A MACHINE LEARNING APPROACH TO BUILDING A DIGITAL MAP OF COVID-19

Thom Hue Huynh[1] **and Son The Pham**[2]

[1]Faculty of Geography
University of Social Sciences and Humanities – VNUHCM
10-12 Dinh Tien Hoang Street, Ben Nghe Ward, District 1, Ho Chi Minh City, Vietnam
Email: 1856080099@hcmussh.edu.vn

[2]Department of Data Science
Faculty of Information Science and Engineering
University of Information Technology - VNUHCM
Quarter 6, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam
Email: sonpt@uit.edu.vn

ABSTRACT

Coronavirus disease 2019 (COVID-19) has become more and more complicated from the end of 2019 until now. Information about the COVID-19 epidemic situation is a hot spot of interest. A problem is how to update the information about the COVID-19 epidemic quickly and accurately, simply and effectively. It can help managers and people capture information the most quickly. Therefore, we choose GIS applications as a powerful tool capable to solve the proposed problem.

The main objective of the project is to build a digital map application to support observing the COVID-19 epidemic situation in Vietnam. The map has the following features: (1) Automatically update new data about the epidemic situation; (2) Detailly present information about the developing epidemic situation by a digital map; (3) Allow users to search information about the epidemic; (4) Basically predict of the number of patients with days using a method of machine learning. Used dataset of this paper was collected until July 31, 2021.

1. INTRODUCTION

From the end of 2019 until now, the COVID-19 epidemic has spread all over the world, including Vietnam. Governments have promoted the development of many applications to limit the spread of the disease. Besides, there are also many types of research that have proposed GIS application solutions to monitor disease developments. Among them, we are especially interested in GIS applications to support building COVID-19 maps, statistical methods, and forecast models to predict the epidemic situation.

Specifically, GIS-based spatial modeling of COVID-19 (Mollalo, Vahedi, & Rivera, 2020) presented the rate in the United States, using maps to illustrate in the state area. Evaluation of map applications and analysis based on COVID-19 data sources in Europe's population (Pászto, Burian, & Macku, 2020). Reviews on spatial analysis and GIS in the research of COVID-19 highlighted important GIS applications to study COVID-19 (Franch-Pardo, Napoletano, Rosete-Verges, & Billa, 2020). Application of GIS map presented the status of COVID19 cases in Maharashtra state of India (Kodge, 2021). The Prophet package was used to predict covid-19 cases in India (Indhuja & Sindhuja, 2020). In general, the studies focused on forecasting and presenting maps showing the epidemic situation of countries.

In the scope of this research, we build a digital map application to support observing the COVID-19 epidemic situation in Vietnam. The map has the automatically update new data about the epidemic situation and detailly present information about the developing epidemic situation. And focus on analyzing epidemic data in the Ho Chi Minh City area to basically predict the number of patients with days using methods of machine learning. Because now the epidemic in the city is the most increasing. This issue is a concern to the whole society.

2. DATASET DESCRIPTION

2.1. Spatial data

Shapefile data of 63-province Vietnam was collected from the Open Development Mekong (ODM) at website (https://opendevelopmentmekong.net/). Shapefile is a type of vector data able to reference position, measurement unit, and spatial relationship. That means describe phenomenal shapes. The data is used to represent the scope of each administrative region on the map. The spatial reference system (SRS) on the project is UTM Zone 48-N to build the map.

2.2. Attribute data

Attribute data of COVID-19 cases is collected at the website of the Ministry of Health. The data daily updated as follows a total number of new cases of each region (province and city) in Table 1. Besides, we also collect detailed data of Ho Chi Minh City about the daily number of COVID-19 cases. The data is used to solve the forecasting problem that we introduced above. Data samples were collected until July 31, 2021. Nextly, we selected 5 data samples to present as follows:

Table 1. The five samples of cases of each region in Viet Nam

Region	Total
TP HCM	90243
Bình Dương	14679
Long An	5443
Đồng Nai	4126
Khánh Hòa	1710

Table 2. The five samples of cases in Ho Chi Minh City

Date	Total
2021-07-31	4180
2021-07-30	4282
2021-07-29	4592
2021-07-28	4449
2021-07-27	6318

