# Emerging dynamics from high-resolution spatial numerical epidemics

Olivier Thomine[1,*], Samuel Alizon[2], Marc Barthelemy[3], Corentin Boennec[2], Mircea T. Sofonea[2]

[1] LIS UMR 7020 CNRS, Aix Marseille University, France
[2] MIVEGEC, CNRS, IRD, Montpellier University, France
[3] Institut de Physique Théorique, CEA, CNRS-URA 2306, F-91191, Gif-sur-Yvette, France
[*] Corresponding author, olivier.thomine@univ-amu.fr

**Simulating nationwide realistic individual movements with a detailed geographical structure is urgently needed to optimize public health policies. However, existing tools have limited resolution or can only account for a limited number of agents. We introduce Epidemap, a new framework that can capture the daily movement of more than 60 million people in a country at a building-level resolution in a realistic and computational-efficient way. By applying it to the case of an infectious disease spreading in France, we uncover hitherto neglected effects, such as the emergence of two distinct peaks in the daily number of cases or the importance of local density in the timing of arrival of the epidemic. Finally, we show that the importance of super-spreading events strongly varies over time.**

***124 words in the abstract. 2352 words in the manuscript (including references and figure captions)***

A major limitation to the applicability of mathematical models is that they need to resort to strong assumptions. In the case of infectious disease epidemiology, for instance, it is easy to point out that the canonical Susceptible-Infected-Recovered model (*1*) neglects stochasticity, misses spatial structure or ignores contact patterns. Although these aspects, and many others, can be incorporated with the appropriate hypotheses and techniques (*2–4*), the resulting models usually remain too abstract to help improve local policies.

Agent-Based Simulation (ABS), where individuals are modelled explicitly, represent a seducing option to achieve a high degree of realism, from a biological and environmental point of view (*5*), but their routine use in public health faces three major limitations. First, ABS are very computationally demanding (*6*), which restrains the total number of agents that can be simulated. The second limitation comes from the model dimensionality and the introduction of numerous parameters, many of which are poorly informed and set in an *ad hoc* manner, to capture individual heterogeneity. A third limit resides in the way geographic structure is implemented into the simulation (but see (*7*)).

Here, we introduce Epidemap, a novel numerically efficient agent-based method that addresses the current limitations of ABS platforms. Epidemap simulates infectious diseases epidemic scenarios at the scale of a whole country by combining detailed geographical structure, demographic information, and mobility statistics. Practically, this is achieved by using High-Performance Computing (HPC) techniques, high-resolution spatial data from the OpenStreetMap (OSM) project (*8*), and sophisticated mobility models (*9*). Overall, these simulations only require 8 parameters (see Supplementary Methods).

To illustrate the power and flexibility of the platform, we study the transmission dynamics of an (uncontrolled) epidemic of respiratory infections in France. The biological features of the epidemiological model originate from a discrete-time SARS-CoV-2 transmission model parameterised with national hospital data (*10*). We analyse the output of 100 stochastic simulations

from epidemic emergence to extinction, *i.e.* approximately one year. A typical simulation generates daily mobility patterns for 65 million individuals at the scale of buildings and lasts less than 2 hours on a 12-core AMD Ryzen 9 CPU, with less than 64 GB of RAM.

In the simulation (Figure 1), each individual is assigned to a residency building following the national density estimates at the district level, the French 'canton' level (*11*). Every day, we assume that individuals visit 0, 1, or 2 buildings other than their own residence (*12*). These buildings are chosen at random based on a distance kernel. If more than one individual visits the building the same day, transmission can occur with a probability $b = 6\%$ (*13*), which is weighted by the number of days since the onset of the infection in the donor host. As in previous studies, this weight is based on the SARS-CoV-2 serial interval (*10, 14*). The simulation tracks the age of the individuals, which is distributed geographically according to national demographics data (*15*). For infected individuals, disease progression follows that described in (*10*) and a fraction $\theta$ of infected individuals develop severe symptoms requiring hospitalisation, which is simulated by moving them to the nearest hospital building from their residence. For all recovered individuals, we assume perfect immunity for the duration of the simulation. All simulations are initialised with 15 infected individuals in Paris to avoid premature random epidemic extinction.

