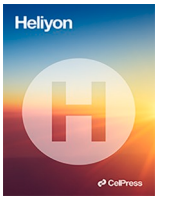
Heliyon 9 (2023) e17625

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



A new RNN based machine learning model to forecast COVID-19 incidence, enhanced by the use of mobility data from the bike-sharing service in Madrid

Mario Muñoz-Organero *, Patricia Callejo, Miguel Ángel Hombrados-Herrera

Telematic Engineering Department, Universidad Carlos III de Madrid, Leganes, 28911, Madrid, Spain

ARTICLE INFO

Keywords:
COVID-19 short term forecast
LSTM machine learning models
Mobility enhanced models
Open data driven models

ABSTRACT

As a respiratory virus, COVID-19 propagates based on human-to-human interactions with positive COVID-19 cases. The temporal evolution of new COVID-19 infections depends on the existing number of COVID-19 infections and the people's mobility. This article proposes a new model to predict upcoming COVID-19 incidence values that combines both current and near-past incidence values together with mobility data. The model is applied to the city of Madrid (Spain). The city is divided into districts. The weekly COVID-19 incidence data per district is used jointly with a mobility estimation based on the number of rides reported by the bike-sharing service in the city of Madrid (BiciMAD). The model employs a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) to detect temporal patterns for COVID-19 infections and mobility data, and combines the output of the LSTM layers into a dense layer that can learn the spatial patterns (the spread of the virus between districts). A baseline model that employs a similar RNN but only based on the COVID-19 confirmed cases with no mobility data is presented and used to estimate the model gain when adding mobility data. The results show that using the bike-sharing mobility estimation the proposed model increases the accuracy by 11.7% compared with the baseline model.

1. Introduction

The COVID-19 virus has affected millions of people worldwide [1] since the first cases found in China by the end of 2019. Far from disappearing, the COVID-19 virus is unceasingly causing new cases every day and the emergence of highly transmissible viral variants that partially escape antibodies is still a major challenge [1]. Understanding the mechanisms governing COVID-19 infections and trying to estimate the evolution of the pandemic has been the subject of many previous studies [2,3]. Anticipating the scalation of infections may help in providing in time response and optimal resource utilization [4]. Being able to anticipate the propagation of the virus is an important element together with early detection and diagnoses [5,6] in anticipating the demand of heath care services which allows states to use their resources effectively [7].

Different approaches can be used to generate accurate predictions for the propagation of the COVID-19 virus. Masum et al. [8] performed a comparative analysis of mathematical epidemic models, statistical models and machine learning models, showing with experimental results that deep machine learning models were able to outperform the other methods in predicting accuracy as

* Corresponding author.
E-mail address: munozm@it.uc3m.es (M. Muñoz-Organero).

https://doi.org/10.1016/j.heliyon.2023.e17625
Received 6 April 2023; Received in revised form 22 June 2023; Accepted 23 June 2023
Available online 24 June 2023

compared with the other studies approaches. The mathematical models, however, showed more interpretable hints about the virus spread [8]. A systematic review of mathematical epidemic prediction models and public health intervention strategies dealing with COVID-19 new infections was carried out by Yue Xiang et al. [9]. The results showed that COVID-19 epidemic models focus on the infectivity of the virus (providing estimations for the basic reproduction number), and incorporate the key time periods of the infection. Epidemic models provide estimations for the short and long-term evolution of the virus evaluating the different impacts of alternative public health interventions [9]. Statistical models use probabilistic representations for the variables modelling the propagation of the virus. The time for incubation and hospitalization are examples of stochastic variables [10] commonly used. Trainable functions are used by machine learning methods in order to learn how to estimate unknown variables (such as the upcoming values for new infections) based on input data that can be observed (such as the current incidence and mobility values). Machine learning (ML) models need to learn from previous data in order to optimize their internal parameters. The trained models are able to provide estimations for new data. Machine learning (ML) models are also popular models mainly designed to generate short term forecasts for the evolution of new COVID-19 infections [11,12], in order to support the diagnoses [13] or provide estimates for other COVID-19 influenced variables [14]. A positive aspect for machine learning (ML) methods is that they do not use a simplified representation of the underlying processes (as epidemic models do). ML models observe data samples and are trained to learn from them as a mechanism to estimate the propagation of the virus. The combination of both epidemic and machine learning models has also been previously studied as hybrid methods which use ML models to estimate some parameters which are then used by the epidemic models [15]. Since hybrid models are based on epidemic models, they maintain similar simplifications from the virus spreading processes but their parameters are fitted to the underlying data. Apart from the previous models, other approaches based on simulation tools have also been used to forecast the spread of the COVID-19 virus. Simulation based models use behavioral rules to mimic the human to human interactions and use propagation rules based on the characteristics of the virus in order to imitate the behavior in the real world. Agent-Based environments are examples of simulation-based models [16] based on the utilization of autonomous software agents to replicate the behavior of the members of a population. The agents are based on rules including mobility and social aspects which try to capture the behavior of the people in the area of study in which the spread of the virus is simulated.

COVID-19 is a virus propagated based on human interactions which have both spatial and temporal components [14,15]. The virus is spread when people are exposed to it for a certain time. Human mobility carries the virus from one region to another [14,15]. This paper uses the data from a bike-sharing service in order to estimate the people's mobility in space (from one district to another) and uses a novel ML model to estimate the short-term spreading of the COVID-19 infections. The model uses a Recurrent Neural Networks (RNN) to analyze the combined influence of human mobility and COVID-19 infections. The design of the ML model includes both the analysis of space and time to find combined patterns. The results are validated using data for a complete year for the City of Madrid (Spain). The main contributions of this manuscript are:

- Defining a new method to get a proxy variable to capture the mobility of the population among different areas of the city (districts) based on the mobility data from a bike-sharing service. The heterogeneous distribution and the density of the bike stations per zone has been considered in order to estimate the probability of using the bike sharing service for human mobility.
- Enhancing the accuracy of time-based Recurrent Neural Network (RNN) methods for estimating the short-term COVID-19 incidence by combining the temporal patterns of different zones (districts) with a spatial analysis that considers the people's mobility among different zones, using a combined spatio-temporal model based on the results presented in Ref. [18].
- Using different validation approaches to measure the impact of mobility data when estimating one-week ahead COVID-19 incidence for different zones (districts) and different waves of infections of the same district. Injecting the spatial mobility data into the short term COVID-19 incidence estimation machine learning model, is able to generalize better over space (estimating the incidence for a new zone) than over time (estimating a new wave of infections for a particular zone).

The article is organized into 5 sections. The first section, this section, provides motivation for the study carried out in this manuscript and presents the objectives. The related previous research studies are presented in section 2. We focus on machine learning methods for predicting the short-term evolution for new COVID-19 infections. Section 2 shows a gap in the availability of previous studies that try to enhance predictions using mobility data. Section 3 describes the proposed mobility aware model together with the datasets that are used to assess the results. Section 3 also captures baseline model from previous research that will be used to assess the improvement in accuracy introduced by using mobility data. Section 3 also presents the method used to estimate human mobility based on the information obtained from a bike-sharing service. The results and discussion are captured in section 4. Finally, section 5 presents the principal conclusions of the work in this manuscript.

2. Related research

Since the outbreak of the COVID-19 pandemic, machine learning methods have been used to model the evolution of the COVID-19 virus. Some of the first studies used shallow machine learning techniques such as [16,17]. The research in Ref. [19] showed that even shallow machine learning models are able to generate promising results. The study concluded that machine learning models can show effective COVID-19 forecasting behavior paving the way to compare future studies [19]. Majhi et al. [20] used models such as decision trees and random forests in order to estimate the number of infections. Data from the initial weeks of the COVID-19 virus propagation in China was used to train the models which were then applied to the data in India. The authors assessed that their ML models were able to estimate the upcoming number of COVID-19 infections accurately.