3. PROPOSED METHOD

In this paper, we based on two main library packages: (1) The Arcpy package is used to support building a digital map representing the number of covid cases by areas. The digital map has the ability to automatically update the map state when the attribute data changes over real-time. In this section, the digital map will continuously automatically show the epidemic situation of provinces and cities so that managers can know the epidemic situation across the country. When the epidemic data changes, the map also updates. In addition, the map presents additional statistics on the epidemic situation. (2) The Prophet package supports building a machine learning model to predict the number of covid cases in the near future. In this forecast, we only focus on the Ho Chi Minh City area. Because this area has seriously been affected by the Covid-19 epidemic currently. The number of daily infections is increasing rapidly. Therefore, we are very interested and want to build a forecasting model

for that area. According to exploratory data analysis, we found the dataset to be very suitable for the Prophet forecast model we proposed.

3.1. ArcPy package

ArcPy is a package in Python programming that is able to perform geographic data analysis, map automation, data conversion, and data management (Zandbergen, 2020b, 2020a). ArcPy is a powerful and useful tool that is a product developed by ESRI. In this section, we apply ArcPy to build a digital map presenting an overview of the epidemic situation in Vietnam. The map can update itself when the attribute data changes. During the mapping process, we parallelly combine ArcPy and ArcGIS Pro to optimize the operation process. Some strengths of ArcGIS Pro as follow spatial analysis, image processing, and remote sensing, build maps and displays, real-time GIS data, data collection and management, and 3D GIS made us chosen it to develop the map (Allen, Znamirowski, & Chandler, 2021).

In the ArcPy package, there are many support modules for programmers to develop GIS applications, including the following: Charts module (arcpy.charts), Data Access module (arcpy.da), Geocoding module (arcpy.geocoding), Image Analysis module (arcpy.ia), Mapping module (arcpy.mp), Metadata module (arcpy.metadata), Network Analyst modules (arcpy.nax and arcpy.na), Sharing module (arcpy.sharing), Spatial Analyst module (arcpy.sa), Workflow Manager (Classic) module (arcpy.wmx). The modules covered other areas of ArcGIS. Besides, in the process of building the map, we have designed and built more tools embedded in ArcGIS Pro to automate operations with the following 5 processing tools: Add Data, Add Table, Join Table, Update Data, View Map.

3.2. Prophet package

The Prophet package in Python and R programming language is a machine learning model that is able to forecast many values in the future (Taylor & Letham, 2018). The forecast is based on time-series data collected and preprocessed in advance from the news website of the respiratory disease COVID-19 (Ministry of Health). The predictive model was installed with the 5 stages.

Firstly, preparing full data and follow/match the format of the model. They must be correct for the time series data type. The data of this step is the number of infections per day in Ho Chi Minh City. The city was a very serious Covid-19 epidemic now. Up to now (July 31, 2021), we have collected 96 statistical samples of the number of cases per day. Specifically, the five collected samples recently were presented in Table 2.

Secondly, based on the existing dataset to train the predictive model. As a result of this stage, we get a predictive model organized into an object in the programming language. Then set the number of predicting periods from the trained model. For the predicting periods of future cases, the periods are set up with 5 days. That is mean to predict the number of cases in the next five days. The 5-day period is just enough days to compare the actual values to calibrate and re-train the model. The results of the training model and the forecast results are visualized in Figure 1. In Figure 1, there are two visualization components: (1) visualization for the training data part and (2) visualization for the forecasting part. The forecasting section is marked with a red square in Figure 1. Figure 2 shows the forecast trend of the dataset calculated on a daily and weekly basis. In the weekly forecast trend, there is a negative value, because it is an overall value to represent the model. In the reality, the value will not happen.

