Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Overview and cross-validation of COVID-19 forecasting univariate models

Mintodê Nicodême Atchadé, Yves Morel Sokadjo

a National Higher School of Mathematics Genius and Modelization, National University of Sciences, Technologies, Engineering and Mathematics, Abomey, Rep. Benin
b University of Abomey-Calavi/International Chair in Mathematical Physics and Applications (ICMPA: UNESCO-Chair), 072 BP 50 Cotonou, Rep. Benin

Received 15 March 2021; revised 10 June 2021; accepted 14 August 2021
Available online 21 August 2021

KEYWORDS
COVID-19;
Cross-validation;
Forecast;
Statistical modeling;
Time series

Abstract Researchers have been working with different models to forecast COVID-19 cases. Many of their estimates are not accurate. This study aims to propose the best model to forecast COVID-19 cumulative cases using a machine learning technic. It is a work that focused on time series univariate models because there are too many debates about the quality of the pandemic data. To increase the likelihood of the findings, we avoided many variables modeling and proposed a robust process to forecast COVID-19 cumulative cases. It will help international institutions to take optimal decisions about the world economy and response to the pandemic. Consequently, we used the data titled “Coronavirus Pandemic (COVID-19)” from “Our World in Data” about cases from 22 January 2020 to 30 November 2020. We computed Error Trend Season (ETS), Exponential smoothing with multiplicative error-trend, and ARIMA on the training data sets. In addition, we calculated the Mean Absolute Percentage Error (MAPE) per model. Among those models, we notice that ETS (with additive error-trend and no season) has the smallest MAPE statistics compared to the others. The findings revealed that with the ETS model we need at least 100 days to have good forecasts with a MAPE threshold of 1%.

1. Introduction

In China, precisely in Wuhan city, COVID-19 was declared as a world pandemic by WHO Emergency Committee on January 30th, 2020 [1]. Since that day, we are observing rapid increases [2,3] of cases that made authorities give daily government communication to update the information of the day before. In this context, one of the main issues is to know the future number of cases with the least possible bias, to define appropriate policies to control COVID-19 spread. Thus, it is easy to notice that numerous papers [4–8] have proposed different forecast models of COVID-19 using univariate or multivariate time series modeling or complex models considering each data distributions [9–13].
The multivariate time series models have advantages because they can reveal the influence of many parameters such as face mask wearing, social distancing, hand washing, airport screening, quarantine, and treatment protocols on COVID-19 spread. However, as they require numerous parameters, the goodness of fit is affected when any parameter gets to be not correct. Besides, the quality of data related to COVID-19 is debatable because many countries were obliged to substract some cases that they reported to the World Health Organization (WHO) [14,15]. Actually, those issues happened because it has been noticed that some cases were reported without any respect of the official WHO technical guidance for laboratory testing [16]. Consequently, in data sets we can get negative numbers and it is an indicator of not high-quality data sets. Recently, due to data quality doubts, the authors were obliged to increase twenty times the cases, forty times the recovered patients to approach real figures [5]. Multivariate time series models also face data availability issues because during confinement, the data collected were not exhaustive. Moreover, most of the compartmental models in epidemiology depend on estimated inputs such as case fatality and case recovery that are very sensitive to data quality. In the paper [5], it is easy to notice that the change in data due to uncertainty quite influenced the estimates of the Susceptible-Infectious-Recovered-Dead models. Considering the fact that the more complex model, the more the need of data, and the lesser the quality of COVID-19 data sets; we think that a good univariate model that can give better results will help to handle the bias in forecasts.

Concerning univariate time series, there are many works that use different methods. We have the computation of forecasts with an exponential smoothing family [6] that has appreciable forecast accuracy and can fit short series. However, the choice of multiplicative trend and error in the context of time series cross-validation is much debatable. Even the multiplicative trend model takes into account the asymmetric risk, we think that a robust model should take into account an additive trend because short-term multiplicative trend is to be additive in the long run. In that same paper, the author used 90% prediction interval and we think it can be improved. Another recent study [7] has compared time series models in predicting COVID-19 cases, but it was not exhaustive because it just focused on Auto Regressive models (1–3) using Maximum Likelihood, Conditional Least Squares, and Unconditional Least Squares. Additionally, there is just a work in Nepal [17] about COVID-19 forecasts that uses ETS models, Auto Regressive Integrated Moving Average (ARIMA), and Susceptible Infectious-Recovered-Dead models. Considering the fact that the more complex model, the more the need of data, and the lesser the quality of COVID-19 data sets; we think that a good univariate model that can give better results will help to handle the bias in forecasts.

2. Materials and methods

2.1. Materials

As introduced, we focused on COVID-19 total cases from 2020–01–22 to 2020–11–30. They are time series, and we collected each day total cases in the dataset titled “Coronavirus Pandemic (COVID-19)” from “Our World in Data” [18]. The variable of interest is the count of laboratory-confirmed infections and they are indexed by their respective date. The data were accessed on 2020–12-02 and we can make the dataset available if requested. The accuracy and reliability of those numbers are linked to daily verification and change. Actually, in the context of this work, we divided the period and considered the first 30 days and after iteratively added 14 days to the last period to compare the real and predicted values.

2.2. Methods

The classical univariate models in this work are ARIMA, exponential smoothing model with multiplicative error and multiplicative trend components (ESM), and ETS. Each of them will be computed using cross-validation methods, and finally we will select the most appropriate in terms of forecast errors using the mean absolute percentage error (MAPE).

