# Forecasting COVID-19 Infections With the Use of Simple Cellular Automata

EFTHIMIS TSILIONIS, Institute of Informatics & Telecommunications, NCSR "Demokritos", Greece
ALEXANDER ARTIKIS, Institute of Informatics & Telecommunications, NCSR "Demokritos", Greece and Department of Maritime Studies, University of Piraeus, Greece
GEORGIOS PALIOURAS, Institute of Informatics & Telecommunications, NCSR "Demokritos", Greece

The outbreak of the COVID-19 pandemic led policy makers and public health officials around the world to implement non-pharmaceutical interventions to suppress the spread of the virus. The goal was to reduce human mobility and social contacts, considered as the main factors of virus diffusion. However, these containment measures needed to be revised on a frequent basis to avoid serious economic and social costs. Moreover, spatial disparities needed to be taken into consideration, since the behavior of the virus was different according to the geographical context. Therefore, a frequent update on the short-term forecasts of the epidemic course, which considered spatial heterogeneities, was crucial for planning appropriate mitigation strategies. In this paper, we present a simple epidemiological model based on Cellular Automata, that takes into account human mobility and produces short-term forecasts of daily virus infections. Cellular Automata allow the discretization of time and space and thus, the spatio-temporal dynamics of the disease can be explored at the desired scale. We apply our model on real daily infection and mobility data from Spain and show that it is reliable in predicting the short-term daily infections trajectory both at the country level, as well as at the regional level of Autonomous Communities. Furthermore, compared against four state-of-the-art methods, the proposed method achieves comparable forecasting performance with significantly lower computational resources.

CCS Concepts: • **Information systems** → **Spatial-temporal systems**; • **Networks** → *Network simulations*; • **Mathematics of computing** → Time series analysis; • **Computer systems organization** → Real-time system architecture.

Additional Key Words and Phrases: epidemic modeling, time-series forecasting, mobility patterns, COVID-19

**ACM Reference Format:**
Efthimis Tsilionis, Alexander Artikis, and Georgios Paliouras. 2018. Forecasting COVID-19 Infections With the Use of Simple Cellular Automata. *J. ACM* 37, 4, Article 111 (August 2018), 31 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

The coronavirus, SARS-CoV-2, was first identified in Wuhan, China, in 2019, and since then it spread in many countries all over the world. In March 2020 the World Health Organization (WHO)[1] declared the COVID-19 disease a pandemic. COVID-19 is characterized by high transmissibility rates and can lead to severe symptoms and even death. This resulted in high numbers of hospitalizations and a high pressure on public health systems. In order to control the rapid spread of

[1]https://www.who.int/europe/emergencies/situations/covid-19

Authors' Contact Information: Efthimis Tsilionis, eftsilio@iit.demokritos.gr, Institute of Informatics & Telecommunications, NCSR "Demokritos", Athens, Greece; Alexander Artikis, a.artikis@unipi.gr, Institute of Informatics & Telecommunications, NCSR "Demokritos", Athens, Greece and Department of Maritime Studies, University of Piraeus, Piraeus, Greece; Georgios Paliouras, paliourg@iit.demokritos.gr, Institute of Informatics & Telecommunications, NCSR "Demokritos", Athens, Greece.


Manuscript submitted to ACM

the virus, public officials implemented multiple non-pharmaceutical interventions (NPIs), including closure of schools and shops, travel restrictions and even national lockdowns [19, 52]. These NPIs varied across countries and aimed to reduce human activity and social contacts, considered as the main factors of virus diffusion [4, 29, 49]. However, even though the measures taken had a significant effect on decreasing the daily new infections, they also had a serious impact on the economy and great social costs [9, 46]. Therefore, policymakers and health officials needed to impose restrictions with great caution and update the intervention policies on a frequent basis.

To support efficient decision making, officials rely on several indicators that can capture the course of an epidemic. The *daily case counts* is one of these indicators. To predict future infections scientists rely on epidemiological models. The most popular ones are *compartmental models*, which distinguish the population under investigation in one of the possible states (see [8, 10] for two surveys). With the use of appropriate differential equations, these models express the rate of change of individuals in each of the possible states. However, compartmental models assume a homogeneous and well-mixed population, meaning that all individuals are equally likely to have contacts. As a result, such models cannot capture the spatial disparities of the disease, such as population density in a specific area, or the mobility networks that strongly influence the social contacts. These characteristics are essential in forecasting the temporal and spatial evolution of the virus [44].

Improved versions of compartmental models add more compartments-states, in order to take into consideration specific aspects of a virus, such as incubation period, latency period, hospitalization, etc. [31]. However, this leads to a significant increase in the state space and the necessary variables and as a consequence increased computational complexity. Furthermore, such models are more complex and more difficult to interpret [36].

The need to consider the spatial distribution of the population, as well as the mobility patterns, led to the development of *metapopulation* networks. In a metapopulation network, agents occupy different patches which are connected through links. Agents are allowed to travel to other patches if there is a connection link with their residence patch. Therefore, the disease propagation can be examined not only inside a patch but also across patches. Metapopulation models highlight the spatial evolution of the virus through the mobility of agents [1].

Taking this idea of spatio-temporal modelling one step further, we propose the use of simple cellular automata to forecast daily virus infections. Cellular automata discretize naturally time and space and have been applied successfully in predicting the spread of viruses in the past [12, 51]. Each cell of the automaton corresponds to a specific geographical area, while special functions are used to model the transition of each cell-area from one state to another. Our method utilizes a simple transition function that takes into account the mobility of people and tracks the infection trajectory of each cell-area. We use mobile-phone geolocation data, in order to capture human mobility within and across cells. We perform forecasts with a horizon of 14 days, over a sliding window, and show that the proposed method can accurately predict observed case counts. The model offers quick detection of infections and facilitates decision-making by allowing public health officials to reconsider the containment policies already applied and suggest new interventions.

In particular, the contributions of this paper are the following:

- We present a model based on cellular automata that takes into account human mobility and makes accurate forecasts of virus infections.
- Our model is simple, easily interpretable, and favors quick detection of the course of the epidemic, allowing healthcare officials to obtain a quick overview of the virus spread.
- We evaluate our model using real mobility and COVID-19 infection data from Spain and compare it against four state-of-the-art methods, i.e., Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory

(LSTM) neural network, Graph Neural Network (GNN), and Microscopic Markov Chain Approach (MMCA). The experiments show that the proposed method achieves very good performance overall and especially in short-term forecasts.

- Additionally, its processing time is orders of magnitude lower than that of the other methods.

The structure of the paper is as follows: Section 2 summarizes epidemic models and cellular automata. Section 3 elaborates on the details of the presented method, while Section 4 presents our empirical analysis. Section 5 summarizes our work and Section 6 discusses related work. Finally, the article concludes in Section 7, where we outline potential future directions.

## 2 Background

In this section we introduce basic concepts of epidemiology and present the different approaches in epidemiological modeling.

### 2.1 Basic Concepts of Epidemics

In epidemiology the main concern is the study and understanding of infectious diseases, in order to control them and prevent the wide spread in the population. To achieve these goals, epidemiologists rely on factors that permit them to quantify the various aspects of an infectious disease. The population is usually divided into subcategories, according to an individual's relation with a disease. *Susceptible* population refers to those individuals that have not been infected or more broadly they have not developed immunity against the virus and thus can acquire the virus. *Infected* individuals are those who are infected and therefore can transmit the virus to susceptible individuals. Depending on the virus under investigation the infected individuals can be further divided, taking into account whether they exhibit symptoms or not and whether they are considered infectious. When the infectious period is over, the infected individual has either recovered or passed away. A *recovered* individual may possess immunity forever or for a certain period of time, depending on the virus. When the immunity period ends, the individual becomes again susceptible.

The most fundamental epidemiological quantity scientists try to monitor is the *effective reproduction number*, $\mathcal{R}$ [16, 27, 47]. $\mathcal{R}$ represents the average number of infections caused by an infected individual, during the course of that individual's infectious period. This number informs about the transmissibility of the virus and is used to control the epidemic. According to the value of $\mathcal{R}$, and more specifically when $\mathcal{R} > 1$, scientists propose interventions and containment measures to set the epidemic under control. The goal of the measures is an $\mathcal{R}$ that is below 1 and close to 0.

The calculation of the effective reproduction number at specific points in time is called instantaneous reproductive number, $\mathcal{R}^{(t)}$, and is usually computed in two ways [27]. The basic calculation of $\mathcal{R}^{(t)}$ is as follows:

$$\mathcal{R}^{(t)} = \beta^{(t)} \frac{S^{(t)}}{N} \tau \tag{1}$$

where $\beta^{(t)}$ is the transmission rate at time $t$, i.e., the probability of a susceptible individual to become infected after a contact with an infected individual, $S^{(t)}$ is the number of susceptible individuals at time $t$, $N$ the size of the population, and $\tau$ the mean duration of the infectious period. An enhanced variant of this calculation takes into consideration the observed *infected cases* and is the following [16]:

$$\mathcal{R}^{(t)} = \frac{I^{(t)}}{\sum_{s=1}^{t} I^{(t-s)} w_s} \quad , \tag{2}$$

where $I^{(t)}$ is the number of infected individuals at time $t$ and $w_s$ is the *infectivity profile*. The infectivity profile $w_s$ is a probability distribution that expresses how likely an individual is to be still infectious. The infectivity profile $w_s$ may be approximated by the distribution of the *generation interval*, i.e., the time between the infection of a primary case and the infection of a secondary case. However, time of infection is hard or even impossible to observe with accuracy, since a new disease case is usually reported after the onset of symptoms, which differs from the time a virus was acquired. Therefore, instead of the generation interval, the *serial interval* is preferred. The serial interval denotes the time between the onset of symptoms of a primary case and the onset of symptoms of a secondary case. The serial interval can be computed by the daily observed counts of a disease. It is usually assumed to follow a gamma distribution and it provides an approximation of the infectivity profile $w_s$. The interested reader can refer to [27] and [16] for more information about the effective reproduction number and the difficulties in computing it.

### 2.2 Compartmental Models

The most common models used when studying epidemics are the so called compartmental models [8]. They are mathematical models that divide the population under investigation into compartments or states. The goal is to calculate the degree of change of each class over time, captured by appropriate differential equations.

A popular compartmental model is the susceptible-infected-recovered (SIR) model [34]. The SIR model assumes a closed population, $N$, meaning that the population remains stable through time (i.e., number of births equals number of deaths). Each individual is assigned to one of three compartments at each time-point $t$: susceptible-$S^{(t)}$, infected-$I^{(t)}$ and recovered-$R^{(t)}$. Only two transitions are allowed, i.e., from $S$ to $I$ and from $I$ to $R$. The three states change dynamically over time and this is captured by the following three equations:

$$\begin{aligned} \frac{\partial S^{(t)}}{\partial t} &= -\beta \frac{S^{(t)}}{N} I^{(t)} \\ \frac{\partial I^{(t)}}{\partial t} &= \beta \frac{S^{(t)}}{N} I^{(t)} - \gamma I^{(t)} \\ \frac{\partial R^{(t)}}{\partial t} &= \gamma I^{(t)} \end{aligned} \tag{3}$$

where $\beta$ is the transmission rate, i.e., the probability of a susceptible individual to get infected, and $\gamma$ is the recovery rate. Note that since a closed population is assumed, $S^{(t)} + I^{(t)} + R^{(t)} = N$.

The family of compartmental models includes many variations of the SIR model that consider factors such as the exposure to a virus, temporary immunity and as a consequence reinfection, therapeutic measures such as hospitalization, vaccination etc [10]. The different models differ mainly in the number of compartments. These models are simple and explainable, providing a straightforward solution to study the spread of a disease. However, compartmental models assume homogeneous mixing of the population, meaning that each individual has the same probability of meeting and transmitting the disease to any other individual of the population. This simplifying assumption may be suitable for small communities, where the local propagation of a disease is being considered. At larger scales, such as national or global, additional factors should be taken into account to describe and analyze the virus spread [1, 4, 13].

### 2.3 Metapopulation Models

Metapopulation models have been developed to address the issues of spatial disparities observed during the spread of a virus in a geographically distributed heterogeneous population. A key aspect that these models incorporate is human mobility [1]. Traveling between locations is a determining factor that shapes the course of a disease. Infected

individuals that commute from their home location to another location (e.g. work) may act as infection seeds at the latter location [44]. Therefore, mobility plays a significant role in the evolution of a disease.

Metapopulation models assume a network of interconnected locations. This network may be seen as a graph, where each location is a vertex and the edges represent the connections between locations. The locations are usually called patches. Individuals may travel to any other patch different from their home-patch (if a connection edge exists) or move inside their home-patch. Usually, the mobility patterns are represented with a transfer matrix, expressing the probabilities of commuting from one patch to another. Furthermore, this class of models employ a compartmental model in each patch, where a closed population may be assumed, and describe the temporal evolution of each state with differential equations.

Other aspects that this class of models incorporate are social and demographic, such as economic or age-related details that influence an epidemic [2, 13, 28]. These models may be either deterministic or stochastic and are considered the most appropriate ones for capturing the dynamics of a disease at large geographical spaces [2]. The incorporation of the various parameters and the segmentation of space lead to increased computational complexity and make these models unsuitable in situations where predictions concerning the epidemic spread are needed in a timely manner [36].