Figure 2 shows the output of the dynamics at the national level for 100 stochastic simulations. For optimal readability, the dates are aligned based on the day where ICU-occupancy reaches a value of 700. With our minimal parameterisation, we see that the basic reproduction number, which is denoted ($R_0$) and corresponds to the average number of secondary infections caused by an infected individual (*16*), is in the order of 3, which is consistent with estimates for the French epidemic (*10, 17*). Note that this value is here computed directly at the individual level, by counting how many infections an individual causes. Furthermore, the daily estimates for the temporal reproduction number calculated in the same manner are very similar to those esti-

mated using the daily case incidence data (dashed blue curve) and the method from (*18*). These uncontrolled epidemics last 308 days (95% confidence interval (CI): [286;345] days) and the final total epidemic size is $q = 61.1\%$ (95%CI [60.1%;61.9%]) of the initial susceptible population. As described by earlier studies (*19*), this proportion is lower than the prediction from a mean-field model that is given by the well-known equation $qR_0 + \log(1 - q) = 0$ (*1*), which yields $93\%$ for $R_0 = 3$. This shows that geographical structure greatly impacts the unfolding of the epidemic.

The national prevalence data uncovers a bimodal structure of the epidemic, which can be understood by moving to the regional scale (Figure 2b). The first peak corresponds to the spread in the region where the outbreak emerges (here the Ile-de-France), whereas the second corresponds to the sum of the epidemic peaks in the other regions. This bimodal structure is particularly pronounced given the population density in the region of origin but it is a direct consequence of the detailed geographical information in the simulation platform.

The resolution of the Epidemap simulations allows us to perform analyses at the district level (see Supplementary Videos 1 and 2). In Figure 3a, we show that the date of onset of the epidemic in an area strongly depends on its distance from the origin (here assumed to be Paris). Furthermore, there is an additional effect of density such that denser areas are infected first. Interestingly, at the departmental level, these trends are not significant, further showing the importance of a fine-grain simulation level. Furthermore, the total proportion of inhabitants of a district who have been infected at the end of the epidemic strongly increases with density. The pandemic propagation velocity also increases with time. This can be explained by the fact that, when incidence is high, long-distance dispersion events are not rare anymore, which biases the mean distance of contamination towards higher values. For densely-populated districts (Figure 3b), this proportion converges towards the mean-field prediction from well-mixed models mentioned above (*1*). However, if the population density is low, this proportion is more variable

showing the limit of classical assumptions.

The individual-based nature of our simulations allows us to follow transmission chains (Supplementary Video 3), which has direct applications. For instance, we can count, at the end of each infection, how many secondary infections were caused. The distribution of these individual reproduction numbers is particularly important in the context of emerging epidemics because the more disperse, the more the spread relies on super-spreading events (*20*). Early in the epidemic, the distribution is tightly centred around the $R_0$ value (Figure 4a). As the epidemic unfolds, the distribution changes with a mode that decreases towards 0, and a wider dispersal. This pattern can be formalised by assuming that the distribution of individual reproduction numbers follows a negative binomial distribution (*20*). In Figure 4b, we show that the mean of this distribution (in blue) follows the pattern estimated in Figure 2a. The dispersal parameter ($k$) indicates that super-spreading events reach a peak at the end of the first national epidemic wave. During the end of the epidemic, stochasticity is strong but there is a general trend towards a decrease in super-spreading events. Note that in these simulations we do not introduce host heterogeneity, which means we only capture the dimension of superspreading that originates from spatial heterogeneity. This is already sufficient to show that studies attempting to quantify the importance of superspreading events should account for the stage of the epidemic they analyse.