2.2.1. ARIMA modeling

ARIMA or autoregressive integrated moving average is a form of statistical modeling that uses time series data to either predict a future trend or to output latent information to understand how a variable of interest changes within a period. It has three parameters and the first one \(p\) is about the order of the Autoregressive (AR) model, the second one \(d\) concerns the level of differentiating, and the third one \(q\) shows the Moving average (MA) order. Thus, ARIMA model is denoted by ARIMA \((p,d,q)\) and it is frequently used to represent actual time series data. It is an improved version of an ARMA model, that is, a stationary series without difference. For non-stationary series, ARIMA can be very useful. Its functional is the Eq. (1).

\[
(1 - \rho_1B - \ldots - \rho_pB^p)\eta_t = (1 - \theta_1B - \ldots - \theta_qB^q)e_t
\]

(1)
With $\epsilon_t \sim \text{whitenoise}(0, \sigma^2)$, $By_t = y_{t-1}$ and $B^d = y_{t-d}$, $\nabla^d = (1 - B)^d$ is the differentiation parameter of order $d$ ($d \geq 0$), $(\rho_1, \ldots, \rho_p), (\theta_1, \ldots, \theta_q)$ are the coefficients and $\sigma^2$ the residual variance to be estimated.

2.2.2. ETS modeling

ETS model is a modeling that captures different components (Error, Trend, Season) and makes short-term forecasts, which is appropriate in the case of strong dynamics. This model focuses on trend, seasonal components of different traits [19]. The possible combinations of the trend and season give the 15 following models in the Table 1. Consequently, 30 different models are possible (15 with additive errors and 15 with multiplicative ones). In other words, in combination with the error that can be Additive or Multiplicative, the models in the Table 1 can be extended to 30 models in total. We can recall that the paper [6] used ETS (M,M,N) that is already included in the Table 1. For instance, ETS (A,A,N) is defined by:

\begin{align}
    y_t &= l_{t-1} + b_{t-1} + \epsilon_t \\
    l_t &= l_{t-1} + b_{t-1} + \alpha \epsilon_t \\
    b_t &= b_{t-1} + \beta \epsilon_t
\end{align}

Besides, with $h$, a step ahead forecast parameter, the particular case related to damped trend has as a recurrence form [20], the following equations:

\begin{align}
    \hat{y}_{t+h} &= l_t + (\phi \epsilon_t^1 + \ldots + \phi^h \epsilon_t^h) \\
    l_t &= \lambda y_t + (1 - \lambda)(l_{t-1} + \phi b_{t-1}) \\
    b_t &= \beta(l_t - l_{t-1}) + (1 - \beta) \phi b_{t-1}
\end{align}

where $l_t$ and $b_t$ are, respectively, the level and trend components at time $t$, $\epsilon_t$ the error term with the smoothing parameter $0 < \lambda < 1$ for level, and $0 < \beta < 1$ for trend, and $\phi$ the damping parameter. The initial coefficients $l_0$ and $b_0$ and the latter smoothing parameters are obtained with the maximum likelihood estimate (MLE) method [21].

2.2.3. Mean absolute percentage error

The mean absolute percentage error (MAPE) is also called mean absolute percentage deviation (MAPD), and it is a statistic that quantifies whether a forecast is accurate or not. To make that estimate easier of interpretation, MAPE is set as a percentage of the errors and it is the formula (4).

\begin{equation}
    \text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100
\end{equation}

Table 1 Different models in ETS modeling.

| Trend | Season |
|-------|--------|
| N     | A      | M      |
| N     | N,N    | N,A    | N,M    |
| A     | A,N    | A,A    | A,M    |
| AD    | AD,N   | AD,A   | AD,M   |
| M     | M,N    | M,A    | M,M    |
| MD    | MD,N   | MD,A   | MD,M   |

Note: N: None; A: Additive, M: Multiplicative, D: Damped, AD: Additive Damped, MD: Multiplicative Damped.

Fig. 1 Study analysis process.
Such that \( n \) is the number of predicted values, \( y_t \) the actual value at time \( t \), and \( \hat{y}_t \) the forecast. Actually, in the context of this work, the retained threshold of MAPE is 1%.

2.2.4. Analysis process

There are many works about latency periods and we got 11.5 days (CI: 8.2 to 15.6 days) from [22], 7–21 days from [23], 2–14 days from [24]. Although this period might change, we kept the latency period of 14 days [25] because it seems to be a common estimation in the aforementioned studies. We also used R software [26] (version: 4.0.0) for the whole work and the library “forecast” with the function ets() for ETS model, auto.arima for ARIMA model, and ets() with the parameters (model = “MMN”, damped = False) for the exponential smoothing model with multiplicative trend and error. The models use the algorithm of Hyndman & Khandakar [21] that combines unit root tests, minimisation of the AICc, and MLE to propose the best model. Besides, we have some main points in the analysis process before stepping in the final forecast. We selected the appropriate model among the ETS, ARIMA, Exponential smoothing (Multiplicative trend and error). Actually, for each training data set, we computed 2 weeks forecasts to check what model of short-term prediction is good enough to forecast COVID-19 cumulative cases. Every training data set was increased of 2 weeks until the end of the data. It helps to check how many days we can trust for a robust forecast. To select the most appropriate model, we checked each model assumption and used the MAPE. Generally, the analysis process in the work is in Fig. 1.

