Predicting The Trend of COVID-19 Cases in Ho Chi Minh City Using The Prophet Model

Minh-Khoi Nguyen-Nhat
Honors Program, Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
19120020@student.hcmus.edu.vn

Hoang-Long Vu-Dao
Honors Program, Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
19120022@student.hcmus.edu.vn

Duy-Hoang Do-Nguyen
Honors Program, Faculty of Information Technology
University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
19120077@student.hcmus.edu.vn

***Abstract*—COVID-19 has become a pandemic due to its rapid spread. Our daily lives have been severely disrupted as a result of this, both directly and indirectly. We aim to utilize Machine Learning model to forecast the trend of the disease in Vietnam. Through a plenty of solution using Susceptible-Infectious-Recovered model with good result, we choose Prophet model in attempt to find a better way to predict the number of infected case and death case in the future. The model is then fitted with a trusted data from the Vietnamese National Cyber Security Center. The result show that the model prediction is not match with the infected data, but has some positive result in the death data. In conclusion, we state that the model is not good enough to be trusted as a forecast tool for support important decision.**

***Keywords*—Prophet, COVID-19, forecast, infected case, death case, Vietnam**

I. Introduction

COVID-19 is the ongoing global pandemic which have infected and killed millions of people, caused by the SARS-CoV-2 virus. As time goes on, there appears to be variants with various changes compared to the original virus first found in China, which includes the infection rate and fatality of the disease. These variants have been affecting the world as a whole. Particularly in Vietnam, one of the countries before known with the best policies against the pandemic and the lowest cases per capita, saw a severe outbreak when the Delta variant hit, and turned everything upside down. As a way to battle against COVID-19, analysis from the data and using them to predict the future trend of the pandemic can be crucial for a country's government to decide its own policies which best fit the situations.

Thus, there have been cases of analysis of the pandemic, many of which use SIR (Susceptible-Infectious-Recovered) as prediction model. However there are disadvantages of SIR, therefore in this article we aim to find a better way to tackle the problem of analyzing the trend of cases and deaths of COVID-19 in a particular region. We use the Prophet model instead of SIR to forecast the number of cases and deaths in Ho Chi Minh City - the largest city in Vietnam, which is also the region that got hit the hardest by the pandemic, with around 50% of cases of the entire country within the city. There are challenges regarding approaches to the problem. For instance, Prophet is strongly affected by the seasonality, which can be a strong obstacle for data analysis and prediction. However, it gives the ability to make the data more flexible using holiday effects and change points where there is a potential difference in the growth of the pandemic [1].

The content of this paper is as follows:

First, in the section I, we define our motivation for doing the research, the challenge as well as the scope in our problem.

In section II, we discuss the main idea of the Prophet model by breaking down its formulas, explaining the trend, seasonality and the holiday component as well as how the model fits the data and gives reasonable results.

All of our experiments and results are detailed in section III, where we will talk about the dataset and computing environment, show the test result, then extract some insight about the COVID-19 pandemic trend in Vietnam as well as the model's efficiency.

Finally, in section IV, we come to conclusion and outline some potential future works.

II. Method

A. The Prophet model

Prophet is a forecast procedure created by the Facebook's Core Data Science team, which is best used in predicting trends in a time series data where there is contribution of impact from seasonality and holiday events.

The core of Prophet is a decomposable time series model with three main components: the trend function $g(t)$, the seasonality $s(t)$ and holiday effects $h(t)$, all of which are functions of time t. Combined, they form the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

B. The trend

The number of deaths related to COVID-19 can be easily correlated to the number of cases. Therefore while we only consider the number of cases in this section, the same principles should also apply to the number of deaths.

