Short-term prediction model of the number of COVID19 cases: Case of Senegal

Ousseynou Mbaye and Siriman Konare

Université Alioune Diop de Bambey, Bambey, Senegal
Universite Gaston Berger de Saint Louis
ousseynou.mbaye@uadb.edu.sn
konare.siriman@ugb.edu.sn

Abstract. Background COVID-19 is an infectious respiratory disease caused by the last coronavirus discovered. This new virus was unknown before the outbreak began in Wuhan, China. In December 2019, this disease very quickly became a public health problem around the world. Thus it has become necessary to see even urgent to find means of prevention but also of prediction to fight effectively against this disease.
Methods: We use time series with Holt's smoothing method that we modify by adding in input a logarithm function for a good modeling of the starting values.
Results and conclusions: The different methods used give striking results, in particular the one with linear tend. By experimenting with our model on data sets from other countries such as France, Italy, China or even countries bordering Sengal such as Mali, we obtain practically the same performances. This proves that the method can be generalized to all countries in the world where the pandemic exists.

Keywords: COVID19 · Times Series · Prediction Model, Holt method, forcast.

1 Introduction

In December 2019, a cluster of pneumonia cases, caused by a newly identified -coronavirus, occurred in Wuhan, China. This very virulent virus has taken on a global scale, causing thousands of deaths in the space of a month. Hot spots have appeared in Asia, Europe, America, Africa to tell you that the level of international spread is alarming. This coronavirus was initially named coronavirus of 2019 (2019-nCoV) on January 12, 2020 by the World Health Organization (WHO) [2]. WHO has officially named the disease coronavirus 2019 (COVID-19) and the Coronavirus Study Group (CSG) of the International Committee has proposed to name the novel coronavirus SARS-CoV-2, both published on February 11, 2020. This monster will end up being declared as a pandemic given the number of deaths and its speed of spread. Today in these uncertain times it doesn't take a genius to recognize that this pandemic puts all societies in uncomfortable situations.
This COVID-19 crisis affects us at different levels. Adaptability, psychological risk, social functioning, daily management or projection into the future are put to the test. We must therefore better understand it to allow us to prepare for new completely unforeseen episodes and even more to avoid their occurrence, thanks to changes in human

paradigms. Faced with this situation and aware of the role that falls to the intellectual in the process of transformation of our societies, all scientists have put themselves at the service. By examining the situation, they try to find models that best describe the spread, the mode of contamination, the speed of contamination but also to map the patients to better define the people in contact with the confirmed patients. Indeed, the question of knowing what the need for statistics arises, especially when the situation deteriorates day by day with cries of alarm from all sides.

So this work that we present to you is a scientific study of the dynamism of the pandemic at the intercontinental and continental level and in particular in Senegal. The main objective of this study is to predict the cumulative number of cases of the disease based on historical data. Its use will be purely for statistical results intended for governments. The data is reliable and in any case the objectivity will not be shaken in this prediction and can serve as a very good basis for decision-makers.

We will use time series with Holt's smoothing method that we modify by adding a logarithm function as input for a good modeling of the starting values. In fact, screening tests are carried out daily and therefore data updates are carried out every day in all the countries affected by the crisis.

The rest of the article is organized as follows. In section 2, we present some results relating to time series but also to COVID19, followed by mathematical tools and methods in section 3. The forecasting model that we proposed is detailed in section 4 and the results of our experiment and their interpretations in section 5. Finally, in section 6, we draw conclusions.

