meet) that exactly express the contact probabilities determining the couplings.

More precisely, Meta Colocation Maps give the probability that a randomly chosen person from region r and a randomly chosen person from region s are both located in the same $600m * 600m$ square during a randomly chosen five-minutes time bin in a given week. This probability can be computed in few steps. Take first X_{ijr} the number of people from region r who are located in space tile i (a $600m * 600m$ square) during time bin j. We can obtain the number of colocations of people from regions r and s as:

$$m_{rs} = \sum_{ij} X_{ijr} X_{ijs}, \tag{3.1}$$

with the sum over i iterating over the whole set of space tiles in the country and the one over j scanning all the 2016 five-minutes time bins in a week. Note that for $r = s$:

$$m_{rr} = \sum_{ij} X_{ijr}(X_{ijr} - 1) \tag{3.2}$$

to avoid counting a user as colocated with themselves. The probability of a colocation of people from regions r and s is given by the ratio between m_{rs} and the total number of possible colocations:

$$p_{rs} = \frac{1}{n.\ time\ bins} \frac{m_{rs}}{n_r n_s} \tag{3.3}$$

with n_r and n_s the total number of people tracked from regions r and s and *n. time bins* $= 2016$ the number of time bins in a week. In the case $r = s$:

$$p_{rr} = \frac{1}{n.\ time\ bins} \frac{m_{rr}}{n_r(n_r - 1)}. \tag{3.4}$$

We now make an additional approximation. Let $m_{rs,k}$ be the number of contacts between people of patches r and s in the time bin k of duration Δt. We assume that $m_{rs,k} \simeq m_{rs}\Delta t/week\ duration$, implying $m_{rs,k} \simeq p_{rs} n_r n_s (n.\ time\ bins)(\Delta t/week\ duration)$. What this assumption basically implies is that p_{rs} is a time invariant and time-bin duration independent measure of the coupling between two sub-populations, ranging from 0 to 1. It is, in other words, a correctly formulated measure of the between-communities time integrated expected infectivity A_{rs}, except for some constants.

To give some intuition, the SIR set of dynamical equations describing the evolution of subpopulation r can be written in terms of colocation data as:

$$\begin{aligned} &\tfrac{\overline{S}_r}{dt} = -\beta \overline{S}_r \textstyle\sum_s p_{rs} \overline{I}_s \\ &\tfrac{\overline{I}_r}{dt} = -\mu \overline{I}_r + \beta \overline{S}_r \textstyle\sum_s p_{rs} \overline{I}_s \\ &\tfrac{\overline{R}_r}{dt} = \mu \overline{I}_r. \end{aligned} \tag{3.5}$$

Here $\overline{X}_r$ is the expected number of individuals assigned to patch r that are in state X. β is the constant rate at which infection is spread from an infected to susceptible individual while two individuals are colocated, and μ is the constant rate at which a person recovers. The - symmetrical - expected infectivity between two sub-populations can be expressed in the above case of a SIR model as $A_{rs} = \beta p_{rs}/\mu$. Then, the average number of cases generated in r by a case in s assuming a completely susceptible population in patch r is given by:

$$Q_{rs} = \beta p_{rs} n_r / \mu. \tag{3.6}$$

Note that n_r is here the actual population of patch r and not the tracked one. In general, depending on the compartmental model we choose,

$$Q_{rs} = C p_{rs} n_r, \tag{3.7}$$

with C some constant. Until specified, not to introduce new notation, we will refer to Q as the matrix with entries $Q_{rs} = p_{rs} n_r$, with $C = 1$.

3.2 *Stay Put* correction

The large spatial resolution of Colocation Maps was chosen in Facebook's Data for Good program due to privacy concerns. Colocation of two people in a $600m * 600m$ tile shouldn't be considered as sufficient to transmit COVID-19 disease. However, similarly with what already done in the case of time bins, we can suppose that colocation rates in smaller tiles would scale by a constant - the ratio between the area of the small tile and the area of the large tile. p_{rs} are within this assumption space invariant measures, which we can consider to be a proxy for detailed face-to-face proximity rates, even though some features of face-to-face networks are difficult to reproduce [29].

Again due to privacy concerns, too small colocation probabilities were omitted not to allow for people's identification in less populated patches. Even in this case, no major problems arise. The smallest colocation probabilities are not influential in shaping a country's mobility, so that we can safely ignore them when setting couplings of a metapopulation model.

What should instead be a matter of concern is that colocation probability p_{rr} between individuals belonging to the same sub-population may be overestimated due to housing density. People staying at home cannot actually infect nor be infected by people outside their place. We expect this to mainly affect diagonal - within a single patch - colocations. We think in particular to the case of cities with tall residential buildings and high population density in general, where all residents of a neighbourhood of buildings contribute to the colocation probability. We introduce for this reason a correction to diagonal colocation probabilities. We use for this purpose Stay Put data [32], which are also part of Facebook's Data for Good program. Stay Put data or Movement Range Maps give the fraction of inhabitants of a *region* that do not leave a given $600m * 600m$ tile for the whole day. To be precise, data points include observations from 8 pm to 7:59 pm of the next day in local time, in order to include a full night and a full day. Regions do not correspond with patches in Colocation Maps. Both for France and Italy, they are the actual administrative regions.

In order to compute a correction to the reproduction operator based on Movement Range Maps, we first average Stay Put values over one week periods, in order to match Colocation Maps time resolution. We then assume Stay Put values to be equal in all patches in a same region at the same date. The number of Stay Put colocations in a given week between individuals of patch r is given by all the allowed couples of people remaining home in each tile times the number of tiles in that patch:

$$m_{rr}^{(s)} = (sp_r d_r A)(sp_r d_r A - 1) * \ n.\ of\ tiles\ in\ r, \tag{3.8}$$

with sp_r the average Stay Put fraction in patch r in the considered week, d_r the population density in patch r (population divided by surface area of the entire patch) and $A = 0.36\ km^2$ the area of a single tile. Then:

$$p_{rr}^{(s)} = \frac{m_{rr}^{(s)}}{n_r(n_r - 1)}, \qquad Q_{rr}^{(s)} = p_{rr}^{(s)} n_r = \frac{m_{rr}^{(s)}}{(n_r - 1)}. \tag{3.9}$$

In conclusion

$$Q_{rr}^{(s)} = \frac{(sp_r d_r A)(sp_r d_r A - 1) * \ n.\ of\ tiles\ in\ r}{n_r - 1} \tag{3.10}$$

is the value we subtract to the previously obtained Q_{rr} entries to exclude home staying colocations.

3.3 COVID-19 reproduction operators in France and Italy

We now discuss the obtained reproduction operators for COVID-19 spreading in French departments and Italian provinces in 2020 and 2021 and their properties. In Fig. 3.1 we give a graph representation of matrices Q (with $C = 1$) in French and in Italy, with nodes representing patches and edges representing couplings. Two distinct periods are compared. We take first the one ranging from 2020-07-21 to 2020-08-17, that we identify as a typical holidays period. In both countries, no restrictions to mobility were applied at this time. The second time interval chosen is the one ranging from 2020-11-20 to 2020-12-07, during which France was under its second lockdown and Italy was applying regionally specific restrictions to mobility based on evaluated risk in each of them (yellow, orange and red areas existed as an index of increasing risk). In order to allow for a better understanding of figures, an undirected graph was plotted even though the reproduction operator is not a symmetric one. The plotted weight of the edge (i, j) corresponds to $\max\{Q_{ij}, Q_{ji}\}$, $\forall\ (i, j)$ belonging to the set of edges. In addition, only edges with weight greater than an arbitrary threshold 0.25 were plotted. The radiuses of nodes are proportional to the corresponding diagonal elements of Q.

Fig. 3.1 gives us a good qualitative understanding of the mobility network structure in France and Italy as well as of the major and abrupt - only two months separate the considered time ranges - mobility changes that took place in the last years due to COVID-19 response measures. It is evident that French network structure is heavily centralized, presenting a star like structure with Paris department as central node, together with the neighbouring ones in Ile de France. Observing the second lockdown in France, Paris is the only department maintaining significant connections even with non neighbouring departments. We thus expect Paris to be both the department importing the most cases and the most important exporting basin. If we also consider the very high value of the within-Paris reproduction operator element due to the huge population density in the city, we can predict that Paris will have a dominant role in the simulations we will develop starting from the obtained operators.

The Italian mobility network is less centralized, with multiple cities playing an important role. Palermo is typically the province having the highest within-patch reproductive operator element while Milano and Rome are usually the ones that imports the more cases (highest $\sum_{j\neq i} R_{ji}$).

As expected, the reproduction operator Q is *weakly coupled* both in France and Italy during the whole period of availability of data, according to the definition in Eq. (2.19). In terms of the convergence process we want to observe, this means we shouldn't observe oscillations of S around $\mathcal{R}$. As a second point, we

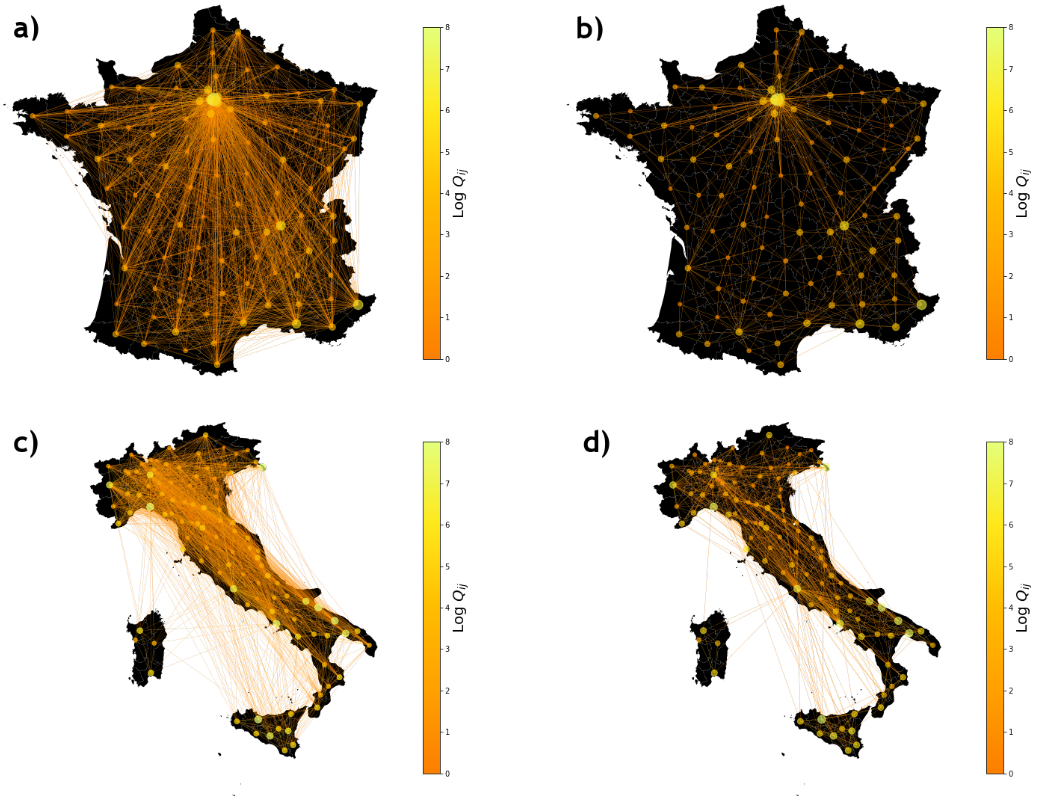

Figure 3.1: *Undirected graph representations of the average reproduction operator Q (with $C = 1$, see Eq. (3.7)) in France (**a** and **b**) and Italy (**c** and **d**) during the periods from 2020-07-21 to 2020-08-17 (**a** and **c**) and from 2020-11-20 to 2020-12-07 (**b** and **d**). The weight associated to the edge (i, j) is $\max\{Q_{ij}, Q_{ji}\}$, $\forall\ (i, j)\ \in E$, with E the set of edges. The radius of node i is proportional to Q_{ii} $\forall i$. Only edges with weight greater than an arbitrary threshold were plotted, and logarithmic values of the operators' elements are considered. The same color scale is kept for all plots. The major variations in mobility due to COVID-19 response restrictions are evident, as well as the dissimilarities between the two countries' network structures.*