### 2.4 Cellular Automata

Cellular automata (CAs) are dynamical systems that have been applied to the study of disease evolution, mainly due to their simplicity and ability to mimic to an extent biological and physical processes [5]. CAs comprise of cells that form a $D$-dimensional lattice, where usually each cell is a finite state automaton. CAs permit the discretization of time and space. The temporal aspect of a CA is captured by the transition of a cell from one state to the next at a predefined time-step. The transition is governed by a local *update function* that takes into consideration the cell's current state and the state of its neighbors. The interaction of a cell with its neighbors to decide its next state models the spatial dynamics of the problem under investigation. The states of the cells of a CA at each time-point define the *configuration of the CA*. Evaluating a CA many times and monitoring the states of the cells allow to observe the physical evolution of a system and understand its behavior.

This characteristic has made CAs very attractive for modeling the spread of infectious diseases [51]. In epidemic studies, CAs are typically used to study the spreading of a disease in a particular geographical area. The area may vary from a country to a small residential area. Each cell may refer to a single individual or to the population of an area, as is the case with the patches used in the metapopulation models (see Section 2.3). An epidemiological model is then applied to each cell, in order to determine the state of the cell. Consider for example, that a SIR model is used in each cell of CA and that each cell represents an individual. Then, the possible states of each cell are $S$, $I$ and $R$ and, according to the update function, an individual transits sequentially from one state to the other. This implies that the state space is discrete, which is true most of the times. However, the state of a cell can also be defined in terms of continuous values. Furthermore, the CAs usually found in the literature are uniform, meaning that all cells employ the same update function.

In addition to the transition function, an important factor that dictates state transitions is the neighborhood of a cell. The neighborhood refers to the cells, with which the target cell can interact. If a cell is not included in the neighbors of the target cell, it cannot influence the transition of the target cell directly but only indirectly (e.g., by affecting a neighbor of the target cell). The definition of a neighborhood is thus very important and depends on the relationship between cells, e.g. geographical proximity. The most popular neighborhoods are the Von Neumann and Moore.

Next, we provide a formal definition of a CA. A Cellular automaton (CA) is a quadruple $(\mathcal{L}, \mathcal{B}, V, f)$, where:

- $\mathcal{L} \subset \mathbb{Z}^D$ is the set of cells forming the $D$-dimensional cellular space. An element of $\mathcal{L}$ corresponds to a $D$-tuple specifying the coordinates of a cell in the $D$-dimensional grid.
- $\mathcal{B}$ is the set of states. A cell at each time-point may be in one of the possible states. Usually, $\mathcal{B}$ consists of discrete values but can also take continuous values. In the former case, the CA is called *discrete*, otherwise is called *continuous*.
- $V(c) = \{c, v_1(c), v_2(c), \ldots, v_{M-1}(c)\}$ is the neighborhood vector of a cell $c$, consisting of $M$ neighbors, including $c$ itself. The value of $M$ is predefined, it may differ for each cell, and each function $v_i : \mathbb{Z}^D \to \mathbb{Z}^D$ returns a neighbor of the target cell $c$.
- $f : \mathcal{B}^{M+1} \to \mathcal{B}$ is the local update function or transition function of the automaton. This rule determines the new state of a cell by taking into consideration the cell's current state and the states of the cell's neighbors. If $b_c^t, b_{v_1(c)}^t, \ldots, b_{v_{M-1}(c)}^t$ are the states of a cell $c$ and its neighbors at time $t$, then the next state of cell $c$ at time $t+1$ is given by $b_c^{t+1} = f(b_c^t, b_{v_1(c)}^t, \ldots, b_{v_{M-1}(c)}^t)$. When the same function is used by all the cells of the CA the CA is called uniform.

A global configuration $\mathcal{B}^{(t)}$ of a CA corresponds to the states of all cells at each time-point $t$, i.e. $\mathcal{B}^{(t)} = \{b_c^t \mid c \in \mathcal{L}\}$. To compute the configuration of the CA at the next time-point $t+1$ we apply the update function to each cell. Therefore, a CA can be seen as a function $G : \mathcal{B}^{(t)} \to \mathcal{B}^{(t+1)}$, where $\mathcal{B}^{(t+1)} = G(\mathcal{B}^{(t)}) = \{f(b_c^t, b_{v_1(c)}^t, \ldots, b_{v_{M-1}(c)}^t) \mid c \in \mathcal{L}\}$. Computing $k$ configurations of a CA means to apply iteratively $k$ times the function $G$, i.e., $G(\mathcal{B}^{(t)}) \mapsto G^2(\mathcal{B}^{(t)}) \mapsto \cdots \mapsto G^k(\mathcal{B}^{(t)})$.

## 3 Proposed Method

In this section, we elaborate on our proposed CA-based approach for the forecasting of daily virus infections.

### 3.1 Cell Characteristics

We use a fixed finite 2-dimensional grid where each cell corresponds to one of the $M$ geographical divisions of a country. In contrast to other studies [17, 25, 42], we choose a cell to represent a whole region instead of a single individual. We employ a *SIR* mathematical epidemiological model on each cell. Each cell maintains variables $S_c^{(t)}$ and $I_c^{(t)}$, which represent respectively the number of susceptible and infected individuals at time-point $t$ at cell $c$. We do not maintain a variable $R_c^{(t)}$ that will keep track of the recovered individuals. However, after a certain period of time we assume that infected individuals exit the infected state and enter the recovered state. Our goal is to compute the number of newly infected people at future time-points at each cell. We assume a homogeneous population at each cell, $N_c$, meaning that the population of each cell remains constant over time. Furthermore, we assume a homogeneous mixing of the population inside each cell, meaning that an individual has the same probability of meeting any other individual inside the cell.

The neighborhood of a cell $c$, regardless of its position on the lattice, includes all the other cells. Thus, $V(c) = \{c, v_1(c), v_2(c), \ldots, v_{M-1}(c)\}$, where $|V(c)| = M$. A cell has interactions with all other cells and this allows us to use mobility data and consider the impact of visitors from other cells to the spread of infections in the target cell.

### 3.2 Cell's Transition Function

As mentioned above, each cell of the CA represents a geographical area. In order to produce forecasts of infections for each cell, we use a transition function that incorporates the movement of people across areas, as well as the movement

inside an area. Additionally, we consider infections that happen without any movement, e.g. household infections. To distinguish the two different aforementioned sources of infection, we use the average number of contacts at each source.

The transition function of a cell informs about the newly infected people at the next time-point. In particular, the transition function of a cell $c$ is the following:

$$\hat{I}_c^{(t+1)} = \underbrace{\frac{S_c^{(t)}}{N_c} \frac{\overline{\beta_c}}{k_m} \sum_{j=1}^{M} \frac{\overline{I_j^{(t)}}}{N_j} m_{jc}^{(t)}}_{\text{Infections caused from mobility}} + \underbrace{\frac{S_c^{(t)}}{N_c} \frac{\overline{\beta_c}}{k_h} \overline{I_c^{(t)}}}_{\text{Infections not caused from mobility}}, \tag{4}$$

where $\hat{I}_c^{(t+1)}$ is the prediction of new infections in cell $c$ at time $t+1$ and $I_c^{(t)}$ is the actual number of newly infected individuals in cell $c$ at time $t$. $\overline{I_c^{(t)}}$ represents all the active cases of a disease in cell $c$ at time $t$. As active cases in a cell we consider the people reported infected in the last $\omega$ time-points, and thus $\overline{I_c^{(t)}} = \sum_{s=0}^{\omega-1} I_c^{(t-s)}$. $\omega$ is the size of a temporal window with a start point at time $t-\omega+1$ and an end point at time $t$. $\omega$ denotes the number of past values to consider. After $\omega$ time-points of infection, we assume that an individual cannot spread the virus any more and is considered recovered. $S_c^{(t)}$ is the number of susceptible individuals in cell $c$ at time $t$. Since a susceptible person is someone who can become infected, we calculate at each time-point $t$, $S_c^{(t)} = N_c - \sum_{s=0}^{t} I_i^{(t-s)}$. Hence individuals that get infected are removed from the susceptible population. $m_{jc}^{(t)}$ denotes the number of trips performed from cell $j$ to cell $c$. Multiplication by the fraction of the population at cell $j$ that is infected, i.e., $\frac{\overline{I_j^{(t)}}}{N_j}$ provides an approximation of the infected people that traveled to cell $c$ and can transmit the virus.

Eq. (4) comprises two terms that account for infections caused due to people moving or not. $\overline{\beta_c}$ represents the average transmission probability of the last $\omega$ time-points at cell $c$ and thus, $\overline{\beta_c} = \frac{\sum_{s=0}^{\omega-1} \beta_c^{(t-s)}}{\omega}$, where $\beta_c^{(t)}$ is the transmission rate at cell $c$ at time $t$. The variables $k_m$ and $k_h$ are parameters that differentiate the transmission probability according to the source. $k_m$ represents the average number of contacts a moving person makes during a day, regardless of the location (indoors or outdoors) these contacts take place, excluding the ones taking place at home. On the other hand, $k_h$ represents the average number of contacts a person makes when staying at home. The two terms of Eq. (4) differ only in the contact parameters and the number of infected people that can transmit the virus (visitors or locals). A graphical representation of the components of Eq. (4) is shown in Figure 1.

Eq. (4) provides a prediction of the newly infected people at the next time-point. The variables in the right-hand side of Eq. (4) are computed from actual data. In order to produce forecasts further ahead in the future, some of the variables in Eq. (4) can be estimated using the predictions of the previous time-points. For example, assume we want to compute $\hat{I}_c^{(t+2)}$. $\overline{I_c^{(t+1)}}$ that estimates the active cases of the last $\omega$ time-points will be equal to $\sum_{s=1}^{\omega-1} I_c^{(t-s)} + \hat{I}_c^{(t+1)}$. Similarly, $S_c^{(t+1)} = S_c^{(t)} - \hat{I}_c^{(t+1)}$. The same process is followed for predictions in time-points that are further to the future. All such predictions together constitute a *forecast operation*.

The forecast operation is displayed graphically in Figure 2. $f_h$ denotes the *forecast horizon*, i.e. the number of predictions in future time-points we will produce, and $\omega$ is the temporal window denoting the number of past values (actual or predicted) to be considered in the calculation of each forecast. For visualization purposes in Figure 2, $f_h$ and $\omega$ are initially set to 3 time-points. Each row in the diagram of Figure 2 corresponds to a forecast at a different time-point. The top row generates a forecast for time-point $t+1$ ($\hat{I}_c^{(t+1)}$), the middle row for time-point $t+2$ ($\hat{I}_c^{(t+2)}$) and the bottom row for time-point $t+3$ ($\hat{I}_c^{(t+3)}$). Notice that in the forecast for time-point $t+1$, $\omega$ contains only actual values of infected

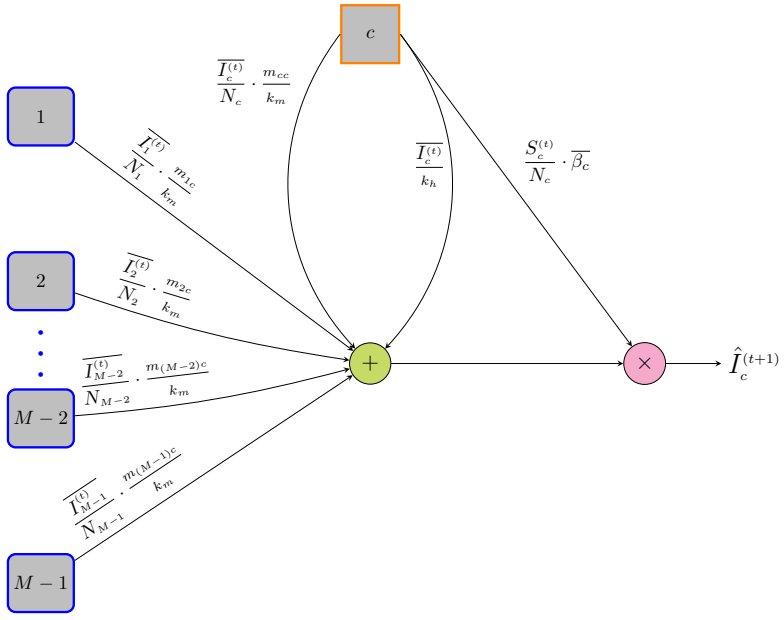



Fig. 1. Graphical illustration of the cell's transition function (Eq. (4)). The rectangle with the orange border represents the target cell $c$ at time $t$. The blue border rectangles represent the remaining $M-1$ cells of the lattice at time $t$. The green circle denotes summation, while the pink circle multiplication.

cases, while actual and predicted values are used for the forecasts of time-points $t+2$ and $t+3$. For the forecasts at times $t+2$ and $t+3$, $\omega$ progresses one time-point each time. In contrast, the forecast horizon $f_h$ does not slide and becomes smaller at each step of the forecast operation.

When a forecast operation is complete, the observations of the next time-point ($t+1$ in our example) become available and a new forecast operation can start. To achieve this, the window $\omega$ is positioned to start one time-point later than the start of the previous forecast operation ($t-2$ in our example). This process is illustrated in Figure 3, where two consecutive forecast operations are shown. Each row of Figure 3 displays the start of a different forecast operation. Notice that in both rows window $\omega$ includes only actual values since only the start of a forecast operation is depicted. Moreover, between consecutive forecast operations the forecast horizon $f_h$ also moves by one time-point.

