

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/364324489>

Improving The Forecasting Of Covid-19 Cases Based On Embedding Re-Weighted And Commitment Features

Article in *Webology* · January 2021

CITATIONS

0

READS

2

2 authors:



Dhuha Hussein

University of Babylon

2 PUBLICATIONS 0 CITATIONS

SEE PROFILE



Eman Salih Al-Shamery

University of Babylon

54 PUBLICATIONS 125 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Disease Prediction Improvement Based on Modified Rough Set and Most Common Decision Tree [View project](#)



churn prediction [View project](#)

Improving The Forecasting Of Covid-19 Cases Based On Embedding Re-Weighted And Commitment Features

Dhuha H. Al-Jubory¹, Eman S.Al-Shamery²

¹software department_colledge of IT_University of Babylon_Babil, Iraq
dhuha.mohamedjawad@student.uobabylon.edu.iq

²software department_colledge of IT_University of Babylon_Babil, Iraq
emanalshamery@itnet.uobabylon.edu.iq

Abstract—

The outbreak of Coronavirus is the most significant global crisis since World War II. Therefore, it is imperative to keep track of the number of infected patients. In -this research, a forecasting technique was performed to forecast the impact of the pandemic in the near future. The study focuses on the daily new affirm cases from badly hit countries which are (The United Kingdom, Germany, Italy, and France).

The data was collected from the online database of the World Health Organization (WHO). We used the 'Seasonal Autoregressive Integrated Moving Average Exogenous' (SARIMAX) model, which is essentially a SARIMA with eXogenous input model to forecast new Covid-19 cases by using historical data from March 11, 2020, to March 30, 2021, to predict the sample in April 2021. Then we tried to enhance the forecasting results by reweighting the new cases feature to add importance to the latest months of the study period that the new infections depend more on them.

After that, to get more enhanced results, we associate an exogenous input to the model which is the factor of citizens' commitment to the recommendation of the World Health Organization to reduce human movements to stay as safe as possible, since the restriction of human mobility is still the only dynamic way to control the spread of the virus.

Human mobility was extracted from daily reports of driving, transit, and walking that are collected from the Apple Company official website. The last part of the study was a comparative study between the performance of the model that fitted to the original dataset and when we reweighted infected cases in addition to the commitment factor. The result shows that the model when using the dataset with the factor of people's commitment generally performs better in terms of Mean Square Error (MSE) and Root Mean Square Error (RMSE) as well as the actual and predicted new infections of the last month.

Keywords— COVID-19 pandemic, epidemic spread, forecasting, time series, SARIMAX model, human mobility, mobility trends.

I. Introduction

Since the Coronavirus was declared a pandemic, many countries around the world have been badly affected by the new disease and various measures have been taken to restrict the virus's spread. For example, the United States, Australia, India and other countries have adopted various preventive measures like the usage of masks, the adoption of stay-at-home orders, lockdowns and social segregation (ArunKumar et al., 2021). Due to the randomness and irregularity nature of the virus, there is greater instability around the choice on ideal time of disappearance of this illness. Furthermore, as the number of verified cases grows, the demand for hospital beds, particularly Intensive Care Unit (ICU) beds has increased, Therefore, for the better management of the economic, societal, cultural, and public health issues, it is important to get an estimate about what would be the daily new cases at least short term forecasting for the upcoming few weeks, ultimately also to help the healthcare systems and governments to prepare in advance for the expected number of the new infections (Yousaf et al., 2020).

Time series forecasting is an important task in a variety of sectors, including finance, economics, health, engineering, meteorology, and science. (Domingos, de Oliveira, and de Mattos Neto, 2019).

Two steps in forecasting that are based on previous data, the learning model step is the initial step consists of training the employed methodology on the training data, while the second step is to predict target attributes (forecasting step). The model represents a forecasting system because it receives and analyzes input data to produce an output called the predicted value. Predictive models are based on assumptions about the relationship between historical data (input) and future values (output). (Al-Shamery and Al-Gashamy, 2018).

Several recent research have used machine learning and deep learning approaches to forecast diseases trends based on reported time-series data in order to estimate virus spread. (Zhang et al., 2013) have used the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model and other four models to forecast typhoid fever in China. Likewise, the Auto-Regressive Integrated Moving Average (ARIMA) model have used widely in forecasting, where (He and Tao, 2018) have used the model to forecast the future and fit the time series of influenza virus among children through the year 2016, their model was able to accurately forecast positive rate of influenza disease in a short period of time in Wuhan, China, while (Duan and Zhang, 2020) applied ARIMA model to predict the daily new cases of the COVID-19 pandemic from April 27, 2020, to May 3, 2020, using South Korean and Japanese data. The dependent structure of the daily time series of newly confirmed cases was well captured by the estimated model of ARIMA. Another Autoregressive (AR) time series, called TP-SMN-AR model have been employed by (Maleki et al., 2020) based on two-piece scale mixture normal distributions to predict the recovered and confirmed cases of COVID-19 around the world.

Furthermore, other models were introduced to handle forecasting by using Regression where (Ghosal et al., 2020) have applied Linear Regression in their analysis study which aims to track trends related to the expected number of deaths of Corona-virus in India. A Logistic Regression model was also utilized by

(Almeshal et al., 2020) to estimate the spread size of Covid-19 in Kuwait, by forecasting the daily new cases, total infected cases, and the expected dates of the start and ending phase of the pandemic, besides that, simulations of (susceptible-infectious-recovered) stochastic individual contact model SIR ICM were performed to examine the impact of the simultaneous changes on the number of people who are susceptible and infected.

In this paper we focused on enhancing the forecasting results for the selected SARIMAX model, thus we compare the performance of the model on the original dataset and the updated dataset to forecast SARS-CoV-2 new confirm cases, as we access mobile data and epidemic data through Apple's daily trend report and WHO data, respectively.