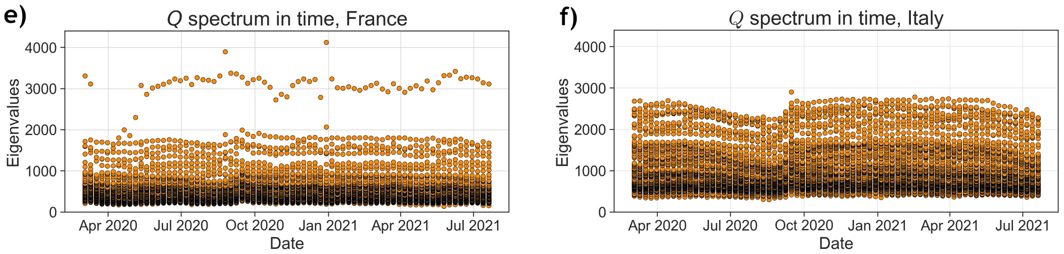

Figure 3.2: *Spectra of reproduction operators Q (with $C = 1$, see Eq. (3.7)) in France (**e**) and Italy (**f**) over time in the period ranging from 2020-03-03 to 2021-07-20. The spectrum in Italy is typically more compact than in France, with in particular a higher ratio $\Lambda_1/\mathcal{R}$. On the other hand, $\mathcal{R}$ is typically larger in France. These are consequences of the network structures depicted in Fig. 3.1 and in particular of the network centrality of the node associated to the department of Paris in France. Seasonal effects are visible for both countries, even though they have a larger impact in Italy. The effect of the first lockdown, that lasted from March 16 to May 11, is clearly visible in the case of France.*

expect convergence times to be different in the two countries because of the major dissimilarities between the spectra of the reproductive operators reported in Fig. 3.2. More precisely, the first eigenvalue $\mathcal{R} = r(Q)$ is typically almost two times the second eigenvalue Λ_1 in France during the observed period. As discussed, this should imply a fast to equilibrium convergence time. In Italy, on the other hand, the spectrum is more compact over the all observed period, with little difference between the first and the second eigenvalue. A longer convergence time of $\mathcal{S}$ to $\mathcal{R}$ is expected. On the other hand, a synthetic epidemic in France should have a faster outbreak than in Italy given the same model parameters due to the generally higher value of the Perron eigenvalue of Q.

Seasonal effects can be observed on the spectrum of the Q operator in Italy. Starting from May, inter-provinces mobility increases, so that the Q matrix becomes less and less diagonal. The matrix is the least diagonal in August. The spectrum consequently becomes more compact, with in particular the highest eigenvalues relevantly decreasing. In September, inter-provinces mobility decreases and the highest eigenvalues grow. The same pattern was observed analysing the reproduction operators in Spain, Portugal and Sweden. For what concerns France instead, the effect of the first lockdown, that lasted from March 16 to May 11, is clearly visible, with the Perron eigenvalue almost halved.

As a last remark, we should discuss the properties of the Perron eigenvectors associated to these operators. As already stressed, the obtained matrices are weakly coupled, with within-patch entries typically 10^3 and up to 10^4 times larger than non diagonal entries. It is actually reasonable that the couplings between inhabitants of a same patch are some orders of magnitude larger than those between inhabitants of different patches. Consider now the case of a fully diagonal matrix D, call Λ_μ its eigenvalues and w_μ the corresponding eigenvectors, with normalisation $\sum_i w_{\mu,i} = 1 \;\; \forall \mu$. Λ_μ will correspond to some diagonal entry, say $\Lambda_\mu = D_{i,i}$ and $w_{\mu,i} = 1$, $w_{\mu,j} = 0 \;\; \forall j \neq i$. This extreme case gives us an intuition about the entries of the Perron eigenvectors associated to the reproduction operators of France and Italy. For all the reconstructed reproduction operators, the Perron eigenvector has one entry $v_k > 0.6$ (typically $v_k > 0.9$), while all other entries are of order $\sim 10^{-2}$ or $\sim 10^{-3}$. Thinking about the epidemic process, we can say that in the long term and in the linear approximation (the fraction of susceptibles stays approximately constant) the exponential growth of the disease will cause the majority of the cases to be located in the patch at highest risk. In terms of the reproductive number, we have $\mathcal{R} \simeq Q_{k,k} = \max_{r,s} Q_{r,s}$. This means that the within-patch highest entry of the reproduction operator will cause the reproduction number in all other patches to grow and eventually reach $\mathcal{R}$ in the long time limit. The specific characteristic times of this process are the matter of study of this work.

Chapter 4

Metapopulation model

In this chapter we test our analytical findings and expand our analysis through a metapopulation model informed on census data in France and Italy, as well as on the discussed reproduction operators. After defining the model and computing some useful quantities, we conduct some experiments in order to analyse the dependence of the dynamical processes of interest on initial conditions, on the network topology and on the transmission rate. Lastly, we study the impact that discontinuities in time of the mobility network have on the dynamical processes.

4.1 Model description

As mentioned, a metapopulation model is one where a population is considered as divided in sub-groups of preferential interaction [24, 30, 31]. We build a stochastic discrete time metapopulation model with a synthetic population based on the National Institute of Statistics and Economic Studies (INSEE) censuses for French departments and the Istituto Nazionale di Statistica (ISTAT) censuses for Italian provinces. Colocation Maps [16], corrected with Movement Range Maps [32], are taken as a measure of the coupling between sub-groups. We choose a compartmental model of the Susceptible-Exposed-Infected-Recovered (SEIR) kind [33]. In addition, we consider two different infectious compartments, namely clinical and sub-clinical infectious. Sub-clinical infectious are less likely to transmit the disease when in contact with another individual. This model is a simple yet effective representation of COVID-19 infection [34]. A scheme of the compartmental model is shown in Fig. 4.1.

The time T spent in each compartment is exponentially distributed according to the rates reported in Fig. 4.1. Exploiting the exponential distribution memorylessness property $Pr(T > s + t|T > s) = Pr(T > t)$, at each time step of duration

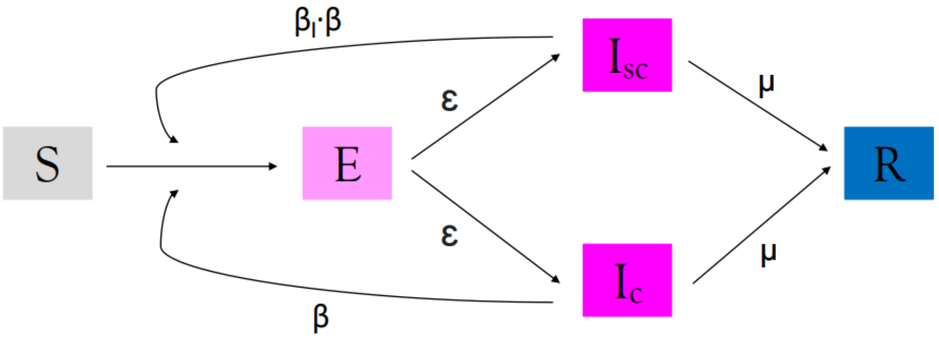

Figure 4.1: *The chosen compartmental model. Five different compartments are present, with transition times from one to another exponentially distributed according to the rates reported on arrows.*

$\Delta t = 1day$, the probability associated to a transition characterised by rate λ is:

$$p_\lambda = Pr_\lambda(T < \Delta t) = 1 - e^{-\lambda \Delta t}. \tag{4.1}$$

In particular, $\epsilon = (3.7days)^{-1}$ is the rate of becoming infectious for exposed individuals, $\mu = (9.1days)^{-1}$ is the recovery rate, $p_{sc} = 0.32$ is the probability of being assigned to the sub-clinical compartment once exposed, β is the transmission rate for infected individuals. The transmissibility of subclinical cases is rescaled by a factor $\beta_I = 0.51$. Values of the parameters ϵ and β_I are taken from Faucher *et al.* [34], and the average recovery time μ^{-1} is obtained as the sum of the average duration of the pre-symptomatic period $(2.1days)$ [34] and the average recovery time for infectious symptomatic individuals $(7days)$ [34]. p_{sc} was computed as the average over age classes of the probabilities reported by Davies [35]. Parameter β is a free parameter, and will be explored.

We track the number of individuals in each compartment over time for all spatial patches. At each time step, we extract from a binomial distribution the number of individuals that transfer from a compartment to another. In particular, let X_i be the number of individuals in compartment X, patch i at time t. The number of new recover events at time $t + 1$ in patch i from the infected classes is extracted from:

$$Bin(n = I_{sc/c,i}\ ,\ p = p_\mu). \tag{4.2}$$

The number of new infected individuals of the sc type is extracted from:

$$Bin(n = E_i\ ,\ p = p_\epsilon p_{sc}). \tag{4.3}$$

For the clinical infected class:

$$Bin(n = E_i\ ,\ p = p_\epsilon(1 - p_{sc})). \tag{4.4}$$

Finally, the number of new exposed individuals in patch i is distributed according to

$$Bin(n = S_i \ , \ p = p_\beta \sum_j p_{ji} I_{c,j} + p_{\beta\beta_I} \sum_j p_{ji} I_{sc,j}),$$

which can be rewritten as

$$\begin{aligned} Bin(n = S_i \ , \ p = p_\beta \sum_j p_{ji} n_j I_{c,j}/n_j + \\ p_{\beta\beta_I} \sum_j p_{ji} n_j I_{sc,j}/n_j), \end{aligned} \tag{4.5}$$

with sums running over all nodes of the graph, i.e., over all spatial patches. In Eq. (4.5) we recognize the reproduction operator Q with $C = 1$. We now want to recover the exact form of the operator, constants included, in order to obtain the exact reproduction number for our metapopulation model. We adopt the method presented by Diekmann, Heesterbeek and Roberts [36].

