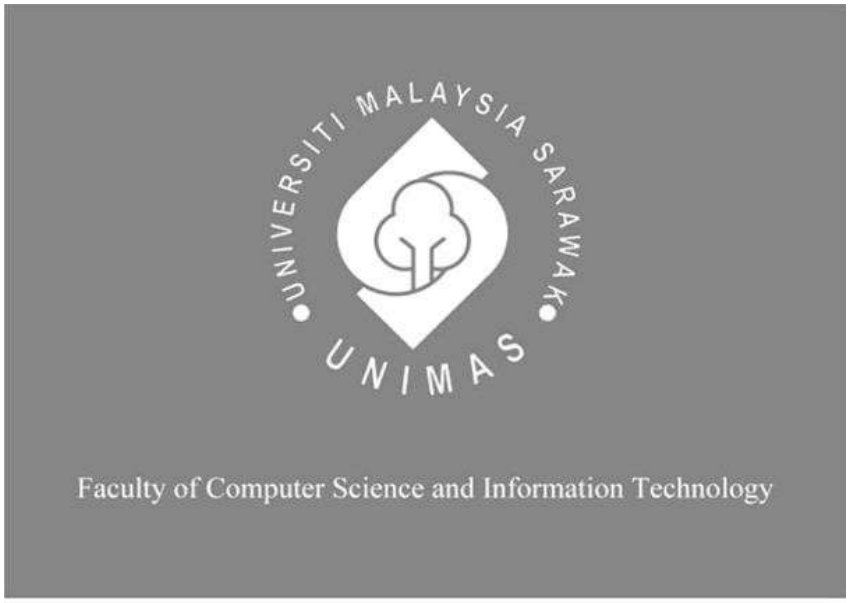

Early Warning Detection For Extreme Events In Time Series Using Topological Data Analysis

KHOO ZI YIK

Bachelor of Computer Science with Honours
(Computational Science)

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE Early Warning Detection For Extreme Events In Time Series Using Topological Data Analysis

ACADEMIC SESSION: 2024/2025

KHOO ZI YIK
(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

☐	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
☐	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
☑	UNRESTRICTED	

Validated by

Khoo
(AUTHOR'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Permanent Address

6378, Lorong Inang 5/1, Taman Ria Jaya, 08000 Sungai Petani, Kedah

Date: 21 June 2025

Date:

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

Early Warning Detection For Extreme Events In Time Series Using Topological Data Analysis

Khoo Zi Yik

This project is submitted in partial fulfilment

of the requirements for the degree of

Bachelor of Computer Science with Honours

(Computational Science)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
(2024)

Pengesanan Awal Untuk Peristiwa Ekstrem Dalam Siri Masa Menggunakan Analisis Data Topologi

Khoo Zi Yik

Projek ini merupakan salah satu keperluan

untuk Ijazah Sarjana Muda Komputer

dan Teknologi Maklumat

(Sains Komputan)

Fakulti Sains Komputer dan Teknologi Maklumat
UNIVERSITI MALAYSIA SARAWAK
(2024)

Acknowledgement

I thank God for His guidance and blessings throughout this journey. He has provided me with strength in moments of difficulty, courage in times of fear, and clarity in moments of uncertainty.

I would like to express my sincere gratitude to my supervisor, Dr. Phang Piau, for the invaluable guidance, support, and encouragement provided throughout this project. Your expertise and insights have greatly shaped my learning and growth.

To my parents, Khoo Kim Hee and Ho Woei Ling, your unwavering love, support, and sacrifices have been my foundation. Thank you for always being there for me, through every challenge and triumph. To my sister, Khoo Zi Xin, your companionship and understanding have meant the world to me.

To my dear friends, Wong Wen Hong and Wong Yu Seng, thank you for your constant support, encouragement, and friendship. Your belief in me has been a great source of motivation.

Lastly, I would like to thank the Faculty of Computer Science and Information Technology for providing such a supportive learning environment for conducting my research project.

Table of Contents

List of Figures

List of Tables

Abstract

This research investigates the application of an approach in Topological Data Analysis (TDA), specifically persistent homology, to detect extreme events in time series data as an early warning signal. Extreme events, often occurring unexpectedly and without prior warning, can have significant repercussions on economic and social systems. These events are frequently derived from time series data, which pose analytical challenges due to their inherent characteristics such as non-linearity, noise, and fluctuations. These complexities complicate the use of traditional statistical methods for effective analysis. The most recent and impactful extreme event in Malaysia was the COVID-19 pandemic. Consequently, this study focuses on COVID-19 as a case study for detecting extreme events. The research outlines a systematic methodology, beginning with data collection and pre-processing through phase plane reconstruction, followed by feature extraction using persistent homology, and culminating in feature selection. This methodological framework aims to provide valuable insights into the underlying structure of time series data and to facilitate the identification of anomalies or extreme events within the data.

Keywords: Early warning, extreme events, persistent homology, time series data, topological data analysis (TDA)

Abstrak

Kajian ini menyelidiki penerapan pendekatan dalam Analisis Data Topologi, khususnya homologi berterusan (persistent homology), untuk mengesan peristiwa ekstrem dalam data siri masa sebagai isyarat amaran awal. Peristiwa ekstrem, yang sering berlaku tanpa dijangka dan tanpa amaran awal, boleh memberikan kesan yang ketara terhadap sistem ekonomi dan sosial. Peristiwa ini biasanya berasal daripada data siri masa, yang menimbulkan cabaran analisis disebabkan oleh ciri-cirinya seperti ketidaklinearan, bunyi, dan turun naik. Kerumitan ini menyukarkan penggunaan kaedah statistik tradisional untuk analisis yang berkesan. Peristiwa ekstrem terkini dan paling memberi kesan di Malaysia adalah pandemik COVID-19. Oleh itu, kajian ini memberi tumpuan kepada COVID-19 sebagai kajian kes untuk mengesan peristiwa ekstrem. Penyelidikan ini menggariskan kerangka metodologi yang sistematik, bermula dengan pengumpulan data dan prapemprosesan melalui pembinaan semula satah fasa (phase plane reconstruction), diikuti oleh pengekstrakan ciri menggunakan homologi berterusan, dan diakhiri dengan pemilihan ciri. Kerangka metodologi ini bertujuan untuk memberikan pandangan mendalam tentang struktur asas data siri masa dan memudahkan pengenalpastian anomali atau peristiwa ekstrem dalam data.

Kata kunci: Amaran awal, peristiwa ekstrem, homologi berterusan, data siri masa, analisis data toopologi

Chapter One

Introduction

1.0 Overview

This research focuses on early warning detection of extreme events in time series by using the topological data analysis. This chapter will offer a brief review of the background of study in section 1.1, problem statement for this project in section 1.2, the research objectives for this project are stated in section 1.3, scope of study were discussed in section 1.4, significant of study in section 1.5, project outcome in section 1.6, a brief methodology in section 1.7 and last but not least, the summary of the chapter in section 1.8.

1.1 Background of Study

With the advancement of technology, the significance of data has become increasingly evident in every aspect of human life (Otter et al., 2017). Data has profoundly impacted various fields, including medicine and healthcare, engineering, and agriculture (Corbet et al., 2019). Bukkuri et al. (2021) stated that by leveraging mathematics particularly statistics combined with technological innovations, the value of data is now more apparent than ever. Through statistical analysis, fundamental metrics can be derived and utilized for further investigation. However, with the exponential growth of data generation across diverse domains, the structure of data has become increasingly complex, surpassing the analytical capabilities of traditional methods.

Data can generally be categorized into two types: time series data and cross-sectional data. Cross-sectional data comprises multiple variables recorded at the same point in time (Wang & Cheng, 2020). Meanwhile, time series data is collected over consistent time intervals (Fatima & Rahimi, 2024). Recent technological advancements, such as sensors and data loggers, have made it possible to collect large volumes of time series data. However, along the occurrence in the time series, the pattern of the data can be considered as nonlinear, non-stationary, and fluctuating (Rhif et al., 2019). This could lead to some extreme events could be break out at the certain point of time

Back in December 2019, COVID-19 has spread rapidly among the residents of Wuhan City, Hubei Province, China and emerged as a global threat, becoming the most dangerous influenza virus (Hashim et al., 2021). On certain days, the daily global infection count reached unprecedented levels, causing widespread panic. Nations worldwide implemented strict measures, including lockdowns, to reduce transmission rates. In Malaysia, the government introduced the Movement Control Order (MCO) to curb the spread of the virus (Shah, 2020).

While effective in slowing the growth of new cases, these measures severely impacted the economy. According to Karim et al. (2020), the nation's tourism industry has been crippled with an estimate loss of RM 3.37 billion in the first 2 months of the year. Within a year of the outbreak, vaccines were developed, and mass immunization campaigns began, leading to initial signs of improvement. However, as time passed, the virus mutated, and by 2022, the emergence of variants such as Omicron triggered a resurgence in cases. Despite the availability of vaccines, the rapid transmission of the mutated virus led to a significant increase in infection rates. The sudden surge in infected cases caught many off guard, leading to its classification as one of the most severe extreme events of the 21st century (Varrelman et al., 2024).

Extreme events are characterized as infrequent, unforeseen occurrences that significantly deviate from typical patterns (Machado & Lopes, 2020). COVID-19 data, as a time series, exhibits characteristics typical of extreme events, including non-linearity, non-stationarity, and noise. The element of non-stationarity causes statistical properties to vary over time, while noise introduces random variations (Ranjai Baidya & Lee, 2024). These factors increase the complexity of data analysis for the researchers failed to detect the extreme events from happening during year 2020 and 2022.

In recent years, a new branch of mathematical analysis, Topological Data Analysis (TDA) has emerged alongside technological advancements. Topological data analysis (TDA) is one of the popular approaches that utilized in data science field that lead to a notable success in the field of science and engineering nowadays. TDA focuses on understanding the shape, connectivity, and structure of data (Corbet et al., 2019). As data grow dramatically across the time especially in the era nowadays, it led to an increase in complexity in data which include the situation where the data is noisy and incomplete. TDA is the relatively new technique that

provide a wealth of new insights into the study of data (Otter et al., 2017). Unlike traditional methods, TDA applies concepts from topology and geometry to capture intricate relationships within complex datasets (Munch, 2017). This offers a novel approach and perspective for gaining deeper insights from the data. This makes TDA particularly suitable for analysing time series data, where patterns are often obscured by noise and fluctuations. One of the key techniques in TDA is called persistent homology which tracks the evolution of topological features such as connected components, loops, and voids across various scales, providing a multiscale view of the data (Schindler & Barahona, 2023). This method allows for the identification of key structures by examining the birth and death of topological features.

In conclusion, data analysis has become a crucial area of focus in both academic and professional fields. Academically, it has been integrated into computer science curricula as an essential discipline. Professionally, data analysis is pervasive across industries. The COVID-19 pandemic in 2020 further underscored the importance of data analytics, particularly in policymaking to address the exponential rise in new cases. For instance, data analysis has provided insights into the impact of non-pharmaceutical interventions, such as lockdowns and mask mandates, proposed by governments. Beyond policymaking, data analysis facilitates resource optimization. By predicting resource demands, such as ICU beds and ventilators, under different scenarios, data analytics has proven invaluable in crisis management. These examples highlight the vast potential of data analysis, extending far beyond initial expectations.

1.2 Problem Statement

Extreme events are characterized by unexpected occurrences that deviate significantly from typical or expected patterns. Data related to extreme events often exists in time series form; however, such data is typically nonlinear, non-stationary, noisy, and highly variable. These characteristics contribute to irregular occurrences and complex patterns within the data, posing significant challenges for analysis. Traditional time series analysis methods often struggle to effectively capture the underlying topology and structure of such data, thereby limiting their capacity to perform in-depth analysis. Consequently, researchers may face difficulties in providing early warnings for extreme events, such as financial crises in stock market data, flood disasters in meteorological data, and disease outbreaks in epidemiological data. Despite the inherent nonlinearity and noise present in time series data, many traditional statistical analysis techniques lack flexibility, as they often rely on the assumption that the analysed data conforms to specific statistical distributions (Yuhei et al., 2019).

1.3 Objective

1. Explore preliminary data pre-processing requirements and transformation techniques for applying topological data analysis on time series.
2. Integrate the commonly used topological features and descriptors in feature extraction of time series data.
3. Carry out the TDA-based time series analysis using on enhancing the early warning for the extreme event in time series.

1.4 Scope of Study

This research project focuses on detecting the extreme events and its early warning signal specifically in Covid-19 cases data within the context of Malaysia which cover 13 states and 3 federal territories. By examining the Covid-19 case trends across different states in Malaysia, this study aims to identify and analyse the anomalies or significant deviations from typical patterns that indicate extreme events such as the sudden spikes or surges case numbers.

The project gathers Covid-19 case data for each state in Malaysia from the start of the pandemic until June 2024, covering the entire range of case numbers over time to provide a thorough overview of the pandemic's progression in the country. In this project, the persistent homology from TDA approach is employed that is well-suited for extracting the shape of data in a complex, and high level of dimensional datasets.

This project outcome is intended for the use by the institute medical research centres Malaysia and public health authorities who monitor, analyse, and respond to Covid-19 cases. By providing early earning of potential extreme events in Covid-19 case trends, the project can support proactive intervention measures, improve pandemic preparedness, and inform resource allocation. The topological approach used here may contribute valuable insights into epidemiological patterns, making it a potential model for detecting extreme events in other infectious disease outbreaks beyond Covid-19.

1.5 Significant of Study

The significance of this project lies in its potential to enhance early warning systems for extreme events in historical Covid-19 case trends, especially within the complex and unpredictable context of infectious disease outbreaks. Traditional methods for analysing time series data often fall short when dealing with nonlinear, non-stationary, and highly fluctuating data, as is typical with pandemic case numbers. This study addresses these limitations by applying TDA specifically through persistent homology, to capture the underlying geometric and topological structures that represent significant changes in case trends.

By employing TDA, this project can identify critical shifts and anomalies in Covid-19 case data that may signal upcoming extreme events, such as rapid surges or decreases in cases. Such insights can be pivotal for public health authorities and medical research centres, allowing them to act proactively. Early detection of these patterns supports better preparedness and faster response times, which are crucial in managing resources, implementing timely interventions, and ultimately, mitigating the impact of Covid-19 surges on public health systems.

Furthermore, this project contributes a novel methodological approach to epidemiological analysis. By demonstrating the effectiveness of TDA in identifying extreme events in Covid-19 data, the findings have potential applicability beyond this pandemic. The framework could be adapted for other infectious disease outbreaks or even other fields with extreme events in time series data, such as finance or meteorology, making it a valuable tool for broad risk mitigation efforts. This innovative application of TDA can provide a new perspective on understanding complex data, aiding in the development of robust, data-driven early warning mechanisms.

1.6 Project Outcome

As a project outcome, several persistence diagrams will be constructed based on the point clouds derived from the datasets. These topological representations will provide a detailed analysis of the structural patterns within the data. Persistence diagrams, in particular, will enable the identification of significant topological features across varying scales. By utilizing these diagrams, persistence landscapes can be visualized and analysed to uncover critical insights into the data. The primary focus of this project is to detect extreme events and its early warning signal related to the COVID-19 pandemic in Malaysia. Through topological data analysis (TDA), the project aims to identify anomalies or irregularities in the time series data of COVID-19 cases, such as sudden surges or drastic declines. These patterns are indicative of extreme events, such as outbreak peaks, rapid declines due to effective interventions, or other significant epidemiological trends. The insights gained from this analysis will not only enhance our understanding of the dynamics of COVID-19 in Malaysia but also contribute to the development of early warning systems. Such systems can aid policymakers and healthcare providers in taking timely action to mitigate the impact of similar events in the future.

1.7 Brief Methodology

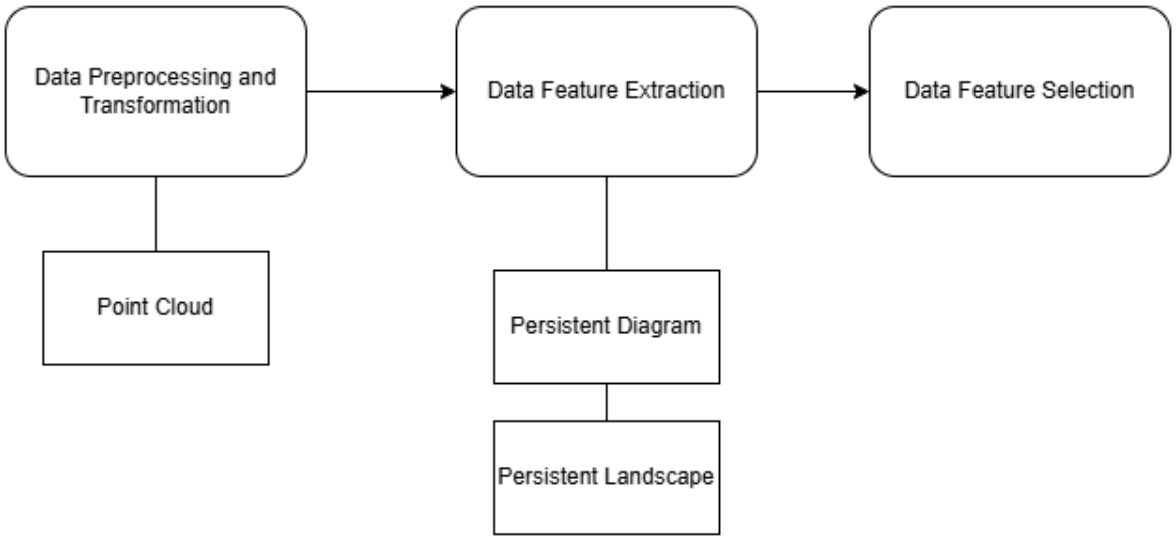

Figure 1.1: Flow Chart of Brief Methodology

The brief methodology for this research project which is shown in figure 1.1 consists of three main stages. In the first stage of this project, it focuses on the data pre-processing and transformation. This is the stage where the time series data are being prepared for topological analysis. This is because in order to apply the topological data analysis (TDA) effectively, the one-dimensional time series data needed to be transformed into a higher-dimensional space by using time delay embedding to reconstruct the state space of the data.

