Short-Term Holt Smoothing Prediction Model of Daily COVID-19 Reported Cumulative Cases

Ousseynou Mbaye[1(✉)], Siriman Konare[2], Mouhamadou Lamine Ba[1], and Aba Diop[1]

[1] Université Alioune Diop de Bambey, Bambey, Senegal
{ousseynou.mbaye,mouhamadoulamine.ba,aba.diop}@uadb.edu.sn
[2] Université Gaston Berger de Saint Louis, Saint-Louis, Senegal
siriman.konare@ugb.edu.sn

Abstract. COVID-19 is the most deadly respiratory diseases worldwide known so far. It is a real public health problem against which contingency measures such as social distancing and lock down are often used to decrease the number of cases when it increases exponentially. These measures along with their impacts are set based on knowledge about the propagation of the disease, in particular the daily reported new and total cases within a given country. To plan in advance efficient contingency measures in order to stop its rapid propagation and to mitigate a possible explosion of the active cases leading to an uncontrolled situation and a saturation of health structures, governments need to have an indication about the potential number of total cases during incoming days; prediction models such as SIR algorithm try to provide such a kind of prediction. However, 'existing models like SIR are complex and consider many unrealistic parameters. This paper proposes, based on Holt's smoothing method combined with a logarithmic function for cold start, a very simple short-term prediction of the daily total number of COVID-19 cases. Our experimental evaluation over various COVID-19 real-world datasets from different countries show that our model, particularly using a linear trend function, gives results with low error rates. We also show that our approach can be generalized to all countries around the world.

Keywords: COVID-19 · Total cases · Prediction model holt smoothing · Trend · Evaluation performance

1 Introduction

In December 2019 [7], a cluster of pneumonia cases due to a newly identified coronavirus, referred to as COVID-19, has been discovered in Wuhan, a city in China. At that time, such an unknown disease showed a high human transmission rate leading to a rapid global propagation so that most of the countries worldwide have reported local or imported new cases within a very short time period.

Y. Faye et al. (Eds.): CNRIA 2021, LNICST 400, pp. 15–27, 2021.
https://doi.org/10.1007/978-3-030-90556-9_2

In few months, various hot spots have quickly appeared in Asia, Europe, America, and Africa pushing the experts of the World Health Organization (WHO) to declare COVID-19 as pandemic. Meanwhile they recommended governments to set contingency measures such as social distancing, systematic wearing of medical masks by citizens, mobility reduction or lock down in order to stop the increasing expansion of the virus. This rapid spread has been accompanied with thousands of deaths worldwide caused by the severe form of the disease with a high saturation of health structures in European countries in particular.

WHO has officially adopted Coronavirus 2019 (COVID-19 in short) as the official naming of this new respiratory disease in February 11, 2020 while at the same time the Coronavirus Study Group, an international committee that gathers several renowned health experts, has proposed a more scientific name which is SARS-CoV-2. This study aims at providing a prediction model of the daily total number of reported cases in a given country which information is critical in order to fight against the spread of the disease with suitable actions or to evaluate the impacts of already taken contingency measures. The number of new or cumulative total cases per day is one of the most important indicator that is monitored by governments and health organizations when dealing with pandemic disease. Till now, COVID-19 crisis is still ongoing despite the development of promising vaccines and still impacts highly the citizens of each country in all aspects of their daily life. Adaptability, psychological risk, social interactions, daily usual activities or projection in the future, economic activities are challenging for everyone. Various contingency measures such as social distancing, confinement and curfew are often used by governments to decrease the number of cases when it increases exponentially. These measures along with their impacts are set based on knowledge about the propagation of the disease, in particular the daily reported new and total cases within a given country. To plan in advance efficient contingency measures in order to stop its rapid propagation and to mitigate a possible explosion of the new cases leading to an uncontrolled situation and a saturation of health structures, governments need to have an indication about the potential number of new cases during incoming days. Several models have been proposed by researchers and scientists to describe the spread, the mode of contamination, the speed of contamination but also to trace back the movements of the population in order to easily find the persons in contact with the confirmed patients. However, 'existing prediction models, e.g. SIR model [6], which predict the number of new or total cases are complex and integrates sometimes unrealistic parameters.

