Correlation of subway turnstile entries and COVID-19 incidence and deaths in New York City

Sina Fathi-Kazerooni a, Roberto Rojas-Cessa a, *, Ziqian Dong b, Vatcharapan Umpaichitrac

a Department of Electrical and Computer Engineering, Newark College of Engineering, New Jersey Institute of Technology, Newark, NJ, 07102, USA
b Department of Electrical and Computer Engineering, College of Engineering and Computing Sciences, New York Institute of Technology, New York, NY, 10023, USA
c Department of Pediatrics, SUNY Downstate Health Sciences University, Brooklyn, NY, 11203, USA

A R T I C L E   I N F O

Article history:
Received 25 August 2020
Received in revised form 14 November 2020
Accepted 16 November 2020
Handling editor, Dr. J Wu

Keywords:
COVID-19
Time-series analysis
New York city subway
SARS-CoV-2
Long short-term memory
ARIMA

A B S T R A C T

In this paper, we show a strong correlation between turnstile entries data of the New York City (NYC) subway provided by NYC Metropolitan Transport Authority and COVID-19 deaths and cases reported by the NYC Department of Health from March to May 2020. This correlation is obtained through linear regression and confirmed by the prediction of the number of deaths by a Long Short-Term Memory neural network. The correlation is more significant after considering incubation and symptomatic phases of this disease as experienced by people who died from it. We extend the analysis to each individual NYC borough. We also estimate the dates when the number of COVID-19 deaths and cases would approach zero by using the Auto-Regressive Integrated Moving Average model on the reported deaths and cases. We also backward forecast the dates when the first cases and deaths might have occurred.

© 2020 The Authors. Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

New York City (NYC) was the epicenter of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in early 2020 (Levenson, 2020). Many people have been hospitalized and also many others have lost their lives to this disease. It is unknown when the first COVID-19 cases occurred in the city, but the first case was reported on Feb. 29, 2020. The COVID-19 cases and deaths in NYC increased rapidly after that date, peaked in mid-April 2020 (NYC Health Department, 2020a), and started to decline after the city implemented a series of containment measures, including closing schools, social distancing, business shut-downs, and shelter-in-place (Nicola et al., 2020). As a result, the COVID-19 pandemic has brought and continues to bring economic hardship to NYC residents and people worldwide, and especially those in the lower economical strata (Wadhera et al., 2020). As city dwellers mainly rely on public transportation and more notably the subway to move around, it is expected that a highly contagious virus such as SARS-CoV-2 would easily...
spread through the NYC population. It is expected that the NYC subway, being an enclosed environment and also a major indicator of people’s social activity, would be directly correlated with the spreading of SARS-CoV-2.

Although this correlation has been described by reported cases throughout neighborhoods, the correlation between the subway ridership and COVID-19 incidence and deaths has not been shown. In this paper, we show this correlation between the public turnstile entry data of NYC subway provided by the NYC Metropolitan Transportation Authority (MTA) (NYC MTA, 2020; Whong, 2020) and the statistics of the confirmed COVID-19 deaths provided by the NYC Department of Health (DOH) (NYC Health Department, 2020b). We used linear regression and tested the turnstile data from a few days before the day of the reported deaths and cases to support this correlation. There are many prediction models for forecasting the spread of COVID-19 (Al-Qaness et al., 2020; Anastassopoulou et al., 2020; Ceylan, 2020, p. 138817; Chakraborty & Ghosh, 2020, p. 109850; Fanelli & Piazza, 2020; Goscé & Johansson, 2018; Kucharski et al., 2020; Kuniya, 2020; Li et al., 2020; Mo et al., 2020; Perc et al., 2020; Petropoulos & Makridakis, 2020; Qin et al., 2020; Ribeiro et al., 2020, p. 109853; Tosepu et al., 2020; Yang et al., 2020; Zhang et al., 2020). Much of the literature focuses on modeling the incidence of COVID-19 in different countries and partially enclosed environments, such as cruise ships (Ceylan, 2020, p. 138817; Fanelli & Piazza, 2020; Kucharski et al., 2020; Li et al., 2020; Ribeiro et al., 2020, p. 109853; Tosepu et al., 2020; Zhang et al., 2020). These models use local and regional data to examine the correlation of COVID-19 cases with environmental or social factors. Many of the approaches use the Susceptible-Exposed-Infected-Recovered epidemiological model to analyze data on this outbreak. These models rely on the accuracy of the reported cases. However, the accuracy of case reports is expected to be low at the early stage of this new outbreak and that may add significant errors in their analysis and forecast. To avoid such errors, some models forecast the number of COVID-19 cases according to the population size (Kermack & McKendrick, 1927; Petropoulos & Makridakis, 2020; Qin et al., 2020). There is an increasing interest in models that estimate the impact of public transit on incidence and the possibilities of contagion by airborne transmissions (Goscé & Johansson, 2018; Mo et al., 2020).

