International University of Ho Chi Minh City

School of Industrial Engineering & Management

CAPSTONE DESIGN

MODELLING THE SPREAD OF COVID-19 IN HO CHI MINH CITY: DATA-DRIVEN ANALYSIS

Group members:

Vo Thi Thien My – IEIEIU18057

Nguyen Huynh Ngoc Que – IEIEIU18073

Pham Ngoc Thu Uyen – IEIEIU18114

Supervisor: Mr.Tran Van Ly

Ho Chi Minh City, 22 November 2021

TABLE OF CONTENTS

ABSTRACT:

The coronavirus disease started in 2019 in Wuhan Province, China. It is found that the COVID-19 comes from a large family of viruses identified in 1965. Since the outbreak in 2020, COVID-19 and its variants have spread and infected millions of people around the world, causing severe damage in terms of economic, social, and healthcare aspects; taking a toll on businesses, transportation, and activities in many cities and countries, this also includes Ho Chi Minh city, the most developed city in Vietnam. The fourth wave pandemic has pushed Ho Chi Minh city to its boundaries, making residents' lives difficult and threatening the overall economic development of the city and the country. However, recently, data analysis, exploration, and comparative analysis were conducted by researchers worldwide to find suitable and optimal models to forecast the infected and confirmed cases of COVID-19, including in Vietnam. This paper shows a quantitative analysis of forecasting confirmed cases of Ho Chi Minh City through the ARIMA model during April 27th, 2021, and November 24th, 2021, while implementing external parameters such as daily vaccination rate, covid-variant incubation period and generation time, as well as social distancing applied for each control policy through Pearson's correlation method. The result shows that the forecasting of daily confirmed cases is similar to those of recorded ones; specifically, the model was able to demonstrate the distribution of daily confirmed cases accurately and provided the correct estimation for the two peaks during the fourth wave between April and November which indicates that the implementation of external data is considered adequate. RMSE estimation of the model proposed in this study is lower than other practices that used the ARIMA model and other conventional methods to forecast the COVID-19 outbreak.

Chapter I. INTRODUCTION

1.1.

Motivation:

The dangers of this virus and the spread of the COVID-19 pandemic are well-known since this pandemic has caused a significant impact on social life, economy, education, technology development, etc. Hence, by forecasting the spread of COVID-19, the effects and risks of the Corona-virus pandemic can be controlled. To cope with the situation caused by the coronavirus, countries' governments have proposed various preventive strategies to minimize the impact the pandemic has on the economy, social aspects, and the countries' healthcare systems, through establishing field hospitals, gathering as many facilities resources as possible to tackle the outbreak in the hope of having the spread under control. Therefore, forecasting COVID-19 confirmed cases could help the government and regional Medical departments raise awareness among the population since it can help inform the danger of the upcoming outbreak and figure out the scale of impact a COVID-19 variant may have on the population.

This research focuses on forecasting the spread of COVID-19 in Ho Chi Minh City (HCMC) during the fourth wave, from April 27, 2021, to November 24, 2021. Through this research, it is beneficial that HCMC will understand the basics of the impact an upcoming wave can have on the social and economic aspects, which assists the city government in establishing more suitable and effective control measures.

Moreover, HCMC can inform the danger of the upcoming outbreak to its residents, prepare enough resources which can deal with the worst-case scenario to minimise the negative impact on the economy and supply chain of the city. Finally, the forecasting result can assist the city in preparing enough resources in terms of medical staff and specialised medicine to minimise the death rate caused by the pandemic while preparing more suitable and efficient vaccination campaigns to increase its community immunity

1.2. Problem Statement:

The problems of the current pandemic situation, particularly results of daily confirmed cases, still exist although temporary epidemic prevention methods have been in process. Concretely, most of citizens reported that there sometimes exists the huge gap of forecasting the number related to exposed individuals, leading to that control measures and plans of preventing the spread of cases of the COVID-19 are not suitable to the present state and citizens are not aware of the dangers of this situation, for example, they do not wear mask frequently, do not accept the

two-metre distance among people, or do not accept to get vaccinated. Additionally, most people could see that sometimes government and front line medical staff teams do not support effectively for inhabitants.

Due to the previous description in the above part and the epidemic problems occurring at the moment, the huge gap between current forecasting data with the real data of COVID-19, especially the Delta variant, the final conclusion must be conducted are reducing the gap, leading to the most suitable control measures and vaccinating obligating rules can be served for inhabitants, particularly citizens in Ho Chi Minh city.

- The problem or the gap needed to fill

Most of the research paper that provides time-series model for forecasting COVID-19 confirmed cases generally depend on only the confirmed, death or recovered cases collected, without considering the external factors that contributing to the spread of COVID-19, such as the control measures proposed, Covid-variant and even vaccination while this factor is one of the most important elements for community immunity towards the coronavirus. Hence, this research offers a way to implement these external factors in order to increase the effectiveness of the proposed forecasting model, as well as enhancing the model forecasting accuracy to establish more appropriate control measures in the upcoming outbreak.

- What needs to be solved or achieved?

This research should not have the same problems as previous ones which is not considering external factors impacting on the spread of the COVID-19 pandemic in Ho Chi Minh city. As a result, it must work on the core element which is COVID-19 confirmed cases, and others from outside to have the more accurate results to support other citizens and government to control the epidemiology spread by stating more necessary measures and information about the pandemic.

1.3. The Design Project Objectives and Requirements

The result of this capstone design project will be applied in controlling plans and preventive measures of the Ho Chi Minh City government and businesses including local business, international corporations, National Corporation, etc. Additionally, the study team will have strong collecting and analyzing data, looking for and researching related topics and previous research. Finally, the most important thing is that this three-people group can complete this Capstone Design subject with three credits and receive background skills for the thesis in the next semester.

After completing this project, medical and epidemiological researchers can apply the result to define and approach the most suitable methods for improving the current situation in HCMC. Additionally, the HCMC government and Department of Health can identify new control measures fit with both states of this city and citizens. So as to receive these applications, the expected results of this research must have the as small as possible gap between the forecasting and the future confirmed cases data and the supporting charts which can be used in future researches by other authors.

Expected outputs and/or applications

- Identify the suitable external parameters than directly affect the number of confirmed cases in HCMC
- Provide suitable model approach for COVID-19 data forecast for the future outbreak
- Accurately predict the COVID-19 outbreak peak during the fourth wave
- Successfully implementing external factors that enhance the forecasting accuracy of the model
- Know how to apply Python and its applications in establishing forecasting models to save time and effort when updating data and forecasting with new models.
- Understand the effectiveness and usefulness of forecasting models can have on minimizing negative impacts the pandemic have on the economy and social aspects

So as to conduct final results for use in those above applications, the study team has to define accuracy in modelling COVID-19 spread in HCMC and provide a suitable model approach for COVID-19 data forecast. In order to receive that model, the team must improve themselves day after day to have ability of implementing certain external parameters into the model to increase the accuracy of the forecast and utilising the historical data to help develop a better numerical simulation, so that forecasting the spread of COVID-19 in short-terms with high accuracy could be in process of this project.

Before conducting data collection and establishing forecasting model, it is worth noted that data collection and model forecasting must perform under the medical standards proposed by Vietnamese Ministry of Health, such as the number of vaccination doses permitted to perform per day by Ho Chi Minh City government; time periods between the first and the second dose of each type of vaccine; types of vaccines are permitted to be combined during the first and second dose; allowable number of swab tests available per day; types of vaccines are permitted to be combined during the first and second dose; types of vaccine are available for children from 11 to 17 year-old; General COVID-19 vaccine clinical trials; health check process before vaccination; groups of people who are allowed and not allowed to receive COVID-19 vaccination.

1.4. Scope and Limitations:

During the progress of researching, the student researcher group had to face with some limitation which are having lack of knowledge and time for researching leading to that the model could have better forecasting results and can be further extended and applies better method to solve, and the data of confirmed cases in HCMC may not be accurately updated so that this makes the team to perform data validation and accept certain deviations in terms of data collected.

The scope of this research and the group of student researchers to achieving is that forecasting the COVID-19 confirmed cases in HCMC as mentioned above, knowing how to combine vaccination rates in the model while collecting vaccination campaigns in HCMC starting from July 2021, when the government decided to perform large-scale vaccinations on HCMC residents. Additionally, the research can show the model performed under the medical standards mentioned in the paper with the time for modelling the spread of COVID-19 mostly focused on the Fourth wave since it provides the most frequent data on confirmed cases and it is also the most complicated outbreak during 2020-2021 in HCMC. Therefore, the result can be applied in providing suitable control measures implemented in the 4th wave.

During research progress, the student team assumes that there does not exist COVID-19 variant consideration, and does not include numerous data of individuals who moved from other regions to HCMC, and from HCMC to others, also citizens living in Saigon collected in national systems in 2019 and 2020. Additionally, collected data are accepted by medical ethics stated by the national government, HCMC government and World Health Organisation (WHO). Besides, the vaccination rate is only considered from the milestone when HCMC decided to perform the first large-scale vaccination campaign for HCMC residents. However, vaccination efficiency deviations are not included into the model, its values are only considered according to research of each type of vaccine. Finally, the model complies with medical standards written in the group's paper.

Some practical constraints

a. Time

The given time to develop such a sophisticated model is limited and this Capstone project appears to be novel and challenging for the whole group members at the very first stage. Consequently, we have to refer to several previous scientific journals and papers to figure out what we should do at the next steps. To keep pace with a realistic timeframe while carrying out our practical research may be the greatest challenge of all times. It is common knowledge that the more robust the research, the more time it will take.

Besides, Ho Chi Minh City is not heavily affected by the three previous waves of COVID-19 pandemic (from 23. January. 2020 to 25. March. 2021). As a result, the record of historical data of confirmed cases at this period is extremely rare and discrete. Due to this unavailability of data for such a long period, it is possible that we may not give the exact insights about the data. If we had enough actual data for over a long period, the forecasting results of the future confirmed case using ARIMA model would be more accurate.

b. Research skill

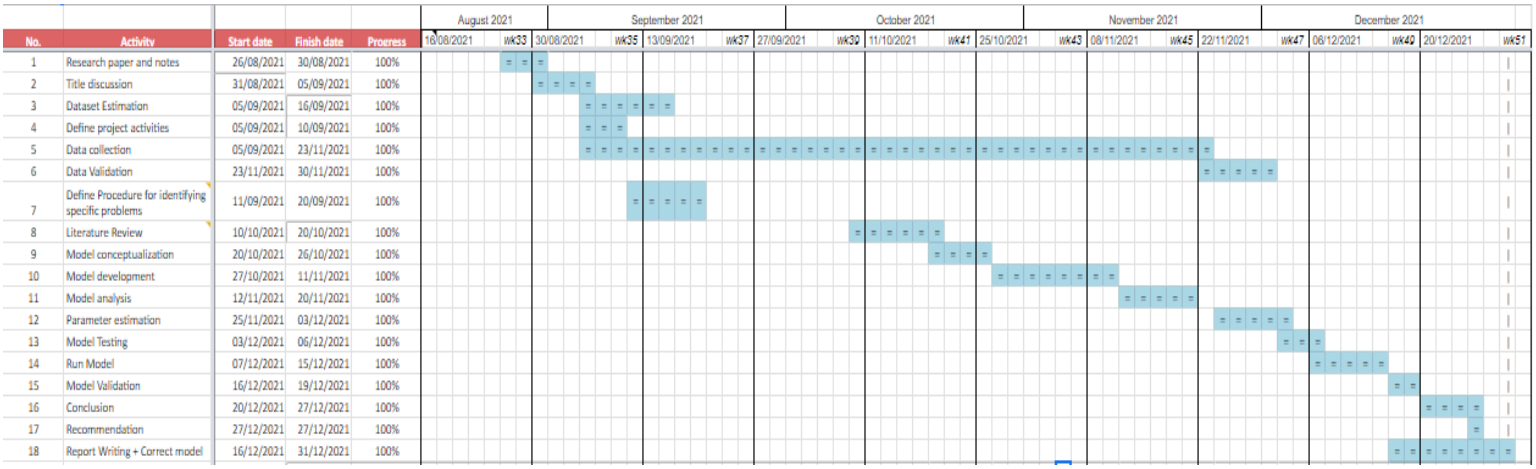
Academic research requires reasoning and critical thinking to determine whether a journal we found is good and persuasive enough for us to take it as a key reference. This is because not all published journals are completely true and can be validated by powerful authorities. Besides, sometimes our personal life, family or other work commitments can interfere and hinder long-term field work. Studies have proved that the success of primary research also depends upon the attitude and skills of the researcher.

1.5. Project Plan

The group divides this study into 8 distinct sections. The first section is the general information about our study. In this section, we introduce the need of the study, which is the need of forecasting new daily confirmed cases in Ho Chi Minh city. We also point out the design requirement of the model, define some model assumptions, specify some practical constraints, and define the limitations of the study. The second section is about the consideration of model design concepts. In this section, we refer to many current COVID-19 forecasting models and several techniques that are widely used in previous study. We make a summary table to describe clearly about these approaches and make judgement about the pros and cons of each approach we have researched. In the third section, we use the multi-criteria decision making (MCDM) method to select the most suitable approach for our study, based on the detailed description as well as the approach pros and cons analysis in the previous part. We come up with the design structure of the selected model and develop the flowchart for the forecasting process. Additionally, we analyze and justify techniques we are about to use in this study. The fourth section of our study is the data collection and validation. We collect all required data for developing the COVID-19 forecasting model and perform the data validation with several hypothesis testing at different confidence intervals. The fifth section is the estimation of all parameters of the model. The sixth section is the demonstration of the modelling process. At this step, we come up with the forecasting results and draw initial analysis towards those results. In the seventh section, we draw our conclusion about the forecasting results and discuss how the

Government can do differently to diminish the impact of COVID-19 spread. The last section is the references and appendix of the whole study.

- Gantt Chart of Capstone Project:



No.	Activity	Start date	Finish date	Progress
1	Research paper and notes	26/08/2021	30/08/2021	100%
2	Title discussion	31/08/2021	05/09/2021	100%
3	Dataset Estimation	05/09/2021	16/09/2021	100%
4	Define project activities	05/09/2021	10/09/2021	100%
5	Data collection	05/09/2021	23/11/2021	100%
6	Data Validation	23/11/2021	30/11/2021	100%
7	Define Procedure for identifying specific problems	11/09/2021	20/09/2021	100%
8	Literature Review	10/10/2021	20/10/2021	100%
9	Model conceptualization	20/10/2021	26/10/2021	100%
10	Model development	27/10/2021	11/11/2021	100%
11	Model analysis	12/11/2021	20/11/2021	100%
12	Parameter estimation	25/11/2021	03/12/2021	100%
13	Model Testing	03/12/2021	06/12/2021	100%
14	Run Model	07/12/2021	15/12/2021	100%
15	Model Validation	16/12/2021	19/12/2021	100%
16	Conclusion	20/12/2021	27/12/2021	100%
17	Recommendation	27/12/2021	27/12/2021	100%
18	Report Writing + Correct model	16/12/2021	31/12/2021	100%

Figure 1: The research project Gantt Chart for conducting a quantitative analysis of forecasting confirmed cases in Ho Chi Minh city from 27th April,2021 to 24th November, 2021

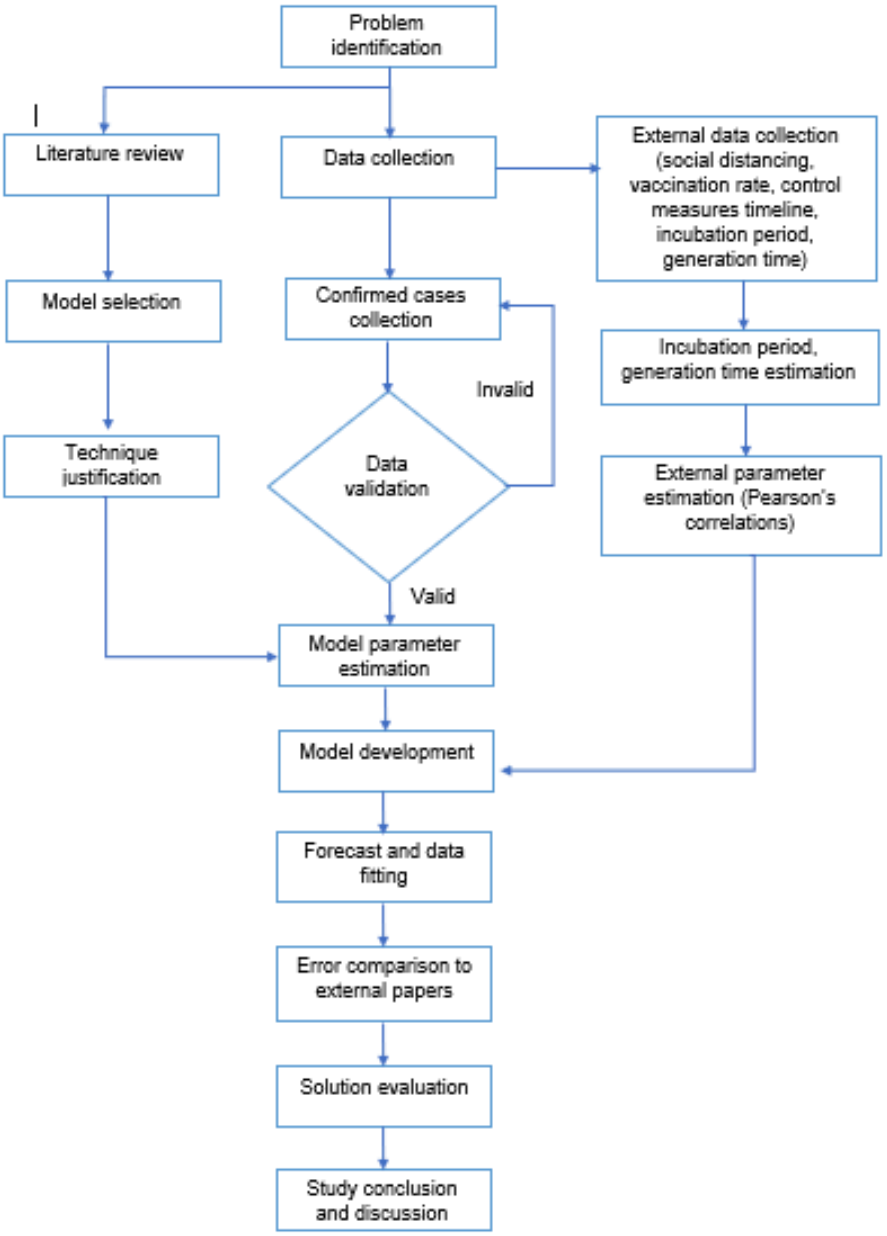


Figure 2: Flow chart of the study

Chapter II. DESIGN CONCEPTS CONSIDERATION

2.1. Overview:

As mentioned in most newspapers, Ho Chi Minh city is one of regions which have a huge number of COVID-19 cases in the fourth wave of the Coronavirus epidemic. Therefore, the study focuses on the pandemic situation of Ho Chi Minh city, specifically the daily confirmed cases during the fourth wave from 27th April, 2021 to 23rd November, 2021

2.2. Literature Review:

This research focuses on the COVID-19 pandemic caused by the Coronavirus, particularly forecasting confirmed cases in Ho Chi Minh City with the expected result for implementing future plans in both academic and practical situations for controlling the dangers of pandemic in the city. Generally, based on the previous research focusing on forecasting COVID-19 cases shown, the team could divide two types of general concepts which are time series models and data assumption models.

In particular, on one hand, when first look on Ogundokun et al.and Wang et al. research, data assumption models were used as models which use assumptions derived from the forecasting situation in order to model the event, usually considering initial values, instead of a set of data from the past (SIR and SEIR models). Additionally, while considering their implementing the result for future academic research, Wang et al. suggested that data assumption models can be added more impacting factors into the models, then researchers could re-forecast by extending other directly impacting factors such as vaccination stand for V in SEIR model. Particularly, in paper [23], Yang et al. applied the SaucIER model which is the extension of the benchmark of the SIR model by dividing the flow of people in the infected state in both confirmed and actual cases to forecast the number of F0 people located in China. Hence, Yang et al. has predicted the size and the spread of the infected cases of COVID-19 in China by using Python from late February to early May 2020, and the numerical analysis validates the high degree of predictability of our proposed SaucIER model compared to existing resemblance. Hence, This drives us to formulate a new infectious dynamic model for forecasting the COVID-19 pandemic within the human mobility network. However, this concept made the research team have to forecast the whole results of all factors mentioned during the model.

On the other hand, time series models are also used in forecasting not only in production but also in epidemic problems. These models solely depend on past data to forecast for future events, which are ARIMA, Regression Analysis, Prophet, Neural Network, Bayesian dynamic linear model. Besides, during using this concept, based on what previous authors conducted, Didi et al. stated that ARIMA models, which is one of methods of the time series models concept, can forecast just the specific result. Focussing on modelling and forecasting the number of cases of COVID-19 pandemic in Turkey and G8 countries including Germany, the United Kingdom, France, Italy, Russia, Canada and Japan from January 22nd to March 22nd in 2020 by implementing the curve estimation models, the Box-Jenkins which is ARIMA model, and the Brown/Holt linear exponential smoothing methods in [1] conducted by Yonar, H. and et., this research provides that there are three groups of results which are statistically significant but clinically unqualified results, reliable results and results by using a specified particular model used in Turkey. Therefore, more accurate evaluations can be made with more data in future and lead countries to guide fitter

measure and intervention as early as possible by extending it in future researches. Particularly, next, in paper [18], Ogundokun et al. used a linear regression model to predict the impact of certain factors on COVID-19 outbreak and take the necessary measures to respond to this crisis. This led to the result that based on the data collected from March 31, 2020, to May 29, 2020, they adopted the estimator to measure the impact of travelling history and contacts on the spread of COVID-19 in Nigeria and made a prediction. The result is that the model was conducted before and after travel restriction was enforced by the Federal government of Nigeria. In paper [39], Kumar et al. had used the ARIMA model combining with the Prophet time series forecasting model to simulate the evolution of the COVID-19 outbreak and perform the prediction based on the information of COVID-19 spread for cumulative cases from the whole world and 10 mostly affected countries which are the US, Spain, Italy, France, Germany, Russia, Iran, United Kingdom, Turkey, and India, spread from January 22, 2020, to May 20, 2020. This forecasting activity helped in understanding the trends of the disease outbreak and provide epidemiological stage information of adopted countries. Hence, the ARIMA model was more effective for forecasting COVID-19 prevalence. As paper [17] showed, Mashel et al. had predicted and analysed, so planned the future approach for COVID-19 pandemic about the recovered and death cases in Saudi Arabia by using the time series dependent facebook prophet model. This result aimed to determine the pandemic prediction of COVID-19 in Saudi Arabia, using the Time Series Analysis to observe and predict the coronavirus pandemic spread daily or weekly. Moreover, paper [40] showed that Sulasikin et al. used time series models such as Holt's exponential smoothing and Auto-Regressive Integrated Moving Average (ARIMA) to forecast the number of COVID-19 cases in Jakarta between March 1 and July 6. Hence, ARIMA has the highest R-Squared (R^2), and lowest (Mean Squared Error) MSE and Root Mean Squared Error (RMSE) is the best model to forecast the upcoming number of daily infected cases of COVID-19 in Jakarta, so they can plan the future approach for their current situation. Marco et al. conducted in paper [7] an out-of-sample comparison with forecasts from an autoregressive integrated moving average (ARIMA) model to forecast confirmed new cases of COVID-19 pandemic in Italy from May 19 to June 2, 2020. An out-of-sample comparison with forecasts from an autoregressive integrated moving average (ARIMA) model is considered. This comparison indicates that our procedure outperforms the ARIMA model with the result that Root Mean Square Error (RMSE) of the ARIMA is always greater than that of our procedure and generally more than twice as high as our procedure RMSE confirm that forecasts from our procedure are significantly more accurate at all horizons. In addition, in paper [38], Khan et al. used the model called Bayesian Dynamic Linear Model' (BDLM) for the forecast of daily new infections, deaths and recovered cases regarding COVID-19 with the recursive Kalman filter in Pakistan. This had resulted in the average number of new infections, deaths and recovered cases being 3,282, 52 and 1,840, respectively, in the upcoming 20 days. Therefore, based on this result, authors can suggest a more suitable way to manage the number of cases that would not be able to

reach that level. Moreover, in paper [26], Melin et al. proposed a multiple ensemble neural network model with fuzzy response aggregation for the COVID-19 time series, which are composed of a set of modules that are simple neural networks, and used to produce several predictions under different conditions. Besides, fuzzy logic is then used to aggregate the responses of several predictor modules, in this way, improving the final prediction by combining the outputs of the modules in an intelligent way. Its results showed the trend in Saudi Arabia would continue growing and may reach up to 7668 new cases per day and over 127,129 cumulative daily cases in a matter of four weeks if stringent precautionary and control after successfully implementing previous methods. Additionally, in paper [2], Alzahrani et al. proposed the Autoregressive Integrated Moving Average (ARIMA) model to forecast the expected daily number of COVID-19 cases in Saudi Arabia in the next four weeks, and they claimed that the ARIMA model outperformed the other models by first performing four different prediction models; Autoregressive Model, Moving Average, a combination of both (ARMA), and integrated ARMA (ARIMA), to determine the best model fit. Results showed the trend in Saudi Arabia will continue growing and may reach up to 7668 new cases per day and over 127,129 cumulative daily cases in a matter of four weeks if stringent precautionary and control measures are not implemented to limit the spread of COVID-19. In addition, ARIMA were also used in paper [41] by Didi et al. covered time series data on the daily confirmed and death cases of COVID-19 in Nigeria, obtained from the Nigerian Centre for Disease Control (NDCD) from 21 March 2020 to 5 May 2020, covering a total of 51 data points, by developing a suitable ARIMA models which can be used to forecast total daily confirmed and death cases of COVID-19 in Nigeria. Based on times series data of COVID-19 in Nigeria covering a total of 51 data points, authors showed the existence of two adequate subset ARIMA (2, 2, 1) and AR (1) models for the confirmed and death cases, respectively, is fitted and discussed. Hence, a forecast of 239 days – from 6th May 2020 to 31 December 2020 was conducted using the fitted models and the COVID-19 pandemic data had observed an upward trend and been best forecasted within a short period. Finally, paper [24] written by Kumar et al. using hybrid technique based on self-organised maps and fuzzy time series (SOMFTS), evaluation of COVID-19 forecasting models based on Multi Criteria Decision Making (MCDM) to firstly forecast a COVID-19 spread in the specific area which it compared with and secondly evaluate itself with seven other conventional methods to choose the most efficiency method to achieve its objectives. Therefore, results of the process were observed demonstrating the efficiency of SOMFTS technique for future forecasting of COVID-19 cases and the utility of MCDM methods for evaluation and selection of COVID-19 forecasting models. Therefore, with the same scope, the team decided to choose some suitable methods in both concepts to make a detailed decision to take which one can be used for conducting the final result which is the forecasted COVID-19 cases in Ho Chi Minh city. In contrast, to take a look generally through previous research using these two concepts, our study team has conducted that authors did not include the external factors as the lower important level to forecast the result of

COVID-19 pandemic. Hence, in this paper, the team will choose the most suitable method for us to gain the desired result.