3. Results, discussion, and conclusion

3.1. Results

In the world, the number of new cases of COVID-19 infection is counted and put at the disposal of everyone by the WHO. Actually, the cumulative number of daily confirmed cases of COVID-19 on a given day is the sum of the new cases on that day and the total number of cases on the eve. It is illustrated in Fig. 2. When we look at Fig. 2, we can understand that there was a flat part of the graph until around the mid of March and after, we noticed a rapid increase in the number of cases. To forecast this time series, we need a model with the least bias in terms of MAPE. Thus, we used different training and test

| TDS  | 2 Weeks MAPE | ARIMA | AC | Norm | Hetero | Stat |
|------|--------------|-------|----|------|--------|------|
| From 1 to 30 | 12.06 | 010 | 0.19 | 0.00 | 0.35 | 0.58 |
| From 1 to 44 | 20.19 | 110 | 0.97 | 0.00 | 0.97 | 0.28 |
| From 1 to 58 | 23.75 | 021 | 0.66 | 0.00 | 0.62 | 0.07 |
| From 1 to 72 | 2.45 | 020 | 0.90 | 0.00 | 0.97 | 0.18 |
| From 1 to 86 | 2.66 | 222 | 0.94 | 0.00 | 0.03 | 0.01 |
| From 1 to 100 | 0.33 | 120 | 0.48 | 0.00 | 0.13 | 0.01 |
| From 1 to 114 | 0.48 | 120 | 0.57 | 0.00 | 0.29 | 0.01 |
| From 1 to 128 | 0.99 | 021 | 0.65 | 0.00 | 0.43 | 0.01 |
| From 1 to 142 | 0.60 | 021 | 0.84 | 0.00 | 0.49 | 0.01 |
| From 1 to 156 | 1.26 | 222 | 0.06 | 0.00 | 0.24 | 0.01 |
| From 1 to 170 | 0.89 | 222 | 0.01 | 0.00 | 0.28 | 0.01 |
| From 1 to 184 | 0.10 | 222 | 0.00 | 0.00 | 0.50 | 0.01 |
| From 1 to 198 | 0.42 | 222 | 0.00 | 0.00 | 0.46 | 0.01 |
| From 1 to 212 | 0.35 | 222 | 0.00 | 0.00 | 0.58 | 0.01 |
| From 1 to 226 | 0.26 | 222 | 0.00 | 0.00 | 0.34 | 0.01 |
| From 1 to 240 | 0.14 | 222 | 0.00 | 0.00 | 0.16 | 0.01 |
| From 1 to 254 | 0.72 | 222 | 0.00 | 0.00 | 0.04 | 0.01 |
| From 1 to 268 | 1.14 | 222 | 0.00 | 0.00 | 0.00 | 0.01 |
| From 1 to 282 | 0.55 | 222 | 0.00 | 0.00 | 0.04 | 0.01 |

Note: TDS: Training dataset, AC: Autocorrelation, Norm: Normality, Hetero: Heteroscedasticity, Stat: Stationarity.
data sets to check the stochastic assumptions and accuracy of the three models to forecast COVID-19 cumulative cases. The results are set in the Tables 2–5. We have in the Tables 2–4 different MAPEs and the verification of each models’ assumptions. To choose the best model, we proceeded as follows:

- **Point 1:** The best fitting model can be selected with the MAPE in the Tables 2–4. We have 2 weeks period of forecast and for each one we output descriptive statistics in the Table 5 to select the models with the smallest MAPE. In terms of range, we can notice that ETS models have the smallest mean (3.62), median (0.64), and maximum (23.48) of MAPE. Although it has the highest minimum (0.15), we can notice the best model when we consider the fitting is ETS.

- **Point 2:** About the assumptions for time series, the residuals autocorrelation $p$-value $> 0.05$ is essential. In the Tables 2–4, we can count 9 autocorrelated issues for ARIMA model, none for ETS, and none for MMN. The residuals do not follow a normal distribution ($p$-value $< 0.00$) and this was

| TDS       | 2 Weeks MAPE | ETS | AC  | Norm | Hetero | Stat |
|-----------|--------------|-----|-----|------|--------|------|
| From 1 to 30 | 12.37        | A,A,N | 0.45 | 0.00  | 0.35   | 0.49 |
| From 1 to 44 | 20.71        | A,A,N | 0.44 | 0.00  | 0.97   | 0.17 |
| From 1 to 58 | 23.48        | A,A,N | 0.86 | 0.00  | 0.96   | 0.10 |
| From 1 to 72 | 2.45         | A,A,N | 0.90 | 0.00  | 0.61   | 0.07 |
| From 1 to 86 | 1.98         | A,A,N | 0.68 | 0.00  | 0.04   | 0.01 |
| From 1 to 100 | 0.45         | A,A,N | 0.73 | 0.00  | 0.13   | 0.01 |
| From 1 to 114 | 0.64         | A,A,N | 0.68 | 0.00  | 0.29   | 0.01 |
| From 1 to 128 | 0.99         | A,A,N | 0.70 | 0.00  | 0.43   | 0.01 |
| From 1 to 142 | 0.61         | A,A,N | 0.80 | 0.00  | 0.49   | 0.01 |
| From 1 to 156 | 0.86         | A,A,N | 0.52 | 0.00  | 0.07   | 0.01 |
| From 1 to 170 | 0.25         | A,A,N | 0.46 | 0.00  | 0.01   | 0.01 |
| From 1 to 184 | 1.12         | A,A,N | 0.99 | 0.00  | 0.01   | 0.01 |
| From 1 to 198 | 0.98         | A,A,N | 0.90 | 0.00  | 0.00   | 0.01 |
| From 1 to 212 | 0.55         | A,A,N | 0.35 | 0.00  | 0.00   | 0.01 |
| From 1 to 226 | 0.30         | A,A,N | 0.58 | 0.00  | 0.00   | 0.01 |
| From 1 to 240 | 0.52         | A,A,N | 0.24 | 0.00  | 0.00   | 0.01 |
| From 1 to 254 | 0.15         | A,A,N | 0.22 | 0.00  | 0.00   | 0.01 |
| From 1 to 268 | 0.30         | A,A,N | 0.38 | 0.00  | 0.00   | 0.01 |
| From 1 to 282 | 0.16         | A,A,N | 0.97 | 0.00  | 0.00   | 0.01 |