Our analysis leverages on the reported number of deaths because their count has higher accuracy than the reported number of cases. However, the data on cases is larger than the one on deaths, so we also analyze the reported cases. We employ linear regression for analyzing the correlation of the reported number of COVID-19 deaths and the NYC subways turnstile entries and Long Short-Term Memory (LSTM) neural network (Hochreiter & Schmidhuber, 1997) for predicting the number of deaths and cases through time-series data analysis to confirm the correlation between these two data sets. We extend this analysis to each of the five NYC boroughs: Bronx, Brooklyn, Manhattan, and Staten Island.

We also use the Auto-Regressive Integrated Moving Average (ARIMA) model to 1) predict the dates when COVID-19 deaths and cases would approach zero according to the reported data, and to 2) estimate the time when the contagion might have started in NYC by identifying the dates when the first case and death might have occurred through analysis of the provided cases and deaths data.

The analysis performed in this paper exclusively uses the mentioned public datasets, and so it leaves out factors that may have affected the actual counts of cases and deaths. For example, the data analyzed in this paper was recorded while no face coverings were widely considered by NYC dwellers.

The remainder of this paper is organized as follows. Section 2 describes the models and methods used in our data analysis. Section 3 describes the obtained results. Section 4 presents a discussion and remaining questions. Section 5 presents our conclusions.

2. Model description

In this section, we present the data sources and analysis tools used in this paper.

2.1. Datasets

The NYC DOH dataset reports the first COVID-19 case on Feb. 29, 2020 and the first death on Mar. 11, 2020. The MTA turnstile entry data includes the number of subway entries and exits per station in NYC, recorded every 4 h each day from a total of 379 subway stations (NYC MTA, 2020; Whong, 2020). We calculate the average daily entries of NYC subway stations per day for our analysis. We also use the number of daily cases and the number of daily deaths of the COVID-19 dataset published by the NYC DOH (NYC Health Department, 2020b).

We consider the reported COVID-19 incubation and symptomatic periods (Centers for Disease Control and Prevention (CDC), 2020; Guan et al., 2020), as a combined period, in the reported subway entry data as features in the analysis. We call these features shifted days. For example, a shifted day is a day prior to the day under prediction. The number of daily subway entries is denoted as $x_t$ for a given day $t$, and the number of turnstile entries of $m$ days before $t_0$ is denoted as $x_{t_0-m}$. We use $x_{t_0}$ and $x_{t_0-m}$ as features to predict the number of deaths and cases for day $t_0$, where $1 \leq m \leq 25$. We use 20% of the NYC DOH data as test data for validation and the rest for training.

2.2. Forecasting models

In our analysis, we use linear regression with L1 regularization to find the correlation between the two datasets and the precedence of the different shifted days. We use LSTM, a recurrent neural network (RNN) (Hochreiter & Schmidhuber, 1997) that learns the time evolution of data sequences (Hochreiter & Schmidhuber, 1997; Ramsundar & Zadeh, 2018), to predict the
number of daily deaths and the number of daily cases and evaluate their accuracy as verification of the correlation in NYC and its five boroughs.

We use ARIMA, a tool to forecast time-series data (Box et al., 2011; Flandoli, 2020), to predict the number of daily deaths and the number of daily cases, evaluate their accuracy, and estimate the dates of when the first case and death may have occurred. For the latter use, we analyze the data in a backward (time) sequence.