Upcoming values for COVID-19 new infections, hospitalizations and deaths have also been estimated by using deep learning models

to try to achieve better predictions. Alassafi et al. [21] modeled the propagation of the virus in Malaysia, Morocco and Saudi Arabia. The study analyzed several Deep Learning (DL) architectures and estimated the number of confirmed infections and deaths using a prediction horizon of seven days. Shahid et al. [22] used the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the r2_score indices to analyze the results achieved by several deep leaning models including the Bidirectional Long Short-Term Memory (Bi-LSTM) anticipating values in temporal COVID-19 series. The study showed that some Deep learning architectures outperformed previous shallow models. Shastri et al. [23] compared similar deep learning models such as stacked LSTM, Bi-directional LSTM and convolutional LSTM in order to forecast COVID-19 infections in India and the USA, showing that convolutional LSTM was able to outperform the other two models.

Meta-studies using ML models designed to compare the accuracy in estimating COVID-19 data have also been published. Dairi et al. [24] captured a study that compared several machine learning (ML) models for anticipating the propagation of the virus. Such methods included the single Convolutional Neural Network CNN model, the single Long Short-Term Memory LSTM model, the combined ML model based on the use of Convolutional Neural Networks-Long Short-Term Memory (LSTM-CNN), the combined Gated Recurrent Unit-Convolutional Neural Networks (GRU-CNN) ML model, and the Restricted Boltzmann Machine (RBM) ML model. Logistic regression (LR) and support vector regression (SVR) were included as baseline references for assessing the results. Data from seven countries: Brazil, France, India, Mexico, Russia, Saudi Arabia, and the US was used to validate the accuracy of the models. Deep learning approaches combining LSTM-CNN and GRU-CNN were able to improve the results in COVID-19 forecasting. Nabi et al. [25] present a different comparative study comparing four deep learning models: LSTM, GRU, CNN, and Multivariate Convolutional Neural Network (MCNN). The convolutional neural networks outperformed the recurrent neural networks in the case of a limited availability of training data.

The COVID-19 virus spreads following spatio-temporal patterns based on human to human interactions. The spatial information capturing the propagation of the virus has been considered in several research studies that have tried to improve the predictions over time for the upcoming COVID-19 incidence values in a particular area by adding the temporal values in connected areas into a machine learning model. Huang et al. [17] made use of incidence data for COVID-19 in three European countries. The selected countries presented high infection rates at the beginning of the pandemic (Germany, Italy, and Spain) to feed a new COVID-19 space-aware machine learning model and were able to extract spatiotemporal features and to predict the number of confirmed cases. The results improved previous models that did not take the spatial information into account. Munoz-Organero et al. [26] added spatial information into a Deep Learning (DL) model that used sequences of COVID-19 incidence images as input. Spatial patterns were extracted using a Convolutional Neural Network (CNN) which was staked with a Long-Short Term Memory (LSTM) Recurrent Neural Network (RNN) to extract temporal patterns. The study validated the model using data from the 286 health zones in the Madrid region in Spain. The results showed improved prediction accuracy than previous models focusing on extracting only temporal patterns. Apart from Convolutional Neural Networks to extract spatial patterns in the spreading of the COVID-19 virus, models based on Graph Neural Networks (GNN) have also been used. GNN are able to define how different regions are interconnected, defining paths for the spatial propagation of the virus. Meznar et al. [27] validated that Graph Neural Networks (GNN) can be applied to estimate the evolution of an epidemic and showed the high utility of GNNs in order to provide a better intuition of the results. Deng et al. [28] designed a cross-location attention-based GNN (Cola-GNN) to extract patterns from time series embeddings for long term predictions showing strong predictive performance.

Human-to-human interactions are an important factor in the transmission of the COVID-19 virus and human mobility is a key element that controls the amount of interactions among people. The initial months in the spread of the COVID-19 virus came together with mobility restrictions to minimize the human-to-human transmission of the virus and highway traffic volumes were used as a proxy of activity and human interaction in Ref. [29] analyzing the drastic changes in human mobility due to the COVID-19. Micro-mobility patterns also showed significant changes in the first moths of the COVID-19 pandemic [30]. COVID-19 has had a major impact on traffic [25,26] and traffic volumes have significantly influenced the spread of the virus [27–29]. Lee et al. [31] found a correlation between traffic volumes and the propagation of COVID-19. The study used the data in South Korea. Similar correlations in impact on COVID-19 data due to the mobility of people in the two largest counties in Wisconsin are shown in Ref. [32]. The movement of people differentiated business foot traffic and considered other variables such as race, ethnicity and age [32]. The authors found characteristic patterns followed by different groups having an impact on the COVID-19 virus propagation speed. Other proxy variables to estimate the mobility of the people have also been proposed in previous research studies. Ayan et al. [33] proposed the use of the mobility of portable devices as a mechanism to estimate human mobility to add spatial information to a ML model that estimated the spread of the COVID-19 pandemic. Ayan et al. [33] used the data from 973 antennas in Rio de Janeiro and its suburbs from the cellular network to estimate human mobility. A Markovian model was used to capture mobility. Adding mobility estimated data to the machine learning model provided better accuracy values [33]. Human mobility was also estimated based on the use of the cellular network in Ref. [34] in order to enhance COVID-19 predictions. Rashed el al [34]. combined COVID-19 infection data with meteorological and human mobility data based on user connections to a major mobile phone carrier and an LSTM deep machine learning model in order to improve COVID-19 forecasts. The average relative error of the proposed model ranged from 16.1% to 22.6% in major regions of Japan. Rashed el al [35]. added COVID-19 variants information into the COVID-19 mobility aware predictive model in order to model the vaccination effectiveness.

In this research study, we propose a new machine learning model that extracts human mobility patterns from the bike-sharing service in a smart city. The mobility patterns are combined with a spatial representation of COVID-19 incidence data to forecast upcoming incidence values. Unlike previous mobility aware models such as [34] or [35], the implications of human mobility among different regions for the spatio-temporal spread of the virus are considered. The proposed model modulates the injection of new cases into a particular region based on confirmed cases in other regions and the estimated mobility of users among such regions. Validation is

based on data for the city of Madrid (Spain).

3. Materials and methods

This article analyses the performance of a new model combining COVID-19 incidence and human mobility data, estimates human mobility based on the information of citizens commuting by bike, validates the model using two open datasets for the city of Madrid (Spain) and estimates the model gain when comparing the results with a similar baseline model which does not take mobility data into account. The datasets used are presented in this section together with the mobility estimation method and the proposed ML model. Both geographical infection numbers and mobility data are combined to estimate predictions with a one-week prediction horizon. The baseline model is also captured.

3.1. Datasets

We use two large open datasets to validate the results in this article: the use of the electric bicycle service provided by the Municipal Transport Company (EMT) of Madrid city [36,37] and the Coronavirus Disease infection numbers for the different zones (districts) of the same city [38,39]. These datasets are provided by the regional government.

The use of the electric bike-sharing service in Madrid can be downloaded from Ref. [36]. The data files comprise the historical data from April 2017 until June 2021 for 264 stations mainly located in the city's central districts. The dataset offers information about all the rides organized in monthly files using a json format. For each ride, the dataset provides information about the user ID; the type of the user; ID of the origin station; ID of the destination station; IDs of the bases (socket numbers) in both the origin; and destination stations and time and duration of the ride.

The locations for the bike stations can be found in Ref. [37]. The latitude and longitude coordinates are provided for each bike station together with descriptive information about the name and postal address of the station and the number of bikes that can be plugged in (number of bases). The id linking to the district where the station is placed is also provided.

For each time period, the bike-sharing service's mobility is estimated by combining the location of each bike station and the number of rides among each pair of stations. The bike-sharing service captures only a small part of the mobility of the people in the city. A method is proposed in the next subsection to estimate the total mobility based on the number of bike rides between districts.

Fig. 1 shows the location of the bike stations in the different districts of the city of Madrid. The districts located in the center of the city have a higher density of stations while the service is not yet offered in the city's suburbs. The district with the highest number of stations is Madrid-Centro. Fig. 1 captures the stations in Madrid-Centro in blue to visually show the difference in density as compared with other districts.

