

Predicting COVID-19 Spread from Large-Scale Mobility Data

Amray Schwabe
schwabea@ethz.ch
ETH Zürich
Zürich, Switzerland

Joel Persson
jpersson@ethz.ch
ETH Zürich
Zürich, Switzerland

Stefan Feuerriegel
sfeuerriegel@ethz.ch
ETH Zürich
Zürich, Switzerland

ABSTRACT

To manage the COVID-19 epidemic effectively, decision-makers in public health need accurate forecasts of case numbers. A potential near real-time predictor of future case numbers is human mobility; however, research on the predictive power of mobility is lacking. To fill this gap, we introduce a novel model for epidemic forecasting based on mobility data, called mobility marked Hawkes model. The proposed model consists of three components: (1) A Hawkes process captures the transmission dynamics of infectious diseases. (2) A mark modulates the rate of infections, thus accounting for how the reproduction number R varies across space and time. The mark is modeled using a regularized Poisson regression based on mobility covariates. (3) A correction procedure incorporates new cases seeded by people traveling between regions. Our model was evaluated on the COVID-19 epidemic in Switzerland. Specifically, we used mobility data from February through April 2020, amounting to approximately 1.5 billion trips. Trip counts were derived from large-scale telecommunication data, i. e., cell phone pings from the *Swisscom* network, the largest telecommunication provider in Switzerland. We compared our model against various state-of-the-art baselines in terms of out-of-sample root mean squared error. We found that our model outperformed the baselines by 15.52 %. The improvement was consistently achieved across different forecast horizons between 5 and 21 days. In addition, we assessed the predictive power of conventional point of interest data, confirming that telecommunication data is superior. To the best of our knowledge, our work is the first to predict the spread of COVID-19 from telecommunication data. Altogether, our work contributes to previous research by developing a scalable early warning system for decision-makers in public health tasked with controlling the spread of infectious diseases.

CCS CONCEPTS

- Computing methodologies → Modeling methodologies;
- Information systems → Data analytics; *Spatial-temporal systems*;
- Mathematics of computing → Time series analysis; *Stochastic processes*.

KEYWORDS

COVID-19, Epidemic Forecasting, Telecommunication Data, Mobility Data, Hawkes Process

1 INTRODUCTION

COVID-19 has evolved into a global pandemic, which, as of January 30, 2021, has been responsible for more than 100 million reported cases [44]. The spread of COVID-19 continues to pose a serious threat not only to public health but also to healthcare institutions and economies as a whole. To control the spread of COVID-19,

decision-makers in public health need accurate predictions of the future case numbers. This allows for early interventions and, on top of that, is important for near real-time resource planning, so that hospital capacities are not exceeded [14].

To make epidemic forecasts in near real-time, leading indicators are needed [16]. Such leading indicators should predict the spread of infectious diseases and, to be of practical value, should be scalable. In epidemiology, spatio-temporal data on mobility has been suggested as a leading indicator that could enable epidemic forecasting as infectious disease are spread through human-human transmissions [13]. In previous work, epidemic forecasting using mobility has been evaluated for dengue fever [23], Ebola [21], and cholera [2]. Yet, in the case of COVID-19, the out-of-sample predictive power of mobility data for epidemic forecasting is subject to research.

Human mobility can be monitored using a variety of methods. One approach is to quantify mobility using point of interest (POI) data obtained from check-ins [3, 8, 11] or smartphone location logs[5, 8, 10, 25]. An alternative is to extract mobility patterns from telecommunication data [35]. Telecommunication data is regarded to provide a more comprehensive coverage of a country’s population [16, 33]. In particular, telecommunication data have, unlike POI data, the advantage of capturing routine activities that might not otherwise be recorded. Telecommunication data originates from cell phone pings and thus track all SIM card movements, regardless of which hard- or software is used by the phone. Motivated by this, several research commentaries have discussed the potential benefits of telecommunication data for monitoring the spread of COVID-19 [16, 33]. However, research actually evaluating the predictive power of telecommunication data for forecasting the spread of COVID-19 is still lacking. This represents the objective and novelty of the present work.

Proposed model (M2H)¹: We propose a novel mobility marked Hawkes model (M2H) for epidemic forecasting based on mobility data. The M2H model consists of three components: (1) A Hawkes process captures the transmission dynamics of infectious diseases. The benefit of a Hawkes process is that each event (infection) may trigger further offspring events, thus reflecting the transmission dynamics of infectious diseases. In our model, each region is represented by a different instantiation of the Hawkes process. (2) A mark modulates the rate of infections, thus accounting for spatio-temporal variation in the reproduction number R . The mark is modeled using a regularized Poisson regression based on mobility covariates extracted from telecommunication data. (3) A correction procedure incorporates new cases seeded by people traveling between regions. The model is thus semi-mechanistic in the sense that

¹Both code and data will be made available via <https://github.com/amrayschwabe/PredictingCovidSpreadMobility>.

its components are based on transmission dynamics and human behavior.

Findings: Our proposed M2H model was evaluated with large-scale mobility data from Switzerland. We derived mobility based on telecommunication metadata from *Swisscom*, the largest telecommunication provider in Switzerland. Mobility data from February through April 2020 was used, consisting of approximately 1.5 billion trips. The M2H model was evaluated over forecast horizons ranging from 5 to 21 days. In particular, we compared the root mean squared error (RMSE) against state-of-the-art baselines [9, 35, 45]. Overall, we found that the RMSE of our model outperformed the baselines by 15.52 %. In addition, we re-estimated our model with POI data (i.e., Google Community Mobility Reports [27]) instead of telecommunication data. This decreased the prediction performance by 10.57 %, thus confirming that the use of telecommunication data as done in our study is superior.

Contributions: Our contributions are as follows:

- (1) We present a novel mobility marked Hawkes model for epidemic forecasting. Compared to previous research, our model captures the effects of between-region transmissions and incorporates the effects of mobility based on telecommunication data.
- (2) We outperform state-of-the-art prediction models when forecasting COVID-19 spread by 15.52 %. The superior performance is consistently demonstrated for forecast horizons between 5 and 21 days.
- (3) We establish the predictive power of large-scale telecommunication data for forecasting COVID-19 spread. We further show that mobility flows from telecommunication data provide higher predictive power than POI data (i.e., Google Community Mobility Reports).

As implications for practice, we help decision-makers in public health by providing a tool for near real-time epidemic forecasting and thus enable better management of the COVID-19 epidemic.

