Autoregressive count data modeling on mobility patterns to predict cases of COVID-19 infection

Jing Zhao

Xi'an Eurasia University

Mengjie Han (✉ mea@du.se)

Dalarna University: Hogskolan Dalarna <https://orcid.org/0000-0003-4212-8582>

Zhenwu Wang

China University of Mining and Technology Beijing Campus

Benting Wan

Jiangxi University of Finance and Economics

Research Article

Keywords: COVID-19, mobility, generalized linear model, autoregressive model, quasi-likelihood

Posted Date: March 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1398399/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Stochastic Environmental Research and Risk Assessment on June 23rd, 2022. See the published version at <https://doi.org/10.1007/s00477-022-02255-6>.

Jing Zhao^{1†}, Mengjie Han^{*2†}, Zhenwu Wang³ and Benting Wan⁴

* Corresponding author, mea@du.se, orcid.org/0000-0003-4212-8582

† The authors contributed evenly to this work.

1 School of Business Administration, Xi'an Eurasia University, Yanta District, Xi'an, China

2 School of Information and Engineering, Dalarna University, Falun 79188, Sweden

3 Department of Computer Science and Technology, China University of Mining and Technology, Beijing 100083, China

4 School of Software and IoT Engineering, Jiangxi University of Finance & Economics, Nanchang 330013, China

Abstract

At the beginning of 2022 the global daily count of new cases of COVID-19 exceeded 3.2 million, a tripling of the historical peak value reported between the initial outbreak of the pandemic and the end of 2021. Aerosol transmission through interpersonal contact is the main cause of the disease's spread, although control measures have been put in place to reduce contact opportunities. Mobility pattern is a basic mechanism for understanding how people gather at a location and how long they stay there. Due to the inherent dependencies in disease transmission, models for associating mobility data with confirmed cases need to be individually designed for different regions and time periods. In this paper, we propose an autoregressive count data model under the framework of a generalized linear model to illustrate a process of model specification and selection. By evaluating a 14-day-ahead prediction from Sweden, the results showed that for a dense population region, mobility data with a lag of 8 days is most reliable way of predicting the number of confirmed cases in relative numbers at a high coverage rate. It is sufficient for both of the autoregressive terms, studied variable and conditional expectation, to take one day back. For sparsely populated regions, a lag of 10 days produced the lowest error in absolute value for the predictions, where weekly periodicity on the studied variable is recommended for use. Interventions were further included to identify the most relevant mobility categories. Statistical features were also presented to verify the model assumptions.

Keywords: COVID-19, mobility, generalized linear model, autoregressive model, quasi-likelihood

Autoregressive count data modeling on mobility patterns to predict cases of COVID-19 infection

Abstract

At the beginning of 2022 the global daily count of new cases of COVID-19 exceeded 3.2 million, a tripling of the historical peak value reported between the initial outbreak of the pandemic and the end of 2021. Aerosol transmission through interpersonal contact is the main cause of the disease's spread, although control measures have been put in place to reduce contact opportunities. Mobility pattern is a basic mechanism for understanding how people gather at a location and how long they stay there. Due to the inherent dependencies in disease transmission, models for associating mobility data with confirmed cases need to be individually designed for different regions and time periods. In this paper, we propose an autoregressive count data model under the framework of a generalized linear model to illustrate a process of model specification and selection. By evaluating a 14-day-ahead prediction from Sweden, the results showed that for a dense population region, mobility data with a lag of 8 days is most reliable way of predicting the number of confirmed cases in relative numbers at a high coverage rate. It is sufficient for both of the autoregressive terms, studied variable and conditional expectation, to take one day back. For sparsely populated regions, a lag of 10 days produced the lowest error in absolute value for the predictions, where weekly periodicity on the studied variable is recommended for use. Interventions were further included to identify the most relevant mobility categories. Statistical features were also presented to verify the model assumptions.

Keywords: COVID-19, mobility, generalized linear model, autoregressive model, quasi-likelihood

1. Introduction

It has been two years since a group of patients with pneumonia were found in Wuhan, China [1]. While displaying symptoms similar to the severe acute respiratory syndrome (SARS) [2], upon investigation it was discovered that they had contracted a novel form of a coronavirus, named SARS-CoV-2. The disease was later named COVID-19. [3]. The basic reproduction number R_0 is a measurement of the average number of secondary cases that one primary case will generate. For COVID-19 this number was estimated to be 2.2 in western Europe and 3.38 in China [4,5]. The R_0 of the Delta and Omicron variants have been higher compared to the ancestral SARS-CoV-2 virus [6,7]. There is currently no medication approved by US Food and Drug Administration that offers a cure for this disease[8]. The death rate for COVID-19 has been reported as 3.4%, but this figure can vary depending on factors such as age, sex, the overall health of a population and the extent or otherwise of a national health system. For example, for those with cardiovascular disease the death rate from COVID-19 is 10.5% while for diabetes it is 7.3% [9]. Without effective control measures, regions with relatively older populations could see disproportionately more cases [10]. Most countries struggle to control the social and economic activities which lie behind many large-scale gatherings, thus making it difficult to isolate the sources of infection and cut off the channels of transmission. The primary control measures continue to be good hand hygiene, social distancing and quarantine.

Any additional preventative measures should consider the local environment and be country dependent.

Large-scale governmental measures, such as national lockdowns and the widespread cancellation of mass gatherings, may be taken quickly and produce instant results. However, the implementation of these strategies is usually *ex post facto* leaving limited time for organizations and the general public to adjust. In contrast, model-based predictions could enable governments to announce soft control measures and take action well in advance. Based on accurate predictions, well-planned control measures could lessen the number of infected individuals as well as minimizing the costs of preventative measures [11]. In addition, control measures could provide the medical system with sufficient time to allocate resources. Models that can handle government interventions are preferred in actual decision-making. A dynamic loop between modeling and measures deployment could be created to establish an equilibrium point, where additional efforts made on improving a model's accuracy will not radically affect a measure's enactment. Thus, predictive models that help to understand the future trends of confirmed COVID-19 cases should serve as the basis for decision support at a national or regional level. The SARS-CoV-2 spreads primarily through respiratory droplets and close contact. Given that the most probable transmission pathway is between humans, even asymptomatic ones [12], of the best ways to chart its spread is by examining the variable of human mobility and the different patterns that can take. Mobility data has two key features: how people move and how long people stay at a place. Due to the characteristics of the data and the reduced opportunities for public transport (see Section 4.1), we only considered the second feature, mobility change at particular locations, to build the models in this paper.

Adopting mobility could generate some uncertainties in the modeling, for example, the choice of lag effect, the impact of population density and the identification of autoregressive terms. In order to handle these uncertainties, the aim of this paper is to investigate how mobility change can be effectively used to predict the number of cases of COVID-19 infection. The empirical findings are illustrated by using a set of regions, taken from a Nordic country, in this case Sweden, with heterogeneous population densities. The modeling framework is thought to be replicable in a similar context. The specific contributions that this paper aims to make are to:

- 1) Propose an autoregressive modeling framework for count data in regions with different population densities;
- 2) Identify the categories of locations where mobility change is decisive for accurate predictions;
- 3) Provide a scheme to deal with the lag effects and autoregressive terms;
- 4) Analyze the effect on modeling by including an intervention;
- 5) Present statistical characteristics for two distributions of the count data: Poisson and negative binomial.

