New SARIMA Approach Model to Forecast COVID-19 Propagation: Case of Morocco

Ibtissam CHOUJA[1]
Engineering Sciences Laboratory
Faculty of Sciences
Ibn Zohr University
Agadir, Morocco

Sahar SAOUD[2]
Technical Research Laboratory
Faculty of Applied Sciences
Ibn Zohr University
Agadir, Morocco

Mohamed SADIK[3]
Engineering Sciences Laboratory
Faculty of Sciences
Ibn Zohr University
Agadir, Morocco

Abstract—The aim of this paper is to avoid any future health crises by analysing COVID-19 data of Morocco using Time Series to get more information about how the pandemic is spreading. For this reason, we used a statistical model called Seasonal Autoregressive Integrated Moving Average (SARIMA) to forecast the new confirmed cases, new deaths, cumulative cases and deaths. Besides predicting the spreading of COVID-19, this study will also help decision makers to better take the right decisions at the right time. Finally, we evaluated the performance of our model by measuring metrics such as Mean Squared Error (MSE). We have applied our SARIMA model for a forward forecasting in a period of 50 days, the MSE reported was 62196.46 for cumulative cases forecasting, and 621.14 for cumulative deaths forecasting.

Keywords—*COVID-19, machine learning, seasonal autoregressive integrated moving average, SARIMA, statistical modeling, time series forecasting*

I. Introduction

Since the announcement of sars-cov2 outbreak as a public health emergency by the world health organization (WHO) in January 2020, The world didn't stop counting its damage due to the spreading of this virus. The incessant stream of new unpredictable cases has caused crises in many countries. For instance, in Italy, the death toll caused by COVID-19 raised dramatically in a short period of time making it the biggest crisis since the Second World War. To control the crisis, many decisions have been made all over the world. However, all these efforts lacked efficiency since we didn't yet predict what the world needs in terms of tools and information to make better decisions, to end the pandemic as fast as possible, or at least prevent some of its dire consequences.

In Morocco, the first case was detected on March 2020, and by July 2021 the total cases reached 552635 confirmed cases and 9427 deaths according to the Moroccan health ministry [1]. The situation has become critical especially when COVID-19 reaches its peak at the critical phase. To manage this, the Moroccan government has taken many decisions regarding the timeline development of the disease, but as the pandemic hasn't been stopped yet, we can expect other peaks that should be predicted beforehand to minimize the consequences.

Many studies have been done to decrypt the spreading mechanism of the mysterious virus that has caused vast damage in different sectors. It was important for mathematicians, epidemiologists, and data analysts to analyze and model the phenomenon using various models and methods such as compartmental models [2, 3, 4, 5], machine learning and artificial intelligence models[6, 7], or statistical methods [8, 9, 10, 11].

In this paper, we present a new SARIMA model to forecast the new cases and deaths in Morocco to help decision-makers to prevent any critical phases. We have chosen to forecast COVID-19 using time series analysis. In Section II, we will present the related work. Then Section III will present the time series approach for forecasting, and the process of this study. Afterward in Section IV, we will describe the materials and methods we used in this work. Section V will be devoted to presenting the numerical results that will be discussed in Section VI. Finally, we will end this article with a general conclusion.

II. Related Work

A. Compartmental Models

Compartmental models are mathematical modelling of infectious diseases. One of the most important and commonly used of these models is the "Susceptible-Infectious-Recovered or Resistant" model known also as the SIR model, a simplest compartmental model based on the research of Anderson Gray McKendrick and William Ogilvy Kermack in 1927[12]. This model divides the population into three compartments or categories and calculates the transition rate between compartments. Other models had been extended from the SIR model like the SEIR[13] model dividing the population into four compartments instead of three, by adding a new one named "Exposed" representing people highly exposed to the infection.