The following is how this article is structured: The next section examines several related works that employ machine learning and data mining techniques to forecast Covid-19. The methodology and data that we used in our research are given in Section 3. The methodology part is followed by the results section, which includes some informative diagrams and tables relevant to our study, and the final section is the challenges and conclusion of this article.

II. Related work

In this part, we'll go through some of the work that's related to Coronavirus time series forecasting. (Ala'raj, Majdalawieh, and Nizamuddin, 2021) proposed model consists of ARIMA model and a modified SEIRD (Susceptible, Exposed, Infectious, Recovered, and Dead) dynamic model to improve the prediction ability of the classical model (SEIR). The method was tested on US statistics from the "COVID Tracking Project". They used MAE, MSE, and MLSE metrics and normalized MAE and MSE to compare prediction results, the model was able to produce accurate forecasts for up to two months in advance. (Khan and Gupta, 2020) used statistics from the Ministry of Health and Family Welfare for India from the 31st of January through the 25th of March, 2020, for the next 50 day predictions, an Auto-Regressive Integrated Moving Average model was used and verified with data collected from March 26, 2020, to April 4, 2020, the authors constructed a nonlinear autoregressive neural network to compare prediction accuracy. The coefficient of determination and RMSE was used to calculate the performance evaluation. According to the report, the country will see exponential and rapid growth in the number of cases in the near future. (Arun Kumar et al., 2021) conducted a study to anticipate Covid-19 epidemiological trends using ARIMA and SARIMA projections for the top-16 nations (USA, Brazil, Russia, India, Peru, Russia, Chile Mexico, the UK, Iran, South Africa, Spain, Pakistan, Italy, Saudi Arabia, and Turkey), the data were obtained from the COVID-19 database, which is available at John Hopkins University. They utilized the ARIMA model to pick the initial model parameter combination and then search for the optimized model parameters based on the best fit between the predicted and test data. To test the model's dependability in generating a 60-day prediction of (confirmed, cumulative, recovered, and deaths cases) the Partial Auto-correlation Function (PACF), Autocorrelation Function (ACF), Bayes Information Criterion (BIC), and Akaike Information Criterion (AIC) were employed, also the best model selection was based on evaluation metrics (MAE), (MSE), (RMSE), and (MAPE). The study's findings demonstrate that SARIMA model predictions are more accurate than ARIMA model predictions. The study explained how the nations listed above should plan their healthcare strategies in light of the current epidemics.

Authors of (Alzahrani, Aljamaan and Al-Fakih, 2020) employed four models to forecast the Coronavirus daily new infections in Saudi Arabia for one month ahead which are (AR, MA, ARMA and ARIMA) the data were collected from the Saudi Ministry of Health official website from March 2, 2020 to April 20, 2020 across 31 influenced cities. The data were splitting into training and testing datasets in the ratio of 70:30. According to their findings ARIMA (2,1,1) model that achieves the smallest RMSE, RMSRE, MAPE values, and highest R2 value performed better than other models in forecasting the daily infections in the kingdom.

In another study (Aditya Satrio et al., 2021) amid to utilize Machine Learning model in attempt to estimate the disease trend in Indonesia within a 30-day timeframe from April 22, 2020 and to May 21, 2020, using dataset obtained from the Kaggle website containing the confirmed, recovered and deaths cases, Facebook's Prophet (FB) and ARIMA models was introduced to compare their performance. They used (R2), (MSE), (MAE) and (MFE) as an evaluation metrics. The results show that the two models were quite inaccurate in handling the forecasting process

III. Methodology

The methodology diagram is shown in fig.1. This study focuses on countries with high rates of Covid-19 pandemic spread to generate a 30-day forecast of the new infections of the disease based on the SARIMAX model fitted to the data.

A. The Data

Covid-19 pandemic data in this study were downloaded from the WHO official website (WHO, 2020-2021) which is a completely open-access database that provides daily data on newly confirmed cases for all countries. Because the start date of Coronavirus disease varies from country to country, we considered the date where the World Health Organization announced the disease as a worldwide epidemic as the start date for our work, thus we used the Covid-19 new cases data from March 11, 2020, till the end of March 30, 2021, for countries with high infection rates to foresee the newly infected people with the disease in April 2021.

Mobility data that we used is provided by Apple Company (Apple, 2020-2021) which consists of daily human movement trends of walking, driving, and public transportation. These reports reflect route requests on Apple Maps. There is existing privacy, Apple has no search history for individual movements.

The data included in our work are:

- The original dataset (OD) comes with the following features:
 - Full date for the study period.
 - Number of people diagnosed with the disease each day.
- The dataset after reweighted infected cases (RD) contains:
 - Full date for the study period.
 - Number of people diagnosed with the disease each day after adding weights to increasing the importance of the latest months.
- The dataset with additional commitment factor (CD) consists of:

- Full date for the study period.
- Daily reweighed infected cases.
- People’s commitment to the movement’s restrictions according to:
 - Driving
 - Transit
 - Walking