The next phase focuses on emphasizing the feature extraction through topological descriptors. In this research project, persistent homology approach is used. Through persistent homology, some of the topological features such as persistent diagram and persistent landscape will be computed on the transformed time series data.

The final stage is featuring selection. Feature selection will be conducted using standard metrics to evaluate the quality and effectiveness of the chosen features in achieving the desired outcomes.

1.7 Summary of Chapter

The chapter introduces the significance of analysing extreme events in time series data, emphasizing their non-linear and non-stationary nature, which complicates traditional statistical methods. Extreme events, such as stock market crises, natural disasters, and pandemics like COVID-19, are rare and impactful. Topological Data Analysis (TDA), specifically using persistent homology, is highlighted as a powerful method for uncovering structural patterns and anomalies in complex datasets. The study aims to enhance early detection of extreme events by leveraging TDA to analyse COVID-19 case trends in Malaysia. The research methodology involves data transformation, feature extraction, and integration of topological features into analytical frameworks to identify early warning signs. This approach offers potential applications across various fields and promises improved resource allocation, policymaking, and risk management.

Chapter Two

Literature Review

2.0 Overview

This chapter gives a detailed exploration of the past literacy of extreme events, topological data analysis, persistent homology, and related studies. In section 2.1, the background of the extreme events was introduced, the impact of extreme events and the ways to prepare for them from happening were discussed in the sub section of 2.1.1 and 2.1.2. In section 2.2, topological data analysis was introduced, followed by the usage of topological data analysis in section 2.2.1, introduction of point cloud in section 2.2.2, and introduction of time delay and embedding dimensions in section 2.2.3. Furthermore, persistent homology, which is one of the approaches in topological data analysis was introduced in section 2.3. In its sub section, persistent diagram, persistent barcode, the benefits and limitations of persistent homology were being discussed in the sequence of section 2.3.1, section 2.3.2, and section 2.3.3. Moreover, three related works were reviewed and discussed in section 2.4.1, section 2.4.2, and section 2.4.3. In the section 2.5, the chapter was summarized and concluded in this section.

2.1 Introduction to Extreme Events

According to the Rai at al. (2024), an extreme event can refer to an occurrence that is highly uncommon under normal circumstances within time series data, natural phenomena, or financial markets, yet has significant impacts when it does occur. Notable examples include the COVID-19 pandemic in 2019, where the sudden emergence of a viral mutation led to a resurgence of cases to unprecedented levels, and the financial crisis in 2008, during which massive stock market crashes caught individuals and institutions off guard. Historical examples of extreme events suggest that they are typically difficult to predict using traditional forecasting methods. This is because such events often do not conform to standard distribution assumptions and may be influenced by the nonlinear dynamics of complex systems.

2.1.1 Impact of Extreme Events

Looking back at historical events, many extreme events that have occurred around us have indeed caused significant negative impacts. Extreme events not only disrupt the social way of life that humans once knew, but they can also have significant effects on the economic health of nations. The impacts of these events depend on their type and duration.

For instances, in medical fields, according Aikebaier (2024), it has brought up mental health issues with significant increase in depression and anxiety worldwide in psychologically way when dealing with the extreme events of Covid-19. Sahimi et al. (2021) have mentioned that 11.1% of the healthcare workers in Malaysia have the current suicidal ideation during the Covid-19 pandemic and lockdown period. This is because being frontline personnel of Covid-19 pandemic, who frequently expose themselves in the life-and-death situation has higher potential of getting the severe psychological distress such as depression and anxiety symptoms

In addition, extreme events bring positive and negative impact in the economic and financial sectors. For example, during the extreme events in financial crisis 2008 some investors able to grab the golden opportunity to make a profit from there (Rai et al., 2024). In contrast, the example of negative impact of extreme events can be seen from the perspective of household income and poverty, the median monthly household income decreased by 11.3% in 2020, dropping from RM 5,873 to RM 5,029 when dealing with the extreme events of Covid-19 (Zakaria et al., 2023) which indirectly attributed to a sharp rise in the unemployment rate, which increased from 3.3% in 2019 to 4.5% in 2020, peaking at 5.1% during the first quarter of the year (Loong & Wan Amirah, 2022)

Moreover, extreme events are still able to bring a huge impact on the education sector. For instance, the school and universities in Malaysia have shifted to the remote online learning platform when dealing with the extreme event of Covid-19 which able to lead to several challenges for the students when the students don't have digital assets or a stable internet connection and lead to an increase in the burden of family and educational inequalities especially in the low-income household and rural areas (Loong and Wan Amirah, 2022).

2.1.2 Ways for Preparing the Extreme Events from Happening

Extreme events are unavoidable. Therefore, proactive preparation is the only effective approach to addressing them. Education and awareness in disaster management can play a crucial role in adequately preparing for future extreme events and mitigating associated risks. According to Nahid et al. (2018), integrating disaster risk reduction topics into school curricula can equip children with essential knowledge about hazards and enable them to respond effectively during disasters. Schools can serve as platforms to educate and cultivate awareness and preparedness among students and young people regarding the prevention of extreme events

(Hodson et al., 2024). This approach ensures that when extreme events occur, societal panic can be minimized.

Furthermore, in today's era of advanced technology, leveraging technological innovations can significantly enhance preparedness for extreme events. Peiman Alipour et al. (2019) explained that the integration of machine learning, blockchain, and database management technologies can generate intelligent hazard scenarios and propose optimal emergency evacuation plans with 3D notifications. Additionally, collaborative technologies can improve the agility of responses to extreme events, effectively addressing the management and coordination challenges inherent in current emergency management practices (Jefferson & Harrald, 2007).

2.2 Introduction to Topological Data Analysis

Topological data analysis (TDA) can be referred to a class of methods that gather the information from topological structures in the data that belong to topological space (Ravishanker & Chen, 2019). Otter et al. (2017) has mentioned that TDA is a field that intersect with various disciplines. It is a combination of data analysis, algebraic topology, computer science, computational geometry, and statistics field. TDA cleverly utilizes methods and results from geometry and topology to create a novel analytical tool for studying the qualitative aspects of data.

The emergence of TDA effectively addresses one of the key challenges faced by traditional statistical data analysis. Yuhei et al. (2019) has explained that in data analysis including with the help of machine learning, the conventional statistical analysis techniques have to make the assumptions where the data that being analysed need to follow some distribution. For example, when analysing the data of test scores, the shape of the data needs to be assumed to follow the shape of bell curve where most of the students obtained the grade at average while few students obtain very low or high scores. However, real-world data often deviates from these idealized patterns. In some cases, the data's structure or distribution may not align with the expected assumptions. Moreover, with the rise of big data, the complexity and patterns within data are becoming increasingly intricate. Traditional statistical data analysis, which relies heavily on fixed assumptions about data patterns, often struggles to fully capture the underlying information.

As Bukkuri et al. (2021) note, TDA focuses on studying the "shape" of data. By leveraging this unique characteristic, TDA offers a novel perspective to extract valuable insights that conventional statistical data analysis methods might overlook (Yuhei et al., 2019).

2.2.1 Usage of Topological Data Analysis

For the past few years, TDA has been applied in various type of sectors such as engineering technology, medical, finance, and economic (Corbet et al., 2019). For instance, TDA has been used as a technique to distinguish the healthy patients and diabetic patients by encoding the retinal image into the persistent diagram (Garside et al., 2019). Besides, TDA can be employed to identify the novel pathological phenotypes of asthma (Siddiqui et al., 2018). In the economic sector, TDA has being used as one of the analysis tools to determine the extreme events in the stock markets in America, Asia, Oceania, and Europe (Rai et al., 2024). Other than that, in engineering and technology field, Uray et al. (2024) have mentioned that different types of TDA such as mapper, persistent homology, and uniform manifold approximation and projection (UMAP) are being organized, and types of input being used according to the specific manufacturing process to improve efficiency.

2.2.2 Introduction to Point Cloud

Point cloud is the data structure that often used to represent the 3D geometry (Bello et al., 2020). According to the definition given by Hyun et al. (2021), a point cloud can be comprehended as a visual presentation of an object geometry. Point cloud can exist in various types of dimension space. For example, by using the advance technologies such as 3D scanners or LIDAR, a 3 dimension of a point cloud can be constructed by the x, y, and z coordinates with an optional feature vector assigned per each of the points. With the existence of point cloud, we are able to provide a rich and clear geometry, shape, and the scale information that complement the 2D images. In some contexts, point cloud sometimes is being referred to the state space reconstruction.

As mentioned above, the point cloud can be constructed with the help of technologies. However, Altındiş et al. (2021) has explained that the point cloud can also be constructed

through a mapping method called time delay embedding method. In general, constructing a point cloud data diagram or state space reconstruction is a necessary process that needs to be carried out before carrying out the data analysis of a time series data sets (Krakovská et al., 2015). This approach is primarily designed to preserve the essential geometric and dynamical invariants, such as the fractal dimension of the attractor, entropies, and sensitivity to the initial conditions.

Krakovská et al. (2015) has explained that technically, by using the time delay embedding method, reconstructing the state space or constructing the point cloud requires the decision making or selecting the optimal time delay, and the dimensional state space. Altındiş et al. (2021) mentioned that the coordinates of the data points which plotted in the point cloud can be shown in the Eq 2.1:

$$y(t) = x(t), x(t + \Delta T), x(t + 2\Delta T), \dots, x(t + (M - 1)\Delta T) \quad (2.1)$$

2.2.3 Introduction to Time Delay and Embedding Dimensions

Krakovská et al. (2015) has defined that the embedding dimension, M is the size of the state space being constructed and time delay, ΔT is the value of time shift between coordinates. Altındiş et al. (2021) has stated that one of the proper way to determine M is using the false nearest neighbor (FNN) method. The determination of M is crucial. This is because if the value of M is lower than optimum value, a denser state space will be constructed which will lead to the overlapping of the state space features and affect the analytic result. In contrast, if the value of M is higher than optimal value, it will cause sparsity in the reconstructed state space as each of the points of the point cloud will be located far from each other.

False nearest neighbor (FNN) is a method for determining the optimal embedding dimension, M . Krakovská et al. (2015) elucidated that the way the FNN method works is based

on the principle that two points in proximity can maintain their proximity relationship within a sufficiently dimensioned embedding space. Technically, FNN starts with the low dimension of 1. By increasing the dimension, the distance between two points is analysed to determine whether the points become more dispersed within the higher dimensional state space when extra dimension is added in. In FNN, the distance between two points is calculated based on Euclidean distance formula (Altındiş et al., 2021), as given in Eq (2.2)

$$Rm = \sqrt{\sum_{k=0}^{m-1} \left(x(t + k\Delta T) - x_{nn}(t + k\Delta T)\right)^2} \tag{2.2}$$

where $x(t + k\Delta T)$ is the coordinates of data point in the point cloud while x_{nn} is the point's coordinates which has the nearest distance to $x(t + k\Delta T)$ points. From that the distance between data points and nearest to their data points, R_m can be obtained, an embedding dimension ratio can be calculated by using the Eq 2.3:

$$\frac{|R_{m+1} - R_m|}{R_m} \tag{2.3}$$

The ratio obtained can be converted into percentage as the optimal embedding dimension of the FNN percentage is below 1%.

2.3 Introduction to Persistent Homology

Persistent homology is one of the methods of TDA approach for measuring the topological features of shape (Ravishanker & Chen, 2019). Otter et al. (2017) have explained that based on the topological view, there is no information that can be extracted from the point cloud data. To address the limitations of point cloud data, persistent homology can be utilized to transform the data points into simplicial complexes, enabling the analysis of the topological structure of a space at different spatial resolutions (Ravishanker & Chen, 2019).

2.3.1 Introduction to Persistent Diagram

According to Ravishanker & Chen, (2019), the persistent diagram can be constructed from the point cloud data. Assume d , which indicates that the dimension of the point cloud is 2, then $P = \{v_i : i = 1, 2, 3, \dots, N\}$ where v refer to a single data points among the point cloud, i indicate that the index of each of the data points, and P is the collection of all the points v_i in the data sets. All the data points in the point cloud data can be measured by using the Euclidean distance and stored in the distance matrix, D_E where D_E is a matrix, $D_E = \{D_E = (v_i, v_j)\} for\ i, j = 1,2, \dots, N$

Altındiş et al. (2021) explained that Vietoris-Rips filtration is one of the popular approaches to extract the homologies of the state space. Around each of the points in the point cloud, a disk radius started to grow from the initial radii zero. In Vietoris-Rips filtration approach, a predetermined value of τ_{max} need to be determined. The disk radii will growth until the τ_{max} is reached. As the disk radii start to grow through filtration, two of the disks may overlap, a new connection may form. However, when one of the disk embraced to the other disk's centre, the connection between them is dead as shown in figure 2.1.

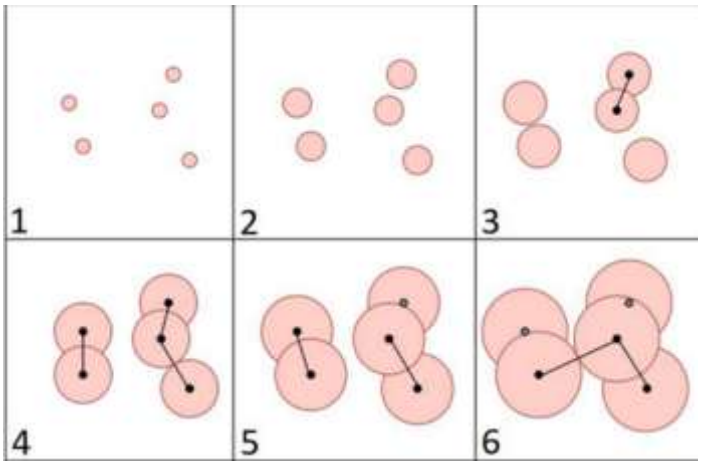

Figure 2.1: Connection Formed Through Filtration (Altındiş et al., 2021)

The value of τ_{max} is critical to determine. This is because according to Altındiş et al. (2021), if the chosen τ_{max} is too large, it will lead to the result of the persistent homology becoming meaningless as the feature related homologies disappear. In contrast, if the value of τ_{max} is too small, the noise might appear as persistent homology. Therefore, Altındiş et al. (2021) mentioned that the value of τ_{max} should be enough large to remove the noise and small enough to capture the homologies features.

The $Betti$ numbers, β_n can be used to represent and count the homologies features that were found after the filtration where β_0 represent the connected components, β_1 represent the number of connections between the points (edges), β_2 indicate the void and surface, and so on (Altındiş et al., 2021).

2.3.2 Persistent Barcode

Through the Vietoris-Rips filtration process, the connection between the points (edge) will emerged and die. To visualize the resolution of the state space, a persistent barcode can be constructed (Altındiş et al., 2021). From the persistent barcode, the lifetime of the features can be observed (Otter et al., 2017). According to Otter et al. (2017), the left endpoint of the interval

can be presented as the birth of the features meanwhile the endpoint at the right of the interval indicates the death of the same topological feature. If the lifetime of a homology is short, this can refer to noise and artefacts. In contrast, if homology has a long lifetime, it can be referred to as persistent homologies.

2.3.3 Benefits and Limitations of Persistent Homology

Based on the Bukkuri et al. (2021), persistent homology approach is stable when handling the noise. Moreover, from the perspective of coverage, the strengths of persistent homology can be further accentuated. A sufficient number of data points is needed to construct the shape structure of an object. If the points are insufficient to cover the object, persistent homology is able to detect it as the short bar which might be considered as noise. In higher dimension, the structure of the data that needed to cover the shape of the object become complicated. This is because more things have to be considered such as direction, and the degree of freedom. From that the difficulty of enhancing the shape coverage using the data will arise. The most quintessential example is the insufficient coverage to recover the topology of torus (a doughnut like shape) characterized by loops or holes. This results in a barcode that differs significantly from the one obtained through uniformly under-sampling points from a sphere, which lacks holes or loops. From here, it shows that persistent homology is useful with dealing with incomplete data which allows us to compare and differentiate the datasets without perfect coverage.