The dataset containing epidemiologic information (infections, hospitalizations and deaths) for the COVID-19 virus in the Community of Madrid can be downloaded from Ref. [38]. The dataset includes data about the new cases reported for each health zone every week (together with information about the number of hospital admissions and deaths). There are 286 health zones in the area of the Community of Madrid (Spain). Each health zone is the area where a primary care center provides health services. Primary care centers used PCR tests to track new infections in the health zone. The dataset contains information about where the different heath zones are located. Health zones are grouped into districts. There are 21 districts covering Madrid city. Fig. 2 captures the locations of the centers for each of the 143 health areas in Madrid and the perimeter of each district in the city. The areas in the suburbs of the city are less populated and the health zones are larger in space. The reported numbers for infections are summed up for each district in Ref. [39] providing the same geographical division as the one reported for electric bicycles [36].

The data files in Refs. [38,39] are divided into three significant periods of time in which different protocols were used to count the number of new infections. The first period extends from the first cases reported (February 25th, 2020) until July 1st, 2020 in which

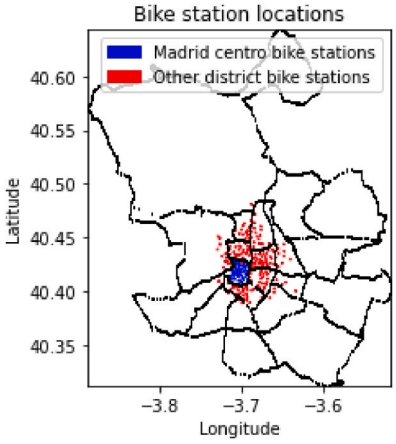


Fig. 1. Placement of the electric bike stations in Madrid.

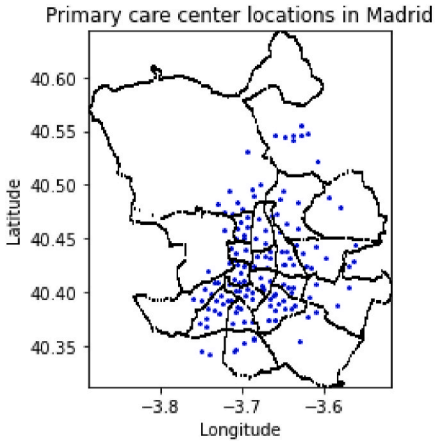


Fig. 2. PCR testing sites in Madrid.

COVID-19 data about new cases and hospitalizations was reported daily. As new information about the virus was known week after week, the procedures for detecting new cases had to be adjusted several times. PCR tests ware not always available in enough quantities to fulfil the requirements in the first months of the pandemic, having an impact in the accuracy of available data for this first period of the pandemic. Incidence data was reported weekly in the second period. This second period goes from July 2nd, 2020 to March 29th, 2022. The procedures for detecting and counting the new COVID-19 infections were more stable in the second period and available data provides a more homogeneous and reliable picture of the evolution of the spread of the virus. Finally, there was a major change in COVID-19 regulations in April 2022, and many restrictions were lifted. The requirements to use PCR tests for identifying new infections were targeted to some specific cases (in particular, for people over 60). The datasets in Refs. [38,39] only capture data for the population over 60 since April 2022.

The period from July 2nd, 2020 until March 29th, 2022 will be used. The information-collecting methods in this period have been stable. From July 2nd, 2020, to June 2021, both COVID-19 incidence and bike-sharing information are available [36]. The size of the bike-sharing service dataset [36,37] used is around 600 Mbytes. The size of the COVID-19 dataset [38,39] for the same period of time is 10.2 Mbytes.

3.2. Human mobility estimation based on the number of bike rides

The spreading of the infections is driven by the mobility of the people in a region. This section proposes a mechanism to estimate the human movements in Madrid by using the electric bicycle rides information.

The city is divided into districts $\{a,b,c\ldots\}$. Each district has a number of bike stations $\{n_a,n_b,n_c\ldots\}$. Let's assume that people move randomly inside the city. For each person i, the probability of moving from one part of the city a to another b using the bike-sharing service can be captured using a random variable p_{iab}. Suppose that p_{iab} depends on the distance from the user to the nearest bike station and the distance from the destination bike station to the destination of the movement. And assume that the origin and destination locations of each movement for user i are also random variables o_i and d_i. We define d_{soi} to the distance between the bike station s and the origin o_i for user I, and d_{udi} to the distance between the bike station u and the destination d_i for user i. Assuming that the probability of using the service is inversely proportional to the distances d_{soi} d_{udi} and considering that o_i and d_i are independent variables, the probability of using the service from o_i and d_i from stations s to u can be approximated as captured in equation (1).

$$p_{iod} \cong \frac{k}{d_{soi}d_{udi}} \tag{1}$$

where k is a constant (in our case, to simplify computations, we assume that k does not depend on the user i).

There are n_a stations in district a and n_b in district b. The probability of using the bike-sharing service to go from the district a to district b (any station at a and any station at b) can be approximated as:

$$p_{iab} = 1 - \overline{p}_{iab} = 1 - \prod_{su}\left(1 - \frac{k}{d_{soi}d_{udi}}\right) \tag{2}$$

Being $\overline{p}_{iab}$ the probability of not using any of the stations to go from a to b.
Assuming that p_{iod} is small, equation (2) can be approximated as shown in equation (3).

$$p_{iab} \cong \sum_{su}\frac{k}{d_{soi}d_{udi}} \tag{3}$$

The number of rides between a and b using the bike-sharing service can be approximated using the probability in equation (3) as the

summation of the probabilities for all the users i that travel from district a to b to use the service (being $m_{iab} = 1$ for users that travel from a to b and 0 otherwise):

$$r_{ab} \cong \sum_i m_{iab} p_{iab} \cong \sum_i m_{iab} \sum_{su} \frac{k}{d_{soi} d_{udi}} \tag{4}$$

Since o_i and d_i are random variables, d_{soi} d_{udi} are also random variables. d_{soi} d_{udi} will be inversely proportional to the number of bike stations in a and b. We can approximate equation (4) as proposed in equation (5).

$$r_{ab} \cong k' \sum_i m_{iab} n_a n_b \tag{5}$$

where k' is a constant, n_a is the number of stations in district $a$and n_b is the number of stations in district b.

The number of people travelling from district a to district b can therefore be estimated as proposed in equation (6).

$$m_{ab} = \sum_i m_{iab} \cong r_{ab} \frac{1}{k' n_a n_b} \tag{6}$$

The number of movements between districts a and b (m_{ab}) will therefore be estimated based on the number of rides reported in the dataset in Ref. [36] divided by $n_a n_b$ in order to normalize the effect of having different number of base stations to measure the mobility of people. The values m_{ab} will also be scaled using a constant k'' different from k' to facilitate the learning of the model described in the following subsection.

Fig. 3 captures an example image generated using Equation (6) for a given day in Madrid. The image has been scaled so that the maximum value is 2. The image contains 21 by 21 pixels. The value for each pixel represents the relative estimation of people travelling from an origin district (row) to a destination district (column). The diagonal of the image captures movements inside each district. Movements from a to b are not necessarily the same as movements from b to a. Since the bike sharing service does not provide stations for all the 21 districts, some values in the image are 0 (those districts have not been included in the model since there is no data to estimate human mobility for them). The number of bike stations in some other districts is so small that the approximation in Equation (6) is not good enough (those districts have also been left apart and not used in the model). We have selected the best eight districts to train the model with significant data to provide mobility estimates: Centro, Arganzuela, Retiro, Salamanca, Chamartín, Tetuán, Chamberí and Moncloa-Aravaca.

One limitation of using the information provided by the bike sharing service as a proxy variable to estimate human mobility is that the service tends not to be used by children and elderly people. The statistical data of utilization of the service for each age group is captured in Table 1. The majority of the users are between 17 and 65 years old. Other proxy variables such as public transportation and private vehicle use will be used in future studies.