**Note:** TDS: Training dataset, AC: Autocorrelation, Norm: Normality, Hetero: Heteroscedasticity, Stat: Stationarity.

| TDS       | 2 Weeks MAPE | MMN   | AC | Norm | Hetero | Stat |
|-----------|--------------|-------|----|------|--------|------|
| From 1 to 30 | 4.39         | M,M,N | 0.57 | 0.00  | 0.40   | 0.13 |
| From 1 to 44 | 17.01        | M,M,N | 0.53 | 0.00  | 0.02   | 0.04 |
| From 1 to 58 | 6.43         | M,M,N | 0.57 | 0.00  | 0.00   | 0.01 |
| From 1 to 72 | 27.83        | M,M,N | 0.54 | 0.00  | 0.00   | 0.01 |
| From 1 to 86 | 9.83         | M,M,N | 0.55 | 0.00  | 0.00   | 0.01 |
| From 1 to 100 | 2.30         | M,M,N | 0.02 | 0.00  | 0.00   | 0.01 |
| From 1 to 114 | 0.79         | M,M,N | 0.44 | 0.00  | 0.01   | 0.01 |
| From 1 to 128 | 0.33         | M,M,N | 0.51 | 0.00  | 0.00   | 0.01 |
| From 1 to 142 | 0.61         | M,M,N | 0.49 | 0.00  | 0.00   | 0.01 |
| From 1 to 156 | 0.33         | M,M,N | 0.46 | 0.00  | 0.00   | 0.01 |
| From 1 to 170 | 0.89         | M,M,N | 0.49 | 0.00  | 0.00   | 0.01 |
| From 1 to 184 | 1.77         | M,M,N | 0.36 | 0.00  | 0.01   | 0.01 |
| From 1 to 198 | 1.37         | M,M,N | 0.36 | 0.00  | 0.00   | 0.01 |
| From 1 to 212 | 0.95         | M,M,N | 0.34 | 0.00  | 0.00   | 0.01 |
| From 1 to 226 | 0.64         | M,M,N | 0.32 | 0.00  | 0.00   | 0.01 |
| From 1 to 240 | 0.63         | M,M,N | 0.33 | 0.00  | 0.00   | 0.01 |
| From 1 to 254 | 0.11         | M,M,N | 0.38 | 0.00  | 0.00   | 0.01 |
| From 1 to 268 | 0.37         | M,M,N | 0.36 | 0.00  | 0.00   | 0.01 |
| From 1 to 282 | 0.13         | M,M,N | 0.34 | 0.00  | 0.00   | 0.01 |

**Note:** TDS: Training dataset, AC: Autocorrelation, Norm: Normality, Hetero: Heteroscedasticity, Stat: Stationarity.
predictable because the dependent variable is a cumulated variable. Besides, for the models, we got stationary residuals with higher TDS and this result is also understandable because with great sample size the variability becomes stationary for the model. MMN and ETS models have most of their residuals that are heteroscedastic ($p$-value $< 0.05$), but ARIMA does not. Actually, the presence of heteroscedasticity in ETS is a point that shows the significant change among the daily numbers of cases about the pandemic. In the context of the current work, it is not important as the autocorrelation of residuals.

The rule of thumb is to keep the best model and after considering the Point 1 and Point 2, now we can also analyze the statistics in Table 5. It is easy to notice that ETS has the smallest median, mean, and maximum. About the minimum, it has the value (0.15) that is the greatest value. Over the four parameters, the model ETS got 3 good MAPE statistics (mean, median, maximum), while ARIMA got just 1 (minimum), and MMN none. Considering those points, we think that from the best choice to the least one, we have ETS, ARIMA, and MMN. Now, as we have chosen the best model, we can now visualize how it works and how to use it for a good forecast. Let us note that in Fig. 3; we should normally have 19 images (the number of TDS), but in terms of commodities (numerous images), we decided to show the beginning A (January-March), B (January-April), C (January-June), D (January-July), E (January-October), and F (January-November). It helps to visualize the good fitting of our selected model. When we look at the Fig. 3, we cannot trust the forecasts in Fig. A because the real data (black line) is in a very big confidence interval. The uncertainty bounds are very big due to the small sample size. However, the remaining graphs (B, C, D, E, and F) showed better forecasts of the model. Considering the MAPE smaller than 1% in the Table 3, we can focus on TDS having at least 100 days with 2 weeks as the period of forecast (PF). Let’s look at Fig. 4 that is the evidence of having MAPE smaller than 1% after the third training period (From 1

| Model  | Min | Mean | Median | Max  |
|--------|-----|------|--------|-----|
| ARIMA  | 0.10| 3.65 | 0.72   | 23.75|
| ETS    | 0.15| 3.62 | 0.64   | 23.48|
| MMN    | 0.11| 4.04 | 0.89   | 27.83|

Table 5 Mimima, means, and maxima of 2-Weeks MAPEs.