Several studies have adapted the SIR and SEIR models to investigate the new coronavirus. Chen et al. [2] created the Bats-Hosts-Reservoir-People (BHRP), a transmission network model describing the transmission chain of the virus, from bats to people passing by the hosts and the seafood market, using the SEIR model. While Euloge Mouvoh et al. [3]described the impact of different intervention strategies on the spread of COVID-19 in Morocco using a contact-structured and age-structured-SEIR model considering the interventions dynamic between different age groups. On the other hand, Ben Hassen et al. [4]combined the Poisson Markov process with a SIR model to build a hybrid model called SIR-Poisson model to estimate deaths number at a certain date t. However, the SIR model failed to forecast COVID-19 in Isfahan, because the

assumptions that the SIR model is based on are not seemingly true for COVID-19 [5].

B. Artificial Intelligence and Machine Learning Models

In addition to compartmental modelization, Artificial Intelligence and machine learning models are highly recommended in such problems it can be used in various aspects. Ahmed et al. [6] suggested some problems related to the pandemic where Artificial Intelligence can act efficiently and expected that studies using those models will increase significantly when more COVID-19 data will be available. In this direction, Mbilong et al. [7] proved the efficiency of Deep Learning models by using six models of machine learning and deep learning to forecast COVID-19 in Morocco during 7 days.

C. Statistical and Time Series Methods

statistical methods have been used to forecast, identify and measure the damage due to the COVID-19 pandemic. Shang et al. [10]used time series analysis to identify any change points using excess in the counts of COVID-19 caused excess death in Belgium. Salgotra et al. [9] have chosen to predict COVID-19 in India using models based on genetic programming and evaluated by statistical methods and metrics. Alsulami et al.[11] presented a statistical method to study the effects of various factors on the deaths due to COVID-19. ArunKumar et al.[8]have forecasted COVID-19 cases using the statistical models SARIMA and ARIMA for 60 days in 16 countries.

In this paper, we have chosen to use the statistical model SARIMA to forecast cases in Morocco. In the next section, we will present the model and the process of this work.

III. COVID-19 PROPAGATION BY USING TIME SERIES APPROACH

Time series analysis is the process of analyzing a sequence of data recorded at a specific interval of time. As its name indicates time is a crucial variable in this process. It shows us how the problem studied is adjusted over time, and it provides an additional source of information extracted from the dependencies established between data and the time variable. Time series analysis is usually utilized for problems that are highly affected by time, it can be used to analyze and forecast trading data, weather data, electricity data, etc. With a representative sample and cleaned data, Time Series analysis can provide a lot of information about the phenomenon in question, or even forecast the future with high accuracy.
The Time series process requires 5 principal steps as shown in Fig. 1:

- **Data collection**: It's the first thing to do when using time series analysis, data can be collected by observation, experimentation , simulation or simply derived.

- **Data-preprocessing**: In this step, we prepare data collected to be used for forecasting, the preprocessing consists of cleaning data and handling the missing values.

- **Selection of Time series model**: Many time series models are available, but their efficiency depends on the data used for the study. In this step, we analyze data to extract its characteristics, to choose the best model for this study.

- **Model tuning**: It's a process that creates a combination of a manually specified subset of the hyperparameter space and evaluates each model to choose the best combination of those parameters. It's called also grid search.

- **Forecasting and interpretation**: After specifying the best parameters of our model, we can finally fit it and forecast our variables.

After explaining briefly the process of a time series analysis, we propose in the next section a comparison between two time series models known as Autoregressive Integrated Moving Average(ARIMA) and seasonal Autoregressive Integrated Moving Average (SARIMA).

A. Time Series Generalities (ARIMA VS. SARIMA)

The Autoregressive Integrated Moving Average model (ARIMA), is one of the models that are most widely used to forecast a time series. It describes the autocorrelations in data by combining the autoregressive model (AR); a linear regression between past and future values; and moving average model (MA) that considers the past forecast error and future values of data. The added value of ARIMA is the Integrated component (I) that describes the order of differentiation used to make data stationary [14].

ARIMA model depends on 3 parameters:
p: trend autoregressive order, **q**: trend difference order, and **d**: trend moving average order.

When time-series data present a seasonality pattern, we use the Seasonal Autoregressive Integrated Moving Average model (SARIMA) instead of ARIMA. SARIMA is an extension of ARIMA that can deal with seasonality to forecast seasonal Time series. It is composed of four parts. In addition to the three components of ARIMA described before, SARIMA is composed of seasonal components describing the seasonality pattern in data [14].