3.3. Mobility enhanced model

The propagation of the COVID-19 virus depends on the current incidence (the higher the number of infected cases, the more likely to be close to them and get infected) and the mobility of the people (higher mobility numbers increase the number of human-to-human

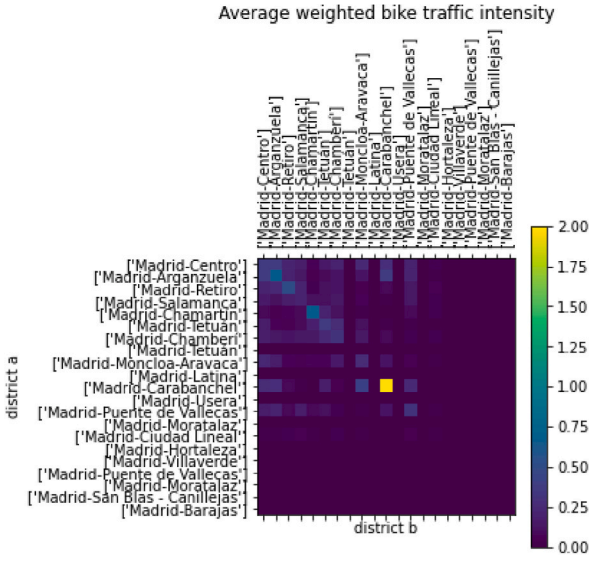


Fig. 3. Scaled image capturing the relative movements between districts.

Table 1
Percentage of people using the bike-sharing service per age ranges.

Age range	% of people using the service
0–16 years old	2.228459736
17 and 18 years old	4.960723613
19–26 years old	11.60148281
27–40 years old	29.94036115
41–65 years old	15.86883815
More than 65 years old	0.723255057

interactions increasing the probability of getting infected). To capture both factors (incidence and mobility) into a model that can extract temporal COVID-19 incidence patterns affected by spatial mobility patterns and predict upcoming incidence values for each spatial location (district), the model in Fig. 4 is proposed. The temporal data for infections ($\bar{x}_a$) for each district a, is used as the input for an LSTM-based RNN to learn temporal patterns. For the districts sharing bike journeys that inject traffic into district a, a similar RNN-based layer is applied to the combined information of past values for COVID-19 incidence ($\bar{x}_i$) and the mobility estimation to district a ($\bar{z}_i$) based-on Equation (6). An independent RNN is applied to each district i to learn the influence of each district i on the upcoming incidence values for district a. All the LSTM-based RNN layers are configured to have the same number of memory units (assuming that the temporal patterns will share a similar complexity from all input signals). The output for each Recurrent Neural Network (RNN) layer is summarized using a dense layer, and the spatial contribution (the influence of districts injecting people and therefore propagating the virus into district a) is captured by combining the information from all the districts using an extra fully connected layer. The output of the model will be trained to forecast infections with a prediction horizon of 7 days.

The model in Fig. 4 introduces novel components for optimizing the estimation of COVID-19 upcoming incidence values as compared with previous models:

- Use of open mobility data provided by a bike-sharing service in order to estimate human mobility flows (not just mobility aggregates) among spatial zones
- Combined temporal and spatial pattern extraction from time series data based on a zonal spatial division that uses estimated mobility flows to modulate the aggregation of inter-zonal COVID-19 incidence data to model the spread if the virus.

3.4. Baseline modelA baseline model is also proposed to evaluate the model's gain obtained when using mobility data. The baseline model is captured in Fig. 5. The model uses a similar structure but only applied to the temporal component of COVID-19 incidence

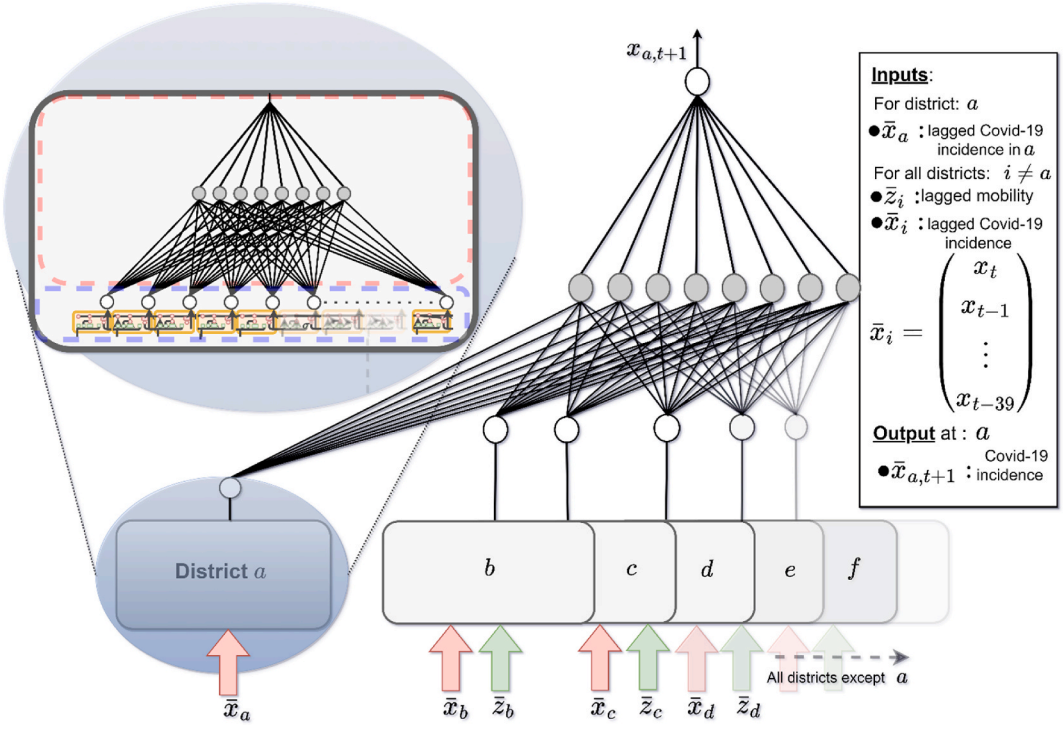


Fig. 4. Proposed model for mobility-enhanced COVID-19 forecasting.

data. No mobility data is considered and the upcoming values for COVID-19 incidence are estimated using only current and past incidence values. The architecture used to learn the time patterns is the same as in Fig. 4. An LSTM-based RNN is used to learn the temporal patterns and the output is also summarized using a similar dense layer as in Fig. 4.

4. Results and discussion

4.1. Validation schemes

The models in Figs. 4 and 5 will be trained and validated using the data in Refs. [36–39] (data for the use of the bike-sharing service [36,37] and data for COVID-19 per district incidence values [38,39]). Only the months from July 2020 and June 2021 will be used. In July 2020, the protocol to measure COVID-19 infections was changed (not compatible with previous data), so only data from July 2020 is useful for training the proposed machine learning models. The bike-sharing service only provides data until June 2021, so the period from July 2020 to June 2021 will be used (being the maximum possible which contains simultaneously data from both datasets).

Three different validation schemes will be used to measure the accuracy of the models:

- A 5-fold cross-validation approach. using 80% of the data for training and 20% for validation. The process will be repeated 5 times using different samples in the validation set so that each sample is used once for validation. The average accuracy values are computed to assess the models.
- COVID-19 has generated waves of infections. During the period used to train and validate the models, 3 different waves have been generated by the propagation of the virus. The information learnt by the model for previous infection waves should be able to generalize for new waves. The second validation scheme consists of a leave-one-wave-out cross-validation. This validation approach is more demanding for the machine learning models compared to the 5-fold cross-validation since the information in the training and validation sets are likely to be less similar.
- Finally, a cross-validation approach based the idea of using the information for all the districts except one to train the model and use the trained model to assess the results for the district left apart (leave-one-district out) is proposed. In the case of the proposed ML model being able to extract the spatial influence of combined incidence and mobility patterns from other districts, this model should behave better than the leave-one-wave-out approach (which tries to perform a generalization over time).