Fig. 3 Different forecasts of the kept ETS (A,A,N) model using each training data set.
to 58). In the case of this paper, we have 314 days in total and the aforementioned checking makes us compute the final model estimates and illustrate the forecasts in Fig. 5.

3.2. Discussion

Daily decisions on COVID-19 have been influencing the spread of the pandemic and an adapted forecasting tool is required for better policy. Many studies have been proposing multi-or univariate models to forecast COVID-19 cases, but most of them failed to predict well the upcoming situation [27]. This study checked among the propositions, the one that is the most appropriate concerning daily realities about COVID-19 in the world. The quality of collected data about the pandemic is still debatable for some countries [14,15,5]; and they are not also exhaustive due to confinement during outbreak periods. To reduce at most as possible the bias in our model, we proposed to avoid complex models due to the issues with current data and focused on univariate time series modeling. Considering the principle of “garbage-in, garbage-out”, using one time series analysis and having good forecasts is advisable.

Although we hypothesized that ETS model might be well adapted because of its capacity to vary in 30 different models and as much adaptive as possible in terms of COVID-19 evolution, we compared classical univariate time series models. Among them, the best one is ETS model because it respects the residuals autocorrelation assumption and has the smallest MAPE in the Table 5. The study in Nepal [17] used 99 days and found ARIMA (MAPE = 4.18) and ETS (M,A,N) (MAPE = 4.55) for 2 weeks forecast. In our case, the MAPE (with 100 days) of ARIMA is 0.33 and ETS (A,A,N) is 0.45, and we can notice that the trend estimate for Nepal is also additive and the difference is about the error that is multiplicative and it can be related to the fact that we are working with the world data. This current study and the one in Nepal are similar in terms of trend type, while the study [6] with at most 50 days of training data set proposed a multiplicative trend. In the Table 3, it is easy to notice that we only got an additive trend and this was our hypothesis. In addition, we used a cross-validation technique and in the work [17], they did not. This point might explain their finding because our process is more robust.

Actually, short-term forecasts (2 weeks) are advised to maintain short-term forecasts because in numerous studies about COVID-19 forecasts [6,17,28,29], the authors proposed a similar forecast period 10–14 days. Although we advise to keep 2 weeks for a forecast and repeat it if needed, people can still consider 3 weeks because it also gave good results in many other works [30–32]. The best forecasts in our model are from the training data having at least 100 days and it is understandable when you look at the Fig. 3. Actually, it has two parts, one that is flat and another one that shows a high

![Fig. 4 MAPE per training period.](image)

![Fig. 5 Best model ETS (A,A,N) actual, fitted, and predicted values.](image)
increase of cases. Especially for TDS (1–58), there are high MAPEs due to the fact that the train data set is at the transition part of the change between the flat part and the high increase part. The forecast model should take into account both parts because in March, new cases started duplicating (becoming additive with time) compared to the past number of cases.

3.3. Conclusion

Time series forecasts of COVID-19 cases became very useful worldwide. Many authorities are concerned to control COVID-19 and it is important to know the future trend of the pandemic. Using multivariable models is labelled with higher bias due to the issue with that data collect. We compared univariate time series (ARIMA, ETS, MMN) models with cross-validation method and found out that ETS is the best one because it got the smallest bias in terms of forecasting. When we assume that future decisions will follow the past structure, we think the best model to forecast COVID-19 cumulative cases in the world is an ETS with additive error and trend without any season. In real life, the international institutions will just need to use ETS models directly or implement with a software to forecast cumulative cases to adjust a policy that affects the worldwide aspects of the pandemic.

In addition, the main limitation of this work is quite related to the predictions of the world figures about COVID-19 because they are aggregated and heterogeneous data sets. These remarks have also been mentioned in [27,33] because this kind of work does not take into account particular changes in small or big countries. However, the final decision with short-term forecasts can help in the decrease of bias in this study and help international institutions to adjust decisions about the pandemic.

The perspective related to this study is to work on each continent and country data to check the robustness of ETS among univariate time series models. It is going to be an essential work that could help each country to get an appropriate tool regarding the pandemic.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to the lecturers at the International Chair in Mathematical Physics and Applications (ICMPA - UNESCO Chair), who were quite helpful in the improvement of this paper. In addition, we are grateful to Mr. Innocent Koovi who have helped to proofread this paper.
Appendix A.

R Script

# You need to set a working directory
#The one down is an exemple of mine
#setwd("/DATA/Article/Submissions")

library(ggplot2)
library(tseries)
library(forecast)
library(lubridate)
library(ggthemes)
library(stargazer)
library(reshape)
library(MLmetrics)
library(data.table)
library(lme4)
library(lmtest)
library("readxl")

#You need to import the data in you working folder
#the source is
#https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.xlsx
#Here is an example of my data import
#
world=read_excel("doto.xlsx") #nouvelle version
world$Days <- seq(as.Date("2020-01-22"), as.Date("2020-11-30"), by=1)

cases <- ts(world$Cases, start = 1, frequency = 40)
world$Cases <- ts(world$Cases, start = 1, frequency = 40)

ggplot(world, aes(x=Days, y=Cases)) +
  geom_point(color="darkred") +
theme_tufte()+theme_bw()

