Statistical Model for Predicting the COVID-19 Pandemic in South and Southeast Asian Regions

Ameen Mhamad

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Research Methodology
Prince of Songkla University
2022

Statistical Model for Predicting the COVID-19 Pandemic in South and Southeast Asian Regions

Ameen Mhamad

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Research Methodology

Prince of Songkla University

2022

Thesis Title	Statistical Model for Predicting the COVID-19 Pandemic in South and Southeast Asian Regions
Author	Mr.Ameen Mhamad
Major Program	Research Methodology

Major Advisor	**Examining Committee :**
....	....Chairperson
(Dr. Nurin Dureh)	(Asst. Prof. Arinda Ma-a-lee)
	....Committee
Co-advisor	(Dr. Nurin Dureh)
....	....Committee
(Asst. Prof. Dr. Areena Hazanee)	(Asst. Prof. Dr. Areena Hazanee)
	....Committee
	(Asst. Prof. Dr. Wandee Wanishsakpong)

The Graduate School, Prince of Songkla University, has approved this thesis as partial fulfillment of the requirements for the Master of Science Degree in Research Methodology

....

(Prof. Dr. Damrongsak Faroongsarng)

Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate’s own investigations. Due acknowledgement has been made of any assistance received.

....Signature

(Dr. Nurin Dureh)

Major Advisor

....Signature

(Mr. Ameen Mhamad)

Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

....Signature

(Mr. Ameen Mhamad)
Candidate

ชื่อวิทยานิพนธ์	ตัวแบบทางสถิติสำหรับการทำนายการระบาดโรคติดเชื้อไวรัสโคโรนา 2019 ในภูมิภาคเอเชียใต้และเอเชียตะวันออกเฉียงใต้
ผู้เขียน	นายอมีน มะหมัด
สาขาวิชา	วิธีวิทยาการวิจัย
ปีการศึกษา	2564

บทคัดย่อ

การระบาดของโรคโควิด-19 (COVID-19) ที่เกิดจากเชื้อไวรัสสายพันธ์ใหม่ Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) ถูกยกระดับให้กลายเป็นภาวะฉุกเฉินทางสาธารณะสุขระหว่างประเทศ ในระยะเวลาเพียงหนึ่งปีนับตั้งแต่วันที่ 24 มกราคม 2564 ภายหลังจากที่องค์การอนามัยโลกได้มีการประกาศให้โรคโควิด-19 เป็นโรคระบาด พบว่ามีการแพร่กระจายอย่างรวดเร็วของเชื้อโรคที่ขยายวงกว้างไปทั่วโลก ประชากรโลกกว่า 150 ล้านคนตรวจพบการติดเชื้อ และกว่า 3.23 ล้านคนเสียชีวิตจากการติดเชื้อไวรัสโคโรน่าสายพันธ์ใหม่ SARS-CoV-2 และในปัจจุบันหลายภูมิภาคทั่วโลกยังคงต้องพบกับวิกฤติการแพร่ระบาดของโรคโควิด-19 โดยเฉพาะในภูมิภาคเอเชียและเอเชียตะวันออกเฉียงใต้ การศึกษาในครั้งนี้มีวัตถุประสงค์เพื่อประยุกต์ใช้ตัวแบบ Natural cubic spline ในการทำนายจำนวนผู้ติดเชื้อไวรัสโควิด-19 ซึ่งใช้ฐานข้อมูลรายวันจาก ศูนย์ข้อมูลไวรัสโคโรนาแห่งมหาวิทยาลัยจอนด์ฮอพกินส์ ประเทศสหรัฐอเมริกา (Johns Hopkins University coronavirus resource center) ผลการศึกษาพบว่าตัวแบบสามารถทำนายข้อมูลการแพร่ระบาดในประเทศต่างๆได้แก่ ประเทศอินเดีย ประเทศปากีสถาน ประเทศบังคลาเทศ ประเทศเนปาล ประเทศศรีลังกา ประเทศฟิลิปินส์ ประเทศมาเลเซีย ประเทศไทย ประเทศเวียดนาม และประเทศออสเตรเลีย ได้เป็นอย่างดี ซึ่งให้ค่า r-squared ที่ 0.997, 0.981, 0.992, 0.975, 0.995,

0.957, 0.973, 0.939, 0.989, 0.881, 0.943 และ 0.610 ตามลำดับ โดยมีค่าเฉลี่ยของ r-sqared เท่ากับ 0.936 ทั้งนี้มีการใช้ค่า รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เพื่อทดสอบประสิทธิภาพในการทำนายของตัวแบบ โดยเปรียบเทียบค่า RMSE ที่ได้จากตัวแบบโดยใช้ข้อมูลฝึกหัด (Tranining data) และตัวแบบที่ได้จากชุดข้อมูลทดสอบ (Testing data) พบว่าค่า RMSE ของตัวแบบระหว่างข้อมูลฝึกหัดและข้อมูลทดสอบไม่มีความแตกต่างกันมากนัก จากผลการศึกษานี้ จึงอาจจะสรุปได้ว่าตัวแบบมีความเหมาะสมกับชุดข้อมูลการติดเชื้อโควิด-19 ในภูมิภาคเอเชียและเอเชียตะวันออกเฉียงใต้ นอกจากนี้ตัวแบบยังสามารถพยากรณ์จำนวนผู้ติดเชื้อโควิด-19 ที่เปลี่ยนแปลงไปในแต่ละสัปดาห์ (Increasing cases per week) ซึ่งสามารถติดตามและประเมินความรุนแรงของโรคเพื่อออกแบบมาตรการป้องกันได้ทันการณ์ โดยสรุปแล้วตัวแบบ Natural cubic spline สามารถนำไปประยุกต์ใช้กับข้อมูลโควิด-19 ได้ทั่วภูมิภาคทั่วโลก โดยสามารถเพิ่มตัวแปรทำนาย อาทิเช่น ตัวแปรทางด้านสิ่งแวดล้อม ตัวประทางด้านสังคม และประชากรศาสตร์ เพื่อเพิ่มความครอบคลุมและความแม่นยำในการทำนายการแพร่การจายของโรคติดเชื้อไวรัสโควิด-19

Thesis Title	Statistical Model for Predicting the COVID-19 Pandemic in South and Southeast Asian Regions
Author	Mr.Ameen Mhamad
Major Program	Research Methodology
Academic Year	2021

Abstract

The disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been put in the list of public health emergency of international concern (PHEIC) according to the rapid spreading and expantion of the diseases around the world, later named the coronavirus disease 2019 (COVID-19). With in a year, more than 150 million people were identified as SARS-CoV-2 infected cases, and over 3.23 million deaths were reported from COVID-19 weekly update, 24st January, 2021, since the pandemic announcement from the World Health Organization (WHO). Nowadays, many regions world wide have been facing the crises due to the pandemic of COVID-19, also South and Southeast Asian regions. This study aims to construct a model for predicting COVID-19 using natural cubic spline function with equi space knot. The data used in this study were obtained from publicly available databases form Johns Hopkins University coronavirus resource center updated daily and located at GitHub run by Microsoft. The results found that the model fits the data extremely well, defined as that maximizes the r-squared value in India, Pakistan, Bangladesh, Nepal, Sri-Lanka, Philippines, Malaysia, Thailand, Vietnam, and Australia were 0.997, 0.981, 0.992, 0.975, 0.995, 0.957, 0.973, 0.939, 0.989, 0.881, 0.943 and 0.610, respectively. Moreover, model provided mean r-squared 0.936. To access the performance of the natural cubic spline model, apart from using the r-squared values, we compared the RMSE for training and testing data set. We found that the RMSE between these models was not much different. This might be an evidence that the models were not over-fitting. Moreover, the model provides forecasts of daily changes, which signaled when action is needed. Moreover, this model is routinely applicable to all such regions in the world and can be extended to accommodate additional predictors such as environmental and demographic variables. Conclusion, this model is routinely applied to all such regions in the world and can be extended to accommodate additional predictors such as environmental and demographic variables.

Acknowledgements

Bismillahirrahmanirrahim, in the name of Allah (Subhanuahu wa ta' ala), the most merciful, the beneficent. Sholawat and Salam send to his big messenger prophet Muhammad Salla Allaahu 'alayhi wa salaam (peace be upon him) for being the best role model for all humankind.

First of all, I would like to express my sincere gratitude to my advisor Dr.Nurin Dureh and my co-advisor Asst.Prof.Dr Areena Hasanee for all of the support of my master study and related research, for patience, motivation, and immense knowledge. Also Prof. Don McNeil for supporting. For all guidance helped me in all the time of research and writing of this thesis.

Besides my advisor and co-advisor, I would like to thank the rest of my thesis committee'chair Dr.Arinda Ma-a-lee, and additional external committee from Department of statistics, Faculty of science, Kasetsart University Asst. Prof. Dr.Wandee Wanishsakpong, for their in rightful comments and encouragement, but also for their hard questions which incented me to widen my research from various perspectives.

My sincere thanks also go to Assoc.Prof Winai Dahlan, Director of the halal science center Chulalongkorn University, Assist. Prof.NikFarid Radenahmad deputy director of the halal science center Chulalongkorn University, and all my collauge from the halal science center Chulalongkorn University, who are always at my side in any circumstances. Without their moral support it would not be possible to conduct this research.

Last but not the least, I would like to thank my family, my parents and , brothers, and sister for supporting me spiritually throughout writing this thesis and my life in general. This accomplishment would not have been possible without them. Thank you.

Ameen Mhamad

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

This study aims to identify the patterns and trends of the COVID-19 pandemic and predict the COVID-19 confirmed cases in South and Southeast Asian countries. This chapter explains the background and rationale for this study and ends with definitions of the term. The topics will be discussed as follows

1.1 Background and rationale
1.2 Objectives of research
1.3 Expected advantages
1.4 Scope of the research
1.5 Literature reviews

1.1 Background and rationale

The disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been put in to the list of public health emergency of international concern (PHEIC) by World Health Organization (WHO) according to the rapid transmission and expantion of the diseases around the world, later named the coronavirus disease 2019 (COVID-19). Since the pandemic was announced, more than 150 people were identified as SARS-CoV-2 infected cases, and over 3.23 million deaths were reported from COVID-19 weekly update, 24st January, 2021, reported by WHO. The challenge for public health is to confront the emerging and reemerging disease to protect the population.

The South and Southeast Asian (SSEA) region is one of the regions which suffer from the pandemic of COVID-19. There are seven countries in SSEA bordering China, including Bhutan, India, Laos, Myanmar (Burma), Nepal, Pakistan, and Vietnam. SSEA region also reported the first COVID-19 cases outside of China was founded in Thailand on 13th January 2020. Furthermore, the countries in SSEA are in a state of widespread economic and social despair, as seen by high rates of poverty and poor health facilities. As a result, it has posed significant issues to public health organizations in terms of quickly implementing pandemic-prevention measures.

COVID-19 has advocated for a variety of measures from the government and public sectors to combat the epidemic, including school closures, public gathering bans, travel restrictions, public transportation shutdowns, and international movement restrictions (Hale *et al.,* 2020). COVID-19 has had a significant impact on human life, as well as the global community and economic development. As a result, a long-term and dependable strategy is required to stop the virus from spreading. The decision on the best moment for the disease to vanish, on the other hand, is filled with uncertainties.

Despite the availability of effective vaccines or treatments, mathematical modeling studies and unambiguous data suggest the need for public health measures. The COVID-19 pandemic is relevant even in the tiniest hint for short-term forecasting, allowing better social, economic, cultural, and public health issues to be managed in the future month. Many studies used mathematical and statistical approaches to deal with the COVID-19 pandemic, including time series models (Ilie *et al.,* 2020; Iftikhar *et al.,* 2020; Yousaf *et al.,* 2020), deep learning methods (Zeroual *et al.,* 2020; Arora *et al.,* 2020; Wieczorek *et al.,* 2020), and machine learning methods (Kavadi *et al.,* 2020). However, analyzing and interpreting the outcomes of those methods appears to be difficult.

Therefore, this study aimed to determine the pattern and trend of the COVID19 pandemic and predict COVID-19 confirmed cases in South and Southeast Asian countries using a simple statistical model.

1.2 Objectives of the research

1.2.1 To identify the patterns and trends of COVID-19 pandemic confirmed cases, recoveries cases, and death cases in South and Southeast Asian countries.