2 BACKGROUND

2.1 Modeling the Spread of COVID-19

Predicting COVID-19 spread is a challenging task. Previous research has used a range of explanatory models, which describe the spreading dynamics retrospectively (e.g., [1, 35]). In contrast, few models have been developed (and evaluated) for forecasting. Epidemic forecasting is, however, crucial in practice as such forecasts inform decision-makers in public health (e.g., when timing interventions or opening strategies, and when managing the utilization of healthcare resources [14]). In the following, we provide a brief review of predictive models for epidemic forecasting.

Generally, models for epidemic forecasting can be distinguished into stochastic and mechanistic models. Stochastic models mainly rely on the predictive power of data, while mechanistic models explicitly utilize epidemiological theory and empirical evidence. An example of a stochastic model often used in the early stages of disease spread is an exponential model [4]. Another example of a stochastic model is a Gaussian model, which is used to predict the peak of COVID-19 cases [39]. Both models predict future cases solely from past cases. Recent efforts have incorporated additional predictors. In the case of COVID-19, predictors have been derived

from behavioral data such as visitor counts for POIs or call data, and used been used in regression models [38] and machine learning models [20, 45]. However, machine learning models are of limited use in emerging epidemics due to their need for large historical training data. Yet, in epidemic forecasting, such large historical training data are typically not available.

One type of mechanistic model is the compartmental model. In a compartmental model, each individual belongs to some compartment (e.g., susceptible, infected, recovered) and has a certain probability of transitioning to another compartment. The probability of a transition can either be derived from case data or modeled as being dependent on additional predictors. Both approaches have been used in previous research on COVID-19 [4, 8, 10, 28, 34, 36, 49]. However, a downside of compartmental models is, that they cannot fully account for uncertainty. For instance, a fixed value has to be used for the incubation time.

In between stochastic and mechanistic models are point processes, such as the Hawkes process [17]. The self-exciting structure of point processes enables the modeling of how events (i.e., infections) trigger further offspring events. This property reflects human-human transmission in which infected individuals transmit disease to other people [4]. Under some conditions, Hawkes processes have been shown to provide equivalent estimates to compartmental models [37]. Moreover, Hawkes processes can be extended to include a distribution over the incubation time. Due to these benefits, Hawkes processes have previously been used for modeling COVID-19 spread [4, 9, 29]. Yet, to the best of our knowledge, we are the first to incorporate mobility flows from telecommunication data.

2.2 Mobility Data

Mobility data has been extensively used for explaining urban phenomena. Previous applications include predicting crime patterns [18] and detecting the impact of social investments [50]. On an urban scale, previous research has also explored the predictive power of mobility data in forecasting chronic diseases such as cancer or diabetes [43].

In the case of COVID-19, previous work has empirically confirmed the association between mobility and COVID-19 cases based on explanatory analysis [35, 47], while predictive modeling is scarce. To obtain mobility data, one approach is to back-trace human movements from POI data (static data). POI data provide visitor counts to specific locations. Such counts can be collected manually through check-ins or automatically (e.g., Google Community Mobility Reports [27]).

Alternatively, mobility data can be derived from movements across space and time (dynamic data) resulting in origin-destination pairs. In contrast to static data, dynamic data have the advantage of tracing the underlying trajectory and, therefore, are likely to provide a more comprehensive description of mobility. Dynamic data might thus be able to predict COVID-19 spread better than static data.

2.3 Hawkes Processes

A Hawkes process is a self-exciting point process characterized by an intensity function depending on all previous events [17]. The

intensity function describes processes where one event is likely to trigger a series of offspring events. Previous work models disease spread as a naïve Hawkes process [9, 21, 23, 40], yet none of these works leverages telecommunication data to model the spatio-temporal spreading dynamics.

A specific variant of a Hawkes process is the marked Hawkes process. A so-called “mark” measures the “impact” of an event. A marked Hawkes process has been used in applications such as forecasting earthquakes [32]. In this work, we develop a novel, tailored version of the marked Hawkes process for epidemic forecasting based on mobility data.

Research gap: Mobility has been suggested as a leading indicator of the epidemic, which can be inferred from telecommunication data. However, research assessing the predictive power of telecommunication data is lacking. To close this gap, we introduce a novel mobility marked Hawkes process for epidemic forecasting.

3 THE PROPOSED MODEL

Let $\mathcal{I} = \{i = 1, \dots, N\}$ be a finite set of mutually exclusive regions (e.g., states). Further, let $t = 1, \dots, T$ denote a fixed number of days. Our goal is to predict the daily count of new cases $y_{it} \in \mathbb{N}_{\geq 0}$ in each region over a forecast horizon. We will do so by modelling the observed trajectories of reported cases, $\{y_{i1}, y_{i2}, \dots, y_{iT}\}_{i=1}^N$, as realizations from underlying discrete-time stochastic processes $\{Y_i(t) : t = 1, \dots, T\}_{i=1}^N$.

3.1 Model Components

Our M2H model consists of three components. A discrete-time **Hawkes process** for the underlying transmission dynamics, a modeled **mark** for the reproduction rate, and a **correction for between-region mobility**, incorporating new cases seeded by infected people traveling between regions. Later, the mark accommodates within-region mobility, while the correction is responsible for incorporating between-region mobility. In the following, we describe each component.

(1) Hawkes process: Here we describe the standard Hawkes process [17], which we then adapt to our discrete-time setting. A Hawkes process is a self-exciting point process. It models disease spread as a time series of cases where every case can produce offspring cases. Additionally, so-called immigrant cases are triggered by external sources and contribute to disease spread. A standard Hawkes process is defined by the intensity function

$$\lambda(s) = \lambda_0 + \sum_{s_j < s} \phi(s - s_j). \quad (1)$$

The intensity function $\lambda(s)$ represents the expected number of cases at time s . It depends on the background rate λ_0 (rate of immigrant cases) and $\sum_{s_j < s} \phi(s - s_j)$, the rate of offspring cases. This depends on all past cases s_j where $\phi(s - s_j)$ regulates the change that a case at time s_j has on the intensity function at time s . In our case, the kernel function ϕ models the distribution of incubation times. Note that the Hawkes process is a particular case of a non-homogeneous Poisson process, in which the intensity is stochastic and explicitly depends on previous events through the kernel function.