2 Related work

Since the confirmation of the first case of COVID-19 in Senegal on March 02, the number of infected people continues to increase day by day. The spread of the virus is very dangerous and requires a number of measures to be taken. It is therefore very important to anticipate confirmed cases in the coming days to implement tailor-made protection plans. To gain better visibility into the spread of the virus, many studies have been done to predict the number of cases or deaths of COVID-19. It is in this sense that Zhao et al.[8] proposed a mathematical model to estimate the actual number of COVID-19 cases in China in the first half of January 2020, which had gone unreported. They concluded that there were 469 unreported cases between January 1 and January 15, 2020. Karako K et al[5]. developed a stochastic transmission model by extending the SIR (Susceptible-Infected-Removed) epidemiological model with additional modeling of the individual action on the probability of staying away from crowded areas. In Iran, Zareie B et al[7]. used the SIR epidemiological model to estimate the number of COVID-19 cases. The analysis was done on data between January 22 and March 24 and the prediction was made until April 15. The authors have come to the following conclusion that approximately 29,000 people will be infected between March 25 and April 15. However in Senegal the authors of '[6] proposed a SIR epidemiological model combined with machine learning models to predict the course of the disease. Their results predicted the end of the pandemic in many countries by April at the latest. Time series are also often used in disease prediction tasks. Thus the authors of [4] used prophet to predict the number of covid cases in India, they observed that their fit model is accurate within a certain range, and extreme prevention and control measures are suggested in an effort to avoid such a situation. Alexandre Medeiros et al.[4] proposed a time series modify to study the incidence of the disease on mortality, then the authors of [2], in their Covid prediction study combined time series with neural networks through the LSTM algorithm.

3 Material and method

3.1 Simple exponential smoothing

Exponential smoothing[3] is actually a very simple forecasting technique at $t+\ 1$. It applies to time series without trend. The principle is to give more importance to the last observations. In other words, the more recent the observation, the greater the weight will be. For example, it is reasonable to assign higher weights to observations yesterday than to observations made 7 days ago. We do not extend a series as we would, for example, with a simple regression, but we try to obtain a smoothed value at t to simply transfer it to $t+1$.
It has only one component called level with a smoothing parameter denoted by $alpha$. This is a weighted average of the previous level and the current observation.
Mathematical formulation

$$y_{t+1} = \alpha * y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \alpha(1-\alpha)^3 y_{t-3} + ... \tag{1}$$

where $0 \leq \alpha \leq 1$ is the smoothing parameter.

The rate of weight reduction is controlled by the smoothing parameter $alpha$. If $alpha$ is large, more weight is given to more recent observations.
There are 2 extreme cases:

- $\alpha = 0$: the forecast of all future values is equal to the average of the historical data, which is called the Average method.
- $\alpha = 1$: just set all predictions to the value of the last observation, called the naive method.

3.2 Smoothing Holt's method

Holt[1] made an extension of simple exponential smoothing to allow forecasting of data with trends.

Holt's method involves a prediction equation and two smoothing equations
Mathematics formulation
Prediction equation:

$$y_{t+h=l_t+hb_t} \tag{2}$$

Smoothing level equation:

$$l_t = \alpha * y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{3}$$

Trend equation:

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{4}$$

0 $leq\ alpha\ leq$1 is the exponential smoothing parameter

0 $leq\ beta\ leq$1 is the smoothing parameter of the trend.

4 Proposed model

This section presents the proposed **Holt COVID** method. This is a chronological method for predicting confirmed cases of COVID-19, as shown in Figure 1 above. Consider the time series $X = (t_i, y_i)_{1\leq i\leq n}$ for n natural integer.

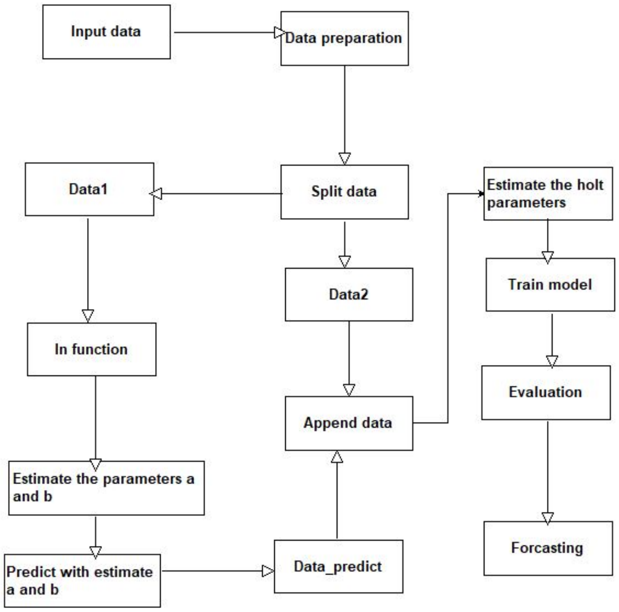


Fig. 1: Proposed model

Let $X = \{(t_1,y_1),(t_2,y_2),(t_3,y_3),(t_4,y_4),..........(t_n,y_n)\}$.
Let f the function defined from $\mathbb{N}^*$ to value in $\mathbb{R}^+$ defined by