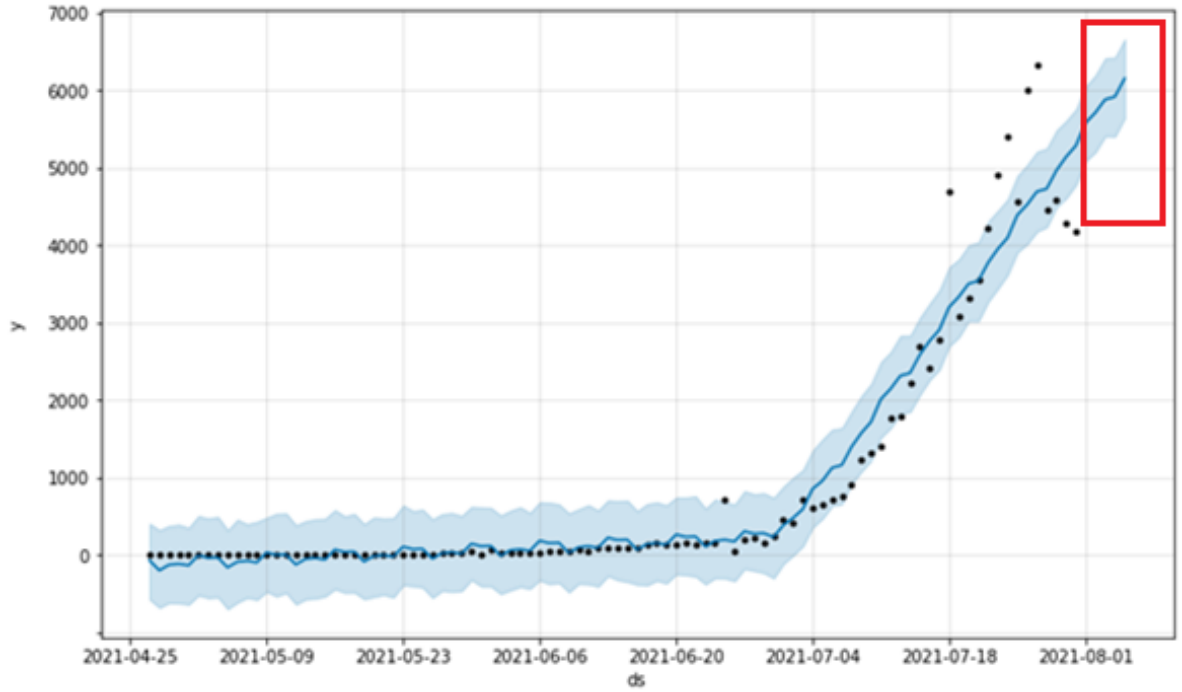

Figure 1. Result of build-in forcasting model.