We adopt multiple linear regression with \( x_j \) independent variables (Montgomery et al., 2012). The estimated value \( \hat{y} \) in the linear regression model follows:

\[
\hat{y} = \beta_0 + \sum_{j} \beta_j x_j + \epsilon
\]

where \( \beta_j \) represents the model weights or regression coefficients and \( \epsilon \) is the error. L1 regularization is the absolute sum of the model weights multiplied by a shrinkage value \( (\lambda) \). It is formulated as \( \lambda \sum_{j} |\beta_j| \) and the regression model goal is to minimize:

\[
\sum_{i} (y_i - \hat{y})^2 + \lambda \sum_{j} |\beta_j|
\]

where \( y_i \) is the actual data point and \( x_{ij} \) is the value of the independent variables for each data point.

We evaluate the accuracy of the prediction performed through LSTM as an indicator of the correlation extent between the two datasets. We use a Savitzky–Golay filter to smooth out the prediction results of LSTM (Schafer, 2011).

ARIMA consists of three models: auto-regressive (AR), integration, and moving average (MA). It uses three hyper parameters: \( p, d, \) and \( q \), where \( p \) is the order of auto-regressive model, \( d \) is the degree of difference, and \( q \) is the order of moving average (Box et al., 2011). The ARIMA\((p, d, q)\) is defined as

\[
(1 - \sum_{k=1}^{p} \alpha_k L^k) (1 - L)^d X_t = (1 + \sum_{k=1}^{q} \beta_k L^k) \epsilon_t
\]

where \( \alpha \) is the coefficient of discrete time linear equation of AR, \( L \) is a time lag operator defined as \( LX_t = X_{t-1} \), \( X_t \) is the observed value at time \( t \), \( \beta \) is the coefficient for the noise term, \( \epsilon_t \) in MA (Flandoli, 2020).

### 2.3. Performance metrics

The performance of the prediction models is evaluated by the \( R^2 \) score, mean absolute error (MAE), and root mean square error (RMSE). The \( R^2 \) score indicates how close the data are from the obtained regression (Montgomery et al., 2012). \( R^2 \) score is calculated as

\[
R^2 = 1 - \frac{SS_R}{SS_T}
\]

where \( SS_R \) is the residual sum of squares and \( SS_T \) is the total sum of squares. \( SS_R \) is calculated as \( \sum (y_i - \hat{y}_i)^2 \), where \( y_i \) is the actual values and \( \hat{y}_i \) is the predicted values. \( SS_T \) is calculated as \( \sum (y_i - \bar{y})^2 \), where \( \bar{y} \) is the mean of the actual values. \( R^2 \) shows the variance of residuals (or prediction errors) in the predicted data points divided by the variance of data points from the average of all data points. \( R^2 \) represents the performance of a model as compared to randomly guessed predictions that are equal to the average of all data. The closer \( R^2 \) is to 1, the predictions are closer to the actual values.

MAE is defined as the average of sum of residuals, or

\[
\frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}
\]

where \( n \) is the number of observations. An MAE that approaches 0 means that the model predicts data points with high accuracy. We use MAE to measure the training and validation losses of LSTM and, in turn, to show whether the prediction is a good fit and not overfitting or underfitting occurs. If the training loss is significantly lower than the validation loss, the model is overfitting the training data, which means the model is learning complex patterns of training data that may not generalize in predicting unseen test data and it results in poor performance. If the validation loss is significantly lower than training loss, the model is underfitting the training data, which means that the model is unable to learn important patterns of the training data, and it results in poor model performance. The training is stopped and the model is improved to avoid overfitting and underfitting if the losses grow (although in different directions) as the training progresses (Abadi et al., 2015).

RMSE is defined as
In other words, RMSE is the standard deviation of residuals. With normally distributed data points, one can expect 68% of predicted points to be within one RMSE from the mean and 95% to be within two RMSE from the mean. In LSTM predictions, we consider the combination of $R^2$ scores and RMSE to assess accuracy.