Other than that, Bukkuri et al. (2021) have mentioned about the limitation of the persistent homology which is the curse of dimensionality. The noise bring up a huge effects on the Euclidean distance when used in the construction of point cloud and persistent diagram in a high dimensional space. For instance, if the coordinate of data point is deviated 0.01 units

from the origin where it suppose should be in due to the noise, then in total the data point is$\sum_{i=1}^{n} 0.01^2$ deviated in the space state. If there is 10000 points in the state space, the data points is perturbed by the distance$\sum_{i=1}^{10000} 0.01^2 = 1$ (Bukkuri et al., 2021). From the example, even a small unit in the perturbations in a single unit of data point, it still able to lead to massive distortions in overall distance.

2.4 Related Studies

This section examines three related studies to analyse and compare the methodologies employed and the results achieved.

2.4.1 Related Work 1: Using persistent homology as pre-processing of early warning signal for critical transition in flood

This research is conducted by Syed Musa et al. (2021) in which they proposed the use of persistent homology to act as a pre-processing procedure to achieve the flood early warning system (FLEWS) through critical slowing down theory (CSD). Flooding can be considered as one of the destructive natural disasters. A severe flood can lead to significant loss of innocent citizen lives and extensive damage of property. There are several previous studies that have mentioned that the presence of early warning signal can indirectly serve as indicator of the system's critical tipping point (Scheffer et al., 2009). According to Wissel (1984), these indicators can be related to the theory of CSD.

The flow chart of the methodology is shown in figure 2.2. They collected data of daily water levels from the Guillemard Bridge station, Kelantan River, Malaysia started from 1st January 2020 until 13th October 2010. As a result, the FLEWS created six false alarms which can be referred to the signal is hardly approach to the critical point. Thus, in this research, persistent homology will be used to extract the topological features of the water level of Kelantan River to stimulate a new signal. By using the CSD indicators, the new signal was calculated to obtain the flood early warning signals. Then quantile estimation was carried out to obtain the dates for flood signal where quantile estimation is one of the techniques for researching the uncommon occurrences and extreme values. In the end, the result of FLEWS

by using the persistent homology as a pre-processing step and the result of FLEWS directly using the water level data were compared.

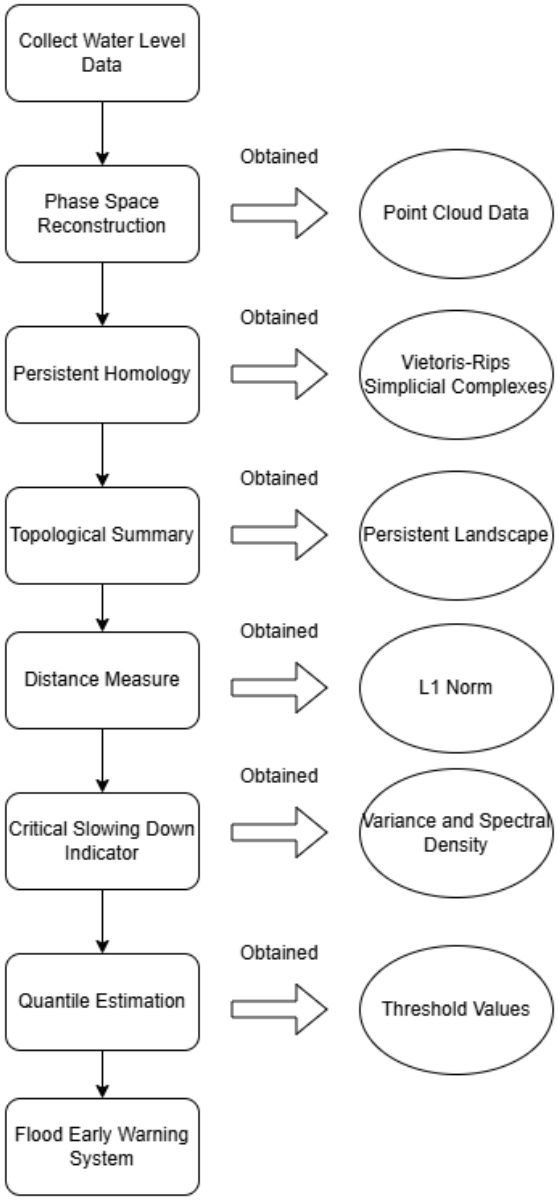

Figure 2.2: Flow Chart of Methodology in Related Work 1 (Syed Musa et al., 2021)

In this work the daily water level data were collected and prepared via Takens' embedding theorem as the phase space can be reconstructed by using the time series data and produce the point cloud data. The constructed phase space can be represented as:

$$x_n(m, \tau) = x_n, x_{n+\tau}, \ldots, x_{n+(m-1)\tau}$$

where m and τ represent the embedding dimension and time delay parameter. They stated that the value of m is set to be 2 while the value of τ is considered as 1 due to this value provides a good analytical result based on their experience. From that, a 2-dimensional point cloud is obtained.

The simplicial complexes can be constructed after obtaining the point cloud data diagram through filtration (Edelsbrunner, 2022). Due to the difficulty of persistent diagram to work with machine learning and statistics, persistent landscape is chosen as a suitable topological summary for carrying out further analysis (Bubenik & Dlotko, 2017). This is because a persistent landscape leverages the vector space structure of its underlying function space.

According to Bubenik (2015), the persistent landscape achieves a stable state with the respect of L^p distance when the $1 \leq p \leq \infty$. This indicates the small changes in the input will only cause small changes in the persistence landscape. In this research, Syed Musa et al. (2021) stated that they only focused on L^1 norm for further computation to measure the difference in persistence landscape by summing up all the absolute differences between values.

In their result analysis, ten days of sliding windows and one day of sliding step is used to calculate the variance and average spectral density at low frequencies of the L^1 norm time series. This finding verifies that the signals obtained from persistent homology exhibits CSD since the time series of variance and average increase around the flood events. After that, quantile estimation is used to determine the dates.

As a result, the optimum quantile is considered as 12% as the threshold as it is the lowest quantile that produces the same result. The optimum quantile of 12% provide the variance of 4.8991 and 2.8203 for the average spectral density of the twelve actual flood events while the optimum quantile of FLEWS directly using the water level, the variance obtained is 0.4742 and 0.7909 for the average spectral density. By applying the persistent homology approach in FLEWS, it creates fewer false alarms where FLEWS stimulate four false alarm which is 25% in overall when compared to FLEWS with water level which stimulate 6 false alarms which is 33% in overall (Syed Musa et al., 2020).

2.4.2 Related Work 2: Identifying Extreme Events in the Stock Market: A Topological Data Analysis

The research conducted by Rai et al. (2024) is mainly focus on detecting extreme events happening in the stock market by using TDA. The method of L^1, L^2 norms and Wassertein distance, W_b i were used in identifying the extreme events of the stock markets of 2008 financial crisis and the pandemic of Covid-19. During the year 1929, the stock market has experienced the depression, followed by the tech bubble of 2000, the financial crisis in 2008, and pandemic of Covid-19 in 2020. They have mentioned that these incidents have provide the opportunity for those investors to make profits, however it also has the risk for investor to lose their capital. Therefore, the study of extreme events in stock market is crucial.

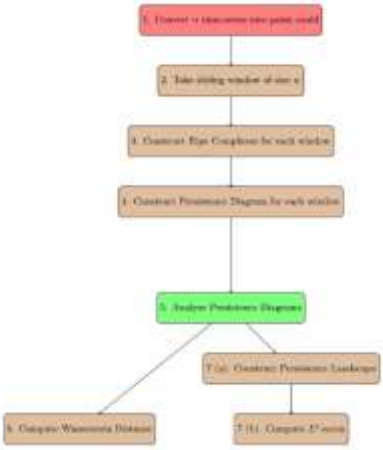

Figure 2.3: Flow Chart of Methodology in Related Work 2 (Rai et al., 2024)

The methodology used in the research is shown in figure 2.3. Initially, a point cloud dataset is being constructed by using a set of data points collected in the dimensional Euclidean space, R^d. Vietoris-Rips complex can be formed and is used to capture the connectivity of the data points at different ε. By increasing the value of ε through filtration, the simplicial complexes can include more number of simplices (Chazal & Michel, 2021). From that, the homology group can capture different dimensional topological features at different ε such as the 0-dimensional homology group, H_0 can be represent as connected components, 1 dimensional homology group,$H_1$is a loop, and 2 dimensional homology group,H_2 can be a void. From that, the birth and death of the topological features can be identified and the persistent diagram can be constructed.

Once obtained the persistent diagram, the Wassertein Distance, W_b can be calculated. They stated that the Wassertein Distance, W_b is able to provide the similarity between the two persistent diagrams for comparing the topological features. Besides, that the L^1 norms and L^2 norms can be calculated from the persistent landscape where persistent landscape can be constructed based on persistent diagram.

There are four continents being selected which are North-South America, Europe, Asia, and Oceania to identify the extreme events. Moreover, the impact of the Covid-19 to different sectors such as pharmaceuticals, banking, metals, automobiles, and fast-moving consumer goods are being analysed. All the data utilized in this study were sourced from the Yahoo Finance website. For the 2008 financial crisis, the dataset comprises daily closing prices from January 1, 2006, to December 31, 2010. In the case of the market crash caused by the COVID-19 pandemic, the data span the period from January 1, 2019, to December 31, 2022.

The result obtained from this research can be divided into three parts which are extreme events of stock markets due to the financial crisis in 2008, the extreme events of stock markets due to the Covid-19 in 2019, and impact of Covid-19 on different Indian sectors. They explained that with the help of threshold of W_d from the W_d diagram and threshold of L^p norms from L^p diagram, extreme events can be detected by using the formula $\mu + 4\sigma$ where μ is the mean for W_d or L^p norms and σ is the standard deviation of W_d or L^p.

Table 2.1: Table represent the threshold of W_d, L^1 norms, and L^2 norms of different continents

(Rai et al., 2024)

Continents	2008 Financial Crisis		
	L^1 norms	L^2 norms	W_d
America	6.68×10^{-5}	5.02×10^{-4}	0.41
Asia	1.00×10^{-4}	6.33×10^{-4}	0.65
Europe	2.84×10^{-5}	2.71×10^{-4}	0.48
Oceania	3.00×10^{-6}	5.41×10^{-5}	0.37

From table 2.1, a clear spike detected between May 2008 until July 2009 from both L^1 norms graph and L^2 norms graph where the norms values exceed predefined threshold values which are 2.84×10^{-5} and 2.71×10^{-4}. Meanwhile, extreme events in America were also

detected from both L^1 norms graph and L^2 norms graph in the interval of July 2008 to October 2009. Asia and Oceania also obtained a similar result.

For the W_d, the threshold of W_d has shown in the table 3.1 above. The W_d values of America from February 2008 to March 2009 has exceed 0.41 which can be considered as extreme events meanwhile for the Europe is fall between December 2008 untili December 2009.

Table 2.2: Table represent the threshold of W_d, L^1norms, and L^2 norms of different continents (Rai et al., 2024)

Continents	2019 Covid-19		
	L^1norms	L^2 norms	W_d
America	4.27×10^{-5}	3.65×10^{-4}	0.47
Asia	4.27×10^{-5}	3.70×10^{-4}	0.38
Europe	9.66×10^{-6}	1.26×10^{-4}	0.43
Oceania	1.17×10^{-5}	1.36×10^{-4}	0.34

For the second category of result which shown in table 2.2 which is the extreme events of stock market due to Covid-19, they have found out that both norms' values of Asia have exceeded the threshold L^1 norm and L^2 norm that stated in the table 2 and reached at the peak during March 2020. Meanwhile for Europe, from January 2020 until January 2021, both of the norm values have exceeded the 9.66×10^{-6} and 1.26×10^{-4} which can detect as extreme event. Similarly, America and Oceania also obtained the same result.

For W_d of Asia has overstepped the threshold value, 0.38 when come to December 2019 until August 2020 while in Europe, it fall be as the same range as Asia where W_d values of Europe exceeded its threshold, 0.43.

Table 2.3: Table represent the threshold of W_d , L^1 norms, and L^2 norms of different continents (Rai et al., 2024)

Sectors	Covid-19		
	L^1 norms	L^2 norms	W_d
Bank	3.01×10^{-5}	2.72×10^{-4}	1.48
Pharmaceutical	3.99×10^{-5}	3.38×10^{-4}	1.86
Automobile	2.75×10^{-5}	2.54×10^{-4}	1.17
Metal	5.54×10^{-5}	4.14×10^{-4}	1.44
FMCG	3.00×10^{-5}	2.46×10^{-4}	1.37

From table 2.3, a spike is observed in norms from January 2020 until January 2021 in Bank sector and pharmaceutical sector. They pointed out that the spike of Bank sector is slightly higher than pharmaceutical sector, this indicates that Bank sector has been experiencing in s stress period for a long time than pharmaceutical sector which take shorter time to recover from the market crash,

2.4.3 Related Work 3: Nonlinear time series analysis of state-wise COVID-19 in Malaysia using wavelet and persistent homology

The research conducted by Phang et al. (2024) is to identify the qualitative characteristic that underpins the daily infection of Covid-19 positive cases in Malaysia's 13 states and three federal territories. They mentioned that although the COVID-19 pandemic has subsided, this does not preclude the possibility of its resurgence. This reflects that analyzing the new daily infection cases remains crucial. By utilizing the time series analysis approach, it is able to facilitate the generating of summaries, the identification of patterns and similarities, and identify the unusual trajectories.

This research focused on analyzing the Covid-19 daily new cases in Malaysia from January 2020 until June 2024 by using two data analysis method which are wavelet and

topological data analysis. These two approaches can provide a clear and comprehensive view on the trends of Covid-10 and variances in Malaysia.

After collecting the data sets from GitHub, some pre-processing steps need to be conducted which is the point cloud reconstruction via taken embedding theorem. There are 2 parameters needed to be considered in that method which are time delayed, τ and embedding dimension, m . Determination of the value of time delayed, τ is critical and various methods can be used to determine the value such as autocorrelation, average mutual information (AMI), and correlation integral (Sivakumar & Deepthi, 2021). In this research, the AMI approach is used.

Moreover, the embedding dimension is selected to be 2 as it is not feasible to visualize the point cloud in the dimension higher than 3. Once the data point cloud was being constructed, topological data analysis can be applied for further analysis. Due to most of the disease time series data having cyclical characteristics, persistent homology is used instead of Mapper algorithm.

Furthermore, Wasserstein distance, W_d is used to compare the state-wise time series Covid-19 data's topological properties during the research of the investigation. If the W_d obtained is large, this indicates that the two-point cloud does not share enough topological features.

Moreover, further analysis can be performed from the persistent diagram of each state in Malaysia to perform hierarchical clustering. Four features were selected for each state which are the emergence of single remaining connected component encompassing all data points in the phase space reconstruction, the longest lifetime of one-dimensional loop, the second longest lifetime of one-dimensional loop, and the sum of all lifetimes of one-dimensional loop.

Nur et al. (2022) mentioned that the topological-based clustering is validated based on 3 performance metrics which are rand index (RI), normalized mutual information (NMI), and normalized variation of information (NVI). All these three metrics are measured in the range of 0 to 1. For RI and NMI, 0 indicates that the worst clustering while 1 represents good clustering meanwhile for NVI is used to measure the dissimilarity between two clustering result.

In their result, the calculated time delay of each state using the data without smoothing falls in the range [9, 19] meanwhile using smoothing data is within [16, 30] except for Pahang. The analysis is further conducted by using a 7-day rolling average to smooth the datasets. They concluded that not all the state in Malaysia was suitable to use 7-day rolling average to smooth the datasets. For instance, three of the states (Labuan, Perlis, and Kedah) showed the circular path in the point cloud diagram and two significant one dimensional topological circles in the persistent diagram which indicates the two dominants pandemic waves in Covid-19. However, when come to Melaka, Selangor, Perak, and Terengganu, it shows less effectively when using the 7-days rolling average. Thus 15-days rolling average applied as a modification to smooth the dataset, enabling the other states to identify the dominant pandemic waves more effectively in point cloud and persistent diagram.

Furthermore, the Wasserstein distance is measured among all the 13 states and 3 federal territories in Malaysia. They found that Selangor has the largest Wasserstein distance meanwhile Perlis, Labuan, Melaka, and Putrajaya have the lowest Wasserstein distance. This can be related to four of the states have the low population, Selangor has the largest population density in Malaysia. From that, it indicates a substantial relationship between the population density and the spread of virus which was examined by Wong et al. (2023).

Last but not least, 13 states and 3 federal territories can be categories into 4 clusters. Due to the largest population density and number of Covid-19 cases, Selangor is categories as one group. The second group includes Sabah, Sarawak, Johor, and Kuala Lumpur due to their high population states. For Labuan, Perlis, Putrajaya, Terengganu, Pahang, and Melaka were categorized in third group which has the intermediate population and the rest of states were categorized as forth group due to low population size.