2.3. Current System Investigation:

As shown above, this paper has selected some forecasting methods for implementing in forecasting the spread of COVID-19.

SIR and SIER model with model extensions

This type of model is easy to adapt, to build and to operate because of its simplicity. Additionally, this model is also easy to over complicate and to try on different ideas or hypotheses. Besides, as mentioned in its name, this kind of model also can add extension conveniently.

However, this model still has its own problem. Despite the prediction of infectious cases in short-term intervals, the constructed SIR model was unable to forecast the actual spread and pattern of the epidemic in the long term. Remarkably, most of the published SIR models developed to predict COVID-19 for other communities suffered from the same conformity. The SIR models are based on assumptions that seem not to be true in the case of the COVID-19 epidemic. Additionally, the appearance of asymptomatic individuals contributes more uncertainty to the model.

Neural Network Model

This type of model can be used with numerous data. That means the neural network model is able to give better results when they receive and gather all data and information about the problem. Therefore, users can realise that adding more data during using this model does not improve or impact on the performance.

In contrast, this model still has its gap: Large number of samples are required to achieve desirable result and many parameters to fine-tune which complicating the model

Bayesian dynamic linear model

This model can provide a mathematically coherent framework if users know how to apply productively. Besides, its gap is that Bayesian inferences require skills to translate subjective prior beliefs into a mathematically formulated prior. If the team does not proceed with caution, you can generate misleading results. Additionally, it can produce posterior distributions that are heavily influenced by the priors. From a practical point of view, it might sometimes be difficult to convince subject matter experts who do not agree with the validity of the chosen prior.

ARIMA model

ARIMA (Auto Regressive Integrated Moving Average) model is a general class of Times Series model forecasting concept. It can be viewed as a "filter" trying to separate the signal which is then extrapolated into the future to obtain forecasts.

However its gap is not automatically updating the forecast: as new data become available, the entire modelling procedure must be repeated, especially with the stationarity and diagnostic checking stage.

Regression Analysis

This method determines the factors that matter most, the factors that can be ignored, and how they interact with each other. Therefore, it provides a powerful role allowing users to examine the relationship among variables.

In contrast, the problem of this method is that it is susceptible to over-fitting, linear regression assumes a linear relationship between dependent and independent variables + not a complete description of relationship among variables.

Prophet model

This method can accommodate seasonality with multiple periods. It is also resilient to missing value and fit of the model is fast. Finally, Prophet has intuitive hyper parameters which are easy to tune.

However, the gap of this model is higher forecasting error than other conventional time series models such as ARIMA, SOMFTS etc.

SOMFTS

This type of method is not widely used. Hence, the team was not able to fully assess the accuracy and the applicability of the model and complicate to implement as it required the combination of many models to forecast

2.4. Design Concepts Consideration

The research group decided to suggest areas of improvement in this model, the whole team included more parameters into the model to provide better forecast accuracy, which was not mentioned in most of the previous works, this might include vaccination related data (vaccination rate, vaccination efficacy) and the control measures proposed by the government.

The student team decided to use time series model to forecast the spread of COVID-19 based on different control measures and vaccination factors combining with external factors such as control measures and vaccination rate daily are included in the model based on divided data are divided, in order to forecast according to different control measures proposed by the Ho Chi Minh City's board of director and the government.

Chapter III. MODEL DESIGN AND ANALYSIS

3.1. Approach comparison and selection

3.1.1. Approach consideration:

There are four potential approaches for making a COVID-19 forecasting model that these following models made the team interested, which are: Autocorrelation Integrated Moving Average model (ARIMA), multiple neural network with fuzzy response aggregation, self-organized maps and fuzzy time series (SOMFTS), and Bayesian dynamic linear model.

Autoregressive integrated moving average model (ARIMA)

ARIMA is actually a class of models that explains a given time series based on its own past values. That is, its own lags and the lagged forecast errors can be used to forecast future values. The ARMA models can further be extended to non-stationary series by allowing the differencing of the data series. The general non-seasonal model is known as ARIMA (p, d, q), where p is the order of autoregressive, d is the degree of differencing, and q is the order of moving average.

In several previous studies, ARIMA provided high forecast accuracy in conditions with high frequent data. Specifically, the model requires at least 50 and preferably 100 observations to be built properly. It is less sensitive to the underlying assumptions of the nature of the data fluctuations than many other systems. Besides, this model often outperforms more sophisticated structural models in terms of short-run forecasting ability. However, the research team hardly automatically updated the forecast as new data became available, the entire modelling procedure must be repeated, especially with the stationarity and diagnostic checking stage.

Multiple neural network with fuzzy response aggregation

Neural networks are composed of a set of modules, which are used to produce several predictions under different conditions. Fuzzy logic is then used to aggregate the responses of predictor modules to manage the uncertainty of the individual networks. In the case of Mexico city, prediction errors of the multiple ensemble neural networks are significantly lower than using traditional monolithic neural networks.

The advantages of a neural network with fuzzy response is that there are no prerequisites for the data we use in the model. The non-linear function can be modelled with neural networks too.

The only drawback of the neural network is that it requires a lot of data (big data) and many parameters to achieve desirable results.

Self-organized maps and fuzzy time series (SOMFTS)

Self-organized maps and fuzzy time series is a forecasting model based on hybrid self-organizing map (an unsupervised learning neural network) and fuzzy time series. This approach uses a log-polynomial model together with a first-order integer-valued autoregressive (INAR(1)) model. The log-polynomial model is used to forecast the ratio number of daily new diagnosed cases/number of swabs and the INAR(1)) model is used to forecast the number of swabs. The forecast of the number of daily new diagnosed cases is then obtained by multiplying these two forecasts.

The advantage of the approach is that it explicitly takes into account the number of swabs. In fact, it is necessary to consider the number of swabs to capture the fluctuations of the number of daily new confirmed cases around the trend. Nonetheless, it was unable to fully assess the accuracy and the applicability of this model approach since it is not widely used. It is also complicated to implement as it requires the combination of many models to forecast.

Bayesian dynamic linear model

The Bayesian framework allows for updating prior knowledge about the quantity of interest using the observed data and calculates posterior distribution for the quantity of interest. Bayesian inferences are derived from the posterior distributions of quantities of interest, which are used for projections and their corresponding credible intervals.

The Bayesian framework provides computational power via the Markov chain Monte Carlo methodology to provide an exact estimate of the quantity of interest, rather than using approximate optimization algorithms. On the other hand, Bayesian inferences require skills to translate subjective prior beliefs into a mathematically formulated prior. Therefore, if you do not proceed with caution, you can generate misleading results. Furthermore, it can produce posterior distributions that are heavily influenced by the priors. From a practical point of view, it might sometimes be difficult to convince subject matter experts who do not agree with the validity of the chosen prior.

3.1.2. Selection of the final approach by quantitative comparison

Fuzzy analytic hierarchy process (fuzzy AHP) is used to determine the best forecasting approach among four alternatives (A1, A2, A3, and A4). The student group took into account three criteria: the accuracy (C1), the flexibility (C2), and the simplicity (C3) of the forecasting model. Regarding the accuracy criteria, the smaller the forecast error in short-term horizon, the higher score of AHP will be given for the alternative [43]. In respect of the flexibility criteria, the

capability of the model in terms of modification and combination with different methods and COVID-19 parameters were considered. Lastly, the simplicity criteria represents how effortlessly the model is implemented. The fuzzy scales are defined in detail in the table below.

C1	Accuracy
C2	Flexibility
C3	Simplicity

A1	ARIMA (autoregressive integrated moving average)
A2	Multiple neural network with fuzzy response aggregation
A3	SOMFTS (self-organized maps and fuzzy time series)
A4	Bayesian dynamic linear model

Fuzzy Scale	Definition Fuzzy	Triangular Scale
1.00	Equal	(1, 1, 3)
3.00	Moderate Important	(1/5, 1/3, 1), (1, 3, 5)
5.00	Important	(1/7, 1/5, 1/3), (3, 5, 7)
7.00	Very important	(1/7, 1/5, 1/9), (5, 7, 9)
9.00	Extremely	(1/9, 1/9, 1/7), (7, 9, 9)

Table 1: fuzzy scales for model selection using fuzzy Analytic Hierarchy Process (AHP)

Using fuzzy AHP techniques [44], the team decided to choose ARIMA to be the final approach to forecast the confirmed case in the future. This is because ARIMA ranks first place among the four alternatives and it outstandingly satisfies all criteria defined at the beginning of fuzzy AHP analysis.

3.2. System design description

3.2.1. Flow chart of the forecasting model

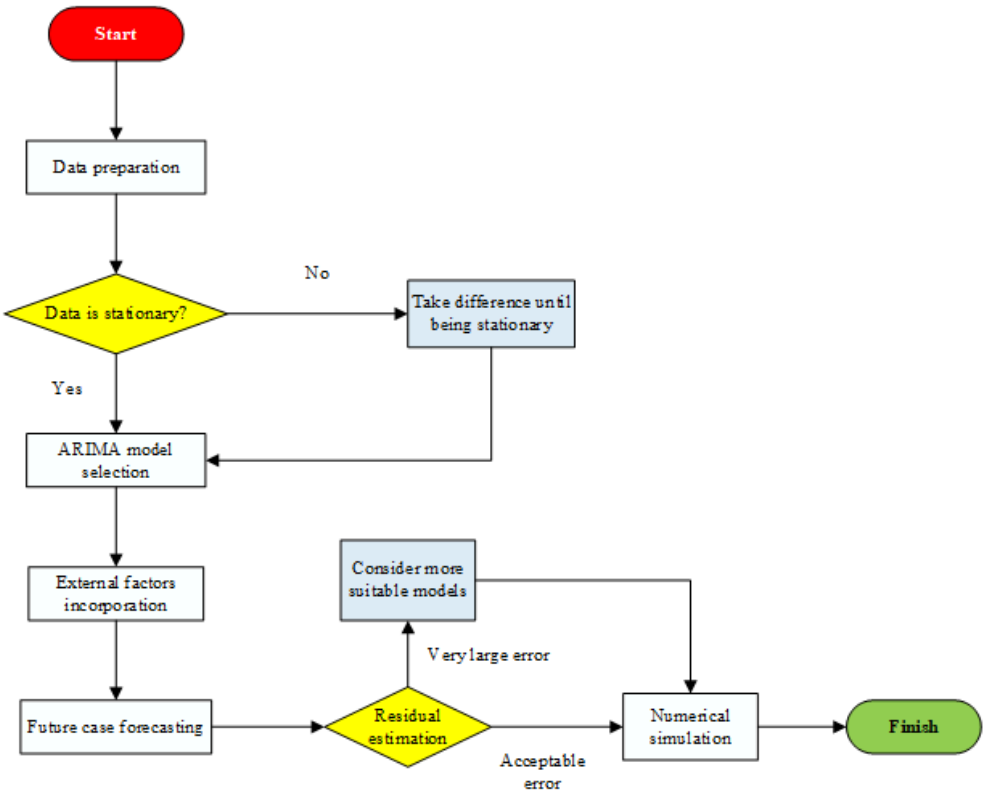


Diagram 1: flowchart of the forecasting process established in this paper

3.2.2. Design structure of the proposed system

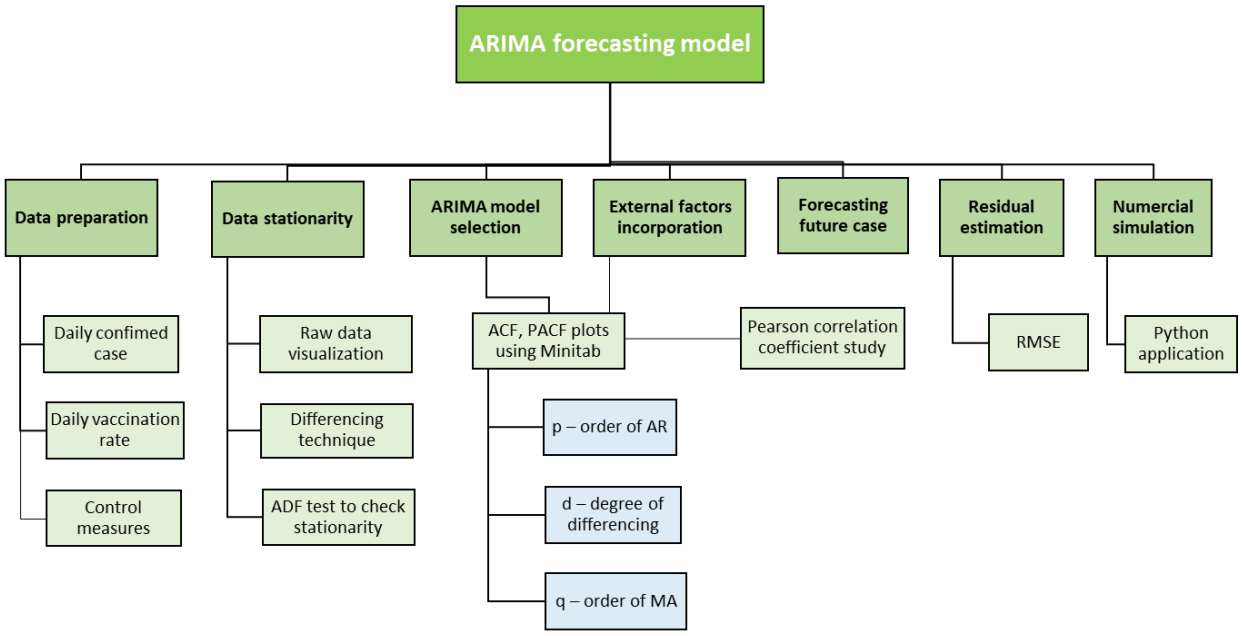


Diagram 2: Detail diagram of ARIMA forecasting model used in this paper

The implementation of ARIMA model used to forecast the future number of confirmed cases consists of seven major steps as following:

Data preparation

In this model, the team used the historical data of confirmed cases in Ho Chi Minh city as an input to generate the output of the model, which is the future prediction of the confirmed case of the city. According to our observations, the number of new infected cases dwindled since the government issued stricter guidelines such as Directive No. 16 or Directive No. 15. Therefore, along with the daily confirmed case, the student group were also interested in the control measures of the government with different pandemic situations and the daily vaccination rate associated with each control measure. The latter two factors will be included into the ARIMA model as external factors that contribute to the spread of COVID-19 in the Ho Chi Minh community. Due to the information resource constraints, the daily data were collected from various websites, and the research group synthesised them into a merged file to ensure the availability of data needed for the forecasting model. For the daily new confirmed case data, the group mainly gathered information from the official portal of Vietnam Ministry of Health and Ho Chi Minh city Health Department – Centres for disease controls. In case, the data is unavailable from those two official websites, gathering information was implemented from online newspapers like Baomoi, Thanhnien, or Tuoitre. Before using the data for forecasting, several hypothesis tests to check the validity of the data such as the Benford law distribution test, Chi-square test, Kolmogorov – Smirnov test, etc were conducted. During the implementation of the model, the daily confirmed case data will be updated regularly.

Data stationarity checking

One of the big deals in time series forecasting models is to make sure the time series, stationary modelling were applied. In other words, the properties of the time series such as the mean and the standard deviation are stable over time. Firstly, the team used the Minitab application to visualise the data over time. That time series plot basically gives us the overview of the data. In this case, the time series exhibits some particular patterns namely trend, seasonal factor, differencing techniques to eliminate those patterns and transform the time series into a stationary one were used. Although the team could judge the stationarity of a time series through its plot, it was necessary to perform a robust formal test in order to better verify it. In this model, the Augmented Dickey-Fuller test (ADF test) was used to investigate whether the historical daily confirmed case the group collected was stationary or not [42]. For each time the student team took the difference, an ADF test was applied to judge the stationarity of the data until it satisfies the initial condition, which is unchanged over time.

ARIMA model selection

In the paper, they use the Box-Jenkins approach for ARIMA methodology. Once the historical time series data was successfully transformed into a stationary one and its validity was checked

by the ADF test, the data is ready for modelling. At this step, Minitab was used to generate the auto correction function (ACF) and partial autocorrelation function (PACF) figures. By applying rules, the research team could spot the special characteristics of the ACF and PACF plots to identify the appropriate ARIMA (p, d, q) model, where p is the order of autoregressive, d is the degree of differencing, and q is the order of moving average. It is possible that there are more than one tentative ARIMA models associated with the given time series. In this case, it could be going to test all possible models and choose the one that gives the highest forecast accuracy.

External factors incorporation

Beside the coefficients of the function, the external factors were also estimated such as the vaccination rate and the control measures to include into the forecasting model. Those parameters improve the forecasting accuracy since they are critical factors that determine the number of new confirmed cases. When estimating the external factors in the ARIMA model, a Pearson correlation coefficient study [19] was used as most of the time there will be a linear relationship between these two components. Pearson's correlation coefficient measures the linear relationship between two variables. Its value ranges from +1 to -1, with 1 representing total positive linear correlation, 0 representing no linear correlation, and -1 representing entire negative linear correlation. The implementation of external factors using Pearson correlation coefficients consists of four stages: propose external factors, Pearson correlation study, choose optimal timeframe & lag and algorithm application. Detailed calculation for each step will be discussed in the model development section.

Forecasting future case

When all coefficients and parameters are estimated, an ARIMA function could be built that represented the time series data and used it to forecast the future case. In terms of y, the general forecasting equation is:

$$y_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}.$$

Considering the external factors by Pearson correlation coefficient, the team could generally get the model for ARIMA (p, d, q) as following:

$$y(t) = ARIMA(p, d, q)(t) + \beta 1 \cdot x1(t-1) + \beta 2 \cdot x2(t-1) + \beta 3 \cdot x3(t-1) + \beta 4 \cdot x4(t-1)$$

Residual estimation

After using the model to forecast the future confirmed case, the group was going to check the accuracy of the forecast results by four different methods. At this step, RMSE (root mean square

error) was used to measure the prediction accuracy [45] of the forecasting model. This is not the only measure method for all time series issues but it is a suitable fit for our situation because our prediction is interested in a numerical end goal value. In case, the team developed more than one potential ARIMA model, the most suitable ARIMA model to be used to forecast would be the one that has the lowest forecast deviation and fits well with the observed data. If the residual of the model is too large, it is vital that the order p, d, q of ARIMA and estimate the relevant parameters were reconsidered. The research team was going to consider including more significant factors into the model or eliminate the redundant ones to enhance the forecast results.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{\left(\hat{y}_i - y_i\right)^2}{n}}$$

Numerical simulation

In this study, a Python code was applied to conduct the numerical simulation of the ARIMA model's forecasting results. The aim of the simulation is to visualize the fitness level of the ARIMA forecast with the actual historical data in Ho Chi Minh city over the last pandemic waves. The simulation enables us to quickly grasp the trends of COVID-19 spread as well as the accuracy of different ARIMA models. Using the simulation result, combined with the calculated forecast errors in the previous part, it could be concluded which model is the best model for Ho Chi Minh city.

3.2.3. Analyze and justify the techniques to be used

A. *Techniques for assessing data stationarity and finding p, q parameters of ARIMA model:*

Autocorrelation function (ACF) & partial autocorrelation function (PACF)

The autocorrelation function (ACF) is a statistical approach for determining how closely values in a time series are connected to one another. The ACF depicts the correlation coefficient versus the lag, which is expressed in terms of periods or units. The first value in the time series is observed after a lag, which corresponds to a certain point in time. Partial autocorrelation (PACF) is a statistical measure that captures the correlation between two variables after controlling for the effects of other variables. In other words, a PACF captures a "direct" correlation between time series and a lagged version of itself. The reason for using this technique [46] is that many previous papers, books and other academic materials have widely applied ACF and PACF to estimate the AR(p) and MA(q) parameters of the model. It is claimed that the techniques help avoid clustering and increase forecasting accuracy.

ADF test

ADF (Augmented Dickey-Fuller) test is a statistical significance test which means the test will give results in hypothesis tests with null and alternative hypotheses. As a result, the team would have a p-value from which would be needed to make inferences about the time series, whether it is stationary or not. So if a time series is non-stationary, it will tend to return an error term or a deterministic trend with the time values. If the series is stationary, then it will tend to return only an error term or deterministic trend. In a stationary time series, a large value tends to be followed by a small value, and a small value tends to be followed by a large value. And in a non-stationary time series the large and the small value will accrue with probabilities that do not depend on the current value of the time series.

Hypothesis of the test:

- **H0**: the time series has unit root (unstationary)
- **H1:** the time series is stationary

The augmented dickey- fuller test is an extension of the dickey-fuller test [42], which removes autocorrelation from the series and then tests similar to the procedure of the dickey-fuller test. The augmented dickey fuller test works on the statistic, which gives a negative number and rejection of the hypothesis depends on that negative number; the more negative magnitude of the number represents the confidence of presence of unit root at some level in the time series. The ADF test is used in other papers to identify the stationarity of a certain set of data, and to eliminate the case of unit root, where the non-stationary data cannot be identified by looking at the data plot.

B. Data validation techniques:

Benford's law – First- & Second-digit test

Benford's law, which can be called Newcomb-Benford Law, was initially observed by Newcomb, an astronomer-mathematician in 1881 and popularised by Benford half a century later. Benford's law is applied in many aspects of our daily life and in businesses, as well as in research and development to assess the authenticity of the data in these fields. Most of these data are for financial assessment, accounting fraud, politics and designing [29], [34]. The Law provides the mathematical approach in which it specifies the frequency distribution of leading digits or significant digits. The data is considered to be suspicious if the observed frequencies of the data deviate greatly from Benford's distribution [32].

Specifically, with base b ≥ 2 In numbers collected, Benford's law provides an expected frequency or distribution of the first leading digits **d {1, 2, . . ., b − 1}.** The observed frequencies of the data will then be collected and test their authenticity of the data using goodness-to-fit tests such as Chi-square, Kolmogorov Smirnov test and statistical measures to identify how large of the deviation in the observed frequencies from the predicted distribution according to Benford's law. The predicted distribution is provided using the following equation [30], [31], [32].

$$P_b(d) = \log_b\left(1 + \frac{1}{d}\right), d = 1,2,3, \ldots,9$$

Where b is the base (b ≥ 2), d is the leading digit considered in the data

In real life application, Benford's law has been a reliable technique to assess data authenticity in many fields, especially in social science and various disciplines such as finance and accounting, politics, and pandemics [31]. Researchers also rely on Benford's law and its application to identify data fraud in non-financial fields as well, including electoral fraud, fraudulent scientific data, suspicious social media activity, falsified law enforcement statistics, and fraud in international trade. Hence, it is a good starting point to use Benford's law in fraud detection in epidemiology. Additionally, it is a good reference to consider applying Benford's law to evaluate the data reliability of public health surveillance systems since a reliable epidemiological surveillance system is a crucial factor in coming up with suitable control measures and responses to minimize the impact of the ongoing pandemic [32].

To perform data assessment and determine its reliability, several digit tests were performed based on Benford's law and its distribution. Here, **only the first-order test was considered since this method provides information on data manipulation**. The first-order tests include the first-digit, second-digit and the first two digits tests. The first-order tests are usually run on the positive or negative numbers, but not both in the same analysis. The reason is the incentive to data manipulation is opposite in positive and negative data collected. The first digit test computes the frequencies of the first digit in each data (or observation) in which the first digit d ranges from 1 to 9 and compares the observed frequencies with the expected ones appeared in Benford's distribution to detect anomalies. Additionally, **the second digit test, also a high-level test, is designed to perform data assessment for conformity or reasonableness to Benford's distribution**. The second-digit d ranges from 0 to 9. It is important to note that the expected frequencies in second-digit are less skewed than the expected ones of first-digit [28].

The following table is the distribution of frequencies of first- and higher-order significant digits. In this paper, only the probabilities of first and second significant digits were considered.