The standard Hawkes process is not directly applicable to modeling infectious diseases, since case numbers are usually available

on a daily basis and, therefore, in discrete time. Hence, instead of a standard Hawkes process, we change the definition to one in discrete time:

$$\lambda(t) = \lambda_0 + \sum_{t_j < t} y_{t_j} \phi(t - t_j), \quad (2)$$

where y_{t_j} is the sum of cases reported on day t_j . Instead of summing over all past events $s_j < s$ occurring in continuous time, we now sum over all past days $t_j < t$ where, on day t_j , y_{t_j} events happened. Note that, even in this formulation, the events themselves need not be in discrete time, but are simply aggregated over discrete time intervals corresponding to days. Accordingly, $\lambda(t)$ now denotes the intensity function on day t .

(2) Mark: We introduce a mark R to account for the spatio-temporal variation in the number of offspring cases per case, commonly called the reproduction rate. The mark is a function of covariates and denotes the “impact” of an event. For instance, a mark of two corresponds to two offspring cases being caused by every past infection. The mark is incorporated into the Hawkes process via

$$\lambda(t) = \lambda_0 + \sum_{t_j < t} y_{t_j} R(\mathbf{x}_t) \phi(t - t_j), \quad (3)$$

where $R(\mathbf{x}_t)$ is the realization of the mark given some covariates \mathbf{x}_t , which we will introduce later. In comparison to Eq. 2, the influence of a case at time t_j on the intensity function $\lambda(t)$ may now not only depend on the time difference $t - t_j$, but also on the day t_j itself. The resulting model is a marked Hawkes process in discrete-time. We model the mark with a regularized Poisson regression, explained in the following.

We estimate the mark in region i on day t with regularized Poisson regression. We chose a Poisson regression for two reasons: (1) Poisson regression is robust against heteroskedastic error terms, whereas estimation with ordinary least squares is not and would give inconsistent parameter estimates [46]. (2) Empirical research in epidemiology has estimated the reproduction rate as Poisson distributed [26, 48].

Two types of covariates are included in the regression: a vector $\mathbf{m}_{it} = [m_{it1}, \dots, m_{itz}]$ of mobility covariates that vary by region and day, and $w_{it} \in \mathbb{R}$ as a summary measure of the weather in region i at day t . In our case, the mobility covariates are the count of within-region trips for different purposes (commuting vs. non-commuting) and using different modes of transport (highway, road, and train). Within-region mobility is added in the Poisson regression, while between-region mobility is part of the correction in component (3). For the weather covariate, we use the maximum temperature per region and day.

Let $\mathbf{x}_{it} = [\mathbf{m}_{it}, w_{it}]$ and $\boldsymbol{\theta} = [\boldsymbol{\beta}, \psi]$ be the covariate and parameter vectors, respectively. The regression model is

$$\mathbb{E}[R(X_{it}; \boldsymbol{\beta}_0, \boldsymbol{\theta}) | X_{it} = \mathbf{x}_{it}; \boldsymbol{\beta}_0, \boldsymbol{\theta}] = \exp(\boldsymbol{\beta}_0 + \boldsymbol{\theta}^\top \mathbf{x}_{it}). \quad (4)$$

The parameters are estimated by fitting the regression on the inferred reproduction rate. We use a lasso regularization [42] to avoid overfitting and facilitate model interpretation by shrinking non-predictive covariates towards zero. Note that the regularization term is not part of the regression model itself but added as a constraint in the estimation. Details are provided in Sec. 3.3.

(3) Correction for between-region mobility: We introduce a correction procedure that incorporates new cases seeded by people

traveling between regions. The procedure is motivated by the following observation: An offspring case in any region can be caused by one of two things: (1) a past case in the same region, or (2) a past case in any other region. The standard Hawkes process only accounts for the effect of (1) while implicitly attributing the effect of (2) to the background rate λ_0 . However, λ_0 is restricted to be constant over time with effects on the intensity function independent of the mark. This is problematic for modeling infectious diseases as the rate of immigrant cases likely depends on (a) the number of cases in other regions, (b) the inflow of individuals from other regions, and (c) the reproduction rate within the same region, all of which may change over time. To capture these dynamic dependencies, we use the following three-step correction procedure:

- (1) Estimate the number of infected individuals $n_i(t)$ traveling into region i on day t from any other region with

$$\hat{n}_i(t) := \sum_{i' \in \mathcal{I} \setminus \{i\}} m_t(i', i) \frac{y_{i'} t}{q_{i'}}, \quad (5)$$

where $m_t(i', i)$ is the count of trips from region i' to region i on day t , $y_{i'} t$ is the sum of infected individuals at time t in any region i' , and $q_{i'}$ is the population size of region i' .

- (2) Multiply $n_i(t)$ by a weighting factor α to allow for a differential effect of spreading dynamics between regions compared to within regions.
- (3) Replace the background rate λ_0 with $\alpha n_i(t)$ and add it to $y_{i'} t$ in the marked Hawkes process.

Altogether, the procedure captures the increase in the intensity function due to transmission of the disease caused by between-region trips of infected people. Note that, in our model, there is no need to include the background rate λ_0 (which is otherwise present in standard Hawkes processes). The reason is that immigrant cases are already captured by the added term $\alpha n_i(t)$.

3.2 The M2H Model

Combining the above components gives our M2H model. Specifically, its intensity function for region i and day t is given by

$$\lambda_i(t) = \sum_{t_j < t} (y_{it_j} + \alpha \hat{n}_i(t_j)) \cdot R(\mathbf{x}_{it}; \beta_0, \theta) \cdot \phi(t - t_j), \quad (6)$$

where $\hat{n}_i(t_j)$ and $R(\mathbf{x}_{it}; \beta_0, \theta)$ are given by Eq. 5 and Eq. 4, respectively. The model is a marked Hawkes process, in which for each region, disease transmissions occur due to people within the region and those traveling into the region. The future spread is modulated by the reproduction number (i.e., mark), which itself is modeled as a function of the amount and type of mobility and weather. We model the kernel function ϕ as the incubation time distribution by adding the constraint

$$\sum_{l \in \mathbb{N}^+} \phi(l) = 1. \quad (7)$$

This defines ϕ as a distribution function. Now, ϕ only determines the timing of offspring cases and thus separably identifies effects of the incubation period to those of the reproduction number. Note that a standard Hawkes process does not model these effects.

3.3 Model Estimation

We estimate our M2H model by developing an expectation-maximization (EM) algorithm [12] that is tailored to our task. This approach is computationally efficient and thus enables fast model estimation and prediction. The algorithm is given in Alg. 1.