Simulating the daily activity of millions of individuals at a build-level resolution with a realistic mobility model significantly improves our understanding of how epidemics unfold. Early stages appear to be consistent with stochastic and deterministic mean-field models. However, once local saturation effects cannot be neglected anymore, Epidemap reveals striking patterns with an unprecedented resolution. First, we show that at a national level, a two-waves epidemic pattern is to be expected, largely because of a wide regional structure. Second, we find that districts are affected by the epidemic depending on how far they are from the epicentre, but also depending on their density. The latter effect is absent at the departmental level, which shows the

added value of a detailed geographical structure. Furthermore, we show that, as expected, the simplistic estimate of the final epidemic size as a function of $R_0$ does not apply at the national level. However, we also show that it does yield relevant results at the district level if the density is sufficiently high. Finally, individual follow-ups allow us to show how super-spreading events becoming increasingly important as the epidemic unfolds, but that their role decrease as the role of stochastic processes increases.

Being able to perform these detailed simulations for so many agents at a national scale has to be traded-off against some simplifying assumptions. The major one is that individual movements are based on a distance kernel centred in their residency home. As shown by earlier work on influenza dynamics, although this assumption is relevant for France, it might be too simplistic for countries like the United States, where air traffic represents a greater proportion of the travels (*21*). A second limitation is that, because of the lack of relevant data, we assumed all individuals to have the same average behaviour, *e.g.* in terms of the number of buildings visited per day. However, it can readily be made a function of the age of infection status (or of the individual) in the simulations.

These results have immediate applications for public health. For instance, they allow authorities to derive a risk-factor per district to help control an epidemic and prevent outbreaks. This can also be instrumental to optimise vaccine coverage, or other control strategies (mask-wearing, lockdown, vaccination). Many insights can also the gained from the ability to follow individual trajectories and transmission chains (Supplementary Video 3), and superspreading events. Further developments in Epidemap will incorporate building meta-data, for instance, to simulate children attending the closest school from the residency home or study the local saturation of hospital emergency care services. Finally, we focused here on SARS-CoV-2 but an asset of this platform is its versatility. It can readily be deployed to study other infectious agents to investigate a variety of control scenarios with high resolution at a national level.

# References and Notes


1. W. O. Kermack, A. G. McKendrick, *Proc. R. Soc. Lond. A* **115**, 700 (1927).

2. L. Pellis, *et al.*, *Epidemics* **10**, 58 (2015).

3. S. Lion, *J theor Biol* **405**, 46 (2016).

4. T. Britton, G. Scalia Tomba, *Journal E Soc Interface* **16**, 20180670 (2019).

5. S. Abar, G. K. Theodoropoulos, P. Lemarinier, G. M. P. O'Hare, *Computer Science Review* **24**, 13 (2017).

6. S. Eubank, *et al.*, *Nature* **429**, 180 (2004).

7. P. Taillandier, *et al.*, *GeoInformatica* **23**, 299 (2019).

8. M. Haklay, P. Weber, *IEEE Pervasive Computing* **7**, 12 (2008).

9. H. Barbosa, *et al.*, *Physics Reports* **734**, 1 (2018).

10. M. T. Sofonea, *et al.*, *Epidemics* in press (2021).

11. INSEE, Populations légales 2017. recensement de la population régions, départements, arrondissements, cantons et communes, `https://www.insee.fr/fr/statistiques/4265511` (2019).

12. C. Schneider, V. Belik, T. Couronné, Z. Smoreda, M. Gonzalez, *J R Soc Interface* **10** (2013).

13. Variant Technical group, *Public Health England. Technical briefing 3.* (2021).

14. H. Nishiura, N. M. Linton, A. R. Akhmetzhanov, *Int J Infect Dis* **93**, 284 (2020).

15. INSEE, Population totale par sexe et âge au 1er janvier 2020, france métropolitaine, https://www.insee.fr/fr/statistiques/1892088?sommaire=1912926 (2020).

16. R. M. Anderson, R. M. May, *Infectious Diseases of Humans. Dynamics and Control* (Oxford University Press, Oxford, 1991).

17. H. Salje, *et al.*, *Science* **369**, 208 (2020).

18. J. Wallinga, M. Lipsitch, *Proc R Soc Lond B* **274**, 599 (2007).

19. M. J. Keeling, *Proc R Soc Lond B* **266**, 859 (1999). Publisher: Royal Society.

20. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, *Nature* **438**, 355 (2005).