Fig. 1 shows the number of subway turnstile entries and the reported daily COVID-19 deaths and cases from Mar. 1, 2020 to May 15, 2020. The trend of deaths trails the cases by a few days. The turnstile data include variations of ridership before the city executed a lockdown. We tested the correlation between NYC subway turnstile entry data and reported daily COVID-19 deaths with turnstile entry data of the predicted day $t_0$ and $m$ shifted days ($t_0 - m$) for $1 \leq m \leq 40$, using linear regression.

Fig. 2 shows the calculated $R^2$ score for $t_0$ and each shifted day. The score of each considered shifted day is about or larger than 0.2, showing a strong correlation between these two data sets. Next, we predict the cases and deaths using LSTM to verify the correlation of the data sets. From the different number of shifted days shown in Fig. 2, we selected those that correspond to the possible reported incubation period of SARS-CoV-2 in a person plus possible symptomatic days. We used three feature sets (Feature sets A, B, and C, as shown in Table 1). Feature set A uses the entries of the predicted day and the 14th prior day ($x_{t_0}, x_{t_0 - m}$; where $m = 14$). Feature set B uses entries of the predicted day and five prior days: 10th, 12th, 13th, 14th, and 17th prior days ($x_{t_0}, x_{t_0 - m}$, where $m = 10, 12, 13, 14, 17$). Feature set C includes the entries of the predicted day and prior-day entries of each of the considered 25 days ($x_{t_0}, x_{t_0 - m}$; where $m = 1, \ldots, 25$). This set is selected for completeness and as a reference.

3. Results of data analysis

3.1. Correlation between NYC subway turnstile entries and COVID-19 deaths and cases

We used the feature sets described in Section 2 on LSTM to predict the number of deaths and the number of cases. We compared the predicted numbers using the different features with the actual reported numbers.

Fig. 3(a) and (b) show the MAE of LSTM training and validation losses for the prediction of the number of deaths using Feature sets A and C, respectively. Feature set A produces an $R^2$ score of 0.61 and an RMSE of 68.15. Feature set C produces an $R^2$ score of 0.98 and an RMSE of 17.29. Fig. 3(c) shows the prediction of the number of deaths using Feature sets A and C, respectively. Feature set C is the top performing feature set and Feature set B is the bottom performing one for the prediction of deaths, as shown in Table 2. The results show that the correlation is strong with one shifted day and the correlation strengthens by adding more shifted days to the feature set as the scores show for the three considered feature sets. We also tested the use of the three feature sets on LSTM to predict the number of cases. Fig. 4(a) and (b) show the MAE of training and validation losses in the prediction of cases for Feature sets A and C, respectively. The $R^2$ score on the test data is 0.74 and the
RMSE is 1,510.43 for Feature set A, and the $R^2$ score and the test data is 0.99 and the RMSE is 389.71 for Feature set C. The LSTM predicted cases using Feature set A (Fig. 4(c)) follows the actual data with, however; small discrepancies. The prediction using Feature set C also follows the actual data; however, it is closer than that predicted with Feature set A. Table 2 summarizes the $R^2$ scores and RMSE for each feature set in the prediction of cases. The RMSE of the predictions of the number of cases is much higher than that of the prediction of the number of deaths because the number of cases each day is larger than deaths. These losses further confirm the lower accuracy of the reported cases despite the high $R^2$ scores.

3.2. Correlation of data set per NYC borough

We analyzed the LSTM predictions of turnstile data and COVID-19 cases and deaths for each NYC borough to show the correlation for the different boroughs of NYC. Each borough may have different socioeconomic status. Table 3 lists the $R^2$ and RMSE for each feature set and borough.

**LSTM prediction of the number of deaths and cases per borough data**

We also evaluated a prediction of the number of deaths and cases per borough considering the turnstile entry data of the boroughs.

**Prediction of the number of deaths per borough.** Table 3 summarizes the $R^2$ and RMSE of the prediction of deaths per borough. Each of these scores was estimated using NYC-subway turnstile entry data per borough for each of the three feature sets used; A, B, and C. The table shows that $R^2$ is the highest for Queens for all feature sets.