1.2.2 To predict the COVID-19 confirmed cases in South and Southeast Asian countries.

1.3 Expected advantages

1.3.1 Provide a method for implementing a large dataset that researchers can utilize to perform statistical modeling.

1.3.2 Provide the government with useful information on the pattern and trend of the COVID-19 pandemic so that it can determine when it is safe to relax social distance measures and preparation for the pandemic.

1.4 Scope of the research

This study focuses on 12 countries in South and Southeast Asia with varying levels of the epidemic, based on the dataset's completeness. India, Pakistan, Bangladesh, Nepal, Sri Lanka, and the Philippines, Malaysia, Thailand, Vietnam, and Australia are among the Southeast Asian countries. Data was collected during a 365-day period, from January 22nd, 2020 to January 21st, 2021, in the form of the total number of confirmed, recovered, and fatalities cases.

1.5 Literature reviews

This subsection provides information related to the topic of this study. It comprises two parts; the first part contains the information about the covid-19 disease, and the second part mentions the related studies that applied the statistical modeling.

1.5.1 COVID-19 disease

Coronaviruses (CoV) are a large family of viruses that cause illnesses ranging from the common cold to severe infections. They primarily affect the respiratory system, but they can also affect other functional organ systems such as the central nervous and neurological systems, as well as the hepatic and gastrointestinal (GI) systems (Ahmed *et al.,* 2020; Aleem *et al.,* 2022). In 2003, a form of coronavirus known as severe acute respiratory syndrome-associated coronavirus (SARS-CoV) was spread from animal to animal for the first time in two decades. Following that, in 2012, a new coronavirus called Middle East Respiratory Syndrome-associated coronavirus (MERS-CoV) was discovered; the majority of cases were reported from the Middle East, particularly Saudi Arabia (Petrosillo *et al.,* 2020). Many cases of respiratory sickness and abrupt respiratory failure were recorded in mainland China late in 2019.

Furthermore, those who went to a seafood market in Wuhan, China, were contacted in the majority of reported cases. The 2019-n-CoV coronavirus was initially discovered in Wuhan and was given the name 2019-n-CoV. As a result, the "coronavirus disease 2019" was shortened to "COVID-19" on February 11, 2020.

The SAR-CoV-2 infection has resulted in a variety of symptoms, from asymptomatic to severe sickness and death. Fever, dry cough, shortness of breath, tiredness, nausea, vomiting, and diarrhea are frequent symptoms in persons hospitalized with COVID-19 infection. In hospitalized COVID-19 patients, complications such as pneumonia, acute respiratory distress syndrome, acute liver injury, cardiac injury, acute heart failure, acute kidney injury, and shock have been linked to an elevated mortality rate. Approximately 5% of COVID-19 patients and 20% of COVID-19 hospitalized patients required ventilatory assistance.

Viruses are constantly evolving, and the genome of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is fast evolving, as is the virus inside each infected person. As a result of natural selection, random genetic drift, and greater transmissibility, each infected individual's genetic heterogeneity will rise. By-products of viral replication include mutations, which are changes in the genetic material RNA of a virus and changes in the genomic sequence. Variants are genomic sequences that differ from one another. This will enhance the disease's transmissibility, severity, and vaccine resistance, as well as make it more difficult to control. These findings will alter the disease's severity, and we must address this by developing a mathematical model to forecast the COVID-19 pandemic.

The COVID-19 is a fully potential distributed human-to-human system. According to the Centers for Disease Control and Prevention, the virus spreads throughout the environment into objects, air, and contacting surfaces, and is transmitted by touching viral particles (CDC). COVID-19 has a 14-days incubation period and can spread to others during that time. This is what makes COVID-19 such a serious threat. The public health sector worker agencies have recognized COVID-19 patients as early as possible to measure the severity of the sickness, preventing the disease from spreading and reducing morbidity and mortality. The COVID-19 transmissibility was shown to be influenced by the country's limited space and high population density.

A hospital must have enough staff and equipment to function properly. To treat COVID-19 patients, an isolation room was required. For healthcare workers, personal protective equipment (PPE) is required. Insufficient medical intervention resources in relation to population size. The rapid rise of new COVID-19 cases in a short period of time can have negative effects on individuals and the healthcare system. Then, out of nowhere, a measurement to manage COVID-19 patients was published, which is time-consuming for the healthcare system and workers. The government and public agencies pushed a variety of approaches to combating the COVID-19 outbreak. Schools, academies, and institutions were shuttered; public gatherings and socializing were prohibited; and travel was restricted. The government's impact on socioeconomic development is measured in all of these ways.

The COVID-19 healthcare system has an impact on a variety of factors, particularly in countries where the majority of people live in poverty. They were afflicted by the major treat associated with the transmission of disease. To manage the epidemic, herd immunity and establishing lockdown or restricting population movement were essential. South Asia, Southeast Asia, and Australia have been affected by the COVID-19 pandemic. While the number of confirmed cases fluctuates at comparable rates, several countries' populations are fast growing, despite the fact that their governments have already implemented political laws to combat the COVID-19 crisis (Sarkar *et al.,* 2020). It's critical to forecast the number of COVID-19 confirmed cases in South and Southeast Asian countries so that public awareness may be raised and appropriate disease control measures can be put in place.

1.5.2 Statistical modeling

Statistical modeling is a method of generating and testing hypotheses and statistical assumptions in order to offer predictions for real-world data, as well as a simple explanation and description of the data behavior. Statistical modeling has been utilized for causal explanation in a variety of areas, with the assumption that models have good predictive and explanatory ability (Shmueli *et al.,* 2010). Several modeling strategies have been employed to analyze data since the COVID-19 disease arose in China at the end of 2019. The compartments with the population-based model, deep learning methods and machine-learning models, and the time series model were the most commonly employed.

The compartment with population-based models is assumed to have a homogeneous sub-population with the same characteristics as the entities being modeled, such as individuals or patients (Porgo *et al.,* 2019). Statistical models, in contrast to compartmental models, often model one state (one compartment at a time) and do not often consider the flow of individuals between states. Furthermore, the population element aids in the resolution of the problem. The population-based model divides the population into several compartments, such as different health conditions (e.g., Susceptible, Exposed, Infected, Quarantined, Recovered, and death). Many research used the traditional SEIR model, but the accepted and enhanced SEIR model delivers better measures than the old SEIR approach. Deep learning algorithms such as the simple recurrent neural network (RNN), long short-term memory (LSTM), bidirectional LSTM (BiLSTM), gated recurrent units (GRUs), and variational autoencoder (VAE) have been used to predict the number of COVID-19 cases around the world (Zeroual *et al.,* 2020; Arora *et al.,* 2020; Wieczorek *et al.,* 2020). Machine learning methodologies such as a progressive partial derivative linear regression (PPDLR) and a nonlinear global pandemic machine learning (NGPML) model were utilized to provide reliable modeling predictions (Kavadi *et al.,* 2020). In addition, time series models such as ARIMA, VAR, and exponential smoothing regression models were used in various research. Furthermore, the most prevalent approaches are linear and polynomial regression, which examine the relationship between models and variables (Progo *et al.,* 2019; Chauhan *et al.,* 2020; yang *et al.,* 2020).

Aside from these, COVID-19 data was modeled using parametric distribution fitting methods, exponential decay models, least-square error models, AI-based models, the Bayesian approach, and hybrid models. To predict the number of COVID-19 cases and study-related outcomes, the researchers devised or used existing mathematical and statistical approaches. COVID-19 modeling has been the subject of numerous studies in various aspects. Some investigations are conducted with the goal of making predictions. Some research focuses on COVID-19 infection estimates and epidemiological forecasting, such as mortality rates. For example, the Bayesian dynamic linear models were used to analyze the future patterns of the COVID-19 pandemic (Feroze, 2020). The time series with an autoregressive integrated moving average (ARIMA) model was used to estimate the prevalence trend in Pakistan and Central European countries (Ilie *et al.,* 2020; Iftikhar *et al.,* 2020; Yousaf *et al.,* 2020).

Various modeling strategies have been employed by a number of researchers to forecast COVID-19 cases in various countries. Short-term forecasting models such as the time series model employed in the autoregressive integrated moving average (ARIMA) model and Holt's exponential smoothing model were utilized in the COVID-19 epidemic in India, for example (Sharma *et al.,* 2020). The COIVD-19 pandemic in Italy (Hao *et al.,* 2020) and the COVID-19 pandemic in France (Hao *et al.,* 2020) were predicted using the basic mean-field model and the susceptible-infected-recovered-death model, as well as the Gompertz model, the logistic model, and the Bertalanffy model (Fanelli *et al.,* 2020). Furthermore, the logistic growth model has been utilized all over the world, particularly in 29 Chinese regions where the pandemic originated, to model the outbreak of COVID-19 and forecast its global expansion (Wu *et al.,* 2020). Similarly, population growth and the ARIMA model were used to predict COVID-19's final size and spread in Italy (Perone *et al.,* 2020). It's also used to predict the total number of confirmed COVID-19 cases in Thailand, Italy, South Korea, Iran, and Mainland China (Dehesh *et al.,* 2020). Policymakers can use a reliable forecast for infectious disease pandemics to make decisions about how to use the country's resources.

Various mathematical models have recently been employed to predict the outbreak of COVID-19 in the countries of South and Southeast Asia. To predict the COVID-19 pandemic in India, they applied the exponential logistic, Gompers growth, and autoregressive integrated moving average (ARIMA) models (Mangla *et al.,* 2021). The akaike information criterion for ARIMA model selection and the mean absolute percentage error and root mean square error comparative were used to assess the growth models' goodness-of-fit. They found that the ARIMA model is the best-fitting model for COVID-19 cases in India and its most impacted states based on available data. The statistical modeling was given information collected from the data, such as estimating and projecting the number of cases, which is required for calculating daily outbreak levels and giving meaningful evidence to support public health measures and COVID-19 infection estimates. It can assist policymakers in managing the direction of measurement in order to deal with infections and allocating the resources needed to keep the virus from spreading. The statistical models outlined above, on the other hand, are difficult to assess and generalize in many places.

Apart from those approaches, we suggest an alternative method that is more accessible and can be applied to any data or region. Natural cubic spline functions with equi-spaced knots were relatively smooth curves that fit the underlying smooth functions automatically (McNeil *et al.,* 2011). With finite second derivatives fitted to a time series defined on a time axis range, these functions have a desirable optimality attribute. Over this range, they minimize any specified linear combination of the data's squared error and the integral of the squared second function's derivative (Sousa *et al.,* 2020). Because the natural cubic spline function has advantages in short-term forecasting of the COVID-19 pandemic, it is critical to develop mathematical models and predictions to determine the degree of the daily outbreak and assist policymakers in managing the disease. A suitable technique predicts the COVID-19 spread and allocates the resources needed to stop the virus from spreading.

Chapter 2

Research Methodology

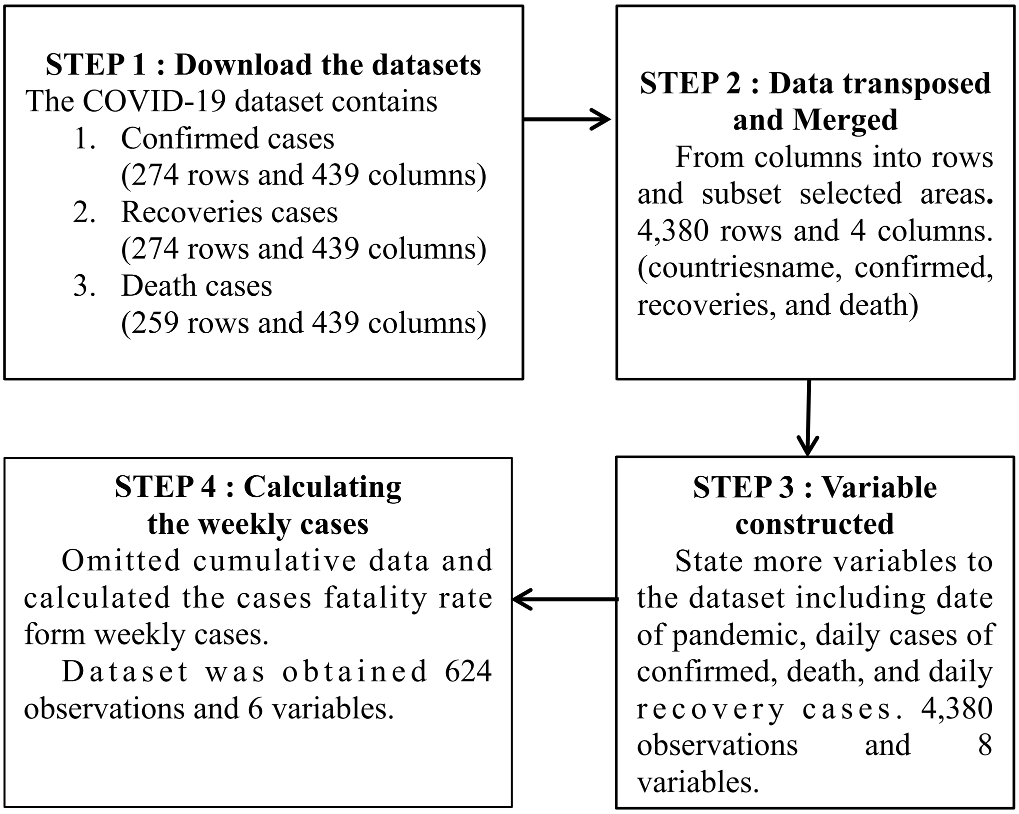
This chapter provided the research methodology used to identify the patterns and trends of the COVID-19 pandemic, the outcomes variables was used consists of confirmed cases, recoveries cases, and death cases. Moreover, the COVID-19 predicting model was performed by using confirmed cases. There are five parts, including data source, data management, study area, path diagram, and data analysis, which were described as follows.