21. P. Crépey, M. Barthelemy, *Am J Epidemiol* **166**, 1244 (2007).

22. The authors thank the Région Occitanie and the ANR for funding (PhyEpi grant).

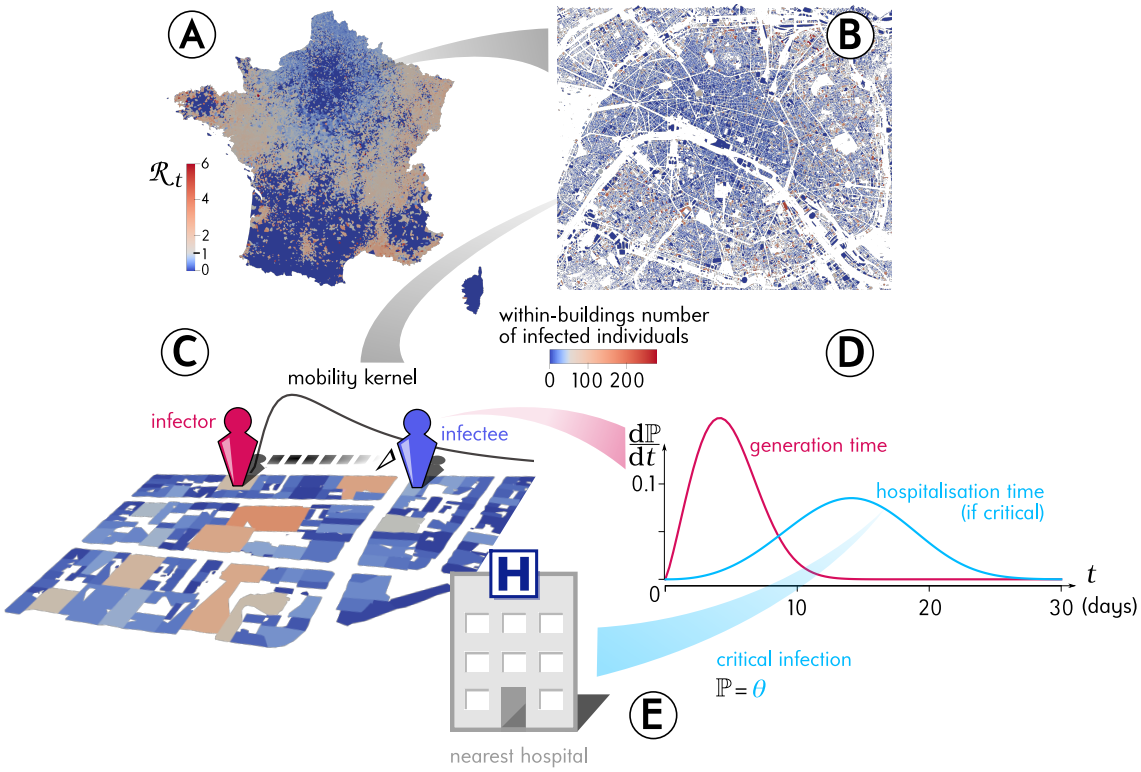


Figure 1: **Outline of the Epidemap simulation framework. A)** The 65 millions inhabitants of metropolitan France are explicitly mapped to housing buildings following cartographic and demographic data. **B)** At each time point of the simulation, the number of infected individuals in each building of the country is recorded, as well as the time past since each got infected (the panel shows the Paris area).**C)** Every day, individuals can randomly move from their home to other buildings according to a mobility kernel and meet other people. If an infected individual encounters a susceptible host, a transmission event can occur. **D)** The contagiousness of a infected individuals varies depending on the time since infection. **E)** A small fraction $\theta$ of infected individuals develop a critical form of the disease that requires their hospitalisation to the nearest facility.