The rest of the paper is organized as follows. Section 2 briefly reviews the research that has studied mobility as a means of analyzing COVID-19. Section 3 presents the generalized linear model for the autoregressive count data and the quasi-likelihood for parameter estimation. Section 4 outlines the nature of both the mobility data and the daily confirmed cases of COVID-

19 in Sweden. Modeling implications and model applicability based on optimization results and predictions are given in Section 5. Section 6 concludes the work and points out some limitations.

2. Existing research on mobility patterns and COVID-19

The modeling methods and data resources used for understanding the spread of COVID-19 are important components for building a framework for prediction. The same method may produce different results on different datasets. To provide a basis for the modeling approach taken in this paper, this section will provide a brief overview of the research on these components, as well as a short discussion of the lag effect.

2.1. Empirical models for forecasting

Certain control measures, such as the suspension of transport infrastructures and the placing of widespread bans on public gatherings, can ensure a significant degree of social distancing that can have a knock-on effect on disease transmission [13,14]. However, the effect of these measures can vary between region and time period, making them stochastic factors in the accurate and long-term prediction of COVID-19 cases. Mobility pattern, on the other hand, is a more sensible proxy for the interpersonal contact rates that contribute to COVID-19 transmission because it not only reflects the consequences of measure execution, but also quantifies how people gather and move. Reduced mobility can simultaneously reduce the peak number of cases and delay the peak, which could help to alleviate the demand burden in the medical system. For example, a 20% reduction in mobility can cut down by 33% the peak number of cases and delay the arrival of that peak by almost two weeks. The reduction of the peak and the delay in its arrival could reach 91% and 14 weeks with a 60% reduction in mobility [15].

There are many data-driven models that associate mobility patterns to case prediction. Basic statistical models that are easy to implement, such as multiple regression, least absolute shrinkage and selection operator, as well as ridge regression (LASSO), have been tested for short-term prediction [16]. As the number of predicted days-head increased, however, the number of errors grew. Another simple and transparent statistical model that has been tried was to estimate the effect of non-pharmaceutical interventions on mobility by using basic machine learning methods to generate a 10-day ahead forecast. The indications are that mobility data on its own is sufficient to forecast meaningfully at all geographic scales [17]. Similarly, basic machine learning methods, such as regression tree, random forest and artificial neural network, have been combined in a linear model to deal with data in a non-linear relationship and with a non-normal distribution [18]. The authors recommended that more sophisticated models for better prediction accuracy need to be explored.

Some efforts have been made to capture inherent temporal dependencies and spatial correlations. Using a LSTM network, for example, the combination of mobility and meteorological data was found to be the primary factor in case prediction [19]. A susceptible to infectious transition rate, as a function of mobility and social behavior with time-dynamics parameters, was modeled by a deep LSTM, which was further integrated into a susceptible-exposed-infectious-recovered (SEIR) model for case forecast [20]. A graph neural network is another attempt to capture the spatio-temporal dynamics, where spatial edges represent mobility-based inter-region connectivity and temporal edges represent node features through time [21].

2.2. Lag effect

A crucial issue in modeling is to determine the lag in days between the mobility data and the confirmed cases. From the time an individual comes into contact with the COVID-19 virus and contracts the disease, confirming the diagnosis can be a complex process. The length of time it can take for a case to be confirmed can be affected by incubation period, testing speed and report delay. The incubation period can vary, but most cases manifest themselves between 3 and 7 days after initial infection [22]. The magnitude of the lag effect be even larger depending on the reporting variations in medical systems in different countries. A lag of 4 days was used to investigate the intervention policies across 33 provincial regions in China as a means of identifying both internal and external transmission effects [23]. An optimal lag of 11 days was found to achieve the highest correlation between the mobility and COVID-19 growth rate ratios for a single all-county model in the United States [24] while Noland used 7-day and 14-day lags to estimate the reproduction number R_0 by a log-linear model [25]. For the modeling of multiple countries, it would appear that a generic 7-day lag can indicate that a 10 percentage point reduction in mobility is associated with a 0.04-0.07 reduction in R_0 [26].

2.3. The Google data

Using smart devices and digital transactions is the most rapid and convenient way to collect human mobility data for forecasting the COVID-19 pandemic [27–29]. While the smart devices are GPS-tracked, the locations of the transaction data can be found at retail outlets, leisure facilities and other public amenities. Among the public datasets currently available, the Google COVID-19 Community Mobility Reports have been widely used for forecasting cases of infection and providing insights on how to use mobility characteristics efficiently [30–34]. Sufficient mobility records in both spatial and temporal dimensions enable the training of machine learning models that require large amount of data. Some research has adopted a number of statistical and machine learning models based on a recurrent neural network as well as an ensemble approach in order to predict trend changes in the 14-day cumulative incidence [34]. In this work, two datasets with similar training periods but different testing periods were used to compare the models. It is not surprising that the dataset with the shorter prediction period and stable trend outperformed the other one and demonstrated a greater predictability. The research also suggested that a 14-day cumulative incidence is predictive of mobility variables with the lag of seven days. In another work, a Network Inference-based Prediction Algorithm (NIPA), a combination of machine learning and phenomenological epidemiology, was proposed to predict the cumulative number of infected cases up to six days ahead [32]. The research that has been done using network-based approaches has considered the interactions between different regions. Including a time-varying or static prior close to the true contact network may improve the performance of a NIPA across logistic function, sigmoid curves and LSTM. A mobility-marked Hawkes model was proposed for case prediction in the early stages of the pandemic [33]. The study retooled to adopt a Hawkes process to capture the transmission dynamics. A mark, estimated by a regularized Poisson regression, was used to modulate the rate of infections and to account for the variations of R_0 across region and time. Finally, a correction procedure incorporated new cases seeded by people traveling between regions. With a forecasting period between 5 and 21 days, the model achieved outstanding performance when compared with six other baseline methods. Using the same data and the assumption of a 14 day incubation period, a recursive training-and-predicting

process was conducted by characterizing the spatio-temporal dynamics [30]. A Markov-Chain Monte-Carlo (MCMC) model was developed to estimate the spread of COVID-19 across a selection of 11 European countries and to further model the number of deaths from negative binomial distribution [31]. For some unknown reason, when a 3-week forecast was reached, the model tended to overestimate the impact of non-pharmaceutical interventions in Sweden and Denmark. These modeling outcomes have provided us with ample resources for benchmarking new modeling methods and suggesting a range of policy implications.

2.4. Indications

Various models and data sources have been identified in order to predict and manage the outbreak and spread of COVID-19 at different locations and at different geographical scales [35,36]. It is still challenging to build a common modeling framework for COVID-19 case prediction that fits all regions and time periods. In addition to the inherent and complex dependencies in both spatial and temporal dimensions, individual behaviors may also affect the predictions. For example, people wearing a mask spread the virus to a smaller number of people and thus contribute to a flattening of the infection curve, even in a scenario without mobility restriction [37]. Travel restrictions may be less effective once the outbreak is more widespread, although they were highly effective measures in the early stages of the pandemic [38]. However, the time boundary for distinguishing early and post-early stages is still unclear. Thus, any potential models need to be built separately in the study area and then compared across different scenarios before being deployed for policy making.