2.1 Data source

The data used in this study derived from databases maintained by the Johns Hopkins University coronavirus resource center and hosted on GitHub by Microsoft. The cumulative number of confirmed cases, total cases, and total deaths are all included in the dataset. The data was collected between January 22nd, 2020 and January 21st, 2021.

2.2 Data management

In this research, data management is an essential procedure. This section aims to clean raw data by removing variables that aren't relevant to the research. Figure 1 illustrates the four phases involved in data management. The data management process began with the download of the COVID-19 dataset from GitHub (https://github.com/)-restored online databases, which was divided into three datasets: confirmed, recoveries, and death cases. Then choose from 12 countries in South and Southeast Asia as your region of interest. The dataset was then transposed and combined, with data from columns into rows. The variable was also created using the dataset, which included the pandemic's start date, daily confirmed, death, and recovery cases. Finally, the total number of confirmed deaths and recoveries was excluded. The case fatality rate was computed in the table after the data was gathered into a 7-day bin or weekly data collection. The full data set was ready for data analysis, which yielded 624 observations and six variables: country names, pandemic date, weekly confirmed cases, weekly deaths, weekly recoveries, and case fatality rate. The R program was used to process all of the data management procedures outlined above.



STEP 1 : Download the datasets
The COVID-19 dataset contains
1. Confirmed cases
   (274 rows and 439 columns)
2. Recoveries cases
   (274 rows and 439 columns)
3. Death cases
   (259 rows and 439 columns)

STEP 2 : Data transposed and Merged
From columns into rows and subset selected areas**.** 4,380 rows and 4 columns. (countriesname, confirmed, recoveries, and death)

STEP 3 : Variable constructed
State more variables to the dataset including date of pandemic, daily cases of confirmed, death, and daily recovery cases. 4,380 observations and 8 variables.

STEP 4 : Calculating the weekly cases
Omitted cumulative data and calculated the cases fatality rate form weekly cases.
Dataset was obtained 624 observations and 6 variables.

Figure 1 Data management diagram

Figure 1 shows the diagram of data management, which includes four steps. Starting from downloading datasets from available online databases, transposed and merged data, constructing the interested variable and calculative the weekly cases and calculating another variable, and omitted the cumulative data which are not interested in this study,which describe as following paragraph.

2.2.1 STEP 1 Download dataset from available online databases

Cumulative confirmed cases, recoveries cases, and death cases were downloaded separately, comprised 274 rows and 439 columns of confirmed and death cases and 259 rows and 439 columns of recoveries cases. Each dataset comprises five variables, including countries names, state names, latitude, longitude, and day of observation. Table 2.1 was shown the original data structure of COVID-19 confirmed cases. The countries namaes variable represent the name of the countries which record the number of COVID-19 confirmed cases. Some countries, such as Australia, divide the number of records by state, such as the Australia Capital Territory, New South Wales, and Northern Territory. The United Kingdom displayed the number of confirmed cases in the subregion in Turks & Caicos Islands. The number of COVID-19 confirmed in the other countries, as stated in table 2.1, was documented in full countries, including Afghanistan, Albania, Palestine (West Bank and Gaza), Yemen, Zambia, and Zimbabwe.

Table 2.1 The original data structure

No.	State names	Countries names	Latitude	Longitude	1/22/20	1/23/20	1/24/20	. . .	01/20/21	01/21/21
1	-	Afghanistan	33.939	67.710	0	0	0	. . .	2354	2363
2	-	Albania	41.153	20.168	0	0	0	. . .	1291	1296
3	Australian Capital Territory	Australia	-35.474	149.012	0	0	0	. . .	54	54
4	New South Wales	Australia	-33.869	151.209	0	0	0	. . .	0	0
5	Northern Territory	Australia	-12.463	130.846	0	0	0	. . .	6	6
.	.	.	.	.	.	.	.	. . .	.	.
.	.	.	.	.	.	.	.	. . .	.	.
271	-	Palestine (West Bank and Gaza)	31.952	35.233	0	0	0	. . .	1751	1757
272	-	Yemen	15.553	48.516	0	0	0	. . .	612	614
273	-	Zambia	-13.134	27.849	0	0	0	. . .	585	597
274	-	Zimbabwe	-19.015	29.155	0	0	0	. .	879	917

2.2.2 STEP 2 Data transposed and merged

The data is subsetted in specific areas in the second step of data management. The data was chosen based on the completeness of the COVID-19 recorded in South and Southeast Asia by the interested area. India, Nepal, Bangladesh, the Philippines, Indonesia, Singapore, Nepal, Malaysia, Thailand, Sri Lanka, Vietnam, and Australia are among the 12 countries represented. The variables of state names, latitude, and longitude were omitted.

The confirmed, deaths, and recovery cases datasets were then converted from columns to rows, resulting in a wide to long data table. Each data structure was transformed from 12 rows and 365 columns to 4380 rows and 2 columns (365 columns represent the number of recorded cases). Following that, the data was combined. The integrated dataset ended up with four variables: country names, confirmed cases, recovered cases, and fatalities cases. As illustrated in table 2.2, the dataset in this step will include 4,380 rows and 4 columns.

Table 2.2 The restructured data for the selected areas

No.	Countries Name	Confirmed cases	Deaths cases	Recoveries cases
1	Australia	0	0	0
2	Australia	0	0	0
.	.	.	.	.
370	Bangladesh	0	0	0
371	Bangladesh	0	0	0
.	.	.	.	.
1009	Malaysia	0	0	0
1010	Malaysia	0	0	0
.	.	.	.	.
.	.	.	.	.

Table 2.2 The restructured data for the selected areas

No.	Countries Name	Confirmed cases	Deaths cases	Recoveries cases
4374	Vietnam	1531	35	1369
4375	Vietnam	1536	35	1380
4376	Vietnam	1537	35	1380
4377	Vietnam	1537	35	1380
4378	Vietnam	1539	35	1402
4379	Vietnam	1540	35	1402
4380	Vietnam	1544	35	1406

2.2.3 **STEP 3** State more interesting variables into the dataset

To prepare the dataset for analysis, the daily number of cases was obtained by extracting the number of cumulative cases from all variables, including confirmed cases, death cases, and recovery cases. The day of the pandemic was created as the other variable. The number of days will begin on January 22, 2020, and end on January 21, 2021, for a total of 365 days, marking the anniversary of the pandemic in South and Southeast Asia. The dataset now contains 4,380 observations and nine variables, as shown in table 2.3

Table 2.3 The dataset structured by generated new variables

No.	Countries Name	Date of Pandemic	Confirmed cases	Death cases	Recoveries cases	Daily confirmed cases	Daily death cases	Daily recoveries cases
1	Australia	1	0	0	0	0	0	0
2	Australia	2	0	0	0	0	0	0
.	.	.	.	.	.	.	.	.
1500	Malaysia	40	29	0	18	4	0	0
1501	Malaysia	41	29	0	18	0	0	0
.	.	.	.	.	.	.	.	.
3001	Singapore	81	2299	8	528	191	1	36
3002	Singapore	82	2532	8	560	233	0	32
.	.	.	.	.	.	.	.	.
4378	Vietnam	363	1539	35	1402	2	0	22
4379	Vietnam	364	1540	35	1402	1	0	0
4380	Vietnam	365	1544	35	1406	4	0	4

2.2.4 STEP 4 Construct a data into 7-days bin (total weekly cases)

The daily number of COVID-19 cases may not be representative of COVID-19 behavior in the real world due to the unpredictability of the daily number of COVID-19 cases due to delayed data processing and database updates in different countries. As a result, rather than daily cases, we estimated total weekly cases (7 days each week). Furthermore, we computed the case fatality in each country based on the number of confirmed and dead cases.

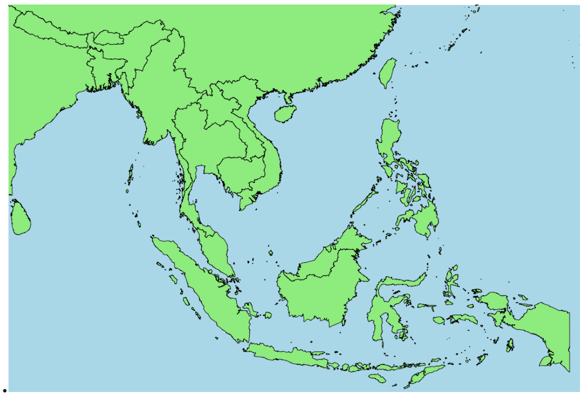
Finally, as shown in table 2.4, the dataset will yield 624 rows with six variables: the country name, the date of the pandemic, the weekly total of confirmed cases, the weekly total of death cases, and the weekly total of recovered cases.

Table 2.4 Final dataset of COVID-19

No.	Countries Name	Date of Pandemic	Weekly Cases	Weekly Death	Weekly Recoveries	Cases Fatality Rate
1	Australia	8	6	0	0	0
2	Australia	15	7	0	2	0
.	.	.	.	.	.	.
259	Malaysia	358	19080	50	11000	0.38957
260	Malaysia	365	24861	67	16084	0.37195
.	.	.	.	.	.	.
441	Singapore	176	1580	1	1665	0.0576
442	Singapore	183	1866	0	1807	0.05539
.	.	.	.	.	.	.
621	Vietnam	344	35	0	42	2.40385
622	Vietnam	351	49	0	30	2.32558
623	Vietnam	358	16	0	16	2.30112
624	Vietnam	365	23	0	37	2.26684

2.3 Study Area

This study focused on 12 countries in South and Southeast Asia, including India, Pakistan, Bangladesh, Nepal, and Sri Lanka, as well as Southeast Asian countries such as the Philippines, Malaysia, Thailand, Vietnam, and Australia, due to the completeness of the dataset. All confirmed cases, recovered cases, and death cases were counted. The data was collected for 365 days, from January 22, 2020, to January 21, 2021



2.4 Path Diagram

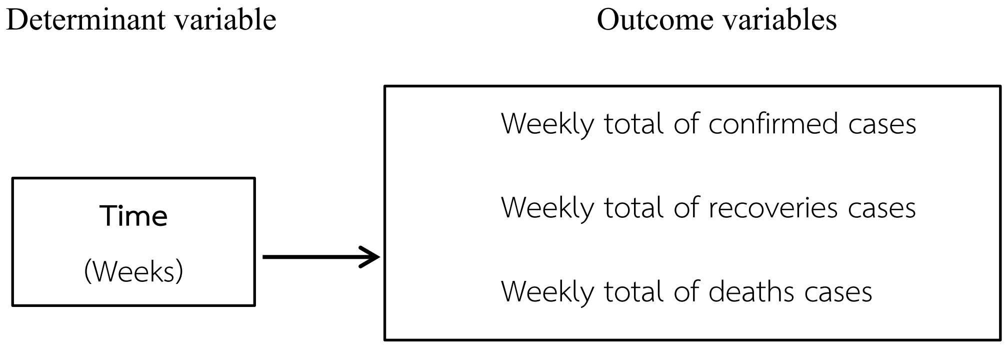

Determinant variable
Outcome variables
Time
(Weeks)
Weekly total of confirmed cases
Weekly total of recoveries cases
Weekly total of deaths cases

Figure 2 Path diagram for outcome and determinant variables of this study

Figure 2 shows the path diagram for the study's outcome and determinant variables, which are the weekly total of confirmed cases, weekly total of recoveries cases, and weekly total of deaths cases. The determinant variable was time, which was shown in pandemic week.