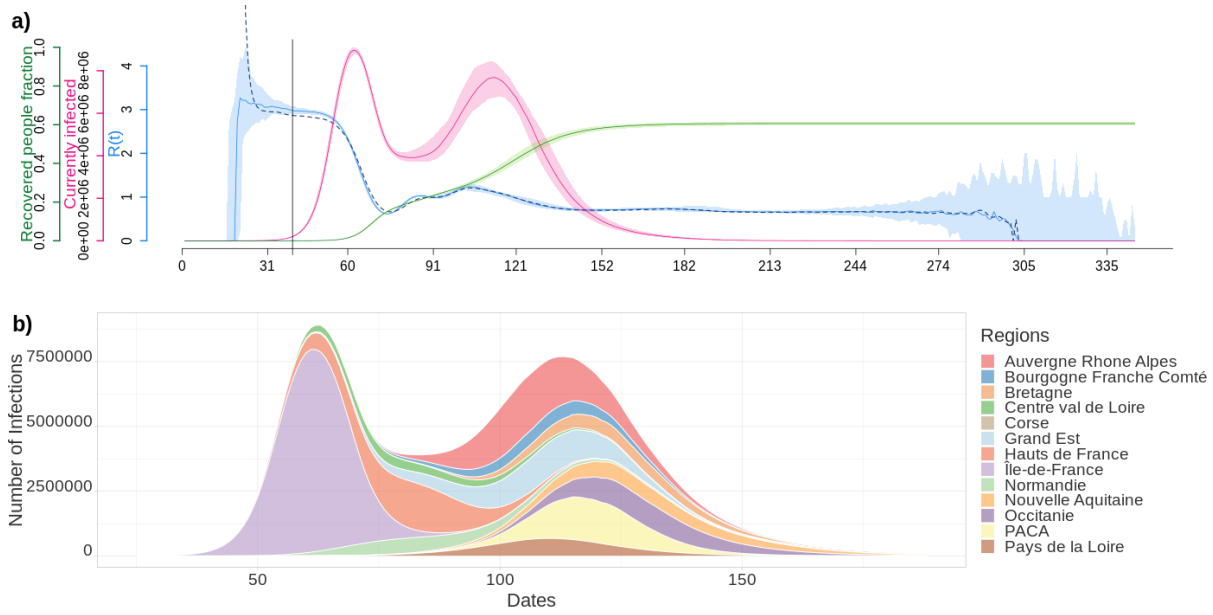


Figure 2: **Epidemiological dynamics at the national (a) and regional (b) level.** a) The daily prevalence (number of infected individuals) is in red, the temporal reproduction number ($R_t$) in blue, and the cumulative number of recovered individuals in green. Shaded areas show the 95% sample quantiles of 100 stochastic simulations. The dashed line shows the median $R_t$ calculated on the case incidence data. b) Each color shows the prevalence in a French region.