3. Modeling method

3.1. Models for count data

The number of daily confirmed COVID-19 cases can be modelled as count data from a Poisson [39,40] or a negative binomial distribution [41]. The conditional Poisson model, denoted by $Y_t = y | \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$, considers a set of historical factors, $\mathcal{F}_{t-1}$, at time $t - 1$ to imply the distribution at time t :

$$\mathcal{P}(Y_t = y | \mathcal{F}_{t-1}) = \frac{\lambda_t^y e^{-\lambda_t}}{y!}, y = 0, 1, 2, \dots \quad (1)$$

where Y_t is a random variable with conditional expectation and variance $E(Y_t = y | \mathcal{F}_{t-1}) = \text{Var}(Y_t = y | \mathcal{F}_{t-1}) = \lambda_t$. Supposing that the expectation can be characterized by $Z_t \lambda_t$ in a mixed Poisson model with a positive i.i.d. random variable Z_t from the gamma distribution, the marginal distribution of $\tilde{Y}_t = y | \mathcal{F}_{t-1}$ then becomes a negative binomial distribution when $E(Z_t) = 1$ and $\text{Var}(Z_t) = \sigma^2$ [42]. A dispersion parameter $\phi = 1/\sigma^2$ is introduced to formulate its probability mass function

$$\mathcal{P}(\tilde{Y}_t = y | \mathcal{F}_{t-1}) = \frac{\Gamma(y + \phi)}{y! \Gamma(\phi)} \left(\frac{\lambda_t}{\phi + \lambda_t} \right)^y \left(\frac{\phi}{\phi + \lambda_t} \right)^\phi, y = 0, 1, 2, \dots \quad (2)$$

denoted by $\tilde{Y}_t = y | \mathcal{F}_{t-1} \sim \text{NB}(\lambda_t, \phi)$. In this setting, $E(\tilde{Y}_t = y | \mathcal{F}_{t-1}) = \lambda_t$ and $\text{Var}(\tilde{Y}_t = y | \mathcal{F}_{t-1}) = \lambda_t + \lambda_t^2 / \phi$, which indicates that the Poisson variable is a limiting case when $\sigma^2 \rightarrow 0$.

3.2. Autoregressive model

In addition to the distributional features, the autoregressive relationship is also reflected in the time series data. In this paper, the Poisson autoregression is built upon an integer-valued generalized autoregressive conditional heteroskedastic (INGARCH) approach to linearly regress on a process of lagged values and conditional means [43,44]:

$$\lambda_t = \beta_0 + \sum_{i=1}^p \gamma_i Y_{t-i} + \sum_{j=1}^q \delta_j \lambda_{t-j}, \quad (3)$$

where β_0 , γ_i and δ_j are parameters to be estimated. The model facilitates the selection of lagged periods and the capture of seasonal effects by determining the values of γ_i and δ_j . One of the limitations for INGARCH is that the inclusion of time varying covariates can only provide positive effects. Negative effects lead to unreasonable regressors. To allow for negative effects, this work thus adopts a more general modeling form of conditional count time series data under the framework of a generalized linear model (GLM) [45,46], but extends this GLM by regressing on past terms such as Y_{t-i} and λ_{t-j} in Eq. (3) [47]. The *autoregressive count data* (ACD) model used in this paper is:

$$g(\lambda_t) = \beta_0 + \sum_{i=1}^p \gamma_i h(Y_{t-i}) + \sum_{j=1}^q \delta_j g(\lambda_{t-j}) + \boldsymbol{\eta}^\top \mathbf{X}_t, \quad (4)$$

where $g: \mathbb{R}^+ \rightarrow \mathbb{R}$ is a link function that maps the expectation to a linear regressor and relaxes the positive constraint of the regressor in Eq. (3). The function $h: \mathbb{N}_0 \rightarrow \mathbb{R}$ ensures that the lagged values are transformed to real values. $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_m)^\top$ is another set of parameters on the covariate vector $\mathbf{X}_t$. Similar to the GLM setting, the ACD model gives consistent maximum likelihood estimators when $g(x) = \log x$ and $h(x) = \log(x + 1)$ [48].

When an intervention, such as the introduction of a new travel policy related to COVID-19, occurs prior to t , the magnitude of its effect, $\xi \Delta^{t-\tau} \mathbb{1}(t \geq \tau)$, can be added, where ξ is the parameter, Δ determines the decay rate of the intervention effect and τ is the time of occurrence. $\mathbb{1}$ is an indicator taking value 1 for $t \geq \tau$ and 0 for $t < \tau$. When multiple interventions generate effects on λ_t for $0 < \Delta \leq 1$, a linear superposition of r , such interventions will form an additional component for the ACD model:

$$g(\lambda_t) = \beta_0 + \sum_{i=1}^p \gamma_i h(Y_{t-i}) + \sum_{j=1}^q \delta_j g(\lambda_{t-j}) + \boldsymbol{\eta}^\top \mathbf{X}_t + \sum_{k=1}^r \zeta_k \Delta^{t-\tau_k} \mathbb{1}(t \geq \tau_k), \quad (5)$$

where ζ_k is a set of parameters reflecting the power of intervention at τ_k . Thus, $\mathcal{F}_{t-1}$ is a collection of $\{Y_{t-1}, Y_{t-2}, \dots, Y_1, \lambda_{t-1}, \lambda_{t-2}, \dots, \lambda_1, \mathbf{X}_t, \mathbf{X}_{t-1}, \dots, \mathbf{X}_1, \Delta^{t-\tau_r}, \dots, \Delta^{t-\tau_1}\}$. The parameters to be estimated will be $\boldsymbol{\theta} = (\beta_0, \gamma_1, \dots, \gamma_p, \delta_1, \dots, \delta_q, \eta_1, \dots, \eta_m, \zeta_1, \dots, \zeta_r)^\top$. Some of the elements in $\boldsymbol{\theta}$ will equal 0 if the corresponding factors are not supposed to have an effect on λ_t . For the logarithm link g , constraints $|\beta_0|, |\gamma_1|, \dots, |\gamma_p|, |\delta_1|, \dots, |\delta_q| < 1$ and $|\sum_{i=1}^p \gamma_i + \sum_{j=1}^q \delta_j| < 1$ are further introduced [47]. The dispersion parameter ϕ in Eq. (2) will be discussed and estimated separately in next section.

3.3. Quasi-likelihood and parameter estimation

It is not uncommon for count data to be overdispersed, meaning that the variance can be larger than its mean value. Investigation into the variance-mean relationship can help with model selection [49]. For a Poisson model, for example, it might be too strict to have an identical mean and variance. One practical way of modeling is to relax the assumption on the random component to allow for the function $vV(\lambda_t) = \text{Var}(Y_t|\mathcal{F}_{t-1})$ to represent the randomness instead of characterizing an exponential distribution, where v is the dispersion parameter and $V(\lambda_t) = \lambda_t$ implies the Poisson model. This relaxation leads to a more general expression of the likelihood function, that is, a *quasi-likelihood* [50].