4.2. Optimizing the model internal parameters

Both models, in Figs. 4 and 5, are designed with two major parameters (the number of memory units in the LSTM layer and hidden neurons in the fully connected layer) that can be modified to adapt the complexity of the model to the complexity of the input data. The nature of the input signals in the different districts is similar and we assume that their complexity will also be similar so the same number of memory units is used for all the LSTM layers. A 20% summarization has been used to relate the amount of memory cells and hidden neurons of the fully connected layer in both models (Figs. 4 and 5).

Both the baseline and the proposed model have been validated using different values for the amount of memory cells. The data from the last 4 weeks is used to feed the models in order to estimate one-week ahead incidence values. The three validation schemes presented in section 4.1 are implemented, and the average results are captured in Figs. 6 and 7. The Root Mean Square Error (RMSE) is

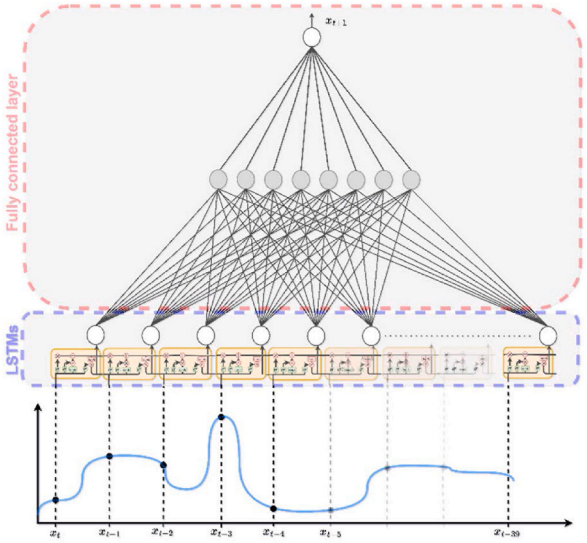


Fig. 5. Baseline model.

used to assess the models' accuracy. Fig. 6 shows the RMSE average values for the mobility-aware model in Fig. 4. Fig. 7 presents similar results for the baseline model in Fig. 5. The difference in the accuracy for both models for the different values in the number of memory units in the LSTM cells (gain of the proposed model in Fig. 4) is captured in Fig. 8. Both COVID-19 incidence and bike mobility values have been scaled using a linear scaler (scaling signals to lie between 0 and 1) to facilitate the learning of the models in Figs. 4 and 5. The model in Fig. 4 performs better than the baseline model for all the cases, except when the number of memory cells is 10 in which results are similar. The optimal results for both models are achieved when the number of memory units is 40. Using a small number of memory units generates configurations for the machine learning model, which are not able to optimally learn the complexity of the patterns in the data. Using a high number of memory units is likely to cause overfitting of the models. In the case of using 40 memory units in the LSTM cells, the RMSE for the proposed model in Fig. 4 is 0.0205, and the RMSE for the baseline model in Fig. 5 is 0.0229. The proposed model in Fig. 4 outperforms the base-line model in Fig. 5 by a 11.7% (model gain). The concomitant use of human mobility data and COVID-19 incidence, together with the method proposed in Section 3 to estimate human mobility, enable our model (Fig. 4) to outperform a mobility-agnostic (baseline) model in one-week ahead predictions of COVID-19 incidence values.

The total number of trainable weights for the model in Fig. 4 configured with the optimal parameters (40 memory units in the LSTM cells) is 82849. The time complexity for the evaluation of a test sample can be approximated following [40] by $O(n_d * (n * ((4d+4n+3)+h_d * (n+1))$ where n (40) is the number of LSTM memory units, h_d is the number of hidden units in the dense layers, n_d the number of districts, and d the number of the input signals (2 in our case, for the COVID-19 incidence data and mobility data) For an uncluttered expression, we have excluded the biases. The configuration of the model in Fig. 4 for the optimal results in Fig. 6 are captured in Table 2.

4.3. Results for each validation scheme

Figs. 6–8 show the average results for all the validation schemes described in sub-section 4.1. Each validation scheme has its particularities. Fig. 9 captures the RMSE values for each validation method for the optimal configuration of the proposed model in Fig. 4. The best RMSE values are achieved for the 5-fold cross-validation method, in which validation data is more similar to the training data. 20% of the dataset is selected in the 5-fold cross-validation approach for validation and 80% for training the model (the process is repeated five times so that each data sample is used once for validation). The training set contains information from all the districts and all the waves in the entire period (the same as the validation set). The patterns learnt from the training set are more likely to be present in the validation samples.

Fig. 9 also shows that the RMSE values achieved when using a leave-one-district-out approach are better than those reached by the leave-one-wave out approach. The proposed model in uses mobility information from other districts to capture the spatial patterns in the disease's dissemination. These mobility patterns have a similar influence on all the districts. The proposed mobility-enhanced model (Fig. 4) generalizes across districts since it performs equally well in 5-fold-cross and leave-one-district-out validation settings. However, in the case of leaving-one-wave-out, the mobility information (spatial component of the model) has a more limited capacity for generalizing over time. Thus, the RMSE values are worse.

Fig. 10 shows the one-week ahead predictions for the Madrid-Centro district and the 5-fold cross-validation approach. The dataset contains information from July 2020 to June 2021. The model in Fig. 4 uses the last four weeks of data to generate one-week ahead predictions. The first predictions are therefore generated for the second week of August, as shown in Fig. 10.

4.4. Results for each district

The model in Fig. 4 uses the mobility data to improve the estimations for the upcoming values of COVID-19 incidence. The mobility is estimated based on the number of rides between districts [36], as presented in section 3. The density of bike stations [37] is different for each district. This section uses the leave-one-district-out validation approach to measure the model's accuracy for each district. The intuition is that districts with a higher number of bike stations will be able to provide better estimations for the mobility of people according to the method presented in section 3. Better estimations for the mobility data should positively impact the accuracy of the

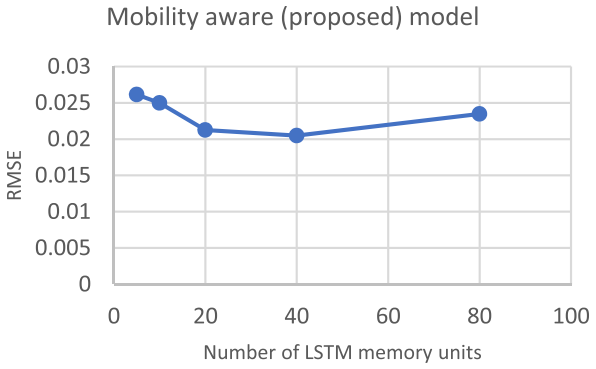


Fig. 6. RMSE values for the model in Fig. 4 for different numbers of memory units.

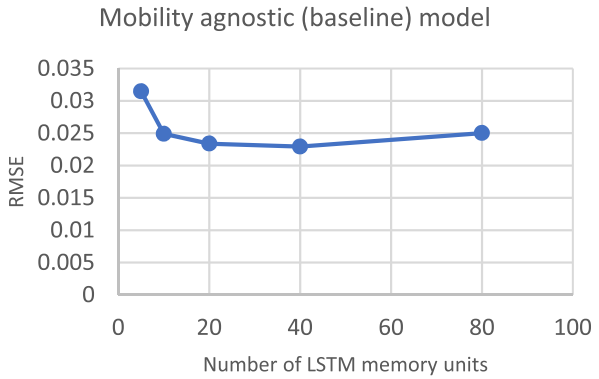


Fig. 7. RMSE values for the model in Fig. 5 for different numbers of memory units.