SARIMA depends on 4 other hyper-parameters (P D Q s) in addition to the ARIMA's parameters (p d q):

- **P**: Seasonal autoregressive order

- **D**: Seasonal difference order

- **Q**: Seasonal moving average order

- **s**: The number of time steps for a single seasonal period.

SARIMA model is described mathematically as :

$$\varphi_p(B)\Phi_p(B^s)\nabla^d \nabla_s^D y_t = \theta_q(B) \Theta_{\mathcal{Q}}(B^s)\varepsilon_t \qquad (1)$$

where:
y_t: is the forecast variable
$\varphi_p(B)$: is the regular AR polynomial of order p
$\theta_q(B)$: is the regular MA polynomial of order q
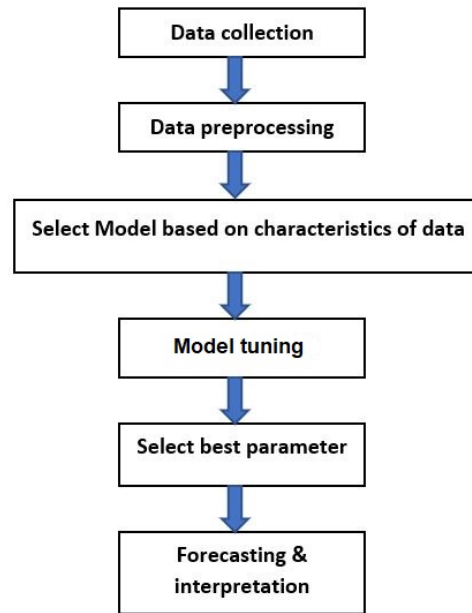$\Phi_p(B^s)$: is the seasonal AR polynomial of order P

Fig. 1. SARIMA Process.

$\Theta_Q\left(B^s\right)$: is the seasonal MA polynomial of order Q
∇^d is the differentiation operator
∇^D_s is the differentiation operator
ε_t white noise process B back shift of y_t

To optimize our model and make sure that the parameters chosen are optimum, we used a process that creates a combination of a manually specified subset of the hyper-parameters space and evaluates each model to choose the best combination of those parameters. This process is known as grid search or model tuning.

IV. MATERIALS AND METHODS

A. Methods and Data Preprocessing

We collected data from [15], it is a recording of corona cases in the whole world, for the period going from 24/02/2020 to 18/07/2021 with 103348 rows and 59 columns.

The data preprocessing in our case consists of creating new data for Morocco by using the global data collected, then cleaning this one by detecting the aberrant values in order to correct them, and finally dealing with the missing values by deleting some useless rows or filling them by fitting values. The specificity of time series data preprocessing is that we should also handle the trend and stationarity by de-trending data and making it stationary especially when using models like ARIMA or SARIMA.

The final dataset is new data recorded from 02/03/2020 (the date of the first case in Morocco) to 18/07/2021 with 504 rows (observations) and 29 columns. The new dataset is shown in Fig. 2

The new data was divided into two categories:

- Training data: is used to train and fit the model to find the best parameters.

- Testing data: is used for forecasting and assessing the performance of our model.

B. Time Series Analysis

In this section, we analyse the components of our data to extract the different characteristics of our time series data.

1) Trend: The first thing to check on time series data is the presence of a long-term movement called a trend in a cyclical context [16].

By plotting the data (Fig. 3), we observe the presence of a trend, but this should be proved by some tests. To test the presence of a trend in this time series, we used "pyMannkendall" a Python implementation of nonparametric Mann-Kendall trend analysis .

The mann-kendall test confirmed the existence of an upward trend on the data of new cases and new deaths with a significant p-value of p=5.329070518200751e-15 for new cases and 1.4821224705308111e-05 for new deaths. Before applying any model of forecasting, we should de-trend our data. There are many ways to do this, the most common method is differentiation.