The RMSE is the smallest for Manhattan and Staten Island for all feature sets. These values combined with their $R^2$ shows that the correlation is high. The RMSE is highest for Brooklyn and Queens in general. During the analyzed period of time, these two boroughs reported the highest number of deaths.

**Predictions of the number of cases per borough.** Table 3 also shows $R^2$ and RMSE for the number of cases per borough. Different from deaths, cases shows high $R^2$ for the five boroughs; and reaching 0.97 for Queens with Feature set C. However, their RMSE is also higher than those for deaths. These scores present close similarities to those obtained at NYC level (i.e., all boroughs together).

**Predictions of the number of deaths by LSTM per borough.** Fig. 5 shows the predicted number of deaths for each of the five NYC boroughs. This figure shows the bottom and top performing feature sets; Feature sets A and C, respectively. The figure shows that for Feature set A, Queens shows the highest similarity to the actual values, and Manhattan and Staten Island the lowest similarity to the actual values. As for Feature set C, the prediction of deaths is very similar to the actual data for the five
Table 1
Feature sets of training data derived from average daily entries of NYC subway stations.

| Feature set   | Days before predicted day |
|---------------|---------------------------|
| Feature set A | Predicted day and 14 days ago |
| Feature set B | Predicted day, 10, 12, 13, 14, and 17 days ago |
| Feature set C | Predicted day, 1, 2, 3, 4, 5, ..., and 25 days ago |

Fig. 3. MAE of LSTM training and validation losses in prediction of deaths using (a) Feature set A and (b) Feature set C. (c) Prediction of COVID-19 deaths using Feature sets A and C on LSTM.

Table 2
RMSE of all data points and $R^2$ scores of test data for LSTM predictions of deaths and cases in NYC.

| Features - Deaths | $R^2$ | RMSE   |
|-------------------|-------|--------|
| Set A             | 0.98  | 17.29  |
| Set B             | 0.61  | 68.15  |
| Set C             | 0.58  | 60.28  |

| Features - Cases  | $R^2$ | RMSE   |
|-------------------|-------|--------|
| Set A             | 0.99  | 389.71 |
| Set B             | 0.84  | 616.63 |
| Set C             | 0.74  | 1510.43|
boroughs. Manhattan and Staten Island reported the smallest number of cases during this period. Therefore, Feature set C confirms the high correlation of deaths and turnstile data for these boroughs.

Predictions of the number of cases by LSTM per borough. Fig. 6 shows the predicted number of cases for each of the five NYC boroughs. Bronx, Brooklyn, and Queens reported the larger number of cases during this period. Different from the predicted...
Fig. 5. LSTM predictions of deaths per borough: (a) Bronx, (b) Brooklyn, (c) Manhattan, (d) Queens, and (e) Staten Island.
Fig. 6. LSTM results of cases predictions in (a) Bronx, (b) Brooklyn, (c) Manhattan, (d) Queens, and (e) Staten Island.
number of deaths, Feature sets A and C closely follow the actual data on reported cases for the five boroughs. Such predictions are very similar to the one obtained for NYC as a whole.

These results show that the LSTM predictions can be transported to smaller segments of the city. They are consistent with those obtained for the entire city.

3.3. Estimation of trends and dates of minimum cases and deaths

Fig. 7(a) shows that considering the time series of the number of deaths on May 16, 2020, the lower bound (with 95% confidence interval) of ARIMA forecast zero deaths on that date. On the other hand, the upper bound does not cross zero, indicating that the number of deaths could also increase. Fig. 7(b) shows the ARIMA backward forecast of the number of deaths to estimate the date when the first death may have occurred in NYC. The graph shows that the forecast date of the first death could have occurred on Mar. 9, 2020, two days earlier than the first reported death. Fig. 7(c) shows the ARIMA forecast of the progression of the number of COVID-19 cases in NYC after May 15, 2020 (last day of reported data). The forecast follows the decreasing trend of the number of cases, reaching zero on Jun. 20, 2020. However, this is a very early date as the number of cases has not reached zero by the time this paper was written. Fig. 7(d) shows the backward forecast of the number of cases to estimate the date of the first case in NYC. The lower and upper bounds of 95% confidence interval of ARIMA forecast shows that the first case of COVID-19 in NYC might have occurred between Jan. 28 and Mar. 20, 2020. The first case was reported on Feb. 29, 2020, but the actual first case is unknown. Table 4 shows the first reported COVID-19 cases (Feb. 29, 2020) and deaths (Mar. 11, 2020) in NYC DOH dataset.