$\beta_c$ is updated at the end of each forecast operation, i.e., when sliding the window $\omega$ to the next time-point. On the other hand, $m_{jc}^{(t)}$ is always based on real mobility data, as these can be known to some extent ahead of time. In order to update the transmission rate $\beta_c$, we use the instantaneous reproduction number $\mathcal{R}^{(t)}$. Consider again Eq. (1) and Eq. (2) presented in Section 2.1. Notice that these equations can be used to calculate the effective reproduction number for each one of the cells, i.e, $\mathcal{R}_c^{(t)}$, using real data. By equating Eq. (1) and Eq. (2), we can calculate the transmission rate in a cell $c$ at time $t$ as follows:

$$\beta_c^{(t)} = \frac{I_c^{(t)} N_c}{\tau\, S_c^{(t)} \sum_{s=0}^{\omega-1} I_c^{(t-s)} w_s} \quad . \tag{5}$$

Recall that $\tau$ is the mean infectious period and is considered a constant; $w_s$ is the infectivity profile and is approximated by a predefined gamma distribution. Notice also that in the denominator of Eq. (5) we consider the infected individuals

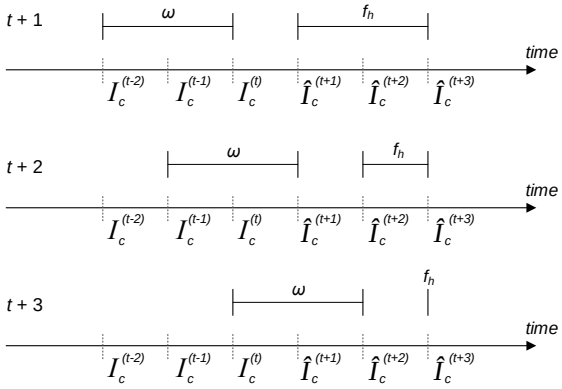



Fig. 2. The forecast operation. The sliding window $\omega$ and the forecast horizon $f_h$ are both initially set to 3 time-points. $\hat{I}_c^{(t)}$ represents predictions, while $I_c^{(t)}$ represents actual data. Each row in the diagram corresponds to a forecast at a different time-point.

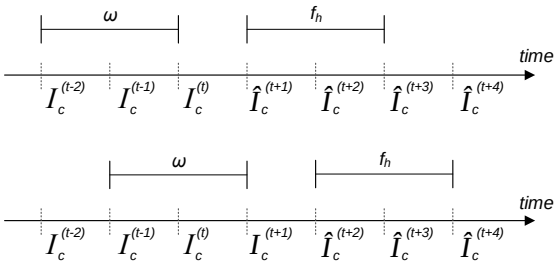



Fig. 3. Consecutive forecast operations. The size of the sliding window $\omega$ and the forecast horizon $f_h$ are both set to 3 time-points. $\hat{I}_c^{(t)}$ represents predictions, while $I_c^{(t)}$ represents actual data. Each row in the diagram shows the start of a different forecast operation.

of the previous $\omega$ time-points. Eq. (5) computes the transmission rate of each cell of the CA at time $t$. This way we can capture differences in the spread of the virus across the cells. By taking the average of the last $\omega$ time-points of $\beta_c^{(t)}$ we also obtain $\overline{\beta_c}$.

## 4 Experimental Evaluation

We present the experimental evaluation of our method and its performance regarding the prediction of COVID-19 infections. Furthermore, we provide a comparison to state-of-the-art systems and show the efficacy of the proposed model.

### 4.1 Application of the CA-based approach to COVID-19

The proposed method has been assessed on the task of predicting the course of the COVID-19 epidemic. At each time-point, we predict the number of new cases in each one of the geographical divisions of a country. The different areas may differ both in geographical area size and population density. Nevertheless, for simplicity we choose to assign each area to a single cell.

As a time-point we use a calendar day and so the predictions refer to daily new case counts. Our main goal is to aid health officials to have a quick overview of the course of the epidemic in the short-term future. In this manner, the model is easily interpretable and can provide updated information once new data are available. The number of forecasts produced depends on the horizon $f_h$ of a forecast operation, i.e., the number of predictions in future time-points computed. In addition to the forecast horizon, a sliding temporal window $\omega$ is used, as explained in 3.2. In our experiments, we employ a forecast horizon of 14 days, $f_h = 14$, and a sliding window of 7 days, $\omega = 7$. We use $f_h > \omega$ in order to examine the behavior of the CA-based approach both in short-term and long-term forecasting.

### 4.2 Data Description and Integration

The data used in the present study come from various sources and are integrated to produce predictions of daily COVID-19 cases. We combine data that concern daily confirmed COVID-19 cases, mobility records, as well as population counts at the level of Spanish provinces. All the data are open-access and can be downloaded through an Application Programming Interface (API) [40]. In the following paragraphs we provide some information on the different types of data used. The interested reader can refer to [40] for more detailed information on the collection, processing and consolidation of the data, as well as different ways to download them.

The COVID-19 records we use in the present study include daily cases for Spain, reported at different levels of spatial resolution, such as autonomous communities and provinces. Additional information, such as accumulated incidence, number of cases per 100,000 inhabitants, hospitalizations etc. is also provided. We utilize only the daily reported new cases of COVID-19 for each province of Spain. The authors in [40] mention that there may be a bias on the incidences concerning weekends, since these are reported on Mondays for both autonomous communities and provinces. They suggest the application of rolling mean average windows, which we also adopt here. Each province of Spain is matched to a cell of the Cellular Automaton (CA). The records of a single day concerning a province/cell of the CA are matched to the variable $I_c^{(t)}$ of Eq. (4). The provinces of Spain are 52 and thus, $M = 52$ in Eq. (4).

Mobility data records are based on Anonymized Mobile (cell) Phone Data (AMPD), evenly distributed across Spain and coming from a single mobile operator. These records are reported by the Spanish Ministry of Transport, Mobility and Urban Agenda[2] (MITMA, *Ministerio de Transportes, Movilidad y Agenda Urbana*). One of the indicators MITMA provides is the number of trips performed from an origin to a destination zone with hourly resolution. A *trip* is defined as any movement of more than 500 meters that lasts more than 20 minutes. Notice that if an individual takes more than one trip per day, these are included as different entries in the set of trips.

Furthermore, the spatial resolution used by MITMA is much higher than the one we use in the present study. Mobility zones used by MITMA are much smaller areas than provinces. This spatial disparity is resolved by a method proposed in [40]. The authors have implemented an approach to project data among different geographical layers, which is based on linear interpolation. The data projection can be based either on the spatial ratios between the areas or on the distribution of the population. The population-based method is considered more reliable and we opt for this approach here. Using this data projection technique in [40], they construct Origin-Destination (OD) matrices at different levels of spatial and temporal resolution. We select the matrices that refer to trips between provinces on a daily basis. In order to produce daily trips, the authors aggregate the hourly trips reported by MITMA. The mobility data of a specific day are matched to the variable $m_{jc}^{(t)}$ of Eq. (4), which represents the trips from province $j$ to province $c$ at time $t$.

[2]https://www.transportes.gob.es/ministerio/covid-19/evolucion-movilidad-big-data/movilidad-nacional

Table 1. Constant epidemic parameters of the model

| Symbol | Description | Value |
|---|---|---|
| $k_m$ | Average number of contacts outside the house | 10.1 |
| $k_h$ | Average number of contacts at home | 3.2 |
| $\tau$ | Mean infectious period duration in days | 4 |

Population data are estimated based on another MITMA indicator that provides the population in each MITMA mobility zone on a daily basis. The authors in [40] have aggregated the population at the province level and use this estimation to capture the population fluctuations in different regions over the year (e.g. winter vs summer). Their estimations have been validated with the population values reported in the Spanish census of 2019. Even though they provide daily population estimates, we choose to use a constant population for each province (recall variable $N_c$ from Eq. (4)). Thus, for each province we consider a population estimate concerning the period of winter, where population fluctuations are rarely observed, and use this value throughout the whole experimental evaluation.

In the definitions of the transition function (Eq. (4)) and the transmission rate (Eq. (5)), we have introduced some variables that do not depend on real data and are handled as constants. These are $k_m$, $k_h$, $\tau$ and $w_s$, representing the average number of a person's contacts outside the house, the average number of contacts at home, the duration of the mean infectious period and the infectivity profile, respectively. Table 1 summarizes these constants. Their values, except $w_s$, are taken from the study presented in [2], which is one of the competing methods that we use, and are either averaged or rounded. Specifically, $k_m$ is the result of subtracting the "Average number of contacts at home" variable from the "Average total number of contacts" variable for adults. $k_h$ is the "Average number of contacts at home" of adults. $\tau$ arises after rounding the symptomatic infectious period of adults.

Finally, $w_s$ is approximated by a gamma distribution and represents the serial interval distribution mentioned in Section 2.1. To construct $w_s$, we follow an approach that considers uncertainty in the infectivity profile, similar to the one presented in [16] . First, we constrain the mean, $\mu_{w_s}$, and standard deviation, $\sigma_{w_s}$, of the gamma-distributed serial interval to follow normal distributions. Then, we sample $n = 1000$ pairs of $\mu_{w_s}$ and $\sigma_{w_s}$ resulting in: $(\mu_{w_s}, \sigma_{w_s})^1, \ldots, (\mu_{w_s}, \sigma_{w_s})^n$. For each pair, we apply the constraint that $\mu_{w_s} < \sigma_{w_s}$ to ensure that the probability density function is null at $t = 0$. Next, we construct a discrete gamma distribution for each pair of mean and standard deviation (1000 distributions in total). For each distribution, we apply our CA approach for a warm-up period, ranging from 22-2-2020 to 31-3-2020 and select the distribution that minimizes RMSE. The resulting discrete gamma distribution represents the serial interval distribution and it remains unchanged throughout the whole evaluation period. Specifically for the COVID-19 disease we used an average mean serial interval of 4 days (sd 1.5, min 1, max 7), and an average standard deviation of 1.5 days (sd 0.5, min 0.5, max 2.5).

### 4.3 Competing Methods

To demonstrate the effectiveness of our approach we compare it against methods coming from different fields of epidemiological modeling. The first one is the Autoregressive Integrated Moving Average (ARIMA), a classical time-series forecasting method that is data-driven and independent of mobility interactions. The second one is the long short-term memory (LSTM) neural network [30], which due to its memory capacity is able to capture time dependencies in the data. The third one is a graph neural network (GNN) [35], which, through the message passing scheme, updates the representation of each node of the graph according to the messages received from its adjacent neighbors. The fourth

one is a metapopulation framework, based on a method known as Microscopic Markov Chain Approach (MMCA), that combines compartmental dynamics and mobility networks to assess the spread of the virus. The methods have been selected, among others, in order to assess the effect of using or not mobility data in disease forecasting. All models are applied on data concerning the epidemic wave in Spain.

*4.3.1 ARIMA.* ARIMA is a regression model denoted as ARIMA($p$,$d$,$q$), where $p$ is the order of the autoregressive part, $d$ is the degree of differencing applied and $q$ is the order of the moving average part. The order of the model ($p$) indicates the history of past values used in forecasting, i.e. how many past values of the variable will be used. Differencing is used in ARIMA to achieve stationarity, i.e. a time-series without trend or seasonality. Stationarity facilitates the forecasting process. The value $d$ denotes how many times we should subtract past values of the variable, to make the time-series stationary. For instance, a value of 1 means the difference between the current time period and the previous one. The last parameter, $q$, denotes the number of past forecast errors that should be taken into account in the forecast operation. We will refer to $p$,$d$ and $q$ as hyper-parameters of the ARIMA model.

In the context of the present study, we apply an ARIMA model to each one of the 52 provinces of Spain. Each ARIMA model is used to forecast the daily new COVID-19 infections inside a province. Initially, each model is trained with daily cases before the forecasting period, in order to tune the hyper-parameters, minimizing the Akaike Information Criterion (AIC). Once the hyper-parameters are defined, they remain unchanged during the whole forecasting process. The forecasting process follows the forecasting operation definition from Section 3.2, using a sliding window. First, a series of forecasts is performed according to the size of the forecast horizon $f_h$ defined. We use an $f_h$ of 14 days similarly to our method. Next, the sliding window progresses one time-point and the COVID-19 data are used to update each model. We employ a sliding window of 1 month.

*4.3.2 LSTM.* The LSTM framework [30] is a deep learning method that has been applied for forecasting the spread of COVID-19 [11, 22, 37, 41, 48, 50]. LSTMs are capable of learning long-term dependencies in the data and are suitable for time-series prediction problems [48]. In particular, LSTMs are a type of Recurrent Neural Network (RNN) that avoids the short-term memory problem, where short-term information has a greater influence over long-term information [37]. An LSTM unit (cell) is equipped with three gates (forget, input, output) that control the information flow inside the unit, by determining what information from the past should be kept, what should be updated, and what should be given as output [30]. This structure of LSTM units allows the discovery of important correlations in the data over arbitrary time intervals.

We adopt the architecture of the LSTM outlined in [50]. The authors of [50] have demonstrated the efficiency of this architecture by predicting the daily new cases in various countries. The LSTM network consists of four hidden layers and 130 hidden units. The learning rate is 0.005 and the optimizer is Adam. Similarly to ARIMA, an LSTM model has been applied to each one of the 52 provinces of Spain. However, the authors of [50] have used a large training period (194 days) to learn the parameters of the model and only a very small period (14 days) to test the model's prediction accuracy. In the prediction phase, the model is fed with the real cases of the previous day and, then, the testing period proceeds day-by-day. Similarly to [50], we normalize the daily case data using MinMaxScaler. An input to the model consists of the actual cases of the previous 7 days, i.e. the sliding window is $\omega$=7, as in our CA-based method. In the training period, the LSTM is trained for 300 epochs as suggested in [50]. Then, in the testing period we follow the forecasting operation discussed in Section 3.2, where a forecast horizon $f_h$ of 14 days is used. Note that during the forecast operation, the input may consist of actual and/or predicted values. After a 14-day forecast operation is performed, the window slides 1 day (new data become available) and a new 14-day forecast begins.