Probabilities of first- and higher-order significant digits				
Digit	Probability of first significant digit	Probability of second significant digit	Probability of third significant digit	Probability of fourth significant digit
0	not applicable	11,97%	10,18%	10,02%
1	30,10%	11,39%	10,14%	10,01%
2	17,61%	10,88%	10,10%	10,01%
3	12,49%	10,43%	10,06%	10,01%
4	9,69%	10,03%	10,02%	10,00%
5	7,92%	9,67%	9,98%	10,00%
6	6,70%	9,34%	9,94%	9,99%
7	5,80%	9,04%	9,90%	9,99%
8	5,12%	8,76%	9,86%	9,99%
9	4,58%	8,50%	9,83%	9,98%
Source: Using Benford's Law to detect data error and fraud: An examination of companies listed on the Johannesburg Stock Exchange, 2006				

Table 3: Probabilities of first- and higher-order significant digits

Chi-square test

To evaluate the suitability of a potential input data to a certain distribution, goodness-to-fit test is a tool that is most commonly used. However, in real life situation, data distribution will never follow the exact distribution that the group wanted them to be, therefore, it was advisory that the test should only be taken as references and in the case of rejecting a certain hypothesis that the data is distributed according to certain distribution, it should only be taken as one piece of evidence in favour of that choice. Additionally, it is crucial to understand the impact a sample size can have on a result of a goodness-to-fit test. According to [35], if the sample size of data is too small, meaning little data are available, the test would likely to accept any candidate distribution; but when the sample size is very large, a goodness-to-fit test will be more likely to reject all candidate distributions.

Among goodness-to-fit tests available such as Kolmogorov–Smirnov and Chi-Square test, Chi-square test is more likely to be implemented when it comes to evaluating the conformity of data with Benford's distribution. In Chi-square test, the hypothesis is tested in which a random sample with size n of the random variable x follows a specific distributional form, in this case in Benford's distribution. The test is performed by comparing the histogram of the data to the shape of the candidate density or mass function. The test is considered valid for large sample size, usually n and for both discrete and continuous distribution. The test statistics is given as the following, in the context of Benford distribution [30], [35]:

$$X^2 = \sum_{i=1}^{9} \frac{(\tilde{P}(i)-P(i))^2}{P_i} \text{ for first significant digit}$$

$$X^2 = \sum_{i=0}^{9} \frac{(\tilde{P}(i)-P(i))^2}{P_i} \text{ for second significant digit}$$

Where P̃(i) is the observed frequency of the sample size at digit i, P(i) is the expected Benford frequency at digit i. The expected frequency of Benford's first and second significant digit is demonstrated above in table 1.

MAD and SSD

In addition to the evaluation of data manipulation, MAD and SSD are applied into this study besides goodness-to-fit Chi-square test. These statistical measures were performed in several papers when it comes to assessing data authenticity. Farhadi and Lahooti [30] used MAD and SSD to evaluate data manipulation of COVID-19 confirmed cases, deaths and recover cases of 182 countries. MAD is the Mean Absolute Deviation, which is the average absolute deviation of the observed from the expected frequencies while SSD is the Sum of Square of Deviation in each frequency of digits. MAD and SSD are expressed as the following calculation:

$$MAD = \frac{1}{9} \times \sum_{i=1}^{9} |O_i - E_i| \; for \; first - digit \; test$$

$$MAD = \frac{1}{10} \times \sum_{i=1}^{10} |O_i - E_i| \; for \; second - digit \; test$$

$$SSD = \sum_{i=1}^{9} (O_i - E_i)^2 \times 10^4 \; for \; first - digit \; test$$

$$SSD = \sum_{i=1}^{10} (O_i - E_i)^2 \times 10^4 \; for \; second - digit \; test$$

Where O_i are the observed frequencies of first- and second-digit tests, E_i depicted as the expected frequencies in first- and second-digit tests of Benford's distribution. d is depicted as the number of leading digit bins (for first-digit, d = 1, 2,…9 and for second-digit, d=0,1,2,…9). For MAD and SSD measures. It is emphasized that a MAD < 0,015 and an SSD < 100 indicate conformity with Benford's Law [30].

Euclidean distance

When implementing Chi-square test as a goodness-to-fit test of Benford's distribution, Euclidean distance (d*) should be considered as it measures the distance between the observed and the expected frequencies so that researchers can have a more overview of the data's leading digits distribution. d* distance quantifies the distance between the sample and the cumulative Benford distribution for first and second digits after normalisation by 1.03606, which is the maximum

possible distance of the measured and expected distribution. Specifically, d* equals to 0 means the data suggest full-conformity with Benford's law, while d* equals 1.0 depicts full non-conformity with the law. In epidemiological study, especially in the COVID-19 pandemic, d* < 0.2 indicates data conformity with Benford's distribution [29], [30]. Euclidean distance is calculated as the following,

$$d^* = \frac{\sqrt{\sum_{d=1}^{9}\left(\tilde{P}(d) - P(d)\right)^2}}{1.03606} \; for\ first-digit\ test$$

$$d^* = \frac{\sqrt{\sum_{d=1}^{10}\left(\tilde{P}(d) - P(d)\right)^2}}{1.03606} \; for\ second-digit\ test$$

$\tilde{P}(d)$ stands for the probability distribution of each first digit in real datasets, P(d) depicts the cumulative distribution of Benford's law [29].

3.2.4. Key advantages of the proposed system design

ARIMA model has the advantage of being less sensitive to the underlying assumptions of the nature of the data fluctuations than many other systems. For short-run forecasts with high frequency data the results may be hard to beat. The most important advantage of ARIMA includes dealing with small data. This is critical since the data the research group were going to use in the forecast of Ho Chi Minh city's confirmed case is considered to be small. Data from the fourth wave of the pandemic were collected, which roughly has more than 200 observations.

Secondly, the ARIMA model is simple to implement with no parameter tuning. Consequently, it is quick to run the forecast, compared to other possible approaches. Lastly, the ARIMA model is easier to handle multivariate data and handle several external factors that contribute to the number of daily confirmed cases in Ho Chi Minh city.

Chapter IV: PROTOTYPE DEVELOPMENT AND IMPLEMENTATION

4.1. Model Development

ARIMA Parameter identification:

This model will integrate the ARIMA forecasting models with other direct external factors to generate better forecasting results. Therefore, there are two kinds of parameters the team was going to estimate: pure ARIMA model parameters and the related external factors parameters. Detailed descriptions of each kind are clearly discussed in the subsection below

ARIMA stands for Autoregressive Integrated Moving Average model, a statistical analysis model that is widely used in either manufacturing or service sectors. The major objective of the ARIMA model is to predict the future values of a particular variable through investigating its historical data. The ARIMA model can only be implemented when the time series is stationary or unchanged over time. An ARIMA model can be understood by outlining each of its components as follows:

- The Autoregressive (AR) represents the difference between an observed value of the time series and its adjacent value in a defined time lag.
- The integration (I) represents the number of differencing times to transform the non-stationary time series into a stationary one.
- The Moving average (MA) represents the difference between an observed value and a residual error in a defined time lag.

Three components of the ARIMA model are notated with three distinct notations p, d, q, which are corresponding with the order of AR, differencing, and MA respectively. While the order of d term can be easily estimated by counting the number of differencing times, the order of AR and MA terms are determined by using statistical tools such as Microsoft Excel or Minitab application. In particular, the team used Minitab to generate the ACF and PACF plots of the stationary time series. Subsequently, the group estimated the order of p and q with the validated rules as shown in the table below.

	ACF plot	**PACF plot**
White noise	All zero	All zero
MA (q)	Drop off after lag q	Die down
AR (p)	Die down	Drop off after lag p
ARMA (p, q)	Die down	Die down

Table 9: rules for estimation of p and q order in ARIMA model [46]

External factor parameter estimation

In this part, the research group was interested in defining the key external variables that significantly contribute to the number of daily new confirmed cases in Ho Chi Minh city. In our

case, the team considered four direct external factors, namely daily vaccination coverage, social distancing policy in accordance with control measures, the Covid variant incubation time and Covid variant generation time. The group used the Pearson correlation coefficient methodology to seek the relationship between external factors and the main forecasting variables, which is the daily confirmed cases. The external factors data were divided into different time frames and time lags to be analysed. The optimal time frame and time lag was the one that gives the highest average correlation coefficient of the four external factors. The Pearson correlation coefficient is calculated using the following formula:

$$\rho\,(x, y) = \frac{cov\,(x,y)}{\sigma x^{*}\sigma y}$$

where:

- (x, y) is a pair of random variables used in correlation study
- Cov is the covariance
- σx is the standard deviation of variable x
- σy is the standard deviation of variable y

Based on the correlation coefficients (ρ) of each external factor, the team can come up with the suitable weights (β) of each factor. The weight of each external factor is calculated by dividing its Pearson correlation coefficient with the sum of all external factor correlation coefficients. The sum of all weighted parameters of external factors must be equal to 1. The general external factor model can be described as the equation below:

$$y_{external}\,(i) = \beta_1\, x_1\,(i - t) + \beta_2\, x_2\,(i - t) + \beta_3\, x_3\,(i - t) + \beta_4\, x_4\,(i - t)$$

where

- y(i) is the predicted value for the four external factors
- β1 is the weighted parameter for the daily vaccination coverage variable
- β2 is the weighted parameter for the social distancing policy variable
- β3 is the weighted parameter for the Covid variant incubation time variable
- β4 is the weighted parameter for the Covid variant generation time variable
- t is the time lag of Pearson correlation coefficient study

External parameter generation:

Considering covid variant, it is true that its transmission rate is one of the most important things to pay attention to in order to understand the characteristics of a certain variant and the seriousness of a certain COVID-19 wave so as to increase effectiveness of

control measures. Specifically, the transmission dynamics of COVID-19 variants are defined by a certain number of key epidemiological measures. They can be considered as the incubation period, generation time, and the reproduction number [36].

In this paper, only 2 epidemiological parameters are considered in transmission rate: incubation period and generation time as they are easy for estimation but still represent the dynamic of COVID-19 variant in Ho Chi Minh City – Delta variant. According to [11],[22],[36], incubation period is the time period in which the exposed individual is infected and starts to present illness onset in a defined period of time. And generation time is considered to be the time interval between infection of the primary case and its secondary cases.

The following flowchart is the steps need to be taken so as to find incubation period and generation time:

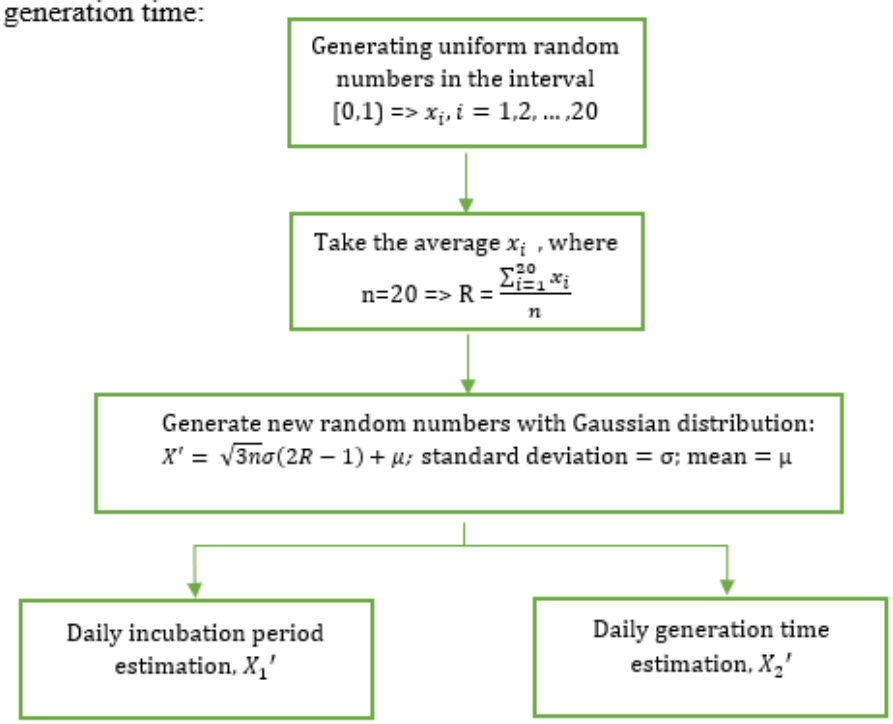


Diagram 3: flow chart of generation time and incubation period estimation process.

According to Pirooz Mohazzabi, Michael J. Connolly [12] has introduced a new algorithm for generating random numbers with a specified normal or Gaussian probability density function which is based on the central limit theorem to generate pseudo-random numbers. The new algorithm efficiency is said to fall between those of the Box-Muller and von Neumann rejection methods.

Central limit theorem:

According to [13], the means of a random sample of size (n), from a population with mean (μ) and variance (σ^2), is normally distributed with mean μ, and variance $\frac{\sigma^2}{n}$. Specifically, as the sample size n from the population increases, its mean gathers more closely around the population mean with a

decrease in variance. Thus, as the sample size approaches infinity, the sample means approximate the normal distribution with a mean, μ, and a variance, $\frac{\sigma^2}{n}$.

Pseudo-random numbers:

According to [16], "Pseudo" is used to imply the act of generating random numbers through a known method that removes the potential for true randomness. When the method used is known, the set of random numbers can be replicated. The goal of this method is to produce a sequence of numbers between 0 and 1 that simulates or imitates the ideal properties of uniform distribution and independence as closely as possible.

Paper [15] provides some well-known Pseudo-random number generators such as mid-square method, linear congruential method (LCM), combined linear congruential generators and random number streams.

A Pseudo–Random Number Generator (RNG) is defined by a structure (S, m, f, U g) where:

- S is a finite set of states.
- μ is a probability distribution on S, called the initial distribution.
- A transition function f: S → S.
- A finite set of output symbols U.
- An output function g: S → U

Then the generation of random numbers is as follows:

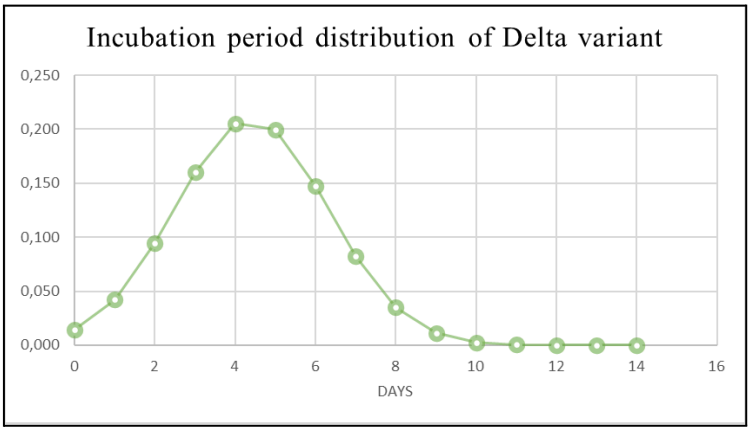
- Generate the initial state (called the seed) s_0 according to μ and compute $u_0 = g(s_0)$
- Iterate for i = 1, 2, 3, …, $s_i = f(s_{i-1})$; $u_i = g(s_i)$

From [12], it is known that the expected value of average random number (from 0 to 1) in n observations that are uniformly distributed is 0.5. The team only needs to generate 20 random numbers x to have the average of x is approximately 0.5. Let's call the average of random number x is R. After which, the team used the below equation to calculate incubation period and generation time, which are normally distributed.

$$X' = \sqrt{3n}\sigma(2R - 1) + \mu$$

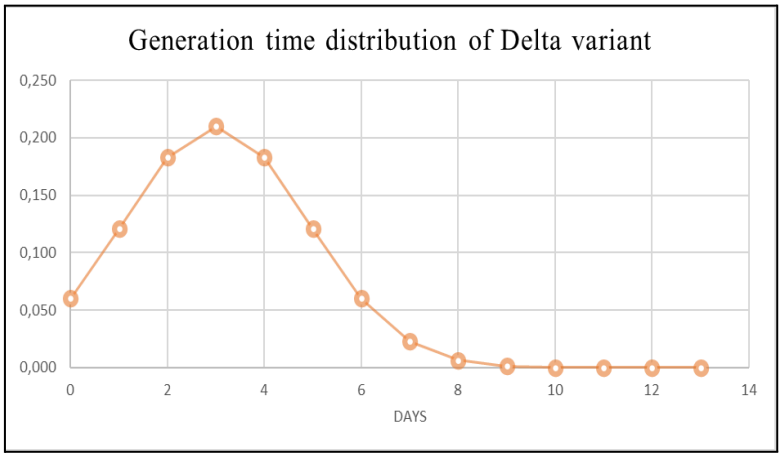
Additionally, [11] has provided the following data of incubation period and generation time of Delta variant through their studies:

Incubation period	Min	Max	Mean	Median	SD	95% CI	
Range	1	14	4,4	4	1,9	3.9-5	days



Graph 2a: Incubation period distribution of Delta variant. Source: [11]

Generation time	Min	Max	Mean	Median	SD	95% CI	
Range	1	14	3	3	1,9	2,4 - 3,3	days



Graph 2b: Generation time distribution of Delta variant. Source: [11]

From the algorithm proposed in the mentioned paper [12], the group obtains the results of incubation and generation time daily estimation, from 27 April, 2021 to 23 November, 2021.

Data collection:

As mentioned in the previous part, the ARIMA model generates the forecast value based on the historical data. Hence, the primary data that needs to be collected is the past daily confirmed cases of Ho Chi Minh city. In this study, the research group was taking into account only the new confirmed cases from the beginning of the fourth wave, which is on 27th April, 2021. This kind of data is the most accessible data yet inadequate and it may pose some confusion between the information sources. The research group collected confirmed case data in a daily time frame and recorded it in a spreadsheet.

The team can gather the daily vaccination rate at Ho Chi Minh city and the team did not discriminate between the first dose and the second dose. The daily vaccination rate was used for the calculation of the vaccination coverage. Thereafter, the vaccination coverage was

used for the Pearson correlation study. This kind of data was also collected from 27th April, 2021 and recorded in the same spreadsheet with the confirmed cases.

Another external factor that can be collected directly is the social distancing factors, which is the allowed distance between people when they are in public places. The numerical value of this factor was derived from the corresponding control measures of each pandemic timeline. Since the government and local authorities changed the control measures several times, depending on the seriousness of the situation, it was necessary to update the notification regularly in order to identify the right value of the social distancing factor.

As the social distancing external factor varies depending on the corresponding control measure issued by the government, the team divides this type of data into several timelines. There have been nine time intervals so far since the first date of the fourth wave of the epidemic in Ho Chi Minh city. Even with the same control measure, there is at least a light difference in the content of each. The detailed timeline for each control measures are listed as the below table.

Control policy number	**Time interval**
15+	27/4/2021 - 30/5/2021
15+	31/5/2021 - 18/6/2021
10	19/6/2021 - 8/7/2021
16	9/7/2021 - 22/8/2021
16+	22/8/2021 - 15/9/2021
11	16/9/2021 - 30/9/2021
18	1/10/2021 - 24/10/2021
New normal, living with the COVID-19 pandemic	25/10/2021 - 17/11/2021
New normal, living with the COVID-19 pandemic	19/11/2021 - present

Table 10: Control policies proposed by Ho Chi Minh city during the fourth wave of COVID-19 pandemic (27th April, 2021 - 24th November 2021)

4.2. COVID-19 Data validation

Our paper only considers confirmed cases in Ho Chi Minh City, Vietnam as data to be evaluated for authenticity with 0,05 significant level [30]. The data is collected according to daily confirmed cases from 27th April, 2021 to 23rd November, 2021 with 211 days (observations)

during the period. Daily confirmed cases in Ho Chi Minh city are collected from COVID-19 Information Portal of Ho Chi Minh city (website: https://covid19.hochiminhcity.gov.vn/web/covid-en). Specifically, two Benford's tests will be performed to test data validity and authenticity: the first- and second-digit tests in order to thoroughly assess the reliability of data. Note that the sample data can be accepted as reliable or partially reliable to be implemented into the forecasting model if it satisfies at least one significant digit test. The team test based on the hypothesis where the confirmed cases data conforms with Benford's law is considered as H0.

Table 4 provide the frequencies in which how many times a certain digit occurs in the dataset

Data value	Second digit									
First digit		0	1	2	3	4	5	6	7	8	9
	1	8	6	5	11	5	5	4	2	1	1
	2	3	1	3	1	4	0	1	1	2	0
	3	1	1	0	4	1	2	4	3	2	3
	4	4	7	2	2	1	2	3	0	0	3
	5	1	3	1	3	5	2	1	1	3	3
	6	1	1	2	1	1	1	2	0	1	1
	7	0	3	0	2	1	1	1	0	1	1
	8	0	0	0	0	1	1	1	0	0	0
	9	3	2	1	1	0	1	3	1	4	0
	sum	21	24	14	25	19	15	20	8	14	12

Table 4: frequencies in times of occurrence of each digit in first- and second-significant digits of daily COVID-19 confirmed cases in Ho Chi Minh city.

From table 4, the group calculates the frequencies in percentage accordingly and the last row and last column in table 5 indicates first- and second-digit observed frequencies of daily confirmed cases in Ho Chi Minh city.

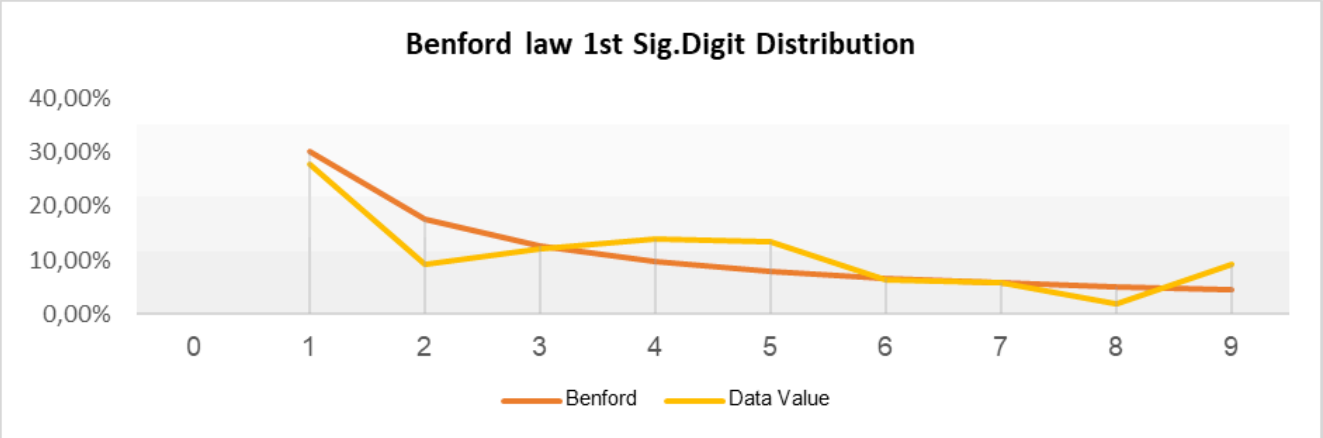
Data value	Second digit										
First digit	0	1	2	3	4	5	6	7	8	9	sum % for 1st digit
1	4,651%	3,488%	2,907%	6,395%	2,907%	2,907%	2,326%	1,163%	0,581%	0,581%	27,91%
2	1,744%	0,581%	1,744%	0,581%	2,326%	0,000%	0,581%	0,581%	1,163%	0,000%	9,30%
3	0,581%	0,581%	0,000%	2,326%	0,581%	1,163%	2,326%	1,744%	1,163%	1,744%	12,21%
4	2,326%	4,070%	1,163%	1,163%	0,581%	1,163%	1,744%	0,000%	0,000%	1,744%	13,95%
5	0,581%	1,744%	0,581%	1,744%	2,907%	1,163%	0,581%	0,581%	1,744%	1,744%	13,37%
6	0,581%	0,581%	1,163%	0,581%	0,581%	0,581%	1,163%	0,000%	0,581%	0,581%	6,40%
7	0,000%	1,744%	0,000%	1,163%	0,581%	0,581%	0,581%	0,000%	0,581%	0,581%	5,81%
8	0,000%	0,000%	0,000%	0,000%	0,581%	0,581%	0,581%	0,000%	0,000%	0,000%	1,74%
9	1,744%	1,163%	0,581%	0,581%	0,000%	0,581%	1,744%	0,581%	2,326%	0,000%	9,30%
Sum % for 2nd digit	12,21%	13,95%	8,14%	14,53%	11,05%	8,72%	11,63%	4,65%	8,14%	6,98%	100,00%

Table 5: frequency in percentage of first- and second-digit of Daily COVID-19 confirmed cases in Ho Chi Minh city

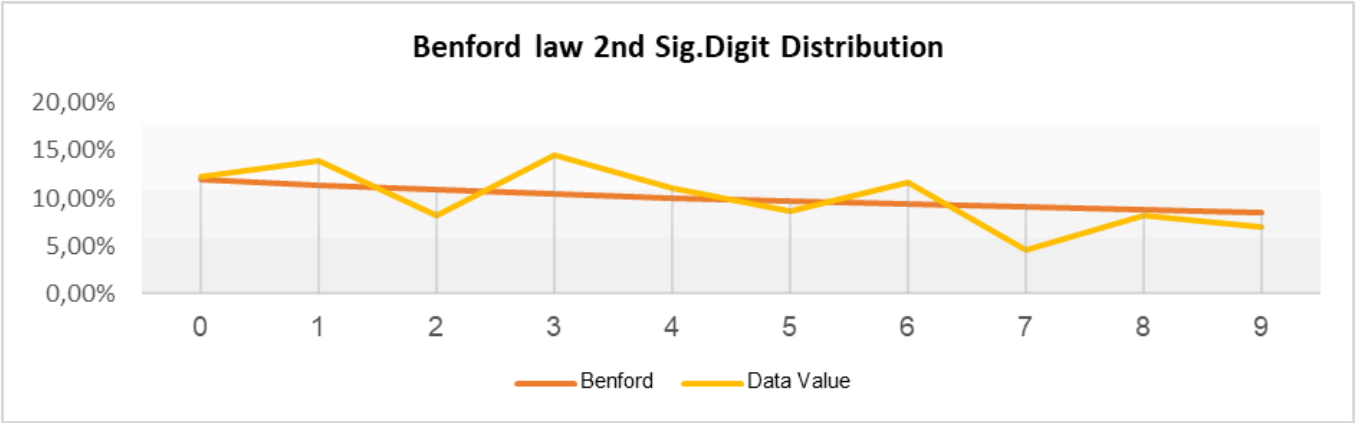
Upon calculating the observed frequencies of the dataset, these frequencies will then be compared to first- and second-digit expected frequencies according to Benford's distribution, depicted in table 4 and graph 1a, 1b.