Let X_i be the number of *infected* individuals in the i-th *infected* compartment. Let's rewrite our epidemic model as:

$$\frac{dX_i}{dt} = F_i(X) - V_i(X) = F_i(X) - [V_i^-(X) - V_i^+(X)].$$

$F_i(X)$ represents the rate of appearance of new infections in compartment i. $V_i^+(X)$ represents the rate of transfer of individuals into compartment i by all other means, and $V_i^-(X)$ represents the rate of transfer of individuals out of compartment i. Let X_0 be the disease free equilibrium, defined as the state $(S = n, E = 0, I_{sc} = 0, I_c = 0, R = 0)$, with n the total population vector. We define the square matrices F and V with elements:

$$\begin{aligned} F_{ij} &= \frac{\partial F_i}{\partial X_j}(X_0) \\ V_{ij} &= \frac{\partial V_i}{\partial X_j}(X_0). \end{aligned} \tag{4.6}$$

Now, the matrix FV^{-1} is the next-generation matrix. The largest eigenvalue in modulus or spectral radius of FV^{-1} is the basic reproduction number of the model.

In the present case, using a shortcut, we can write the dynamical differential equations describing the entire population as:

$$
\begin{aligned}
\dot{S} &= -\beta' S I_c/N - \beta' \beta_I S I_{sc}/N \\
\dot{E} &= -\epsilon E + \beta' S I_c/N + \beta' \beta_I S I_{sc}/N \\
\dot{I_{sc}} &= \epsilon p_{sc} E - \mu I_{sc} \\
\dot{I_c} &= \epsilon (1 - p_{sc}) E - \mu I_c \\
\dot{R} &= \mu (I_c + I_{sc}).
\end{aligned}
\tag{4.7}
$$

In the above equations $\beta' = \beta\, r(K)$, with K such that $K_{ij} = p_{ij} n_i$ is the reproduction operator with constants set to one. Then:

$$
\begin{aligned}
R_0 = r(FV^{-1}) &= \frac{\beta'}{\mu}(1 - p_{sc} + \beta_I p_{sc}) \\
&= r(K)\frac{\beta}{\mu}(1 - p_{sc} + \beta_I p_{sc}).
\end{aligned}
\tag{4.8}
$$

In other words,

$$
Q_{ji} = p_{ji} n_j \frac{\beta}{\mu}(1 - p_{sc} + \beta_I p_{sc}) \tag{4.9}
$$

is the exact reproduction operator in our model.

4.2 The R-package EpiEstim and the generation interval for our model

The aim of the synthetic epidemic experiments we will design will be to show the measurement error we commit when estimating the basic reproduction number of an epidemic from surveillance data. In order to make our conclusions more reliable, we will compute the observed reproduction number adopting two different methods. First, we will compute an *analytical* $\mathcal{S}$ according to the already introduced formula $\mathcal{S} = \sum_{ij} Q_{ij} x_j$. Then, we will use the R-package `EpiEstim` [20], which is a widely used tool to estimate $\mathcal{R}$ from surveillance data.

Some concepts now need to be introduced. *Incidence* is defined as the number of new cases of a disease in a given time period over the total population [37]. It should not be confused with *prevalence*, which is the proportion of active cases in a population at a given time [37]. We also introduce the generation interval as the time interval between a primary and a secondary infection, i.e., the time interval between the infection event for a case and for their infector [38]. Equivalently, the serial interval is defined as the time interval between the onset of symptoms in a case and in their infector [38].

EpiEstim estimates the time varying reproduction number $\mathcal{R}$ over time from incidence time series and given the generation interval probability distribution. Incidence over time will be obtained as an output of our simulations, but we should derive the generation interval distribution for our model. Let's first consider the probability distribution for the time of transmission of the disease. The number of susceptible individuals that a case infects is drawn from a binomial distribution with patch dependent probability. Globally, the probability of transmission we name γ is given by:

$$\gamma = \mathcal{R}_0\mu = r(Q)\mu. \tag{4.10}$$

If γ is small, as in our case, this infection process can be approximated by a Poisson process. Then, the number of new infected that a case produces given the duration τ of their infectious period follows:

$$P(I = n|\tau) \simeq Poisson(\gamma\tau) = \frac{(\gamma\tau)^n e^{-\gamma\tau}}{n!}, \tag{4.11}$$

with τ in units of the model time step. The time distance between two consecutive events of such a Poisson process is an exponential random variable X with parameter γ, hence the probability distribution for the time required for an infectious to transmit the disease is:

$$p(x|\tau, trans) = \gamma e^{-\gamma x}, \tag{4.12}$$

where the conditioning $trans$ specifies that this is the probability distribution assuming that transmission occurred (this may not be the case). Now:

$$\begin{aligned} p(x|trans) &= \int_0^\infty d\tau\ p(x|\tau, trans)p(\tau|trans) \\ &= \int_0^\infty d\tau\ p(x|\tau, trans)p(\tau) \\ &= \int_0^\infty d\tau\ \gamma e^{-\gamma x}\mu e^{-\mu\tau} \\ &= \gamma e^{-\gamma x}. \end{aligned} \tag{4.13}$$

Let now $Y \sim \epsilon e^{-\epsilon y}$ be the random variable for the time for an exposed individual to become infected. The probability density function for $Z = X+Y$ can be obtained from a convolution:

$$
\begin{aligned}
p_Z(z|trans) &= \int_0^z dx\ p_X(x|trans)p_Y(z-x) \\
&= \int_0^z dx\ \gamma e^{-\gamma x}\epsilon e^{-\epsilon(z-x)} \\
&= \gamma\epsilon e^{-\epsilon z}\int_0^z dx\ e^{(\epsilon-\gamma)x} \\
&= \frac{\gamma\epsilon}{\epsilon-\gamma}(e^{-\gamma z}-e^{-\epsilon z}),
\end{aligned} \tag{4.14}
$$

where we have introduced subscripts to the pdf for clarity. The obtained $p_Z(z|trans)$ represents the generation interval distribution we were looking for. This distribution is null in zero, asymmetrically bell shaped and has average $1/\epsilon + 1/\gamma$, i.e., the sum of the average time for the two distinct processes.

We specify that a discrete distribution is required by the EpiEstim package. We choose to compute it over 50 bins in the interval $[0, 49]$. The time window over which to estimate $\mathcal{R}$ is set to be a week. In all plots, we assign the estimate for $\mathcal{R}$ returned for the week interval $[t, t+6]$ to the day $t+3$. Also, given the smaller precision of early estimates, as reported in the documentation [20], we arbitrarily choose not to plot EpiEstim points associated to the first two weeks from the start of the synthetic epidemic.

As a last remark, note that our derivation of the generation interval is not exact. Being our metapopulation model a discrete one, transition times between compartments are not exactly exponentially distributed. As an example, consider the case of an individual passing from compartment S to compartment E at time step n, i.e., at some time $t \in [(n-1)\Delta t, n\Delta t]$. The time required to transition to compartment I_c or I_{sc} is systematically shifted by $n\Delta t - t$. We hence expect the EpiEstim estimate for $\mathcal{R}$ curve not to perfectly match our analytical observed reproductive number.

4.3 Exp. I: dependence of the dynamical process on initial conditions

As a first experiment, we want to study the dependence of the dynamical process of convergence of $\mathcal{S}$ to $\mathcal{R}$ - and consequently of the measurement error for $\mathcal{R}$ - on the initial conditions, precisely the initial distribution of cases amongst departments or provinces. We know from Eq. (1.6) that neither the equilibrium distribution nor the equilibrium reproduction number depend on initial conditions, even though the expected number of cases at equilibrium does. On the other hand, from Eq.

(2.14, 2.15, 2.16) we know that the equilibrium convergence time does depend on initial conditions. In particular, the weight h_1 associated to Λ_1 when decomposing the initial conditions vector on the basis of eigenvectors of Q is the most influential one.

We first consider France and we select a reproductive operator Q we will keep fixed. We consider the period ranging from 2021-26-01 to 2021-16-02 and take an average of the available Q. Mobility didn't undergo major changes during this period so that Q is approximately constant. On the other hand, averaging over one month reduces the effect of weekly fluctuations or anomalous entries. We also fix the value of the transmission rate to $\beta = 7.5 \cdot 10^{-5}$. This value was arbitrarily chosen as matching the simple needs to be above the threshold value for the epidemic to spread ($\mathcal{R}_0 > 1$) and to reach the desired order of magnitude of incidence (10^5 cases) at the peak in not too long a time. We stress that our model is a simplified theoretical representation of the spread of a directly-transmitted, airborne pathogen, similar to SARS-CoV-2, as a case study for the analysis of the along time measurement error committed by estimates of $\mathcal{R}$.

Initial cases are drawn from a multinomial distribution with probabilities $p = (p1, ..., p_M)$, with M the number of nodes and $\sum_i p_i = 1$. The p are chosen as reported in Tab. 4.1.

Multinomial distribution p, Exp. I
I.C. 0
$p_{Paris} = 1$, $p_i = 0 \quad \forall i \neq Paris$
I.C. 1
$p_i \propto n_i \quad \forall i$
I.C. 2
$p_{Paris} = 0$, $p_i \propto n_i \quad \forall i \neq Paris$
I.C. 3
$p_j = 0 \quad \forall j \in \mathcal{I}$, $p_i \propto n_i \quad \forall i \notin \mathcal{I}$

Table 4.1: *p vectors of the multinomial distributions from which initial conditions for Exp. I in France are drawn. Note that $\mathcal{I} = \{$Paris, Hauts-de-Seine, Val-de-Marne, Seine-Saint-Denis, Yvelines, Val d'Oise, Essonne, Seine et Marne$\}$ is the set of Île de France departments.*

Initial conditions in Tab. 4.1 were chosen with the aim of varying the number of initial cases in Paris and in Île de France. These departments actually have high coupling with each other and form an enlarged single sub-population leading the epidemic behaviour in France. Paris, in particular, has the highest reproduction operator within-patch entry. Formally, these initial conditions differ for the weight g_0 of the Perron eigenvector and $h_{0,\mu}$ of the eigenvectors associated to the remaining highest eigenvalues. In *I.C. 0* initial cases are all in Paris. In *I.C. 1* the initial vector of infectious individuals is drawn from a multinomial distribution with p proportional to departments populations. The same holds for *I.C. 2* and *I.C. 3*, but initial cases are set to zero in Paris and in all Île de France respectively. In all cases the number of draws is equal to 500, so that the total number of initially infectious individuals is $I_{tot} = 500$. They are assigned to the sub-clinical or clinical class with probability p_{sc}, $1 - p_{sc}$ respectively. The same number of exposed individuals $E_{tot} = 500$ is set. Each simulation lasts 630 time steps and 100 runs were made for each initial conditions. Medians and 95% confidence intervals were computed for each observed quantity.