*4.3.3 GNN.* The GNN [26, 35] is a deep learning method that operates on graph data and is able to learn efficient representations of nodes by leveraging information from their neighbors. A graph consists of nodes, which are connected/related through edges. The edges may have weights. The graph is encoded by an adjacency matrix. The set of nodes with which a node is connected constitute its neighbors. The network contains a series of aggregation layers, called neighborhood layers. At each iteration, the node representations of a layer are updated by following a two-phase procedure. First, information is propagated along the neighbors, and second, the information is aggregated in order to obtain the updated representations. This procedure is called message passing [26]. GNNs have recently been used for spatio-temporal predictions of the spread of COVID-19 [21, 33, 38].

We follow the approach described in [38], which aimed to predict the daily new cases in different locations of several countries. The authors employ a GNN with two aggregation neighborhood layers, each layer comprising 64 hidden units. Additionally, they utilize skip connections from each layer to the output layer. The model is trained for a maximum of 300 epochs with an early stopping criterion after 100 epochs. The batch size is 8, the optimizer used is Adam and the learning rate is 0.001. The output of every layer goes through batch normalization and dropout, with a dropout ratio of 0.5.

To apply the model to the data from Spain, we constructed a graph for each day. The nodes of each graph are the 52 provinces of Spain. As node features we take the actual cases of the previous 7 days, i.e., $\omega$=7, similarly to our method. A node/province has an outgoing edge to another node/province if there are mobility data from the former to the latter at this date. Recall from Section 4.2 that the mobility data provide the number of trips from one province to another on each day. The number of trips is used as a weight on an edge. During training, the model is trained with the graphs of each day that constitute the observed history. Then, the forecasting operation discussed in Section 3.2 starts, with $f_h$=14. At each step of the forecasting operation, the input consists of the graph of the previous day, where the edges of the graph correspond to the actual mobility data and the node features include actual and/or predicted cases. When the forecast operation is completed, the window slides by 1 day, the model is trained over the whole history and a new series of 14 predictions begins. Our evaluation differs from the one in [38], in that their predictions always refer to a single time-point in the future while we employ the forecast operation and produce predictions for the next $f_h$ days. Thus, their input always consists of actual data, as opposed to the input used in the forecast operation that may include actual and/or predicted values.

*4.3.4 MMCA.* The MMCA model [2] is a complex framework that utilizes a multitude of variables and parameters that incorporate not only the specific characteristics of SARS-CoV-2 transmission but also the particularities of the population affected by the pandemic. Examples of such parameters of the model are the demographic distribution, the complex patterns of social contacts, and the geographic mobility networks. Furthermore, this model can estimate the impact of different non-pharmacological interventions (NPIs) on the trajectory of a disease.

The MMCA model is a metapopulation framework, where the population is distributed across patches. In each patch, an epidemiological model with ten compartments is built to capture the epidemiological and clinical status of the individuals. The authors introduce an asymptomatic state (infectious individuals without symptoms) to emphasize the covert effect of asymptomatic individuals on the virus spread. The model also examines the fatalities and thus the clinical compartments refer to hospitalization. The compartmental dynamics are expressed through equations that represent the probability of individuals at a patch to be in a specific state. These equations also indicate how these probabilities evolve over time.

Additionally, the population inside each patch is split into three age strata (young, adult, and elderly) to address the significant differences observed in the behavior of the virus across different age groups. Different age groups also have different commuting habits and this is reflected on their social contacts. The mobility of population across and within patches is considered to be the main reason of transmission of the virus. The model uses mobility matrices that denote the probability of each age stratum to move from one patch to another. This way, mobility is incorporated and its impact can be investigated. Additionally, the implementation of containment measures is modeled by applying reductions in the mobility of the population.

In the work presented here, MMCA was employed to predict new daily COVID-19 infections. The MITMA mobility zones (Section 4.2) represent the patches of the metapopulation framework. We have executed 1000 runs of the model, each time altering the values of the following parameters:

- Infectivity of symptomatic $\beta_I$
- Infectivity of asymptomatic $\beta_A$
- Exposed rate $\eta^g$
- Asymptomatic rate $\alpha^g$
- Infectious rate $\mu^g$
- Household permeability $\phi$
- Social distancing $\delta$
- Initial asymptomatic population $A_o$

In [2] there is a detailed description of the above parameters, as well as the ones that remain unchanged through all model runs, e.g., the hospitalization rates of each age group.

Having executed 1000 runs, we effectively get 1000 different models and one of them is selected to make a forecast each time. The selection is done with the use of a 1-week rolling window, similar to the sliding temporal window $\omega$ presented in Section 3.2. Let $\hat{I}_i^{(t+1)}$ represent the forecast of model $i$, $1 \leq i \leq 1000$, at time-point $t+1$. The model $i$ achieving the lowest RMSE in the forecasts of the seven days preceding time-point $t+1$ is selected to forecast $\hat{I}^{(t+1)} \ldots \hat{I}^{(t+f_h)}$, where $f_h$=14 is the horizon of the forecast operation, as defined in Section 3.2. After the forecast operation, the window moves by one day and the operation is repeated until the end of the forecasting period. Notice that this means a different model may be selected as the best one in each forecast operation.

### 4.4 Evaluation Metrics

The evaluation metrics used to assess the performance of the five methods are the **mean absolute percentage error (MAPE)** and the **rooted mean-square error (RMSE)**. Their definitions follow:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|I_c^{(t)} - \hat{I}_c^{(t)}|}{I_c^{(t)}} \times 100 \quad , \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left( I_c^{(t)} - \hat{I}_c^{(t)} \right)^2} \quad , \tag{7}$$

where $I_c^{(t)}$ is the actual value of infections in area/cell $c$ at time $t$, $\hat{I}_c^{(t)}$ is the forecast value of infections in area/cell $c$ at time $t$ and $n$ is the total number of observations (days). RMSE computes the standard deviation of prediction errors (residuals) and thus, it estimates the average model prediction error. RMSE penalizes big errors, meaning that few significant errors have a great impact on its value. MAPE calculates the average of the percentage errors, where each

absolute error is divided by the actual value. This way the errors are normalized. However, over-forecasts, i.e. predicting more infections that actual, have a greater effect on MAPE, compared to under-forecasts.

Furthermore, in order to obtain a progressive evaluation of all the models, we use windowed versions of the aforementioned metrics, named **wMAPE** and **wRMSE**, defined as follows:

$$wMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega}\sum_{t=1}^{\omega}\frac{|I_c^{(t)} - \hat{I}_c^{(t)}|}{I_c^{(t)}} \times 100\% \quad , \tag{8}$$

$$wRMSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{\omega}\sum_{t=1}^{\omega}\left(I_c^{(t)} - \hat{I}_c^{(t)}\right)^2} \quad . \tag{9}$$

The windowed versions of the metrics compute a MAPE or RMSE value for each sliding window, i.e., the value of $n$ now corresponds to the total number of windows used to cover the whole forecasting period. The window size is $\omega$ and it progress one point at a time, leading to overlapping consecutive windows, until the end of the forecasting period is reached.

### 4.5 Experimental Results

We present our empirical analysis on real daily COVID-19 incidences from Spain. We compare the five methods in a period spanning from 1 April 2020 to 9 September 2020. A forecast always refers to a single day. In what follows, when we mention the forecast step we refer to the number of days in the future a forecast corresponds to. In the analysis that follows, the minimum forecast step we use is 1 day and the maximum is 14 days, since we employ a forecast horizon $f_h$ of 14 days, i.e., $f_h = 14$. In the windowed versions of the metrics the size of the window is 7 days since $\omega = 7$.

Regarding our CA-based approach and the GNN, the spatial resolution concerns the provinces of Spain. A cell $c$ in the CA model represents a province and daily cases are predicted by applying the transition function at each cell. Similarly, in GNN, the nodes of each graph correspond to provinces, their features consist of daily new COVID-19 infections of the past 7 days ($\omega$=7), and future daily cases are predicted for each node/province. After producing as many predictions as the forecast horizon $f_h$ dictates, the sliding window moves and one more time-point is added to the training set. For the CA this means updating the transmission rate of each cell $\overline{\beta_c}$ (see Section 3.2), while the GNN is re-trained with the whole history.

Regarding ARIMA, a model is trained for each province of Spain, where the training data consists of daily new COVID-19 infections that span a period from 1 January 2020 to 31 March 2020. After producing predictions according to the forecast horizon $f_h$, the sliding window moves one time-point and the real data of the additional day are included in the fitting process. The fitting process concerns the update of the internal parameters of the ARIMA model, using the last month's daily incidents reported for the particular province.

Similar to ARIMA, the LSTM is build for each province of Spain. However, it cannot be re-trained after the end of each forecasting operation due to the computational cost of the operation. Therefore, we decided to split the dataset to 5 different training periods and corresponding testing periods (see Table 2). After the completion of each training period, the corresponding testing period is covered by a series of $f_h$ forecasts, during which no re-training takes place. Recall from section 4.3.2, that the input consists of the past 7 days, i.e., $\omega$=7.

Finally, MMCA does not require re-training, since we have generated 1000 models and the best model is selected as described in Section 4.3.4. However, the patches used by MMCA correspond to MITMA mobility zones, significantly smaller areas compared to provinces. To overcome this issue of spatial granularity, we aggregated the results of all

Table 2. Training and testing periods for LSTM

| Training | Test |
|---|---|
| 01/01/2020 — 31/03/2020 | 01/04/2020 — 30/04/2020 |
| 01/01/2020 — 01/04/2020 | 01/05/2020 — 31/05/2020 |
| 01/01/2020 — 31/05/2020 | 01/06/2020 — 30/06/2020 |
| 01/01/2020 — 30/06/2020 | 01/07/2020 — 31/07/2020 |
| 01/01/2020 — 31/07/2020 | 01/08/2020 — 09/09/2020 |

methods to the level of autonomous communities. Autonomous communities is the first level of Spain division and there are 19 communities in total. We selected the autonomous communities to facilitate the visualization process. Moreover, we aggregated the results at the country level to gain a national perspective of the course of the epidemic.

Figure 4 presents the daily new incidents as predicted by the five methods, as well as the actual reported cases. Figure 4(a) depicts predictions one day ahead, while Figures 4(b) and Figures 4(c) display results of the more difficult tasks of 7-day and 14-day predictions, respectively. All diagrams display aggregated results concerning new daily COVID-19 infections over the whole country of Spain. At this level of granularity, all five methods, seem to predict closely the infection trajectory in the forecasting tasks of one and seven days ahead, with the exceptions of LSTM and GNN in the 7-days predictions. The performance of these two methods deteriorates significantly in forecast steps greater than 7, and we opted not to present their results so that the differences in accuracy among the remaining methods remain visible. As expected, in the 14-days forecast step the performance of the three remaining methods also deteriorates. MMCA achieves the smallest deviation from the actual data. Hence, the CA and ARIMA seem more suitable for short-term forecasts.

The differences in the forecasting accuracy on a specific date, as depicted in Figure 4, remain small. For this purpose Figures 5(a) and 5(b) aggregate the errors of the different models across the time period, and use RMSE and wRMSE to measure the difference of each method from the true values. As shown in the Figures 5(a) and 5(b), ARIMA achieves the lowest RMSE for the forecast of the first four days (steps 1 to 4). LSTM has the worst performance and GNN the second worst one. The performance of our CA approach is closer to that of ARIMA and exceeds it in the harder tasks of predicting 6 to 11 days ahead, in terms of RMSE, and 6 to 13 days ahead, in terms of wRMSE. In both Figures, MMCA exhibits the best performance in steps 12 to 14, highlighting the method's long-term forecasting ability. Looking at the MAPE results (Figures 5(c) and 5(d)) we arrive at similar conclusions, but now our CA-based approach is clearly the best method in the majority of the forecast steps. Similar to Figure 4, to avoid the distortion of the diagrams in Figure 5, we chose not to present results of LSTM and GNN for steps greater than 7.

Next, in Figures 6 and 7 we present the RMSE and MAPE scores of four methods for each autonomous community of Spain. We exclude LSTM, since it performs much worse than the other methods. Figures 6(a) and 7(a) correspond to the forecast of the next day (step 1), Figures 6(b) and 7(b) correspond to the 7-th day forecast, while Figures 6(c) and 7(c) correspond to the 14-th day forecast. As before, in Figures 6(c) and 7(c) we do not present the results of GNN to avoid distorting the diagrams. The results are illustrated on a map of Spain where the borders of each autonomous community are shown. ARIMA achieves the lowest RMSE score in all the communities when the forecast step is set to 1 day (see Figure 6(a)). As the forecast step increases, the accuracy of all methods decreases (see Figures 6(b) and 6(c)). In the 7-th day forecast (see Figure 6(b)), ARIMA provides the best predictions in 9 out of the 19 autonomous

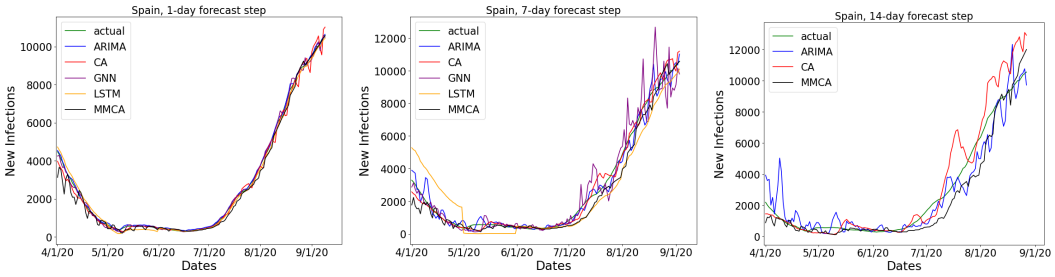



Fig. 4. Daily new incidents of COVID-19 in Spain as predicted by each of the competing methods. In diagram (a) the forecast step is 1 day, in diagram (b) 7 days and in diagram (c) 14 days.