# ARIMA MODELING

days = seq(30, length(world$Days), by=14)

base2 = data.frame(matrix(NA, ncol=10, nrow = length(days)))
colnames(base2) = c("Days", "1 Week MAPE", "2 Weeks MAPE", "3 Weeks MAPE", "ARIMA", "Auto Correlation", "Normality", "Heteroscedasticity", "Stationarity", "AICc")
cases <- ts(world$Cases, start =1)
z = 0

for (i in days[-c(20,21)]) {
  z = z + 1

  fit = auto.arima(world$Cases[1:i])
  fc7 = forecast(fit, 7)
  fc14 = forecast(fit, 14)
  fc21 = forecast(fit, 21)

  MAPE7 = MAPE(as.numeric(fc7$mean), world$Cases[(i+1):(i+7)]) * 100
  MAPE14 = MAPE(as.numeric(fc14$mean), world$Cases[(i+1):(i+14)]) * 100
  MAPE21 = MAPE(as.numeric(fc21$mean), world$Cases[(i+1):(i+21)]) * 100
  Auto = Box.test(fit$residuals, type = "Ljung-Box")
  Normality = shapiro.test(fit$residuals)
  hetero = bp.test((fit$residuals)^2+1+c(1:i))
  Stat = adf.test(fit$residuals)

  base2[z,1] = paste("From 1 to ", i, sep = " ")
  base2[z,2] = MAPE7
  base2[z,3] = MAPE14
  base2[z,4] = MAPE21
  base2[z,5] = paste(arimaorder(fit), collapse = "")
  base2[z,6] = Auto$p.value
  base2[z,7] = Normality$p.value
  base2[z,8] = hetero$p.value
Overview and cross-validation of COVID-19

```r
base2[z, 9] = Stats$ p.value
base2[z, 10] = fit$ aicc

}

base2 = base2[-c(20, 21),]
base2

# ETS MODELING

days = seq(30, length(world$Days), by = 14)
base = data.frame(matrix(NA, ncol = 10, nrow = length(days)))
colnames(base) = c("Days", "1Week MAPE", "2Weeks MAPE", "3Weeks MAPE", "ETS",
"Auto Correlation", "Normality", "Heteroscedasticity", "Stationarity", "AICc")
cases <- ts(world$Cases, start = 1)

z = 0

for (i in days[-c(20, 21)]) {
  z = z + 1
  caset = world$Cases[1:i]
  fit = ets(caset)

  fc7 = forecast(fit, h = 7)
  fc14 = forecast(fit, h = 14)
  fc21 = forecast(fit, h = 21)

  MAPE7 = MAPE(as.numeric(fc7$mean), world$Cases[(i + 1):(i + 7)]) * 100
  MAPE14 = MAPE(as.numeric(fc14$mean), world$Cases[(i + 1):(i + 14)]) * 100
  MAPE21 = MAPE(as.numeric(fc21$mean), world$Cases[(i + 1):(i + 21)]) * 100
  Auto = Box.test(fit$residuals, type = "Ljung-Box")
  Normality = shapiro.test(fit$residuals)
  hetero = bptest((fit$residuals)^2 + c(1:i))
  Stat = adf.test(fit$residuals)

  base[z, 1] = paste("From 1 to ", i , sep = " ")
  base[z, 2] = MAPE7
```
base[3,3]=MAPE14
base[3,4]=MAPE21
base[3,5]=paste(f21$model$components[1], f21$model$components[2],
              f21$model$components[3], sep = " ", )
base[3,6]=Auto$p.value
base[3,7]=Normality$p.value
base[3,8]=hetero$p.value
base[3,9]=Stat$p.value
base[3,10]=fit$aicc
}
base=base[-c(20,21),]
base

#MAN MODELING

days=seq(30, length(world$Days), by=14)
base3=data.frame(matrix(NA, ncol=10,nrow = length(days)))
colnames(base3)=c("Days","1 Week MAPE","2 Weeks MAPE","3 Weeks MAPE", "MMN", "Auto Correlation","Normality","Heteroscedasticity","Stationarity","AICc")
cases<-ts(world$Cases, start =1)
z=0

for (i in days[-c(20,21)] ) {
    z=z+1
    case=world$Cases[1:i]
    fit=ets(case, model="MMN",damped = F)
    fc7= forecast(fit,h=7)
    fc14= forecast(fit,h=14)
    fc21= forecast(fit,h=21)

    MAPE7=MAPE(as.numeric(fc7$mean), world$Cases[(i+1):(i+7)])*100
    MAPE14=MAPE(as.numeric(fc14$mean), world$Cases[(i+1):(i+14)])*100
    MAPE21=MAPE(as.numeric(fc21$mean), world$Cases[(i+1):(i+21)])*100
    Auto=Box.test(fit$residuals, type = "Ljung-Box")
Overview and cross-validation of COVID-19

Normality = shapiro.test(fit$ residuals)
hetero = bptest((fit$residuals)^2 ~ 1 + c(1:i))
Stat = adf.test(fit$residuals)

base3[z,1] = paste("From 1 to ", i, sep = " ")
base3[z,2] = MAPE7
base3[z,3] = MAPE14
base3[z,4] = MAPE21
base3[z,5] = paste(fcv21$model$components[1], fcv21$model$components[2],
                   fcv21$model$components[3], sep = " ", )
base3[z,6] = Auto$p.value
base3[z,7] = Normality$p.value
base3[z,8] = hetero$p.value
base3[z,9] = Stat$p.value
base3[z,10] = fit$aicc