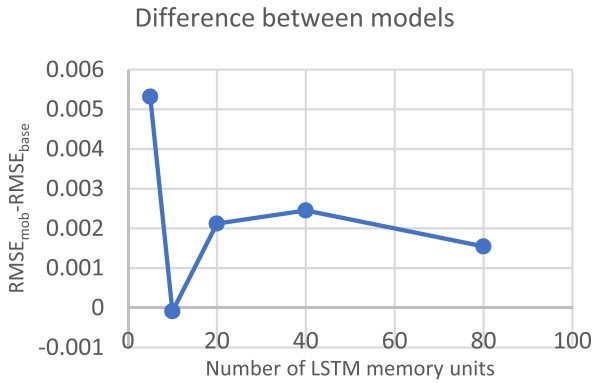


Fig. 8. Accuracy gain when adding mobility data.

predictions for those districts.

Fig. 11 presents de RMSE accuracy results for each district when leaving one district out for validation. The model provides optimal results for the Centro and Salamanca districts which are the ones with a higher number of bike stations. The worse results are achieved for Tetuán and Moncloa-Aravaca districts which are those with a lower number of bike stations. Fig. 12 captures the relation between the number of stations and the accuracy of the proposed model in Fig. 4 when leaving one district out when training the model.

A second parameter that could impact the accuracy of the estimation of human mobility based on the use of bike-sharing service is the average number of rides starting from each district. Districts with low use of the bike-sharing service may get worse estimations for total human mobility for the people of those districts. Fig. 13 captures the model's accuracy depending on the average number of daily rides for all the districts. As in the previous case in Fig. 12, the two districts with higher use of the service obtain optimal results (from the model in Fig. 4). The two districts with the lowest use of the service achieve the worst RMSE values when using the proposed model in Fig. 4 to estimate the upcoming values for COVID-19 incidence one week ahead.

4.5. Results from models in previous studies

In order to compare the results for the proposed human mobility aware model in this paper with previous models presented in previous related studies, the MAPE (Mean Absolute Percentage Error) as defined by equation (7) is used. COVID-19 incidence values depend on data such as population density (the spread of the virus is dependent on the area of study) and observation times (COVID-19 generates temporal waves of different magnitudes depending on the virus variant and vaccination campaigns). The MAPE error provides a figure that compensates the population size in different areas as provided by previous studies and has been selected for this section in order to provide a comparison with previous studies. However, the MAPE error has some limitations since it is influenced by the temporal shape of the COVID-19 incidence waves, assigning a greater importance to the time windows when the incidence values are smaller.

$$E = \frac{1}{N}\sum_{i}\frac{|y_i - \hat{y}_i|}{y_i} \tag{7}$$

Three major studies have been selected from previous literature. The model in Ref. [18] presents a machine learning model that is applied to the same region of study (although to a different time window) and has been compared with previous models in literature

Table 2
Optimal model configuration parameters.

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 4, 1)]	0	[]
input_2 (InputLayer)	[(None, 4, 2)]	0	[]
input_3 (InputLayer)	[(None, 4, 2)]	0	[]
input_4 (InputLayer)	[(None, 4, 2)]	0	[]
input_5 (InputLayer)	[(None, 4, 2)]	0	[]
input_6 (InputLayer)	[(None, 4, 2)]	0	[]
input_7 (InputLayer)	[(None, 4, 2)]	0	[]
input_8 (InputLayer)	[(None, 4, 2)]	0	[]
input_9 (InputLayer)	[(None, 4, 2)]	0	[]
lstm_1 (LSTM)	(None, 40)	6720	['input_1[0][0]']
lstm_2 (LSTM)	(None, 40)	6880	['input_2[0][0]']
lstm_3 (LSTM)	(None, 40)	6880	['input_3[0][0]']
lstm_4 (LSTM)	(None, 40)	6880	['input_4[0][0]']
lstm_5 (LSTM)	(None, 40)	6880	['input_5[0][0]']
lstm_6 (LSTM)	(None, 40)	6880	['input_6[0][0]']
lstm_7 (LSTM)	(None, 40)	6880	['input_7[0][0]']
lstm_8 (LSTM)	(None, 40)	6880	['input_8[0][0]']
lstm_9 (LSTM)	(None, 40)	6880	['input_9[0][0]']
dense_1 (Dense)	(None, 32)	1312	['lstm_1[0][0]']
dense_2 (Dense)	(None, 32)	1312	['lstm_2[0][0]']
dense_3 (Dense)	(None, 32)	1312	['lstm_3[0][0]']
dense_4 (Dense)	(None, 32)	1312	['lstm_4[0][0]']
dense_5 (Dense)	(None, 32)	1312	['lstm_5[0][0]']
dense_6 (Dense)	(None, 32)	1312	['lstm_6[0][0]']
dense_7 (Dense)	(None, 32)	1312	['lstm_7[0][0]']
dense_8 (Dense)	(None, 32)	1312	['lstm_8[0][0]']
dense_9 (Dense)	(None, 32)	1312	['lstm_9[0][0]']
concatenate_1 (Concatenate)	(None, 288)	0	['dense_1[0][0]', 'dense_2[0][0]', 'dense_3[0][0]', 'dense_4[0][0]', 'dense_5[0][0]', 'dense_6[0][0]', 'dense_7[0][0]', 'dense_8[0][0]', 'dense_9[0][0]']
dense_10 (Dense)	(None, 32)	9248	['concatenate_1[0][0]']
dense_11 (Dense)	(None, 1)	33	['dense_10[0][0]']

Total params: 82,849.
Trainable params: 82,849.
Non-trainable params: 0.

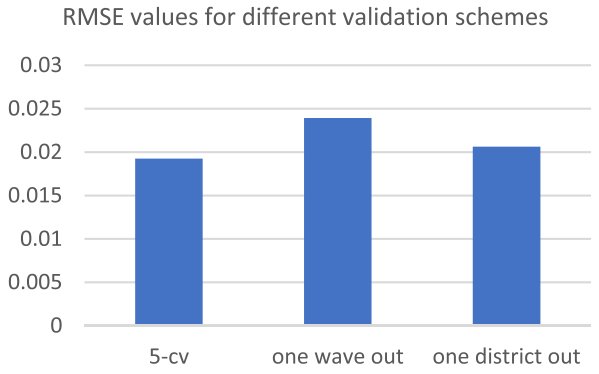


Fig. 9. RMSE values for the model in Fig. 3 for all the validation schemes.

showing better performance values. The study in Ref. [24] presents a review of different models applied to different parts of the World for which MAPE figures are provided. We have selected the best performing model based on an LSTM-CNN architecture in Ref. [24] in order to provide an optimal performing model for comparison. The model in Ref. [34] has also been included since it added the information related to user mobility to the input of the model. The MAPE error of the model in Ref. [34] ranged from 16.1% to 22.6% in major regions of Japan. The proposed model in this paper also includes human mobility estimation data in order to optimize the

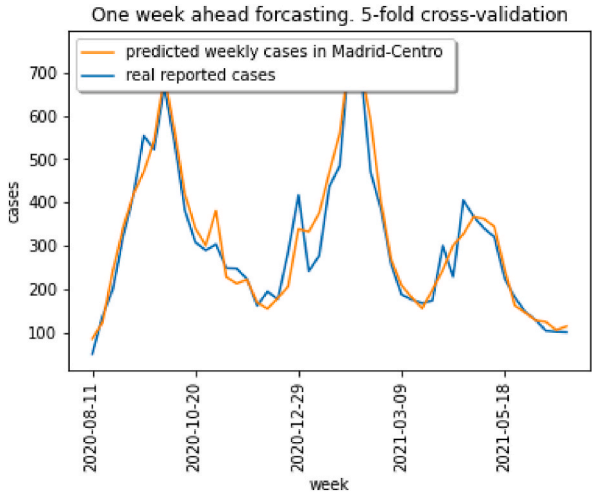


Fig. 10. One-week ahead predictions for the Madrid-Centro district.