The construction of quasi-likelihood is based on a quasi-score function $S(\lambda)$ that is analogous to the score function of the ordinary likelihood. If we let

$$S(\lambda) = \frac{Y - \lambda}{vV(\lambda)}, \quad (6)$$

it can be easily determined that $E(S) = 0$ and $\text{Var}(S) = \frac{1}{vV(\lambda)} = -E(S')$. Thus, the quasi-likelihood function $Q(Y; \lambda)$ can be defined as [51]:

$$Q(Y; \lambda) = \int_{-\infty}^{\lambda} \frac{Y - u}{vV(u)} du + f(Y), \quad (7)$$

where $f(Y)$ is a real-value function of Y . The dispersion parameter v (ϕ for the negative binomial in this paper) usually does not affect the estimation of θ and can be estimated in a separate step. For the negative binomial distribution, one of the approaches is to solve Eq. (8) [42,52]:

$$\sum_{t=1}^n \frac{(\tilde{Y}_t - \hat{\lambda}_t)^2}{\hat{\lambda}_t(1 + \hat{\lambda}_t/\phi)} = n - (p + q + m + r + 1), \quad (8)$$

where, given the estimation of θ , $\hat{\lambda}_t = \lambda_t(\hat{\theta})$ is the fitted value of λ_t and n is the sample size.

In GLM, a quasi-Poisson model represents a class of distributions rather than a single quasi-Poisson distribution. The negative binomial model, one of the most convenient examples of the mixed Poisson process, is usually employed to account for the extra variations in the Poisson model [42]. Despite the fact that overdispersion is a more frequently observed feature [53], quasi-Poisson can also be used for modeling underdispersion. Thus, the conditional Poisson-like quasi-log-likelihood for the ACD model is specified by plugging $V(u) = u$ and $f(Y) = 0$ into Eq. (7):

$$l(\theta) = Q(y_t; \lambda_t(\theta)) = \sum_{t=1}^n (y_t \ln(\lambda_t(\theta)) - \lambda_t(\theta)). \quad (9)$$

It has been shown [47] that $\partial \lambda_t(\theta) / \partial \theta$ can be obtained recursively and thus come to form the differential calculations in the score function $\partial l(\theta) / \partial \theta$. The quasi-likelihood estimator is thus expressed as:

$$\hat{\theta} = \arg \max_{\theta} l(\theta). \quad (10)$$

To solve Eq. (10), the first step is to use an adaptive barrier algorithm [54] to enforce the constraints of the non-intervention parameters. A quasi-Newton BFGS is then applied for optimization without constraints [55–58].

4. Data and workflow

In this paper we illustrate the relationship between mobility and COVID-19 transmission by examining the case of Sweden, a country with the fifth largest land area in Europe and one characterized as having a low population density. Low population density implies there is a relatively large distance between people compared to other European countries and that the extent of unobserved close contact is also low.

4.1. Mobility data

Substantial efforts have been made to create data sets that reflect mobility change since the outbreak of the COVID-19 pandemic, where starting points between January 2020 and February 2020 have been selected to provide a baseline [59–61]. The baselines enable the user to compare the effect of policy interventions or mobility changes during the COVID-19 period. The mobility data usually reflects where people move from and to and how long they stay at their destinations. Due to fact that many travelers in our study region significantly changed their mobility patterns by abandoning public transport after the pandemic broke out [62], our assumption is that the highest risk of COVID-19 transmission occurs, not when people travel by public transport, but when they remain in a location for a period of time. In this paper, the mobility data collected by Google [63] will be divided into six different types of locations that represent similar categories of mobility behavior. These categories are *retail and recreation* (e.g., restaurants, cafes, museums, libraries, and movie theaters), *groceries and pharmacies* (e.g., food warehouses, farmers markets, specialty food shops, and pharmacies), *parks* (e.g., local parks, national parks, public beaches), *transit stations* (e.g., subway, bus, and train stations), *workplaces*, and *residential locations*.

The data presents the daily change of visits and length of stay, tracked by mobile devices, for each location category in Sweden from February 15, 2020 and compares it to the baseline days. The relative change in percentage is reported for Sweden as a whole as well as for each region and municipality. For the smaller municipalities without sufficient observations to aggregate, however, missing values are reported thus making it pointless to interpolate synthetic values. On the regional and national levels, the proportion of missing values is marginal for most of the categories. The baseline days are taken from five weeks at the beginning of 2020, where seven median values for each day of a typical week are selected as the baseline values.

4.2. Confirmed cases of COVID-19 in Sweden

There are 21 regions (administrative counties) in Sweden. The majority of daily moving activity in Sweden occurs within these regional borders because the places that most people visit each day, such as schools and hospitals, are regionally organized and tend to be located in the vicinity of residential areas and workplaces. The data produced by the Swedish Public Health

Agency provides a daily report of the total number of confirmed cases of COVID-19 for each region [64]. To control for the effect of regional population differences, we also consider the relative number of confirmed cases per hundred thousand inhabitants. A map of the 21 Swedish regions showing both calculations is shown in Figure 1.

Figure 1 Regions of Sweden and confirmed cases

In Figure 1, regional population density is also shown because it is an important local feature that may have a latent impact on modeling performance. Figure 1 shows, per region, the total number of confirmed COVID-19 cases and cases per thousand inhabitants from the date the first Swedish case was officially reported, February 4, 2020, to the date of writing, December 9, 2021. It is no surprise that the regions with the highest population density have the most confirmed cases. Stockholm, Sweden's capital city and the region with the country's highest population density, for example, has also recorded the highest number of cases, over 295,000. For the relative figures in the right panel of Figure 1, population density does not seem to be an obvious determinant of COVID-19 frequency.

4.3. Workflow

Based on the modeling framework and data described in previous sections, the workflow of the paper is summarized in Figure 2. The categories in the mobility data are covariates that were fixed in both the modeling and prediction stages. The COVID-19 data was fused with the administrative division data and was treated as a random variable. In order to evaluate the model's performance, five factors that might generate discrepancies in the modeling were considered. A holdout set was also selected to evaluate the model's performance. Point prediction was computed for all possible scenarios and intervals from the best models were taken to further examine the coverage. Finally, model scrutinization was carried out to investigate the statistical meaning of the models where intervention was included.

Figure 2 Modeling workflow

5. Results

In this section we present the five factors (and their values) that allowed us to observe and summarize a wide range of prediction results from multidimensional angles. Both point predictions and interval coverages were used to assess the performance of the models with different scenarios. After including interventions, the analytical results provided a number of insights into the statistical features.

5.1. Factors

The study period selected for modeling was January 1, 2021 to December 7, 2021, where the last two weeks in the series were taken to form a holdout set for testing the predictions from different models. Since mobility data was unavailable for two of these days, they were removed from the set. The five factors that were used to evaluate the model's performance in Figure 2 were lag, region, model, Y_{t-i} and λ_{t-i} . The number of elements and the specific values for each factor are presented in Table 1.

Table 1 Factors considered in modeling

Factors	No. of elements	Values
Lag (days)	11	2-12
Region ¹	21	20 regions plus whole country
Y_{t-i}	2	$i = 1$ and $i = 7$
λ_{t-i}	2	$i = 1$ and $i = 7$
Model	24	I, II, ..., XII for both Poisson and neg. bin.