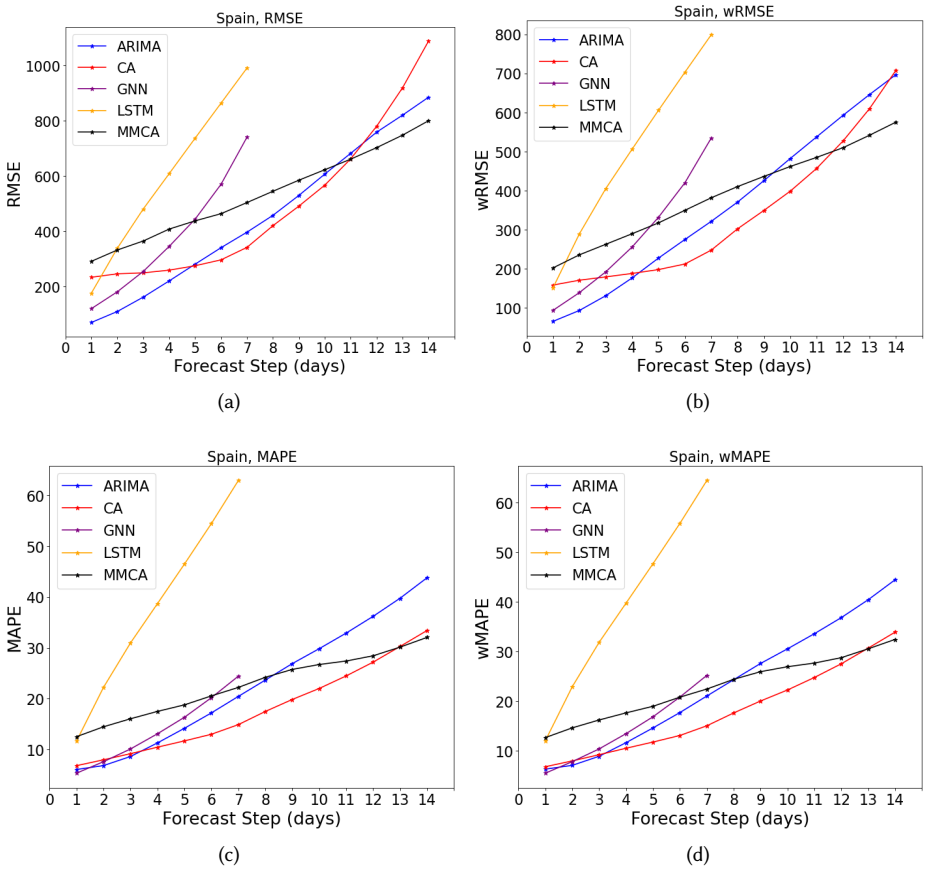



Fig. 5. RMSE (a), wRMSE (b), MAPE (c) and wMAPE (d) for Spain. The forecast step ranges from 1 to 14 days.

communities, while CA and MMCA achieve the best results in 4 communities each. In the 14-th day forecast, ARIMA is more accurate in 7 communities, CA in 3, and MMCA in 8 communities.

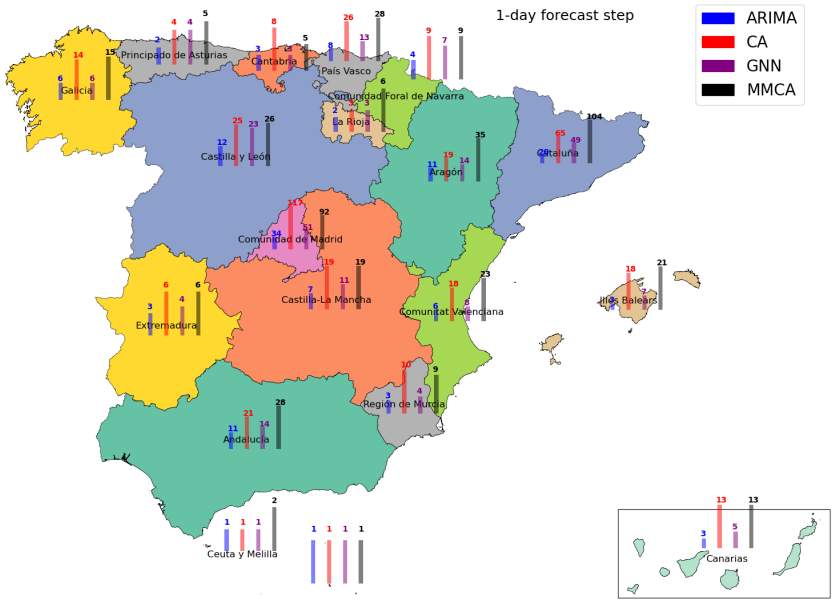
Figures 6(b) and 6(c) illustrate also the effect of the population size on the performance of the models. Looking at the largest two communities, MMCA seems to be doing best in "Comunidad de Madrid" in forecast step 7 and in "Cataluña" in forecast step 14. Regarding the CA-based method, the opposite behavior is observed, that is, it does best in "Cataluña" in step 7 and in "Comunidad de Madrid" in step 14. As expected, the RMSE of all methods is much larger for those two communities, due to the much higher population.

Using the MAPE score (Figure 7), the population size of different communities is normalized. This allows a more direct comparison across communities. According to the MAPE of 1-day forecast (see Figure 7(a)), the CA-based method achieves the best overall result in 9 communities, while ARIMA and GNN are the best methods in 5 and 2 communities, respectively. In the case of the 7-th day forecast (see Figure 7(b)), CA achieves the best MAPE score in 7 communities, while MMCA achieves the best score in 11 communities. GNN performs best in "Comunidad de Madrid". Looking at Figure 7(c), concerning the 14-th day forecast, we arrive at the same conclusions. Interestingly, our CA approach achieves a very good MAPE score in some of the most populous communities, like "Comunidad de Madrid" and "Cataluña". As expected, the performance of all methods is getting worse as the forecast step increases. However, GNN suffers the greatest reduction in performance, followed by ARIMA and then, the CA approach. MMCA demonstrates a more stable behavior, without abrupt changes.

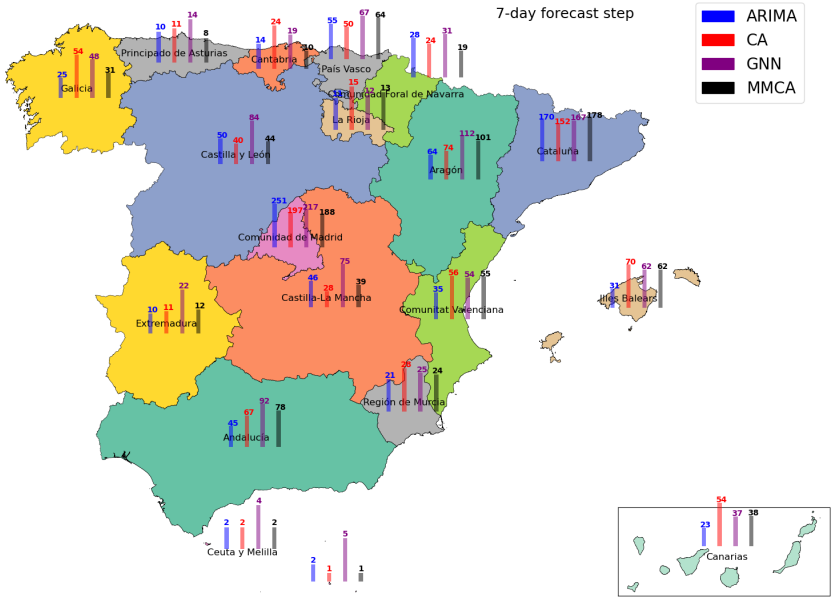
In Table 3, we present the processing time in seconds for the five methods. For ARIMA, as processing time we considered the learning phase, during which the hyper-parameters were estimated, and the time to complete the forecasts of the whole forecasting period, including the time required for parameter fitting at each move of the sliding window. Recall that progressing the window makes new observations available. Similarly, for GNN, the time to reach the end of the forecasting period, including training and forecasting, was considered as processing time. Recall that GNN is trained over the whole history each time a new operation of $f_h$ forecasts is requested. LSTM was evaluated by splitting the dataset in five training and testing periods (see Table 2). When calculating its processing time, we considered the sum of the elapsed times of the five training and testing periods. For MMCA, we considered the time to produce the predictions for each one of the 1000 models, as well as the time to select the predictions of the best model in each employment of the sliding window until the end of the whole forecasting period. For the CA-based method we considered as processing time the completion time of the whole forecasting process, including the time to select the serial interval distribution ($w_s$). As can be seen in Table 3, the computational cost of our CA approach is orders of magnitude lower than that of the other four approaches. This highlights the simplicity of the CA-based method, which can produce reliable forecasts almost instantaneously. In contrast, the high complexity of the other methods translates to increased computational requirements.

Table 3. Processing time in seconds

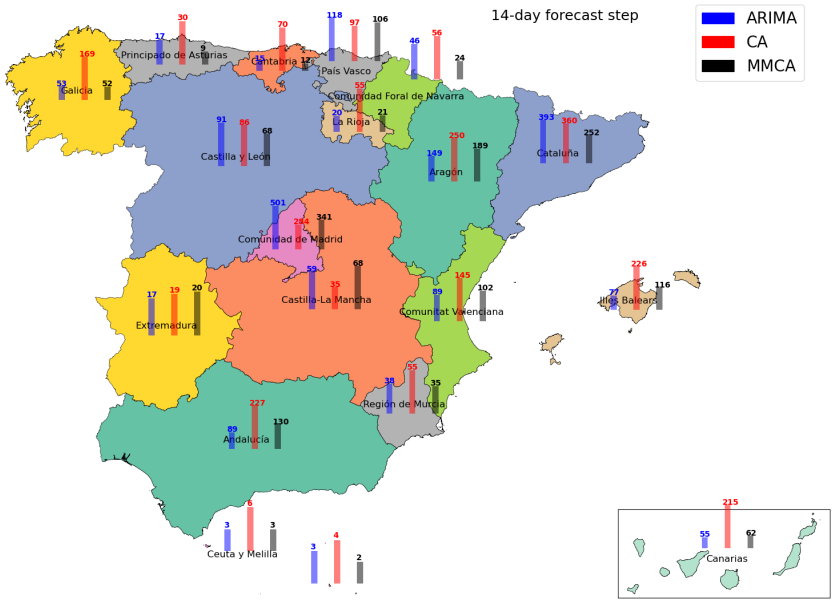
| Method | Time (seconds) |
|---|---|
| LSTM | 211339.1 |
| ARIMA | 23698.8 |
| GNN | 15679.4 |
| MMCA | 3727.8 |
| CA | 57.8 |


1-day forecast step
ARIMA
CA
GNN
MMCA
Principado de Asturias
Cantabria
País Vasco
Comunidad Foral de Navarra
La Rioja
Galicia
Castilla y León
Cataluña
Aragón
Comunidad de Madrid
Castilla-La Mancha
Comunitat Valenciana
Illes Balears
Extremadura
Región de Murcia
Andalucía
Ceuta y Melilla
Canarias

(a)


7-day forecast step
ARIMA
CA
GNN
MMCA
Principado de Asturias
Cantabria
País Vasco
Comunidad Foral de Navarra
La Rioja
Galicia
Castilla y León
Cataluña
Aragón
Comunidad de Madrid
Castilla-La Mancha
Comunitat Valenciana
Illes Balears
Extremadura
Región de Murcia
Andalucía
Ceuta y Melilla
Canarias

(b)



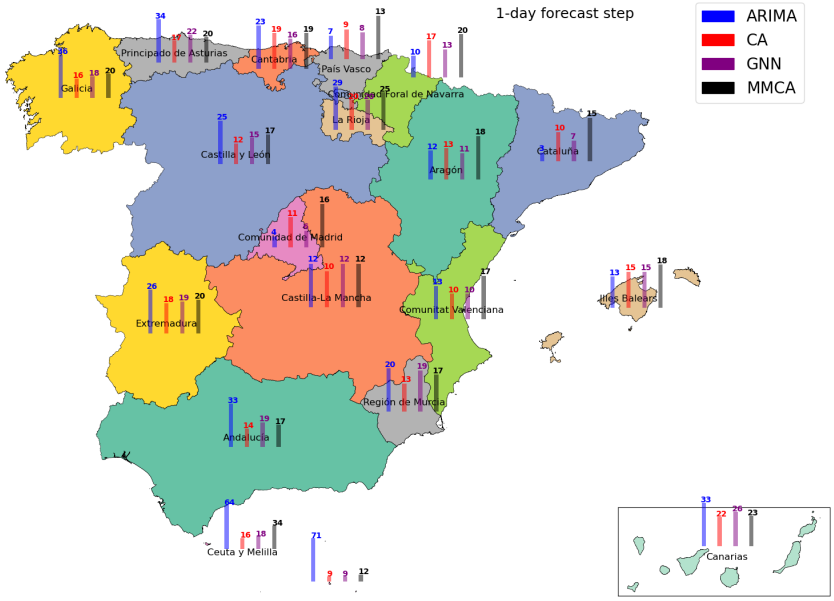

(c)

Fig. 6. RMSE for each autonomous community of Spain. Forecast for the next day (a), the 7-th day (b), and the 14-th day (c).