4.3.1 Exp. I results

Simulations' results for Exp. I are plotted in Fig. 4.2. Looking first at Fig. 4.2, **a**, as intuitively expected, we notice that the fewer initial cases are in the departments at highest risk, the longer is the time before an outbreak occurs. Notably, with the actual choice of parameter β, the epidemic only spreads in Ile de France. More precisely, β is high enough to cause an outbreak in Paris only, due to the large gap between the first and second eigenvalues of Q visible in Fig. 3.2, **e**. Due to the strong coupling between Paris and the other departments in Ile de France, the epidemic then diffuses in the whole region. Even though we could obtain a non localized epidemic by simply raising β, this spectral gap seems unrealistic. Most probably, some further corrections should be applied in order to reduce Q's entry in Paris. Due to its very high population density, we can imagine that many individuals are counted as colocated even though they are not staying in the same place, e.g., they are separated by a wall or they are on different floors of a same building. With approximately 21000 inhabitants/km^2, Paris is indeed the most densely populated city in Europe according to Eurostat. On the other hand, this particular case of a localized synthetic outbreak allows us to observe more clearly some interesting properties of the dynamical process under study.

Consider the process of convergence of the infected fraction x to the Perron eigenvalue v of Q represented in Fig. 4.2, **b**. We choose one minus cosine similarity, defined as $1 - cs(A, B) = 1 - cos(\theta) = 1 - \frac{A \cdot B}{\|A\| \|B\|}$, A and B vectors, as a measure of similarity between x and v. Note that since both x and v have non negative entries only, cosine similarity is bounded in $[0, 1]$, with $1 - cs(x, v) = 0$ if and only if $x = v$.

For *I.C. 0* $cs(x_0, v) \ll 1$ already initially, since the Perron eigenvector of the reproduction operator has $v_{Paris} \gg v_j \quad \forall j \neq Paris$. Looking at Fig. 4.2, **b**, a

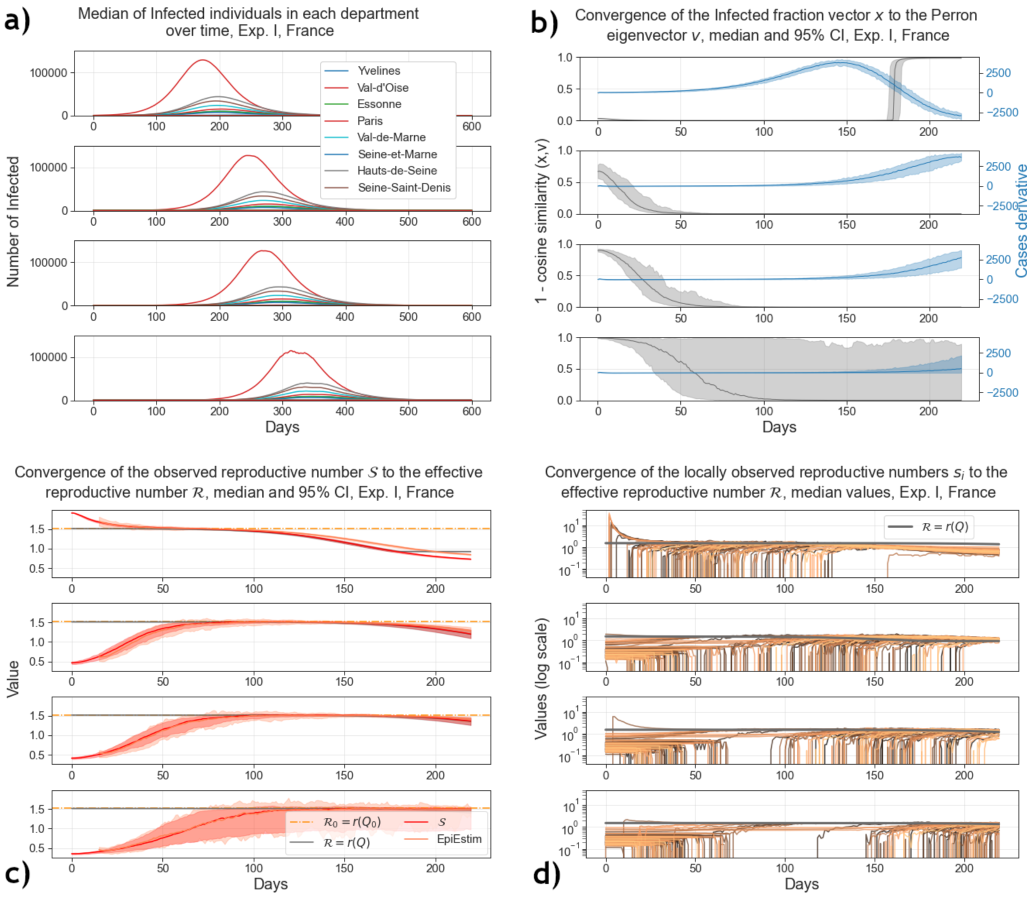

Figure 4.2: *Results of the simulations of Exp. I, with initial conditions in the same order as in Tab 4.1, i.e., for each of* ***a****,* ***b****,* ***c*** *and* ***d****, I.C.s 0, 1, 2, 3 are respectively in the top, second from top, third from top and bottom panels.* ***a*** *Number of infected individuals (prevalence) in time in each department. Only labels for the main departments for number of cases are shown in legend.* ***b*** *Convergence of x to v, with one minus cosine similarity as a metric.* ***c*** *Time evolution of the observed global reproductive number $\mathcal{S}$. The measurement error is evident for both $\mathcal{S}$ and the EpiEstim estimate during the transient period. We remember that EpiEstim estimates relative to the first two weeks since the start of simulations were discarded.* ***d*** *Time evolution of the observed local reproductive numbers s_i, logarithmic scale. In all cases, the chosen initial conditions lead to increasingly slower dynamical processes.*

few time steps before the cases derivative goes to zero the linear approximation expires. Depletion of suceptibles in Île de France and particularly in Paris reduces gradually the corresponding entries of $Q(S)$, until a new eigenvector abruptely establishes as the leading one. $1 - cs(x, v)$ consequently goes to one. In *I.C.s 1, 2, 3* $1 - cs(x_0, v)$ is gradually increasing. The convergence time is also increasing, according to Eq. (2.14). Actually, what makes convergence slower is g_0 decreasing between *I.C.s 1* and *2*. What instead most relevantly differs from *I.C.s 2* and *3* are the values of weights associated to the competing eigenvectors with largest eigenvalues, which have high entries in the indices corresponding to departments at highest risk out of Île de France. More results are shown in Fig. 4.3, where eight different initial conditions are considered and the corresponding convergence time (median and 95% C.I.) is plotted. Apart from the already discussed ones, we consider the uniform case where $p_i = 1/M \quad \forall i$ and the case where $p_j = 0 \quad \forall j \in \mathcal{J}$, $p_i \propto n_i \quad \forall i \notin \mathcal{J}$, with $\mathcal{J} = \{Paris,\ Hauts{-}de{-}Seine,\ Val{-}de{-}Marne,\ Seine{-}Saint - Denis\}$ the set of departments in the inner ring in Île de France. We also take the case where p for the multinomial distribution are drawn from a Dirichlet distribution of the form:

$$p(x) = \frac{1}{B(\alpha)} \prod_{i=1}^{M} x_i^{\alpha_i - 1}, \tag{4.15}$$

with α the vector parameter of the distribution, M the number of categories (here the number of provinces) and $B(\alpha)$ the normalization constant. The support for Dirichlet distribution is $[0, 1]$ and $\sum_i x_i = 1$, which allows us to correctly pick a set of probabilities. We choose $\alpha_i = n_i \ \forall i$ (condition *b* in Fig. 4.3) and $\alpha_i = n_i/N \ \forall i$ (condition *h* in Fig. 4.3). We get $E[X_i] = n_i/N \ \forall i$ for both *I.C.s*, while $Var_h[X_i] \simeq N \cdot Var_b[X_i]$, with $Var_{h/b}[X_i]$ the variance in *I.C. b/h*. In Fig. 4.3, initial conditions are ordered by increasing convergence time, except for the case of the high variance Dirichlet distribution, plotted as last point.

Fig. 4.2, **c**, shows the convergence of $\mathcal{S}$ to $\mathcal{R}$. Most importantly, our prediction is confirmed: for all initial conditions, both estimates - the analytical $\mathcal{S}$ and the EpiEstim estimate - either underestimate or overestimate $\mathcal{R}$ during the transient period leading to equilibrium. The duration of this measurement error depends on initial conditions. Also, as expected, the observed reproductive numbers measured analytically and with EpiEstim are almost equal, even though differences come up in the long time due to the mentioned approximations concerning the epidemic's generation interval. We remember that EpiEstim estimates related to the first two weeks since the start of the simulations are not plotted.

I.C. 0 represents the only case in which $\mathcal{S} > \mathcal{R}$ initially. Then, $\mathcal{S}$ rapidly converges to $\mathcal{R}$. After ~ 100 timesteps $\mathcal{R}$ starts decreasing due to depletion of susceptibles and departs from $\mathcal{R}_0$, with $\mathcal{S}$ following. The epidemic is actually only developing in Ile de France, causing the corresponding Q entries only to decrease