The SARS-CoV-2 virus, essentially like any other infectious disease, has an exponential growth in infection rates in an environment where there is an infinite number of potential hosts. However, its growth is limited to the number of hosts in a community which has achieved herd immunity, which does not necessarily translate to all people in the community. Therefore realistically the growth of SARS-CoV-2, and consequently the number of COVID-19 cases, has a growth which should follow the standard logistic model. Therefore in our experiment, the trend function of the COVID-19 cases is determined as

$$g(t) = \frac{C(t)}{1 + e^{-k(t-m)}}$$

with $C(t)$ being the total capacity, meaning the number of potential hosts of SARS-CoV-2; k being the growth rate and m an offset parameter.

As the formula suggests, the capacity $C(t)$ is a function of time [1]. As stated, the maximum number of hosts does not always equal the total number of people in the community since immunization and infection limiting methods exist. Therefore, the maximum number of infected people should change over time. We compute $C(t)$ of day t by subtracting the number of cases in the previous day $t-1$ from the current population of the city (for deaths, the capacity is strictly correlated to the number of cases for obvious reasons).

The growth rate is also a variable over time. While, like the capacity, it is also affected by people's reactions to the disease and the policies of the authorities, for COVID-19 the growth rate is also affected by the various variants of the virus. For instance the Delta variant, which was first discovered in India, has a much higher infection rate (and fatality) than the original version of the virus. We use changepoints to determine where the growth rate potentially changes.

C. Seasonality and holiday effects

Various early studies have shown that seasons are also a factor when it comes to the infection rate of COVID-19. Within the limitation of our knowledge, we suspect there are reasons for such correlations, such as the fact that the virus is more aggressive in certain climates and different human traffic in different points in time, for example people tend to travel more in the summer. In many regions, holidays can also contribute significantly to the changes of the infection rate. These yearly changes can be a significant factor since the effects for each year should be similar.

To generate the effect of seasonality, by default, Prophet uses a Fourier series with the coefficients randomly generated from a normal distribution with mean 0

$$s(t) = \sum_{n=1}^{N} \left(a_n \sin \frac{n2\pi t}{P} + b_n \cos \frac{n2\pi t}{P} \right)$$

where P is the duration of the regular period [1]. Since we want to observe, and therefore predict, yearly changes throughout the course of two years since the first COVID-19 infection in Ho Chi Minh City, we decided to set $P = 365.25$.

For the holiday effects, we simply use the default formula of the Prophet. Suppose we have L holidays to be considered from D_1 to D_L, the formula for the holiday effects is:

$$h(t) = Z(t)\kappa$$

where $Z(t)$ is the matrix of regressors $Z(t) = [\mathbf{1}(t \in D_1), \mathbf{1}(t \in D_2), ...\mathbf{1}(t \in D_L)]$ and κ is a randomly generated vector, picked from the normal distribution with mean 0 and some standard deviation v.

D. Model fitting

Maximum A Posteriori(MAP) is a method to estimate model parameter. We are given a piece of data $X = \{x_1, x_2, ..x_n\}$ which assumed to follow a distribution p. We have θ as a parameter, maximum a posteriori help we find value of parameter θ for X that maximize a posteriori:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta | x_1, x_2, ..., x_n)$$

In Prophet, after combining seasonality and holiday effects into matrix $\mathbf{X}$ and change points in matrix A, we can rewrite model like this the STAN code 1, STAN is a programming language that is used for probability models, having statistical algorithms that help fitting models, making predictions, estimate posterior and evaluate the result [2]. With STAN, we will use L-BFGS to fitting parameter of the model by MAP method. The L-BFGS is used to optimize the performance of the model on low-computational hardware [3].

III. Experimentation

A. Dataset and configuration

Dataset: We conduct an experiment to forecast the upcoming trend of the increase in the cumulative number of COVID-19 cases and deaths in Ho Chi Minh City using the Prophet model as stated. Using the predicted result, we calculate the forecast number of daily infected cases and deaths case of which to determine how many cases higher (or lower) in one day compared to the day before in percentage. The data used is from the Vietnamese National Cyber Security Center (NCSC) page for COVID-19, last updated in January 20, 2022 [4]. The population of Ho Chi Minh City is also directly extracted from the data of COVID-19 cases by NCSC.