The data about the total number of confirmed cases in a given country form a time series data and a prediction should be suitable for this kind of data. As a result, we introduce in this paper, based on Holt's smoothing method combined with a logarithm function for cold start, a very simple short-term prediction of the cumulative total number of confirmed COVID-19 cases every day. This also enables to deduce the number of daily new cases. The Holt algorithm is an extension of the exponential smoothing model for time series data with trends. In the case of COVID-19 the number of cases can increase or decrease depending on the country, the time period, the respect of social distance by citizens, etc.

To capture such an unknown evolution trend for a given country we evaluate with three different functions (linear, exponential and dampened) in our Holt's smoothing prediction approach. For the experimental evaluation we consider various real-world datasets from Senegal, Mali, France, USA, Australian, and Brazil. The choice of countries is far from being random. Indeed, the study is part of Senegal before being generalized to other countries such as Mali chosen on the basis of its proximity to Senegal. It was also necessary to be interested in what is happening in the European continent strongly impacted by COVID19, hence the choice of France. In addition the study was applied to the United States and Brazil to see its behavior in the American continent but also in Australia. As performance metrics, we measure the mean square error and mean absolute error of our model. Such an experimental evaluation over various COVID-19 real-world datasets from different countries show that our model, particularly using linear trend, gives results with low error rates. We also show that our approach can be generalized to all countries around the world.

The remaining of this study is organized as follows. First we review the related work in Sect. 2. We then introduce mathematical tools and methods used throughout this paper to build our proposed model in Sect. 3. Section 4 details the formalism of our prediction model. We present the results of the performance evaluation of our approach conducted on real-world datasets in Sect. 5 before concluding in Sect. 6 with some further research perspectives.

2 Related Work

Since the confirmation of the first case of COVID-19 in Senegal on March 02, 2020 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 during incoming days to implement suitable 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. [9] have proposed a mathematical model to estimate the actual number of officially unreported COVID-19 cases in China in the first half of January 2020. They concluded that there were 469 unreported cases between January 1 and January 15, 2020. Karako K et al. [5] have 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. [8] have also used the SIR epidemiological model to estimate the number of COVID-19 cases. The analysis has been done on data between January 22 and March 24, 2020 and the prediction was made until April 15, 2020. The authors have come to the conclusion that approximately 29,000 people will be infected between March 25 and April 15, 2020. In Senegal, the authors of [6] have proposed a SIR epidemiological model combined with machine learning models to predict the evolution of the disease. Their results predicted the end of the pandemic in many countries by April 2020 at the latest.

Time series are also often used in disease prediction tasks. Indeed, the authors of [4] used prophet to predict the number of COVID-19 cases in India. They have observed that their fitted model is accurate within a certain range, and extreme prevention and control measures have been suggested in an effort to avoid such a situation. Alexandre Medeiros et al. [4] proposed a time series modified to study the incidence of the disease on mortality. At last, the authors of [2], in their COVID prediction, have studied the combination of time series with neural networks through the LSTM algorithm.

3 Background

This sections introduces the definition of the concepts underlying our proposed prediction approach. We start by the exponential smoothing function.

3.1 Simple Exponential Smoothing

Exponential smoothing [3] is a rather very simple prediction technique that tries to infer the value at $t+1$ from historical data. It applies to time series without trend[1]. The intuition is to give more importance to the last observations. In other words, the more recent the observation is, the greater is the weight that is associated to it. For example, it makes sense to assign higher weights to observations made yesterday than to observations made seven days ago. We do not extend a series as we would like, for instance as we can do with a simple regression. However, we try to obtain a smoothed value at t to simply transfer it to $t + 1$.

Let us define the notion of time series as follows.

Definition 1. *A time series is a series of data points indexed (or listed or graphed) in time order, i.e., a sequence taken at successive equally spaced points in time.*

Then, the exponential smoothing can be defined informally as follows.