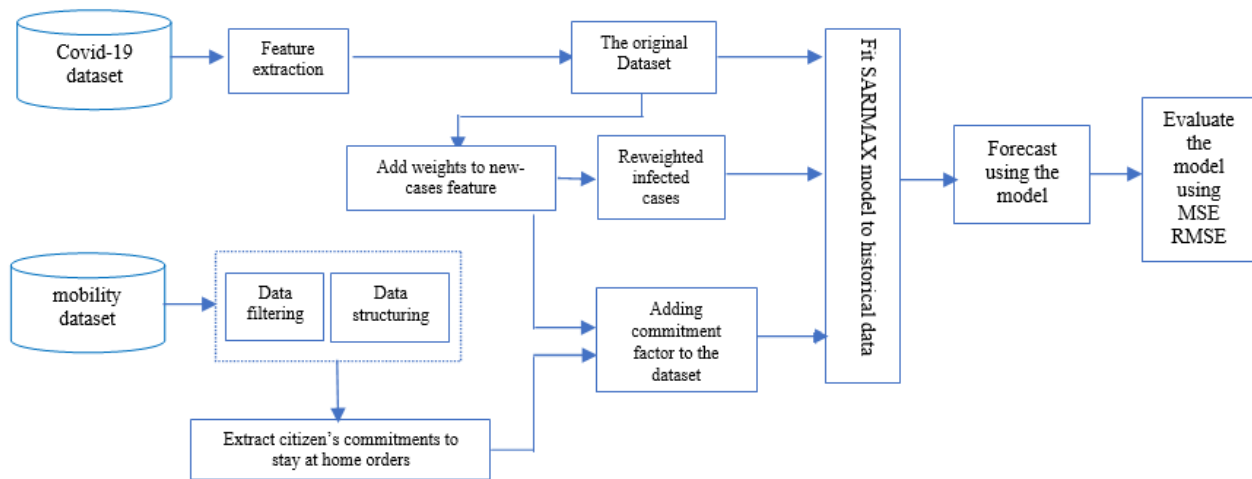


Fig.1 The methodology diagram

B. Data Analysis

This research carried out the following steps in data analysis:

- Feature Extraction (FE): to reproduce the right information to be forecasted, the outbreak data requires exceptions for the cases of recovery and death because the primary interest of our work is just the new infections of the disease.
- Weight normalization: we used statistical weight which is an amount given to increase or decrease the importance of an item to the new cases feature according to the date feature to add more importance to the recent months. This step is followed by normalization, which is a scaling technique helpful for forecasting and prediction (Gopal, Patro, and Kumar Sahu, 2015), There are several types of data normalization including Z-score and min-max, here we used the min-max where the weight values are normalized in the range of [0,1] using the equation 1. (Jain, Shukla and Wadhvani, 2018)

$$\hat{Z} = \frac{Z - \min_A}{\max_A - \min_A} (new_max_A - new_min_A) + new_min_A \tag{1}$$

Where: \max_A is the maximum value of any feature.

\min_A is the minimum value of any feature.

new_max_A and new_min_A represent the maximum and minimum interval of values.

\hat{Z} and Z represent the new feature value and old feature value, respectively.

After that, we multiply the new cases feature by the normalized weights to generate a reweighted infection case of the disease to use it with the SARIMAX model to compare the results of forecasting when using the original dataset and the updated ones.

- Data filtering: Our filtering process involves selecting a specific region of interest from Apple's vast mobile trends which consist of data from most regions and countries in the world. We also needed to aggregate the mobility data for all regions within a country to be compatible with WHO's data that contains countries' data.
- Data structuring: Organizing our data revolves around Manage some gaps in Apple's data by taking the average of the two dates surrounding the missing one to obtain complete data that clearly explains daily trends. Then, extracting people's commitment to stay at home orders through several statistical steps on mobility data using Microsoft Excel Spreadsheets. Then, it is followed by the collecting of the epidemic and the new citizen's commitment data together in a new CSV file format for each country as preparation for the forecasting step with the SARIMAX model.

C. SARIMAX model

This model is a multivariate variant of the SARIMA model with eXogenous factors, which is used to improve the performance of time series forecasting. Time-series forecasting and analysis have long been a hot topic in academia. Time series data is a sequence of values, each of which includes a timestamp, and may be classified into two types: stationary and nonstationary data. A stationary time-series data has statistical properties such as variance, mean, autocorrelation, and so on that remain constant throughout time, whereas seasonality time-series data has patterns. (Arun Kumar et al, 2021)

The model referred to as (p, d, q) (P, D, Q) Whereas p, d, q, and P, D, Q is nonnegative numbers that represent the polynomial order of the three parts of the non-seasonal and seasonal components of the model, which are autoregressive (AR), integrated (I), and moving average (MA), respectively. Where p is the order of the AR (Auto-Regressive) term and denotes the number of previous data points that should be used as predictors, q is the order of the MA (Moving Average) term and denotes the number of lagged errors in the forecast that must be entered into the model, and d is the number of differentiations required for the time series to be stationary. (Khan and Gupta, 2020) c.

the Seasonal ARIMA with eXogenous factor is described mathematically as:(L, 2016)

$$\varphi_p(B) \Phi_P(B^s) \nabla^d \nabla_s^D y_t = \beta_k x_{k,t}' + \theta_q(B) \Theta_Q(B^s) \varepsilon_t$$

Whereas,

$\varphi_p(B)$ is the regular Auto-Regressive polynomial of order p.

$\Phi_P(B^s)$ is the seasonal Auto-Regressive polynomial of order P

$\theta_q(B)$ is the regular Auto-Regressive polynomial of order p

$\Theta_Q(B^s)$ is the seasonal Moving-Average polynomial of order Q

∇_s^D is the operator of seasonal differentiating that eliminates the seasonal non-stationarity.

∇^d is the operator of differentiating that eliminates non-seasonal.

B is the operator of the backshift, which shifts the y_t observation by one point in time.

ε_t follows a white noise process.

s is the seasonal period.

$x_{k,t}'$ is the vector containing the kth explanatory input variables at time t.

β_k is the coefficient of the k-th exogenous input variable value.

Our work steps to implement the SARIMAX model:

- The first step of building the model is to load the dataset.

- Preprocessing the dates feature to ensure that the values are understood by the programming language which we used.
- Check for stationarity of the data as shown in fig.2.
- Parameter selection and fit SARIMAX model.

For the SARIMAX time series model we take values [0, 1] that's mean $p = [0,1]$, $d = [0,1]$, $q = [0,1]$, and take monthly periodicity thus $s = 12$ as well as, we set the endogenous factor to "new_cases" which indicates output (y) first we just learn the intercept thus x is none. After that we reweighted the new infections feature to enhance the results of predictions then we employed the exogenous factor which stands for citizen's commitment to stay at home recommendations (x).