$$f(t) = btln(1+at) \tag{5}$$

where t denotes the index of the date considered to the index i. The coefficients a and b satisfying the conditions:
$0 \leq a \leq 1$ and $0 \leq b \leq 1$ are parameters to be estimated.

Then for any $0 < i < k$ with k natural integer we define a new series X_i^k as follows: $X_i^k = (t_i, f(t_i)), 1 \leq i \leq k$.
To be able to make predictions with Holt's method we define a new series obtained from the series $X_i = (t_i, y_i)_{1 \leq i \leq n}$ and $X_i^k = (t_i, f(t_i))$ which we call X_2 in the following way:

$$X_2 = \{(t_1, f(t_1)), (t_2, f(t_2)), (t_3, f(t_3)), ...(t_k, f(t_1)), (t_k + 1, y_k + 1)..........(t_n, y_n)\} \tag{6}$$

By making the following change of variable:

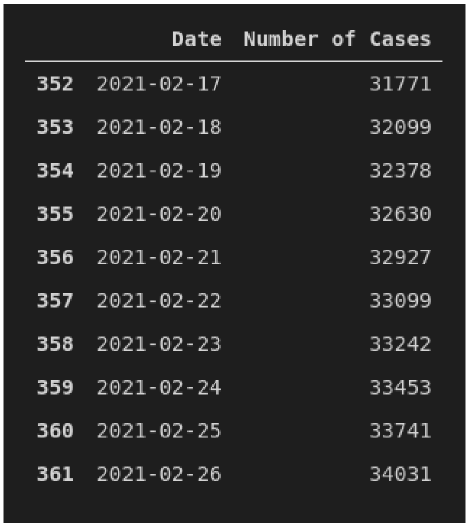
$$Z_i = f(t_i) for \quad 1 \leq i \leq k \quad and \quad Z_i = X_{k+i} \quad for \quad k+1 \leq i \leq n \tag{7}$$

we get the series $Z = (t_i, Z_i)_{1 \leq i \leq n}$ on which we will apply the different methods of **Holt**

5 Experimentation and results

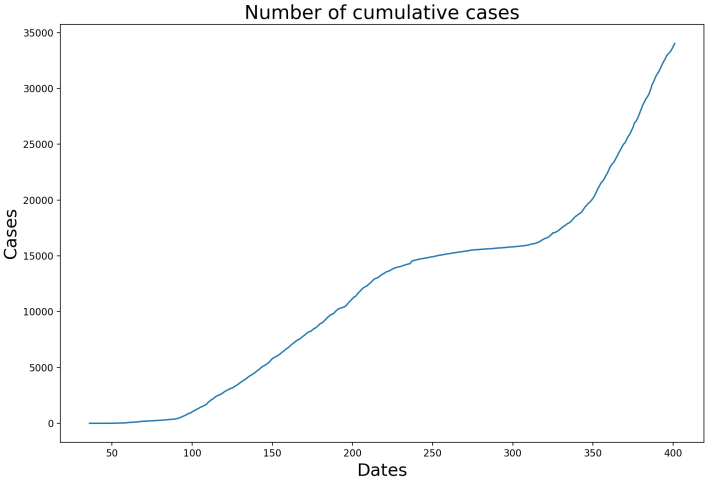
5.1 Description and data preparation

Descripion. The data used in this study comes mainly from daily reports from the Senegalese Ministry of Health and Social Action. We chose the cumulative number of confirmed cases as the target rather than the number of new cases per day. However, it should be noted that this target variable may be biased due to the fact that Senegal has decided not to carry out mass screening. That is, only patients with symptoms of covid19 are tested. The dataset (in figure 2 below) is a time series whose records are made daily between 02/03/2020 and 26/02/2021.