Definition 2. *Exponential smoothing is a rule of thumb technique for smoothing time series data using the exponential window function.*

The aim of smoothing is to give a general idea of relatively slow changes of value with little attention paid to the close matching of data values, while curve fitting concentrates on achieving as close a match as possible. The exponential smoothing has only one component called *level* with a smoothing parameter denoted by α. It is formally defined as a weighted average of the previous level and the current observation.

$$y_{t+1} = \alpha \times y_t + \alpha(1-\alpha) \times y_{t-1} + \alpha(1-\alpha)^2 \times y_{t-2} + \ldots + \alpha(1-\alpha)^n \times y_1 \quad (1)$$

where $0 \leq \alpha \leq 1$ is the smoothing parameter. The rate of weight reduction is controlled thanks to the smoothing parameter α. If α is large, more weight is given to more recent observations. We have the two following extreme cases.

[1] https://zhenye-na.github.io/2019/06/19/time-series-forecasting-explained.html.

- If $\alpha = 0$ we obtain the *average method* which corresponds to the case where the prediction of all future values is equal to the average of the historical data.
- If $\alpha = 1$ we obtain the *naive method* that just set all predictions to the value of the last observation.

3.2 Smoothing Holt's Method

Holt [1] is an extension of the simple exponential smoothing to allow the predictions of time series data with trends. The time series data with trend is informally defined as follows.

Definition 3. *A time series with trend is a set of ordered data points in which there is a long-term increase or decrease in the data. Such a trend does not have to be necessarily linear.*

Holt's method is based on a prediction function and two smoothing functions whose formalism are given below.

Prediction Function.

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

Smoothing Level Function.

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

Smoothing Trend Function.

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

In the functions above, α ad β have the following semantics.

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

4 Short-Term Holt Smoothing Based Prediction Function

This sections presents the formalism of the model we propose for the prediction of the total number of reported cases in a given day. We consider the set of total number of reported COVID-19 in a country in a period of time as time series data and we propose a Holt's smoothing based prediction model as we will detail it next. The proposed model based on the Holt smoothing is a chronological approach whose different steps are sketched in Fig. 1 for the training, validation and prediction tasks.

Let us assume the time series data $X = (t_i, y_i)_{1 \leq i \leq n}$ for n natural numbers and a function $f : N^* \mapsto R^+$ defined by

$$f(t) = b \times t \times ln(1 + at) \tag{5}$$

where t denotes the rank of the date considered to the index i, ln corresponds to the logarithm function and y_i denote the number of COVID19 cases at date t_i. The coefficients a and b of the function f satisfy the following conditions:

- $0 \leq a \leq 1$; and
- $0 \leq b \leq 1$.