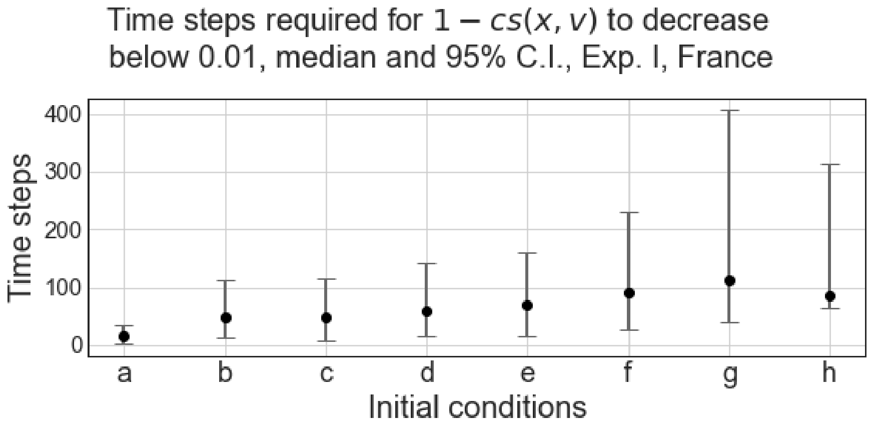

Figure 4.3: *Time steps required for the convergence of x to v with eight initial conditions. Here: a) $p_{Paris} = 1$, $p_i = 0 \quad \forall i \neq Paris$; b) $p \sim Dir(\alpha_i = n_i \ \forall i)$; c) $p_i \propto n_i \quad \forall i$; d) $p_i = 1/M \quad \forall i$; e) $p_{Paris} = 1$, $p_i = 0 \quad \forall i \neq Paris$; f) $p_j = 0 \quad \forall j \in \mathcal{J}$, $p_i \propto n_i \quad \forall i \notin \mathcal{J}$, with $\mathcal{J} = \{Paris, \ Hauts-de-Seine, \ Val-de-Marne, \ Seine-Saint-Denis\}$; g) $p_j = 0 \quad \forall j \in \mathcal{J}$, $p_i \propto n_i \quad \forall i \notin \mathcal{I}$, with $\mathcal{I}$ corresponding to Île de France; h) $p \sim Dir(\alpha_i = n_i/N \ \forall i)$.*

until a new Perron eigenvalue establishes. We will now have $\mathcal{R} = r(Q(S)) = \Lambda_1$. We may say we have some sort of phase transition, with the first derivative of $\mathcal{R}$ showing a discontinuity. At the discontinuity point both $\mathcal{S}$ departs from $\mathcal{R}$ and the Perron eigenvector changes establishing a new equilibrium. We could now assume to be in linear approximation again and on a depleted network, since the epidemic has consumed the susceptibles in some of the network nodes, excluding them from the next incoming virus diffusion. The dynamical process will repeat itself, eventually departing from the new equilibrium and possibly setting another one.

Concerning *I.C. 1, 2* and *3*, we see that the convergence time progressively increases. Consequently, estimates of $\mathcal{R}$ are not able to reproduce the correct value for longer times. For these *I.C.*s, and in general for all initial conditions tested apart from *I.C. 0*, $\mathcal{S} < \mathcal{R}$ during the out-of-equilibrium period.

Consider now Fig. 4.2, **d**. As expected, the locally observed reproductive numbers converge to $\mathcal{R}$ even though an exponential outbreak only took place in Ile de France. Exportation of cases ensures that all the s_i reach the equilibrium value. However, they all have different *I.C.*-sensitive convergence times, so that there exist some *I.C.* such that some of the s_i do not reach $\mathcal{R}$ before the linear approximation falls.

Exp. I helps us get an understanding of the dynamical process. The expected behaviours are observed, as well as the expected dependence on *I.C.s*. As also clear from the obtained Eq. (2.14, 2.16, 2.22), the dynamics of x, $\mathcal{S}$ and s_i all have different characteristic times.

4.4 Exp. II: dependence on the network topology and transmission rate

We have seen in Exp. I the simplified case of an epidemic spreading in a strongly connected set of sub-populations of the entire network, as an effect of Q's large spectral gap. We replicate Exp. I in the case of the Italian population and reproducing the Italian mobility network, in order to observe the impact of the network topology on the dynamical process. The parameters of the model do not change with respect to Exp. I. Again, 100 simulations are run for each *I.C.*, each of them lasting 630 timesteps. Medians and 95% confidence intervals are computed. For brevity, we show in Fig. 4.4 the results obtained for two *I.C.s* only. Namely, we draw the initial number of cases in each province from a multinomial distribution with parameter vector p. Components p_i are in turn drawn from a Dirichlet distribution with $\alpha_i = n_i \; \forall i$ for *I.C. 0* and $\alpha_i = n_i/N \; \forall i$ for *I.C. 1*. Again, we get $E[X_i] = n_i/N \; \forall i$ for both *I.C.s*, while $Var_1[X_i] \simeq N \cdot Var_0[X_i]$, with $Var_{0/1}[X_i]$ the variance in *I.C. 0/1*.

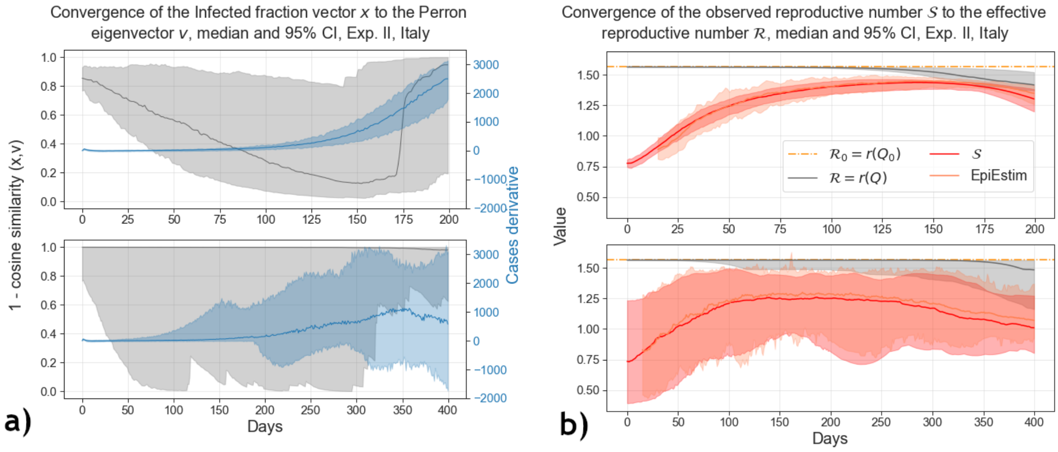

Figure 4.4: *Results of the simulations of Exp. II, with both in **a** and **b** I.C. 0 in the top panel and I.C. 1 in the bottom panel. **a** Convergence of x to v, with one minus cosine similarity as a metric. **b** Time evolution of the observed systemic S. Convergence to the equilibrium quantities is not observed in any case, with in particular $\mathcal{S}$ and EpiEstim always underestimating $\mathcal{R}$.*

These two initial conditions alone give us the possibility to explore the basic case of an initial infectious distribution almost exactly proportional to the provinces' population, together with a wide range of variations of this basic case. Most of the provinces involved in the metapopulation model observe an outbreak in this experiment, due to the more compact Q spectrum. Notably, Fig. 4.4 shows that equilibrium cannot be reached starting from a population-proportional initial distribution, given the chosen β and the reconstructed mobility network for Italy. For none of the runs in the 95% confidence interval we could observe $x \simeq v$, nor $\mathcal{S} \simeq \mathcal{R}$. Rephrasing, in all the plotted runs the time for the linear approximation to vanish was shorter than the convergence to equilibrium time of x or $\mathcal{S}$. Most interestingly, the same happens for *I.C. 1*. The confidence interval of $1 - cs(x, v)$ does not reach zero and the confidence interval of $\mathcal{S}$ does not reach $\mathcal{R}$. This tells us that whatever initial condition drawn, v should be considered as a fictitious limit and $\mathcal{R}$ as an upper bound for the observed global reproductive number. Also, the median of none or only few of the local s_i is able to reach $\mathcal{R}$ in the linear approximation interval. One exception exists, even though not represented in the 95% confidence interval. Starting the epidemic with cases only in the province with the highest within-patch Q entry brings to a situation similar to the one seen for the first initial condition of Exp. I. However, no other exception was found, meaning that in all other cases estimates of $\mathcal{R}$ never match the true value.

We should now discuss the dependence of the dynamical process on the transmission rate β. From Eq. (2.16) we know that the convergence of $\mathcal{S}$ to $\mathcal{R}$ is faster for higher β. Precisely, given the dependence of R_0 on β (Eq. (4.8)), we have that Δ_t grows linearly with β in the linear approximation regime. On the other hand, we expect that the fraction of recovered individuals grows faster when β is higher (assuming $\beta > 1$), so that the linear approximation validity time will be shorter. In particular, the number of recoveries grows as $\sim \beta^t$ with the transmission rate. In conclusion, as β grows the linear approximation validity time decreases faster than the convergence time of $\mathcal{S}$ to $\mathcal{R}$.

We want to explore parameter β to detect for which values $\Delta_t \simeq 0$ for some t. Following the above reasoning, we should consider values of β smaller than the one so far used in Exp. II. We test values of β equal or greater than the threshold value such that $\mathcal{R}_0 = 1$, i.e., $4.8 \cdot 10^{-5}$, and spaced $0.1 \cdot 10^{-5}$ with each other. We see that for $\beta \geq 5.2 \cdot 10^{-5}$ the observed reproductive number never reaches the true one. However, below this value an outbreak only occurs in the province with the highest within-patch entry (Palermo). We may conclude that given the Italian network structure reconstructed from Meta Colocation Maps, $\mathcal{R}$ estimates can never reach the true value in the linear approximation regime if the epidemic spreads in a minimum of two provinces.

4.5 Exp. III: impact of discontinuities in time of the mobility network

In this third experiment we evaluate the effect of introducing a discontinuity in the mobility pattern, i.e., in the coupling between subpopulations and in the within-patch mixing. We design transitions that mimic the beginning or the release of a national lockdown, with mobility abruptly significantly reduced or incremented starting from timestep (day) t^*. We choose France as a case study again, $\beta = 7.5 \cdot 10^{-5}$ as in Exp. I, and a single initial condition. We draw 500 initial cases from a multinomial distribution with $p_i \propto n_i \, \forall i$. The reason for these choices is the following: we found that equilibrium is reached within the linear approximation regime simulating an epidemic in France with this β and starting with a population-proportional initial condition. Hence, this setting enables us to consider the interesting cases of t^* being smaller and larger than the $\mathcal{S}$ to $\mathcal{R}$ convergence time. The aim is to evaluate how the introduction or release of a lockdown affects the dynamical process of convergence to equilibrium, and how these measures perturb equilibrium once reached, if they can.

Specifically, we choose the two averaged reproduction operators of Fig. 3.1, **a** and **b**. The first one is the average of the available Q in the period from 2020-07-21 to 2020-08-17, that we take as an example of intensive mobility, with no restrictions. The second time interval chosen is the one ranging from 2020-11-20 to 2020-12-07, during which France was under its second lockdown. We design, using these two operators, four different transitions. Given that the time needed to reach equilibrium is $t_{eq} \sim 80 days$, we refer to *a* as a transition from non restricted to restricted mobility at time $t^* = 30 < t_{eq}$; *b* is a transition from non restricted to restricted mobility at time $t^* = 100 > t_{eq}$; *c* is a transition from restricted to non restricted mobility at time $t^* = 30 < t_{eq}$; *d* is a transition from restricted to non restricted mobility at time $t^* = 100 > t_{eq}$. Results for *a, b, c* and *d* are reported in Fig. 4.5 **a**, **b**, **c** and **d** respectively.

First, we see that as the reproduction operator Q shaping mobility changes, $\mathcal{R} = r(Q)$ adjusts to a new value, lower than the initial one in the case a lockdown is starting, higher than the initial one when a lockdown is released. We should analyse case by case how does $\mathcal{S}$ behave in response.