2) Stationarity: Stationarity is one of the most important things to check before dealing with a time series. Stationary data means that the variance and the mean are constant over time. In other words, the data is stable [17]. However A non-stationary data is unpredictable data unless we make it stationary which is possible using some transformations. To test stationarity, we used two tests ADF and KPSS from the statsmodels package:

Fig. 2. Moroccan Data.

Fig. 3. A- Daily New Cases of COVID-19 in Morocco. B- Daily New Deaths due to COVID-19 in Morocco.

- Augmented Dickey-Fuller test or ADF test: in which the null hypothesis admits the existence of unit root which means that the data is non-stationary [18].

- Kwiatkowski Phillips Schmidt Shin or KPSS test: unlike ADF the null hypothesis that we are testing in the KPSS is that our data is stationary [19].

The results of these two tests are shown in Table I.

Both ADF and KPSS tests confirm the non-stationarity of our data for the reasons below:

- For ADF:
  - The p-value is higher than the significance level of 0.05.
  - The ADF statistical test is higher than all the critical values.

- For KPSS:
  - The p-value is less than the significance level of 0.05.
  - The KPSS statistical test is higher than all the critical values.

It's indispensable to make data stationary before using it on a forecasting operation to make it predictable and to get relevant results. In our case, we have chosen the differentiation method to make it stationary. Fig. 4A and Fig. 4B show the data after 1st and 2nd order differencing, we can observe that our data had become stationary with the first differencing order and there is no need for a second one. To prove the stationarity of the new data, we have retested it with KPSS and ADF.
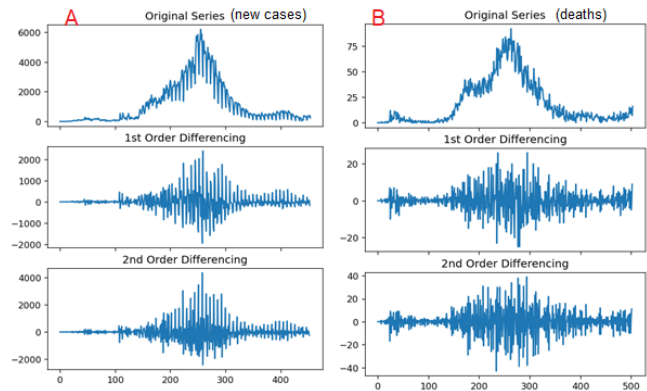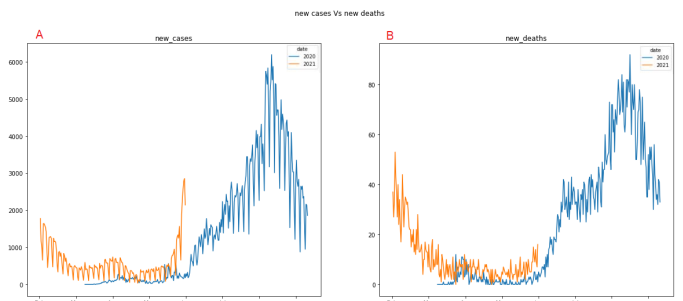

Fig. 4. A- Differencing Results of Original Series of New Cases. B- Differencing Results of Original Series of New Deaths.

Fig. 5. A- Daily New Cases per Year in Morocco. B- Daily New Deaths per Year in Morocco.

3) Seasonality: The last component that we should verify is seasonality which means the presence of a pattern. In the Fig. 5A and Fig. 5B, we observe the existence of some pattern between 2020 (year of the appearance of COVID-19 In Morocco) and 2021. For example, the number of new cases increases between June and August, probably because of the school and summer vacations in which most people travel a lot and don't respect restriction policies. In contrast, the number of new cases decreases between November and February, and it is somehow stationary between February and January.

TABLE I. KPSS VS. ADF TEST RESULTS

		KPSS				ADF		
New cases	**Stastical test**	0.449026				-1.813390		
	p-value	0.01				0.373834		
	critical values	10 %	5 %	2.5 %	1 %	10 %	5 %	1 %
		0.119	0.146	0.176	0.216	-2.570297	-2.868161	-3.445368
New deaths	**Stastical test**	0.449026				-1.813390		
	p-value	0.01				0.373834		
	critical values	10 %	5 %	2.5 %	1 %	10 %	5 %	1 %
		0.119	0.146	0.176	0.216	-2.569954	-2.867518	-3.443905

V. NUMERICAL RESULTS

In this section, we present the results of COVID-19 daily new cases, new deaths, cumulative cases, and cumulative deaths forecasting using our proposed models.