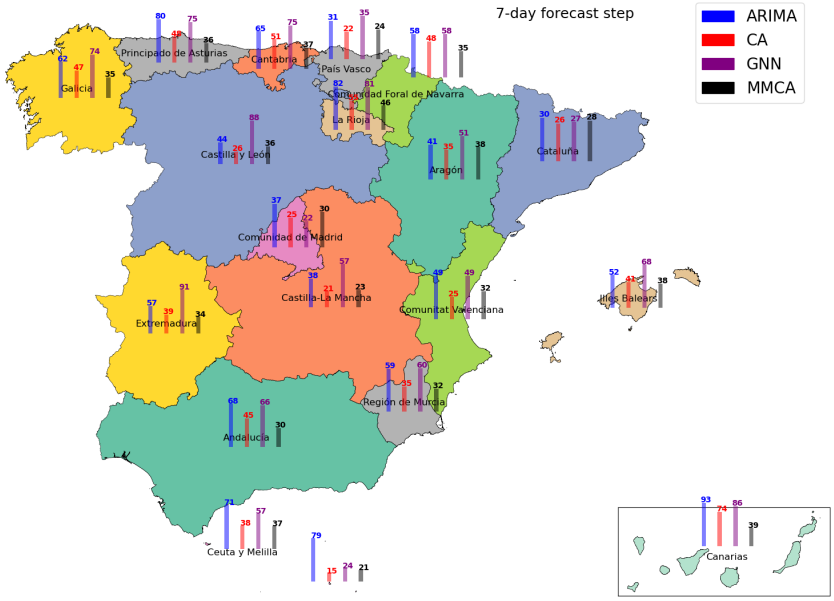
Delving deeper into the computational benefits of our CA-based method, in Fig. 8, we provide an illustration of the trade-off between predictive accuracy and computational performance. We present results for each forecast step size (x-axis) for the whole of Spain. We have included only the methods that were able to produce forecasts of 14-days, that is, ARIMA, MMCA, and the CA. At each step, we first present the improvement of CA in terms of accuracy, based on the MAPE score (red y-axis). Specifically, we compare the CA approach against the best method or the second best one, if CA achieves the best accuracy (positive y values). To establish the trade-off, we also show the factor by which run time is improved by CA (blue y-axis), against the same method that we calculated the accuracy improvement. Run time is calculated as the time needed for the complete experiment by each method (as per Table 3), and it varies very little between forecast steps. As Fig. 8 shows, our method achieves MAPE improvement in most of the steps, while at the same time it improves the computational cost by orders of magnitude. In the cases where our CA approach achieves lower accuracy, one needs to consider if this decrease in accuracy can be tolerated, given the speed at which the estimates are generated.

## 5 Discussion

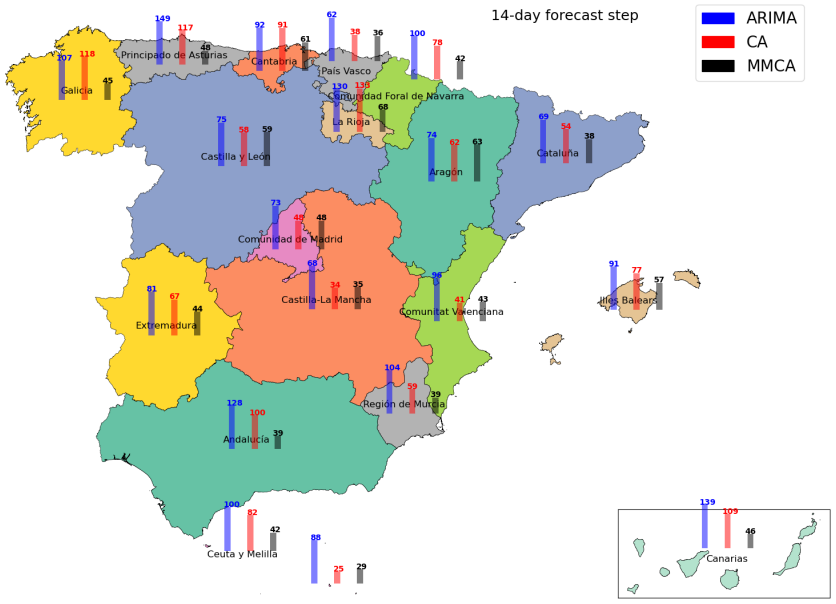
Our proposed CA-based method accurately predicts the daily new COVID-19 cases. It provides a good fit to the actual data, both at the country and at a regional level. The spatial resolution employed in our implementation may be modified. The method is applicable to any scale if an appropriate division of the spatial area under investigation exists and is accompanied by the appropriate mobility and infection data.


1-day forecast step
ARIMA
CA
GNN
MMCA
Principado de Asturias
Cantabria
País Vasco
Galicia
Comunidad Foral de Navarra
La Rioja
Castilla y León
Cataluña
Aragón
Comunidad de Madrid
Castilla-La Mancha
Comunitat Valenciana
Illes Balears
Extremadura
Region de Murcia
Andalucía
Ceuta y Melilla
Canarias

(a)


7-day forecast step
ARIMA
CA
GNN
MMCA
Principado de Asturias
Cantabria
País Vasco
Galicia
Comunidad Foral de Navarra
La Rioja
Castilla y León
Cataluña
Aragón
Comunidad de Madrid
Castilla-La Mancha
Comunitat Valenciana
Illes Balears
Extremadura
Region de Murcia
Andalucía
Ceuta y Melilla
Canarias

(b)




(c)

Fig. 7. MAPE for each autonomous community of Spain. Forecast for the next day (a), the 7-th day (b), and the 14-th day (c).

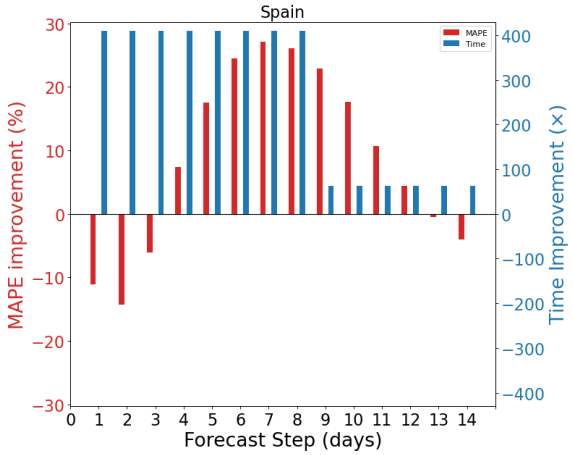
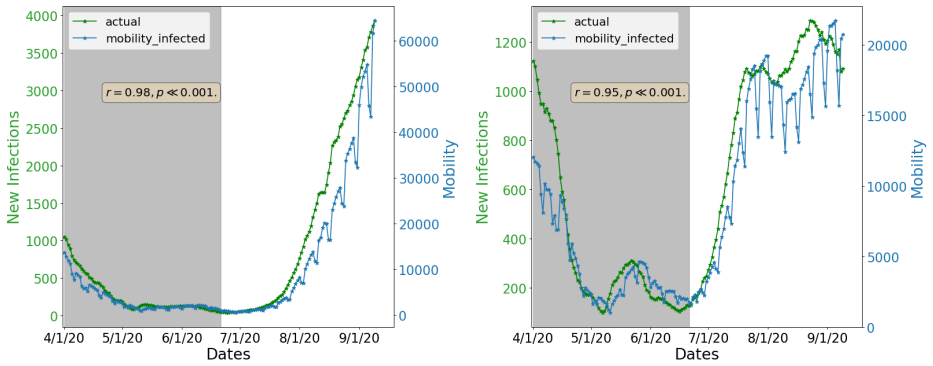



Fig. 8. Trade-off between prediction accuracy based on MAPE and computational performance with respect to the competing approaches.

The transition function of the proposed method (recall Eq. (4)) takes into account both infections caused by human movement (first term of Eq. (4)), as well as those caused by people staying at their home location (second term of Eq. (4)). In respiratory syndromes, like COVID-19, human movement is a decisive factor for disease spread [40, 44].




(a) "Comunidad de Madrid" (b) "Cataluña"

Fig. 9. Daily COVID-19 new cases and mobility of infected people. The grey shaded area corresponds to the days during which, the whole of Spain was under lockdown. The yellow box shows the Pearson correlation coefficient ($r$).

To further highlight the importance of mobility and show how the transition function captures movement across and inside cells/regions, we present in Fig. 9 the actual daily new COVID-19 cases along with the proportion of infected people moving, in two of the largest autonomous communities of Spain: "Comunidad de Madrid" and "Cataluña". Apart from the daily cases (green y-axis), the plot additionally shows the number of infected people that traveled to the target communities the previous day and can transmit the virus (blue y-axis). This number is the sum $\sum_{j=1}^{M} \frac{\overline{I_j^{(t-1)}}}{N_j} m_{jc}^{(t-1)}$, where the number of trips performed from a source community to a target community the previous day ($m_{jc}^{(t-1)}$), are multiplied by the fraction of the population at the source community that was reported infected the last 7 days ($\frac{\overline{I_j^{(t-1)}}}{N_j}$). This is actually the sum appearing in the first term of Eq. (4), capturing the effect of mobility on infections. As it can be seen in both figures, the mobility of infected people is highly correlated to the daily new cases. This is confirmed also by the Pearson correlation coefficient ($r$) shown in the yellow boxes. Similar results are also observed for the remaining communities. This high degree of correlation illustrates how the proposed transition function for the CA captures mobility across and inside regions, and in addition underlines the importance of combining mobility and infections in the forecasting process.

One of the main advantages of the proposed method is its simplicity. The discretization of time and space is comprehensible and the transition function of each cell is simple, compared to many state-of-the-art models. Health officials or policy makers require methods that are interpretable and can aid in efficient planning of strategies to confront and reduce the spread of the virus. The proposed method also produces results quickly, at low computational cost. In contrast, many state-of-the-art methods, like ARIMA, LSTM, GNN and MMCA, require high computational resources and their processing time is orders of magnitude higher. The computational efficiency of our method, along with the fact that it is extensible and flexible, provide the opportunity to incorporate additional factors known to influence the evolution of a disease.

Despite its advantages, the proposed method can be improved in various ways. Commuting and social mixing vary a lot among different population groups. For example, working individuals are more probable to get out of their residency, compared to more senior citizens. Hence, the probability of a working individual to make contacts and acquire the

virus is higher than that of older people. In the presented study we could not study the social dynamics of different age groups. Similarly, the effect of changes in the population over time, e.g. due to births, deaths and tourism, could be studied, if data were provided for a longer period of time. Another important factor to be added to the study is the response of the authorities to the pandemic. In particular, the containment measures applied by authorities to set the epidemic under control have a direct impact on the number of contacts an individual can have, which in turn map onto specific parameters of the CA model, i.e., $k_m$ and $k_h$ in Eq. (4). By varying these parameters, one could study the effect of different disease management actions on daily new infections.

The epidemic model utilized in our study is the usual SIR model. The employment of different compartmental models could also be investigated. Immunity, due to recovery from the disease or due to vaccination, is a parameter that should be taken into account if the model is used in more recent periods. Similarly, considering the incubation period to account for asymptomatic individuals or the hospitalization rates that provide a better understanding of the long-term evolution of a disease [2], might also be helpful. Furthermore, reinfections should be accounted for, and for these a SIRS epidemic model might be more appropriate, since all individuals after a period of time stop being immune and return to the susceptible state. However, any of the above extensions of the model should be done with caution, as it will increase complexity and compromise explainability.

Finally, we need to acknowledge that we aimed at a short-term forecasting (up to 7 days) method, which was not designed to capture time dependencies that extend far into the past. The long-term simulation of disease spread may need to consider other important parameters, such as asymptomatic individuals, and/or a different temporal aggregation of existing parameters. We believe this is very difficult to achieve without affecting the simplicity and complexity of the method.

## 6 Related Work

The goal of epidemiological models is to predict the course of an epidemic, in order for appropriate measures to be taken to control the spread in the society. Regarding airborne diseases, like COVID-19, one of the most significant causes of the spread of the virus is the mobility of humans [44]. Several studies have been published that correlate the spread of the SARS-CoV-2 virus with human mobility patterns [2, 4, 13, 28, 49]. Considering movement habits and patterns of individuals is thus a crucial factor to comprehend the dynamics of fast-spreading pandemics, and the use of real mobility data can push research forward [13, 40, 44]. A study that investigates different movement models and their impact on epidemiological outcomes can be found in [15].

In particular, metapopulation epidemic models address the issue of human movement by overlaying an epidemiological compartmental model on top of locations called ‘patches’, where a number of individuals reside. The spatial interaction among patches, i.e., the movement of individuals from one patch to another, facilitates the exploration of the spatial dynamics of a disease. This is in contrast to the usual compartmental models that only consider the temporal evolution of a disease. The authors in [13] propose a dynamic mobility network, which models the movement of people from small places to various points of interest (POIs), such as restaurants, stores etc, which are assumed to be the main centers of COVID-19 transmission. They use a Susceptible-Exposed-Infected-Recovered (SEIR) model along with mobile phone data to capture mobility patterns and succeed to reproduce the observed case counts. Additionally, their approach allows to estimate the effects of several Non-Pharmacological Interventions (NPIs) on the spread of the virus, including effects on disadvantaged racial and socioeconomic groups. The performance though comes at the cost of large computational resources and especially processing time. Crucial parameters of the method require accurate fine-tuning, a time-consuming process, making the method hard to deploy in practice.

