Webology (ISSN: 1735-188X)
Volume 18, Number 1, 2021

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

Dhuha H. Al-Jubory[1] , Eman S.Al-Shamery[2]

[1]software department colledge of IT University of Babylon Babil, Iraq
dhuha.mohamedjawad@student.uobabylon.edu.iq

[2]software 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 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 orginal 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 employs 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 artical .

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 un 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

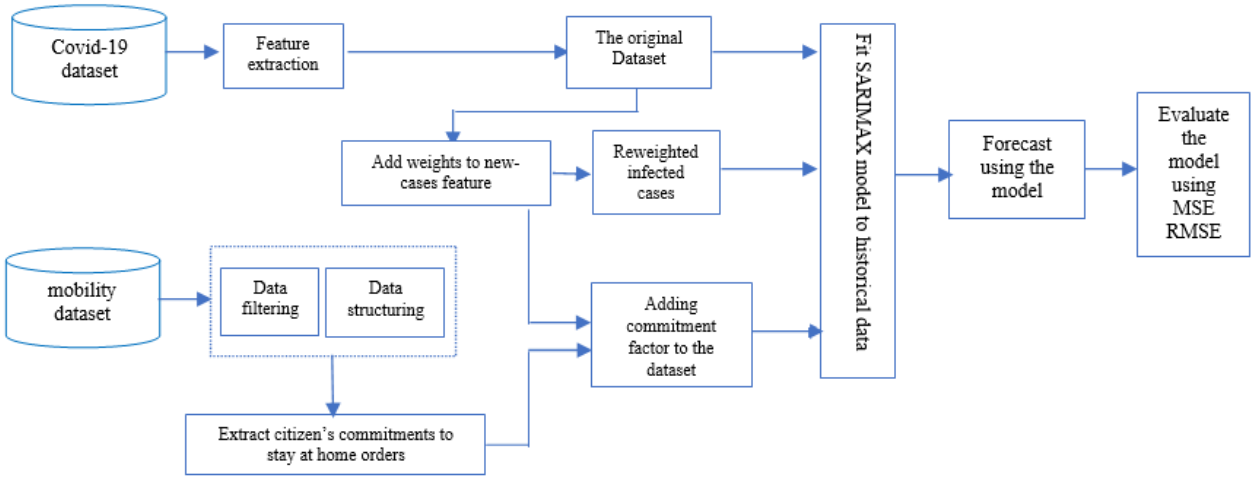


Fig.1 The methodology diagram

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)

$$\hat{Z} = \frac{z - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (1)$$

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:(I., 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 yt 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.
βk 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)

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

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)

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

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

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

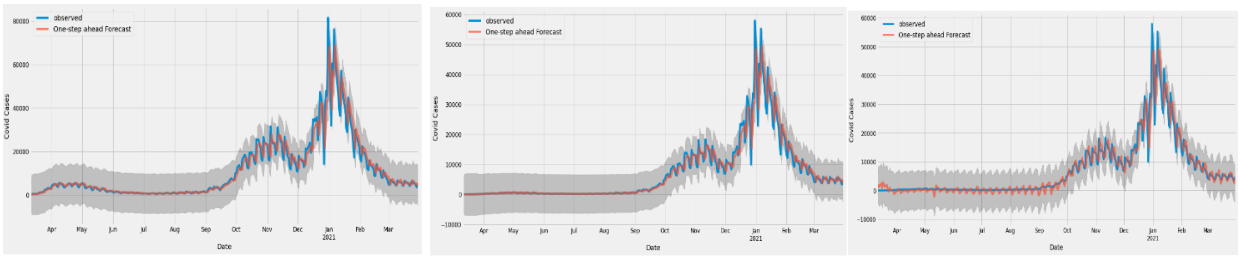
can see the performance of the model with the reweighted infected cases and the people's commitment factor was able to overcome the performance on the original dataset in terms of MSE and RMSE for the test data. Thus, these eXogenous features are of significance to the forecasting process where the P-value 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

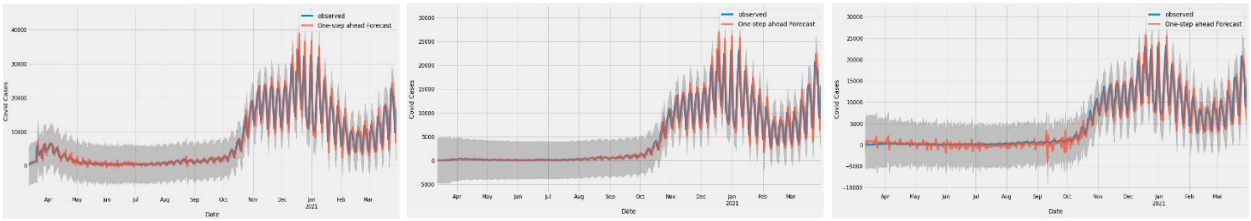
Table II. a comparison between evaluation metrics used in the study