The reason why we choose the total number instead of the number each day is based on observing Figure 1. We notice that the data has many outlier, especially in June and July 2021. The data is updated each two or three days, especially on July 23, there were over 16,000 new cases added. Those outliers make our data unstable and hard to predict. Therefore, we decided to use cumulative number of infected/death cases, which can ease this problem.

Configuration: The time period of forecast is 120 days, from November 18, 2021 to May 17, 2020.

Listing 1: Example STAN code for fitting model parameter [1]

```
model{
     k ~ normal(0, 5);
     m ~ normal(0, 5);
     epsilon ~ normal(0, 0.5);
     delta ~ double exponential(0, tau);
     beta ~ normal(0, sigma);
     y ~ normal(C ./ (1 + exp(-(k + A * delta).*(t - (m + A*gamma)))) + X*beta
                                        , epsilon);
}
```

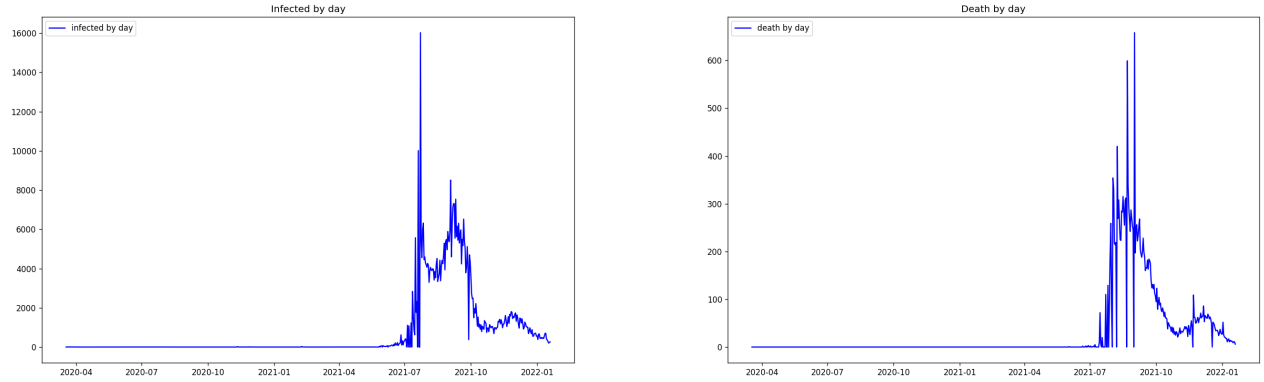

Figure 1: Number of infected case by day and number of death case by day from March 18, 2020 to January 20, 2021.

We also consider these change points, which were significant points in time where there could be a change in the trend of the infection.

- May 1, 2020: When Vietnam started to loosen restriction after a 15-day nationwide lockdown after a surge started in February.
- July 28, 2020: Outbreak in Da Nang city, which led to a surge of cases in HCMC as a result.
- January 28, 2021: Outbreak in Hai Duong province, which led to another surge of cases.
- April 20, 2021: This is around the time when the Delta variant was first detected in Vietnam, which led to a severe outbreak in HCMC.
- July 26, 2021: HCMC imposed the first ever mandatory curfew in the country.
- October 1, 2021: Cases of infection per day was on the trend of decline.
- January 10, 2022: Around the time when HCMC was declared a safe zone based on Vietnam's COVID-19 safety measures.

B. Result

Using the dataset from NCSC with records created back from March 2020, we achieved the predicted case plot at Figure 2. Using the predicted infected cases above as capacity for death cases, we have Figure 3. Combine two plot above, we have Figure 4.