The MMCA method [2], which was used also in our experimental evaluation, proposes an even more extensive compartmental model than the one in [13], calculating the probabilities of state transition for each patch, by taking into consideration mobility. Recall from Section 4.3.4 that the model is sensitive to a number of parameters, combinations of which result in 1000 different variants of the model. Even though several aspects of the COVID-19 pandemic are considered, this results in a highly complex and computationally expensive model. On the other hand, our CA-based approach is simple and lean, depending only on daily cases of the previous week to update the transmission rate of each cell. Hence, it provides quick and accurate results at a low computational cost.

A computationally light metapopulation model for COVID-19 is presented in [36]. This approach acknowledges the importance of bidirectional commuting, where individuals moving to another patch are expected to return home. This way the method avoids diffusing infected individuals from one patch to all other patches and compartmental population variations are modeled more accurately. Furthermore, the method is not so complex and relies on solving ordinary differential equations. However, the model has not been experimentally validated with real mobility data and thus, no definite conclusions can be drawn about its performance.

A stochastic compartmental epidemic model for COVID-19 forecasting, which is not a metapopulation one, is presented in [20]. As in our study, the authors employ a simple SIR model to describe the states of the disease and combine it with a Bayesian sequential estimation and forecasting algorithm, based on daily reported cases. One important aspect of their method is the estimation and application of a time-varying transmission rate, during the forecasting operation as opposed to our time-invariant parameter model. The transmission rate is subject to changes throughout the whole course of a disease, due to factors such as the active infected cases, mitigation strategies, etc. Thus, by updating it during forecasting, its dynamic nature is considered and the accuracy of the predictions can be improved. In [43] sequential Bayesian learning is further combined with the mean-agnostic sequential test (MAST) [7] statistic to forecast COVID-19 infections. MAST is a method to predict transitions between a controlled phase, where there is no need for interventions, and a critical phase, where there is an exponential growth of infections and policy makers must adopt containment measures for the epidemic. MAST trades off decision delay, i.e., the time to declare a phase transition with the risk to declare a transition too soon. However, all these approaches neglect the spatial dimension of the epidemic, since they do not incorporate mobility factors, which are a significant cause of disease spread.

On the other hand, time-series models have been used widely to forecast the course of the COVID-19 pandemic [11, 14, 23, 32, 39]. In [11] and [32] different time-series models are used to forecast infected cases in several countries, showing that a single method cannot fit accurately the data in all countries. In our experimental evaluation, we included ARIMA, a popular forecasting approach, and showed that it can produce reliable short-term forecasts. However, in order to make accurate predictions, ARIMA uses a significant amount of historical data for parameter fitting, making the method data and computationally expensive, compared to the CA-based approach. Recall from Section 4.3.1 that the incidences of the previous month are used to update the internal parameters (fitting process) of each ARIMA model applied to a province.

In the spirit of data-driven approaches, deep learning techniques have also been used to study the development and spread of the epidemic [11, 22, 37, 41, 48, 50]. These methods employ a number of features to train the neural networks, including the daily confirmed cases, containment policies such as face covering, gathering restrictions, statistical properties of the data, etc., and have succeeded in accurately predicting the infection trajectory in the short-term future. In our empirical analysis, we included as a competing method, the LSTM. LSTM has been shown in the literature to perform better than time-series methods (including ARIMA) [41, 53] and other machine learning methods [50, 53]. In [48], the authors use a rolling mechanism, where an LSTM network is re-trained with its own predictions to estimate

the epidemic trend 150-days ahead in different countries. In [22], the authors perform transfer learning by training an LSTM network in the data of two countries and use the trained networks to forecast the daily cases in other countries. In [50], an LSTM, a convolutional neural network (CNN), and a hybrid approach CNN-LSTM, have been used to predict the daily COVID-19 cases in three countries. In addition to the reported cases, various mitigation policies were included as features in the training of the three models.

However, our experimental results did not confirm those presented in the literature for LSTM networks. Furthermore, the accuracy of the networks was shown to deteriorate quickly as the forecast step increased. This is mainly due to the limited size of the training set and highlights the dependence of neural networks on large amounts of data. In the first stages of an epidemic, or in small geographical regions, the lack of sufficient training data makes deep learning techniques unsuitable for forecasting. In addition, the training of the networks requires substantial computational resources and cannot be used for real-time decision making. In contrast, the CA-based approach can produce reliable short-term forecasts, using few data and limited resources. Furthermore, the CA-based method is simple and easily interpretable, which is not the case for LSTM and ARIMA. It is worth-mentioning that we tried to improve LSTM's accuracy by incorporating the mobility data as additional features, but an improvement was not observed.

In order to study the spatio-temporal spread of a disease, a natural choice is to represent the problem as a graph. Our CA-based approach assumes a 2D grid, where the connections among cells vary over time and space, depending on human movement. These dynamics can be captured by spatio-temporal graphs. The literature includes some deep learning approaches that use graphs to study COVID-19 [21, 33, 38]. These approaches are typically based on GNNs and show improved performance compared to statistical and other deep learning approaches, including ARIMA and LSTM. In [21, 33], GNNs are used to predict COVID-19 daily cases in all counties of the USA. In both studies, each node of the graph corresponds to a county and has static features, such as population size, population density, etc., and dynamic features, such as number of active cases, total cases, etc., falling inside a predefined window. In [21], the authors construct a model that combines a graph attention network (GAT) to extract spatio-temporal features and a gated recurrent unit (GRU) to learn temporal features. The model has multiple outputs for predicting the cases for a fixed number of days in the future. The spatial edges of the graphs are based on demographic similarity and geographical proximity between the counties, and thus, the effect of mobility on disease spread is not studied. In [33], a simple GNN is used to predict the cases of the next day in each county of the USA, but now the edges represent the human mobility between regions. A similar model to [33] that is based on mobility to construct the graphs is proposed in [38]. The authors implemented a GNN that exhibits good performance in predicting future COVID-19 cases at the regional level of several countries. In addition, to account for the limited amount of training data and the fact that different countries may be on different stages of the epidemic, they developed an efficient transfer learning technique. Due to their mobility-based graph representations and the use of different forecast steps, we included their model in our experimental comparison. Notice that in all of the aforementioned approaches, the GNNs that are used are not the typical type, where the network is trained over a single graph. Training data consist of temporal snapshots, where each snapshot is a graph with different features for the nodes and different weights for the edges.

The authors in [38] report results for up to 14 days ahead in the future, where they train a network each time new data become available and for each forecast step separately. This means that trainable parameters must be learned for 14 different networks each time the window slides. This is a very time-consuming process, not favoring real-time decision making. Additionally, regardless of the step used, future predictions in [38] were always based on real infection data. This is not feasible when monitoring the epidemic spread, and a model should be able to use its own predictions to produce reliable forecasts further ahead in the future. Aiming at a fair comparison, we trained a single GNN for

the next day each time and then applied the forecast operation discussed in Section 3.2 to generate forecasts. Our evaluation demonstrated that the GNN has a very poor performance as the forecast step increases and it uses its own predictions as input. This result highlights once more the need of deep learning methods for sufficient training data. Despite the similarity in the representation of the task between the GNN and our CA-based method, the former has to learn many parameters only to produce accurate results for small forecast steps. In contrast, the CA-based approach is capable for short-term forecasting by tuning a single parameter (the infectivity profile $w_s$, see Section 4.2), which is kept stable during the whole forecasting period. Moreover, it can exploit its own predictions to generate forecasts without sacrificing significantly the accuracy. This demonstrates the cost-effective use of resources of the CA-based approach and its ability to provide a quick overview of the ongoing epidemic. Another practical problem of the GNN is the difficulty in interpreting the results of each node/region. In contrast, the transition function we employ on the cells of the CA is simple and it makes clear how spatial and temporal factors affect the outcome. Finally, we have also evaluated a model proposed in [38], which combines the GNN with an LSTM to capture temporal dependencies in the data. We have observed a significant drop in accuracy compared to the simple GNN, in agreement with the results in [38].

The spatio-temporal dynamics of epidemic spread have also been explored using CA [3, 6, 17, 18, 24, 25, 42, 45]. Various types of CA have been proposed in the literature, with the main objective being to observe complex macroscopic behaviors from the interaction of the local cells (see [5] for a CA survey). These include the use of novel interaction neighborhoods [3, 17, 42, 45], instead of the classical Von Neumann and Moore neighborhoods of radius 1, deterministic [3, 18, 45] or probabilistic [17, 24, 25, 42] transition functions, non-uniform functions [18], where each cell may have a different transition function. Furthermore, CAs that incorporate individual heterogeneity, such as age and sex information [6, 17], the effect of population density [6, 17, 25], the application of several NPIs [25] or the pressure to the healthcare systems [42], have been proposed.

Specifically for COVID-19, a probabilistic CA is used in [42], along with a detailed epidemiological model capturing many of the states an individual can be in, after acquiring the virus. The method explores the effects of social isolation and healthcare infrastructures on the evolution of the epidemic. The simulation employs real data from Brazil and a huge lattice is used to cover the whole population of the country, each cell referring to a single individual. The complexity of the method makes it impractical for quick prediction of the spread dynamics. Probabilistic CAs have also been used in [25] to explore the effects of population density, movement restriction, lockdowns, virus testing efficiency and different state transition probabilities. That method is combined in [24] with a genetic algorithm to estimate from real data the initial parameters of the model and use it for forecasting. The authors in [24] present results from the disease evolution in many countries with great accuracy. However, the cells represent again individuals and the use of small interaction neighborhoods cannot capture the effects of mobility on disease spread.

Another CA model for simulating and analyzing COVID-19 evolution is proposed in [17], where heterogeneity parameters, including sex ratio, age structure, individual immunity, incubation and treatment period, and population movement are considered. The number of confirmed COVID-19 cases is predicted in the context of different control strategies. Each cell corresponds to an individual and the interaction neighborhood is combined with a maximum moving step to simulate human movement. Once more, it is not clear whether such approaches can simulate accurately the mobility influence on the epidemic course, where traveling among regions is a decisive factor for the diffusion of the virus. The CA model that relates more closely to our approach is the one presented in [18], where each cell represents a whole area and continuous variables are used to record the number of susceptible, infected and recovered individuals at each time step. The lattice used overlays the China map divided into administrative regions, where each region may

consist of more than one cell. The neighborhood utilized is the Von Neumann of radius 1 and thus, long distance human movement, i.e. traveling among regions that do not abut, is not considered.

Finally, a subtle difference of all the CA approaches outlined in the present section, as compared to our approach, is that they try to simulate human mobility through small interaction neighborhoods without considering real mobility data. In contrast, our CA-based method explores the spatio-temporal dynamics of COVID-19 by incorporating real mobility data in the transition function.

## 7 Conclusions and Future Work

In this paper, we presented a method based on Cellular Automata (CA), tailored to short-term forecasting of daily new COVID-19 cases. Our approach employs a simple SIR epidemiological model in each cell of the CA, where each cell represents a geographical region, and avoids the utilization of traditional neighborhood schemes. The transition function of each cell takes into account the impact of human mobility in the spread of the virus. The neighbors of a cell are all other cells from which there is incoming human movement. The flow of incoming individuals determines to an extent the number of infected individuals at each time-step. Furthermore, we adopt a unique transmission rate for each cell, computed directly from data, in order to capture disparities that appear in different spatial areas. Our method has been applied to Spain, using daily COVID-19 incidences and mobility data, and was evaluated retrospectively on the prediction of future infections. In addition, we compared our method against four state-of-the-art methods from the fields of time-series analysis, deep learning, and metapopulation epidemic modeling. The results are very promising and show that our method produces accurate forecasts, exceeding in most cases the performance of more complex methods. One of the main advantages of the proposed method is that it makes use of limited computational resources, favoring real-time decision making. Thus, it may assist policy makers to adapt quickly the measures they are taking for attenuating the epidemic.

In future work, we intend to overlay different epidemiological models over each cell of the CA, in order to include more specific features of the virus and assess the effect of this additional complexity on prediction accuracy. Furthermore, we intend to test the presented approach on a higher spatial resolution and apply it to other countries and infectious diseases.

## Acknowledgments

We would like to thank Miguel Ponce-de Leon and Javier Valle not only for their help and assistance in downloading and interpreting the various experimental data but also for providing the necessary data and parameters to run the MMCA model used for experimental comparison.