	First digit			second digit	
Digit	Benford	Data Value	Digit	Benford	Data Value
0			0	11,97%	12,21%
1	30,10%	27,91%	1	11,39%	13,95%
2	17,61%	9,30%	2	10,88%	8,14%
3	12,49%	12,21%	3	10,43%	14,53%
4	9,69%	13,95%	4	10,03%	11,05%
5	7,92%	13,37%	5	9,67%	8,72%
6	6,69%	6,40%	6	9,34%	11,63%
7	5,80%	5,81%	7	9,04%	4,65%
8	5,12%	1,74%	8	8,76%	8,14%
9	4,58%	9,30%	9	8,50%	6,98%

Table 6: Observed and Expected frequencies summary for first- and second-significant digit (expected frequencies are highlighted in yellow)



Graph 1a: Observed and Expected distribution of first significant digit



Graph 1b: Observed and Expected distribution of second significant digit

Goodness-to-fit Chi-square test is performed and table 7 provides a result summary of the test. It is concluded that the daily confirmed case has conformity with Benford's law when performing a second-digit test, which is acceptable since either test can be used to evaluate the authenticity of the dataset. Table 8 indicates that again, second-digit tests prove the conformity of the dataset with Benford's distribution where SSD and Euclidean distance statistical values satisfy the conformity with Benford's value limits of these two measures, which are indicated with yellow highlights. MAD measures in the two tests did not satisfy the limit value, which is less than 0.015.

Goodness-to-fit test		1st dgit	2nd digit
		95% CI	95% CI
Chi-square:	test value	28,934	10,008
	critical value	15,500	16,900
		reject H0	accept H0

Table 7: Chi-square goodness-to-fit test result summary of first- and second-digit tests.

Statistical measures		1st dgit	2nd digit
d-factor (Euclipean distance)	< 0.25 => conformity to Benford dist.	0,120	0,075
MAD	< 0.015 => conformity to Benford dist.	0,032	0,020
SSD	< 100 => conformity to Benford dist.	155,592	60,125

Table 8: Statistical measures result summary of first- and second-digit tests.

4.3. Model Parameters Estimation

Part 1: Construction of ARIMA model

The general process flow chart of modelling ARIMA is given as the following figure:

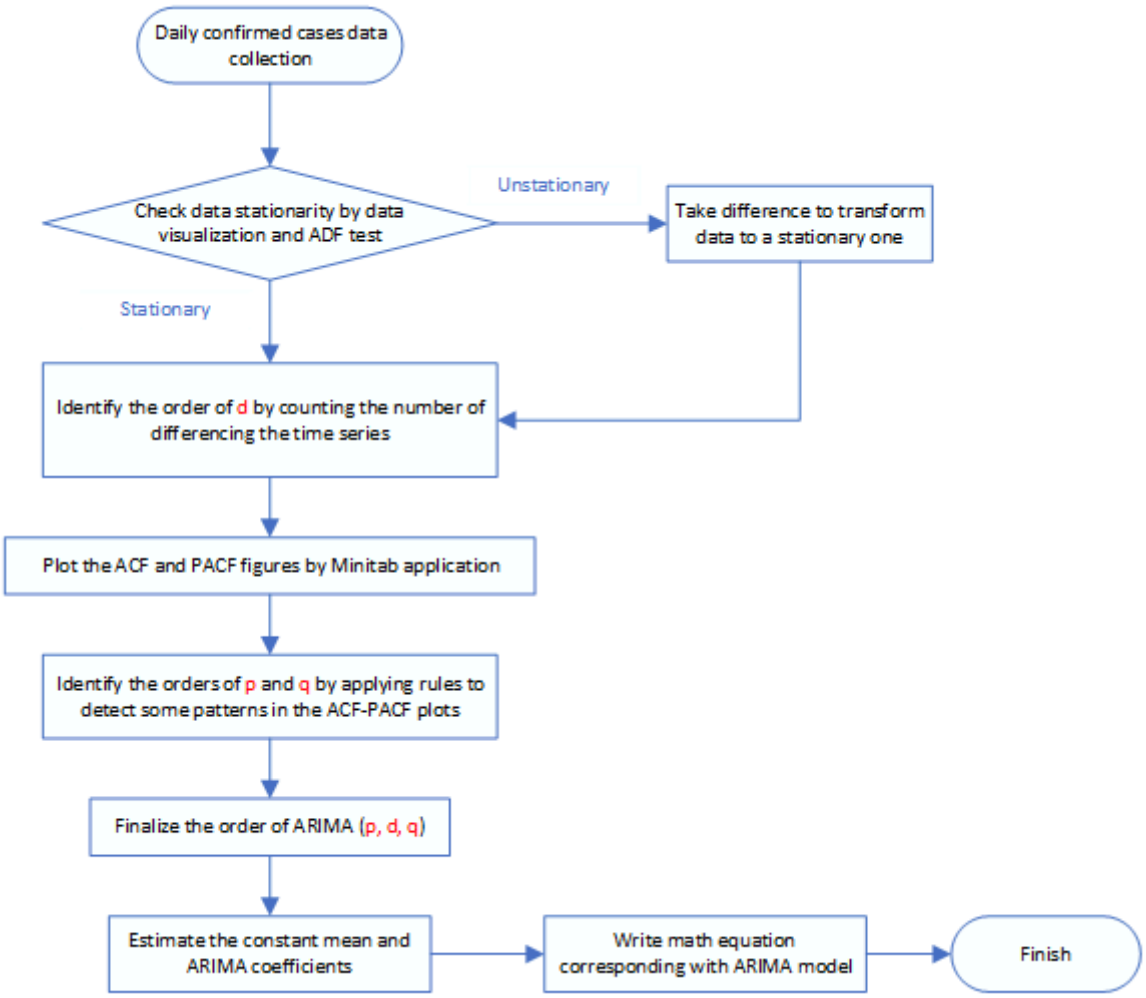
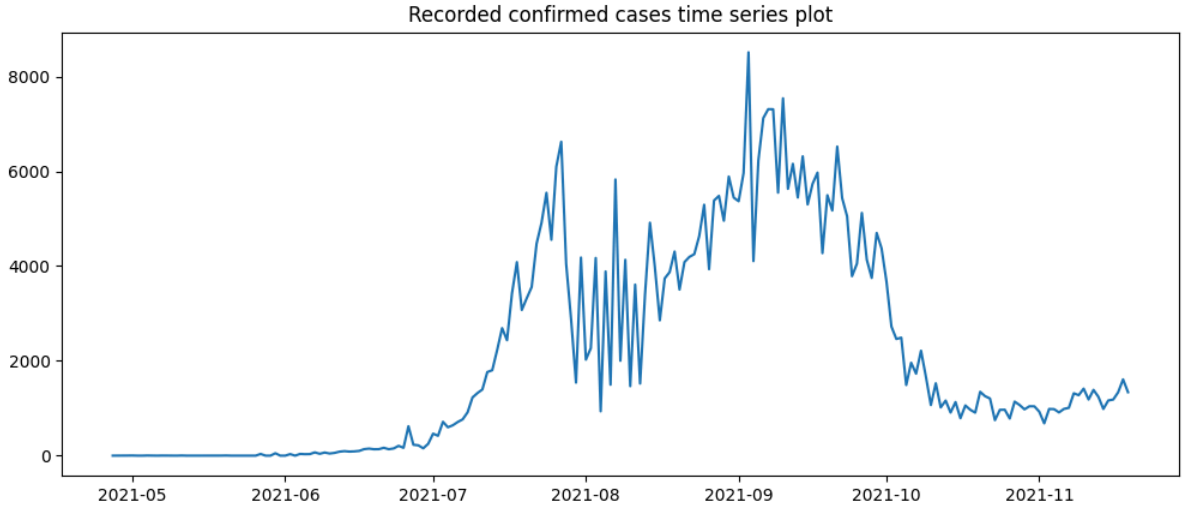


Diagram 4: Flow chart of ARIMA modelling process

Data stationarity checking

A collection of confirmed case data in a daily time frame from 27 April, 2021 was performed and the research team continued to update the data daily into the data tracking file. The figure below represents the time series plot of the confirmed case data from 27 April, 2021 to 24 November, 2021. There were 212 observations in total.



Graph 3: Daily confirmed cases during the fourth wave of COVID-19 pandemic in Ho Chi Minh city

As shown in the figure, the data was changing over time, which indicates non-stationarity in the data. Therefore, this data is not ready for implementing an ARIMA forecasting model. To ensure the judgement was correct, the team performed an ADF test and the result of the test was shown in the picture below. The test statistics were greater than the critical values of three different confidence levels. Hence, the confirmed case data was statistically non-stationary.

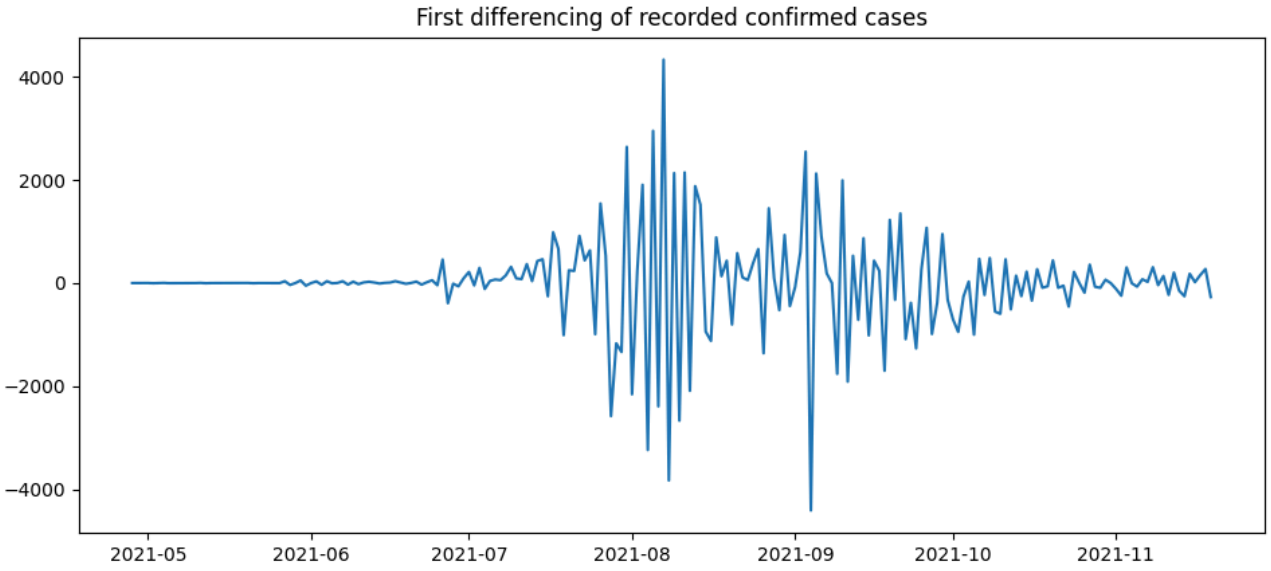
Null Hypothesis: SER01 has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 5 (Automatic - based on t-statistic, lagpval=0.05, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.229351	0.9012
Test critical values:	1% level	-4.004365	
	5% level	-3.432339	
	10% level	-3.139924	

*MacKinnon (1996) one-sided p-values.

Figure 2: ADF test results for stationarity of 0 differencing in confirmed cases

The group then took the first difference to transform the data and did an ADF test again to check the stationarity of the first-difference data. The time series plot and ADF test both indicated the data was now stationary. The results were shown as follows.



Graph 4: Daily confirmed cases distribution in the first differencing

Null Hypothesis: D(SER01) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 4 (Automatic - based on t-statistic, lagpval=0.05, maxlag=14)

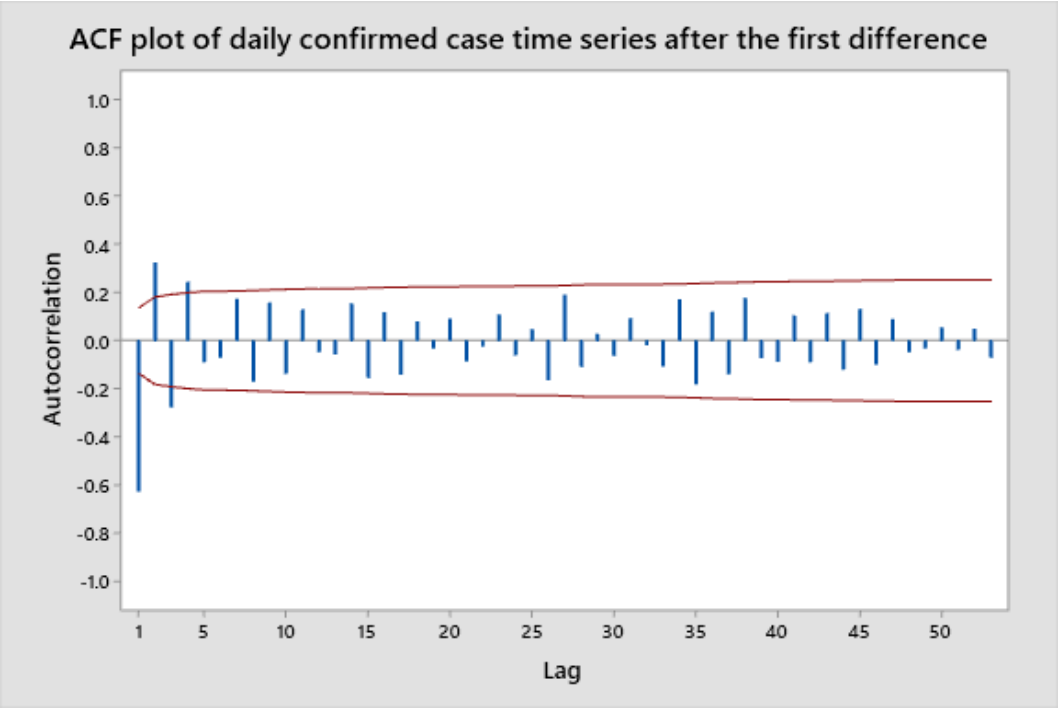
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.664411	0.0000
Test critical values:	1% level	-4.004365	
	5% level	-3.432339	
	10% level	-3.139924	

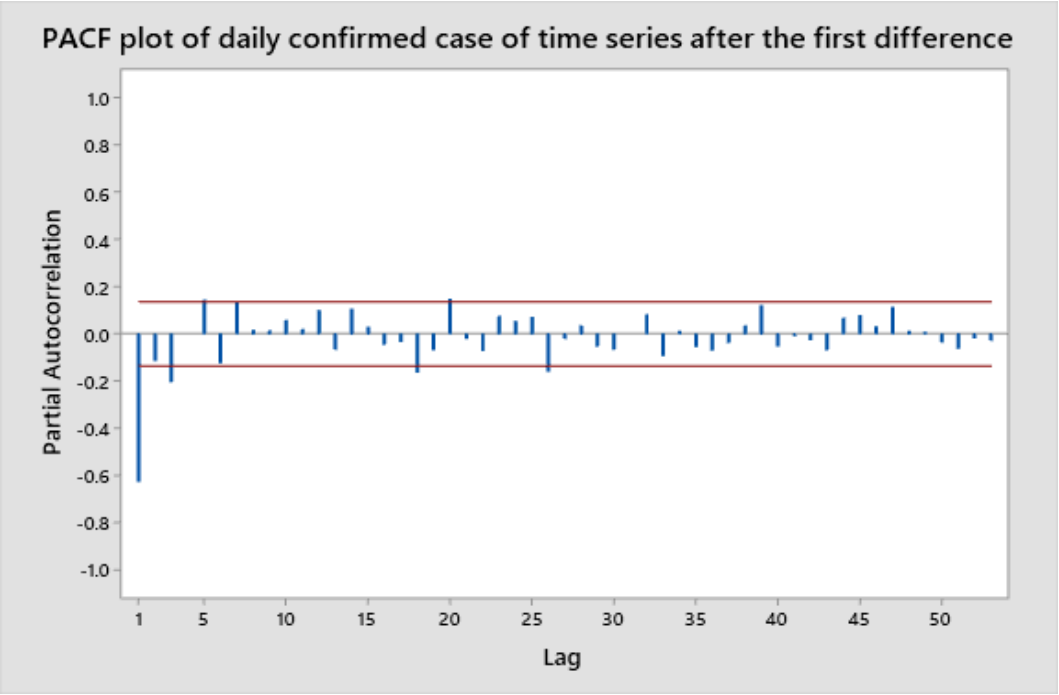
*MacKinnon (1996) one-sided p-values.

Figure 3: ADF test results of confirmed cases stationarity of the first differencing

ARIMA model selection

The order of the ARIMA model was identified based on the ACF and PACF of the stationary time series, which was the first-difference confirmed case data. The ACF and PACF of the data were shown in the following figures.





Graph 5a, 5b: ACF and PACF plot of daily confirmed cases during the first differencing

The research group used rules defined in part V to determine the order of p and d. In this case, the ACF plot died down and the PACF plot dropped off after lag 1. Thus, the ARIMA model would be ARIMA (1, 1, 0).

ARIMA(1,1,0) is the difference first-order autoregressive model. This would yield the following prediction equation:

$\hat{Y}_t - Y_{t-1} = \mu + \phi 1(Y_{t-1} - Y_{t-2})$ and $\hat{Y}_t - Y_{t-1} = \mu$

which can be rearranged to

$$\hat{Y}_t = \mu + Y_{t-1} + \phi 1\,(Y_{t-1} - Y_{t-2}) = 10.4 + Y_{t-1} - 0.628(Y_{t-1} - Y_{t-2})$$

where μ is the constant mean and $\phi 1$ is the coefficient of the autoregressive term.

Part 2: Construction of External factor model:

The general process flow chart of modelling external factor is given as the following figure:

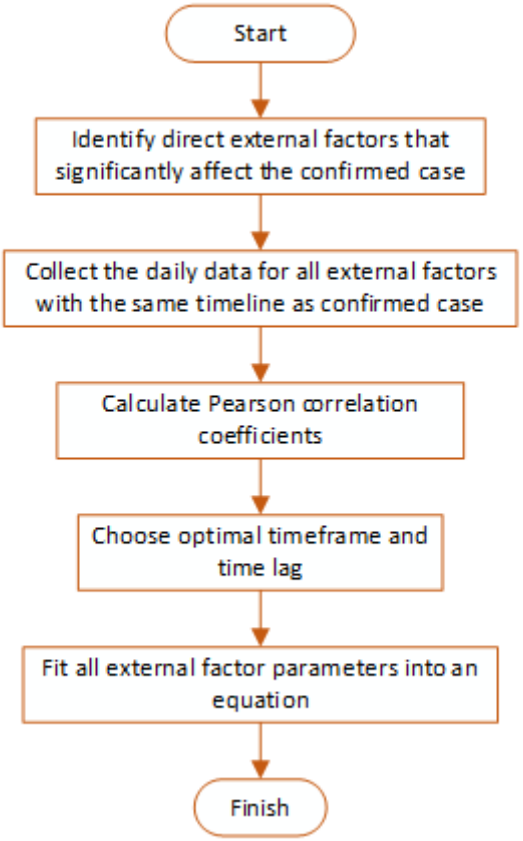


Diagram 5: Flow chart of estimation of external factors and implementation into forecasting model

Regarding external factors that have an impact on the number of the daily new confirmed cases, the team considered three main factors. They were the vaccination factor, the Covid variant factor, and the control measure factor that the Government and Ho Chi Minh city issued with different situations of COVID-19. The group seeked for the direct and measurable characteristics of each factor and eventually came up with four direct external factors to include in our ARIMA model: the daily vaccination rate, the social distancing policy, the incubation time, and the generation time of Covid variant. The external factor layer for our forecasting model was demonstrated in the diagram below.

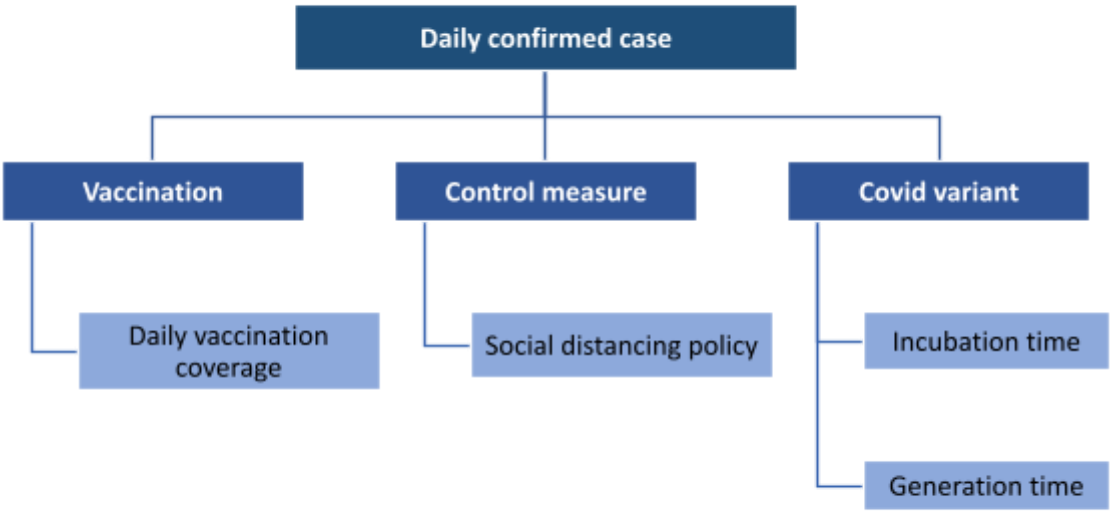


Diagram 6: External factor layer demonstration

The team examined the correlation coefficient of external factors with two time lags. Since the research team only had 212 observations, the allowable time frame is daily only. Regarding the weekly time frame, there were merely 30 observations during the fourth wave of the pandemic, which was not enough for our model study and neither was the monthly time frame. The need of doing correlation research with varied temporal properties stemmed from the fact that time-series data from different applications might behave quite differently. For example, in some applications, a monthly time frame with a lag can give the most appropriate forecast; whereas in other studies, a daily or weekly timeframe with a lag of one or two units is the most appropriate. Thus, we wanted to make sure that our external factors correlation study will reach out to the best forecast accuracy.

The application of Data Analysis ToolPak in Excel was used for Pearson correlation coefficient calculation. The results were shown in the six following tables.

For daily time frame and time lag = 1, the result was defined as the following table:

	Confirmed case	Vaccination coverage	Social distancing	Incubation period	Generation time
Confirmed case	1	0.460	0.824	0.007	0.018
Vaccination coverage		1	0.423	-0.037	-0.047
Social distancing			1	0.115	-0.061
Incubation period				1	-0.037
Generation time					1

For daily time frame and time lag = 2, the result was the following table:

	Confirmed case	Vaccination coverage	Social distancing	Incubation period	Generation time
Confirmed case	1	0.520	0.832	-0.043	0.131
Vaccination coverage		1	0.414	-0.087	-0.041
Social distancing			1	0.054	0.050
Incubation period				1	-0.102
Generation time					1

This table was the summary of Pearson correlation coefficient results with three direct external factors and two daily time lags that was found in the previous section.

Temporal properties			External factor variables				Result
Observation	Time frame	Time lag	Vaccination coverage	Social distancing	Incubation period	Generation time	Average ρ
212	Daily	1	0.460	0.824	0.007	0.018	0.4303
106	Daily	2	0.520	0.832	-0.043	0.131	0.4366

Judging by the result summary table, the best Pearson correlation coefficient was a daily timeframe and a lag of 2. In this case, the calculation considered 106 observations in total, which was large enough to be acceptable to use the forecasting model.

In this case, the Pearson correlation coefficients were 0.52, 0.832, -0.043, 0.131 for daily vaccination coverage (x1), social distancing policy (x2), incubation time (x3), generation time (x4), respectively. Subsequently, the best β parameters for our model are:

$$\beta_1 = \frac{\rho_2}{\rho_1+\rho_2+\rho_3} = \frac{0.52}{0.52+0.832-0.043+0.131} \approx 0.361$$

$$\beta_2 = \frac{\rho_2}{\rho_1+\rho_2+\rho_3} = \frac{0.832}{0.52+0.832-0.043+0.131} \approx 0.577$$

$$\beta_3 = \frac{\rho_2}{\rho_1+\rho_2+\rho_3} = \frac{-0.034}{0.52+0.832-0.043+0.131} \approx -0.03$$

$$\beta_4 = 1 - \beta_1 - \beta_2 - \beta_3 = 1 - 0.361 - 0.577 + 0.03 \approx 0.092$$

With the calculated β, the final forecasting model established in this study was as follows:

$$y(i) = \text{ARIMA}(1, 1, 0)(i) + 0.361\, x_1(i-2) + 0.577\, x_2(i-2) - 0.03\, x_3(i-2) + 0.092\, x_4(i-2)$$

$= 10.4 + Y_{t-1} - 0.628(Y_{t-1} - Y_{t-2}) + 0.361\ x_1\ (i-2) + 0.577\ x_2\ (i-2) - 0.03\ x_3\ (i-2) + 0.092\ x_4\ (i-2)$

Chapter V: FORECASTING RESULT & ANALYSIS

5.1. Data collection

As mentioned in the previous part, the ARIMA model generates the forecast value based on the historical data. Hence, the primary data that needs to be collected is the past daily confirmed cases of Ho Chi Minh city. In this study, the research group was taking into account only the new daily confirmed cases from the beginning of the fourth wave, which is on 27th April, 2021. The research group collected confirmed case data in a daily time frame and recorded it in a spreadsheet.