Algorithm 1 EM algorithm for the M2H model

```

1: procedure EM( $\alpha, \xi, \phi, \{y_{it}, \mathbf{x}_{it}, \hat{n}_i(t) : \forall i, t\}$ )
2:   initialize  $R(\mathbf{x}_{it}; \beta_0, \theta) = 1 \forall i, t$ 
3:   while  $\|\Delta\theta\|_1 > \text{tol}$  do
4:     Expectation step:
5:     for  $i = 1, 2, \dots, N$  do
6:       for  $\forall t, t' = 1, \dots, T$  where  $t < t'$  do
7:          $p_i(t, t') = \frac{R(\mathbf{x}_{it}; \beta_0, \theta) \cdot \phi(t' - t)}{\lambda_i(t')}$ 
8:       end for
9:       for  $t = 1, \dots, T$  do
10:         $r_{it} = \sum_{t' > t} p_i(t, t') \cdot y_{it'}$ 
11:      end for
12:    end for
13:    Maximization step:
14:     $\hat{\theta}, \hat{\beta} = \arg \min_{\beta_0, \theta} -\frac{1}{NT} \log \mathcal{L}(\beta_0, \theta | \mathbf{X}, \mathbf{R}) + \xi \|\theta^\top\|_1$ 
15:  end while
16: end procedure

```

The EM algorithm proceeds as follows: It takes the weight α , the lasso penalty ξ , and the kernel function ϕ as given and initializes the estimated reproduction rates $R(\mathbf{x}_{it})$. In the expectation step, then estimates the probability $p_i(t, t')$ that one case on a day t' is caused by one in day t , for all days in the data. Then, the reproduction rate r_{it} is calculated from the obtained probabilities and the number of cases. This procedure is repeated for all regions.

In the maximization step, the parameters θ of the Poisson regression are estimated by minimizing the negative log-likelihood function subject to the lasso regularization. The regression model is specified as a generalized linear model with log-link. The log-likelihood function of θ given the design matrix \mathbf{X} (with elements \mathbf{x}_{it}) and the reproduction rate matrix \mathbf{R} (with elements r_{it}) is

$$\log \mathcal{L}(\beta_0, \theta | \mathbf{X}, \mathbf{R}) = \sum_{i=1}^N \sum_{t=1}^T \left(r_{it} (\beta_0 + \theta^\top \mathbf{x}_{it}) - \exp(\beta_0 + \theta^\top \mathbf{x}_{it}) \right).$$

The expectation and maximization steps were repeated until convergence (until β_0 and θ vary by less than 10^{-5}). All covariates in the regression were standardized to aid convergence.

Informed by research on incubation times [24], we model ϕ as a discretized Gamma(5.807, 1.055) distribution. Specifically, let F be the cumulative distribution function (CDF) of the Gamma distribution. Then $\phi(l) = F(l + 0.5) - F(l - 0.5)$ for $l \in \mathbb{N}^+$, with $\phi(1) := F(1.5)$ to cover the probability mass on the interval $[0, 0.5]$.

To estimate the weight α and lasso penalty ξ , we conduct a grid search to minimize the prediction error of the reported number of cases in a leave-one-out (here, region) cross-validation setting. The intervals for α and ξ were $[0, 3]$ and $[0, 20]$, respectively.

3.4 Forecasting

To predict the number of future COVID-19 cases in region i at δ days ahead, we proceed as follows:

- (1) Starting at day t , estimate the reproduction rate so far using the values of the covariates until day t and the estimated parameters of the regularized Poisson regression.
- (2) An intensity function estimate $\hat{\lambda}_i(t)$ is then obtained by evaluating the model at the parameter estimates $(\hat{\beta}_0, \hat{\theta})$.
- (3) Draw $Y_{it} \sim \text{Poisson}(\hat{\lambda}_i(t))$ as a prediction for the number of cases in region i at day t .
- (4) Repeat steps (1)–(3) for day $t + d$ with $d = 1, \dots, \delta$.

Steps (1)–(4) are repeated ten times for each region $i = 1, \dots, N$, resulting in ten predictions of future disease spread for all regions. The ten different predictions reflect the uncertainty in the future number of cases in a given region.

3.5 Model Variants

We tested several variants of our model as part of an ablation study (see Sec. 5.4). This allowed us to assess which parts of the model improve prediction performance.

- (1) **M2H model without regularization:** We evaluated a variant of our M2H model without regularization to test whether the penalty is necessary.
- (2) **M2H model without weather covariate:** Weather could be an important determinant of COVID-19 spread, and, to test this, we evaluated our model against one without the weather covariate.
- (3) **M2H model without correction term:** To evaluate if our proposed correction procedure enhances prediction performance, we also considered a model in which it is replaced with the background rate.
- (4) **M2H model with between-region mobility covariates in the mark:** To evaluate if our correction procedure sufficiently accounts for between-region mobility, we added the corresponding covariates to the regression model for R .
- (5) **M2H model with demographic covariates:** Following prior literature [9], we tested if prediction performance is increased by adding demographic covariates (here: population density, population size, and percentage of city population).
- (6) **M2H model with POI data:** To evaluate if telecommunication data offers higher predictive power, we re-estimated our model with POI data derived from the Google Community Reports [27]. We did not apply our correction for between-region travel, as POI data does not provide any information about between-region mobility.

4 SETTING

4.1 Data

The study period was Feb 24, 2020–April 26, 2020. This period covers the first wave of COVID-19 in Switzerland, starting at the date of the first case and ending when the number of new cases again approached zero.

Case data: In Switzerland, COVID-19 cases are reported at the state level (called “cantons”). We model each of the 26 cantons as a separate region. In our results, we refer to regions as “cantons”. The daily number of COVID-19 cases per canton was obtained from the Federal Office of Public Health of the Swiss Confederation (BAG) [30]. No additional preprocessing is applied to the case data.

Mobility data: Telecommunication data were used to infer mobility of SIM-based devices. In our work, telecommunication data was obtained from a collaboration with *Swisscom*, the largest telecommunication provider in Switzerland. The data collection and processing procedures are compliant with data privacy laws in Switzerland. Ethics approval from ETH Zurich was obtained (2020-N-41).