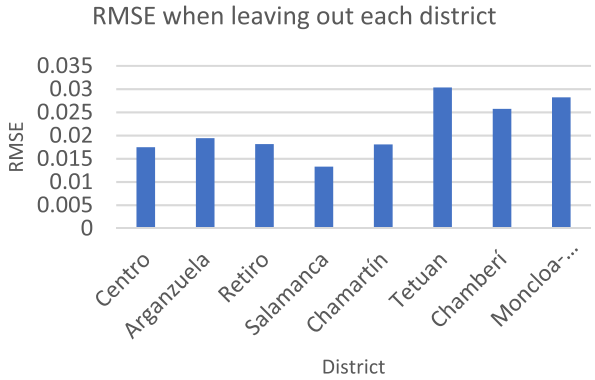


Fig. 11. Accuracy of the proposed model for each district when using the leave one district validation.

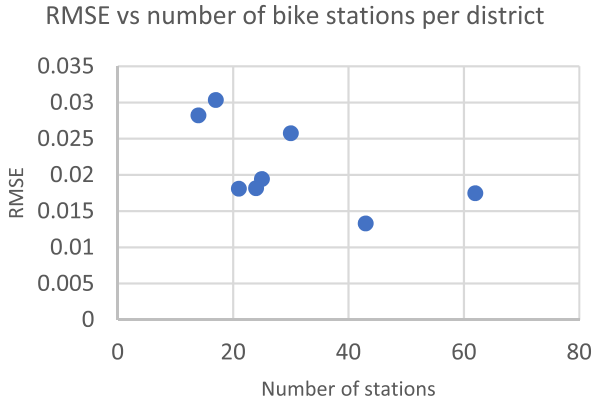


Fig. 12. RMSE vs number of stations in each district.

performance of a space-distributed LSTM model. In our case, the proposed model is able to estimate human mobility flows (not only aggregated mobility values) and trains the model in order to capture the virus spreading patterns among different areas. The comparative results are captured in Table 3.

The model proposed in this paper is able to outperform the study in Ref. [34]. The results for the previous model in Ref. [18] for the same geographical area are also optimized. The comparison of results with the review in Ref. [24] shows that the proposed model is

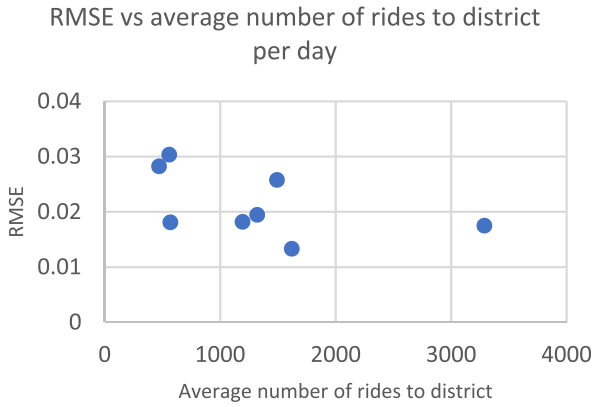


Fig. 13. RMSE vs the average number of rides per day for each district.

able to outperform the best model in Ref. [24] for some areas but not for all of them. The results in Ref. [24] show that the same model provides different results when applied to different regions and different temporal observational windows. The MAPE values for Brazil, India and the US are 6.48, 6.02 and 1.63 respectively. The models in Ref. [24] are trained to estimate the total number of infections instead of the upcoming new infections. Since the MAPE values are higher when estimating small values (used as the denominator in the calculation formula), the results in Ref. [24] would be higher when used in a similar scenario as the one used in this paper.

5. Conclusions

A new model that uses together human mobility and COVID-19 infection information as inputs to provide short-term (prediction horizon of one-week) forecasts for COVID-19 new infections has been proposed and validated in this manuscript. The model combines both spatial and temporal information to optimize predictions. The spatial component combines the reported COVID-19 incidence values in each city's district and the dissemination of the virus among the districts based on the estimation of human mobility. A new method to estimate human mobility based on the use of the bike-sharing service in a city has been proposed. The temporal component uses the time series for both COVID-19 incidence and mobility data. An RNN model based on LSTM cells is used to extract the time patterns, and a dense layer provides the spatial analysis in the proposed model. This model is compared against a baseline version that lacks the spatial component to assess the benefit of factoring in information about mobility through the bike-sharing service. The results of the model have also been compared with similar previous studies.

The datasets in Refs. [36–39] have been used to train both the proposed mobility-aware and baseline models. Three different validation approaches (5-fold cross-validation, leave-one-wave-out validation, and leave-one-district-out validation) have been used. The proposed mobility-enhanced model shows an 11.7% gain compared with the baseline model which uses a similar structure to detect time patterns but no spatial dependencies are used in the baseline model. Adding the mobility information has a similar positive influence on all the districts and provides better results for the leave-one-district-out validation than in the case of leaving-one-wave-out, showing that the mobility information (spatial component of the model) has a more limited capacity for generalizing over time than over space. The results have been validated for one year-data for the city of Madrid. The generalization to other areas and time frames will be done as a future work.

The number of bike stations in each district is key to a better estimate of total mobility values. With the proposed model, districts with higher number of stations achieved better results than districts with lower number of stations. Corresponding the best result to the district with the largest number of stations. Having a low number of stations may discourage people in the district from using the service and use other means of transportation instead. The number of rides per district also shows a negative correlation with the accuracy of the predictions when using the proposed model for estimating upcoming COVID-19 incidence values for a particular district. Districts showing the lowest use of the bike-sharing service achieve the worst predictions and vice versa. A limitation of the proposed method to estimate human mobility based on the use of the bike sharing service is the different use by different age ranges. The majority of the users are between 17 and 65 years old. Other proxy variables such as public transportation and private vehicle use will be used in future studies.

Table 3
Comparison with previous studies.

Ref	Model	Region	MAPE
[18]	LSTM-CNN	Madrid (Spain)	5.5
[24]	LSTM-CNN	Brazil, India, US	4.71
[34]	LSTM	Japan	16.1
Our model	Space-distributed_LSTM	Madrid (Spain)	4.9

Author contribution statement

Mario Muñoz-Organero: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.Patricia Callejo and Miguel Ángel Hombrados-Herrera: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper..

Data availability statement

The data used in this paper is publically available as open data [36–39].

Additional information

No additional information is available for this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is part of the agreement between the Community of Madrid and the Universidad Carlos III de Madrid for the funding of research projects on SARS-CoV-2 and COVID-19 disease, project name “Multi-source and multi-method prediction to support COVID-19 policy decision making”, which was supported with REACT-EU funds from the European regional development fund “a way of making Europe”. This work was supported in part by the projects “Real time social sensor AnaLysis and deep learning based resource EStimations for multimodal transport” MaGIST-RALES, funded by the Spanish Agencia Estatal de Investigación (AEI, doi: 10.13039/501100011033) under grant PID2019-105221RB-C44/AEI/10.13039/501100011033 and “FLATCITY-APP: Mobile application for FlatCity” funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación MCIN/AEI/10.13039/501100011033 and the European Union “NextGenerationEU/PRTR” under grant PDC2021-121239-C33.