This study uses the daily vaccination rate at Ho Chi Minh city from April as an external data while the first dose and the second dose. The data uses Pearson correlation to determine its relationship with other external factors and COVID-19 confirmed cases. Another external factor that can be collected directly is the social distancing factors, which is the allowed distance between people when they are in public places depending on the control measures proposed by the city during the outbreak.

5.2. Residual diagnostics for confirmed cases

Before performing the forecast, it is necessary to perform diagnostics on the residuals that have been provided by model ARIMA (1,1,0). This step is used to evaluate whether there are any significant lags provided in the residue. If a significant lag exists, it is advisable to include it into the model. On the other hand, if there is no significant spike – meaning a white noise process appears, in the correlogram visualised, then the model is adequate and ready to implement and forecast the data series of this study.

Specifically, the figure below demonstrates the correlogram of residuals of model ARIMA (1,1,0), the left panel demonstrates the residuals and the right panel is the probability density function of residuals. It is noticeable that there is no sudden spike or significant lag. Hence, ARIMA (1,1,0) can be used to forecast.

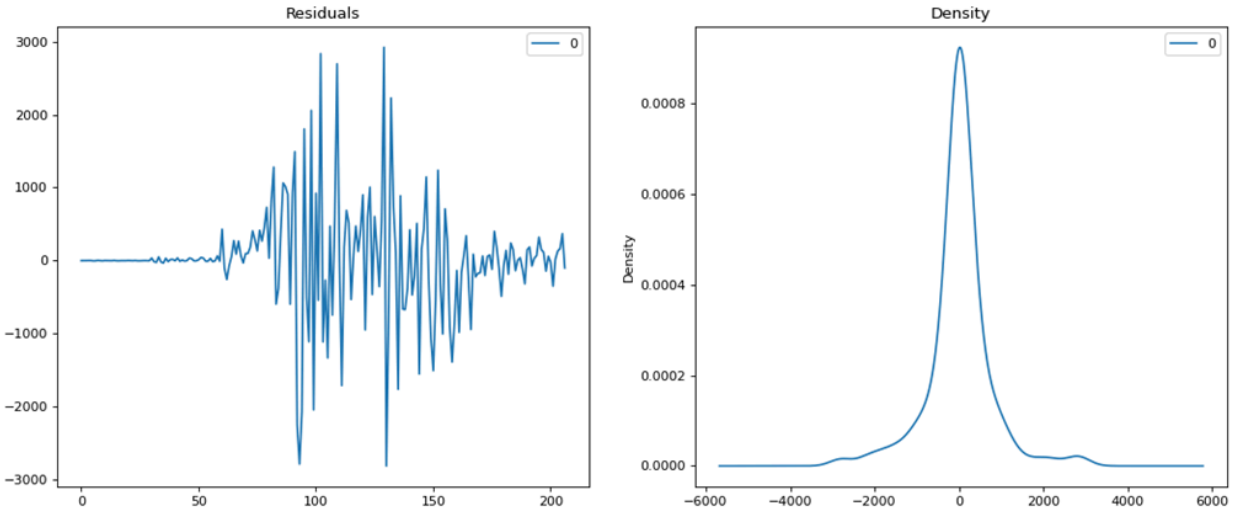
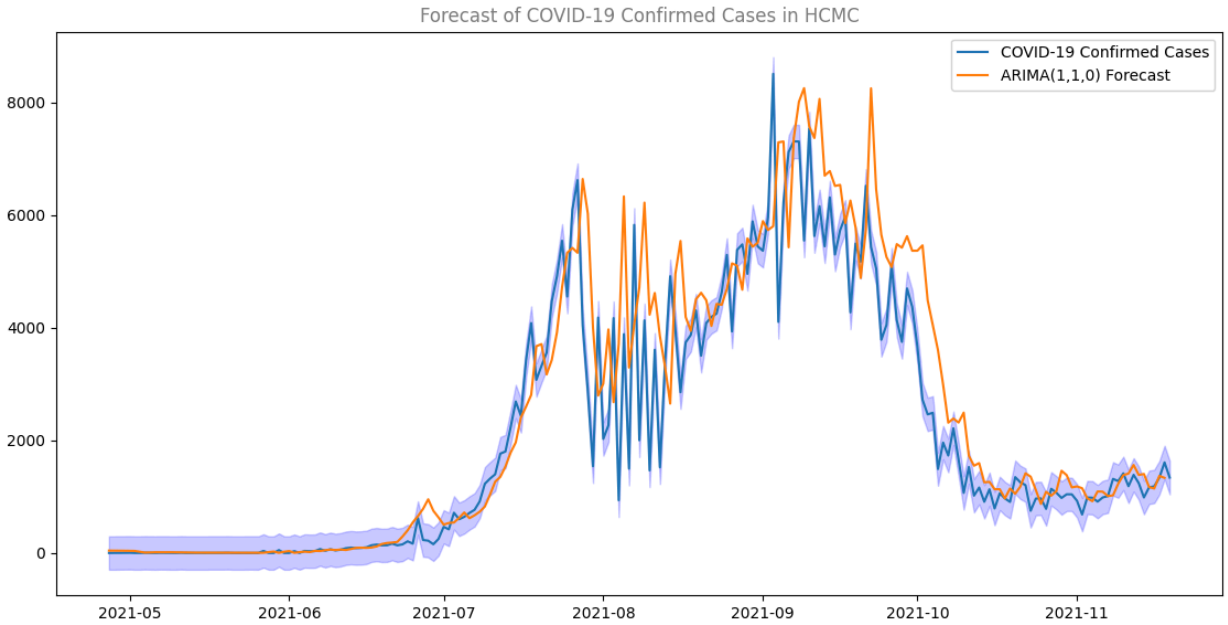


Figure 4: Left panel is residuals of COVID-19 confirmed cases in Ho Chi Minh city. Right panel is the probability density function of residuals.

5.3. Forecasting result and Analysis

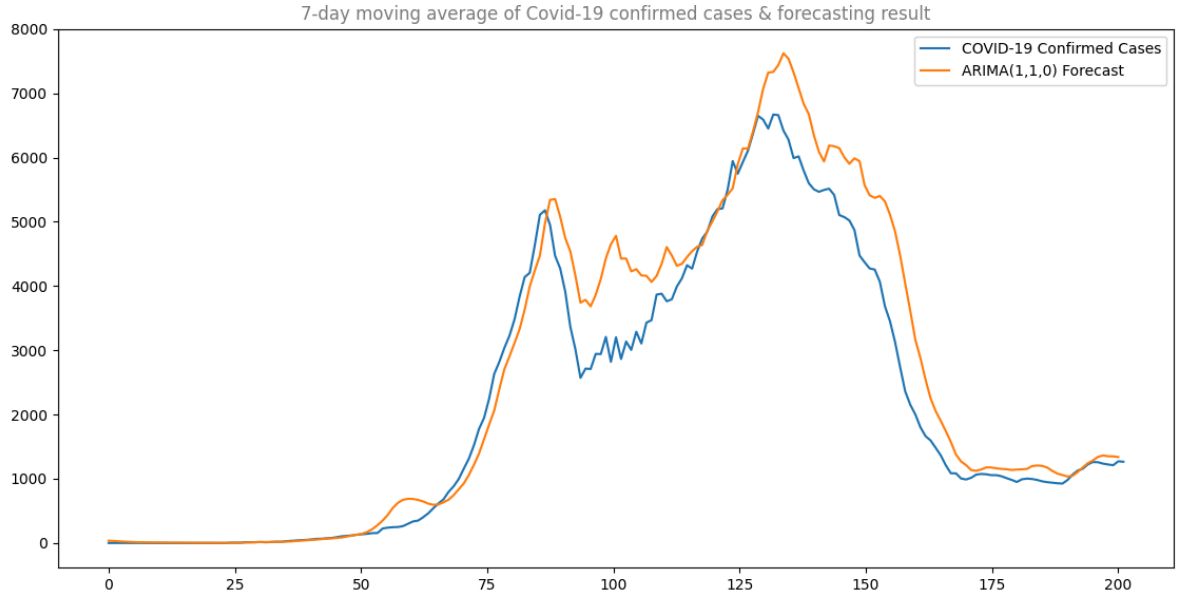
The below figures are the forecasting results of model ARIMA (1,1,0) and implementation of external parameters to forecast the number of confirmed cases in Ho Chi Minh city from 27th April, 2021 to 23th November, 2021



Graph 6: Forecasting result of confirmed cases in Ho Chi Minh city, from 27th April, 2021 to 23rd November, 2021 with 95% confidence level

According to Graph 6, the forecasting results provide a higher value overall, especially during the peaks compared to the collected confirmed cases. This can be seen that control measures used at the time were effective and the model has successfully demonstrated the dynamic of the recent outbreak in HCMC and effectively implemented the external data into the ARIMA model, which increases its accuracy in determining the outbreak peaks. Moreover, the peaks are

estimated accurately, specifically, the forecasting results suggest that the first peak would be from 26th July to 29th July, whereas the actual first peak is on 27th July with 6622 confirmed cases. Additionally, it is forecasted that the second peak would be from 7th September to 12th September, with the highest value during this 6-day period being 8131 confirmed cases. However, the actual second peak appears earlier than predicted, with over 8500 cases on 3rd September. The RMSE parameter in this study poses a high accuracy in the forecasting values compare to the recorded confirmed cases as the error estimation RMSE for the model is considered to be smaller compare to other papers, such as those of [5], where both Prophet and ARIMA were used to forecast confirmed, death and recovered cases. The paper provides a table of RMSE estimation in forecasting confirmed cases using the ARIMA model has the average value at 7015.26 while this study has a RMSE of 986.497 in ARIMA (1,1,0) while successfully introducing the external parameters into the model.



Graph 7: 7-Day moving average of COVID-19 Confirmed cases and model forecasting results

Considering the 7-day moving average between the forecasting result, which is the orange line and the recorded confirmed cases depicted as the blue line in Graph 7, the first peak of the two is nearly at the same time, which is specifically pointed out in Graph 6. Additionally, it is clear that the second peak predicted is much higher and is produced later than the actual data. However, through graph 7, the predicted value of the model is slightly greater than the actual 7-day smoothing confirmed cases. Which can be considered as a good sign for the model. Specifically, predicted results which are greater than the actual value can help organisation plan and distribute resources more effective which can deal with the worst scenario of the outbreak or any other issues depicted as the forecasting values since the purpose of forecasting is for planning and preparing for the upcoming events or situation, in order to successfully tackle with uncertainties and risks. Additionally, it is worth noticing that according to [37], the local official newspaper on the situation that would happen to Ho Chi Minh city after 28th August, 2021, depicted that the

actual infected cases during the fourth wave in Ho Chi Minh city can consider to be four times higher than the recorded cases, therefore, the confirmed cases proposed in this study might be the tip of the iceberg. However, if the actual cases were much higher than the recorded ones, it can also be a good sign of community immunity to the virus before vaccination is accessible to every citizen.

Chapter VI: CONCLUSION AND DISCUSSION

This paper aims to provide a suitable and easy-to-implement forecasting model for Ho Chi Minh city in predicting the incoming spread of pandemic while trying to include the external data into the model so as to have better description and more accurate prediction of the next wave of the pandemic.

Through this study, the forecasting model can be used by governments and organisations to come up with more effective control policies and measures to minimize the negative impact this outbreak or the future outbreak can have on society. In order words, the study can assist organisations in recognizing the pandemic wave's peaks in different scenarios, where different levels of control policies and the characteristics of different covid-variants are applied. Through the model, the dynamic of the upcoming outbreak can be visualized, and therefore, appropriate distribution of resources and measures can be applied, with the main purpose of minimizing the impact each outbreak has and actions to take in order to adapt with the new normal effectively.

Additionally, it is clear that forecasting can propose many uncertainties and is not always applicable in many situations. Hence, it is worth understanding that when should we forecast, at what frequencies should the forecast be considered according to each pandemic situation, and how far away from the situation should we forecast. It is advisable to forecast before the next wave appear from 1-2 months, since at that point, information on the characteristics of the potential covid-variant and its transmission rate, as well as the economic and social situation of the city or country is considered adequate since the economic and social situations in 1 to 2 months from the next outbreak does not pose any significant changes, at that point, before the outbreak, the forecasting frequencies should be in days, which means the forecasting results should visualized predicted values in days, to fully understand the impact a certain covid-variant have on the population. Regarding the new normal situation, the forecasting frequencies should be in weeks, instead of days, since the new normal can be seen as living side by side with the pandemic while striving for community immunity through vaccinations. Hence, it is not necessary to predict by days, since at this point, people who are infected with the virus usually have little or no symptoms at all, and their recovery is faster than those who have not been

vaccinated. Therefore, forecasting by weeks should be applied, in order to monitor the development of the pandemic to take timely protections from future threats.

Effectiveness of forecasting COVID-19 outbreak has on economic and social aspects:

In terms of economic aspects, it is clear that COVID-19 forecasting model can timely inform firms and organisations of the potential outbreak to quickly make suitable plans in order to maintain production and supply chains while changing their approach to business. The main purpose is to not interrupt the logistics and supply chains. Additionally, service industries such as travelling, tourism and transportations will have enough information to adapt and readjust their resources to cope with the next outbreak, minimizing the number of layoffs across organisations and firms. Appropriately distributing the money to strengthen the healthcare system and resources for potential outbreak is one of the considerations successful forecasts of pandemic can have on the economy. Governments and organisations can also readjust their investments on social projects to ensure an effective cash flow while providing a steady foundation to maintain the supply chains and fulfil export targets during the outbreak, in order words, providing a suitable means for safe production and transportation. Moreover, providing adequate encouragement for firm formations and startups during difficult times of the pandemic through accurate and effective pandemic forecasts.

When it comes to social aspects, forecasting for COVID-19 possible outbreak helps school and educational organisations to prepare better plans in case the outbreak occurs and ensure students still receive proper education programs while keeping them safe from the pandemic. Forecasting for the next possible outbreak can also help schools adjust and ensure minimum layoffs among teachers and human resources needed to provide educational programs to students. Healthcare systems can distribute their resources more effectively and provide adequate support to deal with the worst-case scenario when the outbreak appears. Thus, this study can act as a platform for further development of forecasting models in order to minimizing the negative effect a possible pandemic wave has on the quality of life of any social class, especially the vulnerable ones. Forecasting confirmed cases can support the government making better plans to ensure every citizen and population is protected and cared for, since quality of life can have major correlation with the population's economy and environment.

However, it should be noticed that forecasting can be considered ineffective when the cost for forecasting is too high, its uncertainty is too critical to accept. Additionally, forecasting models and techniques do not provide timely results as expected is also considered unhelpful. Governments and organisations should provide a more in-depth study in model selection and combine with new technologies to produce a timely, more convenient and cost-efficient while providing adequately accurate results of the situation.

Upon the research team's work on forecasting and modelling the spread of the forth COVID-19 pandemic wave in Ho Chi Minh city, some limitations must be listed as they may have some impacts during the forecasting process, in addition, these limitations can act as a platform for performing better forecasting results in future research. Specifically, the team was not able to fully express the effectiveness of vaccination, since it takes time for vaccines to be effective against COVID-19 variant and it also depends on the vaccination coverage in a community. Moreover, the model does not automatically update data forecasting. So when people want to produce a new forecast of a different time period, they have to start from assessing the stationarity of the new dataset, find new values for (p,d,q) parameters and produce another forecast with ARIMA which results in time consuming and effort. Daily incubation period and generation time are estimated parameters, their accuracies are uncertain as the group generated the daily values based on the distribution provided by other research. Additionally, the group did not rely on the third party to assess the effectiveness and accuracy of considered models at the beginning (in model selection) when implementing Multi-criteria decision-making methods into the paper. Hence, the group provides linguistic terms to make pairwise comparisons between models, given the information provided in other research where researchers compare two or more models to evaluate their accuracy in forecasting the COVID-19 pandemic. Therefore, this work may result in biassed consideration toward certain models. Last but not least, the team only did data validation on the recorded confirmed cases since the data is the most critical element for forecasting. However, collected data such as daily vaccination should be considered in data validation as well.

It is advisable that the team needs to perform hypothesis tests for all data that were implemented into our model to evaluate the reliability of the data and confirm their confidence interval. In our study, COVID-19 data is dynamic and there is a strong possibility that it is unable to access the exact data of the Ho Chi Minh city epidemic situation. The ineffective epidemiological surveillance systems and insufficient medical strength are two possible reasons for the COVID-19 data shortage. In most cases, all required data for developing the ARIMA time series forecasting model are not fully available and the team had to collect data from different sources, which leads to the data variation. Since the ARIMA model is a historical data based model, unless it was able to judge the reliability of data, the student team might make wrong conclusions about the effectiveness of our model.

Furthermore, in this study, the research team integrated the external factor into our forecasting model, which is believed to upturn the forecast accuracy. It is critical that the group can spot the direct factor that gives a significant contribution to the behaviour of the main forecasting variable and add it into the ARIMA model. Another noteworthy point when implementing external factors into a forecasting model is that all the external factors must be independent with

each other. In other words, there should be no strong correlation between external factors. Hence, the process of scanning and selecting external factors is worth careful consideration. Any invalid and redundant external factor adding can lead the model to the inappropriate direction and expand the gap between the forecasting value with the actual ones.

Chapter VII: APPENDIX AND REFERENCE

REFERENCE:

1. Harun, Y., Aynur, Y., Mustafa, A. T., Melike, T. (2020). *Modelling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods.*
2. Saleh, I., Alzahrani, I., Aljamaan, E., Fakih, A. (2020). *Forecasting the Spread of the COVID-19 Pandemic in Saudi Arabia Using ARIMA Prediction Model Under Current Public Health Interventions.*
3. Andi, S., Yudhistira, N., Juan, K., Alex, L. (2020). *Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta.*
4. Essi, I., Nwuju, K., Etuk, E. (2021). *ARIMA Modelling and Forecasting of COVID-19 Daily Confirmed/Death Cases: A Case Study of Nigeria.*
5. Naresh, K., Seba, S. (2020). *COVID-19 Pandemic Prediction using Time Series Forecasting Models.*
6. Christopher, Bennett., Rodney, S., Junwei, Lu. (2014). *Autoregressive with Exogenous Variables and Neural Network Short-Term Load Forecast Models for Residential Low Voltage Distribution Networks.*
7. Marco, T., & Umberto, T. (2021). *Forecasting the number of confirmed new cases of COVID-19 in Italy for the period from 19 May to 2 June 2020.*
8. William W.S. Wei (2006). *Time series analysis – univariate and multivariate methods, second edition.*
9. Peter J. Brockwell Richard A. Davis (2002). *Introduction to Time Series and Forecasting, Second Edition.*
10. George, E., Box, G., JENKINS, G., Reinsel, G. (2009). *Time series analysis forecasting and control fifth edition.*
11. Meng, Z., Jianpeng, X., Aiping, D. (2021). *Transmission dynamics of an outbreak of the COVID-19 Delta variant, Guangdong Province, China.*
12. Pirooz, M., M, J. (2019). *An Algorithm for Generating Random Numbers with Normal Distribution.*

13. Sang, G. Kwak & Jong Hae Kim. (2015). *Central limit theorem: the cornerstone of modern statistics.*
14. Mostafa, A., & Tatiana, M. (2021). *System for forecasting COVID-19 cases using time series and neural networks models.*
15. Behrouz, F., & Rahim, A. (2021). *A novel Pseudo random number generator for cryptographic applications.*
16. Terry Banks (2004). *Discrete event system simulation, fourth edition.*
17. Mashael, K., Kaouther, L., Nada, A., & Maysoon, A., (2019). *Time series Facebook prophet model and Python for COVID-19 outbreak prediction.*
18. Roseline, O., Ogundokun, A., Lukman , Golam, B., Joseph, B., Benedita, A. (2020). *Predictive modelling of COVID-19 confirmed cases in Nigeria.*
19. Jose Maria Vera Barberan (2020). *Adding external factors in time series forecasting case study: Ethereum price forecasting.*
20. Cort J, W., Kenji, M. (2005). *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.*
21. Niko A, K., Carey, L., Vikas, P., Matthew, L., (2021). *A Bayesian mixture model for predicting the COVID-19 related mortality in the United States.*
22. Robbeca, G., Tiffani, C., & Laura, S. (2021). *Impact of Sars-Cov-2 Delta variant on incubation, transmission settings and vaccine effectiveness: Results from a nationwide case-control study in France.*
23. Wang, X., Yang, L., Zhang, H., Yang, Z., & Liu, C. (2020). *Forecasting confirmed cases of the COVID-19 pandemic with a migration-based epidemiological model.*
24. Ajay, K., Kamaldeep, K. (2021). *A hybrid SOM-fuzzy time series (SOMFTS) technique for future forecasting of COVID-19 cases and MCDM based evaluation of COVID-19 forecasting models.*
25. Shahriare, S., Koushik, C, H., Mufti, M., Shamim, K., Sheikh, S, I., Julian, M, W., Salem, A., & Mohammad, A, M. (2021). *Short-term prediction of COVID-19 cases using machine learning models.*
26. Patricia, Melin., Julio, C, Monica., Daniela, S., Oscar, C. (2021). *Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: The case of Mexico.*
27. Noah, F., H, Lahooti. (2021). *Are COVID-19 data reliable? A quantitative analysis of pandemic data from 182 countries.*
28. Association of Certified Fraud Examiners. (2020). *Using Benford's Law to detect fraud.*
29. Anran, W., Andre, E. (2020). *Is COVID-19 data reliable? A statistical analysis with Benford's Law.*
30. Noah, F., Hooshang, L. (2021). *Are COVID-19 Data Reliable? A Quantitative Analysis of Pandemic Data from 182 Countries.*

31. Noah, F. (2021). *Can we rely on COVID-19 data? An assessment of data from over 200 countries worldwide.*
32. Adrian, P, K., Sheung, C, P, Y. (2020). *On the authenticity of COVID-19 case figures.*
33. Alex, E, K. (2021). *On the Mistaken Use of the Chi-Square Test in Benford's Law.*
34. Saville, A, D. (2006). *Using Benford's Law to detect data error and fraud: an examination of companies listed on Johannesburg stock exchange.*
35. Jerry, B., John, S, C, II., Barry, L, N., David, M, N. (2004). *Discrete-event system simulation. Chapter 9: Input model. Page: 287. 4th edition.*
36. Yui, C. L, Tim, K.T. (2021). Joint Estimation of Generation Time and Incubation Period for Coronavirus Disease 2019. *The Journal of Infectious Diseases*. *224*(10). 1664-1671. https://doi.org/10.1093/infdis/jiab424
37. Nguyen, T.N. (2021, August 28). Khả năng diễn biến dịch COVID-19 ở Thành phố Hồ Chí Minh sau ngày 25-8-2021. *People's Army Newspaper. Retrieved from:*

    https://www.qdnd.vn/tren-tuyen-dau-chong-dich/cac-van-de/kha-nang-dien-bien-dich-COVID-19-o-thanh-pho-ho-chi-minh-sau-ngay-25-8-2021-669335

38. Khan, F., Ali, S., Saeed, A., Kumar, R., Khan, A, W. (2021). *Forecasting daily new infections, deaths and recovery cases due to COVID-19 in Pakistan by using Bayesian Dynamic Linear Models*.
39. Kumar, N., Susan, S. (2020). *COVID-19 pandemic prediction using Time Series Forecasting Models*.
40. Sulasikin, A., Nugraha, Y., Kanggrawan J., Suherman, A., L. (2020). *Forecasting for a data-driven using time series policy methods in handling COVID-19 pandemic in Jakarta*.
41. Didi, E., I., Kingdom, N., Harrison, E., E. (2021). *ARIMA modelling and forecasting of COVID-19 daily confirmed/death cases: A cases study of Nigeria*.
42. Alastair,. H. (2012). *Testing for a unit root in time series with pretest data-based model selection.*
43. Da., Y., C. (2010). *Applications of the extent analysis method on fuzzy AHP.*
44. Taylor and Francis (2019). *Measuring healthcare service quality from patients's perspective: Using AHP application.*
45. T. Chai and R. R. Draxler. (2014). *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature.*
46. P. J. Brockwell and R. A. Davis. (1991). *Introduction to time series and forecasting (2nd edition).*

APPENDICES

Appendix A. Data collection

The data was collected daily from two main sources which are Ho Chi Minh City Department of Health and Ministry of Health from the day of April 27th to November 23rd 2021. Besides, other sources such as VNExpress, Tuoi Tre, Nhan Dan, Suc khoe doi song, etc. e-newspapers also were references for the whole team to consider the accuracy of updating data in each specific time.