¹ 1-whole country, 2-Blekinge, 3-Dalarna, 4-Gävleborg, 5-Halland, 6-Jämtland, 7-Jönköping, 8-Kalmar, 9-Kronoberg, 10-Norrbottn, 11-Örebro, 12-Östergötland, 13-Skåne, 14-Södermanland, 15-Stockholm, 16-Uppsala, 17-Värmland, 18-Västerbotten, 19-Västernorrland, 20-Västmanland, 21-Västra Götaland

Lag refers to the number of days prior to the COVID-19 report date that will be selected for modeling the mobility data. COVID-19 has a 4 to 9 day incubation period before symptoms are perceived [65]. When mobility is added to the mix, the lag for accurate prediction can rise to 12 days. [24]. It can also take several days for test results to be reported. Due to these delays, and so as not to miss any potential features, we studied a range of 2-12 days. Region refers to the geographical division upon which the model is built, in this case the Swedish administrative county. Of the 21 regions included in our models, one of them did not have complete mobility data for the whole of the study period. Thus, its data was disregarded and the models were built around the 20 remaining regions. We also included Sweden as a whole country as an additional region for comparison.

The two types of autoregressive terms, Y_{t-i} and λ_{t-i} , were used to determine how the number of confirmed cases for any given day were related to the previous observations and their expected values. Since the number of possible combinations was too great to calculate, we only considered three of them: the most recent ones: 1) $\langle Y_{t-1}, \lambda_{t-1} \rangle$, and one with weekly seasonality: 2) $\langle Y_{t-1}, \lambda_{t-7} \rangle$ and 3) $\langle Y_{t-7}, \lambda_{t-1} \rangle$. This was because the baseline days were chosen from a period prior to the onset of COVID-19 and represent a typical week. The variation in models concerns the specification of different covariates, how Y_t is measured and what distribution is assumed. For the covariates, missing values for the variable *parks* were found for some of the regions and thus this variable could not be included in our tests. The remaining five variables were included as covariates, both individually as well as all together, thus making six different models. For Y_t , two measurements - daily total confirmed cases and daily confirmed cases per hundred thousand inhabitants - were used for modeling. The latter measurement was given a relative value to eliminate the influence of population size. Thus, the Poisson model with six types of covariates to model total cases were denoted as I, II, ..., VI, respectively. Those modeling relative values were denoted as VII, VIII, ..., XII, respectively. The settings were replicated for negative binomial distribution. Since both distributions share the same likelihood function, as shown in Eq. (9), the point estimators of the parameters were the same. The difference between them can be seen in the estimation of standard errors and the intervals of predicted values. In conclusion, while the factors considered here may not be exhaustive, they are sufficient to provide an overview of a general evaluation framework.

5.2. Model evaluation

All of the factors taken together formed $11 \times 21 \times 3 \times 24 = 16\,632$ models that were evaluated on the holdout set. This implies that each model could generate additional 14-day predicted values. The accuracy was defined by the mean absolute relative errors (MARE):

$$MARE = \sum_{t=1}^{14} \frac{|\hat{Y}_t - Y_t|}{Y_t} \quad (11)$$

for each model. Since the case number 0 is not defined for MARE, it was excluded in the evaluation. We will show that this caused only negligible effects on the global evaluation. For each lag number and combination of autoregressive terms, the best MARE was selected to indicate which model-region among all of them was the most reliable. In Figure 3a, for example, the MAREs of the scenario $\langle Y_{t-1}, \lambda_{t-1} \rangle$ and a lag of 8 for all regions and models were normalized and compared. The outcome produced by region 15 and model XII gave the lowest

MARE. Similarly, the configuration for the scenario $\langle Y_{t-1}, \lambda_{t-7} \rangle$ and lag 5 was presented in Figure 3b and $\langle Y_{t-7}, \lambda_{t-1} \rangle$ and lag 10 in Figure 3c. From Figure 3, it seems that model performance tends to have less variance within regions than between regions.

Figure 3 Normalized MARE for three example scenarios where the best models were found: a) $\langle Y_{t-1}, \lambda_{t-1} \rangle$ and lag 8; b) $\langle Y_{t-1}, \lambda_{t-7} \rangle$ and lag 5; c) $\langle Y_{t-7}, \lambda_{t-1} \rangle$ and lag 10

In order to investigate how the top-ranked model-region combinations could be consistently obtained for each scenario, an overall evaluation was conducted and is presented in Table 2. It should be noted that Table 2 gives the results of the top two model-region combinations as well as the MAREs that applies to both distributions. In general, across different scenarios, ten out of twelve models achieved the lowest MAREs. For none of the scenarios did Model I or VII perform best, thus indicating that only measuring mobility change in retail and recreation locations by themselves does not provide reliable predictions. Model III and IV showed good accuracies for most lags given $\langle Y_{t-1}, \lambda_{t-1} \rangle$. Nevertheless, XII (with a lag 8 and 2) ranked in the top two for this setting. Similarly, model V (lag 5) and II (lag 6) achieved the lowest errors given $\langle Y_{t-1}, \lambda_{t-7} \rangle$ while model IV (lag 10 and 2) outperformed the other models for $\langle Y_{t-7}, \lambda_{t-1} \rangle$. With regard to the regions, almost all of the scenarios for $\langle Y_{t-1}, \lambda_{t-1} \rangle$ and $\langle Y_{t-1}, \lambda_{t-7} \rangle$ ended up with selecting the most populous region, 15-Stockholm. Only two scenarios that ranked second (lag 7 and 9 for $\langle Y_{t-1}, \lambda_{t-7} \rangle$) ended up with selecting region 1-whole country. For $\langle Y_{t-7}, \lambda_{t-1} \rangle$, the best models for all 22 scenarios were emerged from the data from region 18-Västerbotten, a region characterized by its low population density. The

following analysis, therefore, will focus on six of the models, the top 2 for each autoregressive setting.

Table 2 MARE for each lag number and combination of autoregressive terms

Lag No.	$\langle Y_{t-1}, \lambda_{t-1} \rangle$		$\langle Y_{t-1}, \lambda_{t-7} \rangle$		$\langle Y_{t-7}, \lambda_{t-1} \rangle$	
	1st	2nd	1st	2nd	1st	2nd
2	XII (0.189)	V (0.196)	IX (0.197)	III (0.202)	IV (0.133)	V (0.147)
3	III (0.195)	V (0.196)	XI (0.202)	X (0.210)	X (0.138)	IV (0.138)
4	III (0.196)	IV (0.200)	X (0.197)	V (0.206)	X (0.160)	IV (0.168)
5	III (0.202)	IV (0.202)	V (0.177)	XI (0.187)	V (0.158)	XI (0.159)
6	III (0.204)	IV (0.210)	II (0.186)	VIII (0.191)	X (0.149)	XI (0.152)
7	III (0.209)	IV (0.210)	VIII (0.205)	IV (0.207)*	VIII (0.157)	II (0.160)
8	XII (0.175)	VI (0.191)	XI (0.188)	XII (0.200)	IV (0.136)	V (0.138)
9	XII (0.203)	III (0.211)	IX (0.203)	XII (0.232)*	V (0.148)	II (0.162)
10	III (0.208)	IV (0.209)	XI (0.199)	X (0.206)	IV (0.130)	V (0.139)
11	IV (0.203)	III (0.207)	X (0.196)	IV (0.203)	X (0.150)	IV (0.158)
12	IV (0.201)	III (0.212)	X (0.188)	V (0.190)	IV (0.136)	X (0.141)