The selection of the seasonal (P, D, Q) and nonseasonal (p, d, q) parameters was based mainly on the score estimator for regression time series models (Akaike information criterion (AIC)) that shows the error in prediction vs actual, less the error the better is the model also we select monthly periodicity to forecast.

equ.2 shows the formula of AIC. (Zhang et al., 2013)

$$AIC = -2 \ln L + 2K \quad (2)$$

Here: L represents the function of likelihood.

K is the number of free parameters ($k = p + q + P + Q$).

n is the residuals number that can be calculated for the time series.

- Validating forecasts to figure out how accurate our model is, we compare predicted cases to the real number of cases of the time series, and we set forecasts to start at 2020-03-11 to the end of the data as shown in fig.2, fig.3, fig.4, and fig.5 for each country.
- Model evaluation measures used in this study are MSE and RMSE to compare the performance of the model, these measures are defined in eq.3 and eq.4 respectively, (Kırbaş et al., 2020)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

Where \hat{y}_i is the model expected value and y_i is the actual value.

- Producing and visualizing forecasts as shown in fig.7
- Compute the MSE and RMSE for test data as shown in table I.

IV. THE RESULTS

We tried several time series forecasting techniques before settling on the SARIMAX model, which had the smallest error between actual and predicted new infections as well as the outcomes of evaluation metrics. We then improved its results by adding other features. The model was fitted to the new infections of the outbreak data from March 11, 2020, to March 30, 2021, and tested by predicting the new affirm cases of the disease for one month ahead for (The United Kingdom, Germany, Italy, and France). The SARIMAX model that we produced was employed on raw data of the new daily cases for the mentioned countries. After that, we used reweighted new cases, then we added people's commitment to public recommendations of staying at home as exogenous factors to the model to enhance the prediction results. Table I shows a comparative study among them, as we

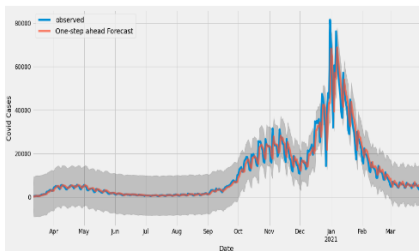
can see the performance of the model with the reweighted infected cases and the people's commitment factor was able to overcome the performance on the original dataset in terms of MSE and RMSE for the test data. Thus, these eXogenous features are of significance to the forecasting process where the P-value for people's commitments according to driving, transit and walking is greater than alpha (threshold = 0.05) as shown in Table II also the actual new cases and SARIMX predictions were presented in Table III and Table IV for all the countries involved in our study.

Table I. P-values for people's commitment according to driving, transit, and walking

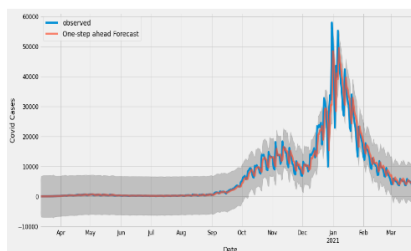
| Country | P-value | | |
|---------|---------|---------|---------|
| | Driving | Transit | Walking |
| UK | 0.232 | 0.001 | 0.480 |
| Germany | 0.757 | 0.002 | 0.257 |
| France | 0.761 | 0.311 | 0.756 |
| Italy | 0.938 | 0.393 | 0.921 |

Table II. a comparison between evaluation metrics used in the study

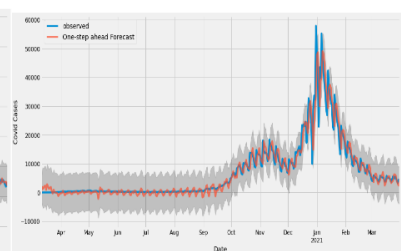
| Countries | The original dataset (OD) | | The dataset after reweighted infected cases (RD) | | The dataset with additional commitment factor (CD) | |
|--------------------|---------------------------|--------------|--|--------------|--|--------------|
| | MSE | RMSE | MSE | RMSE | MSE | RMSE |
| The United Kingdom | 31345 30.82 | 1770 .46 | 312388 6.04 | 1767 .45 | 304793 5.13 | 1745.8 3 |
| Germany | 57315 609.41 | 7570 .71 | 529520 95.52 | 7276 .82 | 506260 30.83 | 7115.2 |
| Italy | 34031 843.27 | 5833 .68 | 283488 85.3 | 5324 .37 | 185105 91.64 | 4302.3 9 |
| France | 24259 0731.6 4 | 1557 5.32 | 195496 892.81 | 1398 2.02 | 194156 625.12 | 13934. 01 |



(a)

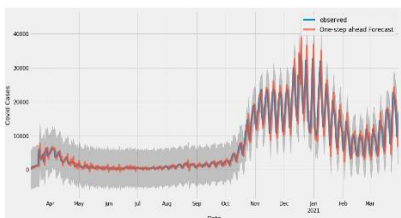


(b)

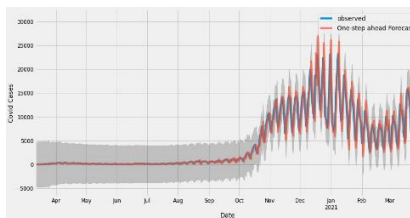


(c)

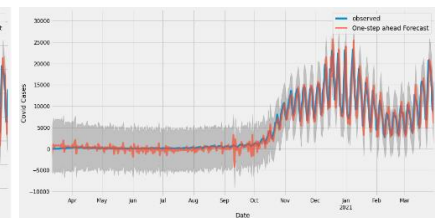
Fig.3 validate the model on training data for UK. (a) OD, (b) RD, and (c) CD