2.5 Data analysis

2.5.1 Descriptive analysis

The data was explored using descriptive analytic methods such as frequency, percent, standard deviation, minimum, maximum, mean, and graphical tools. The case fatality rate (CFR) was used to assess the COVID-19 pandemic's severity. It is the percentage of confirmed cases with multiple mortality cases during a given time period. The formula was written as follows:

$$\text{Case Fatality Rate (\%)} = \frac{\text{(Number of COVID-19 related deaths in a given period)}}{\text{(Number of COVID-19 confirmed cases in a given period)}} \text{ X } 100$$

2.5.2 Statistical modelling

This subsection describes the detail for statistical analysis and modeling. It began with data transformation, followed by the natural cubic spline and linear regression. The brief description for each is as follows.

2.5.2.1 Data transformation

The response variable was transformed using the cube root to satisfy the distributional and variance homogeneity requirements. The cube root transformation was appropriate for COVID-19 data since it avoided over-transforms and handled zero counts. We did not attempt to impute such outliers despite apparent anomalies in the database where records are misplaced and later added. This process provided COVID-19 data that was appropriately prepared and validated.

2.5.2.2 Natural cubic spline function

The interpolation procedure used in COVID-19 data was the spline function for smoothing the regression model. A cubic spline with n knots $X_1 < X_2 <\ldots< X_n$ is the mathematical function provided the smoothing regression with continuous second derivatives comprising piecewise of cubic polynomials between and beyond the location of knots. These functions was written as

$$s(x) = d_0 + d_1 x + d_2 x^2 + d_3 x^3 + \sum_{i=1}^{n} c_i (x - x_i)_+^{\ 3} \qquad (1)$$

Natural cubic spline functions with equispaced knots (McNeil *et al.,* 2011) satisfy the additional requirement that the function is linear for values of x outside the knots. Provided the forecasting of COVID-19 regression value. Since s(x) is linear for x < x_1 if d_2 and d_3 terms in s (x) must also disappear for x < x_n, so to be natural spline, the n + 4 coefficients in the cubic spline must satisfy the following two sets of equations.

$$d_2 = 0\ ,\ \sum_{i=1}^{n} c_i = 0, \qquad (2)$$

$$d_3 = 0\ ,\ \sum_{i=1}^{n} x_i c_i = 0. \qquad (3)$$

When fitting functions to data, natural cubic splines are smoothest in the sense that they minimize the integral of the squared second derivative of the fitted function. For knot x_n where *n* ranges from 1 to *p*, this function has the formula

$$S(x) = a + bx + \sum_{n=1}^{p-2} C_n \{(x - x_n)_+^3 - d_1 (x - x_{p-1})_+^3 + d_2 (x - x_{p-2})_+^3\},$$

Where $d_1 = (x_p - x_n) / (x_p \text{ - } x_{p\text{-}1}), d_2 = (x_{p-1} - x_n) / (x_p \text{ - } x_{p\text{-}1}),$

And x_+ = max (x,0), (The positive part of x).

The knots of the model were placed at an equal interval. Thus, the number of knots suitable for the equation will provide the best fit of the model to the data, which maximizes the r-squared value in the period of observation.

2.5.3 The performance of natural cubic spline model

The correlation coefficient (R^2) value was used to determiand the performance of the natural cubic spline model. The equation was written as.

$$R^2 = \left(1 - \frac{\sum_{i=1}^{n} (P_i - A_i)^2}{\sum_{i=1}^{n} (\bar{A} - A_i)^2} \right) \qquad (1)$$

Where n is the total number of knots of natural cubic spline model, P_i, A_i and $\bar{A}$ are the predicted values, actual values and average of actual values, respectively

The natural cubic spline model was evaluated using cross validation. Cross-validation is a method for testing and training a model that employs different parts of the data.

The Root Mean Square Error (RMSE) is the residuals' standard deviation (prediction error). The distance between the regression line and the data point is measured in residuals. The RMSE is a measure of how evenly distributed the residuals are. To put it another way, how dense the data is around the line of best fit. The root mean square error (RMSE) and the standard deviation are used to determine the model. The formula was written as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_i - A_i)^2} \qquad (2)$$

Where n is the total number of points in cros-validation, Pi and Ai are the predicted values, and actual values, respectively.

2.5.4 The application of natural cubic spline

In practice, the change in the number of new infections per day, which can take both positive and negative values, is a more important statistic than the number of new infections per day. As a result, this research shows how to apply the natural cubic spline to cumulative confirmed cases and growing confirmed cases each day on a linear scale.

Chapter 3
Results

This chapter presents the finding on the patterns and trends of COVID-19 pandemic confirmed cases, recoveries cases, and deaths cases. Furthermore, the natural cubic spline model was used to predicted the COVID-19 confirmed cases in South and Southeast Asian countries. The following results were listed below

1. Descriptive results
2. Fitting the data by natural cubic spline model
3. The performance of the natural cubic spline model
4. The application of the natural cubic spline model to the increasing cases per day

3.1 Descriptive results

The COVID-19 pandemic pattern in South and Southeast Asian countries was investigated using descriptive statistics. Confirmed cases, death cases, recoviries cases, and case fatality rate (CFR) were the variables used.

Table 3.1 The COVID-19 pandemic cases in the selected countries (from 21st January 2020 to 21st January 2021)

Continents	Countries	Confirmed Cases	Death Cases	Recoveries Cases	CFR rate (Percent)	First started date
South Asia	India	10,610,883	152,869	10,265,706	1.44	27/1/2020
	Bangladesh	529,687	7,950	474,472	1.50	7/3/2020
	Pakistan	527,146	11,157	480,696	2.12	26/2/2020
	Nepal	268,310	1,975	262,642	0.74	23/1/2020
	Sri Lanka	55,189	274	47,215	0.50	27/1/2020
Southeast Asia	Indonesia	939,948	26,857	763,703	2.86	1/3/2020
	Philippines	505,939	10,042	466,993	1.98	30/1/2020
	Malaysia	169,379	630	127,662	0.37	25/1/2020
	Singapore	59,197	29	58,926	0.05	23/1/2020

Table 3.1 The COVID-19 pandemic cases in the selected countries (from 21st January 2020 to 21st January 2021)

Continents	Countries	Confirmed	Death	Recoveries	CFR rate	First started
		Cases	Cases	Cases	(Percent)	date
	Thailand	12,795	71	9,842	0.55	13/1/2020
	Vietnam	1,544	35	1,406	2.27	23/1/2020
Australia	Australia	28,749	909	22,716	3.16	25/1/2020

The COVID-19 pandemic in the selected countries is shown in table 3.1. The COVID-19 outbreak in India has more reported cases than the other four countries in South Asia. On January 27, 2020, India and Sri Lanka jointly reported the first case. The COVID-19 epidemic began in Nepal on January 23, 2020, and spread to the other two countries afterwards (February 26th, 2020 in Pakistan and March 7th, 2020 in Bangladesh)**.** During same time span, India recorded 10,610,883 confirmed cases, 10,265,706 recoveries, and 152,869 deaths, respectively. The total number of confirmed recoveries and death cases reported in Bangladesh during this time period were 529,687 cases, 474,472 cases, and 7,950 cases, respectively. The total number of confirmed recoveries and death cases reported in Pakistan during this time period were 527,146 cases, 480,696 cases, and 11,157 cases, respectively. The overall number of confirmed, recovered, and death cases recorded in Nepal during this time period was 268,310 cases, 262,642 cases, and 1,975 cases, respectively. Finally, the total number of confirmed, recoveries, and death cases reported in Sri Lanka were 55,189 cases, 47,215 cases, and 274 cases, respectively.

Thailand is the first country in Southeast Asia to report confirmed cases on January 13th, 2020. Indonesia had the highest confirmed case count in the last month of observation, followed by the Philippines, Malaysia, Singapore, Thailand, and Vietnam. The overall number of verified recoveries and death cases reported in Indonesia during this time period was 939,948, 763,703, and 26,857, respectively. The total number of confirmed recoveries and death cases reported in Philippinse for the time were 505,939, 466,993, and 10,042, respectively. The total number of confirmed recoveries and death cases reported in Malaysia during this time period were 169,379, 127,662, and 630, respectively. The overall number of confirmed, recoveries, and death cases reported in Singapore during this time period were 59,197, 58,926, and 29 deaths, respectively. The overall number of confirmed, recoveries, and death cases reported in Thailand during this time period was 12,795 cases, 9,842 cases, and 71 fatalities. Finally, the total number of confirmed recoveries and death cases reported in Vietnam during this time period were 1,544, 1,406, and 35, respectively.

Despite the fact that Indonesia is the newest Southeast Asian country to report COVID-19 infected cases, the number of confirmed cases continues to rise, with Indonesia having the greatest number of confirmed cases and the highest case fatality rate among Southeast Asian countries. The total number of confirmed cases was 28,749, with 22,716 recovered cases and 909 deaths. In comparison to the other 11 countries, Australia had the highest death rate.

We displayed time series of confirmed, recoveries, and mortality cases data from 12 countries to establish the trend of COVID-19 pandemic in the selected area, as shown in figure 3.

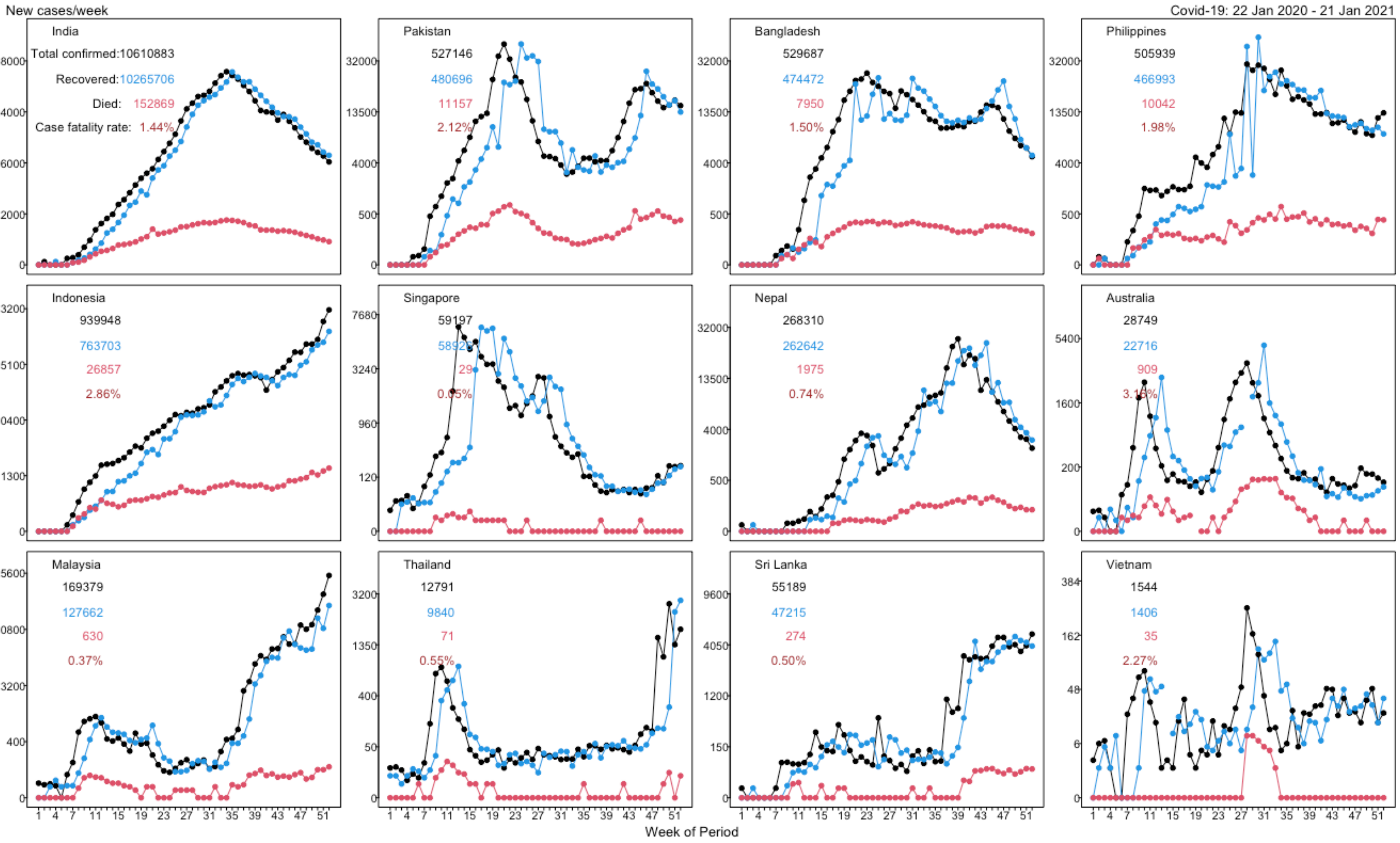


Figure 3 The distribution of COVID-19 pandemic weekly cases from 21st January 2020 to 21st January 2021

The pattern of the COVID-19 pandemic is depicted in figure 3. The number of confirmed cases is represented by the black dots. The number of recovered cases is represented by blue dots, whereas the number of death cases is represented by red dots. Overall, the number of deaths and cases recovered follows a similar trend, with similar-sized increases and decreases. The number of deaths, however, is smaller than the number of confirmed and recovered cases.