After the model tuning, we have chosen the best parameters for each model. So we selected SARIMA (5,1,0)(1,0,1,7) to forecast new cases, and SARIMA (0, 1, 1) (2, 0, 2, 7) for the new deaths forecasting. The Fig. 6A shows the prediction of daily new cases, while figure Fig. 6C shows the prediction of daily new deaths for 104 days.

To measure the performance of our models, we calculated the Mean Equared Error(MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) for 104 days in Table II, and we have compared our results with other persistent models.

In Fig. 6B and Fig. 6D we noted that the model fits very well at the first 50 days of forecasting and start deviating after, so we recalculated the errors for 50 days. Table III shows the metrics reported for the 50 days forecasting.

VI. DISCUSSION

In the present study, we developed a SARIMA model to fit the reported data of COVID-19 in Morocco. The proposed model was prepared to select the best parameters in order to predict cumulative cases and deaths, new daily cases, and deaths.

The results show that the proposed models fit very well at the first 50 days with a reported MSE of 316960.2267400527 for new daily cases forecasting, 8.28 for new daily deaths, 62196.46 for cumulative cases, and 621.14 for cumulative deaths. The model starts to deflect after 50 days and the error starts to become higher. We tested our model for a forecasting period of 104 days and have reported an MSE of 4881.384615384615 for cumulative deaths forecasting, and an MSE of 108689697.75 for cumulative cases.

The 104 days predictions results are good if we compare them to most of the reported results in [8] for 60 days forecasting in 16 countries, where the MSE variates between 6.63E+04 for cumulative cases in Bangladesh and 2.69E+09 for Brazil, while the MSE of cumulative deaths forecasting for the same number of days variates between 3.10E+00 in Spain and 2.24E+06 in India.

VII. CONCLUSION

Since the outbreak of COVID-19 in Morocco, the government has made many decisions regarding the real-time situation, therefore forecasting the COVID-19 impact could help in preventing any critical situation by making a well-informed decision.

In this paper, we used time series analysis, one of the most important aspects of data analytic to analyze Moroccan data by understanding components like trend and seasonality. The ultimate objective of this work was to predict cases (new confirmed cases, new deaths, cumulative cases, and cumulative deaths) for 104 days using the SARIMA model. The results of our proposed models matched with test data especially for the first 50 days of forecasting for both cumulative cases and deaths. Decision-makers should consider forecasting models like SARIMA in preventing any future critical situation.

REFERENCES

[1] "Statistique of covid-19 in morroco," http://covidmaroc.ma/Pages/ AccueilAR.aspxhttp://covidmaroc.ma/Pages/AccueilAR.aspx, accessed: 2021-07-16.

[2] T.-M. Chen, J. Rui, Q.-P. Wang, Z.-Y. Zhao, J.-A. Cui, and L. Yin, "A mathematical model for simulating the phase-based transmissibility of a novel coronavirus," *Infectious Diseases of Poverty*, vol. 9, no. 1, 2020.

[3] A. C. Euloge Mouvoh, A. Bouchnita, and A. Jebrane, "A contact-structured SEIR model to assess the impact of lockdown measures on the spread of COVID-19 in Morocco's population," in *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*. Kenitra, Morocco: IEEE, Dec. 2020, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9314462/

[4] H. Ben Hassen, A. Elaoud, N. Ben Salah, and A. Masmoudi, "A SIR-Poisson Model for COVID-19: Evolution and Transmission Inference in the Maghreb Central Regions," *Arabian Journal for Science and Engineering*, vol. 46, no. 1, pp. 93–102, Jan. 2021. [Online]. Available: http://link.springer.com/10.1007/s13369-020-04792-0