(a)



(b)



(c)

Fig.4 validates the model on training data for Germany. (a) OD, (b) RD and (c) CD

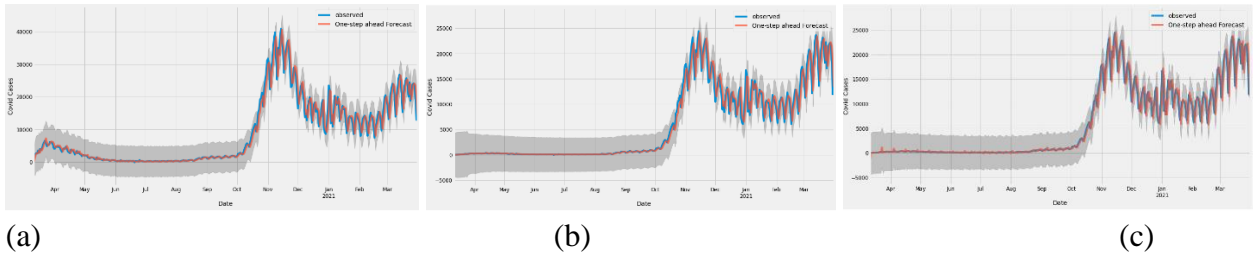


Fig.5 validates the model on training data for Italy. (a) OD, (b) RD and (c) CD

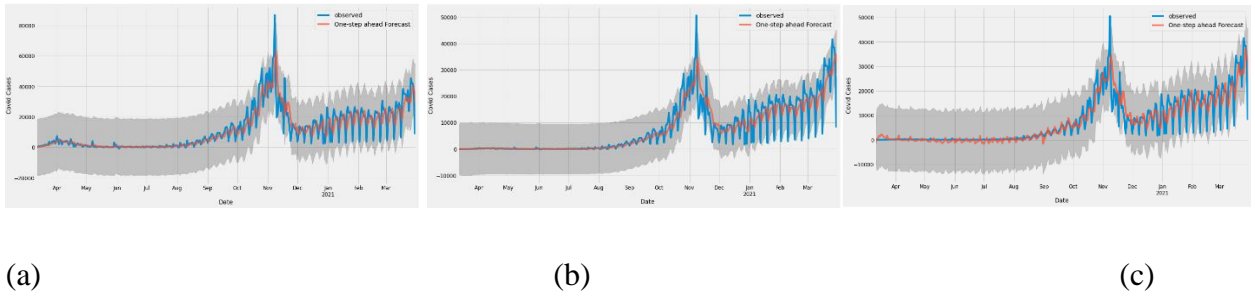


Fig.6 validates the model on training data for France. (a) OD, (b) RD and (c) CD

Table III. Actual and SARIMAX predictions for UK and Germany

| The Date | The United Kingdom | | | | | | Germany | | | | | |
|-----------|---------------------------|-----------|--|-----------|--|-----------|---------------------------|-----------|--|-----------|--|-----------|
| | The original dataset (OD) | | The dataset after reweighted infected cases (RD) | | The dataset with additional commitment factor (CD) | | The original dataset (OD) | | The dataset after reweighted infected cases (RD) | | The dataset with additional commitment factor (CD) | |
| | Actual | predicted | Actual | predicted | Actual | predicted | actual | predicted | actual | predicted | actual | predicted |
| 4/1/2021 | 358 | 3656 | 358 | 3656 | 333 | 4225 | 243 | 20116 | 226 | 18873 | 226 | 19274 |
| 4/2/2021 | 374 | 3428 | 374 | 3428 | 349 | 4279 | 218 | 21166 | 204 | 19589 | 204 | 19786 |
| 4/3/2021 | 346 | 4125 | 346 | 4125 | 323 | 5590 | 181 | 22465 | 169 | 20970 | 169 | 20372 |
| 4/4/2021 | 258 | 4281 | 258 | 4281 | 242 | 5327 | 121 | 22429 | 114 | 20963 | 114 | 20708 |
| 4/5/2021 | 233 | 3813 | 233 | 3813 | 219 | 3584 | 849 | 21266 | 798 | 19568 | 798 | 15922 |
| 4/6/2021 | 223 | 4811 | 223 | 4811 | 210 | 3921 | 688 | 20153 | 648 | 18265 | 648 | 12893 |
| 4/7/2021 | 254 | 4862 | 254 | 4862 | 240 | 4164 | 967 | 19727 | 914 | 18041 | 914 | 11569 |
| 4/8/2021 | 301 | 3832 | 301 | 3832 | 285 | 3695 | 204 | 19947 | 193 | 18131 | 193 | 11178 |
| 4/9/2021 | 285 | 3685 | 285 | 3685 | 271 | 4701 | 254 | 20461 | 241 | 18808 | 241 | 11173 |
| 4/10/2021 | 245 | 4471 | 245 | 4471 | 233 | 5978 | 240 | 22318 | 229 | 20750 | 229 | 13842 |
| 4/11/2021 | 845 | 4404 | 845 | 4404 | 806 | 3158 | 178 | 22520 | 170 | 20935 | 170 | 12295 |
| 4/12/2021 | 173 | 4490 | 173 | 4490 | 165 | 1143 | 132 | 21038 | 126 | 19242 | 126 | 9416 |
| 4/13/2021 | 356 | 3833 | 356 | 3833 | 342 | 1019 | 108 | 20144 | 103 | 18414 | 103 | 9459 |
| 4/14/2021 | 247 | 3464 | 247 | 3464 | 237 | 1188 | 216 | 20289 | 208 | 18443 | 208 | 9110 |
| 4/15/2021 | 249 | 4056 | 249 | 4056 | 240 | 1676 | 294 | 20057 | 283 | 18128 | 283 | 8799 |