2.4.4 Summaries of Finding

Table 2.4: Summary of 3 Related Findings

	Related Work 1	Related Work 2	Related Work 3
Objective of Study	Comparing the result with FLEWS constructed directly from water level and with FLEWS constructed via persistent homology of Kelantan River, Malaysia.	Detecting the extreme events of stock markets of 2008 financial crisis, and the pandemic Covid-19.	Detecting the qualitative properties of Covid-19 infectious cases of 13 states and 3 federal territories in Malaysia.
Methodology Used	• Taken Embedding Theorem • Persistent Homology • Critical Slowing Down Indicators • Quantile Estimation • Flood Early Warning System	• Converted the time series data into point cloud data. • Constructed Rips complexes. • Constructed Persistent Diagram. • Constructed Persistent Landscape. • Computed L^p norms. • Computed Wasserstein Distance	• Conducted Time delay embedding theorem to construct the phase space reconstruction • Performed TDA to obtained persistent diagram • Wasserstein distance is measure based on persistence diagram

			• Hierarchical clustering is performed to categories the groups.
Result	FLEWS constructed via persistent homology stimulates fewer false alarms with a rate of 25% compared to FLEWS constructed directly from water level with higher rate of 33.33%.	The extreme events of America, Asia, Oceania, and Europe detected where the values of Wasserstein distance and norm values exceed the predefined threshold in financial 2008 and Covid-19. Besides, the extreme events of all the sectors, bank, Pharmaceutical, automobile, metal, and FMCG also detected where the Wasserstein distance and norm values exceed the threshold.	Selangor obtained the largest Wasserstein distance meanwhile Perlis, Labuan, Melaka, and Putrajaya have the lowest Wasserstein distance. This can be related to the population density and infectious cases. Moreover, Selangor is categories as one group, Sabah, Sarawak, Johor, and Kuala Lumpur as second group, third group include Labuan, Perlis, Putrajaya, Terengganu, Pahang, and Melaka, meanwhile the last group contain Negeri Sembilan, Perak, Kedah, Pulau Pinang, and Kelantan.

From the summary of Table 2.4, although the research fields differ, the methodologies appear to share similarities. Specifically, the collected data is typically constructed into a point cloud, which serves as the foundation for extracting features using persistent homology. Subsequently, the Wasserstein distance and norm values are computed to facilitate further analysis and derive insights from the data.

2.5 Summary of Chapter

This chapter explored the study of extreme events, TDA, and related methodologies like point clouds and persistent homology. Extreme events, such as the COVID-19 pandemic and financial crises, are defined as rare occurrences with significant impacts that defy traditional forecasting methods. These events disrupt mental health, economies, and lifestyles, emphasizing the importance of proactive disaster preparedness through education and technological innovations. TDA, a method for analysing the topological structure of data, addresses the limitations of conventional statistical techniques, particularly with complex, non-standard data patterns. Applications of TDA span various fields, including healthcare, economics, and engineering. Point clouds, as 3D geometric representations, play a crucial role in state-space reconstruction for time-series data analysis. Persistent homology, a TDA approach, extracts topological features from data, offering tools like Vietoris-Rips filtration and persistent barcodes to distinguish meaningful data structures from noise. While persistent homology is robust to noise and effective with incomplete datasets, it faces challenges such as dimensionality issues and susceptibility to noise in high-dimensional spaces. This review underscores the potential of these advanced techniques in understanding complex data and preparing for extreme events. Last but not least, 3 related works were studied, the methodology used and result obtained were compared.

Chapter Three

Methodology

3.0 Overview

This chapter outlines the research methodology which show in figure 3.1 for detecting the extreme events of the Covid-19 cases in all 13 states and 3 federal territories in Malaysia by using topological data analysis approach. Topological data analysis is known for its ability to provide the insight about the shape of data (El-Yaagoubi et al., 2023). The data collection process is comprehensively detailed in Section 3.2. Subsequently, Section 3.3 elaborates on the pre-processing steps undertaken to manipulate the datasets for point cloud formation. Furthermore, the methodologies for feature extraction are thoroughly discussed in section 3.4, while Section 3.5 focuses on the process of feature selection.

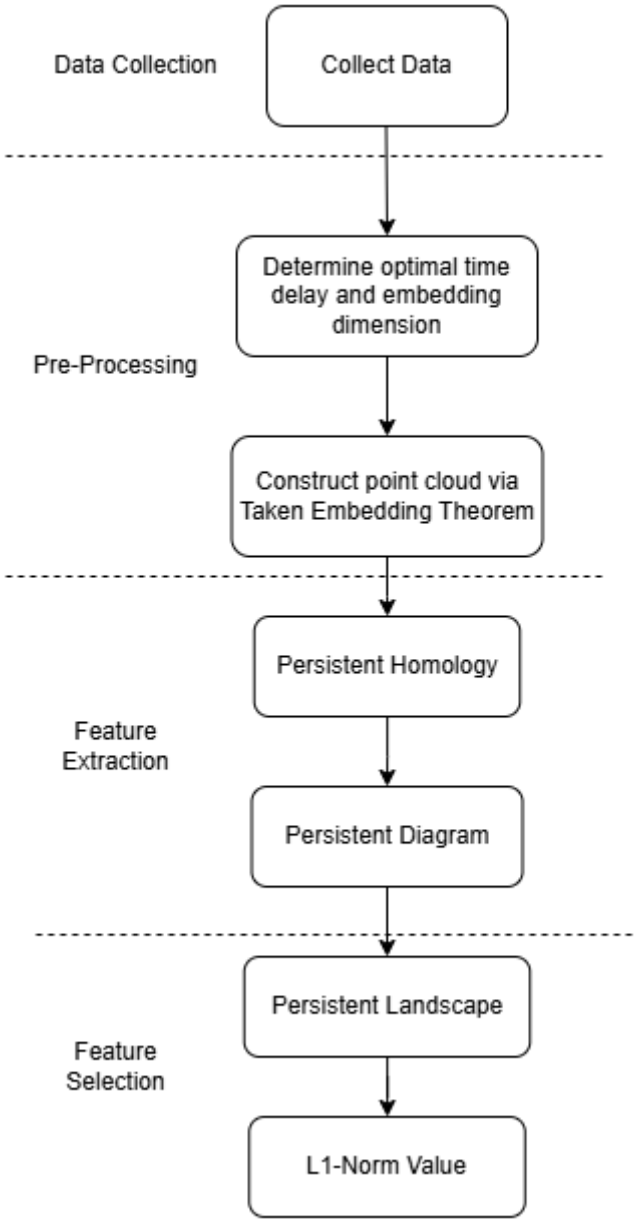

Figure 3.1: Flow Chart of Methodology

3.1 Data Collection

Ministry of Health Malaysia has provided an open-source official Covid-19 data on GitHub platform. Based on the description, it stated that all the datasets are provided by crisis preparedness and response center (CPRC) system, *"Makmal Kesihatan Awam Kebangsaan"* and My Sejahtera. In their GitHub pose, they provide a few datasets. For instance, the

customers check in date and time from MySejahtera, the vaccination of booster received for each of the state in Malaysia, and the new cases of Covid-19 in Malaysia. For this project, the daily new cases of infectious cases are considered. Therefore, the data sets of new cases of Covid-19 of each of the states and federal territories in Malaysia from 25th January 2020 until 1st June 2024 was obtained from the GitHub.

3.2 Data Pre-processing

The data collected from GitHub can be considered as raw data. Therefore, pre-processing steps are essential to manage and refine the data, minimizing noise that could adversely impact the quality of the data and subsequent analyses (Yin et al., 2022). More specifically, the data can be regarded as one-dimensional time series data. For TDA, it requires a minimum of two-dimensional data. Hence, the data dimension needs to be augmented by using the phase space reconstruction.

3.2.1 Phase Space Reconstruction

The data sets downloaded from the GitHub repository are in the non-linear time series form. Tan et al. (2023) explained that this had increased the difficulty of fully observed the entire system due to the hidden dependencies such as delay in symptoms onset and reporting and some latent variables such as contact patterns, virus mutations, and human immunity. From that, the raw data have been segmented into different segments through segmentation with a range of 6 months for all the 13 states and 3 federal territories. The purpose is to focus on the specific behaviours and capture the localized patterns that vary across the segments.

Each segment of one-dimensional non-linear time series data is transformed into higher dimensional data cloud by embedding techniques, $\Psi: R \rightarrow R^m$, where R is the real number of data sets and m is the dimension. In this project, Takens embedding theorem is considered to reconstruct the phase state. The time delay and the dimension need to be considered as the parameter for this theorem (Altındiş et al., 2021). Mutual information approach is used to determine the optimum time delay for each of the segments (Altındiş et al., 2021).

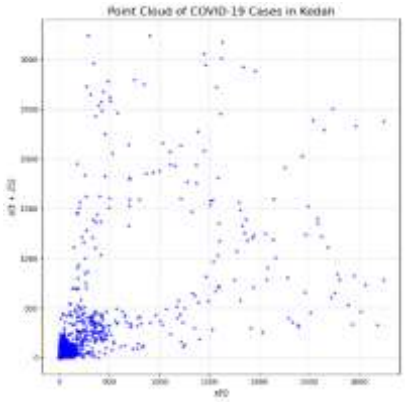

Figure 3.2: Sample of 2D point cloud

For the dimension, m, 2 is chosen to construct the point cloud to simplify the analysis. With these two parameters, the delay vector $\big(x(t),\ x(t+\tau)\big)$ can be obtained and used to construct the 2D point cloud. An example of this 2D point cloud is given in figure 3.2.

3.3 Features Extraction

From the 2-dimensional point cloud, the features extraction can be performed to extract the feature of the data by using persistent homology approach. In topology data analysis, persistent homology is chosen as the approach to analyse the 2-dimensional point cloud among the approaches such as Mapper and Reeb graph due to the time series dataset for this project having the cyclic element (Phang et al., 2024).

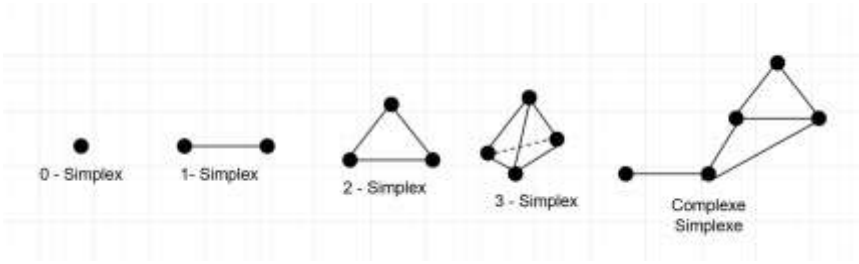

Figure 3.3: Types of simplexes

Persistent homology operates by analysing the evolution of topological features within a point cloud as a parameter, often referred it as a distance value, $\in$ increase. In this project of 2-dimensional point cloud, each of the data point is treated as the centre of a geometric circle. As the radius of these circle increases with the growth of $\in$, the coverage area of each of the circle expands progressively. When two of the edges of the circle intersect overlapped, 1-simplex is formed, representing an edge that connects the two points which shown in figure 3.3.

As $\in$ continues to increase, more intersections occur, leading to the formation of higher-dimensional simplices. For instance, when three edges close into a triangular connection, a 2-simplex is created. This process of systematically growing $\in$ and oberving the formation and evolution of simplices is known as filtration which shown in figure 3.4. The filtration process captures the birth and death of topological features across the scales which is fundamental to the computation of persistent homology.

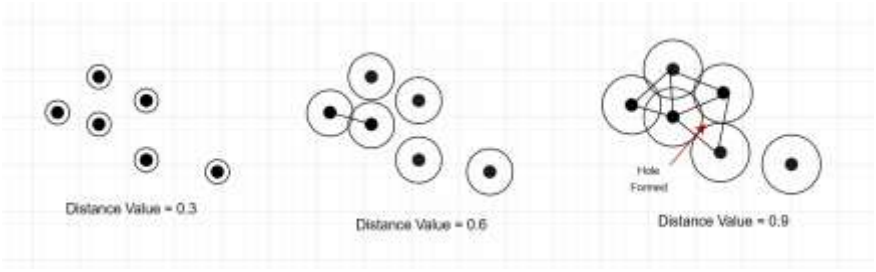

Figure 3.4: Filtration process

During the filtration process, topological holes are formed as the simplicial complex evolves. Specifically, when the simplexes are created, the resulting configurations can give rise to holes within the complex. For instance, when a 1-simplex (an edge) is formed, the H_0 feature corresponding to the connected components is identified. As the parameter $\in$ continues to increase and more complex simplexes are constructed, higher dimensional features, such as H_1 (loops), emerge.

Throughout the filtration, new holes are continuously created, while some existing holes die or disappear as the simplicial complex becomes increasingly connected. The dynamic appearance and disappearance of these topological features are systematically counted and recorded.

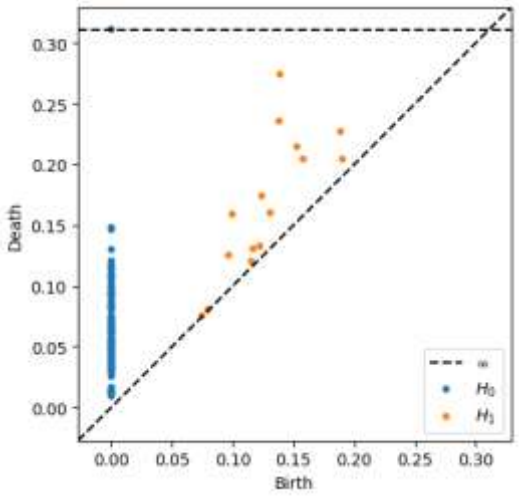

Figure 3.5: Sample of Persistent Diagram

With the help of ripser and persim library in Python, a persistent diagram which shown in figure 3.5 can be constructed from the 2-dimensional point while and persistent barcode which shown in figure 3.6 can be constructed based on the lifespan of each of the hole that formed during the filtration.

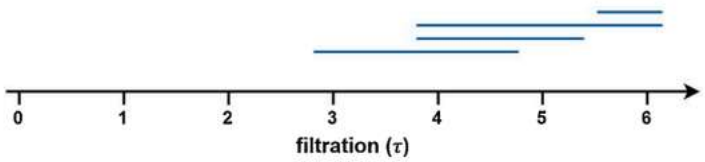

Figure 3.6: Persistent barcode

3.4 Features Selection

Apart from that, the persistent diagram can be transformed into persistent landscapes to perform the statistical analysis (Rai et al., 2024). This is because persistent diagrams are not naturally in a vector space, which make it hard to apply certain mathematical tools on it.

According to Gidea et al. (2020), the persistent diagram can be embedded into a Banach space. For each of the birth-death point in the persistent diagram, $(\alpha, \beta) \in P_n$, where P_n is the persistent diagram, the piecewise linear function can be defined as in Eq (3.1):

$$f(\alpha,\beta)(x) = \begin{cases} x - \alpha, if\ x \epsilon \left(\alpha, \dfrac{\alpha+\beta}{2}\right]; \\ -x + \beta, if\ x \epsilon \left(\dfrac{\alpha+\beta}{2}, \beta\right); \\ if\ x\ \notin (\alpha, \beta\ ). \end{cases} \tag{3.1}$$

To construct the persistent landscape, a sequence of functions $\lambda = \lambda_k$ where $\lambda_k : \mathbb{R} \to [0,1]$ is given by Eq (3.2)

$$\lambda_i(x) = i - \max\{f_{(\alpha,\beta)}(x) |\ (\alpha,\beta) \in P_n\} \tag{3.2}$$

where $i - max$ is the ith largest value of a function. With the Eq (3.2), persistent landscape can be formed as a subset of the Banach space $L^p(\mathbb{N} \times \mathbb{R})$, where the norm of λ is given by,Eq (3.3)

$$\left|\left|\lambda\right|\right|_p = \left(\sum_{i=1}^{\infty} \left|\left|\lambda_i\right|\right|_p^p\right)^{\frac{1}{p}} \tag{3.3}$$

Persistent landscape is the collection of piecewise linear function $\lambda_k(x)$ which derived from the persistent diagram.

3.5 Summary of Chapter

Chapter 3 details the methodology for analysing extreme events in Covid-19 case data across Malaysia using TDA approach. Data was collected from the Ministry of Health Malaysia's GitHub repository, encompassing daily new cases from January 2020 to June 2024. The pre-processing phase segmented the time series data into six-month intervals and transformed it into a two-dimensional point cloud using Takens' embedding theorem. Persistent homology was applied to extract topological features, analysing the evolution of connected components and loops through a filtration process. These features were visualized using persistent barcodes and diagrams, offering insights into the structural patterns of the data. Feature selection was then conducted to identify significant attributes, reduce dimensionality, and improve the interpretability and performance of subsequent analyses. This comprehensive methodology provides a robust framework for detecting extreme events in pandemic data through advanced topological techniques.