	Date	Number of Cases
352	2021-02-17	31771
353	2021-02-18	32099
354	2021-02-19	32378
355	2021-02-20	32630
356	2021-02-21	32927
357	2021-02-22	33099
358	2021-02-23	33242
359	2021-02-24	33453
360	2021-02-25	33741
361	2021-02-26	34031

(a) Cumulative number of cases in Senegal (last 10 days)



(b) Cumulative number of cases in Senega

Fig. 2: Dataset Presentaion

5.2 Performance measures

After having trained a chronological model it will have to be evaluated. Thus in the literature there are several performance measures intended to evaluate the accuracy of a given chronological model [1].
In this study we will use the measures of which we give the mathematical formulations below. Let Z_i be the real number of cases at the index date i and P_i the corresponding predicted value.

- Mean Square Error (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Z_i - P_i)^2 \tag{8}$$

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |Z_i - P_i| \tag{9}$$

The lower value of MSE and MAE refers to the best method.

5.3 Parameters choice

The choice of the parameters α and β remains a major problem for the **Holt** method. In fact, these parameters are generally very subjective and vary depending on the context of the study and / or the type of forecast desired.
In this work, the parameters a and b for the logarithmic part of the model as well as the α and β coefficients of the **Holt** method are estimated for values between 0 and 1 with a step of 0.1.
The figures below give for each of these couples the variations of the MSE. Thus we observe that the best pair of values of (a,b) is $(0.2,0.7)$ with an MSE equal to 0.835. However the values of α and β are $(0.9,0.3)$ with an MSE of 1634 for the linear trend, 1661 for the damped trend and of 1694 for the exponential trend. The values of a and b as well as those of α and β are used for the rest of this work.

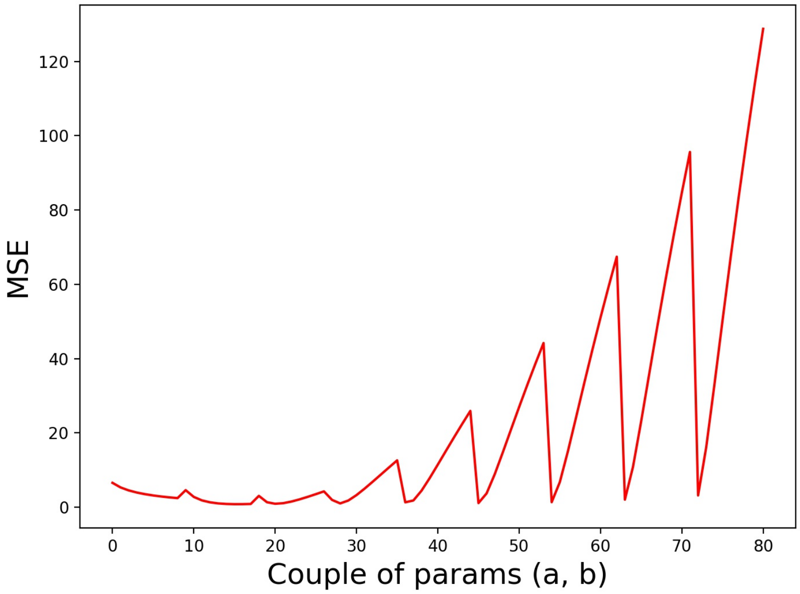
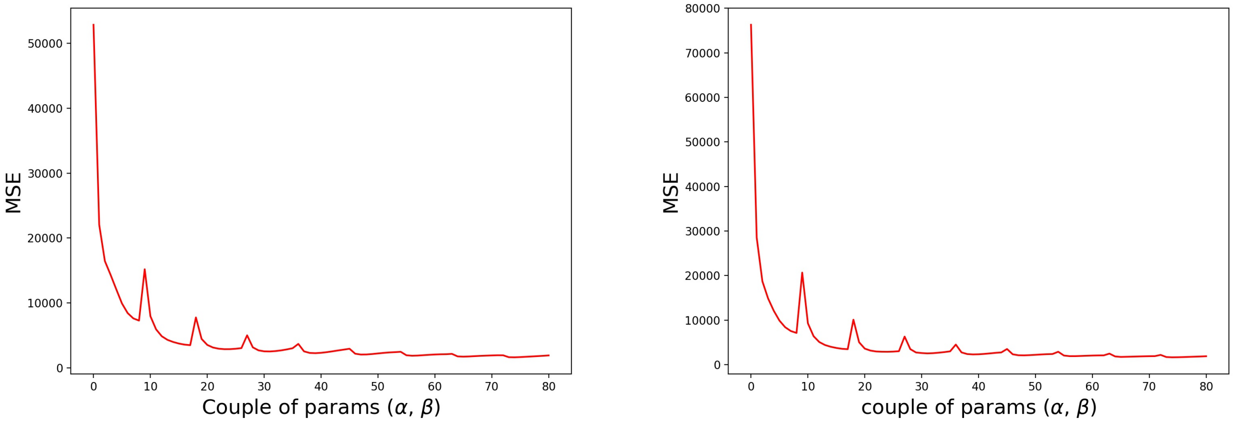