The COVID-19 pandemic might be classified into five patterns, as shown in the figure 3. The first pattern includes countries where the number of COVID-19 pandemics increases dramatically without reaching the peak, such as Indonesia. In the meantime, the second pattern is a graphical pattern that increases until it reaches a peak, which is the largest number of confirmed cases one time before decreasing. India, Bangladesh, Nepal, Singapore, and the Philippines were among the five countries in the second pattern. The third pattern is a graphic pattern in which the steady increases, peaks, and drops. The trend then increases to a new high in Malaysia, Thailand, and Sri Lanka, among other places. The fourth pattern of pandemics is an increase in the number of confirmed cases, which peaks twice, as seen in Australia and Pakistan. The last pattern of the COVID-19 graphical epidemic is in Vietnam, which depicts the oscillations pattern across the observation period.

The number of COVID-19 pandemics increases dramatically in the first pattern, but does not reach its peak. The general trend in Indonesia was classified as the pandemic's first pattern. In the last week of the observation period, the overall number of cases peaked. 939,948 confirmed cases, 763,703 recoveries, and 2,685 deaths were reported. In the last week of the study, the case fatality rate was reported to be 2.86 percent. Between the first and fifth weeks, the number of confirmed cases stayed steady, then increased to a high in the 36th week. Following that, the number of confirmed cases decreased somewhat from the 36th to the 41st week. Following the 41st week, the number of COVID-19 confirmed cases increased dramatically over the next 11 weeks, peaking at 81,905 confirmed cases in the 52nd week. COVID-19 cases and deaths were on the rise, following a similar pattern to verified cases. However, there were differences in the number of confirmed and recovered cases and the size of mortality cases.

The second pattern, which visually depicted confirmed cases and recoveries growing to a peak one time before decreasing to the bottom line of the pandemic, which is the stage COVID-19, was controlled. This pattern included five countries: India, Bangladesh, Nepal, Singapore, and the Philippines. Starting with India, the confirmed cases in India are more severe than in the other three countries, with 652,390 confirmed cases in the 34th week. Between the 6th and 23rd weeks, the number of confirmed cases in Bangladesh reached an all-time high of 26,598 cases. From the 36th to the 39th week, there was a gradual decrease to the lower plateau. There were occasional oscillations from the 40th to the 52nd week, but the overall trend was lower after confirmed cases peaked in the 23rd week. With 30,494 confirmed cases in the 28th week, the Philippines had a high number of confirmed cases, followed by a somewhat decreasing trend. The number of confirmed cases in Singapore peaked in the 13th week, with 6,442 confirmed cases. They gradually returned to baseline in the 39th week of the COVID-19 pandemic. Between the 8th and the 39th weeks in Nepal, the number of confirmed cases progressively reaches a peak of 26,876, but the plot shows some volatility between the 22nd and the 25th weeks. Since the 39th week, the number of confirmed cases has been steadily decreasing after peaking. The Philippines has a larger case fatality rate than the other countries, at 1.98 percent, while Singapore has the lowest rate, at 0.05 percent. As previously said, we can group all four countries into the same pandemic pattern.

The third graph revealed that after the reduction to the baseline, the confirmed and recovered cases peaked. The number of confirmed and recovered cases then surged (the second wave of the COVID-19 was detected). In Malaysia, the index fell softly to the baseline level in the 20th week after attaining its initial peak in the 11th week. The data then revealed that the number of confirmed cases had been steadily increasing, reaching a new high on the last day of monitoring, with 23,861 confirmed cases. Despite having the highest number of confirmed cases, Malaysia's case fatality rate of 0.27 percent is lower than the other two countries. Thailand and Sri Lanka have had similar pandemics. The case fatality rates in Thailand and Sri Lanka were 0.55 and 0.50 percent, respectively.

The fourth graph depicted the number of confirmed and recovered cases increasing and peaking twice at various times. The number of confirmed cases in Australia peaked in the 10th week before rapidly falling between the 14th and 20th weeks to the baseline level. The number of confirmed cases began to rise sharply towards the end of the 20th week, peaking at 3,592 cases in the 28th week. The number of confirmed cases then dropped sharply to the baseline level. In Pakistan, the first peak has a higher number of confirmed cases than the second peak, with 40,582 confirmed cases in the 21st week. The confirmed cases were close to the baseline, peaked at the 46th week, and then began to decline throughout the last observation period. Australia has a higher fatality rate than Pakistan, at 3.16 percent and 2.21 percent, respectively. When all of the countries are considered, Australia has the highest fatality rate. As a result, Australia and Pakistan could be grouped together.

Vietnam has the last pattern of COVID-19's graphical epidemic. Between the first and the 23rd week of the obviated, the epidemic in Vietnam oscillates. The most confirmed cases are 258, which is less than the pandemic's peak in observed countries in the 28th week. The number of recoveries is consistent with the number of confirmed cases. With a case fatality rate of 2.27, the number of deaths in Vietnam was discovered in the 28th week, the same week that confirmed cases reached the pandemic's peak.

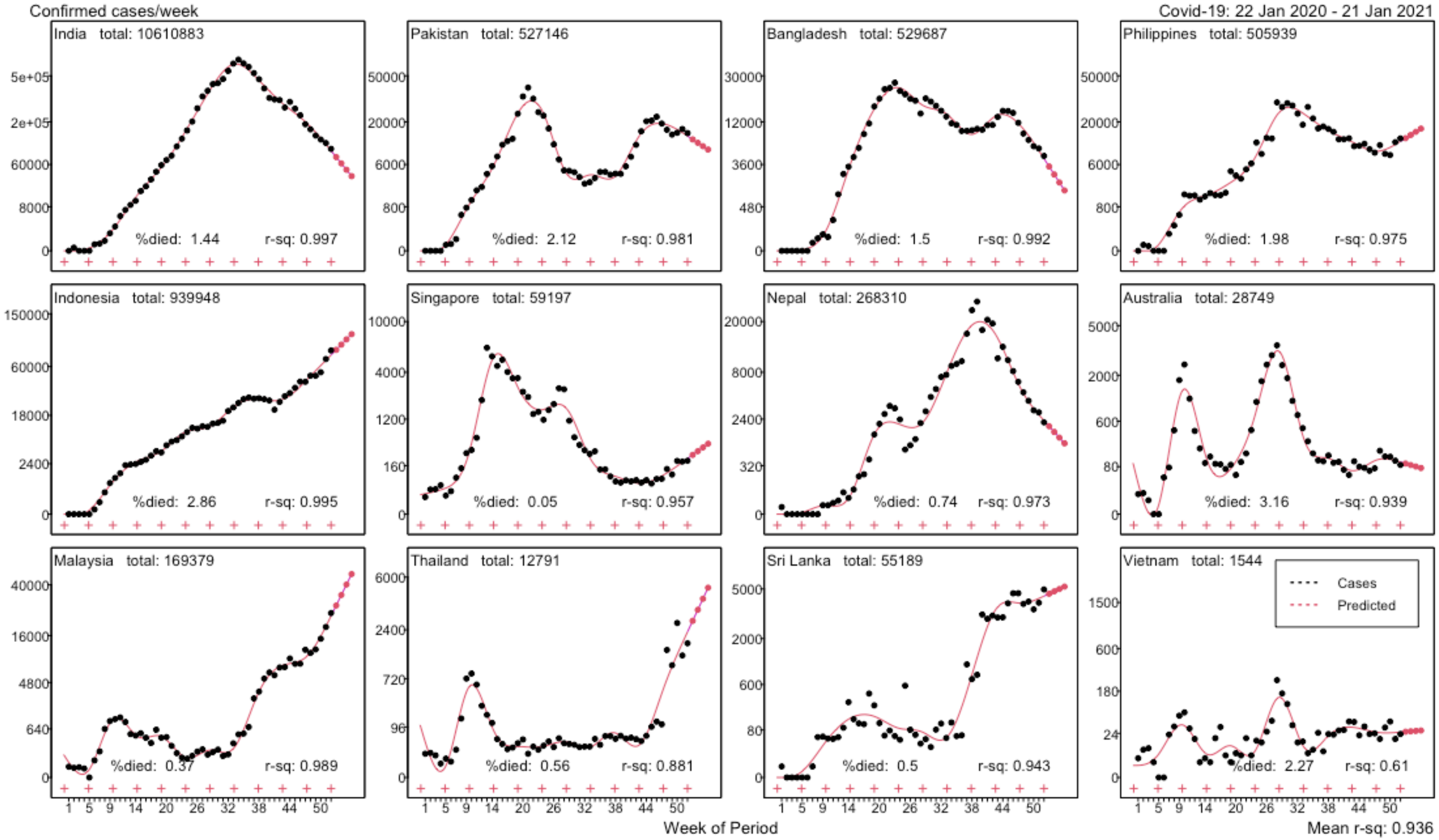


Figure 4 The COVID-19 weekly confirmed cases based on natural cubic spline model

3.2 Natural cubic spline model

A natural cubic spline is linear outside the range of the knots, which can provide the linear forecasting value in the tail of the natural cubic function. So, we employed the natural cubic spline function with equi spaced knots, which were 12 knots, to fit the data and predict the COVID-19 data, as shown in figure 4. The black dot represents the recording of confirmed cases, the pink line is the spline model, and the pink dot is the forecasted data.

The natural cubic spline fitted the data well, with a high r-squared value, according to the findings. The r-squared values varied by country, with India scoring 0.997, Pakistan 0.979, Bangladesh 0.992, the Philippines 0.978, Indonesia 0.995, Singapore 0.961, Nepal 0.975, Australia 0.942, Malaysia 0.989, Sri Lanka 0.945, and Thailand 0.886. As a result, Vietnam's r-squared value, which was 0.598, is poor. In addition, the COVID-19 confirmed cases were forecasted in the short term using the predicting value in the model's tail. Three types of forecasting trends can be identified. The first, confirmed COVID-19 cases, is expected to rise in the next four weeks, starting January 21, 2021, in Indonesia, Malaysia, Thailand, Sri Lanka, the Philippines, and Singapore. Second, the number of COVID-19 confirmed cases in India, Pakistan, Bangladesh, and Nepal declined during the next four weeks. Finally, Australia and Vietnam had a forecasting trend with a consistent trend over the next four weeks.

There were six countries in the first forecast trend, including Indonesia, Malaysia, Thailand, and Sri Lanka, whose graphical plot indicated a rising trend that was associated with the forecasted line. Despite decreasing trends in the Philippines and Singapore over the study's observation period, COVID-19 confirmed cases increased at the end. India, Pakistan, Bangladesh, and Nepal comprised the second category, with a decreasing trend in the short-term future. The third group, which was discovered in Australia and Vietnam, did not show a raised or decreasing trend of confirmed cases in the four weeks after January 21, 2021.