References

[1] D.H. Barouch, Covid-19 vaccines—immunity, variants, boosters, N. Engl. J. Med. 387 (11) (2022) 1011–1020.
[2] Julien Arino, Describing, modelling and forecasting the spatial and temporal spread of COVID-19: a short review, Mathematics of Public Health (2022) 25–51.
[3] A. Hasan, E.R. Putri, H. Susanto, N. Nuraini, Data-driven modeling and forecasting of COVID-19 outbreak for public policy making, ISA Trans. (2021), https://doi.org/10.1016/j.isatra.2021.01.028.
[4] D.M. Kent, J.K. Paulus, R.R. Sharp, N. Hajizadeh, When predictions are used to allocate scarce health care resources: three considerations for models in the era of Covid-19, Diagnostic and prognostic research 4 (1) (2020) 1–3.
[5] M. Ghaderzadeh, M. Aria, Management of covid-19 detection using artificial intelligence in 2020 pandemic, in: 2021 5th International Conference on Medical and Health Informatics, 2021, May, pp. 32–38.
[6] M. Ghaderzadeh, F. Asadi, Deep learning in the detection and diagnosis of COVID-19 using radiology modalities: a systematic review, Journal of healthcare engineering 2021 (2021).
[7] E. Koç, M. Türkoğlu, Forecasting of medical equipment demand and outbreak spreading based on deep long short-term memory network: the COVID-19 pandemic in Turkey, Signal, image and video processing 16 (3) (2022) 613–621.
[8] M. Masum, M.A. Masud, M.I. Adnan, H. Shahriar, S. Kim, Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for COVID-19 forecasting and management, Soc. Econ. Plann. Sci. 80 (2022), 101249.
[9] Y. Xiang, Y. Jia, L. Chen, L. Guo, B. Shu, E. Long, COVID-19 epidemic prediction and the impact of public health interventions: a review of COVID-19 epidemic models, Infectious Disease Modelling 6 (2021) 324–342.
[10] Y. Zhu, Y.Q. Chen, On a statistical transmission model in analysis of the early phase of COVID-19 outbreak, Stat. Biosci. 13 (2021) 1–17.
[11] F. Baldo, L. Dall'Olio, M. Ceccarelli, R. Scheda, M. Lombardi, A. Borghesi, S. Diciotti, M. Milano, Deep Learning for Virus-Spreading Forecasting: A Brief Survey. arXiv, 2021 arXiv: 2103.02346.
[12] Sweeti Sah, R.dhanalakshmi Surendiran, Sachi Nandan Mohanty, Fayadh Alenezi, Kemal Polat, Forecasting COVID-19 pandemic using prophet, ARIMA, and hybrid stacked LSTM-GRU models in India, Comput. Math. Methods Med. 2022 (2022).
[13] Debaditya Shome, T. Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Mu-hammad, Abdullah AlTameem, Yazhou Zhang, Abdul Khader Jilani Saudagar, COVID-transformer: interpretable COVID-19 detection using vision transformer for healthcare, Int. J. Environ-mental Research and Public Health 18 (2021) 21 1–2114.
[14] Sandeep Kumar Satapathy, Shreyaa Saravanan, Shruti Mishra, Sachi Nandan Mohanty, A comparative analysis of multidiemnsional COVID-19 poverty determinants: an observational machine learning approach, New Generat. Comput. 41 (Issue 1) (2023).
[15] L. Wang, T. Xu, T. Stoecker, H. Stoecker, Y. Jiang, K. Zhou, Machine learning spatio-temporal epidemiological model to evaluate Germany-county-level COVID-19 risk, Mach. Learn. Sci. Technol. 2 (2021), 035031.
[16] F. Lorig, E. Johansson, P. Davidsson, Agent-based social simulation of the COVID-19 pandemic: a systematic review, JASSS: J. Artif. Soc. Soc. Simulat. 24 (3) (2021).
[17] C.J. Huang, Y. Shen, P.H. Kuo, Y.H. Chen, Novel spatiotemporal feature extraction parallel deep neural network for forecasting confirmed cases of coronavirus disease 2019, Socio-Econ. Plan. Sci. 80 (2020), 100976.
[18] M. Muñoz-Organero, Space-Distributed Traffic-Enhanced LSTM-Based Machine Learning Model for COVID-19 Incidence Forecasting, Computational Intelligence and Neuroscience, 2022, 2022.
[19] S. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. Varkonyi-Koczy, U. Reuter, T. Rabczuk, P. Atkinson, Covid-19 outbreak prediction with machine learning, Algorithms 13 (2020) 249.
[20] R. Majhi, R. Thangeda, R.P. Sugasi, N. Kumar, Analysis and prediction of COVID-19 trajectory: a machine learning approach, J. Public Aff. 21 (2021), e2537.
[21] M.O. Alassafi, M. Jarrah, R. Alotaibi, Time series predicting of COVID-19 based on deep learning, Neurocomputing 468 (2022) 335–344.

[22] F. Shahid, A. Zameer, M. Muneeb, Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm, Chaos, Solit. Fractals 140 (2020), 110212.
[23] Sourabh Shastri, Kuljeet Singh, Sachin Kumar, Paramjit Kour, Vibhakar Mansotra, Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study, Chaos, Solit. Fractals 140 (110227) (2020), https://doi.org/10.1016/j.chaos.2020.110227. ISSN 0960-0779.
[24] A. Dairi, F. Harrou, A. Zeroual, M.M. Hittawe, Y. Sun, Comparative study of machine learning methods for COVID-19 transmission forecasting, J. Biomed. Inform. 118 (2021), 103791.
[25] NABI, Khondoker Nazmoon, et al., Forecasting COVID-19 cases: a comparative analysis between recurrent and convolutional neural networks, Results Phys. 24 (2021), 104137.
[26] M. Muñoz-Organero, P. Queipo-Álvarez, Deep spatiotemporal model for COVID-19 forecasting, Sensors 22 (9) (2022) 3519.
[27] S. Mežnar, N. Lavrač, B. Škrlj, Prediction of the effects of epidemic spreading with graph neural networks, in: R.M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, M. Sales-Pardo (Eds.), Complex Networks & Their Applications IX, Springer International Publishing, Cham, Switzerland, 2021, pp. 420–431.
[28] S. Deng, S. Wang, H. Rangwala, L. Wang, Y. Ning, Cola-GNN: Cross-Location Attention Based Graph Neural Networks for Long-Term ILI Prediction, Association for Computing Machinery, New York, NY, USA, 2020, pp. 245–254.
[29] S. Parr, B. Wolshon, J. Renne, P. Murray-Tuite, K. Kim, Traffic impacts of the COVID-19 pandemic: statewide analysis of social separation and activity restriction, Nat. Hazards Rev. 21 (3) (2020).
[30] A. Li, P. Zhao, H. Haitao, A. Mansourian, K.W. Axhausen, How did micro-mobility change in response to COVID-19 pandemic? A case study based on spatial-temporal-semantic analytics, Comput. Environ. Urban Syst. 90 (2021), 101703.
[31] H. Lee, S.J. Park, G.R. Lee, J.E. Kim, J.H. Lee, Y. Jung, E.W. Nam, The relationship between trends in COVID-19 prevalence and traffic levels in South Korea, Int. J. Infect. Dis. 96 (2020) 399–407.
[32] X. Hou, S. Gao, Q. Li, Y. Kang, N. Chen, K. Chen, J.A. Patz, Intracounty modeling of COVID-19 infection with human mobility: assessing spatial heterogeneity with business traffic, age, and race, Proc. Natl. Acad. Sci. U.S.A. 118 (24) (2021).
[33] N. Ayan, S. Chaskar, A. Seetharam, A. Ramesh, A.D.A. Antonio, Poster: COVID-19 case prediction using cellular network traffic, in: 2021 IFIP Networking Conference (IFIP Networking), IEEE, 2021, June, pp. 1–3.
[34] E.A. Rashed, A. Hirata, One-year lesson: machine learning prediction of COVID-19 positive cases with meteorological data and mobility estimate in Japan, Int. J. Environ. Res. Publ. Health 18 (11) (2021) 5736.
[35] E.A. Rashed, S. Kodera, A. Hirata, COVID-19 forecasting using new viral variants and vaccination effectiveness models, Comput. Biol. Med. 149 (2022), 105986.
[36] BiciMAD open data for the city of Madrid. Service use, Available on-line at: https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1. (Accessed 31 March 2023).
[37] BiciMAD open data for the city of Madrid. Location of the bike stations, Available on-line at: https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1. (Accessed 31 March 2023).
[38] COVID-19 incidence weekly data for each primary care center for the Comunidad de Madrid region, Available on-line at: https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_zonas_basicas_salud. (Accessed 31 March 2023).
[39] COVID-19 incidence weekly data for each district for the Comunidad de Madrid region, Available on-line at: https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos. (Accessed 31 March 2023).
[40] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.