4. Discussion

The reporting of cases has been affected by the deployment and execution of testing for SARS-CoV-2. The effective and widespread testing offers better identification of the cases and their locations. But because of the time needed to detect a
likely case, the limitation of testing capabilities at the beginning of the pandemic, and the possible occurrence of asymptomatic cases. The estimated number of COVID-19 cases is likely inaccurate.

The estimations performed with ARIMA are based on the provided data, with no additional information. The discrepancies on the estimated number of COVID-19 cases and the reported number of cases are the product of the model using only past data from the two analyzed datasets. Public health suggestions on the use of face covering and social distancing are not considered in the models to reflect the multi-variant effects in linear regression. The use of face coverings adopted at some point in time or neglecting social distancing may affect the actual trends that past data may not be able to forecast.

Looking ahead, comparisons of the incidence of COVID-19 in NYC with cities that experienced an early face-covering measure is of interest for forecasting future incidence or containment of the spread of the virus.

5. Conclusions

We have shown that the NYCturnstile entry data and the COVID-19 deaths and cases data in NYC from March 1, 2020 to May 15, 2020 are highly correlated. We have also shown that the correlation is also detectable for the data of each individual borough of NYC. However, each borough shows a different correlation strength that is probably associated with their geographical location and economical status. The consideration of different incubation–symptomatic phases for subway users as features through Long Short-Term Memory analysis confirms and enhances this correlation. Based on the dataset on COVID-19 deaths and cases in NYC, we have also shown the predicted progression of deaths and cases beyond the dates of the reported data using the AutoRegressive Integrated Moving Average forecast model. This progression shows the possible decline of COVID-19 incidence but also a probable increase. We have also shown a backward forecast of the possible dates when the first case and death might have occurred at the beginning of the pandemic in NYC. The backward forecast of the first death shows high accuracy.

Declaration of competing interest

The authors declare having no conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Viyyal, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/. software available from tensorflow.org.

Al-Qaness, M. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization method for forecasting COVID-19 pandemic. http://users.dma.unipi.it/flandoli/AUJTcap4.pdf.

Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One, 15, Article e0230405.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2011). Time series analysis: Forecasting and control (Vol. 734). John Wiley & Sons.

Centers for Disease Control and Prevention (CDC). (2020). Interim clinical guidance for management of patients with confirmed coronavirus disease (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html.

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. Society of the total environment. Information and communication technology. The authors declare having no conflict of interest.

Declaration of competing interest

The authors declare having no conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Viyyal, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/. software available from tensorflow.org.

Al-Qaness, M. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization method for forecasting COVID-19 pandemic. http://users.dma.unipi.it/flandoli/AUJTcap4.pdf.

Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One, 15, Article e0230405.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2011). Time series analysis: Forecasting and control (Vol. 734). John Wiley & Sons.

Centers for Disease Control and Prevention (CDC). (2020). Interim clinical guidance for management of patients with confirmed coronavirus disease (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html.

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. Society of the total environment. Information and communication technology. The authors declare having no conflict of interest.

Declaration of competing interest

The authors declare having no conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Viyyal, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/. software available from tensorflow.org.

Al-Qaness, M. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization method for forecasting COVID-19 pandemic in China. Journal of Clinical Medicine, 9, 674.

Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One, 15, Article e0230405.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2011). Time series analysis: Forecasting and control (Vol. 734). John Wiley & Sons.

Centers for Disease Control and Prevention (CDC). (2020). Interim clinical guidance for management of patients with confirmed coronavirus disease (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html.