Chapter 4

Result and Analysis

4.0 Overview

This chapter provides a detailed analysis of the results obtained through multiple perspectives from point clouds, persistent diagrams, persistent landscapes to L-norm graph so as to offer a comprehensive understanding of the underlying patterns and structures present within the data analyzed.

4.1 An Overview of Raw Data

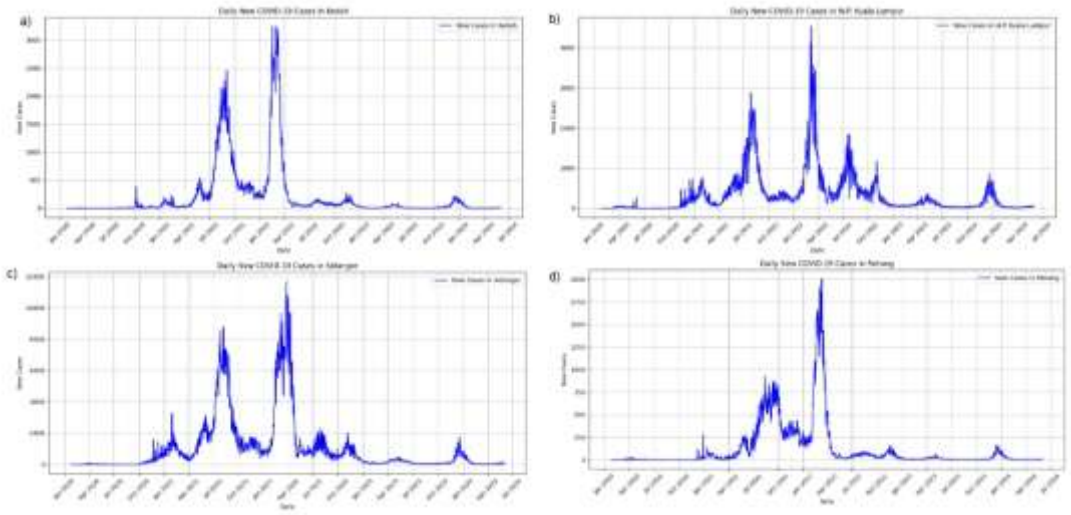

Figure 4.1: Daily new Covid-19 for four selected states

According to Phang et al. (2024), the Covid-19 evolution in Malaysia's 13 states and 3 federal territories can be clustered into four distinct groups. Hence in this study, one representative state was selected from each of these clusters. Their daily new COVID-19 cases were plotted in Figure 4.1. The figures reveal that among the four states, Selangor recorded the highest number of daily cases, with a peak of approximately 12,000 cases in the first quarter of 2022. This was followed by W.P. Kuala Lumpur, which reached a peak of nearly 4,000 cases in early 2022. In contrast, Kedah and Pahang exhibited comparatively lower peaks, at around

3,000 and 2,000 daily cases, respectively, also in early 2022, which are likely attributable to their smaller populations or less densely populated urban areas. These discernible spikes in early 2022 may correspond to the emergence of the Omicron variant.

Figure 4.1 further demonstrates that all four states experienced multiple waves of infection, consistent with global COVID-19 trends. Notably, several other peaks observed in mid-2021—particularly in Selangor and W.P. Kuala Lumpur—coincide with the spread of the Delta variant in Malaysia. From 2023 to 2024, the number of new cases stabilized at lower levels, suggesting a transition to an endemic phase or the buildup of population immunity due to prior infections.

However, presenting the raw data in its unprocessed form, as depicted in Figure 4.1, treats the dataset as a pre-processed time series. This approach overlooks critical topological features—such as cyclical patterns and inherent clusters—embedded within the structural properties of the data. Furthermore, relying solely on raw data makes it difficult to distinguish between natural epidemiological waves (e.g., those driven by the Delta and Omicron variants) and anomalous outliers. To address these limitations, persistent homology is employed to analyse the time series by extracting its underlying topological features. This method enables the identification of extreme events by capturing the data's intrinsic shape and connectivity, thereby providing a more robust framework for distinguishing between expected trends and significant deviations.

4.2 Transforming one dimensional time series to two-dimensional Point cloud

As outlined in Chapter 3, the raw data must first be transformed into a point cloud representation through the application of an embedding theorem. This transformation process is applied across all 13 states and 3 federal territories in Malaysia. The resulting 2-dimensional point clouds typically exhibit one of two distinct geometric configurations which is either a fan-shaped spread pattern or a diagonal structure. These shapes reflect underlying topological characteristics in the data, which may correspond to different epidemiological trends or spatial-temporal dynamics.

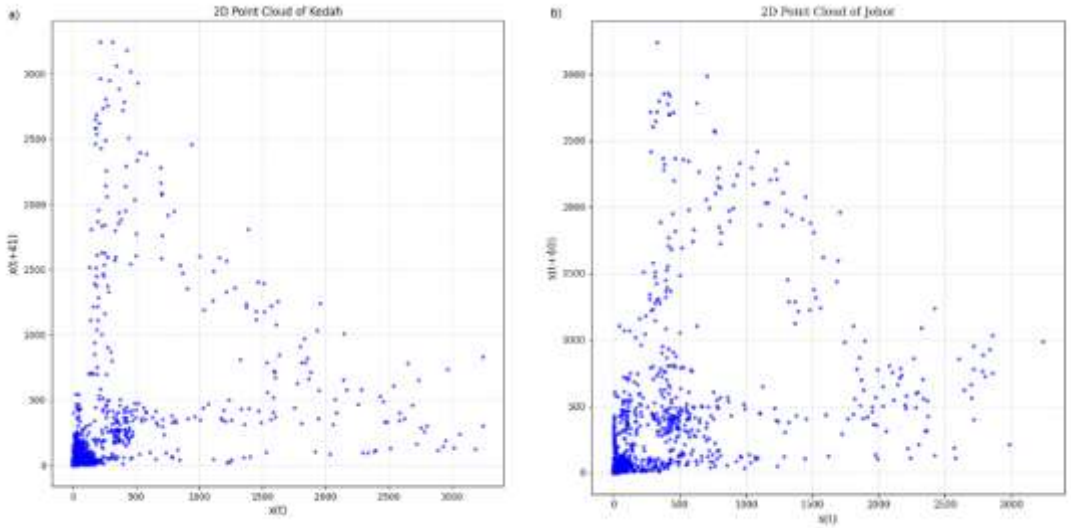

Figure 4.2: 2D point cloud of fan-shaped spread pattern

Among Malaysia's 13 states and 3 federal territories, only Kedah and Johor exhibit a distinctive fan-shaped pattern in their point cloud representations. As illustrated in Figure 4.2, both states demonstrate a high concentration of data points clustered near the origin (lower left quadrant), corresponding to low values of both $x(t)$ and $x(t+\tau)$. This spatial distribution suggests prolonged periods with minimal or zero new COVID-19 case reports, potentially indicative of either the initial stages or post-peak phases of the pandemic.

Conversely, the elongated, high-tailed dispersion of points in both clouds reveals another critical feature: the extension of data points toward elevated values on both axes. This pattern captures days with exceptionally high case counts, a characteristic signature of outbreak waves. The coexistence of these two distinct topological features - dense clustering near the origin coupled with radial dispersion - provides valuable insights into the temporal dynamics of COVID-19 transmission in these regions, highlighting both quiescent periods and explosive growth phases.

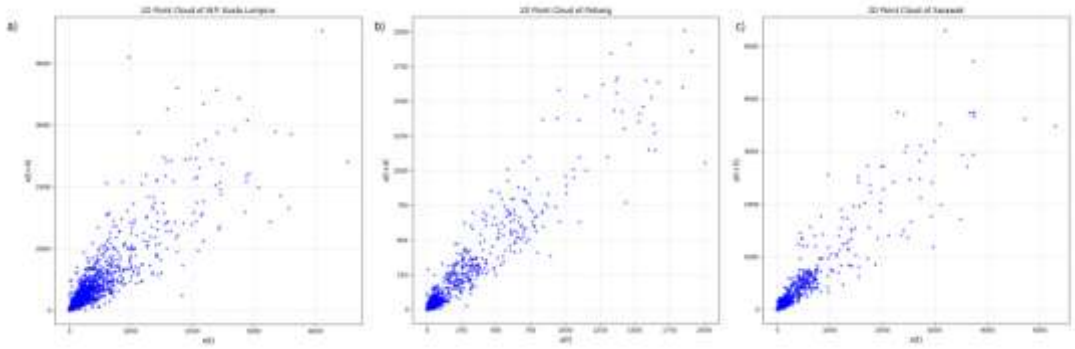

Figure 4.3: 2D point cloud of diagonal structure

The remaining states and federal territories exhibit point clouds demonstrating a distinct diagonal pattern, as illustrated in Figure 4.3. Analysis of these figures reveal particularly tight clustering along the diagonal for Kuala Lumpur, Pahang, and Sarawak. In the case of Kuala Lumpur (see Fig. 4.3 (a)), the majority of data points concentrate within the region ($x(t) < 1000, x(t+4) < 1000$), indicating consistent daily case counts below 1000 with minimal variation across 4-day lag period.

However, these point clouds also display notable outliers extending to approximately $x(t) \approx 3000 - 4000$ daily cases , with corresponding $x(t+4)$ values exhibiting considerable variability - ranging from similarly elevated to significantly reduced case counts. This pattern is consistently observed in both Pahang and Sarawak (see Fig. 4.3 （b） and (c)).

The presence of these divergent trajectories following case spikes suggests fundamentally different epidemiological dynamics during surge periods, where subsequent case counts may either sustain elevated levels or experience rapid declines, indicating unpredictable transmission patterns following outbreak events.

A comparative analysis of the two-point cloud structures reveals distinct geometric characteristics: the diagonal pattern shown in Fig. 4.3 exhibits significantly narrower dispersion compared to the fan-shaped distribution in Fig 4.2. This morphological difference can be primarily attributed to the selection of optimal time delay (τ) parameters implementing in the Takens' embedding theorem.

In this study, the optimal time delay was determined through mutual information analysis, which is essentially a methodological approach that identifies the first local minimum in the mutual information function (see Fig 4.4). This first minimum represents the most appropriate time delay for achieving sufficient reconstruction of the system's phase space while maintaining independence between coordinates. The resulting τ value serves dual purposes which not only governs the point cloud construction but also provides the foundational embedding dimension required for subsequent permutation entropy calculations.

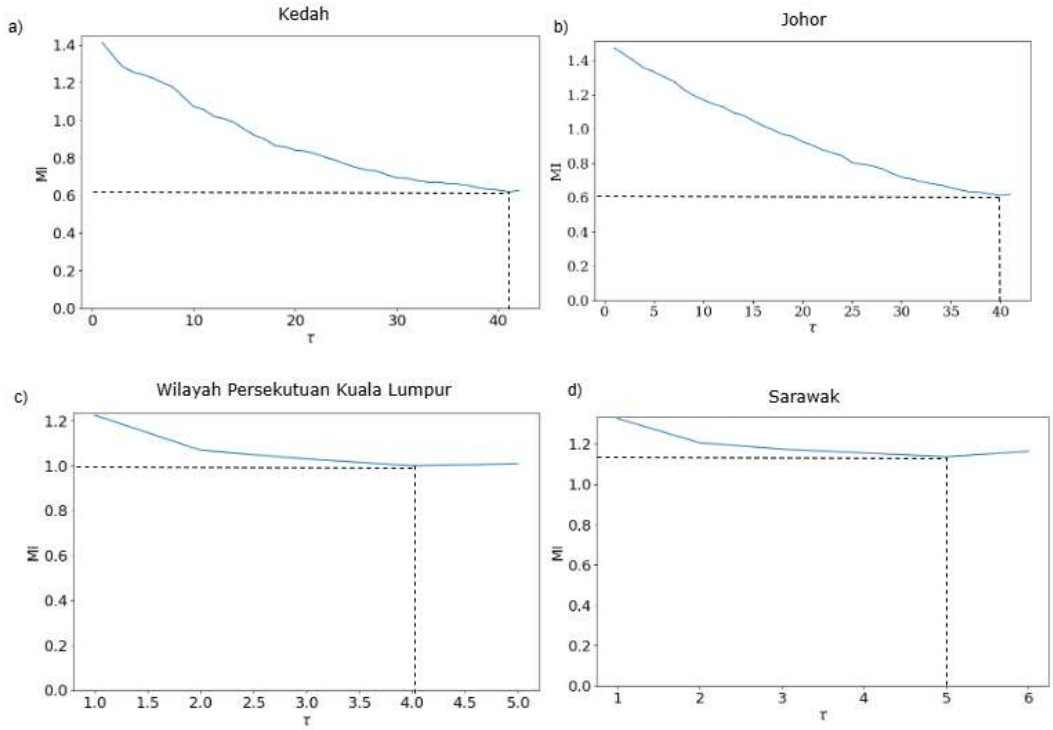

Figure 4.4: Time delay (τ) estimation based on mutual information (MI) of 4 states: Kedah, Johor, Kuala Lumpur, and Sarawak

From Table 4.1, only two states (Kedah and Johor) have exceptionally long delay. These scattered points in the point cloud can form a richer topological structure. This indicate that it might has more persistent of H_0 features and H_1 which indicating the cyclic behaviour. This type of topological signals is ideal for detecting the extreme event and capturing the early warning signal. However, for the states and federal territories that have low optimal time delayed, most of the points appear close to the diagonal line. This indicates that the 0-dimensional topological features, H_0 persist briefly. On the other hand, only a few H_1 features can be captured. As a result, might miss the early transitions as the topology is too smooth. It is still useful for tracking the stable phase, but it is less sensitive to upcoming instability.

Table 4.1: Optimal time delayed of each state or federal territories

States / Federal Territories	Optimal time delayed (days)
Johor	40
Kedah	41
Kelantan	5
Kuala Lumpur	4
Melaka	4
Negeri Sembilan	3
Pahang	4
Perak	4
Perlis	3
Pulau Pinang	5
Sabah	3
Sarawak	5
Selangor	4
Terengganu	4
W.P.Labuan	5
W.P.Putrajaya	5

4.3 Segmentation of time series and construction of Persistent diagram

The time series spanning from 2020 to July 2024 for each state was systematically segmented into eight distinct intervals, each segment spans 180 days temporal window without overlapping. This segmentation enables a detailed examination of the evolving topological features and epidemiological patterns over discrete phases of the pandemic. A comprehensive analysis of the characteristics observed within selected segments (i.e. Segment 1, 3, 4, 5) will be discussed below.

4.3.1 Segment 1

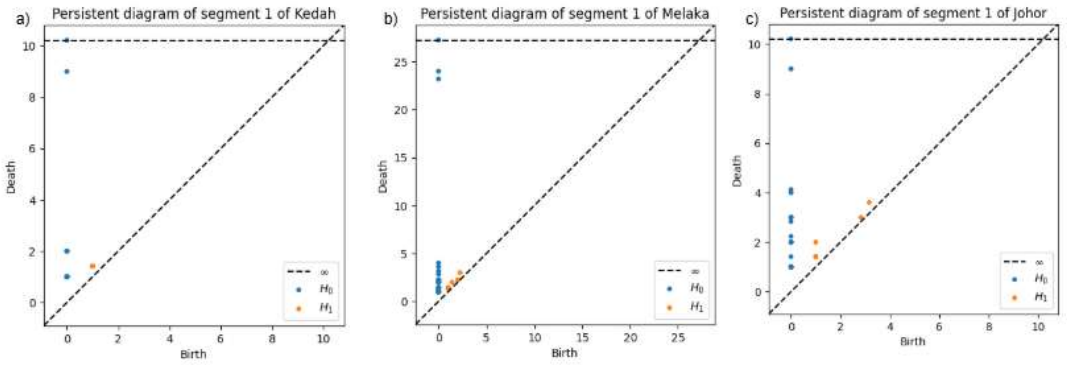

Figure 4.5: First segment 25th Jan 2020 to 25th Jun 2020 of persistent diagram of Kedah, Melaka, and Johor

As illustrated in figure 4.5, the persistence diagram employs a coordinate system where the x-axis (birth) marks the emergence of topological features, and the y-axis (death) denotes their disappearance. The persistence of topological features can be quantitatively assessed by measuring their Euclidean distance from the diagonal. Features close to the diagonal line demonstrate low persistence, while those farther away exhibit greater topological stability.

In the case of H_1 features which representing cyclic structures, their close proximity to the diagonal line in Figure 4.5 indicates minimal persistence that is persistence ≈ 0. This observation suggests an absence of significant periodicity in the outbreak patterns, as genuine cyclic behaviour would manifest as persistent loops across multiple temporal windows. The rapid collapse of these H_1 features in segment 1 implies transient, non-systematic transmission patterns, potentially corresponding to initial epidemic spreading periods where stable cyclic dynamics had not yet emerged.

Conversely, the H_0 features, which are the connected components, display heterogeneous persistence characteristics. Several points exhibit substantial distance from the diagonal, indicating high persistence. This may reflect the incidents of extended periods of

stable disease transmission, or the sustained community spread maintaining consistent case detection rates over the time.