Fig. 3: MSE variations of log function



(a) MSE variations for linear trend

(b) MSE variations for dampened trend

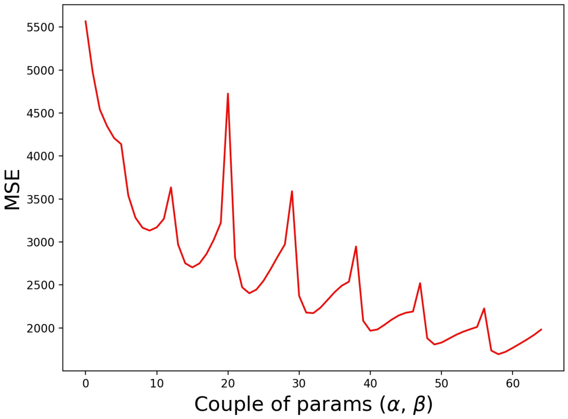


Fig. 4: MSE variations for exp trend

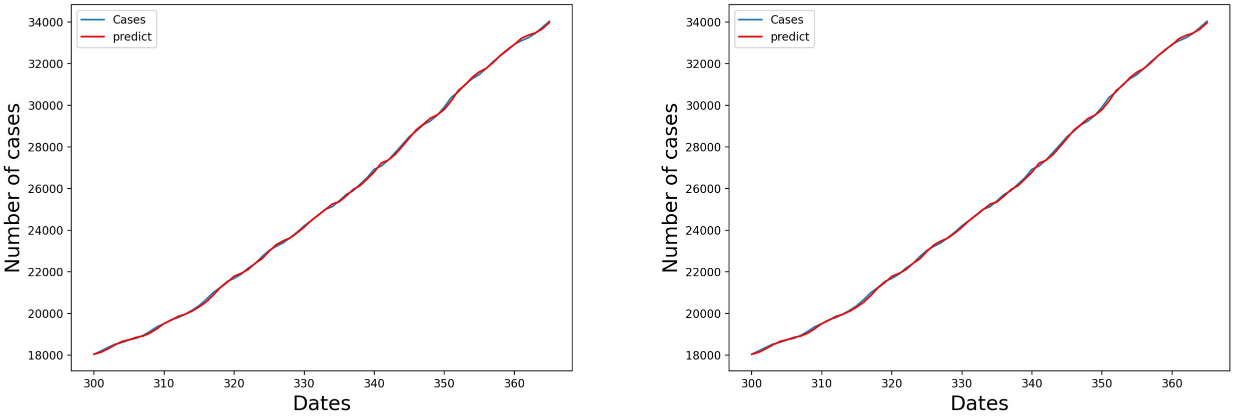
5.4 Presentation of results and Discussion

In this part we present the results for each of the variants of our model. Thus we calculate for each of them The **MSE**, and **MAE**.
Table 1 below gives the respective values of the MSE and the MAE for overall variants of modified Holt's method. Then the curves in Figures 4, 5 and 6 give a comparison between the actual values and the predicted values of the cumulative number of covid cases in Senegal. Looking closely at the results shows that the three methods are quite efficient for the task of predicting the cumulative number of cases of covid in Senegal.