[5] S. Moein, N. Nickaeen, A. Roointan, N. Borhani, Z. Heidary, S. H. Javanmard, J. Ghaisari, and Y. Gheisari, "Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan," *Scientific Reports*, vol. 11, no. 1, p. 4725, Dec. 2021. [Online]. Available: http://www.nature.com/articles/s41598-021-84055-6

[6] A. Ahmed, P. Boopathy, and Sudhagara Rajan S., "Artificial Intelligence for the Novel Corona Virus (COVID-19) Pandemic: Opportunities, Challenges, and Future Directions," *International Journal of E-Health and Medical Communications*, vol. 13, no. 2, pp. 1–21, Jul. 2022. [Online]. Available: http://services.igi-global.com/resolvedoi/ resolve.aspx?doi=10.4018/IJEHMC.20220701.oa5

[7] P. M. Mbilong, A. Berhich, I. Jebli, A. El Kassiri, and F.-Z. Belouadha, "Artificial Intelligence-Enabled and Period-Aware Forecasting COVID-19 Spread," *Ingénierie des systèmes d information*, vol. 26, no. 1, pp. 47–57, Feb. 2021. [Online]. Available: http: //www.iieta.org/journals/isi/paper/10.18280/isi.260105
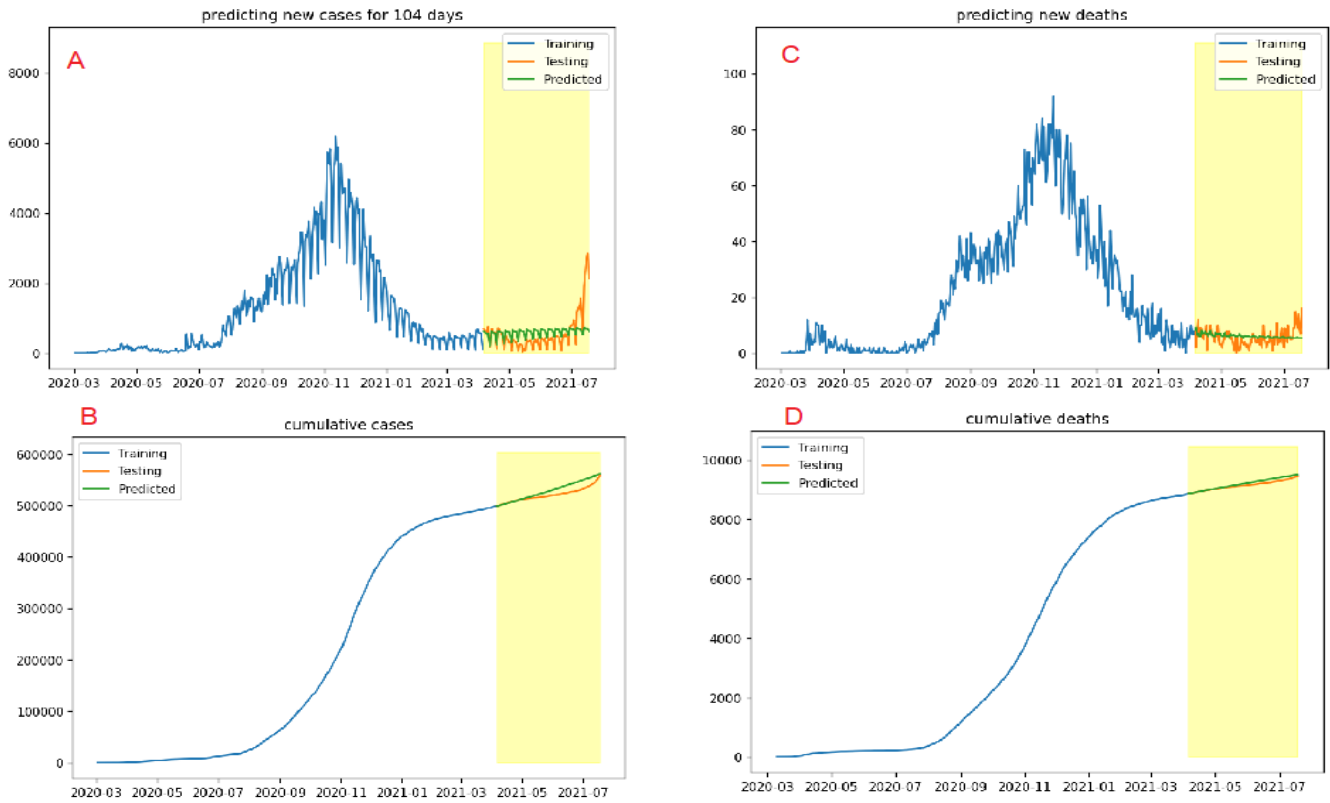
Fig. 6. 104 Days Forcasting. A-Forecasting of Daily New Cases. B- Forecasting of Cumulative Cases.C-Forecasting of Daily New Deaths. D- Forecasting of Cumulative Deaths.