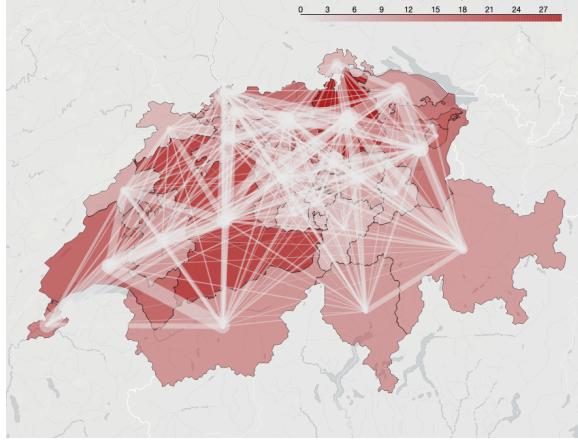
Data collection was conducted in collaboration with *Swisscom* as follows. First, routine signal exchanges (“pings”) between mobile devices and antennas were recorded. The data comprises all pings, regardless of the mobile service provider. Second, a triangulation algorithm was used to approximate the location (longitude, latitude) of each SIM card [19]. The accuracy of the triangulation has been tested and amounts to a median error of 132 meters [19]. Location data was recorded at a resolution of approx. 5 minutes. This ensures that our data capture micro-level movements of individuals of high granularity. Third, the movements were used to compute trips. A trip is registered when a SIM card moves between two different postcode areas and the destination location has been static for at least 20 minutes. Fourth, trips were classified by their mode of transport (“train”, “road”, and “highway”) and their purpose (“commuter” vs. “non-commuter”). The mode of trips is determined via the positions of antennas along train rails, highways, and roads. If a trip uses multiple modes of transport, it is classified as the mode that covered the longest distance. “Commuter” and “non-commuter” trips are differentiated based on extracted home and work locations of individuals. Details are reported in the *Swisscom Mobility Insights* [41]. Fifth, trips were aggregated for each day and region. This results in an origin-destination matrix of mobility flows. We then derive within- and between-region mobility covariates. The within-region mobility covariate is calculated as the reduction in within-region trips at a given day relative to a reference period (here: the average daily trip count during the two weeks before the first Swiss COVID-19 case). The between-region covariates are for each region calculated as the total number of trips into the region per every other region.

Additional predictors: Weather data were retrieved per day and canton from the COVID-19 open data repository². The daily maximum temperature was used as a predictor. For one model variant, we included covariates for the sum of incoming trips per canton. These were obtained for each canton by calculating the change in the sum of incoming trips relative to the reference period. For another model variant, predictors included socio-demographic statistics (canton population, population density, percentage of city population, age group distribution). These were obtained from the Swiss Federal Statistical Office [31].

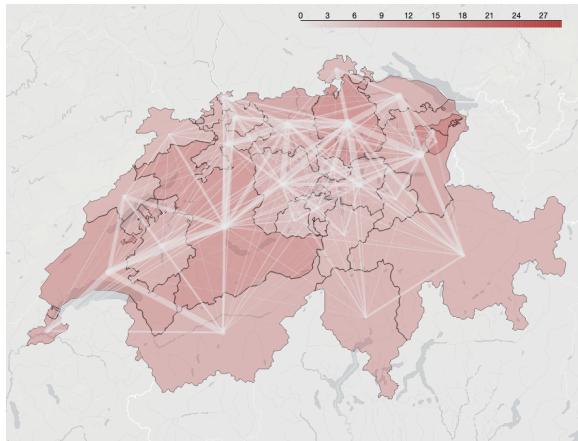
4.2 Summary Statistics

Case data: During our study period, 29,148 COVID-19 cases were registered in Switzerland. COVID-19 spread started on February 24, 2020, with a maximum of 1455 people tested positive on March 23, 2020. COVID-19 spread has varied considerably between regions (i.e., cantons), with only 20 cases registered for the canton of Appenzell Innerrhoden (AI) until the end of our study period (April 26, 2020), but 5308 cases for the canton of Vaud (VD).

²<https://github.com/googleCloudPlatform/covid-19-open-data>, retrieved: January 10, 2020



(a) Mobility flows between cantons on February 21, 2020.



(b) Mobility flows between cantons on March 29, 2020.

Figure 1: Comparison of mobility flows (i.e., origin-destination matrix) on two example dates before COVID-19 (a) and after (b) restrictions have been imposed. Line width denotes volume of between-region trips; color denotes volume of within-region trips (in millions).

Mobility data: Our study period covers approximately 1.5 billion trips. Following the first lockdown in Spring 2020, overall mobility in Switzerland dropped by approximately 50 %. The mobility flows before and after the lockdown are shown in Fig. 1.

4.3 Evaluation Framework

Predictions were evaluated across different forecast horizons (5–21 days ahead). A leave-one-region-out validation was used to test the out-of-sample predictive performance of our model. This ensures that the model does not overfit a specific region. Two performance metrics are reported: the root mean squared error (RMSE) and the mean absolute error (MAE). The former is a statistical measure of model fit whereas the latter quantifies the deviation in the number

of cases and is thus of relevance for decision-making in public health.

4.4 Baselines

Our model is compared against the following baselines:

- (1) **Negative binomial regression** [35]: The negative binomial regression model predicts future case numbers by using lagged mobility indices, time trends, and region-specific factors, and accounts for overdispersion in case numbers.
- (2) **ARMAX model**: The ARMAX(p, q) model is a common baseline for time-series forecasting. It models the daily case count as being dependent on that of the previous p days, the previous q days' estimation errors, and additional covariates. This model should thereby capture time dependence in the cases and prediction errors. The model orders $p = 6, q = 1$ were estimated via a grid search ($p, q \in 1, 2, \dots, 10$).
- (3) **Random forest**: A random forest is a non-parametric machine learning model that forecasts case counts from covariates using an ensemble of decision trees. This baseline is chosen due to its flexibility in data-scarce settings and handling non-linearity.
- (4) **Neural network** [45]: An artificial neural network is used to predict case counts based on past cases and additional covariates. The weather covariate did not increase performance and was thus discarded. The architecture and hyperparameter tuning (via grid search) was done analogous to [45].
- (5) **SEIR mobility model** [22]: A SEIR model predicts future case numbers from past infections in a deterministic manner. Here, we allow the reproduction rate to be modulated by mobility by multiplying the weighted change in mobility with the initial reproduction rate.
- (6) **Naïve marked Hawkes process** [9]: In the naïve marked Hawkes process, the mark is determined from demographic covariates and within-region mobility. Different from our M2H model, regularization for R is not included and the static background rate is used instead of our correction procedure. However, to capture between-region heterogeneity, socio-demographic covariates are additionally included.

We also experimented with a **graph neural network** from [20], yet, for our problem setup, found it inapplicable. Hence, this model was discarded from our comparison. Unless otherwise stated, all models were fed with the same data as our proposed M2H model.