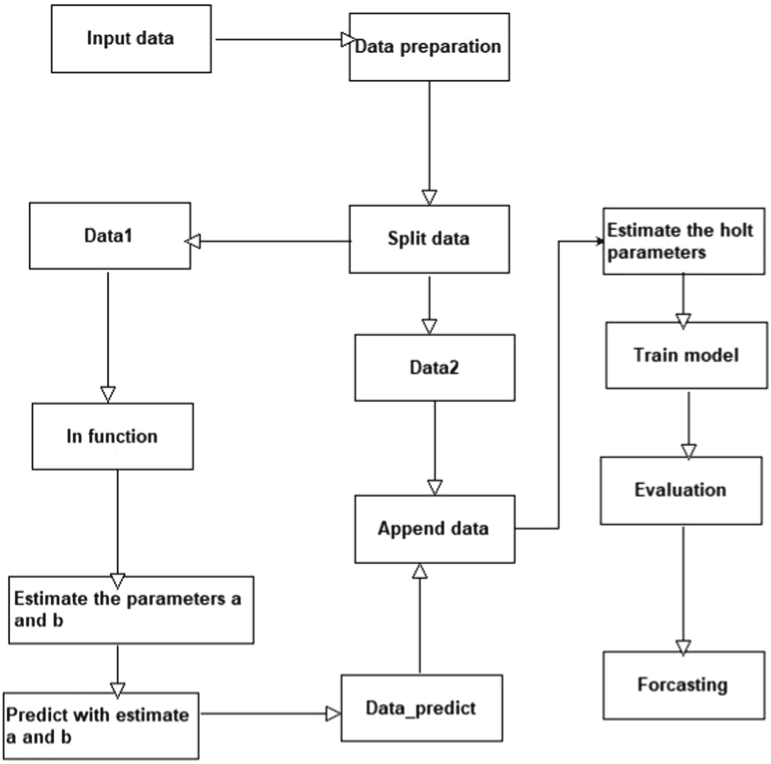


Fig. 1. Training, evaluation and prediction steps of our model

These coefficients have to be estimated from the real data. Then, for any natural number i, $0 \leq i \leq k$, (with k a natural number) we define a new series X_i^k as follows:

$$X_i^k = (t_i, f(t_i)) \tag{6}$$

In order to solve our prediction problem using Holt's method, we introduce a new series, denoted by X_2, obtained from the series $X = (t_i, y_i)_{1 \leq i \leq n}$ and $X_i^k = (t_i, f(t_i))$ in the following manner:

$$X_2 = \{(t_1, f(t_1)), (t_2, f(t_2)), (t_3, f(t_3)), \ldots, (t_k, f(t_k)), (t_1, y_1), \ldots, (t_n, y_n)\} \tag{7}$$

Let us set $Z_i = f(t_i)$ for $1 \leq i \leq k$ and $Z_i = X_i$ for $k+1 \leq i \leq n$ $(k \leq n)$, we simplify Eq. 7 in order to obtain the corresponding time series

$$Z = (t_i, Z_i)_{1 \leq i \leq n} \tag{8}$$

on which we will apply the different trend functions (linear, exponential and dampened) based on the proposed Holt method with trend. We evaluate different trend functions because of the fact that we have no prior knowledge about the actual trend of the evolution of the total number of daily reported cases in a given country. Testing with several functions will help us to choose the optimal one according to the data.

5 Experimentation Evaluation

This section presents the results of our experimental evaluation of the our proposed Holt smoothing based prediction model of daily cumulative cases of COVID-19. Recall that we combine Holt with a logarithm function for solving the cold start problem. We measure the prediction errors of our model with the linear, dampened and exponential trends.

We start by presenting the setting up of our tests with the description of the various datasets used for the experiments, considered performance metrics in order to measure the efficiency of our model and the turning of the optimal values of the parameters of the model. We have implemented our proposed Holt prediction model with the different trend functions using Python programming language. All the implemented source codes as well used datasets are available at: https://github.com/siriman/covid-19-senegal.

5.1 Setting Up Our Experimentation

We first describe used real-world datasets and then we present the performance metrics that have been considered to measure the performance of our model. We end this section by presenting the choice of the optimal values of the input parameters.

Description of the Real-World Datasets. Numerous datasets have been used to evaluate the performance of our model. They come from authoritative sources, mainly from daily reports of each country. We have considered datasets about the cumulative reported cases at each date in Senegal, Mali, France, Australia, USA, and Brazil. Table 1 summarizes the statistics about those datasets by specifying the total cases, the time period of data collection, and the average number of cases per day for each country. For the specific case of Senegal, Fig. 2 shows collected time series data that have been reported from 02/03/2020 to 26/02/2021; the total reported cases for the last ten days and the shape of the curve of evolution of the reported total number of cases given the entire time period. One has to note that, however, these data may be biased due to the

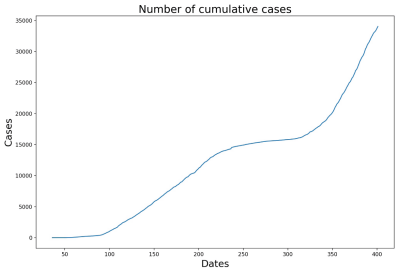
Table 1. Statistics of collected datasets about five countries