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	3134530.82	1770.46	3123886.04	1767.45	3047935.13	1745.83
Germany	5731560 9.41	7570.71	5295209 5.52	7276.82	5062603 0.83	7115.2
Italy	3403184 3.27	5833.68	2834888 5.3	5324.37	1851059 1.64	4302.39
France	2425907 31.64	15575.32	1954968 92.81	13982.02	1941566 25.12	13934.01



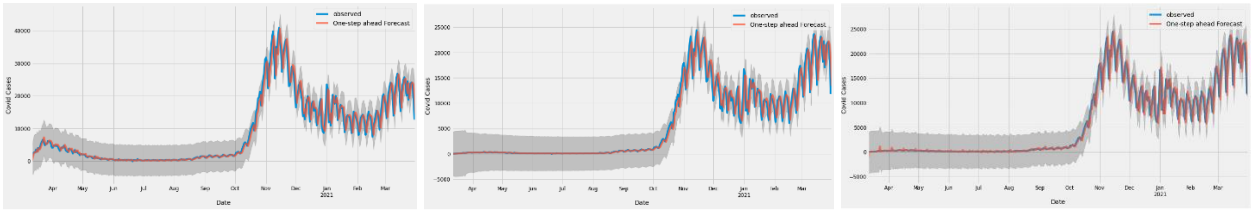
(a) (b) (c)

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



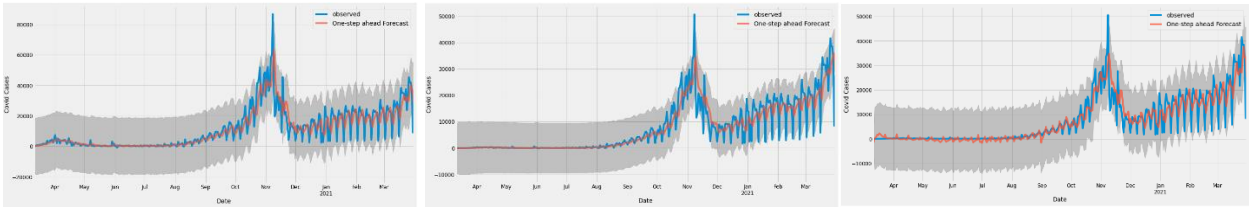
(a) (b) (c)

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



(a) (b) (c)

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



(a) (b) (c)

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	3583	3656	3583	3656	3333	4225	24300	20116	22602	18873	22602	19274
4/2/2021	3743	3428	3743	3428	3490	4279	21888	21166	20411	19589	20411	19786
4/3/2021	3460	4125	3460	4125	3235	5590	18129	22465	16950	20970	16950	20372
4/4/2021	2588	4281	2588	4281	2426	5327	12196	22429	11432	20963	11432	20708
4/5/2021	2338	3813	2338	3813	2197	3584	8497	21266	7985	19568	7985	15922
4/6/2021	2237	4811	2237	4811	2108	3921	6885	20153	6487	18265	6487	12893
4/7/2021	2548	4862	2548	4862	2407	4164	9677	19727	9141	18041	9141	11569
4/8/2021	3011	3832	3011	3832	2851	3695	20407	19947	19325	18131	19325	11178
4/9/2021	2855	3685	2855	3685	2711	4701	25464	20461	24175	18808	24175	11173
4/10/2021	2456	4471	2456	4471	2338	5978	24097	22318	22936	20750	22936	13842
4/11/2021	845	4404	845	4404	806	3158	17855	22520	17038	20935	17038	12295
4/12/2021	1730	4490	1730	4490	1655	1143	13245	21038	12671	19242	12671	9416
4/13/2021	3568	3833	3568	3833	3422	1019	10810	20144	10367	18414	10367	9459
4/14/2021	2471	3464	2471	3464	2376	1188	21693	20289	20857	18443	20857	9110
4/15/2	249	4056	249	4056	240	1676	294	20057	283	18128	283	8799