5 RESULTS

5.1 Overall Prediction Performance

Table 1 reports the macro-averaged prediction performance of the models across the different regions and forecast horizons. Our proposed M2H model achieves an RMSE of 13.20 and an MAE of 11.07. For comparison, the best baseline (the neural network from [45]) has an RMSE of 15.41 and an MAE of 12.87. Hence, our model improves both RMSE and MAE by 15.52 %, and 16.26 %, respectively. Note that all modes are evaluated with the same data, and hence, the performance improvement is due to the model itself.

Table 1: Comparison of overall prediction performance (i.e., macro-averaged across regions and forecast horizons).

Model	RMSE	MAE
Negative binomial regression [35]	19.39	16.77
ARMAX model	28.84	25.14
Random forest	32.78	28.33
Neural network [45]	15.41	12.87
SEIR mobility [22]	36.48	31.37
Naïve marked Hawkes process [9]	22.21	18.64
M2H model (proposed)	13.34	11.07

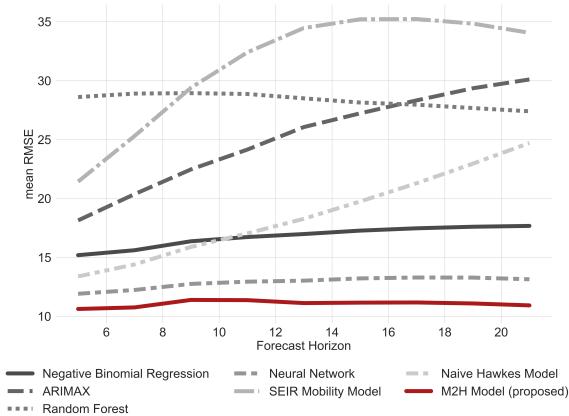
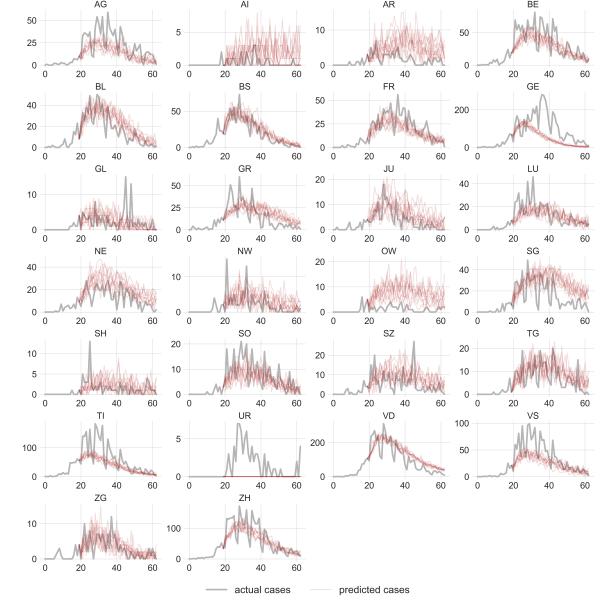
Lower values = better (best values in bold)

5.2 Prediction Performance across Forecast Horizons

Fig. 2 compares the prediction performance across different forecast horizons. The proposed M2H model is superior across all baselines and forecast horizons. Moreover, the prediction performance of our model is largely invariant to the forecast horizon. Interestingly, the improvement over the naïve Hawkes model increases with longer forecast horizons. The results for the MAE are analogous but omitted due to space constraints.

5.3 Actual vs. Predicted Case Counts by Region

We plot predictions of case counts from our model against the actual counts (Fig. 3). Here, the variation in the predictions reflects the uncertainty in future cases. Some cantons experienced only a low number of daily cases, oftentimes below 10 (e.g., Appenzell Innerrhoden (AI), Uri (UR), and Obwalden (OW)). The figure shows that our model yields predictions of similar magnitude for these cantons. Other cantons had over 200 daily cases (e.g., Zurich (ZH)). Again, we find that the predictions fit the actual case numbers well.

**Figure 2: Comparison of prediction error across different forecast horizons macro-averaged by regions.****Figure 3: Predictions (gray) vs. actual case counts (red) per region (i.e., per canton). Results are shown for 21-day ahead forecasts and using 10 predictions to indicate uncertainty. The x-axis denotes the number of days since the first case in Switzerland, and the x-axis denotes the case counts. Names of regions (cantons) are stated using the official two-letter abbreviations.**

For the canton of Geneva (GE), predictions appear too conservative. A closer inspection revealed that case counts in Geneva peaked considerably later than in other cantons, and also much later compared to the reduction in mobility. Nonetheless, the predictions of the best baseline (i.e., the neural network from [45]) were also too conservative. This suggests that the relationship between mobility and cases were different for Geneva as compared to other cantons.

5.4 Ablation Study

An ablation study presented in Table 2 compares the prediction performance of different variants of our M2H model (see Sec. 3.5). All model variants produce predictions inferior to those of the full model, confirming that all parts of the model contribute to its prediction performance. For instance, removing the Lasso regularization reduces the RMSE by 2.62 %; removing the correction for between-region mobility reduces the RMSE by 2.62 %; and removing the weather covariate reduces the RMSE by 3.15 %. Furthermore, adding either demographic information or the incoming trips as covariates did not improve prediction performance. The results are qualitatively the same when examining the change in MAE.

We further compare the predictive power of static (POI data) vs. dynamic data (mobility flows from telecommunication data). For this purpose, we re-estimate our model with POI based mobility data (here: Google Community Mobility Reports [27]). Using POI data increases the RMSE from 13.34 to 14.75, implying a performance drop of 10.57 %. A Wilcoxon signed-rank test (cf. [15] for its use in forecasting) rejects the null hypothesis that the two models'

Table 2: Ablation study.

Model variant	RMSE	MAE
M2H model w/o regularization	13.69	11.28
M2H model w/o correction for travel	13.69	11.38
M2H model w/o weather	13.76	11.47
M2H model with between-region covariates	13.38	11.11
M2H model with demographic covariates	13.56	11.23
M2H model based on POI data	14.75	12.32
M2H model	13.34	11.07

Lower values = better (best values in bold)

prediction errors are from the same distribution (p -value < 0.01). Therefore, the comparison shows that telecommunication data is of higher predictive power.

6 DISCUSSION

Findings: Our model enables accurate epidemic forecasting. We demonstrate its performance over forecast horizons ranging from 5 to 21 days. Using the same mobility data, the M2H model outperforms standard baseline models by 15.52 %.