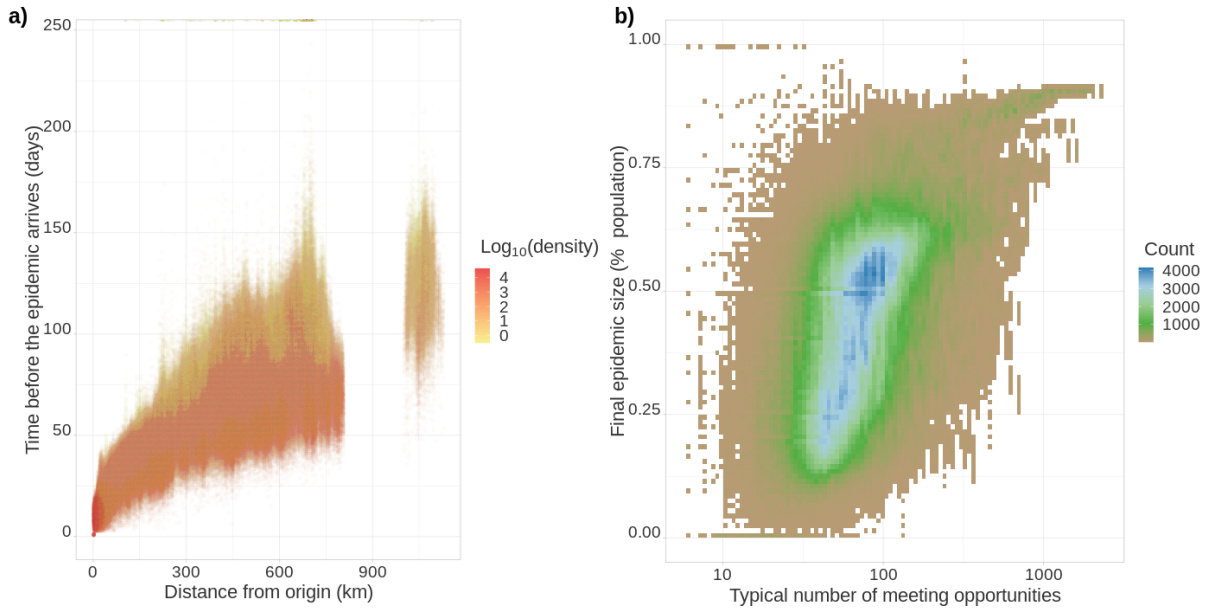


Figure 3: **Epidemic arrival date and size at the district level.** a) Effect of the distance from the epicentre on the number of days until the epidemic begins in a district. The color indicates the population density in the district (number of inhabitants divided by the district's surface). b) District final epidemic size as a function of the characteristic distance between two individuals normalised by the average dispersal distance. The later is computed as $2E[X]\sqrt{\text{density}}$, where $X$ is the log-normal distribution of daily individual covered distance. Both panels show the value for 35,234 French districts and 100 stochastic simulations.

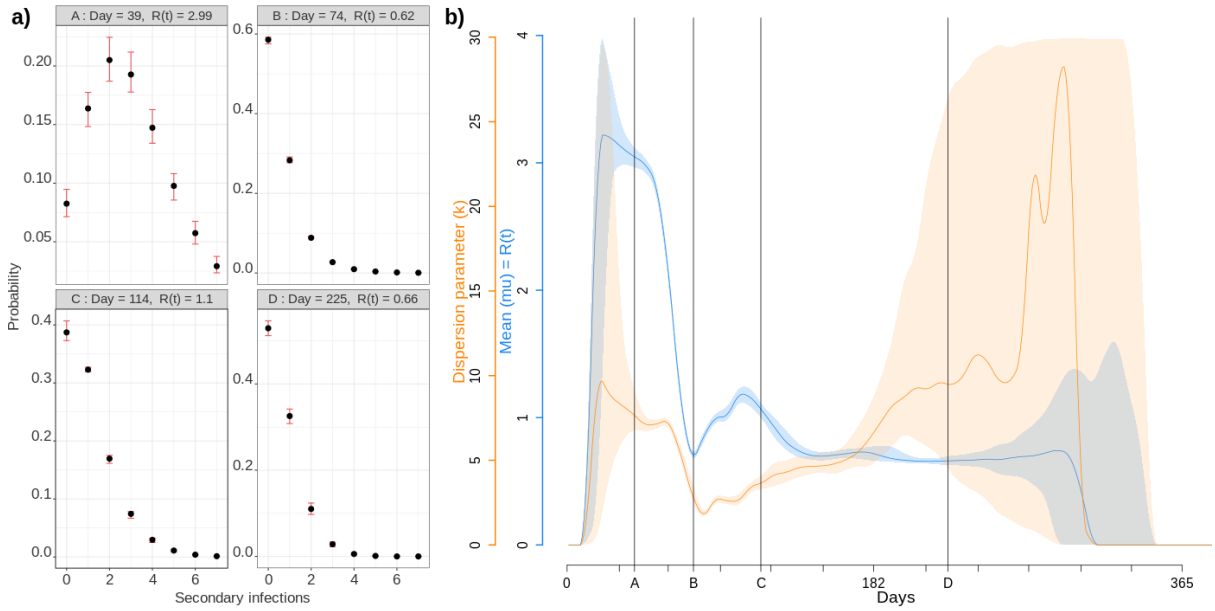


Figure 4: **Individual reproduction number dynamics.** a) Distribution obtained over the whole population on 4 different days post outbreak. b) Daily variation of the mean (in blue) and dispersion (in orange) of the distribution of individual reproduction numbers, which is assumed to follow a Negative Binomial distribution (*20*). Shaded areas show the 95% CI.