Country	Time_Period	Total_Cases	Avg_cases
Senegal	2020-03-02 to 2021-02-26	34031	94
Mali	2020-01-24 to 2021-02-26	8358	24
France	2020-01-24 to 2021-02-26	3746707	9366
Australia	2020-01-24 to 2021-02-26	28965	72
USA	2020-01-24 to 2021-02-26	28486394	70861
Brazil	2020-01-24 to 2021-02-26	10455630	28489

fact that in countries like Senegal there is no massive screening of the populations, i.e. only persons with symptoms have been tested to verify whether or not they suffer from the disease. For turning the parameters of the model and performing the validation of its performance, we have first focused on data from Senegal. Then, we have demonstrated the robustness and the generality of our approach to other countries such as Mali, France, USA, Australia, and Brazil. For the tests, we have considered 90% of each dataset for the training and the 10% remaining for the testing.

	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) Last 10 days



(b) All days

Fig. 2. Cumulative number of reported cases in Senegal

Performance Metrics. After having trained our chronological model, we have then proceeded to its evaluation. In the literature, there are several existing performance measures intended to evaluate the accuracy of a given prediction model[2]. In this study we rely on the *Mean Square Error* and the *Mean Absolute Error* that are two popular metrics to measure the error rate of a chronological

[2] https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/.

model. We provide their definitions below. Let Z_i be the real number of total cases at date i and P_i the predicted number of total cases.

- **Mean Square Error (MSE)** measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the real values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Z_i - P_i)^2 \tag{9}$$

- **Mean Absolute Error (MAE)** corresponds to the arithmetic average of the absolute errors between real values and predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Z_i - P_i| \tag{10}$$

We compare the MSE and MAE values of the three different trend functions (linear, dampened and exponential) when used with the Holt Smoothing model. The best prediction model corresponds to the one which presents the lowest values of MSE and MAE. We will show that our model present very low MSE and MAE rate errors in predicting the number of cumulative cases every day, in particular when used with the linear trend function.

5.2 Parameters Turning

The choice of the optimal values of the parameters α and β are critical for the performance of the Holt algorithm. Such a choice of these parameters are generally very subjective and might vary depending on the context of the study or the type of desired prediction. In this study, 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 based on given training data of the Senegal's dataset.

Figures 3a, 3b, 3c and 3d 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 equals to 0.835. Meanwhile, 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. We choose these values of a and b as well as those of α and β that minimize the MSE for the rest of this evaluation.

5.3 Description and Analysis of the Obtained Results

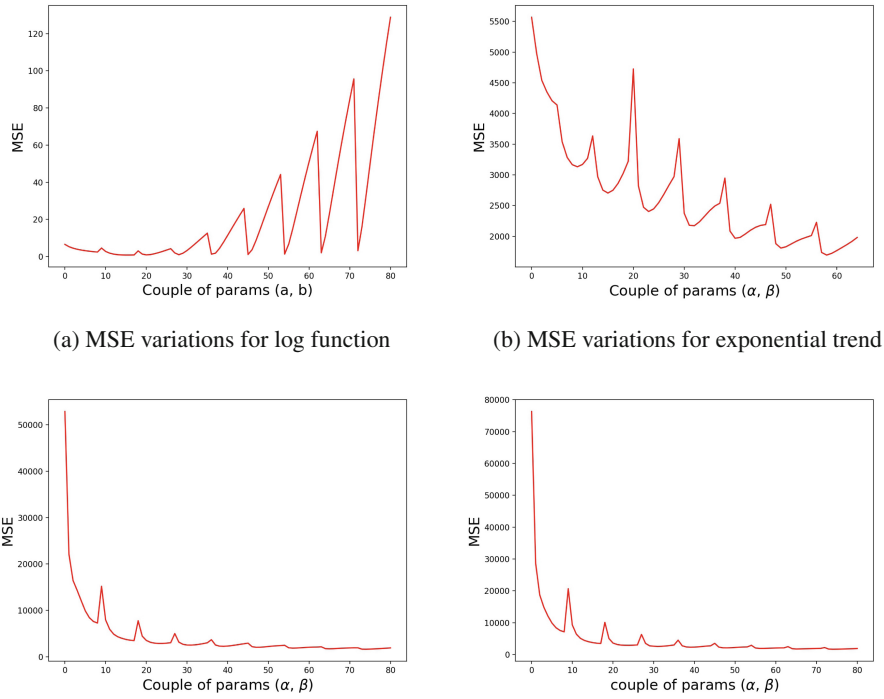
In this section, we present the results of the prediction power of our model with each of the different trend functions. We calculate for each variant its **MSE** and **MAE**.