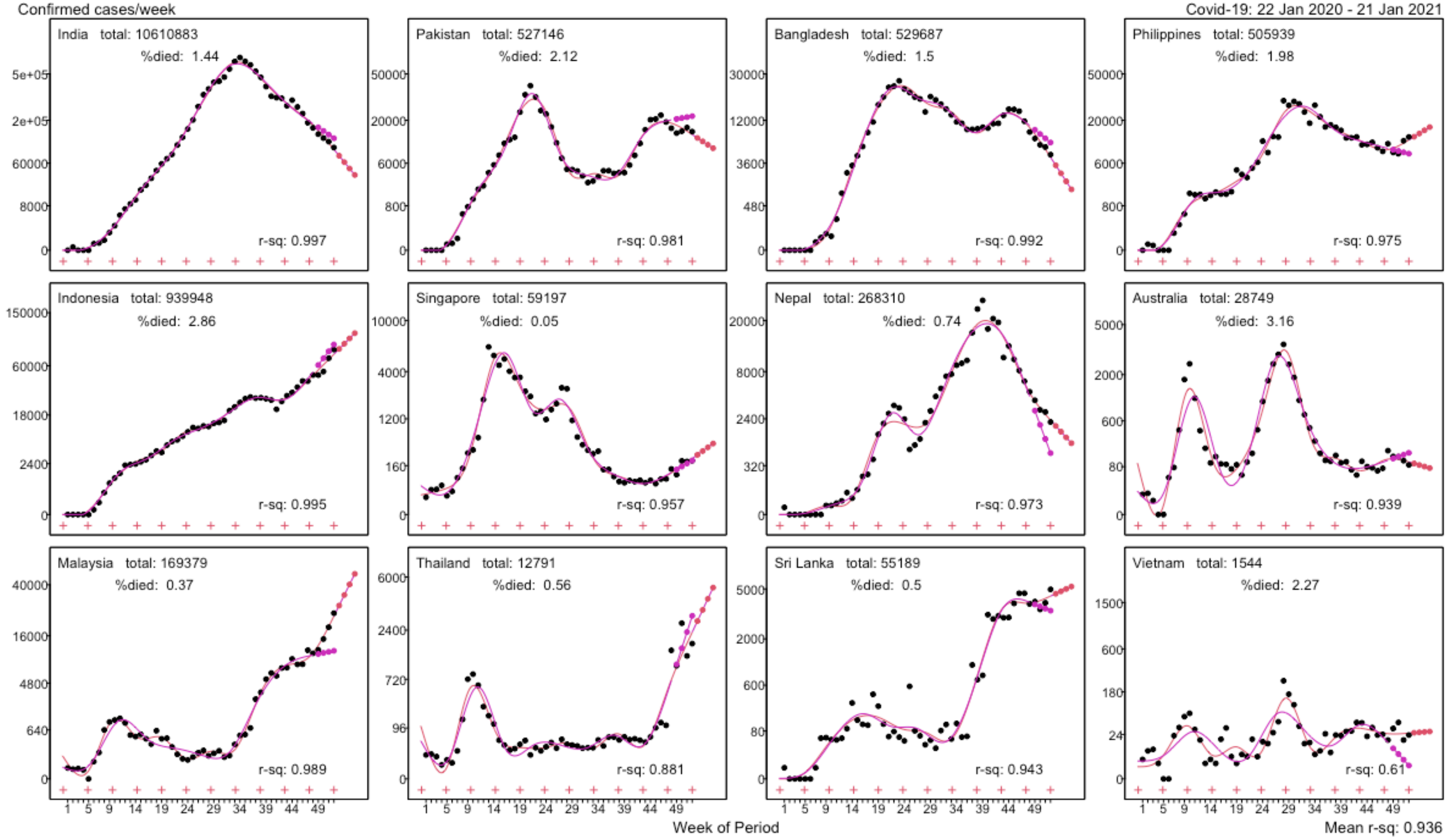


Figure 5. The cross-validation of the natural cubic spline model

3.3 The performance of the natural cubic spline model

Figure 5 shows the model's validity, which was determined by applying the method to reduced data after eliminating the latest four weeks' confirmed cases of observation and then comparing associated forecasts with actual fitted values. The actual data was represented by the black dots. The pink line indicates the training set's fitting spline model (48 weeks of data), while the pink dots reflect the test sample's predicted values (the last four weeks of data).

The predicted and actual values are in agreement based on the results. However, in some areas, such as the Philippines, Singapore, Nepal, Malaysia, and Vietnam, the number of COVID-19 confirmed cases is lower than the actual number of COVID-19 confirmed cases. India, Pakistan, Bangladesh, and Indonesia, on the other hand, have predicted cases that are greater than the actual cases. We calculated the Root Mean Square Error (RMSE) to reflect the model's projected accuracy, as shown in table 3.2.

Table 3.2 Root Mean Square Error (RMSE) for measuring predictive accuracy

Countries Name	**RMSE**	
	Training dataset	**Test dataset**
India	1.505	3.173
Pakistan	1.103	2.626
Banglades	0.833	1.867
Philippines	1.345	1.977
Nepal	1.233	2.829
Sri Lanka	**1.196**	**0.903**
Indonesia	0.686	2.242
Singapore	**0.979**	**0.427**
Vietnam	0.868	1.512
Malaysia	1.089	3.722
Thailand	0.848	2.162
Australia	**1.161**	**0.673**
Average RMSE	**1.071**	**2.001**

The root mean square error (RMSE) of the training and testing data is shown in Table 3.2. When comparing the RMSE of the training and testing models, it was discovered that they were not that dissimilar. This demonstrates that fitting and forecasting the COVID-19 confirmed cases with the natural cubic spline produces credible results that may be used to other datasets.

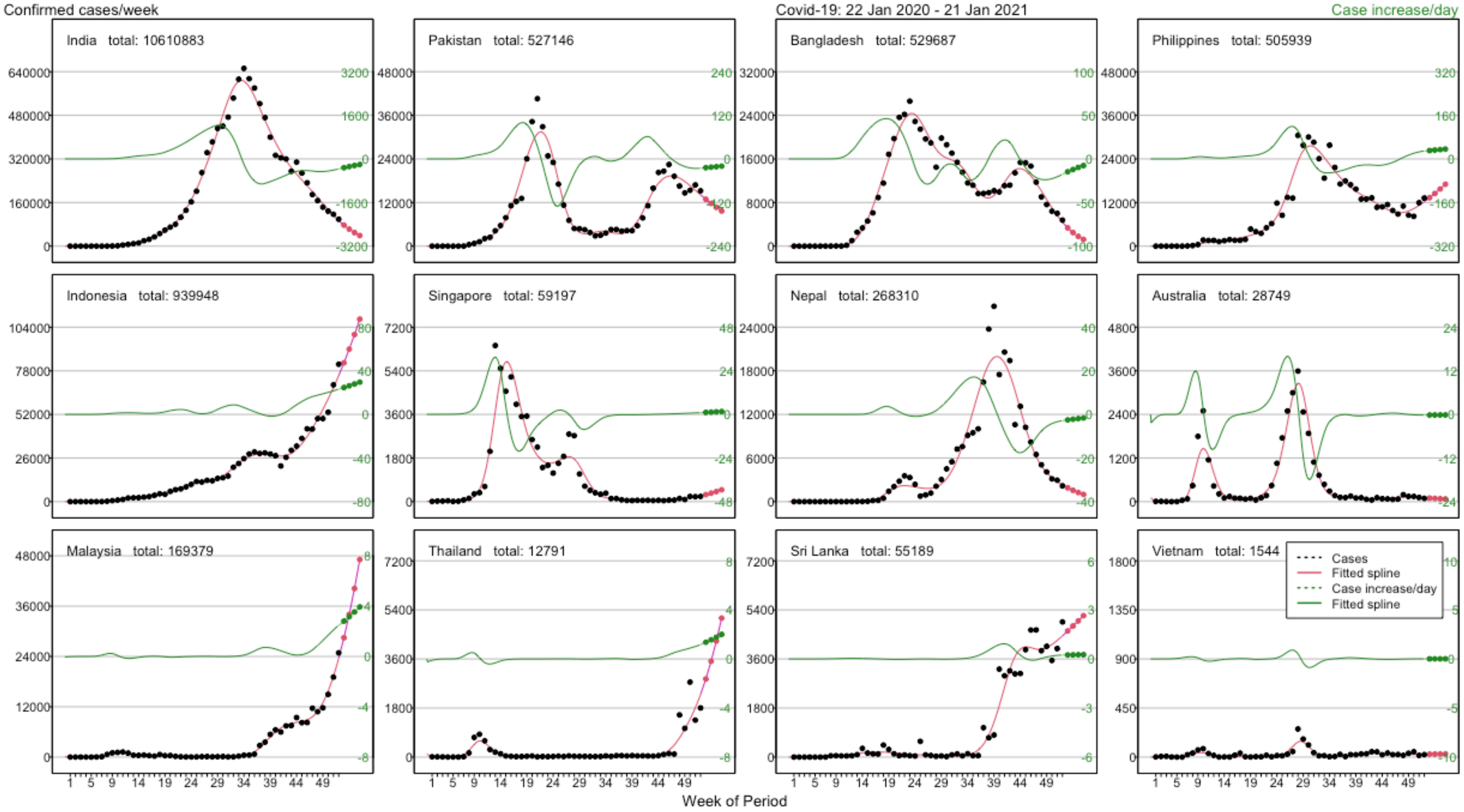


Figure 6 Plotting of the natural cubic spline model to the COVID-19 increasing cases per day

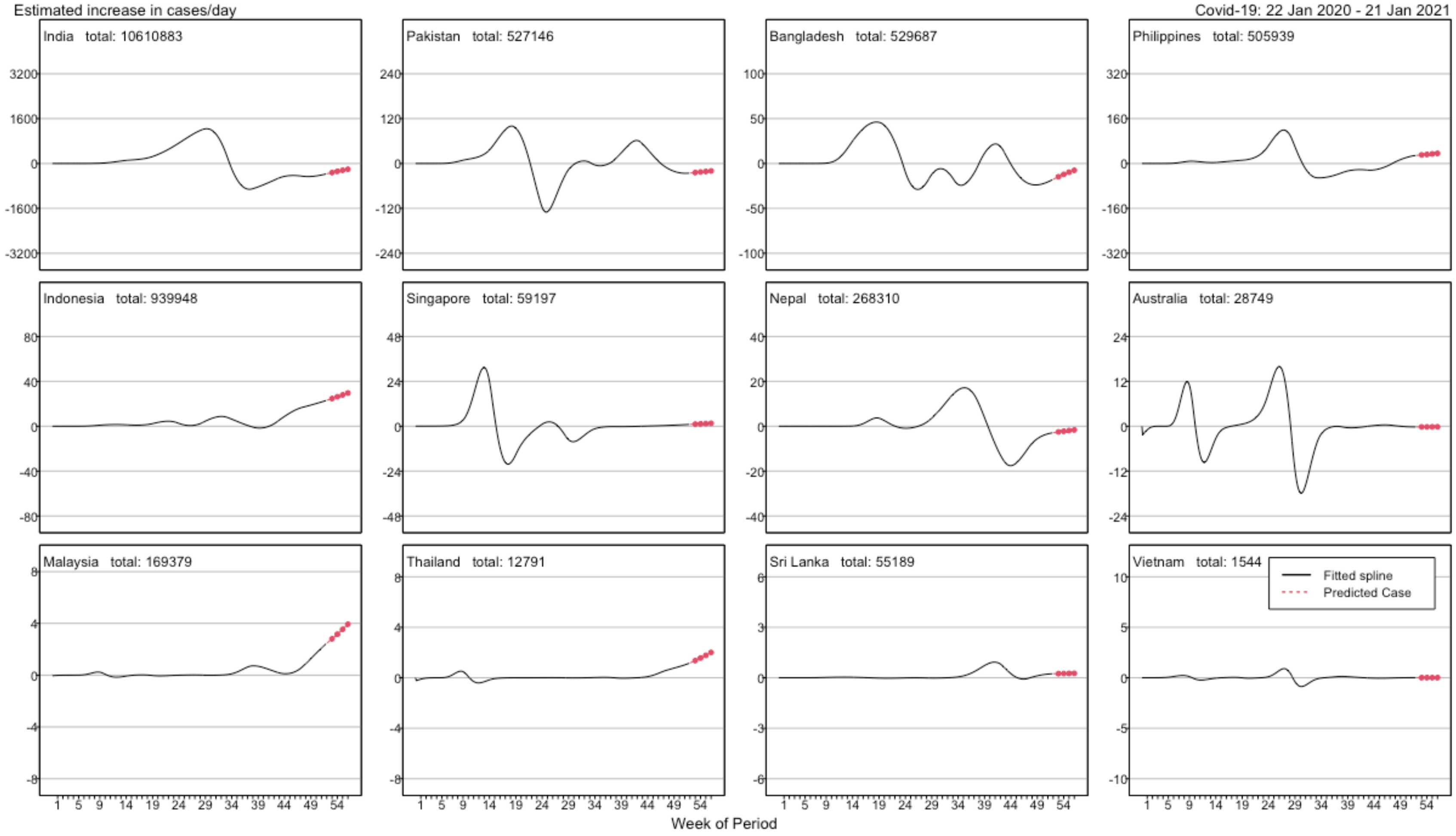


Figure 7 Plots of fitted values (black curves) and forecast values (red dots) of daily increases in numbers of COVID-19 confirmed cases in sampled regions

3.4 The application of the natural cubic spline model

In practice, the change from one day to the next, which can take both positive and negative values, is a more important statistic than the number of new infections every day. As a result, this research demonstrates the practical application of natural cubic spline. On a linear scale, Figure 6 depicts the plot of the natural cubic spline to cumulative confirmed cases and increasing confirmed cases per day. The number of confirmed cases is represented by the Y axis on the left. On the right, the Y axis depicts the growing number of confirmed cases per day. The real data is shown by the black dot, while the fitted spline model is represented by the red curve. The red and green dots represent the forcast values for confirmed cases and increasing confirmed cases per day. For more accessible in interpretable, we extract the plots for the increasing cases per day into another plot with the Y-axis representing the number of increasing cases per day as shown in figure 7.