1. Daily Confirmed Cases in Ho Chi Minh city from 27th April, 2021 to 24th November, 2021

	A	B	C	D	E	F
1	DATA COLLECTION					
2	ĐỢT 4 (27/04/2021 - hiện tại)					
3						
4						
5	Ngày	HCMC Department of Health and Ministry of Health	Other outsources (theo note)	Current Total cases	New cases (Updating)	Note
6	27-Apr-21	0	0	0		https://suckhoedoisong.vn/chieu-27-4-them-5-ca-mac-covid-19-trong-do-1-truong-hop-lay-nhiem-tai-noi-cach-ly-n190799.html
7	28-Apr-21	0	0	0		https://suckhoedoisong.vn/chieu-28-4-viet-nam-them-8-ca-mac-covid-19-the-gioi-ghi-nhan-hon-1485-trieu-ca-n190859.html
8	29-Apr-21	1	1	1		https://nhandan.vn/tin-tuc-y-te/tp-ho-chi-minh-ghi-nhan-mot-truong-hop-nghi-nhiem-covid-19-643968/
9	30-Apr-21	2	2	3		https://suckhoedoisong.vn/chieu-30-4-them-14-ca-mac-covid-19-co-4-ca-ghi-nhan-trong-nuoc-tai-ha-nam-va-ha-noi-n191086.html
10	1-May-21	3	3	6		https://suckhoedoisong.vn/chieu-1-5-bo-y-te-cong-bo-14-ca-mac-covid-19-co-3-ca-trong-nuoc-o-ha-nam-n191165.html
11	2-May-21	0	0	6		https://suckhoedoisong.vn/chieu-2-5-them-20-ca-mac-covid-19-co-8-ca-ghi-nhan-trong-nuoc-tai-ha-nam-vinh-phuc-n191232.html
12	3-May-21	0	0	6		https://ncov.moh.gov.vn/vi/web/guest/-/6847426-2754
13	4-May-21	3	3	9		https://tuoitre.vn/so-gd-dt-tp-hcm-dung-moi-hoat-dong-giao-duc-ngoai-lop-den-het-nam-hoc-20210504183405698.htm https://ncov.moh.gov.vn/vi/web/guest/-/6847426-2831
14	5-May-21	1	1	10		https://baochinhphu.vn/Suc-khoe/NONG-Ghi-nhan-them-18-ca-lay-nhiem-trong-cong-dong/430252.vgp
15	6-May-21	0	0	10		https://giadinh.net.vn/y-te/cong-bo-60-ca-mac-moi-covid-19-rieng-bv-benh-nhiet-doi-trung-uong-co-16-ca-20210506185457643.htm
16	7-May-21	2	2	12		https://suckhoedoisong.vn/chieu-7-5-them-40-ca-mac-covid-19-trong-nuoc-rieng-benh-vien-k-la-11-ca-n191782.html
17	8-May-21	1	1	13		https://baochinhphu.vn/Suc-khoe/NONG-Ghi-nhan-them-65-ca-mac-COVID19-trong-cong-dong/430650.vgp https://giadinh.net.vn/y-te/tphcm-cac-benh-vien-no-luc-that-chat-cac-bien-phap-kiem-soat-lay-nhiem-covid-19-2021050814271284.htm
18	9-May-21	0	0	13		https://tuoitre.vn/toi-9-5-ky-luc-77-ca-covid-19-cong-dong-tai-9-tinh-thanh-20210509191324665.htm

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
19	**10-May-21**	0	*0*	13		
20	**11-May-21**	3		16		https://suckhoedoisong.vn/chieu-11-5-them-30-ca-mac-covid-19-trong-nuoc-co-27-ca-n192195.html
21	**12-May-21**	0	*0*	16		
22	**13-May-21**	0	*0*	16		
23	**14-May-21**	0	0	16		https://suckhoedoisong.vn/toi-14-5-them-59-ca-mac-covid-19-ghi-nhan-trong-nuoc-rieng-bac-ninh-33-ca-n192498.htm
24	**15-May-21**	0	0	16		https://suckhoedoisong.vn/toi-15-5-them-129-ca-mac-covid-19-trong-nuoc-rieng-bac-giang-85-ca-n192558.html
25	**16-May-21**	0	0	16		https://suckhoedoisong.vn/toi-16-5-co-54-mac-covid-19-trong-nuoc-rieng-bac-ninh-24-ca-n192608.html
26	**17-May-21**	2	2	18		https://suckhoedoisong.vn/toi-17-5-them-116-ca-mac-covid-19-trong-nuoc-rieng-bac-giang-va-bac-ninh-la-99-ca-n192728.html
27	**18-May-21**	1	1	19		https://giadinh.net.vn/y-te/ban-tin-covid-19-toi-18-5-them-48-ca-mac-moi-ca-nuoc-co-hon-4-trieu-luot-nguoi-duoc-xet-nghiem-20210518184356163.htm
28	**19-May-21**	1	1	20		https://ncov.moh.gov.vn/vi/web/guest/-/6847426-3710
29	**20-May-21**	3	3	23		https://suckhoedoisong.vn/toi-20-5-them-40-ca-mac-covid-19-trong-nuoc-viet-nam-co-4809-benh-nhan-n193005.html
30	**21-May-21**	0	0	23		https://suckhoedoisong.vn/chieu-21-5-them-57-ca-mac-covid-19-trong-nuoc--n193106.html
31	**22-May-21**	0	0	23		https://plo.vn/suc-khoe/chua-day-1-thang-viet-nam-vuot-2000-ca-nhiem-987108.html
32	**23-May-21**	0	0	23		https://baochinhphu.vn/Suc-khoe/Toi-235-Them-76-ca-mac-COVID19-trong-nuoc/432187.vgpp
33	**24-May-21**	1	1	24		https://suckhoedoisong.vn/toi-24-5-them-95-ca-mac-covid-19-trong-nuoc-bac-giang-va-bac-ninh-chiem-77-ca-n193322.html
34	**25-May-21**	1	1	25		https://suckhoedoisong.vn/toi-25-5-them-284-ca-mac-covid-19-trong-nuoc-rieng-bac-giang-243-ca-n193428.html

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
35	**26-May-21**	0	0	25		https://suckhoedoisong.vn/toi-26-5-them-115-ca-mac-covid-19-rieng-bac-giang-va-bac-ninh-103-ca-n193509.html
36	**27-May-21**	36	36	61		https://suckhoedoisong.vn/toi-27-5-co-150-ca-mac-covid-19-trong-nuoc-rieng-tp-hcm-36-ca-n193619.html
37	**28-May-21**	0	0	61		https://suckhoedoisong.vn/chuoi-lay-nhiem-hoi-thanh-lien-quan-bien-the-an-do-n193726.html https://suckhoedoisong.vn/toi-28-5-them-173-ca-mac-covid-19-trong-nuoc-bac-giang-co-123-ca-n193715.html
38	**29-May-21**	0	0	61		https://suckhoedoisong.vn/toi-28-5-them-173-ca-mac-covid-19-trong-nuoc-bac-giang-co-123-ca-n193715.html
39	**30-May-21**	51	49	112		https://suckhoedoisong.vn/toi-30-5-co-142-ca-mac-covid-19-trong-nuoc-rieng-tphcm-ghi-nhan-nhieu-nhat-voi-49-ca-n193841.html https://suckhoedoisong.vn/d-n193817.html
40	**31-May-21**	0	*0*	112		https://vietnamnet.vn/vn/suc-khoe/tin-tuc-covid-19-chieu-hom-nay-31-5-viet-nam-cong-bo-82-ca-covid-19-trong-nuoc-them-1-tinh-co-ca-nhiem-dau-tien-741665.html
41	**1-Jun-21**	0	*0*	112		https://giadinh.net.vn/y-te/toi-1-6-ha-noi-va-9-tinh-thanh-them-89-ca-mac-moi-covid-19-20210601182838848.htm https://ncov.moh.gov.vn/vi/web/guest/-/6847912-220
42	**2-Jun-21**	33	31	145		https://giadinh.net.vn/y-te/ban-tin-covid-19-toi-2-6-ca-nuoc-them-138-ca-mac-moi-tphcm-tang-31-benh-nhan-20210602183746933.htm
43	**3-Jun-21**	0	*0*	145		
44	**4-Jun-21**	36	26	181		https://suckhoedoisong.vn/toi-4-6-them-87-ca-mac-covid-19-trong-nuoc-ky-luc-157-truong-hop-khoi-benh-n194285.html
45	**5-Jun-21**	31	35	212		https://baochinhphu.vn/Suc-khoe/Them-80-ca-mac-COVID19-trong-nuoc-68-benh-nhan-khoi-benh/433773.vgp
46	**6-Jun-21**	33	*21*	245		https://hcdc.vn/category/van-de-suc-khoe/covid19/ban-tin-hang-ngay/pages-38
47	**7-Jun-21**	69	*46*	314		https://hcdc.vn/category/van-de-suc-khoe/covid19/ban-tin-hang-ngay/pages-38

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4						
5	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
48	**8-Jun-21**	39	39	353		https://suckhoedoisong.vn/bo-y-te-ngay-8-6-ghi-nhan-tong-175-ca-mac-covid-19-40-benh-nhan-khoi-n194587.html https://tuoitre.vn/nguoi-nhap-canh-vao-viet-nam-se-phan-loai-theo-nhom-20210608193622182.htm
49	**9-Jun-21**	66	66	419		https://suckhoedoisong.vn/ngay-9-6-co-407-ca-mac-covid-19-va-87-benh-nhan-khoi-n194676.html
50	**10-Jun-21**	45	*61*	464		
51	**11-Jun-21**	58	30	522		https://suckhoedoisong.vn/ngay-11-6-viet-nam-ghi-nhan-tong-196-ca-mac-covid-19-co-96-benh-nhan-khoi-n194832.html
52	**12-Jun-21**	84	*84*	606		
53	**13-Jun-21**	95	95	701		https://suckhoedoisong.vn/bo-y-te-ngay-13-6-ca-nuoc-them-297-ca-mac-covid-19-rieng-tphcm-95-benh-nhan-n194953.html
54	**14-Jun-21**	86	82	787		https://baochinhphu.vn/Xa-hoi/KY-LUC-Ngay-146-co-238-benh-nhan-COVID-duoc-chua-khoi/434748.vgp
55	**15-Jun-21**	90	90	877		https://baochinhphu.vn/Suc-khoe/Toi-156-them-213-ca-mac-COVID19-303-ca-khoi-benh/434849.vgp
56	**16-Jun-21**	100	*99*	977		
57	**17-Jun-21**	137	137	1114		https://giadinh.net.vn/y-te/ban-tin-covid-19-toi-17-6-them-136-ca-mac-moi-ca-ngay-viet-nam-ghi-nhan-515-benh-nhan-2021061718345643.htm
58	**18-Jun-21**	149	*149*	1263		
59	**19-Jun-21**	135	135	1398		https://ncov.moh.gov.vn/vi/web/guest/-/6847426-4841 https://ncov.moh.gov.vn/vi/web/guest/-/6847426-4841
60	**20-Jun-21**	137	*137*	1535		
61	**21-Jun-21**	166	*166*	1701		
62	**22-Jun-21**	136	136	1837		https://suckhoedoisong.vn/ngay-22-6-viet-nam-co-248-ca-mac-covid-19-va-93-benh-nhan-khoi-n195627.html
63	**23-Jun-21**	152	152	1989		https://suckhoedoisong.vn/toi-23-6-them-85-ca-mac-covid-19-tphcm-co-den-61-ca-n195696.html

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4						
5	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
64	**24-Jun-21**	207	162	2196		https://suckhoedoisong.vn/ngay-24-6-co-285-ca-mac-covid-19-rieng-tphcm-162-truong-hop-n195778.html
65	**25-Jun-21**	165	165	2361		https://suckhoedoisong.vn/ngay-25-6-ca-nuoc-co-305-ca-mac-covid-19-rieng-tp-ho-chi-minh-la-161-ca-n195858.html
66	**26-Jun-21**	621	621	2982		https://suckhoedoisong.vn/toi-26-6-them-123-ca-mac-covid-19-tp-ho-chi-minh-nhieu-nhat-58-ca-n195913.html
67	**27-Jun-21**	230	230	3212		https://suckhoedoisong.vn/toi-27-6-them-197-ca-mac-covid-19-tp-ho-chi-minh-co-den-95-ca-n195981.html
68	**28-Jun-21**	218	218	3430		https://suckhoedoisong.vn/ngay-28-6-viet-nam-ghi-nhan-tong-cong-391-ca-mac-covid-19-rieng-tp-ho-chi-minh-la-218-ca-n196050.htmll
69	**29-Jun-21**	155	155	3585		https://ncov.moh.gov.vn/vi/web/guest/-/6847426-5175
70	**30-Jun-21**	249	249	3834		
71	**1-Jul-21**	464	464	4298		https://baochinhphu.vn/Suc-khoe/Ngay-17-Ca-nuoc-them-713-ca-mac-COVID19-407-nguoi-khoi-benh/436648.vgp
72	**2-Jul-21**	419	419	4717		
73	**3-Jul-21**	714	714	5431		
74	**4-Jul-21**	599	599	6030		https://suckhoedoisong.vn/toi-4-7-them-360-ca-mac-moi-nang-tong-so-benh-nhan-covid-19-trong-ngay-len-887-n196495.html
75	**5-Jul-21**	641	641	6671		
76	**6-Jul-21**	710	710	7381		https://suckhoedoisong.vn/toi-6-7-them-504-ca-mac-nang-tong-so-ca-covid-19-trong-ngay-o-nuoc-ta-len-1029-n196647.html
77	**7-Jul-21**	766	766	8147		https://suckhoedoisong.vn/toi-7-7-them-330-ca-mac-covid-19-nang-tong-so-ca-trong-ngay-len-1007-n196727.html
78	**8-Jul-21**	915	915	9062		https://suckhoedoisong.vn/toi-8-7-them-645-ca-mac-covid-19-nang-tong-so-mac-trong-ngay-vuot-1300-n196810.html
79	**9-Jul-21**	1229	1229	10291		https://suckhoedoisong.vn/toi-9-7-them-591-ca-mac-covid-19-nang-tong-so-mac-trong-ngay-vuot-1600-n196878.html
						https://suckhoedoisong.vn/toi-10-7-them-463-ca-mac-covid-19-tong-so-mac-trong-

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4						
5	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
79	**9-Jul-21**	1229	1229	10291		https://suckhoedoisong.vn/toi-9-7-them-591-ca-mac-covid-19-nang-tong-so-mac-trong-ngay-vuot-1600-n196878.html
80	**10-Jul-21**	1320	1320	11611		https://suckhoedoisong.vn/toi-10-7-them-463-ca-mac-covid-19-tong-so-mac-trong-ngay-vuot-1800-ca-n196947.html https://vnexpress.net/du-bao-ca-mac-covid-19-tp-hcm-tiep-tuc-tang-4307433.html
81	**11-Jul-21**	1397	1397	13008		
82	**12-Jul-21**	1764	1764	14772		https://ncov.moh.gov.vn/vi/web/guest/-/6847426-5637
83	**13-Jul-21**	1802	1797	16574		
84	**14-Jul-21**	2229	2229	18803		
85	**15-Jul-21**	2691	2002	21494		
86	**16-Jul-21**	2436	2420	23930		https://suckhoedoisong.vn/dong-nuoc-mat-hanh-phuc-cua-benh-nhan-covid-19-n197481.html
87	**17-Jul-21**	3420	2786	27350		https://ttbc-hcm.gov.vn/benh-vien-dam-bao-chua-benh-nang-kham-benh-tu-xa-16904.html https://tuoitre.vn/cong-bo-luong-xanh-quoc-gia-cho-xe-di-qua-vung-dich-20210717182836294.htm
88	**18-Jul-21**	4083	4066	31433		
89	**19-Jul-21**	3074	3074	34507		
90	**20-Jul-21**	3322	3322	37829		
91	**21-Jul-21**	3558	3556	41387		
92	**22-Jul-21**	4473	3913	45860		
93	**23-Jul-21**	4913	5087	50773		
94	**24-Jul-21**	5546	5396	56319		
95	**25-Jul-21**	4555	4555	60874		
96	**26-Jul-21**	6097	*6097*	66971		https://giadinh.net.vn/y-te/ngan-hang-mau-can-kiet-vi-covid-19-20210726104141711.htm

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4						
5	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
97	**27-Jul-21**	6622	*6622*	73593		https://suckhoedoisong.vn/hon-1000-benh-nhan-covid-19-o-benh-vien-da-chien-so-3-duoc-xuat-vien-n198372.html https://baochinhphu.vn/Chi-dao-quyet-dinh-cua-Chinh-phu-Thu-tuong-Chinh-phu/Tao-dieu-kien-de-cac-dia-phuong-DN-tiep-can-nguon-vaccine-the-gioi/440185.vgp https://ncov.moh.gov.vn/web/guest/-/6847426-6291
98	**28-Jul-21**	4045	*4045*	77638		
99	**29-Jul-21**	2877	2877	80515		https://suckhoedoisong.vn/toi-29-7-them-4773-ca-mac-covid-19-co-4323-benh-nhan-khoi-benh-n198529.html
100	**30-Jul-21**	1541	1541	82056		https://suckhoedoisong.vn/toi-30-7-them-3657-ca-mac-covid-19-co-3704-benh-nhan-khoi-n198605.html
101	**31-Jul-21**	4180	4180	86236		https://suckhoedoisong.vn/toi-31-7-them-4564-ca-mac-covid-19-nang-tong-so-mac-trong-ngay-len-8624-ca-n198655.html
102	**1-Aug-21**	2025	2025	88261		https://suckhoedoisong.vn/toi-1-8-them-4246-ca-mac-covid-19-ca-ngay-8620-ca-16921080118450266.htm
103	**2-Aug-21**	2267	2267	90528		https://suckhoedoisong.vn/toi-2-8-them-4254-nguoi-mac-covid-19-nang-so-ca-mac-trong-ngay-len-7455-169210802184628943.htm
104	**3-Aug-21**	4171	4171	94699		https://suckhoedoisong.vn/toi-3-8-them-4851-ca-mac-covid-19-ca-ngay-ha-noi-them-gan-100-benh-nhan-169210803180337414.htm
105	**4-Aug-21**	936	935	95635		https://suckhoedoisong.vn/toi-3-8-them-4851-ca-mac-covid-19-ca-ngay-ha-noi-them-gan-100-benh-nhan-169210803180337414.htm
106	**5-Aug-21**	3886	*3886*	99521		
107	**6-Aug-21**	1497	1497	101018		https://suckhoedoisong.vn/toi-6-8-them-4315-ca-mac-covid-19-nang-tong-so-mac-trong-ngay-len-8324-ca-rieng-ha-noi-co-116-169210806181934265.htm
108	**7-Aug-21**	5827	*5827*	106845		
109	**8-Aug-21**	2002	2002	108847		https://suckhoedoisong.vn/toi-8-8-them-4949-ca-mac-covid-19-ca-ngay-tang-9690-rieng-binh-duong-3210-ca-169210808181630595.htm
110	**9-Aug-21**	4132	*4132*	112979		

	A	B	C	D	E	F
1	DATA COLLECTION					
2	ĐỢT 4 (27/04/2021 - hiện tại)					
3						
4						
5	Ngày	HCMC Department of Health and Ministry of Health	Other outsources (theo note)	Current Total cases	New cases (Updating)	Note
111	10-Aug-21	1466	1466	114445		https://suckhoedoisong.vn/toi-10-8-them-3241-ca-covid-19-rieng-ha-noi-60-ca-4428-nguoi-khoi-benh-169210810182530874.htm
112	11-Aug-21	3609	3461	118054		https://suckhoedoisong.vn/toi-11-8-them-3964-ca-mac-covid-19-ca-ngay-8776-ca-169210811180958989.htm
113	12-Aug-21	1521	1521	119575		https://suckhoedoisong.vn/toi-12-8-them-5025-ca-covid-19-binh-duong-dan-dau-voi-2117-ca-169210812180300065.htm
114	13-Aug-21	3399	3531	122974		https://suckhoedoisong.vn/ngay-13-8-ca-nuoc-ghi-nhan-9150-ca-mac-moi-covid-19-tp-hcm-va-binh-duong-chiem-den-6347-ca-169210813181458867.htm
115	14-Aug-21	4915	4231	127889		https://suckhoedoisong.vn/toi-14-8-co-9716-ca-mac-covid-19-rieng-tphcm-4231-ca-169210814172340386.htm
116	15-Aug-21	3975	4561	131864		https://ncov.moh.gov.vn/web/guest/-/6847426-6856
117	16-Aug-21	2855	3341	134719		https://suckhoedoisong.vn/toi-16-8-them-8644-ca-mac-covid-19-tai-tphcm-va-42-tinh-thanh-169210816175318102.htm
118	17-Aug-21	3740	*3396*	138459		
119	18-Aug-21	3873	3731	142332		https://suckhoedoisong.vn/toi-18-8-them-8800-ca-mac-covid-19-va-3751-benh-nhan-khoi-169210818175406672.htm
120	19-Aug-21	4307	4425	146639		https://suckhoedoisong.vn/toi-19-8-them-10639-ca-mac-covid-19-rieng-tphcm-va-binh-duong-7860-ca-169210819180148 81.htm
121	20-Aug-21	3504	3375	150143		https://suckhoedoisong.vn/toi-20-8-them-10657-ca-covid-19-binh-duong-co-so-mac-cao-nhat-voi-4223-ca-169210820181508571.htm
122	21-Aug-21	4084	4084	154227		https://suckhoedoisong.vn/toi-21-8-them-11321-ca-covid-19-binh-duong-tiep-tuc-nhieu-nhat-voi-4505-ca-169210821182332656.htm
123	22-Aug-21	4193	4193	158420		https://suckhoedoisong.vn/toi-22-8-them-11214-ca-mac-covid-19-rieng-tphcm-va-binh-duong-da-gan-8000-ca-169210822174454316.htm
124	23-Aug-21	4251	4251	162671		https://suckhoedoisong.vn/toi-23-8-them-10266-ca-mac-covid-19-tphcm-van-nhieu-nhat-voi-4251-ca-169210823175154355.htm
125	24-Aug-21	4634	4627	167305		https://suckhoedoisong.vn/toi-24-8-them-10811-ca-covid-19-tp-ho-chi-minh-va-binh-duong-da-co-den-8255-ca-169210824174804 31.htm

	A	B	C	D	E	F
1	DATA COLLECTION					
2	ĐỢT 4 (27/04/2021 - hiện tại)					
3						
4						
5	Ngày	HCMC Department of Health and Ministry of Health	Other outsources (theo note)	Current Total cases	New cases (Updating)	Note
126	25-Aug-21	5294	5294	172599		https://suckhoedoisong.vn/toi-25-8-them-1209-ca-mac-covid-19-rieng-tphcm-da-5294-ca-169210825180330532.htm
127	26-Aug-21	3934	3934	176533		https://suckhoedoisong.vn/toi-26-8-them-11575-ca-mac-covid-19-ky-luc-hon-18560-benh-nhan-duoc-chua-khoi-169210826180915 35.htm
128	27-Aug-21	5383	5383	181916		https://suckhoedoisong.vn/toi-27-8-them-12920-ca-mac-covid-19-cao-hon-1345-ca-so-voi-hom-qua-169210827184429047.htm
129	28-Aug-21	5481	5481	187397		https://suckhoedoisong.vn/toi-28-8-co-12103-ca-mac-covid-19-rieng-tp-hcm-va-binh-duong-9530-ca-169210828180600805.htm
130	29-Aug-21	4957	4957	192354		https://suckhoedoisong.vn/toi-29-8-them-12663-ca-mac-covid-19-binh-duong-nhieu-nhat-voi-5414-ca-169210829181052356.htm
131	30-Aug-21	5889	5889	198243		https://suckhoedoisong.vn/toi-30-8-co-14224-ca-mac-covid-19-tang-1467-ca-so-voi-hom-qua-169210830180944 44.htm
132	31-Aug-21	5444	5444	203687		https://suckhoedoisong.vn/toi-31-8-co-12607-ca-mac-covid-19-tphcm-va-binh-duong-da-chiem-den-9974-ca-169210831181214185.htm
133	1-Sep-21	5368	5368	209055		https://suckhoedoisong.vn/toi-1-9-them-11434-ca-mac-covid-19-tp-hcm-nhieu-nhat-voi-5368-ca-169210901183207847.htm
134	2-Sep-21	5964		215019		
135	3-Sep-21	8510	8499	223529		https://suckhoedoisong.vn/ngay-3-9-them-14922-ca-mac-covid-19-rieng-tp-hcm-co-den-8499-ca-169210903180914923.htm
136	4-Sep-21	4104	4104	227633		https://suckhoedoisong.vn/ngay-4-9-co-9521-ca-mac-covid-19-thap-hon-5401-ca-so-voi-hom-qua-169210904181341051.htm
137	5-Sep-21	6226	6226	233859		https://suckhoedoisong.vn/ngay-5-9-them-13137-ca-mac-covid-19-rieng-tp-hcm-va-binh-duong-da-gan-9800-ca-169210905180633 01.htm
138	6-Sep-21	7122	7122	240981		https://suckhoedoisong.vn/ngay-6-9-them-12481-ca-mac-covid-19-rieng-tphcm-7122-ca-169210906182042979.htm
139	7-Sep-21	7310	7310	248291		https://suckhoedoisong.vn/ngay-7-9-them-14208-ca-mac-covid-19-cao-hon-hom-qua-1727-ca-169210907181428422.htm
						https://tuoitre.vn/f0-khoi-benh-vung-chai-noi-tuyen-dau-chong-dich-20210908092

	A	B	C	D	E	F
1	DATA COLLECTION					
2	ĐỢT 4 (27/04/2021 - hiện tại)					
3						
4						
5	Ngày	HCMC Department of Health and Ministry of Health	Other outsources (theo note)	Current Total cases	New cases (Updating)	Note
140	8-Sep-21	7308	7308	255599		https://tuoitre.vn/f0-khoi-benh-vung-chai-noi-tuyen-dau-chong-dich-20210908092340422.htm https://vnexpress.net/f0-xuat-vien-o-tp-hcm-dat-ky-luc-4353205.html?zarsrc=30&utm_source=zalo&utm_medium=zalo&utm_campaign=zalo https://suckhoedoisong.vn/ngay-8-9-them-12680-ca-mac-covid-19-tp-hcm-va-binh-duong-gan-10500-ca-169210908180745819.htmhttps://suckhoedoisong.vn/ngay-8-9-them-12680-ca-mac-covid-19-tp-hcm-va-binh-duong-gan-10500-ca-169210908180745819.htm
141	9-Sep-21	5549	5549	261148		https://suckhoedoisong.vn/ngay-9-9-viet-nam-ghi-nhan-12420-ca-mac-covid-19-va-12523-benh-nhan-khoi-169210909181525359.htm
142	10-Sep-21	7539	7539	268687		https://suckhoedoisong.vn/ngay-10-9-them-13321-ca-mac-covid-19-tp-hcm-va-binh-duong-chiem-den-hon-11100-ca-169210910182529294.htm https://giadinh.net.vn/y-te/f0-vuot-qua-lo-lang-tram-cam-do-mat-nguoi-than-nho-ho-tro-tam-ly-20210910154900451.htm
143	11-Sep-21	5629	5629	274316		https://suckhoedoisong.vn/ngay-11-9-co-11932-ca-mac-covid-19-it-hon-hom-qua-gan-1400-ca-169210911181543947.htm https://suckhoedoisong.vn/bo-y-te-so-ca-nhiem-covid-19-tai-cong-dong-va-tu-vong-giam-so-voi-tuan-truoc-169210911141621468.htm
144	12-Sep-21	6158	6158	280474		https://tuoitre.vn/bo-truong-bo-y-te-so-ca-mac-va-tu-vong-o-tp-hcm-co-xu-huong-giam-ro-ret-20210912115901501.htm https://suckhoedoisong.vn/ngay-12-9-them-11478-ca-mac-covid-19-rieng-tp-hcm-va-binh-duong-ghi-nhan-gan-9350-ca-169210912180815617.htm
145	13-Sep-21	5446	5446	285920		https://suckhoedoisong.vn/ngay-13-9-co-11172-ca-mac-covid-19-tp-hcm-nhieu-nhat-voi-5446-ca-169210913182924132.htm
146	14-Sep-21	6315	6312	292235		https://suckhoedoisong.vn/ngay-14-9-them-10508-ca-mac-covid-19-trong-do-tp-hcm-va-binh-duong-da-gan-8500-ca-169210914182008422.htm
147	15-Sep-21	5301	5301	297536		https://suckhoedoisong.vn/ngay-15-9-co-10585-ca-mac-covid-19-rieng-tp-hcm-da-5301-ca-169210915182952278.htm
						https://suckhoedoisong.vn/ngay-16-9-them-10489-ca-mac-covid-19-rieng-tp-hcm-