* the whole country as the best region

5.3. Intervals

In the 1-step-ahead prediction of $\hat{Y}_{t+1}$, the expected value λ_{t+1} was estimated, first of all, by the exponential operation and plugging in of the covariates at time $t + 1$ using $\hat{\theta}$:

$$\hat{\lambda}_{t+1} = \exp\{\lambda_{t+1}(\hat{\theta})\}. \quad (12)$$

To have a minimal mean square error, $\hat{Y}_{t+1}$ took the value $\hat{\lambda}_{t+1}$. Both of them, as well as the covariates at time $t + 2$, formed the conditional factor $\mathcal{F}_{t+1}$ to get the 2-step-ahead prediction of $\hat{\lambda}_{t+2}$ and $\hat{Y}_{t+2}$. At this point, we did not use any real observations to serve as past information for the whole predictions so that the model uncertainty was close to reality. The update continued until fourteen days were obtained. Meanwhile, we estimated 95% confidence intervals since they reflected the ability of a model to cover real values. The variances were given in Section 3.1. Illustrative intervals for the six Poisson models are shown in Figure 4. Intervals for negative binomial models were also computed, but they were less informative due to larger variance.

Figure 4 95% intervals for top two models for selected scenarios: a) $\langle Y_{t-1}, \lambda_{t-1} \rangle$; b) $\langle Y_{t-1}, \lambda_{t-7} \rangle$; c) $\langle Y_{t-7}, \lambda_{t-1} \rangle$

In Figure 4a, the intervals of the lag 8 model covered all real fourteen day-ahead values in the holdout set and the lag 2 model missed covering only one. This result accords with those in Table 2. In Figure 4b, since the intervals of the globally best model could still fail to include some values, the information provided was thus deemed to be unreliable. Figure 4c resembles Figure 4a because both lag 10 and lag 2 models were able to capture the variations. Model IV, however, was the most suitable for modeling the total number of confirmed cases. Notably, 0 cases were reported on day 5 and day 12. Models for these two days were able to cover or be very close to covering 0, implying that the model performance is resistant to extreme values. Overall, model XII at lag 8 and IV at lag 10 are trustable for modeling densely and sparsely populated regions, respectively. It is worth noting that the accuracy does not decline as the number of predicted days increases. However, there is no model that can better handle the autoregressive terms $\langle Y_{t-1}, \lambda_{t-7} \rangle$ than another.

5.4. Model scrutinization

These two identified models – model XII at lag 8 given $\langle Y_{t-1}, \lambda_{t-1} \rangle$ and model IV at lag 10 given $\langle Y_{t-7}, \lambda_{t-1} \rangle$ – were further investigated by adding an intervention and then comparing the two distributions. The intervention was introduced on September 29, 2021 because in Sweden on

that date many of the pandemic-related restrictions, such as the low maximum attendance numbers for private and public gatherings and events, were removed. Increased social activities could possibly change the mobility pattern. The decay rate Δ was set as 0.5. The MAREs, however, were found to be slightly smaller than those without interventions (Table 3). The models with interventions, therefore, are recommended for making inferences. Fitted values against respective real values were plotted in Figure 5, where smoothed Loess intervals were also given. In both Figure 5a and Figure 5b, the Loess intervals start to become larger at about the 97% quantile. The fitted values tend to underestimate the large real values. In addition, the fitted values for model XII at lag 8 with intervention present a quadratic trend while a linear trend is found for model IV at lag 10 with intervention. From the perspective of modeling, both types of trends can be characterized in GLM.

Figure 5 Fitted values against real values from models with interventions a) model XII at lag 8 given $\langle Y_{t-1}, \lambda_{t-1} \rangle$; b) model IV at lag 10 given $\langle Y_{t-7}, \lambda_{t-1} \rangle$.

Significant parameters, as well as the estimation of $1/\phi$ for both models, are also given at the 0.05 level in Table 3. While all parameters under a Poisson assumption are significant for model XII at lag 8, due to the larger variance, only γ_1 , $\eta_{w.p.}$ and $\eta_{res.}$ are significant for negative binomial distribution. Therefore, we consider that the difference between the two distributions might filter out the variables that depend on others. For example, reduced mobility at bus and train stations could increase the number of people gathered at residential locations. Thus, a direct recommendation to practitioners when making predictions for dense population areas is to collect mobility data from various types of locations. If difficulties restrict the full collection of data, the mobility changes collected from residential locations should be primarily used for modeling. One of the reasons for this is that residential location, in comparison to other location categories, is particularly sensitive to the key transmission mechanism of the COVID-19, that is, the personal contacts between family members and neighbors.

For model IV at lag 10, recommended for modeling regions with low density populations, the values reported for one week before and mobility change at the workplace were the most significant. When the negative binomial distribution is used, the weekly correlation, which also regulates working patterns, should be the one adopted for inference.

Table 3 Model comparison with an intervention

Model	XII, lag 8	IV, lag 10
Autoregressive factors	$\langle Y_{t-1}, \lambda_{t-1} \rangle$	$\langle Y_{t-7}, \lambda_{t-1} \rangle$
Sig. parameters ($\alpha = 0.05$, Poisson)	$\gamma_1 = 0.595; \delta_1 = -0.209;$ $\eta_{ret.} = 5.226; \eta_{pha.} = -2.892;$ $\eta_{tra.} = -4.702; \eta_{w.p.} = 3.294;$ $\eta_{res.} = 11.956$	$\gamma_7 = 0.845; \eta_{w.p.} = 0.228$
Sig. parameters ($\alpha = 0.05$, neg. bin.)	$\gamma_1 = 0.595; \eta_{w.p.} = 3.294;$ $\eta_{res.} = 11.956$	$\gamma_7 = 0.845$
$1/\phi$	1.333	4.282
Intervention date	September 29, 2021 (56 days to Day 1)	
Δ	0.5	
MARE	0.174	0.128

Finally, we computed the difference between the average predictive cumulative distribution function (CDF), $\bar{G}(y) \doteq \lim_{T \rightarrow \infty} \left\{ \frac{1}{T} \sum_{t=1}^T G_t(y) \right\}$, and the true data-generating CDF, $\bar{F}(y) \doteq \lim_{T \rightarrow \infty} \left\{ \frac{1}{T} \sum_{t=1}^T F_t(y) \right\}$, of the threshold values y to make a distributional comparison by the marginal calibration [66], where $G_t(y)$ and $F_t(y)$ are CDFs of sequential time series variables. $G_t(y)$ is said to be marginally calibrated relative to $F_t(y)$ if $\bar{G}(y) = \bar{F}(y)$ for all $y \in \mathbb{R}$ [67]. Thus, the difference between $\bar{G}(y)$ and $\bar{F}(y)$ reflects how close the predicted CDF from the empirical model is to the theoretical one.

Figure 6 Marginal calibration a) model XII at lag 8 given $\langle Y_{t-1}, \lambda_{t-1} \rangle$; b) model IV at lag 10 given $\langle Y_{t-7}, \lambda_{t-1} \rangle$.

Figure 6 provides a comparison of the two selected models. For both distributions in both models, the differences in CDFs are within ± 0.3 . For most threshold values, the Poisson models have the difference in CFD less than 0.1 for model XII and less than 0.05 for model IV. However, the difference given by the negative binomial distribution is relatively larger. For model IV, using Poisson assumption to model a set of values over 100 will largely satisfy the statistical assumptions. For the remaining sets of values, for asserting distributional features, the choice between the two distributions is inconclusive.