Table 2 presents the values of the MSE and the MAE of all the variants of the modified Holt's method. The curves in Fig. 4 compares the actual reported values and the predicted values about the cumulative number of COVID-19 cases

Table 2. Measures of the performance of the Holt model with various trend functions

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

in Senegal for the linear, dampened and exponential trend functions respectively. The evolution of cumulative reported cases is given by the blue line while the cumulative number of cases predicted by our model is given by the red line. An analysis of the obtained results from the experiments of our Holt approach with the three trend functions show that all of them are quite accurate for the task of predicting the daily cumulative number of cases of COVID-19 in Senegal. However, in term of MSE and MAE we observe that combining Holt with the linear trend outperforms the others significantly. Indeed, such a combination provides the smallest MSE and MAE values which are respectively 1634.66 and 25.93. This indicates that our prediction differs, on average, by approximately



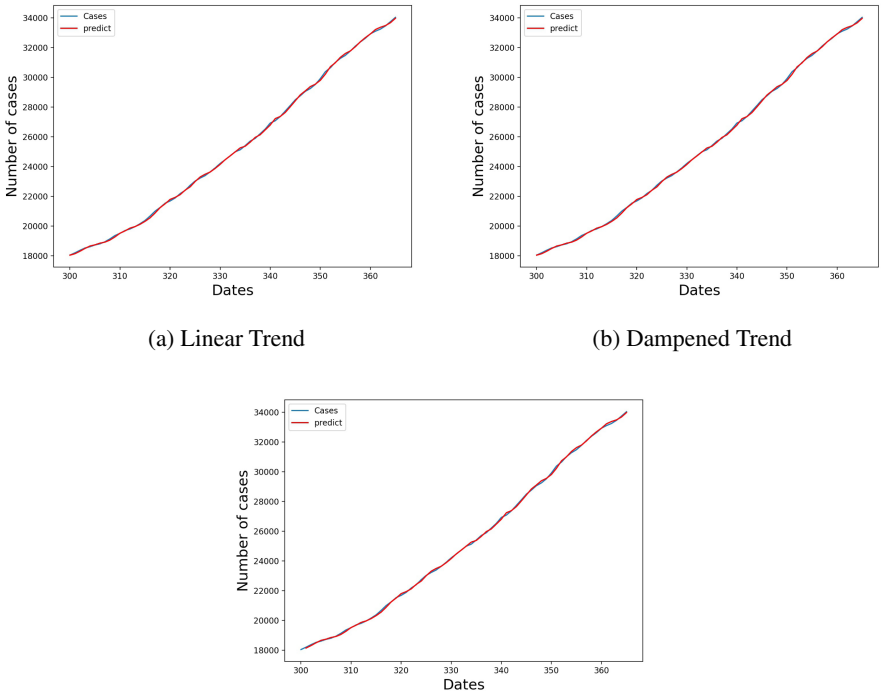
(a) MSE variations for log function

(b) MSE variations for exponential trend

(c) MSE variations for linear trend

(d) MSE variations for dampened trend

Fig. 3. Estimation of the optimal values of the parameters of our prediction model



(a) Linear Trend

(b) Dampened Trend

(c) Exponential Trend

Fig. 4. Number of real cases vs. number of predicted cases

1634.66 from the actual data. In other words, this represents a difference of approximately 25.93 in absolute value from the actual data.

Unlike SIR (Susceptible, Infectious, or Recovered) models which are very complex and which involve a number of parameters that are very difficult to control and whose goal is to predict the basic reproduction number, R_0, our model which makes it possible to predict the number of cases of contamination in the near future can be of a capital contribution in the definition of policies for the fight against the pandemic.