base3 = base3[-c(20, 21), ]
base3

# New dataframe
neww = data.frame(matrix(NA, ncol = 3, nrow = 19))

neww$X1 = base$Days
neww$X2 = base$'2 Weeks MAPE'
colnames(neww) = c("Training Period", "MAPE", "Forecast Period")
neww$'Training Period' = factor(neww$'Training Period'
neww$'Forecast Period' = factor(neww$'Forecast Period'
levels(neww$'Training Period'

neww$'Training Period' = ordered(neww$'Training Period'
levels = c("From 1 to 30", "From 1 to 44", "From 1 to 58", "From 1 to 72", "From 1 to 86", "From 1 to 100", "From 1 to 114", "From 1 to 128", "From 1 to 142", "From 1 to 156", "From 1 to 170", "From 1 to 184", "From 1 to 198", "From 1 to 212", "From 1 to 226", "From 1 to 240", "From 1 to
240", "From 1 to 254", "From 1 to 268", "From 1 to 282")

ggplot(neww, aes('Forecast Period', 'Training Period', fill = MAPE)) +
  geom_tile(colour = "white") +
  scale_fill_gradient(low="lightblue", high="blue")+theme_bw()

#################################################

# You need to change J value by 1, 4, 7, 11, 14, 17, 20
j=19 #1, 4, 7, 11, 14, 17, 20
world$Days[1:(days[j]+21)]

caset=world$Cases[1:(days[j])]
fit=(ets(caset))
fc21= forecast(fit,h=21)
base=data.frame(matrix(NA, ncol = 5, nrow = (days[j]+21)))
colnames(base)=c("Days", "Cases", "Forecast", "F95", "F105")

base$Days=world$Days[1:(days[j]+21)]
base$Cases[1:(days[j]+21)] =world$Cases[1:(days[j]+21)]
base$Forecast[(days[j]+1):(days[j]+21)]=fc21$mean
base$F95[(days[j]+1):(days[j]+21)]=fc21$lower [,2]
base$F105[(days[j]+1):(days[j]+21)]=fc21$upper [,2]

# Here it is sept and the value of J,
# It means when j=4, Sept. should updated into Sept4

sept20=ggplot(base, aes(x=Days)) + theme_bw()+
  geom_line(aes(y = Cases), color = "black", lwd=1.2) +
  geom_line(aes(y = Forecast), color="blue", linetype="twodash", lwd=1.05)+
  geom_line(aes(y = F95), color="blue", linetype="twodash", lwd=1.05)+
  geom_line(aes(y = F105), color="blue", linetype="twodash", lwd=1.05)+
  annotate("rect", xmin =world$Days[(days[j]+1)] , xmax = world$Days[(days[j]+21)],
    ymin = -Inf, ymax = +Inf, alpha = 0.2, fill = "grey")+
annotate("rect", xmin = world$Days[(days[j]+1)], xmax = world$Days[(days[j]+21)],
        ymin = -Inf, ymax = +Inf, alpha = 0.25, fill = "grey")

library(cowplot) #1,4,7,11,14,17, 20
plot_grid(sept1, sept4, sept7, sept11, sept14, sept17, sept20,
          labels=c("A", "B", "C", "D", "E", "F"), ncol = 2, nrow = 3)

###################################################################

fit=auto.arima(world$Cases) ##Arima
fit=ets(world$Cases, model="MMN",damped = F)
fit=ets(world$Cases)

forc=forecast(fit,h=14)

Final=data.frame(matrix(NA, ncol=5,nrow = (2*length(fit$fitted)+14)))
colnames(Final)=c("Cumulative number of daily confirmed cases of COVID-19",
"Data/Fitted/Forecast","Days","LowerIC","UpperIC")

world$date
Final$'Cumulative number of daily confirmed cases of COVID-19'
[1:length(fit$fitted)]=world$Cases
Final$Days[1:length(fit$fitted)]=paste(as.Date("2020-01-22"),
as.Date("2020-11-30"), 1))

Final$'Data/Fitted/Forecast'[1:length(fit$fitted)]=
  "Actual values"

Final$'Cumulative number of daily confirmed cases of COVID-19'
[(length(fit$fitted)+1):(2*length(fit$fitted))]=fit$fitted

Final$'Data/Fitted/Forecast'
[(length(fit$fitted)+1):(2*length(fit$fitted))]="Fitted values"
Final$Days[(length(fit$fitted)+1):(2*length(fit$fitted))]=
paste(seq(as.Date("2020-01-22"), as.Date("2020-11-30"), 1))
References

[1] World Health Organization, et al., Statement on the second meeting of the international health regulations, Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV).[Internet].

[2] B.F. Maier, D. Brockmann, Effective containment explains subexponential growth in recent confirmed covid-19 cases in china, Science 368 (6492) (2020) 742–746.

[3] H. Lau, V. Khorawipour, P. Kochbach, A. Mikolajczyk, H. Ichii, J. Schubert, J. Bania, T. Khorawipour, Internationally lost covid-19 cases, J. Microbio. Immunol. Infect. (2020).

[4] L. Wynants, B. Van Calster, M.M. Bonten, G.S. Collins, T.P. Debray, M. De Vos, M.C. Haller, G. Heinez, K.G. Moons, R.D. Riley, et al, Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal, bmj 369 (2020).

[5] C. Anastassopoulou, L. Russo, A. Tsakris, C. Siettos, Data-based analysis, modelling and forecasting of the covid-19 outbreak, PloS One 15 (3) (2020) e0230405.

[6] F. Petropoulos, S. Makridakis, Forecasting the novel coronavirus covid-19, PloS One 15 (3) (2020) e0231236.