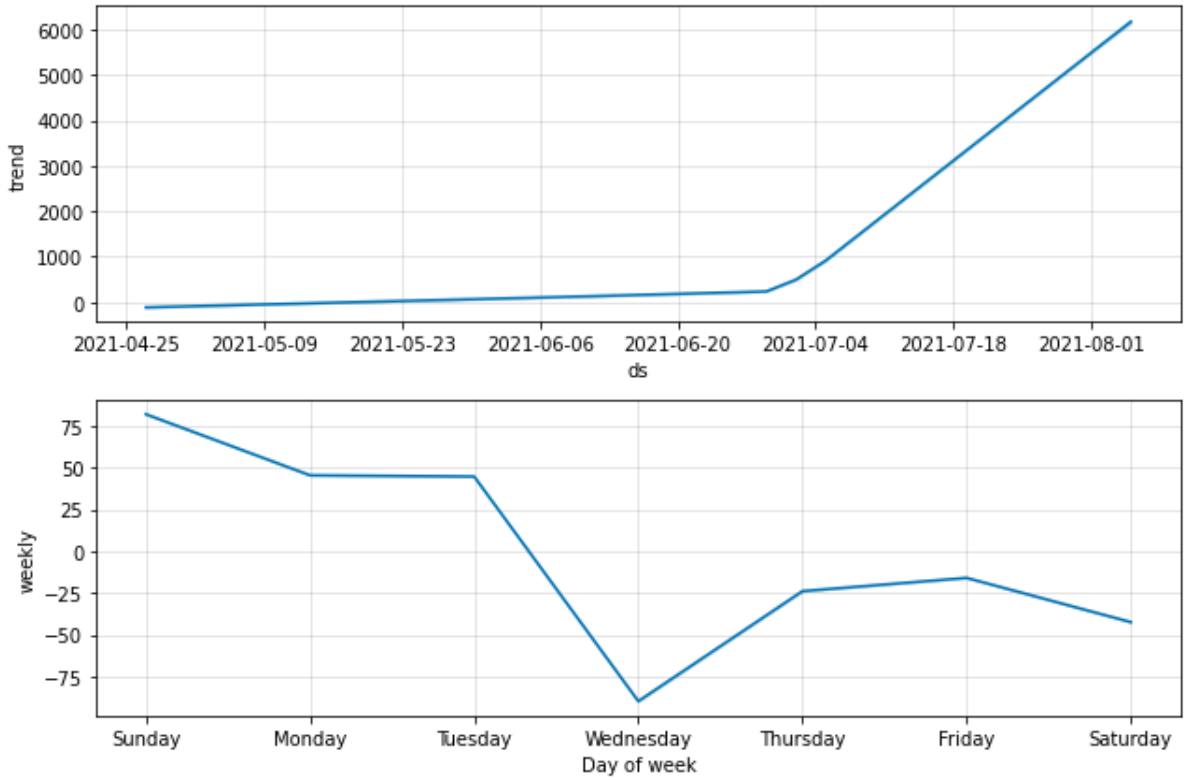

Figure 2. Trend of dataset using build-in forcasting model.

Thirdly, we will make future forecasts. From the obtained training model, we will proceed to get the prediction results. The forecast result will obtain the forecast value (yhat) and the upper and lower error range (called yhat_lower and yhat_upper). The forecast result will be the real number. We will take the integer part to represent the number of COVID-19 cases in Table 3.

Table 3. Results of 5-day prediction in Ho Chi Minh City.

Day	yhat	yhat_lower	yhat_upper
2021-08-01	5575.304875 ~ 5575	5086.763709 ~ 5086	6047.794081 ~ 6047
2021-08-02	5708.581069 ~ 5708	5201.778696 ~ 5201	6195.450087 ~ 6195
2021-08-03	5877.611387 ~ 5877	5406.962430 ~ 5406	6413.376456 ~ 6413
2021-08-04	5913.098525 ~ 5913	5403.379973 ~ 5403	6423.361159 ~ 6423
2021-08-05	6148.652844 ~ 6148	5641.224548 ~ 5641	6658.201478 ~ 6658

Finally, based on actual data, then calibrate and retrain the model. The forecasting model will support predicting the next number of cases. But the model will not indicate when the peaking number of COVID-19 cases, or when the decreasing number of cases, or when there will be a complete decline. So, after the prediction is done and has more data of the new day. The model will be retrained one more so that the next prediction results are likely to be more accurate. Nextly, after retraining, there are 5 new forecasted results from the newly trained model.

4. RESULTS

Map of the situation of the COVID-19 epidemic in Vietnam by the method of representing the quality background. The map shows the epidemic situation distributed over 63 provinces in Vietnam with main information as follows: total number of cases and results of 5-day prediction in Ho Chi Minh City. The overview interface of the results is shown in Figure 3.