We also evaluated and showed the robustness of our prediction model by testing it on datasets from Mali, France, Australia, Brazil, and USA. To this end, for each of these countries we have compared the real cumulative cases during the last five days with the values predicted by our model with the different trend functions in the same period. Tables 3, 4 and 5 contain the results of the comparison and show that Holt with the different trend functions capture well the evolution of the cumulative number of reported cases in these countries regarding the prediction error when comparing real and predicted values, e.g. the maximum deviation is equal to 8 for Mali.

Table 3. Real values vs. Values predicted by Holt with Linear trend

Date	Mali		France		Australia		USA		Brazil	
	Real values	Predicted values	Real values	Predicted values	Real values	Predicted values	Real values	Predicted values	Real values	Predicted
2021–02-22	8306	8311	3669354	3685625	28937	28934	28190159	28963	10195160	10216996
2021-02-23	8324	8316	3689534	3689127	28939	28942	28261595	28254026	10257875	10242316
2021-02-24	8332	8335	3721061	3707712	28947	28944	28336097	28326517	10324463	10302692
2021-02-25	8349	8343	3746475	3740348	28957	28951	28413388	28405129	10390461	10370618
2021-02-26	8358	8363	3746707	3767587	28965	28963	28486394	28486268	10455630	10438595

Table 4. Real values vs. Values predicted by Holt with Dampened trend

Date	Mali		France		Australia		USA		Brazil	
	Real values	Predicted values	Real values	Predicted values	Real values	Predicted values	Real values	Predicted values	Real values	Predicted
2021-02-22	8306	8310	3669354	3684567	28937	28934	28190159	28204446	10195160	10214138
2021-02-23	8324	8316	3689534	3688084	28939	28941	28261595	28252513	10257875	10237408
2021-02-24	8332	8335	3721061	3706683	28947	28944	28336097	28325049	10324463	10299422
2021-02-25	8349	8343	3746475	3739307	28957	28951	28413388	28403632	10390461	10369579
2021-02-26	8358	8360	3746707	3766522	28965	28963	28486394	28484712	10455630	10439238

Table 5. Real values vs. Values predicted by Holt with exponential trend

Date	Mali		France		Australia		USA		Brazil	
	Real values	Predicted values	Real values	Predicted values	Real values	Predicted values	Real values	Predicted values	Real values	Predicted
2021-02-22	8306	8306	3669354	3686297	28937	28934	28190159	28206524	10195160	10217941
2021-02-23	8324	8324	3689534	3689712	28939	28942	28261595	28254376	10257875	10241040
2021-02-24	8332	8332	3721061	3708277	28947	28943	28336097	28326863	10324463	10303092
2021-02-25	8349	8349	3746475	3740981	28957	28952	28413388	28405503	10390461	10373401
2021-02-26	8358	8358	3746707	3768269	28965	28964	28486394	28486679	10455630	10443249

6 Conclusion

In this paper, we have studied the problem of predicting the daily total number of reported COVID-19 cases in a given country. As a solution, we have proposed a Holt smoothing based prediction model combined with a logarithm function for cold start purposes. Our intensive experimental evaluation conducted on various datasets has showed that the proposed model with a linear trend presents very good performances. We have also proved the robustness and an easy generalization of our model to any country in the world. As perspectives, we plan to evaluate our model by using other performance metrics, to perform a deeper comparison with existing prediction models such as SIR and to improve our model so that it can do long term prediction.

References

1. Aragon, Y.: Lissage exponentiel. In: Séries temporelles avec R. Pratique R, pp. 121–132. Springer, Paris (2011). https://doi.org/10.1007/978-2-8178-0208-4_6
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. Int. J. Stat. Appl. Math. **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. World Health Organization: 2019 World COVID-19 Report, December 2019. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
8. 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)
9. Zhao, S., et al.: Preliminary estimation of the basic reproduction number of novel coronavirus ((2019-nco)v) in China from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. Int. J. Infect. Dis. **92**, 214–217 (2020)