[7] M. KoC¸ AK, A comparison of time-series models in predicting covid-19 cases, Turkiye Klinikleri J. Biostat. 12 (1) (2020).

[8] D. Benvenuto, M. Giovannetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the arima model on the covid-2019 epidemic dataset, Data in brief (2020) 105340.

[9] N. Tabassum, Z. Hussain, Mixtue of two generalized inverted exponential distributions with censored sample: properties and estimation, Statistica Applicata-Italian J. Appl. Stat. 30 (3) (2018) 373–391.

[10] T.N. Sindhu, Z. Hussain, M. Aslam, On the bayesian analysis of censored mixture of two topp-leone distribution, Sri Lankan J. Appl. Stat. 19 (1) (2019).

[11] T.N. Sindhu, Z. Hussain, Predictive inference and parameter estimation from the half-normal distribution for the left censored data, Ann. Data Sci. (2020) 1–15.

[12] T.N. Sindhu, A. Shaﬁq, Q.M. Al-Mdallal, On the analysis of number of deaths due to covid-19 outbreak data using a new class of distributions, Results Phys. 21 (2021) 103747.

[13] T.N. Sindhu, A. Shaﬁq, Q.M. Al-Mdallal, Exponentiated transformation of gumbel type-II distribution for modeling covid-19 data, Alexandria Eng. J. 60 (1) (2021) 671–689.

[14] World Health Organization, Coronavirus disease (covid-2019) situation reports, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports, accessed: 2020-05-26 (2020).

[15] Coronavirus (COVID-19) Data Hub, Covid-19 activity.csv, https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=Covid-19+Activity.csv, accessed: 2020-06-08 (2020).

[16] World Health Organization, Global surveillance for human infection with coronavirus disease (covid-19), https://www.who.int/publications-detail/global-surveillance-for-human-infection-with-novel-coronavirus-(2019-ncov), accessed: 2020-05-26 (2020).

[17] S. Bhandary, S.L. Shrestha, R.P. Khatiwada, D.N. Shah, N.N. Munankarmi, M.R. Banjara, R. Thapa-Parajuli, K.D. Manandhar, R. Adhikari, R. Tuladhar, Trend analysis, modelling and impact assessment of covid-19 in nepal, MedRxiv.

[18] E.O.-O. Max Roser, Hannah Ritchie, J. Hasell, Coronavirus pandemic (covid-19), Our World in Data. https://ourworldindata.org/coronavirus,

[19] C.A. Jolipasi et al, Selection for the best ets (error, trend, seasonal) model to forecast weather in the aech besar district, IOP Conference Series: Materials Science and Engineering, vol. 352, IOP Publishing, 2018, p. 012055.

[20] E.S. Gardner Jr, E. McKenzie, Forecasting trends in time series, Manage. Sci. 31 (10) (1985) 1237–1246.

[21] R.J. Hyndman, Y. Khandakar, Automatic time series forecasting: the forecast package for R, J. Stat. Softw. 26 (3) (2008) 1–22, URL http://www.jstatsoft.org/article/view/v027i03.

[22] S.A. Lauer, K.H. Grantz, Q. Bi, F.K. Jones, Q. Zheng, H.R. Meredith, A.S. Azman, N.G. Reich, J. Lessler, The incubation period of coronavirus disease 2019 (covid-19) from publicly reported cases: estimation and application, Annals Internal Med. 172 (9) (2020) 577–582.

[23] Y.-H. Jin, L. Cai, Z.-S. Cheng, H. Cheng, T. Deng, Y.-P. Fan, C. Fang, D. Huang, L.-Q. Huang, Q. Huang, et al, A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-ncov) infected pneumonia (standard version), Military Medical Res. 7 (1) (2020) 4.

[24] N.M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A.R. Akhmetzhanov, S.-M. Jung, B. Yuan, R. Kinoshita, H. Nishiura, Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data, J. Clin. Med. 9 (2) (2020) 538.

[25] C. for Disease Control, Prevention.,et al., Interim clinical guidance for management of patients with confirmed coronavirus disease (covid-19).

[26] RStudio Team, Rstudio: Integrated development environment for r, http://www.rstudio.com/ (2020).

[27] N.P. Jewell, J.A. Lewnard, B.L. Jewell, Predictive mathematical models of the covid-19 pandemic: underlying principles and value of projections, Jama 323 (19) (2020) 1893–1894.

[28] K. Roosa, Y. Lee, R. Luo, A. Kripich, R. Rothenberg, J.M. Hyman, P. Yan, G. Chowell, Short-term forecasts of the covid-19 epidemic in guangdong and zhejiang, china: February 13–23, 2020, J. Clinical Med. 9 (2) (2020) 596.

[29] M.A. Al-qaaness, A.A. Ewees, H. Fan, M. Abd El Aziz, Optimization method for forecasting confirmed cases of covid-19 in china, J. Clin. Med. 9 (3) (2020) 674.

[30] I. COVID, C.J. Murray, et al,., Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months, MedRxiv, 2020.

[31] T. Kuniya, Prediction of the epidemic peak of coronavirus disease in Japan, 2020, J. Clin. Med. 9 (3) (2020) 789.

[32] M. Batista, Estimation of the final size of the covid-19 epidemic, MedRxiv. doi 10 (2020) 16–20023606.

[33] Y.M. Sokadjo, M.N. Atchadé, The influence of passenger air traffic on the spread of covid-19 in the world, Transport. Res. Interdiscip. Perspect. (2020) 100213.