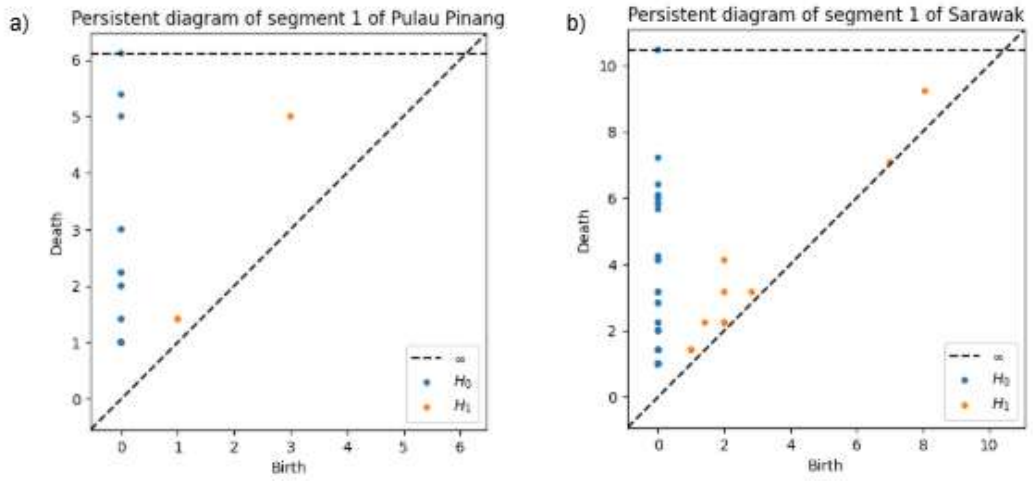

Figure 4.6: First segment of persistent diagram of Pulau Pinang and Sarawak

The first segment id indeed the initial phase of worldwide spread of Covid-19, unlike states in Fig. 4.5, some states experienced relatively stronger periodicity of outbreak with the first six months of pandemic. A significant H_1 feature located substantially distant from the diagonal line, suggests the presence of a persistent cyclic pattern in Covid-19 case dynamics. This topological characteristic was observed specifically in Pulau Pinang and Sarawak for Segment 1 which falls around January to June 2020. Upon a closer investigation on Fig. 4.6, it can be observed that the number of counts of H_1 features in Sarawak's persistent diagram is far more than Pulau Pinang. This may imply multiple loops representing multiple mini waves occurred in Sarawak during that period of time. However, three H_1 features in Sarawak that are mostly short-lived cycles and can be regarded as minor fluctuations which cannot be considered as epidemic waves.

4.3.2 Segment 3

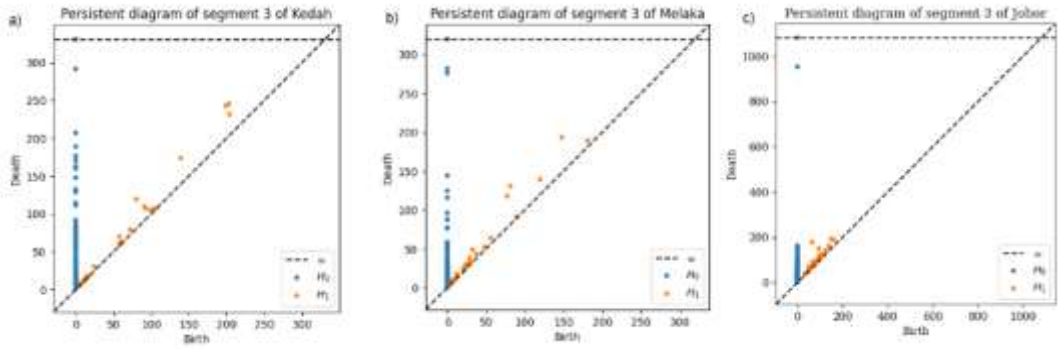

Figure 4.7: Persistent diagram of third segment for Kedah, Melaka, and Johor

The analysis of Segment 3 (early 2021 to Jun 2021) in Figure 4.7 reveals significant epidemiological developments compared to Segment 1 (early 2020 to Jun 2020). Most notably, there is a marked increase in the number of H_1 features, indicative of relatively more emerging cyclic patterns in COVID-19 transmission dynamics within this period of time. This proliferation of topological loops reflects the transition from the initial sporadic outbreak waves observed in Segment 1 to more established periodic infection waves, which is essentially the characteristic of maturing pandemic dynamics. While these H_1 features demonstrate only moderate persistence (positioned at intermediate distances from the diagonal), their increased prevalence suggests developing periodicity in case reporting patterns across multiple states. Concurrently, the persistence of H_0 features remains evident, particularly in Kedah, Johor, and Melaka, signalling sustained community transmission and consistent case detection rates in these states. These observations collectively depict an epidemiological shift from sporadic outbreaks to more predictable wave patterns, with some states exhibiting endemic-like transmission characteristics. The findings align with established models of pandemic progression, where initial outbreak phases typically evolve into more complex cyclical patterns as population immunity develops and viral variants emerge. The topological persistence of

these features provides quantitative evidence of this transition, highlighting the value of computational topology in tracking pandemic evolution.

4.3.3 Segment 4

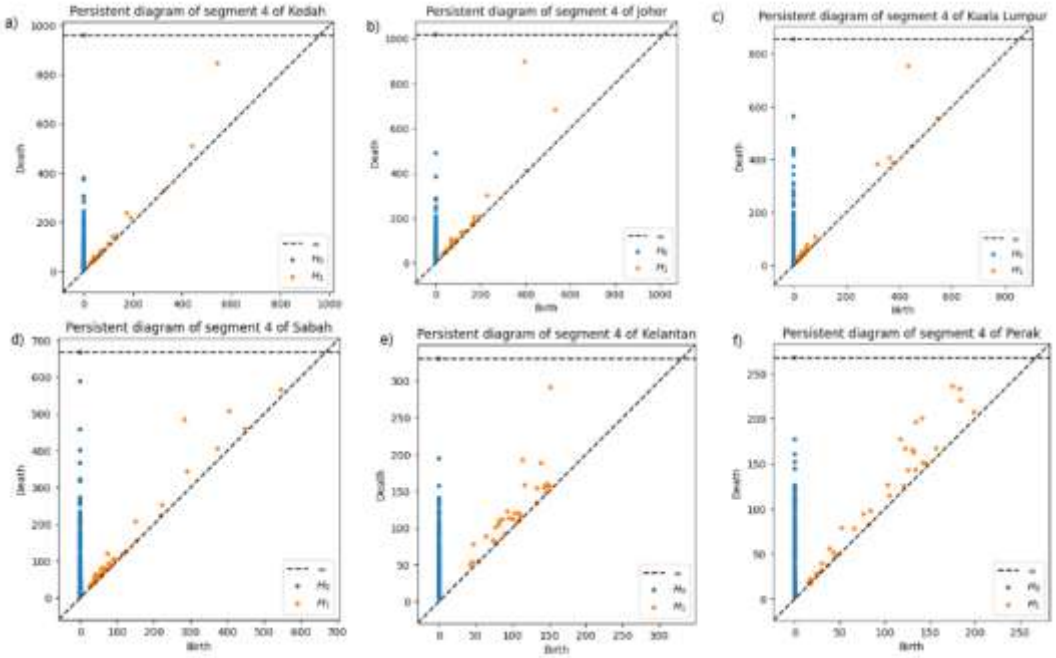

Figure 4.8: Persistent diagram of fourth segment for Kedah, Johor, Kuala Lumpur, Sabah, Kelantan, and Perak

When come to the fourth segment (July-Dec 2021, see Fig. 4.8), we are able to notice that there is a cluster of H_1 features far from the diagonal line when compared to Segment 1 (Fig. 4.5) and Segment 3 (Fig. 4.6). It reflects that some of the states such as Kedah, Johor, Kuala Lumpur, Sabah Kelantan, and Perak strongly exhibit periodic or recurrent structure in their data.

As Segment 4 corresponds to the timeline of middle of 2021, it is highly indicated that the observed high persistent of H_1 features in the persistent diagram might reflect the impact of Delta wave in the Covid-19 pandemic. The Delta variant which peaked in Malaysia between June and September of 2021 was characterized by rapid and pronounced surges in cases and

deaths, forming cyclic rise-and-fall patterns in the data. These patterns are effectively captured by persistent homology as loops, with long-lived H_1 features indicating strong and sustained periodic structures. This is evident in states like Kedah and Johor, where their persistent diagrams show several significant H_1 features far from the diagonal, suggesting intense and prolonged cycles. Meanwhile, states like Kelantan and Sabah display clusters of mid-range H_1 features, further supporting the presence of recurrent oscillations as a typical of behavior pandemic waves.

4.3.4 Segment 5

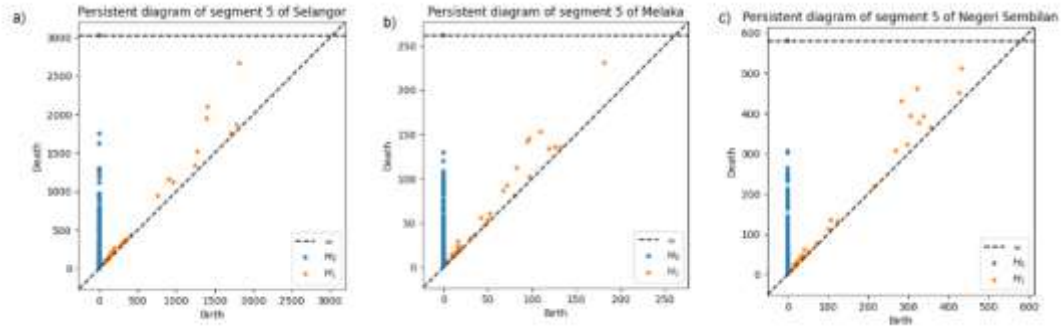

Figure 4.9: Persistent diagram of segment 5 for Selangor, Melaka, and Negeri Sembilan

Similarly, the persistent diagrams for the fifth segment for Selangor, Melaka, and Negeri Sembilan exhibit several highly persistent H_1 features (see the orange points significantly located far from the diagonal line in Fig. 4.9). This indicates that the presence of strong and sustained cyclic behaviour in the data during this period. These long-lived loops likely correspond to complex, periodic oscillations in reported Covid-19 cases or deaths, which are hallmarks of significant pandemic waves.

Given that the fifth segment covers late the period from 2021 to early 2022 when the Omicron variant began to surge throughout Malaysia, figure 4.9 reveal that Selangor, Melaka, and Negeri Sembilan have several prominent H_1 features reaching nearly the top of

the persistent diagram, shows that the signs of prolonged and intense cycles are obvious and these could amplify the dynamic of Omicron.

4.4 Persistent landscape

Following the construction of persistent diagram, a corresponding persistent landscape can be derived. This transformation enables a functional representation of topological features, facilitating further statistical analysis and comparison across the states. In this section, the homology degree is set to be 1 which indicates that only the loop features in the persistent diagram will be captured.

4.4.1 Segment 1

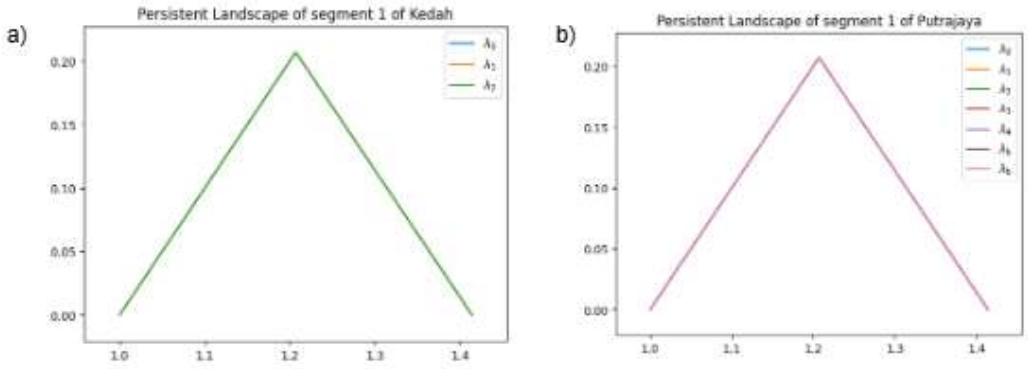

Figure 4.10: Persistent landscape of first segment for Kedah and Putrajaya

The structure of the persistent landscapes for the first segment constructed from each Malaysian state and federal territory generally exhibits similar patterns, as illustrated in Fig. 4.10. In this figure, several landscape functions, denoted by $\lambda_0, \lambda_1, \ldots, \lambda_k$. are presented. Typically, λ_0 corresponds to the most prominent topological features, while λ_1 and subsequent functions represent progressively weaker features. However, for the states of Kedah and W.P. Putrajaya, neither λ_0 nor λ_1 are presented in their respective persistent landscapes, indicating an absence of significant topological features. Instead, Kedah exhibits only λ_2 , while W.P. Putrajaya displays λ_6 suggesting the lack of dominant outbreak waves and the presence of only

low-amplitude, persistent fluctuations. This observation is consistent with the analysis in the preceding section, which noted that most states' persistent diagrams showed low persistence in their H_1 features. Such characteristics may reflect weak cyclic behavior, as the first segment corresponds to the initial phase of the pandemic.

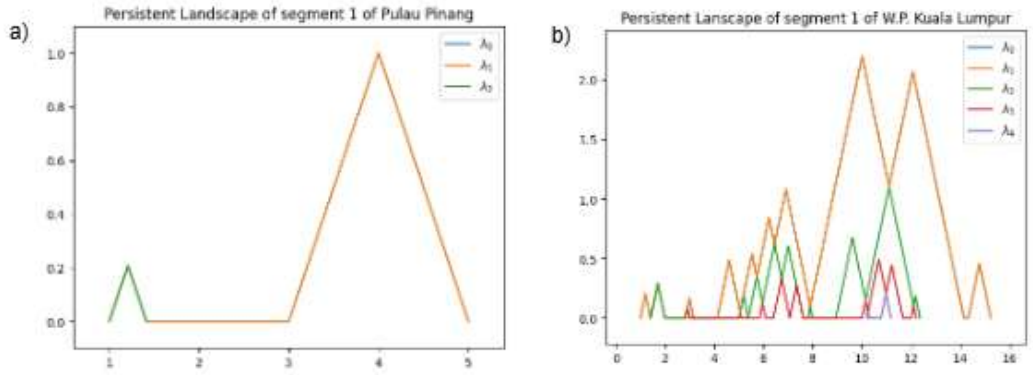

Figure 4.10: Persistent landscape of first segment for Pulau Pinang and W.P. Kuala Lumpur

However, in some states with high population density such as Pulau Pinang and Kuala Lumpur λ_1 can be well captured in their persistent landscape as shown in Fig. 4.10. Pulau Pinang's persistent landscape displays three dominant landscape functions ($\lambda_0, \lambda_1, and\ \lambda_2$), with λ_1 exhibiting the most prominent peak, indicating a strong periodic outbreak pattern. This likely reflects recurrent infection waves driven by urban transmission dynamics in this densely populated state. Similarly, W.P. Kuala Lumpur's persistent landscape shows five landscape functions ($\lambda_0,\ \lambda_1,\ \lambda_2, \lambda_3, and\ \lambda_4$), where the presence of additional higher-order functions (λ_4) suggests more complex outbreak dynamics compared to Pulau Pinang. The extended amplitude range (16 i.e. see the x-axis of Fig 4.10 (b)) in Kuala Lumpur's persistent landscape further supports this interpretation, potentially capturing multiple superimposed transmission cycles or heterogeneous spread patterns across different urban subregions. These persistent landscape features align with the epidemiological context of early 2020, when high-density

urban centres experienced pronounced, repeated outbreaks due to uncontrolled community transmission before intervention measures were fully implemented.

4.4.2 Segment 3

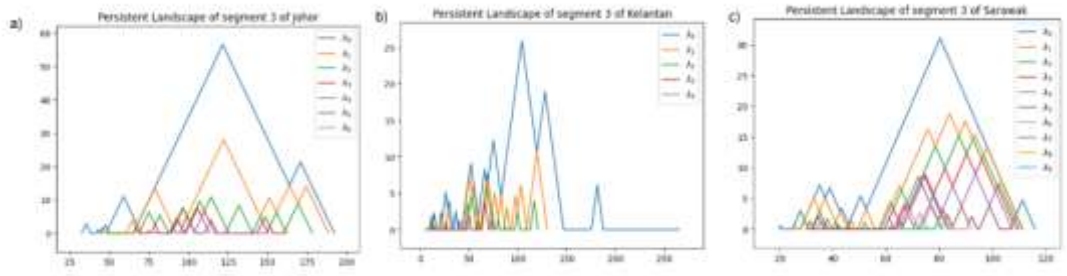

Figure 4.11: Persistent landscape of third segment for Johor, Kelantan, and Sarawak

Beginning from the third segment, which spans approximately from January 2021 to June 2021, all states and federal territories in Malaysia exhibited the presence of the λ_0 feature in their respective persistent landscapes. As illustrated in Fig. 4.11, for three selected states (Johor, Kelantan, and Sarawak), they demonstrate pronounced peaks in their λ_0 curves. These peaks represent the most persistent topological features in the data, manifesting as tall and wide structures in the landscape. Specifically, Johor's λ_0 peak extends approximately from filtration value 60 to 160, Kelantan's from 50 to 150, and Sarawak's similarly from 50 to 150. These substantial peaks indicate the presence of highly significant cyclic structures, likely reflecting prominent wave-like patterns in the COVID-19 case data. Moreover, the wide filtration ranges over which these λ_0 features persist suggest that the detected cycles are robust and not merely the result of noise.