6. Conclusions

Since its first discovery in China back in December 2019, the COVID-19 pandemic has been responsible for at least 5.6 million deaths and has caused unfathomable adjustments to the world's social and economic systems. Every country in the world has introduced some sort of policy to cut off the channels of transmission by controlling human mobility, through such measures as national lockdowns, tight restrictions on public assemblies, self-monitored quarantines and reductions and restrictions on international travel. Despite these restrictions, some degree of mobility is inevitable even just to ensure the meeting of daily needs. Effective epidemic prevention and control, therefore, must take mobility into account. This paper used an autoregressive count data model under the framework of a generalized linear model and with five categories of mobility data collected by Google to explore the construction of a model that could be adopted to predict reliable confirmed cases in 14 days.

Empirical modeling and evaluation of long-term data from Sweden shows that: 1) Mobility change in retail and recreational locations, groceries and pharmacies, transit stations, workplaces and residential locations is the better way to predict confirmed cases in relative numbers for a region with a high population density (367 inhabitants/km² or higher, in this paper), where a lag of 8 days between the date of the observed mobility change and the reported cases is recommended in terms of interval coverage. Mobility in residential locations should be firstly collected with low availability of data. Past observations and conditional means one-period back may be included to provide autoregressive effects. 2) Mobility change at workplaces, as well as weekly seasonality are considered to be the most relevant variables when predicting total confirmed cases in sparsely populated regions (5 inhabitants/km² or lower, in this paper). In this case, a lag of 10 days is recommended, which is consistent with the results studied in other countries. The model used in these cases is also robust when the value zero is reported. 3) Policy interventions can be included when modeling long-term pandemics. In this paper, the effect of the lifting of the ban on public gatherings was used as a national intervention in the modeling with the result that the prediction accuracies were slightly improved. Although both Poisson and negative binomial distributions provided identical point predictions to confirmed cases, the interval estimation of the Poisson model was more compact. Underestimates in large magnitude were only detected for a small proportion (3%) from the fitted values. With regard to parameter estimation, negative binomial models tend to lead to fewer significant parameters due to the introduction of additional variance. This model may be used when the collection of mobility data is too

expensive. 4) A marginal calibration shows, from the statistical perspective, that the choice of distribution for count data is inconclusive except for the values over 100, where both Poisson and negative binomial models can be assumed when making the inference.

Addressing the limitations of this paper in future research may yield additional insights in epidemic modeling. First, as the development of communications technology and the internet of things continues apace, the inclusion of GPS-tracked mobility data during travel will add a new dimension to modeling, but will require more free accesses to the relevant databases. Second, it is still unknown how the models work in different environmental settings. Modeling outcomes from other countries could provide complementary information that would strengthen decision support. Third, other types of data-driven methods that can simultaneously handle time dependency, spatial correlation and count data might be alternative ways of modeling disease transmission.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

All authors contributed to the study conception and design. Jing Zhao and Mengjie Han contributed evenly to this work. Conceptualization, data collection and analysis were performed by Jing Zhao and Mengjie Han. The first draft of the manuscript was written by Jing Zhao and Mengjie Han. Zhenwu Wang and Benting Wan reviewed and edited the manuscript. All authors read and approved the final manuscript.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020;382:727–33.
2. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579:270–3.
3. WHO. Naming the coronavirus disease (COVID-19) and the virus that causes it. [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it) [Internet]. 2020 Feb. Available from: [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)
4. Alimohamadi Y, Taghdir M, Sepandi M. Estimate of the Basic Reproduction Number for COVID-19: A Systematic Review and Meta-analysis. *J Prev Med Public Health*. 2020;53:151–7.
5. Locatelli I, Trächsel B, Rousson V. Estimating the basic reproduction number for COVID-19 in Western Europe. Khudyakov YE, editor. *PLoS ONE*. 2021;16:e0248731.

6. Liu Y, Rocklöv J. The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *Journal of Travel Medicine*. 2021;28:taab124.
7. Nishiura H, Ito K, Anzai A, Kobayashi T, Piantham C, Rodríguez-Morales AJ. Relative Reproduction Number of SARS-CoV-2 Omicron (B.1.1.529) Compared with Delta Variant in South Africa. *JCM*. 2021;11:30.
8. GÜNER R, HASANOĞLU İ, AKTAŞ F. COVID-19: Prevention and control measures in community. *Turkish Journal of Medical Sciences*. 2020;50:572–7.
9. Ahmad S. A Review of COVID-19 (Coronavirus Disease-2019) Diagnosis, Treatments and Prevention. *EJMO* [Internet]. 2020 [cited 2021 Dec 27]; Available from: <https://www.ejmo.org/10.14744/ejmo.2020.90853/>
10. Davies NG, Klepac P, Liu Y, Prem K, Jit M, CMMID COVID-19 working group, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat Med*. 2020;26:1205–11.
11. Mandal M, Jana S, Nandi SK, Khatua A, Adak S, Kar TK. A model based study on the dynamics of COVID-19: Prediction and control. *Chaos, Solitons & Fractals*. 2020;136:109889.
12. Vella F, Senia P, Ceccarelli M, Vitale E, Maltezou H, Taibi R, et al. Transmission mode associated with coronavirus disease 2019: a review. *European Review for Medical and Pharmacological Sciences*. 2020;24:7889–904.
13. Tian H, Liu Y, Li Y, Wu C-H, Chen B, Kraemer MUG, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*. 2020;368:638–42.
14. Vannoni M, McKee M, Semenza JC, Bonell C, Stuckler D. Using volunteered geographic information to assess mobility in the early phases of the COVID-19 pandemic: a cross-city time series analysis of 41 cities in 22 countries from March 2nd to 26th 2020. *Global Health*. 2020;16:85.
15. Zhou Y, Xu R, Hu D, Yue Y, Li Q, Xia J. Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. *The Lancet Digital Health*. 2020;2:e417–24.
16. Wang R, Ji C, Jiang Z, Wu Y, Yin L, Li Y. A Short-Term Prediction Model at the Early Stage of the COVID-19 Pandemic Based on Multisource Urban Data. *IEEE Trans Comput Soc Syst*. 2021;8:938–45.
17. Ilin C, Annan-Phan S, Tai XH, Mehra S, Hsiang S, Blumenstock JE. Public mobility data enables COVID-19 forecasting and management at local and global scales. *Sci Rep*. 2021;11:13531.
18. Kuo C-P, Fu JS. Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions. *Science of The Total Environment*. 2021;758:144151.
19. Rashed EA, Hirata A. One-Year Lesson: Machine Learning Prediction of COVID-19 Positive Cases with Meteorological Data and Mobility Estimate in Japan. *IJERPH*. 2021;18:5736.
20. Bhourri MA, Costabal FS, Wang H, Linka K, Peirlinck M, Kuhl E, et al. COVID-19 dynamics across the US: A deep learning study of human mobility and social behavior. *Computer Methods in Applied Mechanics and Engineering*. 2021;382:113891.