021	1		1		1		26		62		62	
4/16/2021	2671	4220	2671	4220	2581	2176	25831	19980	24960	18076	24960	9109
4/17/2021	2756	3789	2756	3789	2670	2628	23804	20783	23058	18924	23058	11123
4/18/2021	2206	4721	2206	4721	2142	2595	19185	21318	18630	19514	18630	10350
4/19/2021	1882	4793	1882	4793	1832	747	0	20755	0	19111	0	9157
4/20/2021	2963	3775	2963	3775	2892	-139	21046	20702	20539	18940	20539	8952
4/21/2021	2524	3601	2524	3601	2469	340	24884	20427	24344	18705	24344	8779
4/22/2021	2396	4334	2396	4334	2350	1267	29518	20387	28949	18492	28949	8775
4/23/2021	2728	4309	2728	4309	2682	1376	27543	20480	27078	18563	27078	9010
4/24/2021	2678	4367	2678	4367	2639	2411	23392	20859	23054	19032	23054	10591
4/25/2021	2061	3720	2061	3720	2036	1109	18773	20784	18547	19172	18547	9519
4/26/2021	1712	3356	1712	3356	1695	-429	11907	21117	11792	19372	11792	8883
4/27/2021	2064	3951	2064	3951	2049	85	10976	21202	10897	19480	10897	8633
4/28/2021	2685	4114	2685	4114	2672	331	22231	21134	22124	19439	22124	8497
4/29/2021	2166	3682	2166	3682	2161	653	24736	21529	24676	19728	24676	9419
4/30/2021	2445	4616	2445	4616	2445	1371	24329	21723	24329	19846	24329	10345

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)

	Actual	predicted	Actual	predicted	Actual	predicted	actual	predicted	actual	predicted	actual	predicted
4/1/2021	23887	16817	22218	15640	22218	15571	58828	28114	54717	28481	54717	26980
4/2/2021	22184	16684	20687	15715	20687	14966	50659	26993	47241	27813	47241	26223
4/3/2021	21917	17386	20491	16517	20491	14798	45520	26864	42558	27593	42558	28664
4/4/2021	21247	17653	19916	16604	19916	15290	13917	26464	13045	28214	13045	29491
4/5/2021	18017	18153	16932	17208	16932	15185	66794	28198	62770	27943	62770	29706
4/6/2021	10680	18054	10062	17084	10062	15584	12335	26359	11622	28072	11622	27811
4/7/2021	7763	18084	7333	17022	7333	15790	48646	24734	45950	28501	45950	25933
4/8/2021	13686	17924	12960	16889	12960	15690	30785	25747	29153	28786	29153	27262
4/9/2021	17209	17575	16338	16624	16338	15402	23638	24809	22442	28358	22442	26845
4/10/2021	18924	18230	18012	17362	18012	15594	39963	25275	38037	28493	38037	28906
4/11/2021	17551	18900	16747	18160	16747	15581	43283	30361	41301	28787	41301	34888
4/12/2021	15737	19030	15054	18092	15054	15908	34895	27415	33381	29317	33381	29562
4/13/2021	9780	19199	9379	18218	9379	16128	7556	27488	7246	29416	7246	30267
4/14/2021	13439	18533	12921	17696	12921	15564	38956	26492	37454	28748	37454	29378
4/15/2021	16160	18551	15576	17709	15576	15701	43354	25779	41787	28528	41787	29169
4/16/2021	16963	19102	16391	18180	16391	16176	37936	26522	36656	29149	36656	30526
4/17/2021	15923	18887	15424	17937	15424	16047	36360	25900	35221	28878	35221	30755
4/18/2021	15364	18999	14920	18083	14920	16093	34218	25828	33229	29007	33229	30751
4/19/2021	12688	19239	12352	18297	12352	16374	29344	26297	28566	29436	28566	29924

4/20/2021	8863	19116	8649	18217	8649	16274	6636	26635	6476	29721	6476	30418
4/21/2021	12069	18766	11807	17953	11807	15997	42553	25847	41630	29293	41630	29684
4/22/2021	13836	18922	13569	18081	13569	16177	34368	25526	33705	29428	33705	29736
4/23/2021	16046	18846	15775	17962	15775	16232	34034	25599	33460	29722	33460	29543
4/24/2021	14758	19277	14545	18289	14545	16605	32192	26238	31727	30252	31727	31983
4/25/2021	13814	19504	13648	18495	13648	16873	32500	26092	32108	30351	32108	31571
4/26/2021	13157	18890	13030	18050	13030	16267	22849	25093	22629	29683	22629	29670
4/27/2021	8440	18977	8379	18166	8379	16359	49144	24483	4878	29463	4878	29386
4/28/2021	10398	19499	10348	18586	10348	16857	29980	25067	29836	30083	29836	29986
4/29/2021	13382	19356	13350	18454	13350	16652	31444	24818	31368	29813	31368	29681
4/30/2021	14314	19446	14314	18565	14314	17135	26373	24492	26373	29942	26373	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.
Kırbaş, 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).