## References

[1] Andrea Apolloni, Chiara Poletto, Jose Javier Ramasco, Pablo Jensen, and Vittoria Colizza. 2014. Metapopulation epidemic models with heterogeneous mixing and travel behavior. *Theoretical biology & medical modelling* 11 (01 2014), 3. doi:10.1186/1742-4682-11-3

[2] Alex Arenas, Wesley Cota, Jesús Gómez-Gardeñes, Sergio Gómez, Clara Granell, Joan T. Matamalas, David Soriano-Paños, and Benjamin Steinegger. 2020. Modeling the Spatiotemporal Epidemic Spreading of COVID-19 and the Impact of Mobility and Social Distancing Interventions. *Phys. Rev. X* 10 (Dec 2020), 041055. Issue 4. doi:10.1103/PhysRevX.10.041055

[3] Senthil Athithan, Vidya Shukla, and Sangappa Biradar. 2014. Dynamic Cellular Automata Based Epidemic Spread Model for Population in Patches with Movement. *Journal of Computational Environmental Sciences* 2014 (02 2014), 1–8. doi:10.1155/2014/518053

[4] Hamada S Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M Squire, and Lauren M Gardner. 2020. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases* 20, 11 (2020), 1247–1254. doi:10.1016/S1473-3099(20)30553-3

[5] Kamalika Bhattacharjee, Nazma Naskar, Souvik Roy, and Sukanta Das. 2020. A Survey of Cellular Automata: Types, Dynamics, Non-Uniformity and Applications. *Natural Computing: An International Journal* 19, 2 (jun 2020), 433–461. doi:10.1007/s11047-018-9696-8

[6] Sheng Bin, Gengxin Sun, and Chih-Cheng Chen. 2019. Spread of Infectious Disease Modeling and Analysis of Different Factors on Spread of Infectious Disease Based on Cellular Automata. *International Journal of Environmental Research and Public Health* 16, 23 (2019). doi:10.3390/ijerph16234683

[7] Paolo Braca, Domenico Gaglione, Stefano Marano, Leonardo M. Millefiori, Peter Willett, and Krishna Pattipati. 2021. Decision support for the quickest detection of critical COVID-19 phases. *Scientific Reports* 11, 1 (apr 2021). doi:10.1038/s41598-021-86827-6

[8] Fred Brauer. 2008. *Compartmental Models in Epidemiology.* Springer Berlin Heidelberg, Berlin, Heidelberg, 19–79. doi:10.1007/978-3-540-78911-6_2

[9] Jan M. Brauner, Sören Mindermann, Mrinank Sharma, David Johnston, John Salvatier, Tomáš Gavenčiak, Anna B. Stephenson, Gavin Leech, George Altman, Vladimir Mikulik, Alexander John Norman, Joshua Teperowski Monrad, Tamay Besiroglu, Hong Ge, Meghan A. Hartwick, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal, and Jan Kulveit. 2021. Inferring the effectiveness of government interventions against COVID-19. *Science* 371, 6531 (2021), eabd9338. doi:10.1126/science.abd9338

[10] Tom Britton. 2010. Stochastic epidemic models: A survey. *Mathematical Biosciences* 225, 1 (2010), 24–35. doi:10.1016/j.mbs.2010.01.006

[11] Tanujit Chakraborty, Indrajit Ghosh, Tirna Mahajan, and Tejasvi Arora. 2022. *Nowcasting of COVID-19 Confirmed Cases: Foundations, Trends, and Challenges.* Springer International Publishing, Cham, 1023–1064. doi:10.1007/978-3-030-72834-2_29

[12] C.L. Chang, Y. Jing, S. Zhang, G. Stojanovski, M. Stankovski, and G. Dimirovski. 2010. Investigating Epidemic Diseases Characterized by Vertical Transmission and Contract: Cellular Automata Computational Model1. *IFAC Proceedings Volumes* 43, 25 (2010), 155–159. doi:10.3182/20101027-3-XK-4018.00031 13th IFAC Workshop on Supplemental Ways for Improving International Stability.

[13] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. [n. d.]. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589, 7840 ([n. d.]). doi:10.1038/s41586-020-2923-3

[14] Fuad Ahmed Chyon, Md. Nazmul Hasan Suman, Md. Rafiul Islam Fahim, and Md. Sazol Ahmmed. 2022. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. *Journal of Virological Methods* 301 (2022), 114433. doi:10.1016/j.jviromet.2021.114433

[15] Daniel T. Citron, Carlos A. Guerra, Andrew J. Dolgert, Sean L. Wu, John M. Henry, Héctor M. Sánchez C., and David L. Smith. 2021. Comparing metapopulation dynamics of infectious diseases under different models of human movement. *Proceedings of the National Academy of Sciences* 118, 18 (2021), e2007488118. doi:10.1073/pnas.2007488118

[16] Anne Cori, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. 2013. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology* 178, 9 (09 2013), 1505–1512. doi:10.1093/aje/kwt133

[17] Jindong Dai, Chi Zhai, Jiali Ai, Jiaying Ma, Jingde Wang, and Wei Sun. 2021. Modeling the Spread of Epidemics Based on Cellular Automata. *Processes* 9, 1 (2021). doi:10.3390/pr9010055

[18] Puspa Eosina, Aniati Murni Arymurthy, and Adila Alfa Krisnadhi. 2022. A Non-Uniform Continuous Cellular Automata for Analyzing and Predicting the Spreading Patterns of COVID-19. *Big Data and Cognitive Computing* 6, 2 (2022). doi:10.3390/bdcc6020046

[19] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Unwin, Thomas Mellan, Helen Coupland, Charlie Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey Eaton, Mélodie Monod, Azra Ghani, Christl Donnelly, Steven Riley, Michaela Vollmer, Neil Ferguson, Lucy Okell, Samir Bhatt, Pablo Perez-Guzman, and Patrick Walker. 2020. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* 584 (08 2020). doi:10.1038/s41586-020-2405-7

[20] Domenico Gaglione, Paolo Braca, Leonardo Maria Millefiori, Giovanni Soldi, Nicola Forti, Stefano Marano, Peter K. Willett, and Krishna R. Pattipati. 2020. Adaptive Bayesian Learning and Forecasting of Epidemic Evolution—Data Analysis of the COVID-19 Outbreak. *IEEE Access* 8 (2020), 175244–175264. doi:10.1109/ACCESS.2020.3019922

[21] Junyi Gao, Rakshith Sharma, Cheng Qian, Lucas M Glass, Jeffrey Spaeder, Justin Romberg, Jimeng Sun, and Cao Xiao. 2021. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *Journal of the American Medical Informatics Association* 28, 4 (01 2021), 733–743. arXiv:https://academic.oup.com/jamia/article-pdf/28/4/733/36642145/ocaa322.pdf doi:10.1093/jamia/ocaa322

[22] Yogesh Gautam. 2022. Transfer Learning for COVID-19 cases and deaths forecast using LSTM network. *ISA Transactions* 124 (2022), 41–56. doi:10.1016/j.isatra.2020.12.057

[23] Emrah Gecili, Assem Ziady, and Rhonda D. Szczesniak. 2021. Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. *PLOS ONE* 16, 1 (01 2021), 1–11. doi:10.1371/journal.pone.0244173

[24] Sayantari Ghosh and Saumik Bhattacharya. 2020. A Data-Driven Understanding of COVID-19 Dynamics Using Sequential Genetic Algorithm Based Probabilistic Cellular Automata. *Appl. Soft Comput.* 96, C (nov 2020), 12 pages. doi:10.1016/j.asoc.2020.106692

[25] Sayantari Ghosh and Saumik Bhattacharya. 2021. Computational Model on COVID-19 Pandemic Using Probabilistic Cellular Automata. *Sn Computer Science* 2 (2021).

[26] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for Quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 1263–1272.

[27] Katelyn M. Gostic, Lauren McGough, Edward B. Baskerville, Sam Abbott, Keya Joshi, Christine Tedijanto, Rebecca Kahn, Rene Niehus, James A. Hay, Pablo M. De Salazar, Joel Hellewell, Sophie Meakin, James D. Munday, Nikos I. Bosse, Katharine Sherrat, Robin N. Thompson, Laura F. White, Jana S. Huisman, Jérémie Scire, Sebastian Bonhoeffer, Tanja Stadler, Jacco Wallinga, Sebastian Funk, Marc Lipsitch, and Sarah Cobey. 2020. Practical considerations for measuring the effective reproductive number, Rt. *PLOS Computational Biology* 16, 12 (12 2020), 1–21. doi:10.1371/journal.pcbi.1008409

[28] Nicolò Gozzi, Michele Tizzoni, Matteo Chinazzi, Leo Ferres, Alessandro Vespignani, and Nicola Perra. 2021. Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile. *Nature Communications* 12 (04 2021). doi:10.1038/s41467-021-22601-6

[29] Nils Haug, Lukas Geyrhofer, Alessandro Londei, Elma Hot Dervic, Amélie Desvars, Vittorio Loreto, Beate Conrady, Stefan Thurner, and Peter Klimek. 2020. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human Behaviour* (07 2020), 1303–1312. doi:10.1101/2020.07.06.20147199

[30] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (11 1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735

[31] Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. 2021. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology* 19, 3 (2021), 141–154.

[32] Leila Ismail, Huned Materwala, Taieb Znati, Sherzod Turaev, and Moien A.B. Khan. 2020. Tailoring time series models for forecasting coronavirus spread: Case studies of 187 countries. *Computational and Structural Biotechnology Journal* 18 (2020), 2972–3206. doi:10.1016/j.csbj.2020.09.015

[33] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O'Banion. 2020. Examining COVID-19 Forecasting using Spatio-Temporal GNNs. In *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)*.

[34] W. O. Kermack and A. G. McKendrick. 1927. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115, 772 (1927), 700–721. http://www.jstor.org/stable/94815

[35] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl

[36] Azi Lipshtat, Roger Alimi, and Yochai Ben-Horin. 2021. Commuting in metapopulation epidemic modeling. *Scientific Reports* 11 (07 2021), 15198. doi:10.1038/s41598-021-94672-w

[37] Junling Luo, Zhongliang Zhang, Yao Fu, and Feng Rao. 2021. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. *Results in Physics* 27 (2021), 104462. doi:10.1016/j.rinp.2021.104462

[38] George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2021. Transfer Graph Neural Networks for Pandemic Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6 (May 2021), 4838–4845. doi:10.1609/aaai.v35i6.16616

[39] Fotios Petropoulos and Spyros Makridakis. 2020. Forecasting the novel coronavirus COVID-19. *PLOS ONE* 15, 3 (03 2020), 1–8. doi:10.1371/journal.pone.0231236

[40] Miguel Ponce-de Leon, Javier Valle, José Fernández, Marc Bernardo, Davide Cirillo, Jon Sánchez-Valle, Matthew Smith, Salvador Capella-Gutierrez, Tania Gullón, and Alfonso Valencia. 2021. COVID-19 Flow-Maps an open geographic information system on COVID-19 and human mobility for Spain. *Scientific Data* 8 (11 2021). doi:10.1038/s41597-021-01093-5

[41] Zulfany Erlisa Rasjid, Reina Setiawan, and Andy Effendi. 2021. A Comparison: Prediction of Death and Infected COVID-19 Cases in Indonesia Using Time Series Smoothing and LSTM Neural Network. *Procedia Computer Science* 179 (2021), 982–988. doi:10.1016/j.procs.2021.01.102 5th International Conference on Computer Science and Computational Intelligence 2020.

[42] P.H.T. Schimit. 2021. A model based on cellular automata to estimate the social isolation impact on COVID-19 spreading in Brazil. *Computer Methods and Programs in Biomedicine* 200 (2021), 105832. doi:10.1016/j.cmpb.2020.105832

[43] Giovanni Soldi, Nicola Forti, Domenico Gaglione, Paolo Braca, Leonardo M. Millefiori, Stefano Marano, Peter K. Willett, and Krishna R. Pattipati. 2021. Quickest Detection and Forecast of Pandemic Outbreaks: Analysis of COVID-19 Waves. *IEEE Communications Magazine* 59, 9 (2021), 16–22. doi:10.1109/MCOM.101.2001252

[44] David Soriano-Paños, Gourab Ghoshal, Alex Arenas, and Jesús Gómez-Gardeñes. 2020. Impact of temporal scales and recurrent mobility patterns on the unfolding of epidemics. *Journal of Statistical Mechanics: Theory and Experiment* 2020, 2 (feb 2020), 024006. doi:10.1088/1742-5468/ab6a04

[45] Ishant Tiwari, Pradeep Sarin, and P. Parmananda. 2020. Predictive modeling of disease propagation in a mobile, connected community using cellular automata. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30, 8 (2020), 081103. doi:10.1063/5.0021113

[46] Constantine I. Vardavas, Katerina Nikitara, Katerina Aslanoglou, Michele Hilton-Boon, Revati Phalkey, Jo Leonardi-Bee, Gkikas Magiorkinis, Paraskevi Katsaounou, Anastasia Pharris, Ettore Severi, and Jonathan E. Suk. 2021. Effectiveness of non-pharmaceutical measures (NPIs) on COVID-19 in Europe: A systematic literature review. *medRxiv* (2021). doi:10.1101/2021.11.11.21266216

[47] Jacco Wallinga and Peter Teunis. 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology* 160, 6 (2004), 509–516.

[48] Peipei Wang, Xinqi Zheng, Gang Ai, Dongya Liu, and Bangren Zhu. 2020. Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. *Chaos, Solitons & Fractals* 140 (2020), 110214. doi:10.1016/j.chaos.2020.110214

[49] Chenfeng Xiong, Songhua Hu, Mofeng Yang, Weiyu Luo, and Lei Zhang. 2020. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences* 117, 44 (2020), 27087–27089. doi:10.1073/pnas.2010836117

[50] Lu Xu, Rishikesh Magar, and Amir Barati Farimani. 2022. Forecasting COVID-19 new cases using deep learning methods. *Computers in Biology and Medicine* 144 (2022), 105342. doi:10.1016/j.compbiomed.2022.105342

[51] S. Yakowitz, J. Gani, and R. Hayes. 1990. Cellular automaton modeling of epidemics. *Appl. Math. Comput.* 40, 1 (1990), 41–54. doi:10.1016/0096-3003(90)90097-M

[52] Louis Yat, Baoyin Yuan, and Matteo Convertino. 2021. COVID-19 Non-Pharmaceutical Intervention Portfolio Effectiveness and Risk Communication Predominance. *Scientific Reports* (05 2021). doi:10.1038/s41598-021-88309-1

[53] İsmail Kırbaş, Adnan Sözen, Azim Doğuş Tuncer, and Fikret Şinasi Kazancıoğlu. 2020. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos, Solitons & Fractals* 138 (2020), 110015. doi:10.1016/j.chaos.2020.110015