In *a* restrictions end before equilibrium is reached. The analytical $\mathcal{S}$ sharply decreases. If we were observing this reproductive number, we would be able to see an immediate response to the introduction of mobility restrictions. Yet, $\mathcal{S}$ decreases less than $\mathcal{R}$ does, and immediately starts growing again. The EpiEstim estimate, which is computed over a weekly window, is more smooth. The corresponding curve stays approximately constant just after the transition, then starts growing again. In conclusion, looking at measures of $\mathcal{R}$ inferred from cases we would be

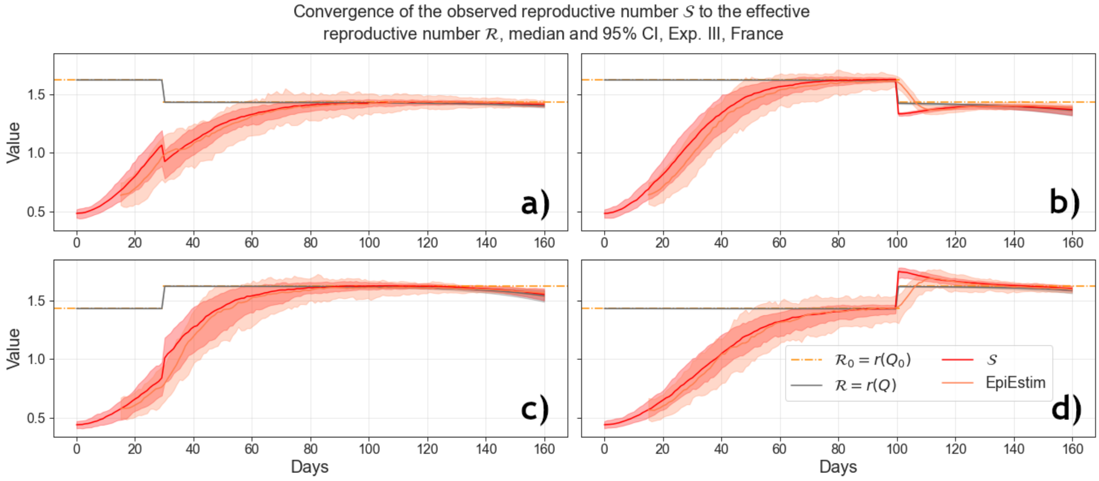

Figure 4.5: *Results of the simulations of Exp. III to evaluate the effect of reproduction operators' discontinuities in time.* ***a*** *From non restricted to restricted mobility at* $t^* = 30$. ***b*** *From non restricted to restricted mobility at* $t^* = 100$. ***a*** *From restricted to non restricted mobility at* $t^* = 30$. ***a*** *From restricted to non restricted mobility at* $t^* = 100$.

barely able to detect some changes, and we would see the reproduction number grow again in a few days. The lockdown measure would seem ineffective, even though the true $\mathcal{R}$ of the system has lowered, as only visible after convergence to equilibrium.

Consider now b, with the lockdown being introduced at $t^* > t_{eq}$. We may say the opposite effect is observed: the impact of the restrictions is initially overestimated, with both $\mathcal{S}$ and the EpiEstim curve going below $\mathcal{R}$, then growing up to the equilibrium value. Also note that this discontinuity in the mobility pattern temporarily drives the system out of equilibrium again.

The case of the release of a lockdown is interesting as well. In *c* the EpiEstim curve is similar to the one we would obtain if no discontinuity was introduced, but the target $\mathcal{R}$ is significantly different. In *d* the system is temporarily driven out of equilibrium and observed reproduction numbers initially overestimate the impact of the restrictions release.

To sum up, not only in none of the analysed cases the two estimates are a good indicator of how changes in mobility patterns influence the reproduction number, but also they may be misleading. The most remarkable case is a, where we see the EpiEstim estimate grow even though $\mathcal{R}$ decreased. In b and d we overestimate the discontinuity impact and in c we are not able to detect any changes.

Chapter 5

COVID-19 data analysis

In this chapter we want to check whether the convergence of the epidemic dynamics to the equilibrium dynamics described analytically and in simulations is visible in real surveillance data, using daily confirmed COVID-19 data. Clearly, identifying in COVID-19 data the signal of the processes that we theoretically analyzed is hard, because there are competing factors that influence incidence of reported cases. Actually, multiple changes that took place in the reproductive operator are not accounted for in our model (e.g.,social distancing, masks, behavioral factors not detectable with colocation data and others [[41]]). Also, detection rate was varying in space and time during the COVID-19 pandemic, and it is common to reconstruct cases from, for instance, hospitalizations, as we will do. Nevertheless, some interesting results may be obtained.

5.1 Data and methods

We consider the usual period from 2020-03-03 to 2021-07-20, during which we have computed reproduction operators for France and Italy at weekly resolution. Let Q_t be these operators, with index t representing the week they refer to. We are then able to compute the Perron eigenvectors v_t of these matrices, that we know to represent the long term equilibrium distribution of cases amongst spatial patches. The idea is to compare the v_t to the COVID-19 cases distribution in French departments and Italian provinces, week by week during our observation period.

We collect data for COVID-19 cases in Italy by province from the dataset of the Italian Civil Protection Department *pcm-dpc/COVID-19*. In this dataset, the Italian Civil Protection Department reports the daily number of currently positive individuals as the sum of hospitalized patients and home-confinement patients [42]. More work is done to obtain positive cases in France. In this case, the French government dataset *data.gouv.fr* only reports incidence of COVID-19 cases (and not prevalence) from May 2020 (later than our desired start date). For what concerns the first issue, considering incidence instead of prevalence should not give major differences in terms of cases distribution amongst departments or, at least, convergence towards the equilibrium eigenvector should be observed anyways. For what concerns the starting date, we decide to solve the issue estimating incidence from hospital admissions data, which are available at the desired dates. Let $H(t)$ be the number of hospital admissions at time t. It is common to use the following simple formula for the incidence at time t: $Incidence(t) = H(t+7)/0.032$, with 0.032 the average fraction of hospitalisations per infectious case [43]. Instead of considering the exact distribution of the time needed for infected individuals to become hospitalized (if they do), we are assuming that individuals become hospitalized after exactly the average required time (7 days) [43]. Both for France and Italy we sum cases over weekly basis to match the reproduction operators' time resolution. We call x_t the normalized cases distribution vector at week t. We adopt, as previously done, one minus cosine similarity as a measure of similarity between x_t and v_t.

We have seen that discontinuities over time of the national mobility pattern have a relevant impact on the convergence process. When analysing the convergence of $\mathcal{S}$ to $\mathcal{R}$ under mobility discontinuities, we stressed that when Q changes the true reproductive number changes as well. In the case of the asymptotic convergence of x to v, if Q changes the Perron eigenvector v also changes. Variations of v may be both of small magnitude, meaning that the leading component does not change, or of greater magnitude, if a new leading component establishes. If the latter is true, the system is likely to be abruptly driven far from equilibrium. We should then contextualize the convergence process of x to v with respect to mobility pattern discontinuities and their intensities. We choose as a metric the Frobenius matrix norm defined as:

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{M} |a_{ij}|^2} = \sqrt{trace(A^T * A)} \tag{5.1}$$

with a_{ij} the elements of a square matrix $A \in \mathbb{R}^{M,M}$. We compute $\|Q_t - Q_{t-1}\|_F$ for subsequent reproduction operators.

5.2 Data analysis results

In Fig. 5.1, **a** and **b**, we show $1 - cs(x_t, v_t)$ and $\|Q_t - Q_{t-1}\|_F$ for France and Italy. Let's first focus on mobility discontinuities. Interestingly, variations of the reproduction operator in the two countries have common patterns. We may identify four main peaks of $\|Q_t - Q_{t-1}\|_F$ shared by France and Italy during the

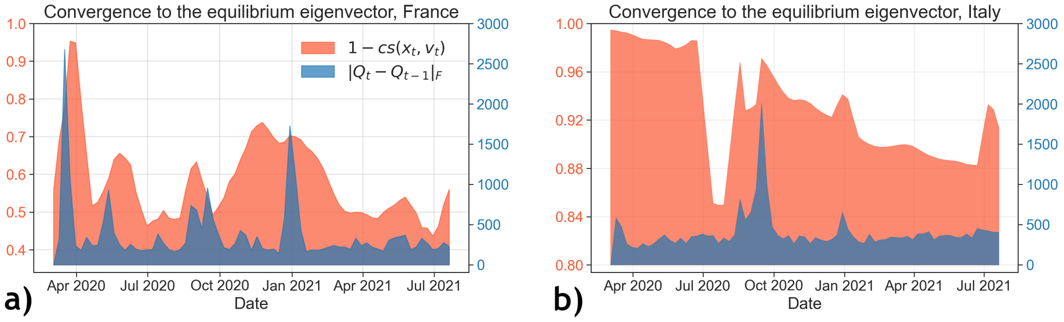

Figure 5.1: *Convergence of the cases distribution vector x (from COVID-19 data) to the equilibrium eigenvector v of Q over time in France (**a**) and Italy (**b**), 2020-03-03 to 2021-07-20. We plot $1 - cs(x_t, v_t)$ as a metric of similarity of the two vectors and $\|Q_t - Q_{t-1}\|_F$ as a metric for mobility patterns variations. Data resolution is weekly, with weeks starting on Tuesdays and ending the following Mondays, following the convention of Meta Colocation Maps. Weekly values are plotted in correspondence of Tuesdays on the date axis. We do a rolling mean of cosine similarity (but not of Frobenius norm). We substitute a generic value at time t with the unweighted average of values at $t-1$, t, $t+1$. Also, we adopt different scales for the two countries to facilitate the figure comprehension. For what concerns $\|Q_t - Q_{t-1}\|_F$, we set the first point to zero.*

observed period, plus a fifth one in France. The first shared peak is in early March and is representative of the mobility reduction due to the introduction of the first lockdown in the two countries. The exact lockdown onset dates are 16 March 2020 for France and 9 March 2020 for Italy [39, 40]. Peaks of $\|Q_t - Q_{t-1}\|_F$ are exactly located at the corresponding weeks. The release of the lockdown is on 11 May in France, with again a visible discontinuity peak. Things are more complicated in Italy, with mobility restrictions gradually lifted in the time span of a month, from 4 May to 3 June, so that we cannot see a single Frobenius norm peak. We then have a double peak, that we found to be present for Spain, Portugal and Sweden as well. $\|Q_t - Q_{t-1}\|_F$ is high in the beginning of August, when many workers go on holidays, and then again in the beginning of September, when holidays end. The last major mobility variation takes place at Christmas, with between patches mobility likely increasing due to people reaching their families. As a general remark, note that the *plateau* values of $\|Q_t - Q_{t-1}\|_F$ are approximately equal to the smallest eigenvalue of Q and the highest peaks are of the order of magnitude of the Perron eigenvalue.