| | | | | | | | | | | | | |
|--------|-----|------|-----|------|-----|------|-----|-------|-----|-------|-----|-------|
| 021 | 1 | | 1 | | 1 | | 26 | | 62 | | 62 | |
| 4/16/2 | 267 | 4220 | 267 | 4220 | 258 | 2176 | 258 | 19980 | 249 | 18076 | 249 | 9109 |
| 021 | 1 | | 1 | | 1 | | 31 | | 60 | | 60 | |
| 4/17/2 | 275 | 3789 | 275 | 3789 | 267 | 2628 | 238 | 20783 | 230 | 18924 | 230 | 11123 |
| 021 | 6 | | 6 | | 0 | | 04 | | 58 | | 58 | |
| 4/18/2 | 220 | 4721 | 220 | 4721 | 214 | 2595 | 191 | 21318 | 186 | 19514 | 186 | 10350 |
| 021 | 6 | | 6 | | 2 | | 85 | | 30 | | 30 | |
| 4/19/2 | 188 | 4793 | 188 | 4793 | 183 | 747 | 0 | 20755 | 0 | 19111 | 0 | 9157 |
| 021 | 2 | | 2 | | 2 | | | | | | | |
| 4/20/2 | 296 | 3775 | 296 | 3775 | 289 | -139 | 210 | 20702 | 205 | 18940 | 205 | 8952 |
| 021 | 3 | | 3 | | 2 | | 46 | | 39 | | 39 | |
| 4/21/2 | 252 | 3601 | 252 | 3601 | 246 | 340 | 248 | 20427 | 243 | 18705 | 243 | 8779 |
| 021 | 4 | | 4 | | 9 | | 84 | | 44 | | 44 | |
| 4/22/2 | 239 | 4334 | 239 | 4334 | 235 | 1267 | 295 | 20387 | 289 | 18492 | 289 | 8775 |
| 021 | 6 | | 6 | | 0 | | 18 | | 49 | | 49 | |
| 4/23/2 | 272 | 4309 | 272 | 4309 | 268 | 1376 | 275 | 20480 | 270 | 18563 | 270 | 9010 |
| 021 | 8 | | 8 | | 2 | | 43 | | 78 | | 78 | |
| 4/24/2 | 267 | 4367 | 267 | 4367 | 263 | 2411 | 233 | 20859 | 230 | 19032 | 230 | 10591 |
| 021 | 8 | | 8 | | 9 | | 92 | | 54 | | 54 | |
| 4/25/2 | 206 | 3720 | 206 | 3720 | 203 | 1109 | 187 | 20784 | 185 | 19172 | 185 | 9519 |
| 021 | 1 | | 1 | | 6 | | 73 | | 47 | | 47 | |
| 4/26/2 | 171 | 3356 | 171 | 3356 | 169 | -429 | 119 | 21117 | 117 | 19372 | 117 | 8883 |
| 021 | 2 | | 2 | | 5 | | 07 | | 92 | | 92 | |
| 4/27/2 | 206 | 3951 | 206 | 3951 | 204 | 85 | 109 | 21202 | 108 | 19480 | 108 | 8633 |
| 021 | 4 | | 4 | | 9 | | 76 | | 97 | | 97 | |
| 4/28/2 | 268 | 4114 | 268 | 4114 | 267 | 331 | 222 | 21134 | 221 | 19439 | 221 | 8497 |
| 021 | 5 | | 5 | | 2 | | 31 | | 24 | | 24 | |
| 4/29/2 | 216 | 3682 | 216 | 3682 | 216 | 653 | 247 | 21529 | 246 | 19728 | 246 | 9419 |
| 021 | 6 | | 6 | | 1 | | 36 | | 76 | | 76 | |
| 4/30/2 | 244 | 4616 | 244 | 4616 | 244 | 1371 | 243 | 21723 | 243 | 19846 | 243 | 10345 |
| 021 | 5 | | 5 | | 5 | | 29 | | 29 | | 29 | |

Table IV. Actual and SARIMAX predictions for Italy and France

| The Date | Italy | | | France | | |
|----------|---------------------------|--|--|---------------------------|--|--|
| | The original dataset (OD) | The dataset after reweighted infected cases (RD) | The dataset with additional commitment factor (CD) | The original dataset (OD) | The dataset after reweighted infected cases (RD) | The dataset with additional commitment factor (CD) |
| | | | | | | |