Model	MSE	MAE
Holt whith linear trend	1634.66	25.93
Holt whith dumped trend	1661.52	26.29
Holt whith exponential trend	1694.13	26.50

Table 1: Measures of performance for whole Holt models

However, in term of MSE and MAE we noticed that the method with linear trend outperform the others significantly. Indeed that method present the smalest MSE and MAE wich are respectively 1634.66 and 25.93 . This indicates that our forecast differs, on average, by approximately 1634.66 from the actual data. This represents a difference of approximately 25.93 in absolute value from the actual data.



(a) Number of real cases and predicted(linear trend)

(b) Number of real cases vs predicted(dampened trend)

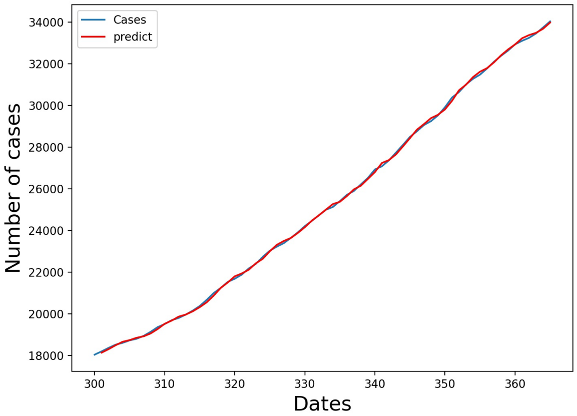


Fig. 5: Number of real cases vs predicted(exp trend)

6 Conclusion

In this paper we have studied Holt's methods in the task of predicting the cumulative number of Covid cases in Senegal. To do this we have introduced a logarithm function for a good modeling of the starting values. The results show our method give striking results in particular the one with linear tend.

By experimenting with our model on data sets from other countries such as France, Italy, China or even countries bordering Senegal such as Mali, we obtain practically the same performances. This proves that the method can be generalized to all countries in the world where the pandemic exists.

However, we also note that the method is very precise over short days, namely a period of 5 days which can be inconvenient.

Thus in perspective we will try to increase the performance of the model over longer durations.

References

1. Aragon, Y.: Lissage exponentiel. In: Séries temporelles avec R, pp. 121–132. Springer (2011)
2. Chimmula, V.K.R., Zhang, L.: Time series forecasting of covid-19 transmission in canada using lstm networks. Chaos, Solitons & Fractals **135**, 109864 (2020)
3. Dufour, J.M.: Lissage exponentiel. Université de Montréal (2002)
4. Indhuja, M., Sindhuja, P.: Prediction of covid-19 cases in india using prophet. International Journal of Statistics and Applied Mathematics (5), 4 (2020)
5. Karako, K., Song, P., Chen, Y., Tang, W.: Analysis of covid-19 infection spread in japan based on stochastic transition model. Bioscience trends (2020)
6. Ndiaye, B.M., Tendeng, L., Seck, D.: Analysis of the covid-19 pandemic by sir model and machine learning technics for forecasting. arXiv preprint arXiv:2004.01574 (2020)

7. Zareie, B., Roshani, A., Mansournia, M.A., Rasouli, M.A., Moradi, G.: A model for covid-19 prediction in iran based on china parameters. MedRxiv (2020)
8. Zhao, S., Lin, Q., Ran, J., Musa, S.S., Yang, G., Wang, W., Lou, Y., Gao, D., Yang, L., He, D., Wang, M.H.: Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. International Journal of Infectious Diseases **92**, 214–217 (2020). https://doi.org/https://doi.org/10.1016/j.ijid.2020.01.050, https://www.sciencedirect.com/science/article/pii/S1201971220300539