All $\|Q_t - Q_{t-1}\|_F$ peaks are accompanied by $1 - cs(x_t, v_t)$ increasing. This tells us that the system was actually leading towards equilibrium and was then driven away from it. After these peaks, $\|Q_t - Q_{t-1}\|_F$ gradually decreases again. The least noisy evidence of the convergence phenomenon we want to observe is the period from January 2021 to April 2021 in France, with $\|Q_t - Q_{t-1}\|_F$ almost completely flat and x gradually approaching v. What may instead look atypical is the abrupt descent of $1 - cs(x_t, v_t)$ in Italy in the beginning of July, followed by a fast growth again. What is actually happening is that a new leading component establishes in v in early July. This new v is nearer to the distribution of cases at the time in Italy. Then the old component establishes again as the leading one. In general, many leading component shifts occur in the observed period in Italy, due to the small spectral gap between the first eigenvalues, and Frobenius norm peaks are always accompanied by these shifts. Conversely, only two shifts occur in France, at the beginning of the observed period. Apart from these exceptions, the Paris component of v is always the major one.

As a last remark, we note that $1 - cs(x_t, v_t)$ is typically higher in Italy than in France and that slopes pointing towards equilibrium are steeper in France. Both these results are in agreement with our simulations: in our model Italy was slower at going towards equilibrium and never reached it, while France was faster and could reach it. This is, again, a matter of spectra. The larger the spectral gap between the first two eigenvalues, the faster the convergence, Eq. (2.14, 2.15, 2.16).

Chapter 6

An alternative estimate of $\mathcal{R}$

We have shown that surveillance based estimates of the reproductive number show a bias in spatially structured populations as far as the system is out-of-equilibrium. We found this result analytically first, then verified it simulating an epidemic through a stochastic model and lastly we brought evidence analysing COVID-19 surveillance data published by France and Italy's governments. In this chapter, we define an alternative method for the estimate of the reproductive number, combining surveillance data and colocation data.

6.1 Measure definition

Our aim is here to define an alternative estimate measure for the reproductive number $\mathcal{R}$ that does not present a bias at the early stages of the epidemic. We will proceed in analogy with the already defined measure $\mathcal{S}$. We remind that when measuring $\mathcal{S}$ we consider the total number of cases at time t, i.e., $I_{t,tot} = F^T I_t = \sum_i I_{i,t}$, with $i = 1, ..., M$ spatial patches. $\mathcal{S}_t$ is simply given by the ratio between the total number of cases at time $t+1$ and the total number of cases at time t:

$$\mathcal{S}_t = \frac{I_{t+1,tot}}{I_{t,tot}} = \frac{F^T Q I_t}{F^T I_t} = F^T Q x_t, \tag{6.1}$$

as in Eq. 2.3. We want to eliminate the dependence of this measure on the spatial distribution of cases, which is time-evolving and responsible for the bias. It is therefore natural to replace the total number of cases by a projection of the cases vector on the equilibrium eigenvector v. We define $J_{t,tot} = M v^T I_t$, where M serves as a normalization constant, assuring the weights sum is conserved $(\sum_i F_i = M = M \sum_i v_i)$. Our new measure will be given by the ratio between subsequent J:

$$\mathcal{S}_t(v - projection) = \frac{J_{t+1,tot}}{J_{t,tot}} = \frac{v^T I_{t+1}}{v^T I_t} = \frac{v^T Q I_t}{v^T I_t}. \tag{6.2}$$

The idea is to observe how the epidemic is evolving on the equilibrium vector subspace only. We actually know that measures of the reproductive number are correct and time independent when restricting to this subspace. Instead of waiting the cases distribution x to converge to the equilibrium distribution v, we can project I_t and obtain a non biased measure even at the early stages of the considered epidemic.

One question may arise, concerning real life applications: assuming the reproduction operator Q is known, why shouldn't we simply compute its largest eigenvalue as a measure of the reproductive number of the system? As already discussed, Q depends on a constant pre-factor (see Eq. 3.7). This does affect the magnitude of the largest eigenvalue, but it does not affect the eigenvalue v. In general, estimating this pre-factor is difficult and would require to consider a large number of variables we are neglecting in our analysis. For this reason, the *v-projection* method is simpler and less subject to errors.

6.2 Testing the measure

We want to test the performance of the defined alternative measure $S(v{-}projection)$. To this aim, we take the metapopulation model of Chapter 4 for France. We set β such that $\mathcal{R} = 1.5$. Analogously with what done in Chapter 4, we accompany analytical estimates of the reproductive number with EpiEstim estimates. In this case, we refer to EpiEstim$(v - projection)$ as the EpiEstim estimates obtained giving as input the projection of cases on v. The reasoning is the same as above: we want to focus on the v-subspace only when monitoring how the epidemic develops. We take 100 initial cases and extract them according to two different multinomial distributions. In a first case the multinomial distribution p are such that $p_i \propto n_i \;\; \forall i$ ($I.C.\ 0$), in the second one $p_{Hauts-De-Seine} = 1$, $p_i = 0 \;\; \forall i \neq Hauts - De - Seine$ ($I.C.\ 1$). 1000 runs were made. Median values and 95% confidence intervals are reported in Fig. 6.1. In both cases reported in Fig. 6.1 $\mathcal{S}(v - projection)$ and EpiEstim$(v - projection)$ are able to estimate the reproductive number with very little or no error, even though the confidence interval for EpiEstim$(v - projection)$ is quite large. $\mathcal{S}(v - projection)$, on the other hand, has very little confidence interval, proving itself to be a very stable and reliable measure. In any case, the proposed estimates have clearly better performances than the canonical ones.

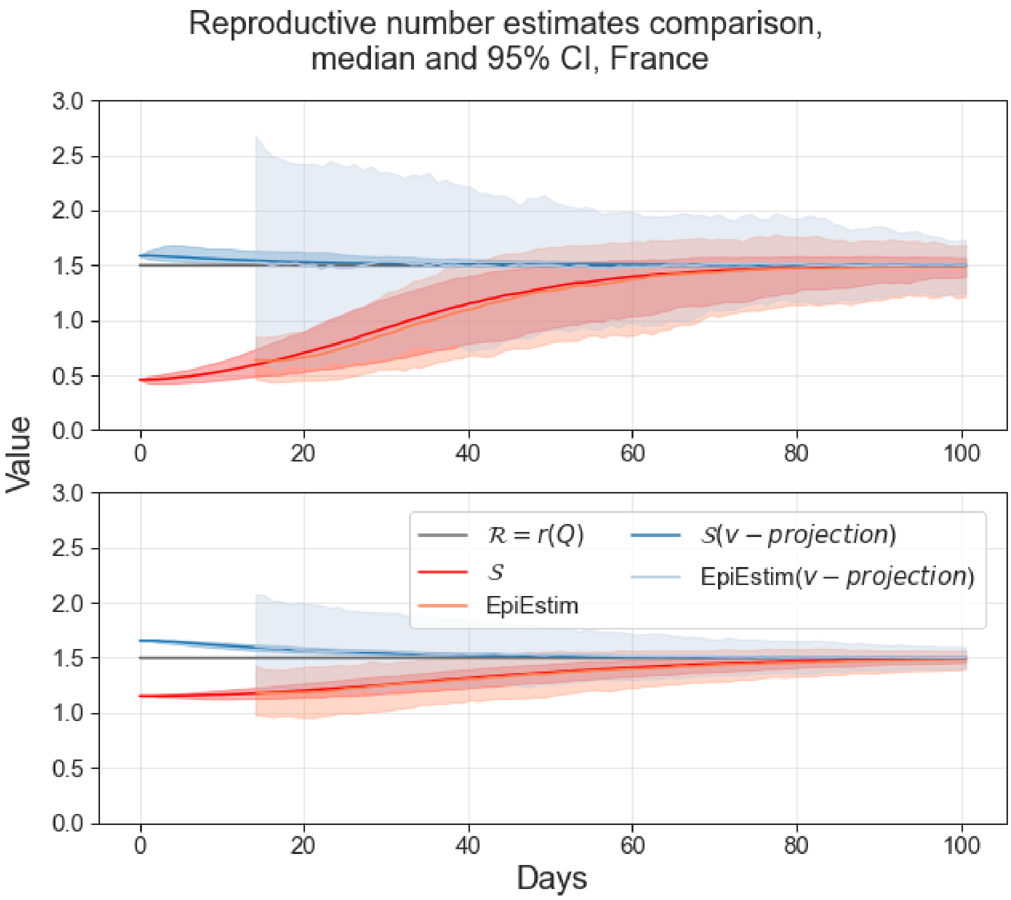

Figure 6.1: *$\mathcal{S}(v - projection)$ and EpiEstim$(v - projection)$ compared to $\mathcal{S}$ and EpiEstim, with I.C. 0 in the top panel and I.C. 1 in the bottom panel. EpiEstim estimates relative to the first two weeks since the start of simulations were discarded. The proposed v-projection measure is not subject to the early stages bias.*

Chapter 7

Conclusions and outlook

In this work we showed that in spatially structured populations surveillance-based estimates of the reproduction number undergo a transient, out-of-equilibrium, period during which they commit substantial error with respect to the true value of $\mathcal{R}$. Then, we showed that non biased measures of the reproductive number may be obtained combining surveillance data and colocation data. Precisely, our receipt consists in replacing incidence with its projection on the Perron eigenvector of the so called reproduction operator when estimating $\mathcal{R}$.

It was already known that in the case of heterogeneous populations the observed growth factor converges to the true reproduction number of the system only after some time, and that a similar process exists for the cases distribution with respect to the Perron eigenvector of the reproduction operator. What we did was first to analytically expand these results focusing on the out-of-equilibrium period. We considered the case of a discrete set of spatial patches in a country. We derived equations for the evolution over time of the global observed reproductive number of the system. We found that $\mathcal{R}$ estimates may oscillate around the true $\mathcal{R}$ if the system is strongly coupled. However, what typically happens is that the system is weakly coupled and oscillations do not occur. This means that depending on initial conditions, surveillance-based estimates systematically either underestimate or overestimate $\mathcal{R}$ during the transient period. We then introduced a definition for local measures of the reproductive number and studied their time evolution as well. In this case we found that oscillations can occur even if the system is weakly coupled. Lastly, we studied how the distribution of cases in patches x behaves in time and found its time evolution equations. We identified the spectral structure of the reproduction operator Q to be a major factor determining the convergence time of x to the Perron eigenvector of Q. We stress that the analytical discussion we developed is in fact valid for generic discrete heterogeneity classes. Motivation for further studies may be to apply the obtained results to the case where classes are - as an example for which data are often available - age classes.