	A	B	C	D	E	F
1	DATA COLLECTION					
2	ĐỢT 4 (27/04/2021 - hiện tại)					
3						
4						
5	Ngày	HCMC Department of Health and Ministry of Health	Other outsources (theo note)	Current Total cases	New cases (Updating)	Note
148	16-Sep-21	5735	5735	303271		https://suckhoedoisong.vn/ngay-16-9-them-10489-ca-mac-covid-19-rieng-tp-hcm-va-binh-duong-hon-8700-ca-169210916180856675.htm
149	17-Sep-21	5972	5972	309243		https://suckhoedoisong.vn/ngay-17-9-co-11521-ca-mac-covid-19-trong-do-tphcm-va-binh-duong-da-gan-10000-ca-169210917181345586.htm
150	18-Sep-21	4273	4273	313516		https://suckhoedoisong.vn/ngay-18-9-co-9373-ca-mac-covid-19-thap-hon-hom-qua-2146-ca-169210918180826807.htm
151	19-Sep-21	5496	5496	319012		https://suckhoedoisong.vn/ngay-19-9-them-10040-ca-mac-covid-19-trong-do-rieng-tp-hcm-co-5496-ca-169210919175648403.htm
152	20-Sep-21	5172	5172	324184		https://suckhoedoisong.vn/sang-20-9-gan-5400-ca-covid-19-nang-dang-dieu-tri-15-dia-phuong-qua-14-ngay-chua-ghi-nhan-f0-trong-nuoc-169210920063247474.htm
153	21-Sep-21	6521	6521	330705		https://suckhoedoisong.vn/ngay-21-9-co-11692-ca-mac-covid-19-tai-tp-hcm-va-33-tinh-thanh-pho-169210921180551315.htm https://vnexpress.net/he-so-lay-nhiem-dang-giam-tp-hcm-co-the-noi-long-gian-cach-4359175.html
154	22-Sep-21	5435	5435	336140		https://tuoitre.vn/dung-coi-thuong-hoi-chung-hau-covid-19-20210922105425001.htm https://suckhoedoisong.vn/ngay-22-9-co-11527-ca-mac-covid-19-rieng-tp-hcm-va-binh-duong-da-ghi-nhan-hon-9600-ca-169210922181558737.htm
155	23-Sep-21	5052	5052	341192		https://suckhoedoisong.vn/ngay-23-9-co-9472-ca-mac-covid-19-giam-2060-ca-so-voi-ngay-hom-qua-169210923175100815.htm https://m.giadinh.net.vn/y-te/khung-hoang-stress-tam-ly-o-nhung-nguoi-mac-covid-19-20210923141609546.htm
156	24-Sep-21	3786	3786	344978		https://suckhoedoisong.vn/ngay-24-9-ghi-nhan-8537-ca-mac-covid-19-thap-nhat-trong-hon-1-thang-qua-cua-dot-dich-nay-169210924181328392.htm
157	25-Sep-21	4050	4046	349028		https://suckhoedoisong.vn/ngay-25-9-co-9706-ca-mac-covid-19-trong-do-tp-hcm-va-binh-duong-da-ghi-nhan-7675-ca-169210925172705503.htm
158	26-Sep-21	5121	5121	354149		https://suckhoedoisong.vn/ngay-26-9-them-10011-ca-mac-covid-19-rieng-tp-hcm-da-5121-ca-169210926182030129.htm
						https://suckhoedoisong.vn/ngay-27-9-co-9362-ca-mac-covid-19-tai-tp-hcm-binh-d

	A	B	C	D	E	F
1	DATA COLLECTION					
2	ĐỢT 4 (27/04/2021 - hiện tại)					
3						
4						
5	Ngày	HCMC Department of Health and Ministry of Health	Other outsources (theo note)	Current Total cases	New cases (Updating)	Note
159	27-Sep-21	4134	4134	358283		https://suckhoedoisong.vn/ngay-27-9-co-9362-ca-mac-covid-19-tai-tp-hcm-binh-duong-va-34-tinh-thanh-pho-169210927174122985.htm
160	28-Sep-21	3749	377	362032		https://suckhoedoisong.vn/ngay-28-9-so-mac-moi-covid-19-chi-4589-ca-trong-khi-so-khoi-nhieu-gap-gan-5-lan-169210928180738217.htm
161	29-Sep-21	4699	4699	366731		https://suckhoedoisong.vn/ngay-29-9-co-8758-ca-mac-covid-19-va-so-benh-nhan-khoi-ky-luc-voi-23568-ca-169210929180546707.htm
162	30-Sep-21	4372	4372	371103		https://suckhoedoisong.vn/ngay-30-9-co-7940-ca-mac-covid-19-so-benh-nhan-khoi-lap-ky-luc-moi-voi-25322-ca-169210930181734537.htm
163	1-Oct-21	3670	1787	374773		https://suckhoedoisong.vn/ngay-1-10-co-6957-ca-mac-covid-19-so-benh-nhan-khoi-dat-ky-luc-moi-voi-27520-nguoi-169211001181857662.htm
164	2-Oct-21	2723	2723	377496		https://suckhoedoisong.vn/ngay-2-10-co-5490-ca-mac-moi-covid-19-thap-nhat-trong-thoi-gian-qua-169211002183753 93.htm
165	3-Oct-21	2461	2461	379957		https://suckhoedoisong.vn/ngay-3-10-co-5376-ca-mac-covid-19-so-benh-nhan-khoi-lap-ky-luc-moi-voi-28859-ca-169211003181621008.htm
166	4-Oct-21	2490	2490	382447		https://suckhoedoisong.vn/ngay-4-10-co-5383-ca-mac-moi-covid-19-rieng-tp-hcm-la-2490-ca-169211004182230114.htm
167	5-Oct-21	1491	1491	383938		https://suckhoedoisong.vn/ngay-5-10-co-4363-ca-mac-covid-19-thap-nhat-trong-khoang-15-thang-qua-169211005180604662.htm https://tuoitre.vn/ca-benh-giam-manh-nhieu-benh-vien-dieu-tri-covid-19-o-tp-hcm-trong-giuong-20211005160241973.htm
168	6-Oct-21	1960	1960	385898		https://suckhoedoisong.vn/ngay-6-10-co-4363-ca-mac-covid-19-tai-tphcm-va-39-tinh-thanh-pho-169211006181056 49.htm
169	7-Oct-21	1730	1730	387628		https://suckhoedoisong.vn/ngay-7-10-co-4150-ca-mac-covid-19-giam-hon-200-ca-so-voi-hom-qua-169211007181929683.htm
170	8-Oct-21	2215	2215	389843		https://suckhoedoisong.vn/ngay-8-10-them-4806-ca-mac-covid-19-rieng-tp-hcm-co-2215-ca-169211008181231204.htm
171	9-Oct-21	1662	1662	391505		https://suckhoedoisong.vn/ngay-9-10-co-4513-ca-mac-covid-19-tai-tp-hcm-va-39-tinh-thanh-pho-giam-261-ca-so-voi-hom-qua-169211009180119826.htm

	A	B	C	D	E	F
1	DATA COLLECTION					
2	ĐỢT 4 (27/04/2021 - hiện tại)					
3						
4						
5	Ngày	HCMC Department of Health and Ministry of Health	Other outsources (theo note)	Current Total cases	New cases (Updating)	Note
172	10-Oct-21	1067	1067	392572		https://suckhoedoisong.vn/ngay-10-10-chi-co-3528-ca-mac-covid-19-nhung-co-den-21398-benh-nhan-khoi-169211010180104264.htm
173	11-Oct-21	1527	1527	394099		https://suckhoedoisong.vn/ngay-11-10-co-3619-ca-mac-covid-19-tai-44-dia-phuong-rieng-tp-hcm-1527-ca-169211011181935125.htmm
174	12-Oct-21	1018	1018	395117		https://suckhoedoisong.vn/ngay-12-10-chi-co-2949-ca-mac-covid-19-tai-43-dia-phuong-thap-nhat-trong-25-thang-qua-169211012180220603.htm
175	13-Oct-21	1162	1162	396279		https://suckhoedoisong.vn/ngay-14-10-co-3092-ca-mac-covid-19-rieng-tp-hcm-909-ca-169211014182809507.htm
176	14-Oct-21	909	909	397188		https://suckhoedoisong.vn/ngay-14-10-co-3092-ca-mac-covid-19-rieng-tp-hcm-909-ca-169211014182809507.htm
177	15-Oct-21	1131	1131	398319		https://suckhoedoisong.vn/ngay-15-10-co-3797-ca-mac-covid-19-tai-tp-hcm-soc-trang-va-45-dia-phuong-khac-169211015181921807.htm
178	16-Oct-21	790	790	399109		https://suckhoedoisong.vn/ngay-16-10-co-3211-ca-mac-covid-19-tai-48-tinh-thanh-pho-giam-578-ca-so-voi-hom-qua-169211016180701583.htm
179	17-Oct-21	1059	1059	400168		https://suckhoedoisong.vn/ngay-17-10-co-3193-ca-mac-covid-19-rieng-tp-hcm-la-1059-ca-so-tu-vong-giam-con-63-ca-169211017175617838.htm
180	18-Oct-21	968	968	401136		https://suckhoedoisong.vn/ngay-18-10-co-3168-ca-mac-covid-19-tai-tp-hcm-soc-trang-va-43-tinh-thanh-khac-169211018182407029.htm
181	19-Oct-21	907	907	402043		https://suckhoedoisong.vn/ngay-19-10-co-3034-ca-mac-covid-19-tai-tp-hcm-va-48-tinh-thanh-khac-giam-132-ca-voi-ngay-qua-169211019180933659.htm
182	20-Oct-21	1347	1347	403390		https://suckhoedoisong.vn/ngay-20-10-co-3646-ca-mac-covid-19-hon-1700-benh-nhan-khoi-169211020181810142.htm https://nld.com.vn/suc-khoe/ngay-20-10-ky-luc-tiem-gan-2-trieu-lieu-vac-xin-so-tu-vong-do-covid-19-o-tp-hcm-thap-nhat-nhieu-thang-qua-20211020161005188.htm
183	21-Oct-21	1255	1255	404645		https://suckhoedoisong.vn/ngay-21-10-co-3636-ca-mac-covid-19-tai-50-tinh-thanh-1541-benh-nhan-khoi-169211021182707088.htm https://tuoitre.vn/ban-do-mau-cap-do-dich-toan-quoc-20211021123000992.htm
184	22-Oct-21	1205	1205	405850		https://suckhoedoisong.vn/ngay-22-10-co-3985-ca-mac-covid-19-va-5202-nguoi-k

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4						
5	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
185	**23-Oct-21**	749	*749*	406599		
186	**24-Oct-21**	966	966	407565		https://suckhoedoisong.vn/ngay-24-10-co-4045-ca-mac-covid-19-tai-47-tinh-thanh-them-386400-lieu-vaccine-astrazeneca-ve-viet-nam-169211024180227579.htm
187	**25-Oct-21**	969	969	408534		https://suckhoedoisong.vn/ngay-25-10-co-3639-ca-mac-covid-19-tai-tp-hcm-va-52-tinh-thanh-1323-benh-nhan-khoi-169211025183826993.htm
188	**26-Oct-21**	783	783	409317		https://suckhoedoisong.vn/ngay-26-10-co-3595-ca-mac-covid-19-tai-tp-hcm-ha-noi-va-47-tinh-thanh-gan-3000-benh-nhan-khoi-169211026183804751.htm
189	**27-Oct-21**	1140	1140	410457		https://suckhoedoisong.vn/ngay-27-10-co-4411-ca-mac-covid-19-tai-47-tinh-thanh-tang-hon-800-ca-so-voi-ngay-qua-169211027182331 95.htm
190	**28-Oct-21**	1069	1069	411526		
191	**29-Oct-21**	977	977	412503		
192	**30-Oct-21**	1042	1042	413545		
193	**31-Oct-21**	1041	1041	414586		
194	**1-Nov-21**	927	927	415513		
195	**2-Nov-21**	682	682	416195		
196	**3-Nov-21**	985	985	417180		
197	**4-Nov-21**	981	981	418161		
198	**5-Nov-21**	912	912	419073		
199	**6-Nov-21**	986	986	420059		
200	**7-Nov-21**	1009	1009	421068		
201	**8-Nov-21**	1316	1316	422384		
202	**9-Nov-21**	1276	1276	423660		
203	**10-Nov-21**	1414	1414	425074		
204	**11-Nov-21**	1185	1185	426259		
205	**12-Nov-21**	1388	1388	427647		
206	**13-Nov-21**	1240	1240	428887		
207	**14-Nov-21**	985	985	429872		
208	**15-Nov-21**	1165	1165	431037		

	A	B	C	D	E	F
1	**DATA COLLECTION**					
2	**ĐỢT 4 (27/04/2021 - hiện tại)**					
3						
4						
5	**Ngày**	**HCMC Department of Health and Ministry of Health**	**Other outsources (theo note)**	**Current Total cases**	**New cases (Updating)**	**Note**
190	**28-Oct-21**	1069	1069	411526		https://vnexpress.net/covid-19/covid-19-viet-nam
191	**29-Oct-21**	977	977	412503		https://suckhoedoisong.vn
192	**30-Oct-21**	1042	1042	413545		https://tuoitre.vn
193	**31-Oct-21**	1041	1041	414586		https://m.giadinh.net.vn/y-te
194	**1-Nov-21**	927	927	415513		https://nld.com.vn/suc-khoe/
195	**2-Nov-21**	682	682	416195		https://baochinhphu.vn/Suc-khoe
196	**3-Nov-21**	985	985	417180		
197	**4-Nov-21**	981	981	418161		
198	**5-Nov-21**	912	912	419073		
199	**6-Nov-21**	986	986	420059		
200	**7-Nov-21**	1009	1009	421068		
201	**8-Nov-21**	1316	1316	422384		
202	**9-Nov-21**	1276	1276	423660		
203	**10-Nov-21**	1414	1414	425074		
204	**11-Nov-21**	1185	1185	426259		
205	**12-Nov-21**	1388	1388	427647		
206	**13-Nov-21**	1240	1240	428887		
207	**14-Nov-21**	985	985	429872		
208	**15-Nov-21**	1165	1165	431037		
209	**16-Nov-21**	1183	1183	432220		
210	**17-Nov-21**	1337	1337	433557		
211	**18-Nov-21**	1609	1609	435166		
212	**19-Nov-21**	1339	1339	436505		
213	**20-Nov-21**	1046	1046	437551		
214	**21-Nov-21**	1265	1265	438816		
215	**22-Nov-21**	1547	1547	440363		
216	**23-Nov-21**	1204	1204	441567		

2. External Parameter

Date	Daily Vaccination rate	Social distancing	Incubation period	Generation time
27/04/2021	-	2	3,41	2,76
28/04/2021	-	2	6,02	3,89
29/04/2021	-	2	7,04	2,51
30/04/2021	-	2	2,28	0,40
01/05/2021	-	2	1,16	4,82
02/05/2021	-	2	5,97	5,16
03/05/2021	8047	2	0,90	2,87
04/05/2021	7996	2	4,92	2,86
05/05/2021	8043	2	6,29	0,90
06/05/2021	7520	2	6,04	4,53
07/05/2021	7055	2	3,08	2,20
08/05/2021	6443	2	2,66	5,15
09/05/2021	3424	2	2,29	2,50
10/05/2021	1042	2	4,33	4,13
11/05/2021	1556	2	1,82	3,23
12/05/2021	1734	2	5,88	4,34
13/05/2021	1906	2	4,74	4,75
14/05/2021	1982	2	3,67	4,66
15/05/2021	1964	2	8,32	4,28
16/05/2021	1787	2	6,80	4,78
17/05/2021	1437	2	8,03	4,72
18/05/2021	971	2	6,33	0,26
19/05/2021	822	2	4,54	1,85
20/05/2021	831	2	2,18	2,90
21/05/2021	899	2	4,22	0,45
22/05/2021	903	2	4,43	1,53
23/05/2021	871	2	2,95	1,56
24/05/2021	632	2	3,66	2,90
25/05/2021	566	2	4,88	1,31
26/05/2021	518	2	3,71	3,57
27/05/2021	318	2	5,25	4,12
28/05/2021	204	2	3,98	0,94
29/05/2021	204	2	3,45	4,77
30/05/2021	204	2	3,05	2,20
31/05/2021	204	2	4,52	3,64
01/06/2021	204	2	1,77	3,18
02/06/2021	204	2	3,87	4,50
03/06/2021	204	2	6,34	3,92
04/06/2021	204	2	5,53	2,92
05/06/2021	204	2	1,89	3,67
06/06/2021	204	2	4,70	5,61
07/06/2021	204	2	3,73	3,83
08/06/2021	204	2	5,77	4,08
09/06/2021	204	2	2,77	3,12

Date	Daily Vaccination rate	Social distancing	Incubation period	Generation time
09/06/2021	204	2	2,77	3,12
10/06/2021	204	2	3,85	0,68
11/06/2021	204	2	3,60	4,50
12/06/2021	204	2	6,05	2,95
13/06/2021	204	2	6,22	3,46
14/06/2021	204	2	6,17	4,77
15/06/2021	204	2	3,05	3,52
16/06/2021	204	2	3,09	2,50
17/06/2021	204	2	2	3,71
18/06/2021	204	2	2,34	0,69
19/06/2021	204	3	8,59	0,54
20/06/2021	204	3	7,79	2,79
21/06/2021	204	3	3,58	1,84
22/06/2021	204	3	2,50	2,84
23/06/2021	204	3	4,38	2,65
24/06/2021	204	3	5,47	2,90
25/06/2021	73.262	3	5,93	3,18
26/06/2021	92.035	3	5,32	3,21
27/06/2021	88.946	3	2,94	1,44
28/06/2021	95.003	3	7,62	2,90
29/06/2021	104.147	3	3,33	2,65
30/06/2021	86.719	3	6,88	1,03
01/07/2021	62.020	3	3,91	4,31
02/07/2021	40.163	3	3,26	4,13
03/07/2021	18.306	3	4,27	4,35
04/07/2021	18.312	3	4,12	1,10
05/07/2021	9.172	3	5,50	3,78
06/07/2021	21.488	3	5,85	2,42
07/07/2021	16.820	3	4,15	4,90
08/07/2021	16.448	3	4,25	1,26
09/07/2021	15.982	4	5,68	2,71
10/07/2021	17.684	4	3,25	1,61
11/07/2021	16.733	4	5,51	3,56
12/07/2021	16.712	4	4,54	2,90
13/07/2021	16.778	4	3,68	3,16
14/07/2021	16.977	4	1,71	3,08
15/07/2021	16.800	4	6,53	4,31
16/07/2021	16.817	4	4,86	2,51
17/07/2021	16.843	4	5,31	3,03
18/07/2021	16.859	4	4,35	2,41
19/07/2021	16.830	4	5,62	3,79
20/07/2021	16.837	4	5,09	1,77
21/07/2021	16.842	4	3,39	4,60
22/07/2021	3.100	4	4,47	4,49

Date	Daily Vaccination rate	Social distancing	Incubation period	Generation time
23/07/2021	3.029	4	1,31	3,18
24/07/2021	16.548	4	0,22	4,44
25/07/2021	36.094	4	5,78	1,28
26/07/2021	47.676	4	2,43	3,20
27/07/2021	39.624	4	1,89	4,34
28/07/2021	69.211	4	3,59	4,58
29/07/2021	72.646	4	6,90	3,47
30/07/2021	78.122	4	3,85	5,64
31/07/2021	82.709	4	4,11	0,32
01/08/2021	95.500	4	4,88	2,90
02/08/2021	118.121	4	6,66	5,56
03/08/2021	111.878	4	6,08	0,47
04/08/2021	145.576	4	4,04	3,18
05/08/2021	671.513	4	3,31	1,99
06/08/2021	250.243	4	5,23	3,69
07/08/2021	210.791	4	2,24	7,31
08/08/2021	319.531	4	2,84	3,46
09/08/2021	363.020	4	3,05	4,40
10/08/2021	285.896	4	5,18	2,51
11/08/2021	294.810	4	2,40	5,36
12/08/2021	315.814	4	5,83	3,21
13/08/2021	93.993	4	4,35	2,99
14/08/2021	85.608	4	7,18	1,50
15/08/2021	197.556	4	5,16	2,13
16/08/2021	194.435	4	3,16	1,46
17/08/2021	126.157	4	2,93	6,76
18/08/2021	150.939	4	3,25	1,37
19/08/2021	142.223	4	3,19	0,39
20/08/2021	116.523	4	7,91	2,20
21/08/2021	95.516	4	4,82	3,86
22/08/2021	61.574	4	6,11	1,87
23/08/2021	58.817	5	0,59	6,76
24/08/2021	37.746	5	6,58	4,15
25/08/2021	57.982	5	5,07	1,90
26/08/2021	51.886	5	4,12	1,42
27/08/2021	62.349	5	4,40	8,09
28/08/2021	39.546	5	2,93	1,82
29/08/2021	32.603	5	7,65	5,98
30/08/2021	30.109	5	4,67	2,84
31/08/2021	40.212	5	6,55	3,27
01/09/2021	33.448	5	6,34	2,90
02/09/2021	63.341	5	6,22	1,50
03/09/2021	41.778	5	5,89	4,65
04/09/2021	74.998	5	7,76	4,78
05/09/2021	87.683	5	3,26	3,21
06/09/2021	105.781	5	6,17	4,13
07/09/2021	161.286	5	2,48	1,42
08/09/2021	163.548	5	6,70	2,90
09/09/2021	188.124	5	5,22	3,44
10/09/2021	183.699	5	1,86	4,15
11/09/2021	214.347	5	2,39	5,56
12/09/2021	246.332	5	7,02	2,71
13/09/2021	174.090	5	6,87	2,90
14/09/2021	177.119	5	4,75	1,42
15/09/2021	148.671	5	4,50	4,38
16/09/2021	119.193	5	2,38	5,09
17/09/2021	78.163	5	3,54	1,62
18/09/2021	86.403	5	5,74	2,71
19/09/2021	87.666	5	4,61	0,12
20/09/2021	30.117	5	4,59	3,91
21/09/2021	71.836	5	4,69	3,02
22/09/2021	513.377	5	2,13	1,77
23/09/2021	67.566	5	3,04	4,70
24/09/2021	70.899	5	3,06	3,69
25/09/2021	134.617	5	1,86	0,88
26/09/2021	237.373	5	4,68	0,64
27/09/2021	205.951	5	7,48	2,38
28/09/2021	132.884	5	1,89	3,26
29/09/2021	325.696	5	5,94	2,90
30/09/2021	251.678	5	7,13	2,48
01/10/2021	157.600	2	2,29	5,07
02/10/2021	269.621	2	2,44	4,18
03/10/2021	233.471	2	3,27	3,72
04/10/2021	283.448	2	4,40	1,44
05/10/2021	225.322	2	3,89	4,19
06/10/2021	173.924	2	4,55	3,28
07/10/2021	128.966	2	2,89	5,36
08/10/2021	102.080	2	7,86	2,88
09/10/2021	80.506	2	4,74	5,01
10/10/2021	96.186	2	5,45	3,54
11/10/2021	56.731	2	0,58	2,14
12/10/2021	61.728	2	4,28	1,65
13/10/2021	51.325	2	-0,81	3,95
14/10/2021	36.191	2	2,86	1,74
15/10/2021	38.759	2	0,39	0,86
16/10/2021	27.716	2	4,45	3,09
17/10/2021	25.451	2	3,50	0,92
18/10/2021	15.216	2	4,60	5,48

Date	Daily Vaccination rate	Social distancing	Incubation period	Generation time
06/10/2021	173.924	2	4,55	3,28
07/10/2021	128.966	2	2,89	5,36
08/10/2021	102.080	2	7,86	2,88
09/10/2021	80.506	2	4,74	5,01
10/10/2021	96.186	2	5,45	3,54
11/10/2021	56.731	2	0,58	2,14
12/10/2021	61.728	2	4,28	1,65
13/10/2021	51.325	2	-0,81	3,95
14/10/2021	36.191	2	2,86	1,74
15/10/2021	38.759	2	0,39	0,86
16/10/2021	27.716	2	4,45	3,09
17/10/2021	25.451	2	3,50	0,92
18/10/2021	15.216	2	4,60	5,48
19/10/2021	23.537	2	3,12	2,09
20/10/2021	20.939	2	2,24	3,76
21/10/2021	21.193	2	1,57	4,26
22/10/2021	20.223	2	5,31	1,81
23/10/2021	23.382	2	4,61	4,08
24/10/2021	16.741	2	3,45	2,90
25/10/2021	7.922	2	2,93	6,54
26/10/2021	23.506	2	4,30	4,63
27/10/2021	23.035	2	4,46	6,74
28/10/2021	36.248	2	4,41	2,90
29/10/2021	68.863	2	1,35	0,51
30/10/2021	68.996	2	4,35	5,02
31/10/2021	32.272	2	4,05	2,81
01/11/2021	27.746	2	3,54	4,97
02/11/2021	29.989	2	6,10	3,59
03/11/2021	26.850	2	5,51	2,24
04/11/2021	24.424	2	2,70	5,23
05/11/2021	21.597	2	3,48	2,90
06/11/2021	26.246	2	5,71	4,19
07/11/2021	13.830	2	7,87	5,17
08/11/2021	5.541	2	6,89	1,89
09/11/2021	20.730	2	3,26	4,69
10/11/2021	15.926	2	1,73	2,49
11/11/2021	16.297	2	1,29	2,81
12/11/2021	46.409	2	4,01	2,16
13/11/2021	24.596	2	4,24	5,73
14/11/2021	13.743	2	5,64	2,17
15/11/2021	4.920	2	5,68	3,40
16/11/2021	18.184	2	2,75	6,07
17/11/2021	39.113	2	3,93	5,19
18/11/2021	18.451	2	8,01	4,84

Appendix B. ADF test

As mentioned in the main part, this ADF test (fully called Augmented Dickey Fuller test) step was conducted for the preparation of the ARIMA process right after being finished. This type of test is a common statistical test used to test whether a given Time series data (in this case is the ARIMA data) is stationary or not. Additionally, it is one of the most common statistical tests for analyzing the stationary series.