| | Act ual | predic ted | Act ual | predic ted | Act ual | predic ted | actu al | predic ted | actu al | predic ted | actu al | predic ted |
|---------------|------------|---------------|------------|---------------|------------|---------------|------------|---------------|------------|---------------|------------|---------------|
| 4/1/20 21 | 238 87 | 16817 | 222 18 | 15640 | 222 18 | 15571 | 588 28 | 28114 | 547 17 | 28481 | 547 17 | 26980 |
| 4/2/20 21 | 221 84 | 16684 | 206 87 | 15715 | 206 87 | 14966 | 506 59 | 26993 | 472 41 | 27813 | 472 41 | 26223 |
| 4/3/20 21 | 219 17 | 17386 | 204 91 | 16517 | 204 91 | 14798 | 455 20 | 26864 | 425 58 | 27593 | 425 58 | 28664 |
| 4/4/20 21 | 212 47 | 17653 | 199 16 | 16604 | 199 16 | 15290 | 139 17 | 26464 | 130 45 | 28214 | 130 45 | 29491 |
| 4/5/20 21 | 180 17 | 18153 | 169 32 | 17208 | 169 32 | 15185 | 667 94 | 28198 | 627 70 | 27943 | 627 70 | 29706 |
| 4/6/20 21 | 106 80 | 18054 | 100 62 | 17084 | 100 62 | 15584 | 123 35 | 26359 | 116 22 | 28072 | 116 22 | 27811 |
| 4/7/20 21 | 776 3 | 18084 | 733 3 | 17022 | 733 3 | 15790 | 486 46 | 24734 | 459 50 | 28501 | 459 50 | 25933 |
| 4/8/20 21 | 136 86 | 17924 | 129 60 | 16889 | 129 60 | 15690 | 307 85 | 25747 | 291 53 | 28786 | 291 53 | 27262 |
| 4/9/20 21 | 172 09 | 17575 | 163 38 | 16624 | 163 38 | 15402 | 236 38 | 24809 | 224 42 | 28358 | 224 42 | 26845 |
| 4/10/2 021 | 189 24 | 18230 | 180 12 | 17362 | 180 12 | 15594 | 399 63 | 25275 | 380 37 | 28493 | 380 37 | 28906 |
| 4/11/2 021 | 175 51 | 18900 | 167 47 | 18160 | 167 47 | 15581 | 432 83 | 30361 | 413 01 | 28787 | 413 01 | 34888 |
| 4/12/2 021 | 157 37 | 19030 | 150 54 | 18092 | 150 54 | 15908 | 348 95 | 27415 | 333 81 | 29317 | 333 81 | 29562 |
| 4/13/2 021 | 978 0 | 19199 | 937 9 | 18218 | 937 9 | 16128 | 755 6 | 27488 | 724 6 | 29416 | 724 6 | 30267 |
| 4/14/2 021 | 134 39 | 18533 | 129 21 | 17696 | 129 21 | 15564 | 389 56 | 26492 | 374 54 | 28748 | 374 54 | 29378 |
| 4/15/2 021 | 161 60 | 18551 | 155 76 | 17709 | 155 76 | 15701 | 433 54 | 25779 | 417 87 | 28528 | 417 87 | 29169 |
| 4/16/2 021 | 169 63 | 19102 | 163 91 | 18180 | 163 91 | 16176 | 379 36 | 26522 | 366 56 | 29149 | 366 56 | 30526 |
| 4/17/2 021 | 159 23 | 18887 | 154 24 | 17937 | 154 24 | 16047 | 363 60 | 25900 | 352 21 | 28878 | 352 21 | 30755 |
| 4/18/2 021 | 153 64 | 18999 | 149 20 | 18083 | 149 20 | 16093 | 342 18 | 25828 | 332 29 | 29007 | 332 29 | 30751 |
| 4/19/2 021 | 126 88 | 19239 | 123 52 | 18297 | 123 52 | 16374 | 293 44 | 26297 | 285 66 | 29436 | 285 66 | 29924 |

| | | | | | | | | | | | | |
|---------------|-----------|-------|-----------|-------|-----------|-------|-----------|-------|-----------|-------|-----------|-------|
| 4/20/2 021 | 886 3 | 19116 | 864 9 | 18217 | 864 9 | 16274 | 663 6 | 26635 | 647 6 | 29721 | 647 6 | 30418 |
| 4/21/2 021 | 120 69 | 18766 | 118 07 | 17953 | 118 07 | 15997 | 425 53 | 25847 | 416 30 | 29293 | 416 30 | 29684 |
| 4/22/2 021 | 138 36 | 18922 | 135 69 | 18081 | 135 69 | 16177 | 343 68 | 25526 | 337 05 | 29428 | 337 05 | 29736 |
| 4/23/2 021 | 160 46 | 18846 | 157 75 | 17962 | 157 75 | 16232 | 340 34 | 25599 | 334 60 | 29722 | 334 60 | 29543 |
| 4/24/2 021 | 147 58 | 19277 | 145 45 | 18289 | 145 45 | 16605 | 321 92 | 26238 | 317 27 | 30252 | 317 27 | 31983 |
| 4/25/2 021 | 138 14 | 19504 | 136 48 | 18495 | 136 48 | 16873 | 325 00 | 26092 | 321 08 | 30351 | 321 08 | 31571 |
| 4/26/2 021 | 131 57 | 18890 | 130 30 | 18050 | 130 30 | 16267 | 228 49 | 25093 | 226 29 | 29683 | 226 29 | 29670 |
| 4/27/2 021 | 844 0 | 18977 | 837 9 | 18166 | 837 9 | 16359 | 491 4 | 24483 | 487 8 | 29463 | 487 8 | 29386 |
| 4/28/2 021 | 103 98 | 19499 | 103 48 | 18586 | 103 48 | 16857 | 299 80 | 25067 | 298 36 | 30083 | 298 36 | 29986 |
| 4/29/2 021 | 133 82 | 19356 | 133 50 | 18454 | 133 50 | 16652 | 314 44 | 24818 | 313 68 | 29813 | 313 68 | 29681 |
| 4/30/2 021 | 143 14 | 19446 | 143 14 | 18565 | 143 14 | 17135 | 263 73 | 24492 | 263 73 | 29942 | 263 73 | 29791 |

V. Conclusion and discussion

The purpose of this study was to compare the SARIMAX statistical model that fitted to different datasets to forecast COVID-19 positive confirmed cases for 30 days. The model is country-specific, and the optimum model parameters were manually selected. Based on the results of this study, it was found that the model with less forecasting error is the model that fitted to the data after reweighting the new infections and adding citizens' commitment to the recommendations of the Ministry of Health as an exogenous factor. The pandemic is still progressing and applying strict restrictions policies such as stopping business, closing schools, and travel, etc., and using vaccines, can potentially reduce the effects of the disease spread.

I. Challenges

The actual number of infections is the key issue in determining the spread of the COVID-19 epidemic.

The number of new daily confirm cases available for mathematical modeling of the dataset are those confirmed by tests. There may be many infected people who never get tested, making affirmed cases only a small fraction of the actual number of infections. Also, despite the new infections are seasonal and time-series data but with very fluctuation patterns since the infections

go up one day and go down significantly in the next day thus making accurate predictions is a hard and challenging process.