Figure 7 depicts the daily increases in COVID-19 confirmed cases in sampled regions as plots of fitted values (black curves) and forecast values (dotted red curves). In different countries, the number of increasing cases per day varied. As a result, the fluctuation in the plot can be taken as a big variation in the number of cases within that time period. The negative values indicate that the cases on that day were lower than the previous day, while the positive values indicate that the cases were higher. Furthermore, the forecast values reveal that some countries, such as Nepal, Indonesia, Singapore, Australia, and the Philippines, are experiencing an increase in the number of cases per day. The rest of the countries, on the other hand, have remained steady throughout time, with increasing cases per day.

Chapter 4
Discussions and Conclusions

This study explored the patterns and trends of COVID-19 pandemic confirmed cases, recoveries cases, and death cases and predicted the COVID-19 confirmed cases in South and Southeast Asian countries. The results challenge statisticians to identify the trend and predict infectious disease. The following were some interesting points emerging from the findings of the study.

4.1 Discussion

4.1.1 Discussion of discriptive

According to the findings, the case fatality rate (CFR) in Southeast Asia ranged from 0.05 to 2.86 percent. Nepal has the highest CFR in South Asia, with 2.12 percent. The lowest CFR was 0.05 percent in Sri Lanka. Indonesia has the highest CFR in Southeast Asia, at 2.86 percent. With a CFR of 0.05 percent, Singapore was the lowest. COVID-19 CFR varies throughout countries, according to this study. There were numerous causes for such a change in the CFR. Some countries have experienced the outbreak earlier than others. The differences in CFR could indicate distinct stages of the disease's spread. As a result, the number of deaths and cases, which are factors in the CFR equation, influenced the CFR variation. Number of COVID-19 deaths varied between countries due to factor increased mortalities risk in population(Rajgor *et al.,* 2020; Spychalski *et al.,* 2020; Yang *et al.,* 2020). Clinical data has revealed various risk factors linked to a poor prognosis at the individual level. The risk of death appears to be considerably increased by age and comorbidities (cardiovascular disease, malignancies, diabetes mellitus, chronic lung disease) (Sorci *et al.,* 2020). The number of COVID-19 cases varied by country and was determined by the testing procedures used in each country.

The graph depicted the pandemic's spreading patterns and severity, as well as the scale of the pandemic in each country. Simultaneously, visualizing the COVID-19 time series data reveals the pandemic pattern, which can be divided into five groups, each indicating a different stage of the epidemic throughout South and Southeast Asia. The easiest method to grasp the ongoing infectious disease was to visualize it using charts. This graphical method depicts the data behavior that revealed the pandemic pattern and trend in South and Southeast Asia. As a result, public health organizations can focus on resource management, such as public measurement and immunizations, in order to combat the epidemic. However, no statistical analysis results were found in this investigation to corroborate the pattern grouping of the COVID-19 data. For data grouping, many statistical methods have been utilized, including time series cluster analysis (Zarikas *et al.,* 2020; Azarafza *et al.,* 2021). Cluster analysis is a way of categorizing data that uses statistical values to find the trend and pattern. The researcher can use such statistical statistics to expand on the graph's findings, making them more robust and accurate.

4.1.2 Discussion of natural cubic spline model

The study analyzed the pandemic trend using a natural cubic spline model and the number of cumulative confirmed cases in each country. When using a natural cubic spline model with equal space knots to fit the data, the model fits well and has good r-squared values. Some countries, however, have a lower r-squared than others, such as Vietnam, which has an r-squared of 0.598. The low r-squared was taken as indicating that the model did not fit the data adequately. The natural cubic spline model was limited in this case (Sousa *et al.,* 2021). Overfitting is another issue with the natural cubic spline model, which may be evaluated using statistical approaches. The cross-validation method, on the other hand, can be used to assess the performance of the natural cubic spline model. The cross-validation procedure is mostly based on the model's projected accuracy. The performance of the natural cubic spline may be related to knot selection, according to earlier research (Kohavi *et al.,* 1995). As a result, knots were changed in countries with low COVID-19 report pandemic cases, and COVD-19 data oscillation was required to increase prediction accuracy (Iddrisu *et al.,* 2021).

4.1.3 Disscussion of the model performance

A situation occurs when a model includes more parameters than are required to monitor a signal, leading to concerns about model overfitting. As a result, we used the cross validation method to validate the model by dividing the data into training and testing data sets. Then we compared the resulting forecasts to the actual fitted values. Apart from r-squared values, we compared the RMSE for training and testing data sets to assess the effectiveness of the natural cubic spline model. The RMSE between these models was not significantly different. This could indicate that the models were not overfit. Cross-validation error was employed to analyze the natural cubic spline model, which was adjusted r-squared to compromise between goofiness of fit and smoothness of the seasonal curve (Wongsai *et al.,* 2017).

However, the model provides a low r-squared for some countries, such as Vietnam, based on the data. This could be due to the natural cubic spline's limitation when compared to the equispace knot, which employs the same number and position of knots. Using the same knot in different datasets can impact the COVID-19's predicting effectiveness. Furthermore, many other aspects of the pandemic's progression were not taken into account in this analysis.

4.1.4 Discussion of the model application (Increasing cases per day)

Due to government policy, which releases several measurements to deal with the disease in order to control its spread, confirmed cases rapidly fall in specific periods. The virus's mutation and a lack of health infrastructure in some countries, on the other hand, were contributors in the rapid increase in COVID-19 cases. To learn from the COVID-19 data, however, we must use biological theory to comprehend the data's behaviors. The epidemiological theory also provides further information on this subject.

The daily change in confirmed cases, as well as the daily change in infections, is an important statistic that can show the number of cases increasing per day, which can be both positive and negative. It provides timely information on the pandemic's spread. This information could be used by related organizations or health policymakers in their work. The increasing number of cases per day indicates a problem that requires immediate attention.

4.2 Conclusion

The COVID-19 epidemic is still spreading over the world. Unstable COVID-19 patterns in South and Southeast Asian countries due to virus genome mutation, which is anticipated to reduce prediction accuracy. The precise mathematical model may aid in the prediction of pandemic waves. The natural cubic spline with equi-space knot was one of several possible techniques for forecasting COVID-19 confirmed cases in the short term that may be used at any moment.

Furthermore, the model provides daily change estimates, indicating when action is required. This model can be be expanded to include additional factors such as environmental and demographic data.

4.3 Limitation of the study

The COVID-19 confirmed weekly cases and growing cases per day fit well with the natural cubic spline model with the equi-space knot. As a result, extra variables, such as health condition, reproductive rate, age, sex and comorbidities, demography, and environmental factors, must be added to the model to improve prediction accuracy. It's also important looking into how the models perform in different places. Furthermore, the equi-spec knot's position and number of knots, which were used the same number and location of knots in different countries, model with an equi-spec knot delivers low accuracy in some countries with an oscillation pattern of data and a small number of cases. The location and number of knots should match to the pattern of change in the data by altering the knot of the natural cubic spline model to fit the varied datasets to increase the model's accuracy.

References

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J.W., Choi, P.M., Kitajima, M., Simpson, S.L., Li, J., Tscharke, B., Verhagen, R., Smith, W.J., Zaugg, J., Dierens, L., Hugenholtz, P., Thomas, K.V. and Mueller, J.F. 2020. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. Science of The Total Environment, 728,1-8. doi.org/10.1016/j.scitotenv.2020.138764

Alabool, H., Alarabiat, D., Abualigah, L., Habib, M., Khasawneh, A. M., Alshinwan, M. and Shehab, M. 2020. Artificial intelligence techniques for Containment COVID-19 Pandemic: A Systematic Review, Research Square, 1-27. doi:10.21203/rs.3.rs-30432/v1

Aleem A, Shah H. 2020. Gastrointestinal And Hepatic Manifestations Of Coronavirus (COVID-19). In StatPearls [Internet]. StatPearls Publishing. Available online: https://www.ncbi.nlm.nih.gov/books/NBK570562/. [February 15, 2022]

Arora, P., Kumar, H. and Panigrahi, B. K. 2020. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos, Solitons & Fractals, 139. doi: 10.1016/j.chaos.2020.110017

Aviv-Sharon, E. and Aharoni, A. 2020. Generalized logistic growth modeling of the COVID-19 pandemic in Asia. Infectious Disease Modelling, 5, 502-509. doi: 10.1016/j.idm.2020.07.003

Azarafza, M., Azarafza, M. and Akgün, H. (2021). Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran. Journal of Applied Science, Engineering, Technology, and Education, 3(1), 1-6. 33(1), 117-120. doi: https://doi.org/10.1101/2020.05.22.20109942

Cucinotta, D. and Vanelli, M. (2020). WHO declares COVID-19 a pandemic. Acta bio-medica: Atenei Parmensis, 91(1), 157-160. doi: 10.23750/abm.v91i1.9397

Feroze, N. 2020. Forecasting the patterns of COVID-19 and causal impacts of lockdown in top five affected countries using Bayesian Structural Time Series Models. Chaos, Solitons & Fractals, 140. doi: 10.1016/j.chaos.2020.110196

Hale, T., Angrist, N., Kira, B., Petherick, A. and Phillips, T. 2020. Variation in government responses to COVID-19 BSG-WP-2020/032, BSG Working Paper Series, Version 6.0. Available online: https://www.bsg.ox.ac.uk/sites/default/ files/2020-05/ BSG-WP-2020-032-v6.0.pdf. [August 15, 2020]

Iddrisu, A. K., Amikiya, E. A. and Otoo, D. (2021). A predictive model for daily cumulative COVID-19 cases in Ghana. F1000Research, 10(343), 343. https://doi.org /10.12688/f1000research.52403.1)

Iftikhar, H. and Iftikhar, M. 2020. Forecasting daily COVID-19 confirmed, deaths and recovered cases using univariate time series models: A case of Pakistan study, MedRxiv, 1-13. doi.org/10.1101/2020.09.20.20198150

Ilie, O. D., Cojocariu, R. O., Ciobica, A., Timofte, S. I., Mavroudis, I. and Doroftei, B. 2020. Forecasting the spreading of COVID-19 across nine countries from Europe, Asia, and the American continents using the ARIMA models. Microorganisms, 8(8). doi.org/10.3390/microorganisms8081158

Karimuzzaman, M., Afroz, S., Hossain, M. M. and Rahman, A. 2020. Forecasting the COVID-19 Pandemic with Climate Variables for Top Five Burdening and Three South Asian Countries. MedRxiv. doi.org/10.1101/2020.05.12.20099044

Kavadi, D. P., Patan, R., Ramachandran, M. and Gandomi, A. H. 2020. Partial derivative nonlinear global pandemic machine learning prediction of covid 19. Chaos, Solitons & Fractals, 139. doi: 10.1016/j.chaos.2020.110056

Kohavi, R. 995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai Vol. 14, No. 2, pp. 1137-1145 doi : 10.5555/1643031. 1643047

McNeil, N., Odton, P. and Ueranantasun, A. 2011. Spline interpolation of demographic data revisited. Sonklanakarin Journal of Science and Technology, 33(1), 117-120.