Particularly, in this paper, EViews which is a software including statistical package for Windows, mainly used for time series econometric analysis orientation, was used as a supporting tool for the group to consider whether the series of data from the end of April to the end of November in 2021 was stationary or not for ARIMA step. Hence, thanks to this support, the team conducted the results mentioned in the above parts, which are that the original data is not stationary, so the team had to make a difference once for this series to make it acceptable for the ARIMA model. The below figures shows the result of these two steps:

Null Hypothesis: SER01 has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 5 (Automatic - based on t-statistic, lagpval=0.05, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.229351	0.9012
Test critical values:	1% level	-4.004365	
	5% level	-3.432339	
	10% level	-3.139924	

*MacKinnon (1996) one-sided p-values.

Figure 4:ADF test result when first added the original data series

Null Hypothesis: D(SER01) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 4 (Automatic - based on t-statistic, lagpval=0.05, maxlag=14)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.664411	0.0000
Test critical values:	1% level	-4.004365	
	5% level	-3.432339	
	10% level	-3.139924	

*MacKinnon (1996) one-sided p-values.

Figure 5: ADF test result when first took first difference order of the original data

Appendix C. ACF + PACF

ACF standing for Autocorrelation function and PACF for Partial Autocorrelation function were used for the research team to form an idea of what ARIMA models fit the current data. These below figures show the ACF and PACF conducted by adding data from April 27th, 2021 to the end of November, 2021 and the support of Minitab computer software which is statistical package on Windows with the function that helping users make better decisions with unparalleled statistical insights and smart víualizations thanks to its predictive tools, visualise data faster with its graph builder display, and boost to solve more challenging problems by its functional algorithm available on Minitab's Predictive Modules.

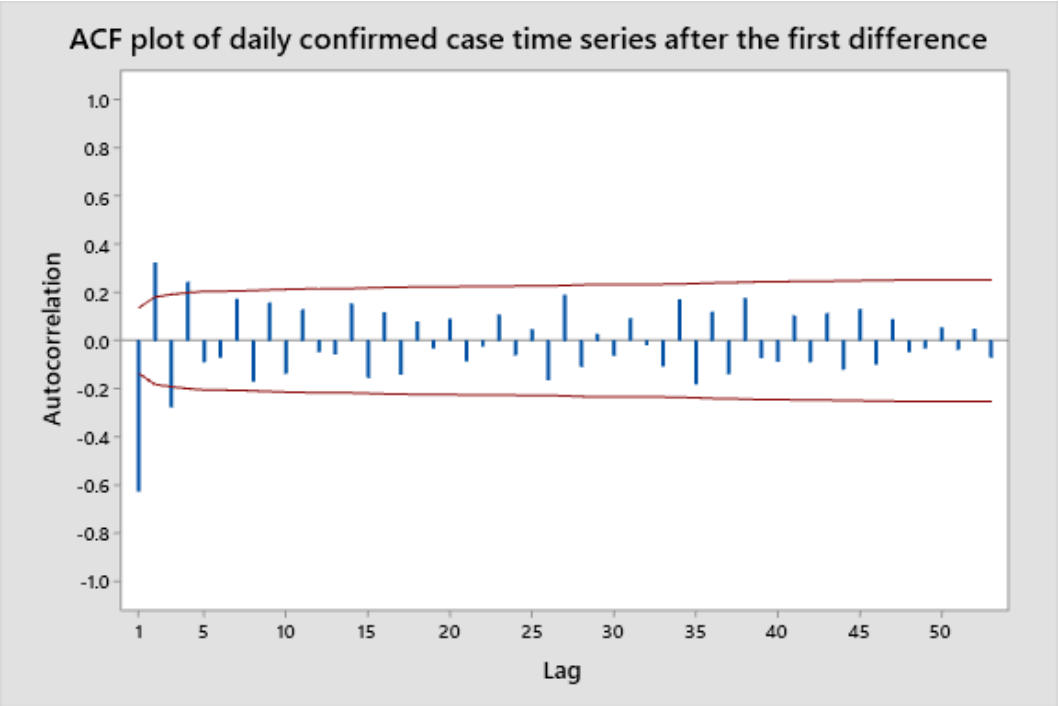


Figure 6: ACF plot used in this paper

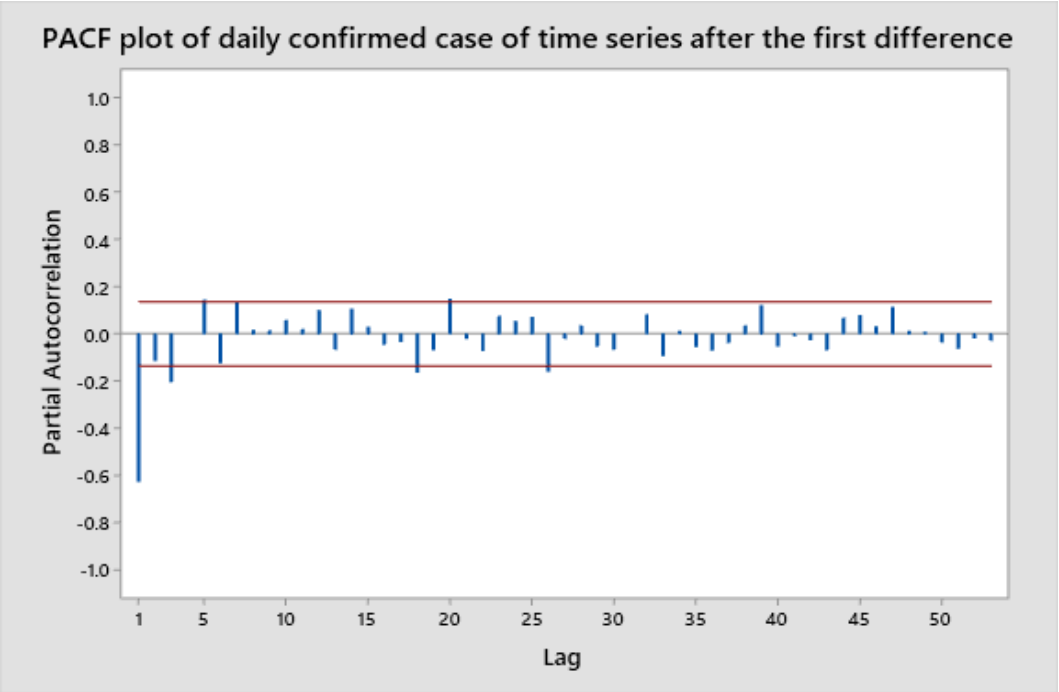


Figure 7: PACF plot used in this paper

Appendix D. Python code

The study considers Python as an application to establish the forecasting model and implement the external parameters such as vaccination rate, covid-variant and control measures into the final model. The following figures describe the detail code, explanation of the code is directly provided next to the coding sentences:

1. Import libraries into Python file:

```
1
2   import pandas as pd
3   import numpy as np
4   import matplotlib.pyplot as plt
5   from statsmodels.tsa.stattools import adfuller
6   from statsmodels.tsa.arima.model import ARIMA
7   from pandas import DataFrame
8   from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
9   import warnings; warnings.filterwarnings(action='once')
10  import math
```

2. Import excel file, this excel file contains the COVID-19 confirmed cases have been recorded daily from 27th April to 24th November 2021. After which, the data was plotted in order to visualize the distribution of the collected data

```
11
12  WB = pd.read_excel("Data import for ADF.xlsx")
13  # Plot data into a graph:
14  y = list(WB['Data value'])
15  x = list(WB['Date'])
16  plt.title('Recorded confirmed cases time series plot')
17  plt.plot(x, y)
18  plt.show()
19
```

3. The first step in model ARIMA development is considering whether the data is stationary. Hence, an ADF test is provided in the code. From which, d parameters will then be determined based on the stationarity of data in which differencing stage. The COVID-19 confirmed cases in this study is considered to be stationary on the first differencing, which means d = 1 in ARIMA model:

```
# ADF test
# Considering the stationarity of the original data
X = WB['Data value'].values
result = adfuller(X)
print('The result of ADF test of the original value is:')
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
    print(key, value)

if result[0] < result[4]["5%"]:
    print("Reject Ho - Time Series is Stationary")
else:
    print("Failed to Reject Ho - Time Series is Non-Stationary")
```

```
# Considering the 1st differencing
WB['1st_Differencing'] = WB['Data value'] - WB['Data value'].shift(1)
plt.plot(x, WB['1st_Differencing'])
plt.title('First differencing of recorded confirmed cases')
plt.show()
result2 = adfuller(WB['1st_Differencing'].dropna())
# Note we are dropping na values because the first value of the first difference is NA
print('-----------------------------------------------------------------')
print('The result of ADF test of 1st Differencing is:')
print('Critical Values:')
for key, value in result2[4].items():
    print(key, value)
print('ADF Statistic: %f' % result2[0])
print('p-value: %f' % result2[1])
if result2[0] < result2[4]["5%"]:
    print("Reject Ho - Time Series is Stationary")
else:
    print("Failed to Reject Ho - Time Series is Non-Stationary")

```

4. AR(p) and MA(q) were determined using ACF and PACF method, the code is as following:

```
# ACF and PACF method
WB['1st_Differencing'] = np.asarray(WB['1st_Differencing'])
fig, (ax1, ax2) = plt.subplots(1,2,figsize=(16,6), dpi= 80)
ACF_plot = plot_acf(WB['1st_Differencing'].dropna(),ax=ax1, lags=60)
PACF_plot = plot_pacf(WB['1st_Differencing'].dropna(),ax=ax2, lags=60, method='ywm')
# Decoration:
# 1.lighten the borders
ax1.spines["top"].set_alpha(.3); ax2.spines["top"].set_alpha(.3)
ax1.spines["bottom"].set_alpha(.3); ax2.spines["bottom"].set_alpha(.3)
ax1.spines["right"].set_alpha(.3); ax2.spines["right"].set_alpha(.3)
ax1.spines["left"].set_alpha(.3); ax2.spines["left"].set_alpha(.3)
# 2.Font size of tick labels
ax1.tick_params(axis='both', labelsize=12)
ax2.tick_params(axis='both', labelsize=12)
plt.show()
```

5. To implement the external parameters into the model, the study uses Pearson's correlation to determine the relationships between Confirmed cases and external data defined. However, since the correlation can easily be calculated through excel, the following code only considers when the Beta parameters of external data have been defined and implement these parameters into the ARIMA model for simplicity:

```
# Implementing external parameters, the coefficients were performed in excel file
WB2 = pd.read_excel("External Parameters.xlsx")
Vaccination = WB2['Vaccination coverage (PPM)'].values
Social_distance = WB2['Social distancing'].values
Incubation = WB2['Incubation period'].values
Generation_time = WB2['Generation time'].values
beta_vaccination = 0.361
beta_distancing = 0.577
beta_incubation = -0.03
beta_generation = 0.092
X1 = beta_vaccination*Vaccination
X2 = beta_distancing*Social_distance
X3 = beta_incubation*Incubation
X4 = beta_generation*Generation_time
WB['External'] = X1 + X2 + X3 + X4
```

6. The ARIMA parameters and external parameters were used to forecast the COVID-19 confirmed cases in this stage, the following code provides a way to plot these data and include the real data for fitting. Noted that in this stage, a 95% confidence level is applied for COVID-19 confirmed cases collected

```
# ARIMA model: provide a forecasting model of ARIMA => ARIMA (1,1,0)
model = ARIMA(WB['Data value'], order=(1,1,0))
RESULT = model.fit()
print(RESULT.summary())

fig2, ax = plt.subplots(1,2,figsize=(16,6), dpi= 80)
residuals = DataFrame(RESULT.resid) # Line plot of residuals
residuals.plot(title="Residuals", ax=ax[0])
residuals.plot(title="Density", ax=ax[1], kind='kde') # Density plot of residuals
plt.show()

# Summary stats of residuals
print(residuals.describe())
# Forecast ARIMA (1,1,0)
WB['ARIMA(1,1,0)'] = RESULT.predict(start=0, end=207, dynamic=False)
WB['Final Forecast result'] = WB['ARIMA(1,1,0)'] + WB['External']
```

```

# 95% Confidence interval:
CI = 1.96*np.std(WB['Data value'])/np.sqrt(len(x)) # since Confidence level = 95%, z-value = 1.96
# Plot
fig3, ax3 = plt.subplots(1,1)
plt.title('Forecast of COVID-19 Confirmed Cases in HCMC', alpha = 0.5)
plt.plot(x, WB['Data value'])
plt.plot(x, WB['Final Forecast result'])
ax3.fill_between(x, (WB['Data value'] - CI), (WB['Data value'] + CI), color = "b", alpha = 0.2)
plt.legend(["COVID-19 Confirmed Cases","ARIMA(1,1,0) Forecast"])
plt.show()
```

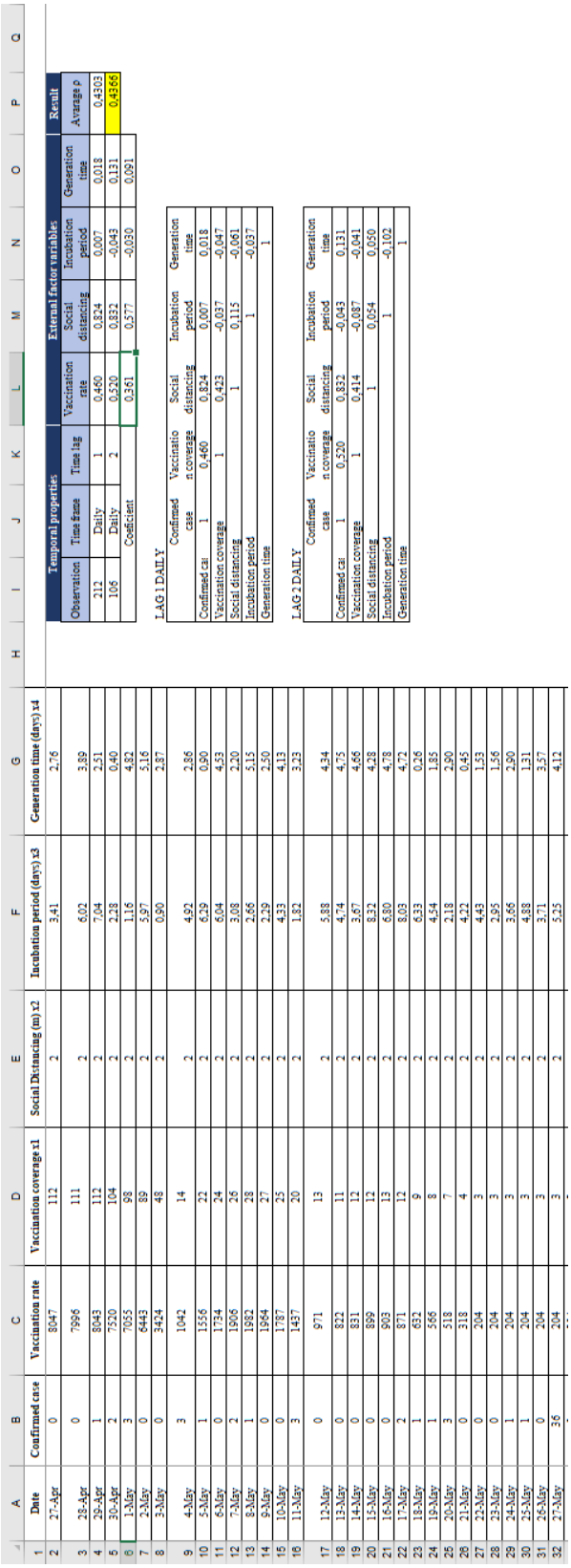
7. After providing forecast results, the study continued with calculating and plotting a 7-day moving average between forecasting results and real data. MSE and RMSE calculation was also included in this part for error comparison in this study:

```
# Calculate the 7-day moving average of both confirmed cases and forecasting results:
# For confirmed cases:
window_size1 = 7
i = 0
moving_avg1 = [] # Provide an empty list to store moving averages
# Loop through the array to consider every window of size 7:
while i < len(WB['Data value']) - window_size1 + 1:
    list1 = list(WB['Data value'])
    window1 = list1[i:i+window_size1] # Store elements from i to i+window_size in list to get the current window
    window_average1 = round(sum(window1) / window_size1, 2) # Calculate the average of current window
    moving_avg1.append(window_average1) # Store the average of current window in moving average list
    i = i+1

# for forecasting results:
window_size2 = 7
j = 0
moving_avg2 = [] # Provide an empty list to store moving averages
# Loop through the array to consider every window of size 7:
while j < len(WB['Final Forecast result']) - window_size2 + 1:
    list2 = list(WB['Final Forecast result'])
    window2 = list2[j:j+window_size2] # Store elements from i to i+window_size in list to get the current window
    window_average2 = round(sum(window2) / window_size2, 2) # Calculate the average of current window
    moving_avg2.append(window_average2) # Store the average of current window in moving average list
    j = j+1

x1 = np.linspace(0,201,201) # Specify the number of observation in 7-day moving average => 201 observations
plt.plot(x1, moving_avg1)
plt.plot(x1, moving_avg2)
plt.legend(["COVID-19 Confirmed Cases","ARIMA(1,1,0) Forecast"])
plt.title('7-day moving average of Covid-19 confirmed cases & forecasting result', alpha=0.5)
plt.show()

# RMSE Calculation:
MSE = np.square(np.subtract(WB['Data value'], WB['Final Forecast result'])).mean()
RMSE = math.sqrt(MSE)
print("Mean Square Error: ", MSE)
print("Root Mean Square Error: ", RMSE)
```

Appendix E. ARIMA model on Excel

1. Pearson's correlations:

Based on vaccination rate, incubation period, generation time of Delta variant, as well as Social distancing according to each control measures proposed, the following figure provides the results of Pearson's coefficients between each external parameters.



Date	Confirmed case	Vaccination rate	Vaccination coverage x1	Social Distancing (m) x2	Incubation period (days) x3	Generation time (days) x4
27-Apr	0	8047	112	2	3,41	2,76
28-Apr	0	7996	111	2	6,02	3,89
29-Apr	1	8043	112	2	7,04	2,51
30-Apr	2	7520	104	2	2,28	0,40
1-May	3	7055	98	2	1,16	4,82
2-May	0	6443	89	2	5,97	5,16
3-May	0	3424	48	2	0,90	2,87
4-May	3	1042	14	2	4,92	2,86
5-May	1	1556	22	2	6,29	0,90
6-May	0	1734	24	2	6,04	4,53
7-May	2	1906	26	2	3,08	2,20
8-May	1	1932	28	2	2,66	5,15
9-May	0	1964	27	2	2,29	2,50
10-May	0	1787	25	2	4,33	4,13
11-May	3	1437	20	2	1,82	3,23
12-May	0	971	13	2	5,88	4,34
13-May	0	822	11	2	4,74	4,75
14-May	0	831	12	2	3,67	4,66
15-May	0	899	12	2	8,32	4,28
16-May	0	903	13	2	6,80	4,78
17-May	2	871	12	2	8,03	4,72
18-May	1	632	9	2	6,33	0,26
19-May	1	566	8	2	4,54	1,85
20-May	3	518	7	2	2,18	2,90
21-May	0	318	4	2	4,22	0,45
22-May	0	204	3	2	4,43	1,53
23-May	0	204	3	2	2,95	1,56
24-May	1	204	3	2	3,66	2,90
25-May	1	204	3	2	4,88	1,31
26-May	0	204	3	2	3,71	3,57
27-May	36	204	3	2	5,25	4,12

	Temporal properties:			External factor variables:			Result
Observation	Time frame	Time lag	Vaccination rate	Social distancing	Incubation period	Generation time	Average ρ
212	Daily	1	0,460	0,824	0,007	0,018	0,4303
106	Daily	2	0,520	0,832	-0,043	0,131	0,4366
	Coefficient		0,361	0,577	-0,030	0,091	

LAG 1 DAILY

	Confirmed case	Vaccinatio n coverage	Social distancing	Incubation period	Generation time
Confirmed cas	1	0,460	0,824	0,007	0,018
Vaccination coverage		1	0,423	-0,037	-0,047
Social distancing			1	0,115	-0,061
Incubation period				1	-0,037
Generation time					1

LAG 2 DAILY

	Confirmed case	Vaccinatio n coverage	Social distancing	Incubation period	Generation time
Confirmed cas	1	0,520	0,832	-0,043	0,131
Vaccination coverage		1	0,414	-0,087	-0,041
Social distancing			1	0,054	0,050
Incubation period				1	-0,102
Generation time					1

2. ARIMA model Forecasting Result:

	A	B	C	D	E	F	G	H	I	J
1	y'(i) = 10.4 + Yi-1 - 0.628(Yi-1 - Yi-2) + 0.361 x1 (i -2) + 0.577 x2 (i -2) - 0.03 x3 (i -2) + 0.091 x4 (i - 2)								3.369	8.131
2		8510	3404							
3	Date	Confirmed case y	Actual case	Vaccination rate x1	Social distancing x2	Incubation period (days) x3	Generation time (days) x4	ARIMA (1, 1, 0) forecast	External factor forecast	Confirmed case forecast
4	27-Apr	0	0	112	2	3.41	2.76			-
5	28-Apr	0	0	111	2	6.02	3.89			-
6	29-Apr	1	4	112	2	7.04	2.51	10	42	52
7	30-Apr	2	8	104	2	2.28	0.40	11	41	52
8	1-May	3	12	98	2	1.16	4.82	12	41	53
9	2-May	0	0	89	2	5.97	5.16	13	39	52
10	3-May	0	0	48	2	0.90	2.87	12	37	49
11	4-May	3	12	14	2	4.92	2.86	10	34	44
12	5-May	1	4	22	2	6.29	0.90	12	19	30
13	6-May	0	0	24	2	6.04	4.53	13	6	19
14	7-May	2	8	26	2	3.08	2.20	11	9	20
15	8-May	1	4	28	2	2.66	5.15	11	10	21
16	9-May	0	0	27	2	2.29	2.50	12	11	23
17	10-May	0	0	25	2	4.33	4.13	11	11	23
18	11-May	3	12	20	2	1.82	3.23	10	11	22
19	12-May	0	0	13	2	5.88	4.34	12	10	22
20	13-May	0	0	11	2	4.74	4.75	12	9	21
21	14-May	0	0	12	2	3.67	4.66	10	6	17
22	15-May	0	0	12	2	8.32	4.28	10	6	16
23	16-May	0	0	13	2	6.80	4.78	10	6	16
24	17-May	2	8	12	2	8.03	4.72	10	6	16
25	18-May	1	4	9	2	6.33	0.26	11	6	17
26	19-May	1	4	8	2	4.54	1.85	12	6	18
27	20-May	3	12	7	2	2.18	2.90	11	4	16
28	21-May	0	0	4	2	4.22	0.45	12	4	16
29	22-May	0	0	3	2	4.43	1.53	12	4	16
30	23-May	0	0	3	2	2.95	1.56	10	3	13
31	24-May	1	4	3	2	3.66	2.90	10	2	13
32	25-May	1	4	3	2	4.88	1.31	11	2	13
33	26-May	0	0	3	2	3.71	3.57	11	2	14
34	27-May	36	144	3	2	5.25	4.12	11	2	13
35	28-May	0	0	3	2	3.98	0.94	24	2	26
36	29-May	0	0	3	2	3.45	4.77	33	2	35
37	30-May	51	204	3	2	3.05	2.20	10	2	13

CONTRIBUTION

The contribution of each member is demonstrated in the following RASIC table:

Responsibility	**Võ Thị Thiện Mỹ**	**Phạm Ngọc Thu Uyên**	**Nguyễn Huỳnh Ngọc Quế**
Literature Review	S	R	S
Data collection	R	R	R
Model development	R	R	R
Data validation	R		R
Model Selection	S	R	
Parameter Estimation		R	R
External parameter implementation			R
Techniques justification	R	R	R
Forecasting & Data fitting	R	R	R
Result Analysis	R		
Discussion	R	S	R
Task distribution	R		
Report Writing	R	R	R
% Contribution	100%	100%	100%