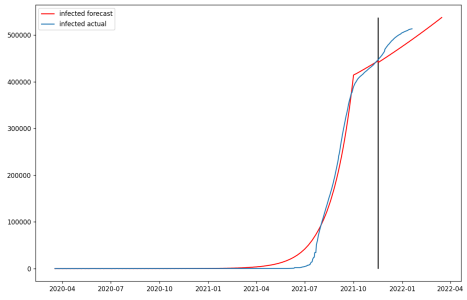

Figure 2: The plot of predicted and actual COVID-19 infected cases in Ho Chi Minh City from March 18, 2020 to May 17, 2022 using NCSC's data from March 2020. The Prophet model is fitted with data before the vertical line.

We use the mean square error and R-squared as the metric for evaluating the model. The test data is extracted from the last 64 days of original data (10% of the original data), from November 18, 2021 to January 20, 2022. The Figure

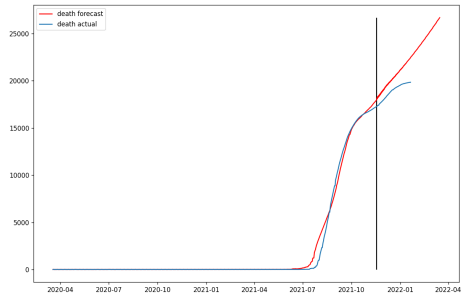

Figure 3: The plot of predicted and actual COVID-19 death cases in Ho Chi Minh City from March 18, 2020 to May 17, 2022 using NCSC's data from March 2020. The Prophet model is fitted with data before the vertical line.

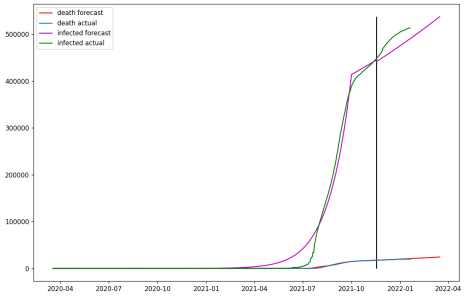

Figure 4: Combine of figure 2 and 3 for better insight

C. Analysis

First, we observe Figure 2. The model is well-fitting to the trained data. However, based on the figure, the model appears to underestimate the increase in the total number of infected patients and not predict the future well. The mean square error and R-square for this case is large. We can explain this by the lack of and inconsistency of data changes since the COVID-19 is hard to predict and Prophet model was firstly designed for seasonality and business-related data.

Moving our focus to the total number of death cases over time at Figure 3, the model is well-fitting to the trained data as with the previous infected data. Though the model still overestimate the trend, its prediction are more accurate than with infected data and the mean square error and R-square score for this case are both acceptable. This is due to the benefit of medicine and the fact that the virus's death rate is significantly lower than the virus's infection rate, allowing people to better control the overall number of death cases than infected cases.

Now let us look at the overall view in the Figure 4 which includes both infected data and death data. We discovered that, despite of underestimate the number of infected cases and overestimate the number of death cases, the model still

	Infected cases	**Death cases**
MSE	608116810.2	303561.5
R-square -0.549		0.535

Table 1: Evaluation for infected case model and death case model

represents a proper trend: the number of death cases is increasing at a slower rate than the number of infected cases, which is a positive feature of the model.

In conclusion of this part, we state that the Prophet model has captured some insight of the COVID-19 data in Vietnam. However, the error in the model still relatively large.

IV. Conclusion

The goal of this research is to determine whether Prophet is a good model to handles time-series data that has no seasonality, random patterns, and few observations, such as COVID-19 instances data. The data is total number of infected cases and total number of death cases from March 18, 2020 to January 20, 2022. Because the data is unreliable, contains numerous outliers, and was not developed for the Prophet model in the first place, the model is insufficient to serve as a government support tool. In this research, we also discover some anomalies in the data. For the future work, we will collect more data and adjust the model for more accurate prediction.

References

[1] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[2] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, no. 1, pp. 1–32, 2017.

[3] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[4] "Ncsc data," https://covid19.ncsc.gov.vn/dulieu.