Why is the M2H model superior? The improved predictive performance of our model can be attributed to its ability to capture infectious disease properties. (1) In infectious diseases, each case triggers new (offspring) cases. These transmission dynamics are captured by using a Hawkes process. (2) The rate of infections (i.e., reproduction rate) varies across space and time and is associated with population mobility. The mark models the reproduction rate using spatio-temporal covariates representing within-region mobility and weather. (3) Between-region mobility leads to spatial spread by seeding new transmission chains in distant areas. This is captured by our correction term.

Model strengths: Our M2H model has several benefits for practical applications: First, the model has a parsimonious structure, simplifying the interpretation and avoiding overfitting. Because of this, our model performs well in data-scarce contexts such as epidemic forecasting. Second, our model captures the prediction uncertainty by probabilistic forecasting. Third, it requires minimal hyperparameter tuning. The model is estimated via leave-one-out cross-validation. Here, the weight α and the lasso penalty ξ were obtained via grid search, and the other parameters are estimated with a tailored EM algorithm.

Static vs. dynamic data: Our study demonstrates the effectiveness of mobility flows derived from telecommunication data for epidemic forecasting. The data in our study were obtained from Swisscom, the largest telecommunication provider in Switzerland. During the COVID-19 epidemic, the use of telecommunication data for epidemic forecasting has been considered by governments in Switzerland and around the world [7]. However, empirical evidence concerning its effectiveness for forecasting is lacking. We find that using POI data results in a 10.57 % higher RMSE, thus reflecting a substantially inferior predictive power. A reason for this is that telecommunication data have larger coverage; it includes every mobile device independent of the operating system and does not

require manual opt-in [6]. Another advantage of telecommunication data is, they are measured with high frequency (e. g., daily), thereby enabling regularly updated forecasts as needed by decision-makers. Future research and practice is therefore encouraged to consider the use of telecommunication data for quantifying mobility.

Limitations and future work: Our work is subject to limitations that should be addressed in future research. First, our model was validated on data from the first wave of COVID-19 in Switzerland. Future work could deploy and validate our model in other settings. Note that our model should be applicable in a straightforward manner given the general model structure. Second, telecommunication data cannot be assumed to provide a perfect measure of mobility, as it only captures the movements of individuals with SIM cards. Nevertheless, telecommunication data is argued to be more comprehensive than other sources [6, 16]. Third, we use various predictors informed by prior literature. The predictive power of other variables could be evaluated in future research.

Implications: Decision-makers in public health need tools for forecasting the spread of epidemics. This helps them in managing interventions (e.g., stay-home orders) and planning scarce healthcare resources (e.g., to avoid over-utilization of ventilators). A challenge is that data on cases are delayed due to incubation times and reporting delays. Hence, scalable tools for near real-time forecasting are needed. For this purpose, research commentaries [4, 16, 33] have called for research using novel stochastic process models and data on mobility flows. Our model should further be of value in managing other diseases (e. g. swine flu, ebola, etc.) and also emerging infectious diseases.

7 CONCLUSION

This work fills a known research gap [16, 33]: leveraging telecommunication data for epidemic forecasting. To achieve this, we propose a novel mobility marked Hawkes process for predicting COVID-19 cases from large-scale telecommunication data. We show that our proposed model outperforms the state-of-the-art in terms of RMSE by 15.52 %. The superior performance is achieved consistently across different forecast horizons. Therefore, our prediction model provides a valuable tool for near real-time monitoring of the COVID-19 spread and thus aids decision-makers in public health.

ACKNOWLEDGMENTS

We thank Swisscom for the collaboration and the unique data access. One author (Stefan Feuerriegel) is a member of a COVID-19 Working Group of the World Health Organisation (WHO) and thanks the members for intensive discussions.

REFERENCES

- [1] Nicolas Banholzer, Eva van Weenen, Bernhard Kratzwald, Arne Seeliger, Daniel Tschernutter, et al. 2020. Impact of non-pharmaceutical interventions on documented cases of COVID-19. *MedRxiv* (2020).
- [2] Linus Bengtsson, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, et al. 2015. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. *Scientific Reports* 5, 1 (2015), 1–5.
- [3] Seth G. Benzell, Avinash Collis, and Christos Nicolaides. 2020. Rationing social contact during the COVID-19 pandemic: Transmission risk and social benefits of US locations. *PNAS* 117, 26 (2020), 14642–14644.
- [4] Andrea L. Bertozzi, Elisa Franco, George Mohler, Martin B. Short, and Daniel Sledge. 2020. The challenges of modeling and forecasting the spread of COVID-19.