TABLE II. 104 Days Forecasting Errors

	MSE	MAE	RMSE
New cases	254695.1826923077	326.72115384615387	504.67334256160956
New deaths	10.192307692307692	2.4615384615384617	3.1925393799149435
Cumulative cases	108689697.75	8267.288461538461	10425.435134803727
Cumulative deaths	4881.384615384615	56.73076923076923	69.86690643920493

TABLE III. 50 Days Forecasting Errors

	MSE	MAE	RMSE
New cases	316960.2267400527	549.3631462639955	562.9922084186003
New deaths	8.28	2.16	2.8774989139876315
Cumulative cases	62196.46	203.38	249.39218111239975
Cumulative deaths	621.14	17.78	24.922680433693323

[8] K. ArunKumar, D. V. Kalaga, C. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "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*, vol. 103, p. 107161, May 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7869631/

[9] R. Salgotra, M. Gandomi, and A. H. Gandomi, "Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming," *Chaos, Solitons & Fractals*, vol. 138, p. 109945, Sep. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960077920303441

[10] H. L. Shang and R. Xu, "Change point detection for COVID-19 excess deaths in Belgium," *Journal of Population Research*, Mar. 2021. [Online]. Available: http://link.springer.com/10.1007/s12546-021-09256-2

[11] M. D. Alsulami, H. Abu-Zinadah, and A. H. Ibrahim, "Machine Learning Model and Statistical Methods for COVID-19 Evolution Prediction," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–6, Dec. 2021. [Online]. Available: https://www.hindawi.com/journals/wcmc/2021/4840488/

[12] G. McKendrick, "A contribution to the mathematical theory of epidemics," p. 22.

[13] M. Y. Li and J. S. Muldowney, "Global stability for the SEIR model in epidemiology," *Mathematical Biosciences*, vol. 125, no. 2, pp. 155–164, Feb. 1995. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0025556495927565

[14] A. Atangana and S. İğret Araz, "Modeling and forecasting the spread of COVID-19 with stochastic and deterministic approaches: Africa and Europe," *Advances in Difference Equations*, vol. 2021, no. 1, p. 57, Dec. 2021. [Online]. Available: https://advancesindifferenceequations.springeropen.com/articles/10.1186/s13662-021-03213-2

[15] "World data of covid-19," https://ourworldindata.org/coronavirus/country/moroccohttps://ourworldindata.org/coronavirus/country/morocco, accessed: 2021-07-18.

[16] M. Hussain and I. Mahmud, "pyMannKendall: a python package for non parametric Mann Kendall family of trend tests." *Journal of Open Source Software*, vol. 4, no. 39, p. 1556, Jul. 2019. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.01556

[17] Badi H. Baltagi, "Time-Series Analysis," in *Econometrics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 355–377. [Online]. Available: https://doi.org/10.1007/978-3-540-76516-5_14

[18] R. Harris, "Testing for unit roots using the augmented Dickey-Fuller test," *Economics Letters*, vol. 38, no. 4, pp. 381–386, Apr. 1992. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/016517659290022Q

[19] D. Lee and P. Schmidt, "On the power of the KPSS test of stationarity against fractionally-integrated alternatives," *Journal of Econometrics*, vol. 73, no. 1, pp. 285–302, Jul. 1996. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0304407695017410