II. The requirements

The requirements used to achieve this work were by the use of Intel® Core™ i5 processors, Windows 10, using Microsoft Excel Worksheet to analyze the data, and Python 3.8 (64 bit) programming language to implement the model.

References

- Aditya Satrio, C. B. et al. (2021) 'Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET', in *Procedia Computer Science*. Elsevier B.V., pp. 524–532.
- Al-Shamery, E. S. and Al -Gashamy, H. A. (2018) Enhanced Evolutionary Sequential Minimal Optimization Model for Inflation Prediction, *International Journal of Engineering & Technology*. Available at: www.sciencepubco.com/index.php/IJET.
- JALIL, A. T., DILFY, S. H., KAREVSKIY, A., & NAJAH, N. (2020). Viral Hepatitis in Dhi-Qar Province: Demographics and Hematological Characteristics of Patients. *International Journal of Pharmaceutical Research*, 12(1).
- Ala'raj, M., Majdalawieh, M. and Nizamuddin, N. (2021) 'Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections', *Infectious Disease Modelling*, 6, pp. 98–111.
- Dilfy, S. H., Hanawi, M. J., Al-bideri, A. W., & Jalil, A. T. (2020). Determination of Chemical Composition of Cultivated Mushrooms in Iraq with Spectrophotometrically and High Performance Liquid Chromatographic. *Journal of Green Engineering*, 10, 6200-6216.
- Almeshal, A. M. et al. (2020) 'Forecasting the spread of COVID-19 in kuwait using compartmental and logistic regression models', *Applied Sciences*, 10(10).
- Jalil, A. T., Al-Khafaji, A. H. D., Karevskiy, A., Dilfy, S. H., & Hanan, Z. K. (2021). Polymerase chain reaction technique for molecular detection of HPV16 infections among women with cervical cancer in Dhi-Qar Province. *Materials Today: Proceedings*.
- Alzahrani, S. I., Aljamaan, I. A. and Al-Fakih, E. A. (2020) 'Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions', *Journal of Infection and Public Health*, 13(7), pp. 914–919.
- Marofi, F., F. Abdul-Rasheed, O., Sulaiman Rahman, H., Setia Budi, H., Jalil, A. T., Valerievich Yumashev, A., ... & Jarahian, M. (2021). CAR-NK cell in cancer immunotherapy; A promising frontier. *Cancer Science*.
- Apple Company Report on data of the mobility trends, 2020-2021. Available at: <https://www.apple.com/covid19/mobility> [Online; accessed: 1 March 2021].
- ArunKumar, K. E. et al. (2021) 'Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive

Integrated Moving Average (SARIMA)', *Applied Soft Computing*, 103.

Saleh, M. M., Jalil, A. T., Abdulkereem, R. A., & Suleiman, A. A. Evaluation of Immunoglobulins, CD4/CD8 T Lymphocyte Ratio and Interleukin-6 in COVID-19 Patients. *TURKISH JOURNAL of IMMUNOLOGY*, 8(3), 129-134.

Domingos, D. S., de Oliveira, J. F. L. and de Mattos Neto, P. S. G. (2019) 'An intelligent hybridization of ARIMA with machine learning models for time series forecasting', *Knowledge-Based Systems*, 175, pp. 72–86.

Widjaja, G., Jalil, A. T., Rahman, H. S., Abdelbasset, W. K., Bokov, D. O., Suksatan, W., ... & Ahmadi, M. (2021). Humoral Immune mechanisms involved in protective and pathological immunity during COVID-19. *Human Immunology*.

Duan, X. and Zhang, X. (2020) 'ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data', *Data in Brief*, 31.

Turki Jalil, A., Emad Al. Qurabiy, H., Hussain Dilfy, S., Oudah Meza, S., Aravindhan, S., M. Kadhim, M., M. Aljeboree, A. (2021). CuO/ZrO₂ Nanocomposites: Facile Synthesis, Characterization and Photocatalytic Degradation of Tetracycline Antibiotic. *Journal of Nanostructures*, (), -.

Ghosal, S. et al. (2020) 'Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks', *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, 14(4), pp. 311–315.

Gopal, S., Patro, K. and Kumar Sahu, K. (no date) Normalization: A Preprocessing Stage. Available at: www.kiplinger.com.

He, Z. and Tao, H. (2018) 'Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study', *International Journal of Infectious Diseases*, 74, pp. 61–70.

I., S. (2016) 'Comparison of SARIMAX, SARIMA, Modified SARIMA and ANN-based Models for Short-Term PV Generation Forecasting', 2016 IEEE International Energy Conference, ENERGYCON 2016. IEEE.

Jain, S., Shukla, S. and Wadhvani, R. (2018) 'Dynamic selection of normalization techniques using data complexity measures', *Expert Systems with Applications*, 106, pp. 252–262.

Jalil, A.T., Kadhum, W.R., Faryad Khan, M.U. et al. Cancer stages and demographical study of HPV16 in gene L2 isolated from cervical cancer in Dhi-Qar province, Iraq. *Appl Nanosci* (2021). <https://doi.org/10.1007/s13204-021-01947-9>

Khan, F. M. and Gupta, R. (2020) 'ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India', *Journal of Safety Science and Resilience*, 1(1), pp. 12–18.

Kırbaşı, İ. et al. (2020) 'Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches', *Chaos, Solitons and Fractals*, 138.

Maleki, M. et al. (2020) 'Time series modelling to forecast the confirmed and recovered cases of COVID-19', *Travel Medicine and Infectious Disease*, 37.

World Health Organization (2020-2021) WHO Coronavirus (COVID-19) Dashboard. Available at:

<https://covid19.who.int/>. [Online; accessed 1 March 2021].

Yousaf, M. et al. (2020) 'Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan', *Chaos, Solitons and Fractals*, 138.

Zhang, X. et al. (2013) 'Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China', *PLoS ONE*, 8(5).