In Chapter 3 we laid the basis for the application of our findings. We reconstructed weekly resolution reproduction operators for France and Italy from Colocation Maps and Movement Range Maps. We analysed how the mobility networks in France and Italy differ from each other and how mobility changed in time due to COVID-19 epidemic response measures in the past two years. We were then able to compute the spectra of the operators along time and the leading eigenvectors corresponding to the equilibrium distribution of cases. We made some predictions concerning how the star-like mobility network structure in France and the more democratic network in Italy would affect our out-of-equilibrium process.

In Chapter 4 we built a stochastic metapopulation model to simulate an epidemic process, validate our theoretical findings and our data-based specific case predictions, to observe the actual convergence times for $\mathcal{S}$, s_i and for x depending on a number of factors, and to evaluate the surveillance-based estimates bias. Given these objectives, we decided to compute observed reproductive numbers not only through our analytical formula, but also through the R-package EpiEstim. Once computed the true reproductive number for our model and the generation interval, we designed three different experiments.

In Exp. I we confirm that the convergence time depends on the initial weights associated to the most important non-leading eigenvectors of Q. We also see that particular conditions exist for which surveillance-based estimates overestimate $\mathcal{R}$ in the transient period, but what typically happens is that $\mathcal{S} < \mathcal{R}$ out-of-equilibrium.

In Exp. II we see that, as expected, the Italian mobility network structure makes convergence times longer. Interestingly, for all transmission rates such that an epidemic outbreak occurs in more than one patch, and given SARS-CoV-2 medical estimates for the other parameters, the linear approximation on which our findings rely expires before equilibrium is reached. This implies that $\mathcal{S} < \mathcal{R}$ for all the duration of the epidemics. A motivation for further works may be to properly study this last empirical finding and to validate it over different network structures. Precisely, interesting research questions may be: is it generally true that the $\mathcal{S}$ to $\mathcal{R}$ convergence time is shorter than the linear approximation time, given that $\mathcal{R}_0$ is above the threshold for the infection of a macroscopic number of subpopulations (see [31])? If not, under which properties of the network does this hold? Is it more specifically true for weakly coupled networks like those describing the interaction of spatially separated communities? Are even more strict conditions necessary? Answering to these questions would allow us to determine whether the surveillance-based estimates bias in structured populations concerns the whole period of existence of a disease, and not only its early stages.

In Exp. III we evaluated the effect of discontinuities in time of the mobility network. We saw that the short-term behaviour of surveillance-based estimates may be misleading after some major discontinuities take place. In particular, considering the introduction of mobility restrictions or their release, we saw that cases exist under which we either overestimate or underestimate the impact of these measures. In some other cases the effect of these measures is not visible from reproductive number estimates. For completeness, we both considered equilibrium and non equilibrium stages of the epidemic. Interesting future research questions arise in this case too. What happens in the case of multiple discontinuities of Q? Is there some frequency (or intensity) in discontinuities that does not allow for equilibrium reaching? The analysis of multiple discontinuities is actually relevant for real-case studies, since human mobility typically undergoes continuous variations of different intensity.

We actually performed a preliminary analysis of how multiple discontinuities shape the convergence process using COVID-19 reported cases data in France and Italy (Chapter 5). We found, qualitatively, that variations $Q \to Q'$ such that the Frobenius norm of $Q - Q'$ is larger than approximately the value of the smallest eigenvalue are those that drive the system away from equilibrium. When variations are of smaller intensity, $1 - cs(x, v)$, with v the Perron eigenvector, decreases along time.

Lastly, in Chapter 6 we defined a new method for the estimate of the reproductive number $\mathcal{R}$ we call $\mathcal{S}(v - projection)$. $\mathcal{S}(v - projection)$ is computed analogously to $\mathcal{S}$ but replacing the cases vector at time t I_t with its projection on v, i.e. J_t. First we motivated this choice, then we tested the accuracy of the method through our metapopulation model. What we claim is that $\mathcal{S}(v - projection)$ is a non biased measure of $\mathcal{R}$, presenting little or no error even out-of-equilibrium. Similarly, we showed that EpiEstim returns better estimates of the reproductive number when giving J_t as an input.

In conclusion, we believe all these contributions and future directions will help to improve our understanding of epidemic dynamics in increasingly realistic and heterogeneously structured models, to improve our methods aimed at estimating the reproductive number for a disease spreading in a complex structured population, and hence to contribute to correctly measure the reproductive number during epidemic outbreaks.

Bibliography

[1] Keeling, M., and P. Rohani, 2007, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, Princeton).

[2] Barrat, A., M. Barthélemy, and A. Vespignani, 2008, *Dynamical Processes on Complex Networks* Cambridge: Cambridge University Press.

[3] Goldstein, E., K. Paur, C. Fraser, E. Kenah, J. Wallinga, M. Lipsitch, 2009, *Reproductive numbers, epidemic spread and control in a community of households* Mathematical biosciences **221** (1): 11-25.

[4] Hartfield, M., S. Alizon, 2013, *Introducing the outbreak threshold in epidemiology* PLoS pathogens **9** (6): e1003277.

[5] Anderson, R. M., and R. M. May, 1992, *Infectious Diseases in Humans*, (Oxford University Press, Oxford).

[6] Kermack, W. O., and A. G. McKendrick, 1927, Proc. R. Soc. A **115**: 700.

[7] Pastor-Satorras, R., C. Castellano, P. Van Mieghem and A. Vespignani, 2015, *Epidemic processes in complex networks*, Reviews of Modern Physics **87**.

[8] Diekmann, O., and J. Heesterbeek, 2000, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, (John Wiley & Sons, New York).

[9] Cox, D. R., 1967, *Renewal Theory* (Methuen, London).

[10] Ross, S. M., 1996, *Stochastic Processes* (John Wiley & Sons, New York).

[11] Van Kampen, N. G., 1981, *Stochastic Processes in Chemistry and Physics* (North-Holland, Amsterdam).

[12] Hethcote, H. W., 2000, SIAM Rev. **42**: 599.

[13] Diekmann, O., J. A. P. Heesterbeek and J. A. J. Metz, 1990, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. of Mathematical Biology **28**: 365-82.

[14] Jagers, P., 1989, *General branching processes as Markov fields*, Stochastic Proc. and their Applications **32**: 183.

[15] Schaefer, H. H., 1960, *Some spectral properties of positive linear operators*, Pacific J. of Mathematics **10**: 1009.

[16] Iyer, S., *et al*, 2020, *Large-Scale Measurement of Aggregate Human Colocation Patterns for Epidemiological Modeling*, medRxiv.

[17] Wang, Y., D. Chakrabarti, C. Wang, and C. Faloutsos, 2003, *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*, in Proceedings of the 22nd International Symposium on Reliable Distributed Systems (IEEE, Los Alamitos, CA), pp. 25–34.

[18] Gómez, S., A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, 2010, *Discrete Time Markov Chain Approach to Contact-Based Disease Spreading in Complex Networks*, Europhys. Lett. **89**: 38009.

[19] Susswein, Z., Valdano, E., *et al*, 2021, *Ignoring spatial heterogeneity in drivers of SARS-CoV-2 transmission in the US will impede sustained elimination*, medRxiv.

[20] Cori, A., *et al*, 2013, *A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics*, American J. of Epidemiology **178**: 1505.

[21] Perron, O., 1907, *Zur Theorie der Matrices*, Mathematische Annalen **64**(2): 248-263.

[22] Horn, R. A., and C. R. Johnson, 1985, *Matrix Analysis*, Cambridge University Press.

[23] Gerschgorin, S., 1931, *Über die Abgrenzung der Eigenwerte einer Matrix*, Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk **6**: 749–754.

[24] Levins, R., 1969, *Some demographic and genetic consequences of environmental heterogeneity for biological control*, Bulletin of the Entomological Society of America, **15** (3): 237–240.

[25] Viboud, C., O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell, 2006, *Synchrony, waves, and spatial hierarchies in the spread of influenza*, science **312**(5772): 447–451.

[26] Balcan, D., V. Colizza, B. Goncalves, H. Hu, J. J. Ramasco, and A. Vespignani, 2009, *Multi-scale mobility networks and the spatial spreading of infectious diseases*, Proceedings of the National Academy of Sciences **106**(51): 21484–21489.

[27] Van den Broeck, W., C. Gioannini, B. Goncalves, M. Quaggiotto, V. Colizza, and A. Vespignani, 2011, *The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale*, BMC infectious diseases **11**(1): 37.

[28] Mazzoli, M., E. Valdano, and V. Colizza, 2021, *Projecting the COVID-19 epidemic risk in France for the summer 2021*, J. of Travel Medicine.

[29] Génois, M., and A. Barrat, 2018, *Can co-location be used as a proxy for face-to-face contacts?*, EPJ Data Science **7**: 11.

[30] Arino, J., 2006, *Metapopulation epidemic models. A survey.*, Fields Institute Communications **48**: 1-12.

[31] Colizza, V., and A. Vespignani, 2008, *Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations*, J. of Theoretical Biology **251**: 450–467.

[32] Data at https://dataforgood.facebook.com/dfg/tools/movement-range-maps

[33] Aron, J. L., and I. B. Schwartz, 1984, *Seasonality and period-doubling bifurcations in an epidemic model*, J. Theor. Biol. **110**: 665-679.

[34] Faucher, B., *et al.*, 2022, Agent-based modelling of reactive vaccination of workplaces and schools against COVID-19, Nature Communications **13**.

[35] Davies, N. G., P. Klepac *et al.*, 2022, *Age-dependent effects in the transmission and control of COVID-19 epidemics* **28**.

[36] Diekmann, O., J. A. P. Heesterbeek and M. G. Roberts, 2009, *The construction of next-generation matrices for compartmental epidemic models*, J. R. Soc. Interface **28**: 873-885.

[37] Rychetnik, L., P. Hawe, E. Waters, A. Barratt, M. Frommer, 2004, *A glossary for evidence based public health*, J. Epidemiol. Community Health, **58**(7): 538–45.

[38] Coale, A.J., 1972, *The Growth and Structure of Human Populations*, Princeton University Press, 18–19.

[39] 2022, *COVID-19 pandemic in Italy*, Wikipedia, The Free Encyclopedia.

[40] 2022, *COVID-19 pandemic in France*, Wikipedia, The Free Encyclopedia.

[41] Funk, S., M. Salathé and V. A. A. Jansen, 2010, *Modelling the influence of human behaviour on the spread of infectious diseases: a review*, J. R. Soc. Interface **7**: 1247–1256.

[42] Italian Civil Protection Department, M. Morettini, A. Sbrollini, I. Marcantoni, L. Burattini, 2020, *COVID-19 in Italy: Dataset of the Italian Civil Protection Department*, Data in Brief **30**.

[43] Pullano, G., L. Di Domenico, C.E. Sabbatini *et al.*, 2021, *Underdetection of cases of COVID-19 in France threatens epidemic control*, Nature **590**: 134–139.