In contrast, the presence of λ_1 and λ_2 features denote additional, but less dominant, cycles within the data. These may correspond to smaller-scale Covid-19 dynamics, such as brief resurgences following initial declines or minor outbreaks. The layered appearance of multiple λ functions for each of the three sample states, as seen in Fig. 4.11, suggests the

existence of multiscale cyclic behaviour. This implies that the progression of Covid-19 cases in Johor, Kelantan, and Sarawak during the third segment was not limited to a single outbreak but rather exhibited recurring structural patterns over time.

4.4.3 Segment 4

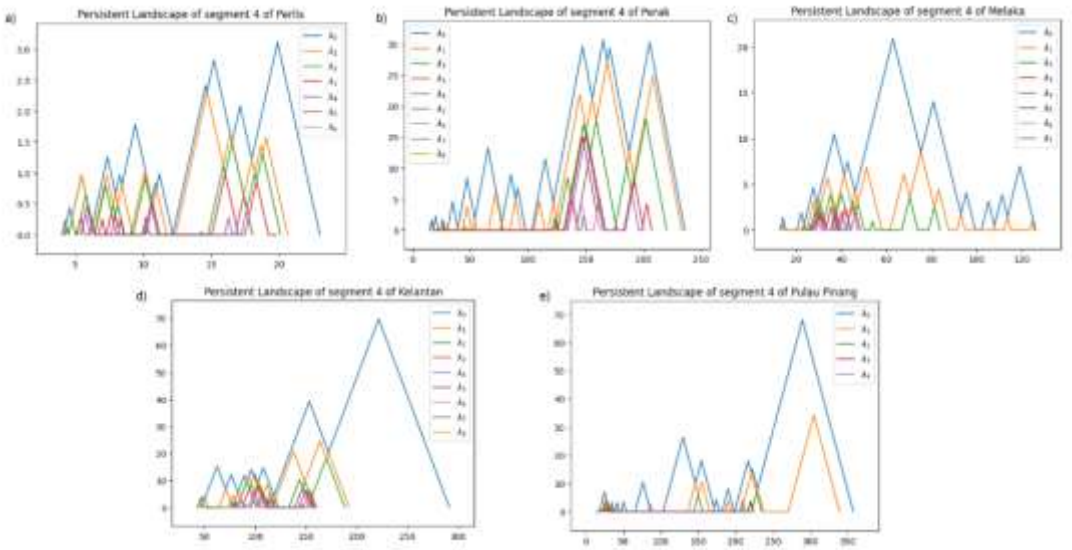

Figure 4.12: Persistent landscape of fourth segment for Perlis, Perak, Melaka, Kelantan and Pulau Pinang

Among the eight segments, the highest amplitude of λ_0 features for all states and federal territory only appeared in either the fourth or fifth segment. This reflects the delta variant surge during the mid of 2021 and the Omicron surge in the end of 2021 or early 2022. Fig. 4.12 shows states including Perlis, Perak, Melaka, Kelantan and Pulau Pinang exhibit prominent λ_0 peaks in this segment, marking this period as epidemiologically significant. The robust λ_0 features, representing the highest amplitude landscape functions which correlate strongly to the Delta variant high-transmission dynamics characteristic. Also, Kelantan's persistent landscape demonstrates particularly wide amplitude range, which is around 300 and hierarchies up to λ_8. This suggests complex and persistent transmission networks during that period in Kelantan. In contrast, Perlis shows a relatively simpler structure ($\lambda_0,\ \lambda_1,\ \lambda_2, \lambda_3,\ \lambda_4,$

$\lambda_5,\ and\ \lambda_6$) with lower amplitude, potentially reflecting its relative isolation and smaller population density. Melaka's intermediate pattern ($\lambda_0,\ \lambda_1,\ \lambda_2, \lambda_3,\ \lambda_4,\ \lambda_5,\ \lambda_6, and\ \lambda_7$) with the range of around 120 indicates more localized outbreak clusters. These topological differences between states mirror known epidemiological variations in Delta variant impact across states in Malaysia, where urban-connected regions experienced more severe and prolonged waves.

4.4.4 Segment 6

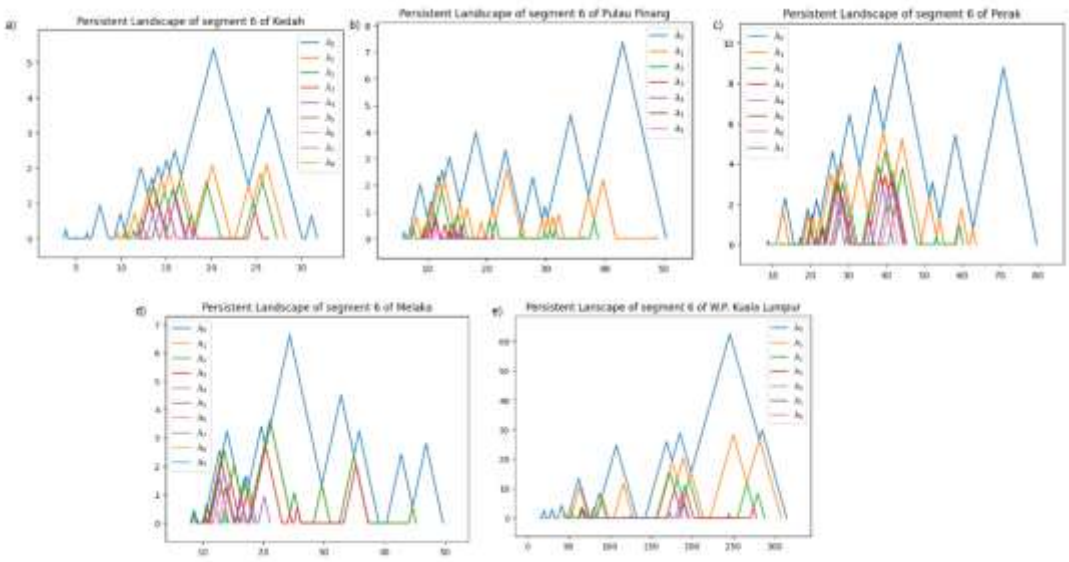

Figure 4.13: Persistent landscape of sixth segment for Kedah, Pulau Pinang, Perak, Melaka, and Kuala Lumpur

In the sixth segment, the presence of the λ_0 feature remains detectable across all the states and federal territories in Malaysia, as illustrated in Fig. 4.13. This indicates the continued presence of topological loops during this period. However, when comparing the amplitude of the λ_0 peaks in this segment to those observed in the fourth segment where large and persistent cycles were evident, it is clearly evidence that the peaks in the sixth segment are significantly smaller. This suggests that the sixth segment reflects a period characterized by localized variability and minor rebounds, rather than strong and widespread cyclic behavior. For instance,

the highest λ_0 feature peak of Pulau Pinang in this segment is approximately 7.5, in stark contrast to the peak of around 70 observed during the fourth segment. This pattern is not unique to Pulau Pinang but is consistent across all states and federal territories. Thus, it can be inferred that the most prominent topological features in the sixth segment were substantially weaker than those in earlier phases of the pandemic.

4.5 L-norm values for persistent landscape

For each persistent landscape, the L_1- norm value can be computed to quantify the overall topological activity. As the data from each Malaysian state and federal territory is divided into eight temporal segments, eight corresponding L_1 norm value are obtained per region. To identify extreme events and detect potential early warning signals, outlier detection technique is employed based on assumptions from the normal distribution. In this study, rare and extreme events are defined as those falling within the upper 5% of the distribution. Specifically, any L_1 norm value exceeding $\mu + 2\sigma$ is flagged as an outlier, where μ denotes the mean and σ the standard deviation of the L_1 norm value. This criterion ensures that only values more extreme than 95% of the dataset are identified as significant.

Table 4.2: Threshold value of each state and federal territories without overlapping segments

States	Threshold values
Johor	5.63×10^4
Kedah	2.04×10^4
Kelantan	1.64×10^4
Kuala Lumpur	4.67×10^4
Melaka	2.74×10^3
Negeri Sembilan	1.33×10^4
Pahang	5.90×10^3
Perak	5.30×10^3
Perlis	1.02×10^3
Pulau Pinang	4.38×10^4
Sabah	1.33×10^5
Sarawak	1.33×10^4
Selangor	3.41×10^5
Terengganu	4.58×10^3
W.P.Labuan	1.79×10^3
W.P.Putrajaya	3.00×10^2

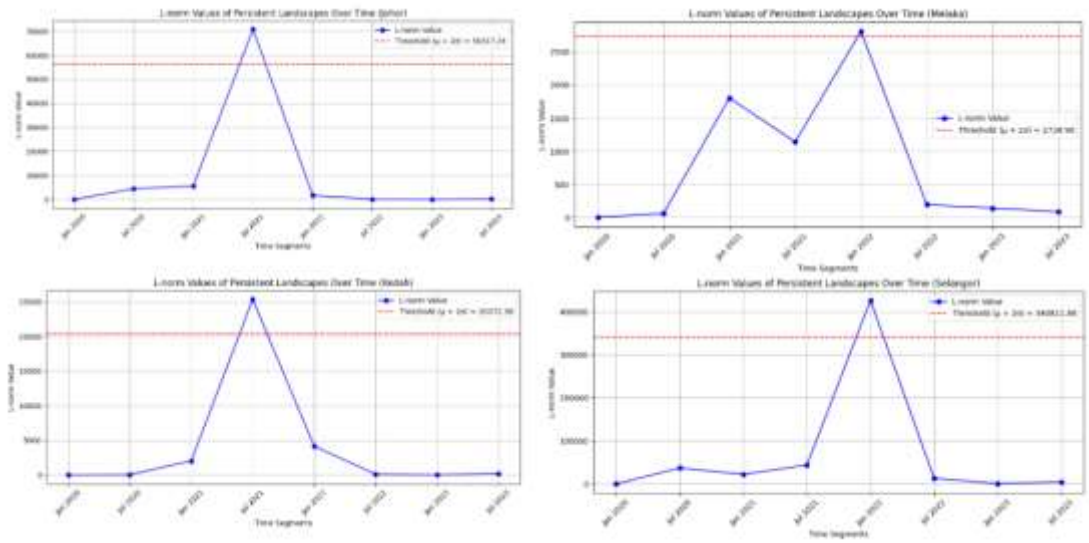

Figure 4.14: The graph of L_1-norm for segmentation without overlapping

Fig. 4.14 presents the L_1 norm values across all eight-time segments for selected states, along with the corresponding threshold values indicated by a red dashed line. This threshold represents the boundary above which L_1 norm values are considered rare and extreme. The results reveal that in certain states, the L_1 norm values in fourth segment exceed the threshold,

while for others, this occurs in fifth segment. These elevated values suggest the presence of significant topological changes in the data, likely corresponding to major epidemiological events. For example, as shown in Fig. 4.14, Johor and Kedah exhibit extreme L_1 norm values during fourth segment, whereas Melaka and Selangor exceed the threshold in fifth segment. This variation may indicate anomalies or irregular patterns in the temporal evolution of Covid-19 cases across different states in Malaysia. The exceedance in fourth segment is likely associated with the Delta variant wave, while the exceedance in fifth segment may correspond to the Omicron wave.

Aside from the segments that exceed the threshold, all other segments with L_1 norm values below the defined cutoff indicate an absence of significant topological deviations during those time periods. In other words, the underlying patterns in the data during these segments are relatively stable and fall within the range of expected or normal behaviour.

However, one limitation of segmenting the time series without overlapping is the potential loss of temporal continuity. Treating each segment as an isolated unit can fragment significant events that span across segment boundaries. In reality, phenomena such as the Delta and Omicron Covid-19 waves unfold continuously over time, regardless of artificial segment divisions. For instance, fourth segment captures only the initial rise of a wave meanwhile fifth segment reflects its decline. As a result, neither segment fully represents the event, causing the corresponding L_1 norm values to fall below the threshold for extreme events. This partial capture dilutes the signal, making it difficult to detect the full impact of the wave. To address this issue, an overlapping segmentation approach was implemented. Specifically, a sliding window with a segment length of 180 days and a step size of 14 days was applied to better capture the structure and dynamics of the data in greater detail.

Table 4.3: Threshold value of each state and federal territories with overlapping segments

States	Threshold values
Johor	5.77×10^4
Kedah	7.74×10^4
Kelantan	1.29×10^4
Kuala Lumpur	4.40×10^4
Melaka	2.64×10^3
Negeri Sembilan	1.39×10^4
Pahang	6.06×10^3
Perak	5.09×10^3
Perlis	6.39×10^2
Pulau Pinang	4.00×10^4
Sabah	1.29×10^5
Sarawak	1.20×10^4
Selangor	3.3×10^5
Terengganu	4.49×10^3
W.P.Labuan	1.48×10^3
W.P.Putrajaya	2.76×10^2

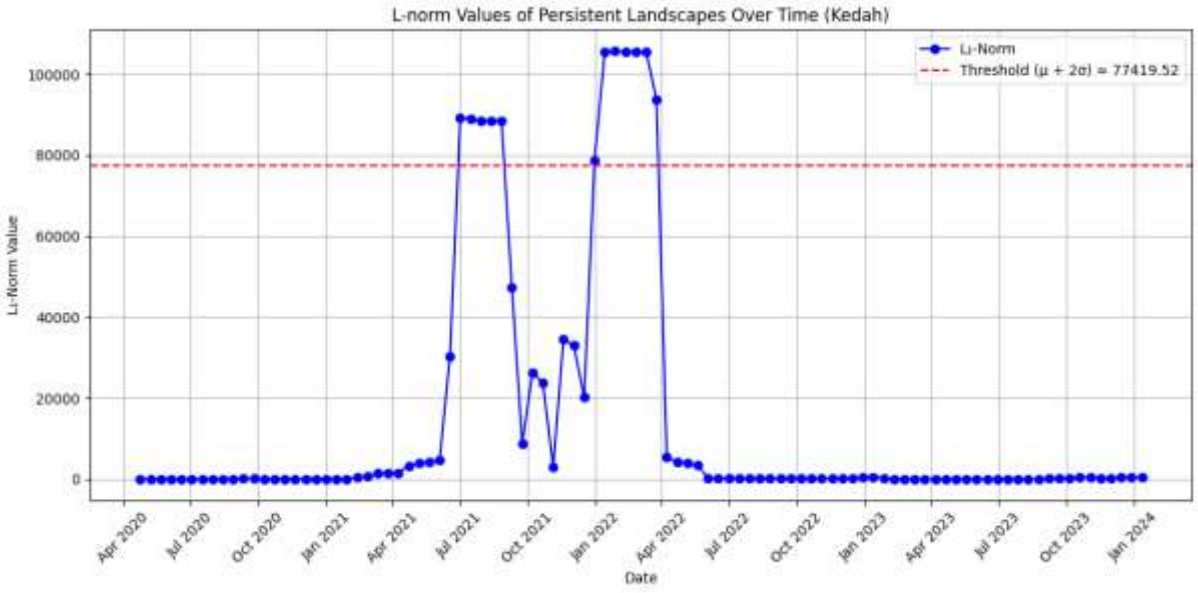

Figure 4.15: The graph of L1-norm with overlapping segmentation for Kedah

Based on Fig. 4.15 illustrates the L_1 norm values of the persistent landscapes for Kedah over time, using an overlapping segmentation approach with a 180-day window and a 14-day step size. The red dashed line represents the threshold $\mu + 2\sigma$ beyond which values are

considered extreme events. Two distinct peaks in the L_1 norm values can be observed, both of which exceed the threshold. These peaks correspond to significant topological changes in the data and can be interpreted as extreme epidemiological events. In the context of this study, these two peaks could represent the Delta and Omicron variant waves respectively.

Moreover, from Fig. 4.15, several L_1 norm values prior to each peak approach but do not exceed the threshold. These points are not classified as extreme but demonstrate a clear upward trend and can be interpreted as early warning signal of an upcoming extreme event. The gradual rise in the L_1 norm suggests increasing complexity or fluctuation in the topological structure of the data, which may precede a major outbreak or wave.

However, it is also noted that not all the states exhibit such a clear dual-peak behavior. In some states, only one peak, either wither Delta or Omicron, is captured as an extreme event. The findings for Kedah highlight the strength of the overlapping method in preserving temporal continuity and enhancing the detection of both extreme events and their early warning indicators.