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. Society of the total environment. Information and communication technology. The authors declare having no conflict of interest.

Declaration of competing interest

The authors declare having no conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Viyyal, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/. software available from tensorflow.org.

Al-Qaness, M. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization method for forecasting COVID-19 pandemic in China. Journal of Clinical Medicine, 9, 674.

Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One, 15, Article e0230405.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2011). Time series analysis: Forecasting and control (Vol. 734). John Wiley & Sons.

Centers for Disease Control and Prevention (CDC). (2020). Interim clinical guidance for management of patients with confirmed coronavirus disease (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html.

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. Society of the total environment. Information and communication technology. The authors declare having no conflict of interest.

Declaration of competing interest

The authors declare having no conflict of interest.

References
Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., Wang, D., Chen, G., Zhang, J., Peng, H., et al. (2020). Propagation analysis and prediction of the COVID-19. *Infectious Disease Modelling*, 5, 282–292.

Mo, B., Feng, K., Shen, Y., Tam, C., Li, D., Yin, Y., & Zhao, J. (2020). Modeling epidemic spreading through public transit using time-varying encounter network. arXiv preprint arXiv:2004.04602.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

Nicola, M., Alsaifi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., & Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (covid-19): A review. *International Journal of Surgery (London, England)*, 78, 185.

NYC Health Department. (2020a). COVID-19 data. https://www1.nyc.gov/site/doh/covid/covid-19-data.page.

NYC Health Department. (2020b). COVID-19 dataset. https://github.com/nychealth/coronavirus-data.

NYC MTA. (2020). Turnstile data. https://data.ny.gov/Transportation/Turnstile-Usage-Data-2020/py8k-a8wg.

Perc, M., Gorisek Miksi, N., Slavinec, M., & Stozer, A. (2020). Forecasting COVID-19. *Frontiers in Physics*, 8, 127. https://doi.org/10.3389/fphy.2020.00127. https://www.frontiersin.org/article/10.3389/fphy.2020.00127.

Petropoulos, F., & Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. *PloS One*, 15, Article e0231236.

Qin, L., Sun, Q., Wang, Y., Wu, K. F., Chen, M., Shia, B. C., & Wu, S. Y. (2020). Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. *International Journal of Environmental Research and Public Health*, 17, 2365.

Ramsundar, B., & Zadeh, R. B. (2018). *TensorFlow for deep learning: From linear regression to reinforcement learning*. O’Reilly Media, Inc.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos: Solitons & Fractals.

Schafer, R. W. (2011). What is a savitzky-golay filter? [lecture notes]. *IEEE Signal Processing Magazine*, 28, 111–117.

Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., Asfian, P., et al. (2020). Correlation between weather and covid-19 pandemic in jakarta, Indonesia. *The Science of the Total Environment*, 138436.

Wadhera, R. K., Wadhera, P., Gaba, P., Figueroa, J. F., Maddox, K. E. J., Yeh, R. W., & Shen, C. (2020). Variation in COVID-19 hospitalizations and deaths across New York City boroughs. *Journal of the American Medical Association*, 323(21), 2192–2195. https://doi.org/10.1001/jama.2020.7197

Whong, C. (2020). Taming the mta’s unruly turnstile data. https://medium.com/qri-io/taming-the-mtas-unruly-turnstile-data-c945f5f96ba0.

Yang, Zifeng, Zeng, Zhiji, Wang, Ke, Wong, Sook-San, Liang, Wenhua, Zanin, Mark, Liu, Peng, Cao, Xudong, Gao, Zhongqiang, Mai, Zhitong, Liang, Jingyi, Liu, Xiaoying, Li, Shiyue, Li, Yimin, Ye, Cheng, Guan, Weijie, Yang, Yifan, Li, Fei, Luo, Shengmei, Xie, Yuqi, Liu, Bin, Wang, Zhoulang, Zhang, Shaobo, Wang, Yaonan, Zhong, Nanhsan, & He, Jianxing (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease*, 12, 165.

Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the diamond princess cruise ship: A data-driven analysis. *International Journal of Infectious Diseases*, 93, 201–204.