Petrosillo, N., Viceconte, G., Ergonul, O., Ippolito, G. and Petersen, E. 2020. COVID-19, SARS and MERS: are they closely related?. Clinical Microbiology and Infection, 26(6), 729-734. doi: 10.1016/j.cmi.2020.03.026

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., Yan, P. and Chowell, G. 2020. Short-term forecasts of the COVID-19 epidemic in

Guangdong and Zhejiang, China: February 13–23, 2020. Journal of clinical medicine, 9(2), 596. doi: 10.3390/jcm9020596.

Sarkar, K., Khajanchi, S. and Nieto, J. J. 2020. Modeling and forecasting the COVID-19 pandemic in India. Chaos, Solitons & Fractals, 139. doi: 10.1016/j.chaos.2020.110049

Sousa, A. R. S., Severino, M. T. and Leonardi, F. G. 2020. Model selection criteria for regression models with splines and the automatic localization of knots. arXiv preprint arXiv. doi: 10.48550/arxiv.org.2006.02649

Wieczorek, M., Siłka, J. and Wozniak, M. 2020. Neural network powered COVID-19 spread forecasting model. Chaos, Solitons & Fractals, 140. doi: 10.1016/j.chaos.2020.110203

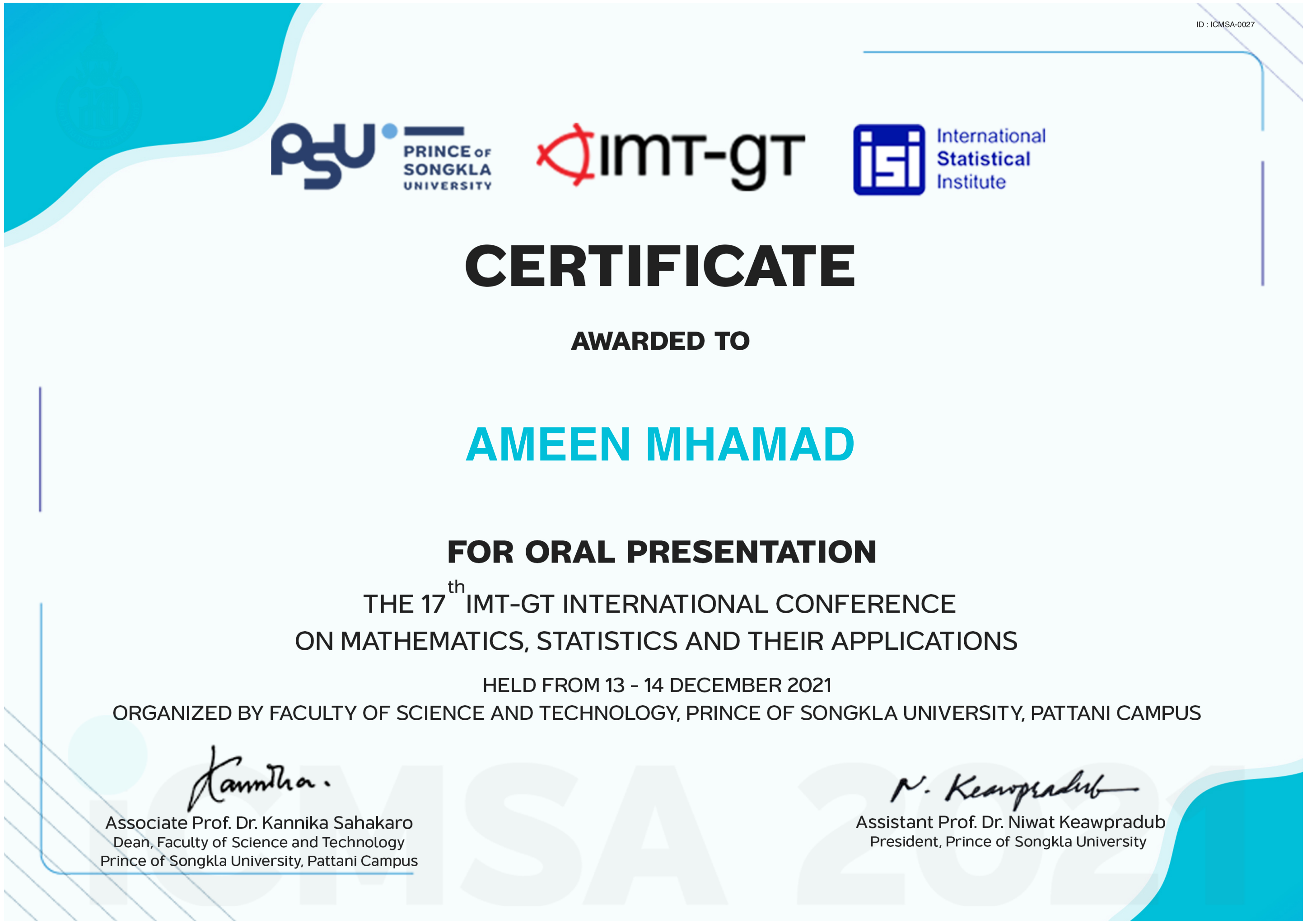
Yousaf, M., Zahir, S., Riaz, M., Hussain, S. M. and Shah, K. 2020. Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan. Chaos, Solitons & Fractals. doi: 10.1016/j.chaos.2020.109926

Zarikas, V., Poulopoulos, S. G., Gareiou, Z. and Zervas, E. 2020. Clustering analysis of countries using the COVID-19 cases dataset. Data in brief, 31, 105787. doi: 10.1016/j.dib.2020.105787

Zeroual, A., Harrou, F., Dairi, A. and Sun, Y. 2020. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. Chaos, Solitons & Fractals, 140. doi: 10.1016/j.chaos.2020.110121

Appendix

Appendix 1 Certification



ID : ICMSA-0027

PSU PRINCE OF SONGKLA UNIVERSITY

IMT-GT

isi International Statistical Institute

CERTIFICATE

AWARDED TO

AMEEN MHAMAD

FOR ORAL PRESENTATION

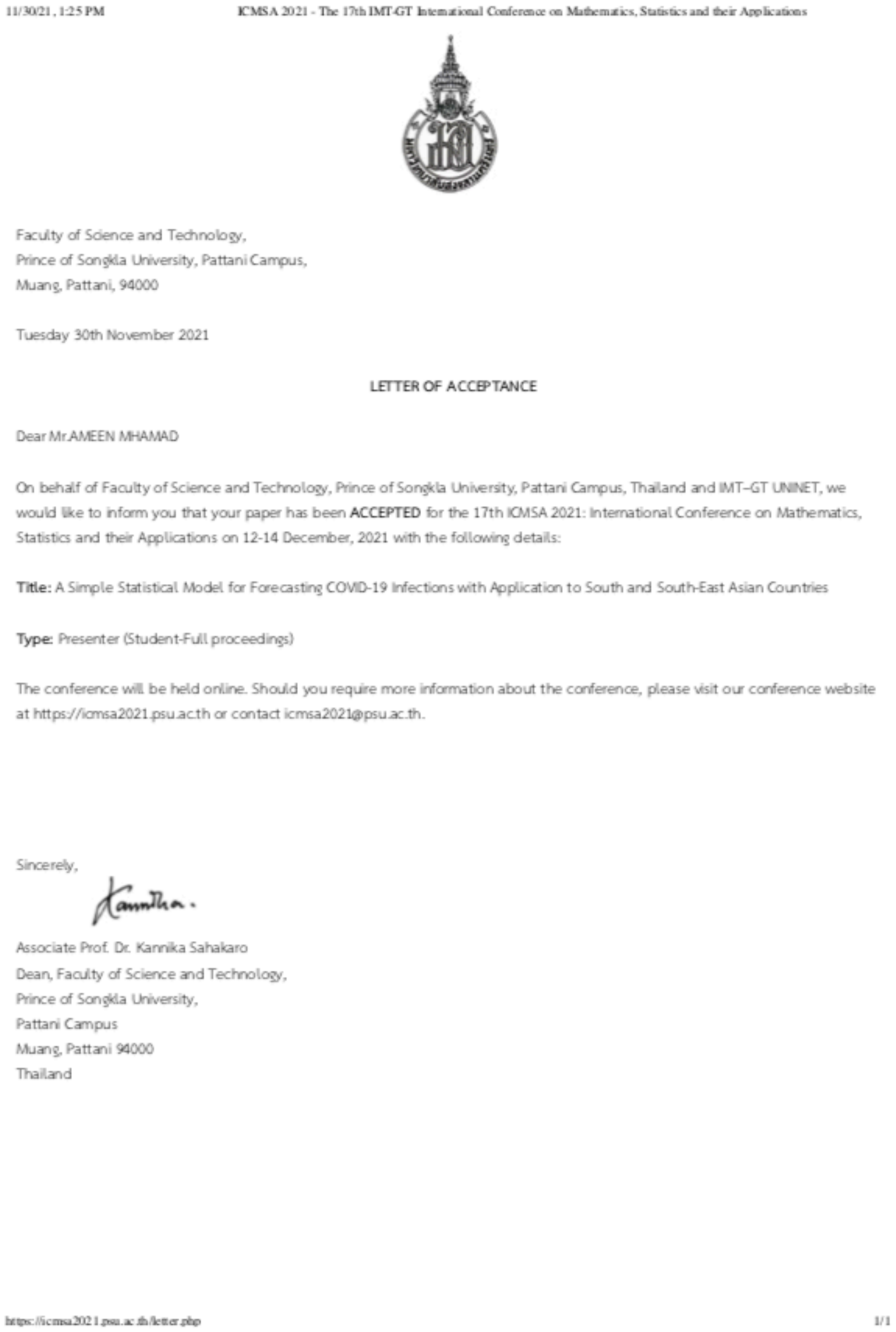
THE 17th IMT-GT INTERNATIONAL CONFERENCE
ON MATHEMATICS, STATISTICS AND THEIR APPLICATIONS

HELD FROM 13 - 14 DECEMBER 2021
ORGANIZED BY FACULTY OF SCIENCE AND TECHNOLOGY, PRINCE OF SONGKLA UNIVERSITY, PATTANI CAMPUS

Kannika.
Associate Prof. Dr. Kannika Sahakaro
Dean, Faculty of Science and Technology
Prince of Songkla University, Pattani Campus

N. Keawpradub
Assistant Prof. Dr. Niwat Keawpradub
President, Prince of Songkla University

Appendix 2 Latter of acceptance



11/30/21, 1:25 PM ICMSA 2021 - The 17th IMT-GT International Conference on Mathematics, Statistics and their Applications

Faculty of Science and Technology,
Prince of Songkla University, Pattani Campus,
Muang, Pattani, 94000

Tuesday 30th November 2021

LETTER OF ACCEPTANCE

Dear Mr.AMEEN MHAMAD

On behalf of Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand and IMT–GT UNINET, we would like to inform you that your paper has been **ACCEPTED** for the 17th ICMSA 2021: International Conference on Mathematics, Statistics and their Applications on 12-14 December, 2021 with the following details:

Title: A Simple Statistical Model for Forecasting COVID-19 Infections with Application to South and South-East Asian Countries

Type: Presenter (Student-Full proceedings)

The conference will be held online. Should you require more information about the conference, please visit our conference website at https://icmsa2021.psu.ac.th or contact icmsa2021@psu.ac.th.

Sincerely,

Kannika.

Associate Prof. Dr. Kannika Sahakaro
Dean, Faculty of Science and Technology,
Prince of Songkla University,
Pattani Campus
Muang, Pattani 94000
Thailand

https://icmsa2021.psu.ac.th/letter.php 1/1

VITAE

Name Ameen Mhamad

Student ID 6220320004

Educational Attainment

Degree	Name of Institution	Year of Graduation
Bachelor of Science	Chulalongkorn University	2017
Mathayom 6	Kanarasadornbumrunf School	2015

Scholarship Awards during Enrolment

Research Grant for Thesiss form Graduate School, Prince of Songkla University, Pattani, Thailand.

Work - Position and Address

Scientist, The Halal Science Center, Chulalongkorn University, Pattani.

List of Publication and Proceeding

Proceeding :

Mhamad, A., Dureh, N., Hazanee, A. (2022): A Simple Statistical Model for Forecasting COVID-19 Infections with Application to South and South-East Asian Countries. The 17th IMT-GT International Conference on Mathematics, Statistics and Their Applications (ICMS2021), Pattani, Thailand.