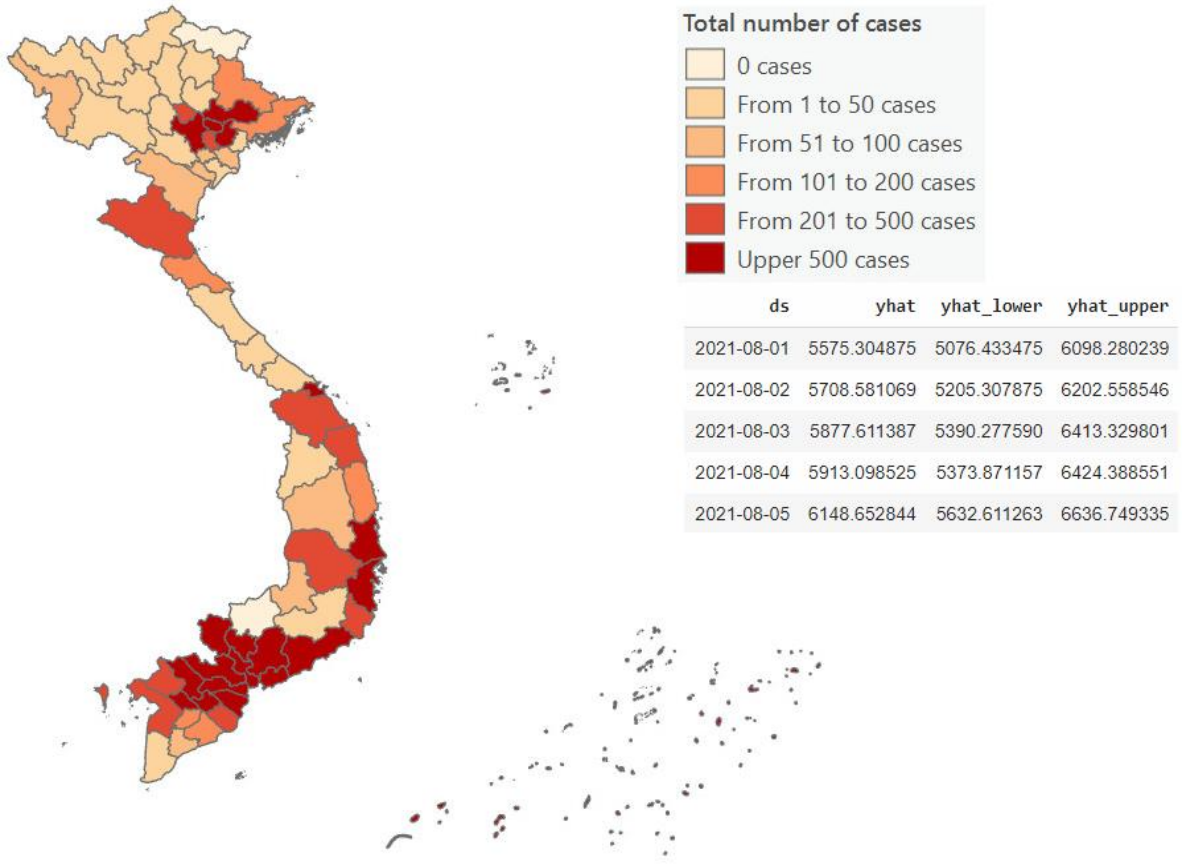

Figure 3. Overall results.

Figure 3 presents the results of the study. The left part is an automated map showing the COVID-19 epidemic situation of regions of Vietnam. The right part presents the scales showing the number of cases in the regions and the forecasted results of the total number of cases at Ho Chi Minh City in the next 5 days.

5. CONCLUSIONS

This paper presented a method for building an automatic digital map to visualize the COVID-19 epidemic situation of administrative regions in Vietnam and a forecast analysis of the number of cases in Ho Chi Minh City. The implementation method is based on two main approaches ArcPy package and Prophet package on Python programming. Firstly, after completing the application, the combination of both ArcGIS Pro and ArcPy for building an automatic digital map not only saves time and effort but also has high accuracy. That has helped us to complete an automated digital map of the covid-19 epidemic extremely smoothly. Secondly, the prediction results of the Prophet machine learning model are also acceptable. This forecast will help managers plan to respond to the growing number of patients. The downside of the Prophet model is that it only correctly predicts when the trend is up or down, does not predict when the number of cases will peak or when the epidemic is likely to end. The results are well applicable, taking advantage of the advances and new features of the software, contributing to improving the position of maps and geographic information systems in many fields.

In future work, we will conduct more automatic data collection from COVID-19 news websites of the Ministry of Health. That means we will build a website scraper and crawler that will automatically get the data when the source site changes. Then the map can perform real-time data updates.

6. REFERENCES

Allen, D. W., Znamirowski, B., & Chandler, M. (2021). Focus on Geodatabases in ArcGIS Pro. *Photogrammetric Engineering & Remote Sensing*, *87*(7), 468–469.

Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of The Total Environment*, *739*, 140033.

Indhuja, M., & Sindhuja, P. P. (2020). Prediction of covid-19 cases in India using prophet. *International Journal of Statistics and Applied Mathematics*, *5*(4), 103–106.

Kodge, B. G. (2021). A review on current status of COVID19 cases in Maharashtra state of India using GIS: a case study. *Spatial Information Research*, *29*(2), 223–229.

Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment*, *728*, 138884.

Pászto, V., Burian, J., & Macku, K. (2020). COVID-19 data sources: evaluation of map applications and analysis of behaviour changes in Europe's population. *Geografie (Utrecht)*, *125*(2), 171–209.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45.

Zandbergen, P. A. (2020a). *Advanced Python Scripting for ArcGIS Pro*. Esri Press.

Zandbergen, P. A. (2020b). *Python Scripting for ArcGIS Pro*. Esri Press.