21. Kapoor A, Ben X, Liu L, Perozzi B, Barnes M, Blais M, et al. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. arXiv:200703113 [cs] [Internet]. 2020 [cited 2022 Jan 13]; Available from: <http://arxiv.org/abs/2007.03113>
22. Wang B, Liu J, Li Y, Fu S, Xu X, Li L, et al. Airborne particulate matter, population mobility and COVID-19: a multi-city study in China. BMC Public Health. 2020;20:1585.
23. Oka T, Wei W, Zhu D. The effect of human mobility restrictions on the COVID-19 transmission network in China. Xue B, editor. PLoS ONE. 2021;16:e0254403.
24. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. The Lancet Infectious Diseases. 2020;20:1247–54.
25. Noland RB. Mobility and the effective reproduction rate of COVID-19. Journal of Transport & Health. 2021;20:101016.
26. Bergman NK, Fishman R. Mobility Levels and Covid-19 transmission Rates [Internet]. medrxiv; 2020 May. Available from: <http://medrxiv.org/lookup/doi/10.1101/2020.05.06.20093039>
27. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. Nature. 2021;589:82–7.
28. Guan G, Dery Y, Yechezkel M, Ben-Gal I, Yamin D, Brandeau ML. Early detection of COVID-19 outbreaks using human mobility data. Gadekallu TR, editor. PLoS ONE. 2021;16:e0253865.
29. Leung K, Wu JT, Leung GM. Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing. Nat Commun. 2021;12:1501.
30. Wang H, Yamamoto N. Using A Partial Differential Equation with Google Mobility Data to Predict COVID-19 in Arizona. Mathematical Biosciences and Engineering. 2020;17:4891–904.
31. Bryant P, Elofsson A. Estimating the impact of mobility patterns on COVID-19 infection rates in 11 European countries. PeerJ. 2020;8:e9879.
32. Achterberg MA, Prasse B, Ma L, Trajanovski S, Kitsak M, Van Mieghem P. Comparing the accuracy of several network-based COVID-19 prediction algorithms. International Journal of Forecasting [Internet]. 2020 [cited 2022 Jan 11]; Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169207020301552>
33. Schwabe A, Persson J, Feuerriegel S. Predicting COVID-19 Spread from Large-Scale Mobility Data. arXiv:210600356 [cs, stat] [Internet]. 2021 [cited 2022 Jan 11]; Available from: <http://arxiv.org/abs/2106.00356>
34. García-Cremades S, Morales-García J, Hernández-Sanjaime R, Martínez-España R, Bueno-Crespo A, Hernández-Orallo E, et al. Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data. Sci Rep. 2021;11:15173.
35. Mohamadou Y, Halidou A, Kapen PT. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. Appl Intell. 2020;50:3913–25.

36. Zhang M, Wang S, Hu T, Fu X, Wang X, Hu Y, et al. Human mobility and COVID-19 transmission: a systematic review and future directions [Internet]. medrxiv; 2021 Feb. Available from: <http://medrxiv.org/lookup/doi/10.1101/2021.02.02.21250889>
37. Lima LL, Atman APF. Impact of mobility restriction in COVID-19 superspreading events using agent-based model. Braunstein LA, editor. PLoS ONE. 2021;16:e0248708.
38. Kraemer MUG, Yang C-H, Gutierrez B, Wu C-H, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*. 2020;368:493–7.
39. Zhang X, Ma R, Wang L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. *Chaos, Solitons & Fractals*. 2020;135:109829.
40. Agosto A, Giudici P. A Poisson Autoregressive Model to Understand COVID-19 Contagion Dynamics. *Risks*. 2020;8:77.
41. Oztig LI, Askin OE. Human mobility and coronavirus disease 2019 (COVID-19): a negative binomial regression analysis. *Public Health*. 2020;185:364–7.
42. Lawless JF. Negative binomial and mixed poisson regression. *Can J Statistics*. 1987;15:209–25.
43. Ferland R, Latour A, Oraichi D. Integer-Valued GARCH Process. *J Time Series Analysis*. 2006;27:923–42.
44. Fokianos K, Rahbek A, Tjøstheim D. Poisson Autoregression. *Journal of the American Statistical Association*. 2009;104:1430–9.
45. Nelder JA, Wedderburn RWM. Generalized Linear Models. *Journal of the Royal Statistical Society Series A (General)*. 1972;135:370.
46. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. Boca Raton: Chapman & Hall/CRC; 1989.
47. Liboschik T, Fokianos K, Fried R. tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. *J Stat Soft* [Internet]. 2017 [cited 2021 Dec 2];82. Available from: <http://www.jstatsoft.org/v82/i05/>
48. Fokianos K, Tjøstheim D. Log-linear Poisson autoregression. *Journal of Multivariate Analysis*. 2011;102:563–78.
49. Ver Hoef JM, Boveng PL. QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA? *Ecology*. 2007;88:2766–72.
50. McCullagh P. Quasi-Likelihood Functions. *Ann Statist*. 1983;11:59–67.
51. Wedderburn RWM. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*. 1974;61:439.
52. Christou V, Fokianos K. QUASI-LIKELIHOOD INFERENCE FOR NEGATIVE BINOMIAL TIME SERIES MODELS: NEGATIVE BINOMIAL TIME SERIES. *J Time Ser Anal*. 2014;35:55–78.
53. Zhu F. Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *Journal of Mathematical Analysis and Applications*. 2012;389:58–71.
54. Lange K. Numerical analysis for statisticians. 2nd ed. New York: Springer; 2010.

55. Broyden CG. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. IMA J Appl Math. 1970;6:76–90.
56. Goldfarb D. A family of variable-metric methods derived by variational means. Math Comp. 1970;24:23–6.
57. Shanno DF. Conditioning of quasi-Newton methods for function minimization. Math Comp. 1970;24:647–56.
58. Fletcher R. A new approach to variable metric algorithms. The Computer Journal. 1970;13:317–22.
59. Hale T, Angrist N, Cameron-Blake E, Hallas L, Kira B, Majumdar S, et al. Variation in government responses to COVID-19. Oxford COVID-19 Government Response Tracker, Blavatnik School of Government. 2020;
60. Aktay A, Bavadekar S, Cossoul G, Davis J, Desfontaines D, Fabrikant A, et al. Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.1). arXiv:2004.04145 [cs] [Internet]. 2020 [cited 2021 Dec 8]; Available from: <http://arxiv.org/abs/2004.04145>
61. Apple. COVID-19 Mobility Trends Reports, <https://covid19.apple.com/mobility> [Internet]. 2021. Available from: <https://covid19.apple.com/mobility>
62. Jenelius E, Cebecauer M. Impacts of COVID-19 on public transport ridership in Sweden: Analysis of ticket validations, sales and passenger counts. Transportation Research Interdisciplinary Perspectives. 2020;8:100242.
63. Google LLC. Google COVID-19 Community Mobility Reports". <https://www.google.com/covid19/mobility/> Accessed: 2021-12-10.
64. The Public Health Agency of Sweden. <https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/bekraftade-fall-i-sverige/>. 2021.
65. Tindale LC, Stockdale JE, Coombe M, Garlock ES, Lau WYV, Saraswat M, et al. Evidence for transmission of COVID-19 prior to symptom onset. eLife. 2020;9:e57149.
66. Christou V, Fokianos K. On count time series prediction. Journal of Statistical Computation and Simulation. 2015;85:357–73.
67. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J Royal Statistical Soc B. 2007;69:243–68.