- PNAS* 117, 29 (2020), 16732–16738.
- [5] Giovanni Bonaccorsi, Francesco Pierri, Matteo Cinelli, Andrea Flori, Alessandro Galeazzi, et al. 2020. Economic and social consequences of human mobility restrictions under COVID-19. *PNAS* 117, 27 (2020), 15530–15535.
- [6] Caroline O. Buckee, Satchit Balsari, Jennifer Chan, Mercè Crosas, Francesca Dominici, et al. 2020. Aggregated mobility data could help fight COVID-19. *Science* 368, e487 (April 2020), 145–146.
- [7] Elvira Pollina Busvine, Douglas. 2020. European mobile operators share data for coronavirus fight. *Reuters* (March 2020). <https://www.reuters.com/article/us-health-coronavirus-europe-telecoms-idUSKBN2152C2>
- [8] Serina Chang, Emma Pierson, Pang Wei Koh, Joline Gerardin, Beth Redbird, et al. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589 (2021), 82–87.
- [9] Wen-Hao Chiang, Xueying Liu, and George Mohler. 2020. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *medRxiv* (2020).
- [10] Matteo Chinazzi, Jessica T. Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, et al. 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 368, 6489 (2020), 395–400.
- [11] Dhaval Dave, Andrew I. Friedson, Kyutaro Matsuzawa, and Joseph J. Sabia. 2021. When Do Shelter-In-Place Orders Fight COVID-19 Best? Policy Heterogeneity Across States and Adoption Time. *Economic Inquiry* 59, 1 (Jan. 2021), 29–52.
- [12] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J R Stat Soc Series B Stat Methodol* 39, 1 (1977), 1–22.
- [13] Angel N. Desai and Payal Patel. 2020. Stopping the Spread of COVID-19. *JAMA* 323, 15 (2020), 1516–1516.
- [14] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, others, and Green. 2020. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Technical Report.
- [15] Benito E. Flores. 1989. The utilization of the Wilcoxon test to compare forecasting methods: A note. *International Journal of Forecasting* 5, 4 (Jan. 1989), 529–535.
- [16] Kyra H. Grantz, Hannah R. Meredith, Derek A. T. Cummings, C. Jessica E. Metcalf, Bryan T. Grenfell, et al. 2020. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications* 11, 1 (2020), 4961.
- [17] Alan G. Hawkes and David Oakes. 1974. A cluster process representation of a self-exciting process. *Journal of Applied Probability* 11, 3 (Sept. 1974), 493–503.
- [18] Cristina Kadar, Stefan Feuerriegel, Anastasios Noulas, and Cecilia Mascolo. 2020. Leveraging Mobility Flows from Location Technology Platforms to Test Crime Pattern Theory in Large Cities. In *ICWSM*. 339–350.
- [19] Mohamed Kafsi. 2019. Quantifying the Accuracy of Mobility Insights from Cellular Network Data. <https://mkafsi.medium.com/quantifying-the-accuracy-of-mobility-insights-from-cellular-network-data-e5b83437a609>
- [20] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, et al. 2020. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. *arXiv preprint arXiv:2007.03113* (2020).
- [21] J. Daniel Kelly, Junhyung Park, Ryan J. Harrigan, Nicole A. Hoff, Sarita D. Lee, others, and Schoenberg. 2019. Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of the Congo using Hawkes point process models. *Epidemics* 28 (2019), 100354.
- [22] 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 (Aug. 1927), 700–721.
- [23] Minkyung Kim, Dean Paini, and Raja Jurdak. 2019. Modeling stochastic processes in disease spread across a heterogeneous social system. *PNAS* 116, 2 (2019), 401–406.
- [24] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, et al. 2020. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine* 172, 9 (2020), 577–582.
- [25] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, et al. 2020. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* 368, 6490 (2020), 489–493.
- [26] Ying Liu, Albert A. Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine* 27, 2 (2020).
- [27] Google LLC. 2020. Google COVID-19 Community Mobility Reports. <https://www.google.com/covid19/mobility/>
- [28] Andrew C Miller, Nicholas J Foti, Joseph A Lewnard, Nicholas P Jewell, Carlos Guestrin, and Emily B Fox. 2020. Mobility trends provide a leading indicator of changes in SARS-CoV-2 transmission. *medRxiv* (2020).
- [29] George Mohler, Frederic Schoenberg, Martin B Short, and Daniel Sledge. 2020. Analyzing the World-Wide Impact of Public Health Interventions on the Transmission Dynamics of COVID-19. *arXiv preprint arXiv:2004.01714* (2020).
- [30] Federal Office of Public Health. 2021. Coronavirus: Situation in Switzerland. <https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien/pandemien/aktuelle-ausbrueche-epidemien/novel-cov/situation-schweiz-und-international.html>
- [31] Federal Statistical Office. 2021. Population. <https://www.bfs.admin.ch/bfs/en/home/statistiken/bevoelkerung.html>
- [32] Yoshihiko Ogata. 1988. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *J. Amer. Statist. Assoc.* 83, 401 (March 1988), 9–27.
- [33] Nuria Oliver, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deleaille, et al. 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science advances* 6, 23 (2020).
- [34] Sen Pei, Sasikiran Kandula, and Jeffrey Shaman. 2020. Differential Effects of Intervention Timing on COVID-19 Spread in the United States. *Science advances* 6, 49 (2020).
- [35] Joel Persson, Jurriaan F. Parie, and Stefan Feuerriegel. 2021. Monitoring the COVID-19 epidemic with nationwide telecommunication data. *arXiv preprint arXiv:2101.02521* (2021).
- [36] Zhaozhi Qian and Ahmed M Alaa. 2020. When and How to Lift the Lockdown? Global COVID-19 Scenario Analysis and Policy Assessment using Compartmental Gaussian Processes. *NeurIPS* 33 (2020).
- [37] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. 2018. SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Populations. *WWW* (2018), 419–428.
- [38] Bahman Rostami-Tabar and Juan F. Rendon-Sanchez. 2021. Forecasting COVID-19 daily cases using phone call data. *Applied soft computing* 100 (2021), 106932.
- [39] Reinhard Schlickeiser and Frank Schlickeiser. 2020. A Gaussian Model for the Time Development of the Sars-Cov-2 Corona Pandemic Disease. Predictions for Germany Made on 30 March 2020. *Physics* 2, 2 (2020), 164–170.
- [40] Frederic Schoenberg, Marc Hoffmann, and Ryan Harrigan. 2019. A recursive point process model for infectious diseases. *Ann. Inst. Stat.* 71, 5 (2019), 1271–1287.
- [41] Swisscom. 2021. Swisscom Mobility Insights. <https://www.swisscom.ch/en/business/enterprise/offer/enterprise-mobile/mobility-insights.html>
- [42] Robert Tibshirani. 1996. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Series B Stat Methodol* 58, 1 (1996), 267–288.
- [43] Yingzi Wang, Xiao Zhou, Cecilia Mascolo, Anastasios Noulas, Xing Xie, et al. 2018. Predicting the Spatio-Temporal Evolution of Chronic Diseases in Population with Human Mobility Data.. In *IJCAI*.
- [44] WHO. 2021. Coronavirus Disease (COVID-19) Situation Reports. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>
- [45] Michał Wieczorek, Jakub Sílka, and Marcin Woźniak. 2020. Neural network powered COVID-19 spread forecasting model. *Chaos, Solitons & Fractals* 140 (2020), 110203.
- [46] Rainer Winkelmann. 2008. *Econometric Analysis of Count Data*. Springer Science & Business Media.
- [47] Chenfeng Xiong, Songhua Hu, Mofeng Yang, Weiyi Luo, and Lei Zhang. 2020. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *PNAS* 117, 44 (2020), 27087–27089.
- [48] Chong You, Yu-hao Deng, Wenjie Hu, Jiarui Sun, Qiushi Lin, et al. 2020. Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *International Journal of Hygiene and Environmental Health* 228 (2020), 113555.
- [49] Zebin Zhao, Xin Li, Feng Liu, Gaofeng Zhu, Chunfeng Ma, et al. 2020. Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, and Kenya. *Science of The Total Environment* 729 (2020), 138959.
- [50] Xiao Zhou, Desislava Hristova, Anastasios Noulas, Cecilia Mascolo, and Max Sklar. 2017. Cultural investment and urban socio-economic development: a geosocial network approach. *Royal Society Open Science* 4, 9 (2017), 170413.