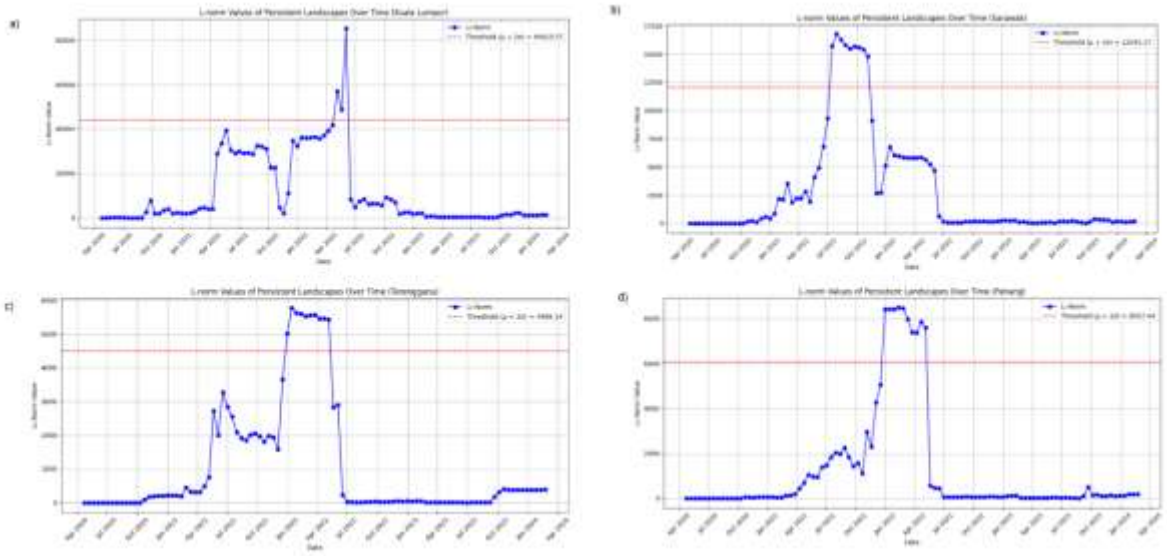

Figure 4.16: The graph L1-norm with overlapping segments for selected states

Fig. 4.16 shows the L_1 norm values of persistent landscape for four selected states using overlapping segmentation. Notably, this approach is able to capture the Delta variant wave in Sarawak, meanwhile Kuala Lumpur, Pahang, and Terengganu, only Omicron waves is identified as an extreme event. Despite the improve resolution provided by the overlapping segments, only one major peak is observed for these regions.

One possible justification for such outcome is there exists substantial variations in the values used for optimal time delay in the time delay embedding step when constructing the point cloud. For instance, states such as Kuala Lumpur, Pahang, and Terengganu exhibit relatively short optimal time delay, approximately 4-5 days meanwhile Kedah has a much longer time delay of approximately 41 days.

This reflects that a short time delay captures rapid changes in the data but it may fail to capture the full cyclic or structural patterns of broader phenomena such as pandemic waves. As a result, the point cloud may lack sufficient geometric complexity to reflect longer-term topological features like loops or persistent structures. This can lead to a suppression of

persistent homology signal, causing one of the major waves, typically the one with lower persistence or the shorter duration to of undetected in the persistent landscape and hence in the L_1 norm analysis.

In contrast, a longer time delay helps to unfold more of the underlying temporal structure, enabling the model to capture long-range correlations in the data. This provides a more robust representation of cyclic behaviors such as repeated waves of infection, thus increasing the likelihood of detecting multiple peaks in the L_1 norm.

4.6 Summary of Chapter 4

Chapter 4 presents a comprehensive analysis of Covid-19 cases data in Malaysia using Topological Data Analysis (TDA), focusing on persistent homology and persistent landscapes to uncover structural patterns not visible in the standard time series analysis. Initially, raw data from selected Malaysia's states revealed multiple infection waves, particularly during the Delta and Omicron surges, but lacked deeper structural insight. To address this limitation, time-series data were transformed into 2D point clouds using Takens' embedding theorem, revealing two key geometric patterns which are fan-shaped structures for Kedah and Johor state which indicate prolonged low case periods punctuated by outbreaks, and diagonal pattern for the rest of the states in Malaysia which suggest more stable case numbers with occasional deviations. The optimal time delay for embedding was determined using mutual information, with longer delays, it able to capture richer topological features. The time series were then segmented into 180-day windows, and persistent diagrams were constructed. Segment 1 shoed weak cyclic patterns, reflecting early, sporadic outbreaks. By segment 3, more prominent loops emerged, signaling the development of recurrent infection waves. Segment 4 and segment 5 exhibited the strongest topological features across many states, aligning with the Delta and Omicron waves respectively. These diagrams were further transformed into persistent landscapes, enabling statistical comparison across states. While early segments showed low persistence, segment 4 and segment 5 displayed high-amplitude features, indicating significant and sustained transmission cycles. L_1 norm value in segment 4 and segment 5 highlighted major epidemiological events. Overlapping segmentation using sliding window was introduced to preserve temporal continuity and better detect the extreme events and early warning signs. As a result, it was found that the states with longer embedding time

delay such as Kedah could capture multiple waves, while those states with shorter delays often detected only one due to less geometric complexity in their point clouds.

Chapter 5

Discussion and Conclusion

5.0 Overview

This chapter provides a comprehensive discussion of the effectiveness of using Topological Data Analysis (TDA), specifically persistent homology, to detect early warning signals in COVID-19 time series data across Malaysian states. The method showed strong alignment with official data, particularly in Kedah, where both Delta and Omicron waves were accurately identified. The chapter also examines how time series characteristics and a 14-day overlapping window influenced detection sensitivity. While the study achieved its objectives, limitations include its retrospective nature and lack of real-time data integration. Suggestions for future work include real-time dashboards and incorporating machine learning for improved prediction.

5.1 Discussion

5.1.1 Validation with data from COVIDNOW website

Based on the results from Chapter 4, among all 13 states and 3 federal territories in Malaysia, only Kedah achieved optimal detection results. The persistent homology method successfully identified both the Delta and Omicron waves for Kedah. In contrast, the other states were only able to detect one pandemic wave, either Delta or Omicron.

Table 5.1: Summary of early warning signal and Delta variant detection

States	Early warning signal for Delta variant	Delta period detected	Highest case during Delta variant
Kedah	Jun 2021	July - August 2021	11th August 2021
Johor	September 2021	September 2021 – November 2021	6th September 2021
Sarawak	July 2021	July 2021 – Nov 2021	16th September 2021
Perak	July 2021	July - October 2021	24th August 2021
W.P. Labuan	March 2021	March 2021	16th June 2021

Table 5.2: Summary of early warning signal and Omicron variant detection

States	Early warning signal for Omicron variant	Omicron period detected	Highest case during Omicron variant
Kedah	Dec 2021	Dec 2021 – March 2022	17th February 2022
Terengganu	Dec 2021	Dec 2021 - March 2022	13th March 2022
Selangor	Jan 2022	Jan 2022 – June 2022	3rd April 2022
Sabah	Dec 2021	Dec 2021 – May 2022	27th February 2022
Pulau Pinang	Nov 2021	Nov 2021 – April 2022	4th March 2022
Pahang	Dec 2021	Dec 2021 – May 2022	12th March 2022
Negeri Sembilan	Dec 2021	Dec 2021 – May 2022	17th March 2022
Melaka	Feb 2022	Feb 2022 - April 2022	14th March 2022
W.P. Labuan	Dec 2021, March 2022	Dec 2021, March 2022 - May 2022	25th February 2022
W.P. Kuala Lumpur	May 2022	May – June 2022	12 March 2022
W.P. Putrajaya	September 2022	September 2022	28th February 2022
Kelantan	Dec 2021, Feb 2022	Dec 2021, Feb – April 2022	22nd February 2022
Perlis	Nov 2021	Nov 2021- Dec 2022	1st March 2022

Based on the data presented in Tables 5.1 and 5.2, the results obtained using persistent homology were compared to official statistics from the COVIDNOW website. For the Delta variant in Kedah, the early warning signals and extreme event periods detected through persistent homology align well with the reported peak case dates, indicating a high level of accuracy. Similarly, for the Omicron variant in Kedah, the highest number of reported cases also falls within the detected range, further validating the method's effectiveness.

In contrast, other states such as Johor, Sarawak, and Perak (see from Table 5.1) were only able to detect the Delta variant. Nevertheless, the detected periods and early warning signals remain consistent with the corresponding peak case dates from the COVIDNOW data.

A similar pattern is observed in Table 5.2 for states including Terengganu, Selangor, Sabah, Pulau Pinang, Pahang, Negeri Sembilan, Melaka, W.P. Labuan, W.P. Putrajaya, and Kelantan, where only the Omicron variant was detected. In all these cases, the peak dates recorded by COVIDNOW fall within the timeframes identified by persistent homology, supporting the method's reliability.

However, notable discrepancies were observed in the cases of W.P. Labuan (see from Table 5.1) and W.P. Kuala Lumpur (see from Table 5.2). For W.P. Labuan, the early warning signal and detected period for the Delta variant occurred around March 2021, whereas the highest recorded number of cases was on 16 June 2021, indicating a significant delay. Similarly, for W.P. Kuala Lumpur, the early warning signal and detected period for the Omicron variant were in May–June 2022, yet the highest number of cases was recorded earlier, on 12 March 2022. These inconsistencies may be attributed to time delay effects, as previously explained in Chapter 4.

5.1.2 Impact of time series characteristics and overlapping time step

The nature of Covid-19 cases data exhibits quasi-periodic and non-stationary behaviour, with fluctuating case numbers driven by external factors such as policy changes, public compliance, and virus mutations. These dynamics pose challenges to detection methods relying on consistent periodic structures.

In this research study, 14 days overlapping segment was employed during the time series segmentation process. This decision was grounded in epidemiological knowledge, particularly the average incubation period of the virus and the reporting delay window, which often spans 10 to 14 days. By using 14 days as step size to overlap, the method increases

sensitivity to local structural changes in the data, thereby enhancing the likelihood of detecting early signals before case surges occur.

The results suggest that the 14 days to overlap provided sufficient temporal resolution to capture meaningful topological features. For instance, this approach successfully captured multiple wave patterns in Kedah state. However, in some cases, such as Kuala Lumpur, the overlapping window might have missed early indicators if the case increases were abrupt or driven by short-term events.

Therefore, while the 14 days overlap strikes a balance between noise reduction and temporal sensitivity, it also introduces a form of time discretization that could delay detection or obscure short-lived signals. This effect reinforces the need to carefully calibrate the segment length and overlap window when applying topological data analysis to epidemiological time series.

5.2 Summary of Chapter

In chapter 5, the performance and practical implications of using Topological Data Analysis (TDA) through persistent homology to detect early warning signals for extreme events in COVID-19 time series data was discussed. By comparing the results of this approach with official data from the COVIDNOW website, the analysis confirms that the method is largely effective, especially for the state of Kedah, where both Delta and Omicron waves were successfully detected. Other states demonstrated partial success—typically identifying one major wave—but still produced detection periods closely aligned with peak case data, indicating good reliability. Moreover, this research study also analysed the impact of the 14-day overlapping segmentation window, chosen to align with the COVID-19 incubation and reporting delay period. These overlaps enhanced sensitivity to evolving patterns and helped

reduce noise, particularly benefiting the detection in states like Kedah. However, it may have limited responsiveness to abrupt spikes in other regions like Kuala Lumpur, highlighting the importance of choosing appropriate segment parameters in TDA.

Table 5.3: Summary of objectives and description of objectives

Research objectives	Description
Explore preliminary data pre-processing requirements and transformation techniques for applying topological data analysis on time series.	The optimal time delay was computed using mutual information and converted the 1-dimensional raw data into 2D point cloud using Taken's embedding theorem. Moreover, 14 days overlapping segmentation was justified based on epidemiological insights, balancing noise filtering and sensitivity
Integrate the commonly used topological features and descriptors in feature extraction of time series data.	Topological features such as persistent diagrams and persistent landscape were extracted and analyzed. These features were used to detect the emergence of waves and compare results across states.
Carry out the TDA-based time series analysis using on enhancing the early warning for the extreme event in time series.	The L_1 norm values were computed from persistent landscape to detect the early warning signals and extreme events for Delta and Omicron waves. In Kedah, both were detected with high alignment to COVIDNOW data. Other states showed consistency in one wave, confirming reliable performance.

5.3. Limitation of study

One of the limitations of this study is this research project is a reprospective analysis only. This study only focuses on the historical analysis of Covid-19 case in Malaysia, analyzing the past data to identify the patterns, topological features and significant epidemiological events. As a result, all the findings which include the detection of extreme events and early warning signals are observed only after the events have already occurred. This limits the practical utility of the analysis in the real-time setting where the early detection and rapid response are critical.

Since the whole analysis in this study is conducted entirely on static, historical datasets, with Covid-19 cases data available only up to July 2024. There is a need to incorporate the real-time data streams or automated pipeline that would allow the model to continuously update its outputs as new information becomes available. This lack of real-time integration significantly reduces the system's responsiveness and limits its practical application in the dynamic public health environment, where up-to-date insights are crucial for timely interventions.

Other than that, this research project only focusses on analyzing univariate time series which is the daily new Covid-19 cases for each Malaysian state and federal territory. This simplifies the modeling and allows the extraction of clear topological patterns, it excludes the other potentially influential variables that could offer a deeper and more nuanced understanding of the pandemic's dynamics. The key multivariate factors such as vaccination rates, mobility trends, lockdowns, and demographic characteristics are not incorporated into the analysis.

5.4. Suggestion for future study

There are several suggestions for future study for this project, the future work can focus on developing a real-time interactive dashboard in the form of a web-based or desktop interface. This platform would connect directly to real-time public health data sources, such as APIs provided by health ministries or open data repositories, enabling automatic updates of persistent diagrams, persistent landscapes, and L-norm graph as new data becomes available. This system would provide significant benefits to various user groups. For instance, public health researchers could use it for timely analysis, policymakers could rely on it for informed and data-driven decision-making, and general users could explore and understand the progression of outbreak dynamic in their states.

Besides, the future studies of this project can integrate machine learning techniques to the persistent landscape. Researchers can develop models capable of classifying or predicting extreme events such as new Covid-19 waves or emerging variants. Other than that, anomaly detection can be automated through supervised or unsupervised algorithms such as support vector machine (SVM), and random forest. The integration of topological data analysis with machine learning model able to improve the predictive analysis, timeliness, and adaptability of outbreak monitoring system, making it more responsive to dynamic and complex epidemiological patterns.

Last but not least, this TDA approach is not limited to the analysis of COVID-19 data. It can also be applied to other fields that involve time series data. By following the same methodology used in this study, which includes forming point clouds, computing persistent diagrams, constructing persistent landscapes, and calculating the L-norm value, TDA is able to reveal hidden patterns and structures embedded within the data.

5.5 Conclusion

This study demonstrates the potential of Topological Data Analysis (TDA), particularly persistent homology, in identifying early warning signals for extreme events within the complex time series data. By applying TDA to Covid-19 daily case data from Malaysia's 13 states and federal territories, the research effectively transformed one-dimensional time series into high-dimensional point clouds, enabling the detection of significant topological features such as loops and connected components through persistent diagrams. The results affirm that persistent homology can capture subtle cyclic structures and abrupt deviations in the data, which are often obscured in raw time series formats. These structures provide valuable indicators of impending surges or critical transitions in pandemic trends. In particular, states

with longer optimal time delays (Kedah and Johor) exhibited richer topological patterns, enhancing sensitivity to early instability signals, whereas shorter delays in other states reflected more stable but less informative structures. This research contributes a geometry-based analytical framework for detecting extreme events in epidemiological time series. The methodology not only addresses the limitations of traditional statistical models when dealing with non-linear and noisy data but also offers a transferable approach for broader applications in domains such as finance, environmental science, and engineering as far as time series data are concerned. With further integration into real-time system, this topological framework holds promise as a foundational tool for future early warning systems and risk mitigation strategies.

References

Aikebaier, S. (2024). COVID-19, new challenges to human safety: a global review. *Front. Public Health*, *12:1371238*. https://doi.org/10.3389/fpubh.2024.1371238

Altındiş, F., Yılmaz, B., Borisenok, S., & İçöz, K. (2021). Parameter investigation of topological data analysis for EEG signals. *Biomedical Signal Processing and Control*, *63*, 102196. https://doi.org/10.1016/j.bspc.2020.102196

Bello, S. A., Yu, S., Wang, C., Adam, J. M., & Li, J. (2020). Review: Deep Learning on 3D Point Clouds. *Remote Sensing*, *12*(11), 1729. https://doi.org/10.3390/rs12111729

Bubenik, P., & Dlotko, P. (2017). A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, *78*, 91–114. https://doi.org/10.1016/j.jsc.2016.03.009

Bukkuri, A., Andor, N., & Darcy, I. K. (2021). Applications of Topological Data Analysis in Oncology. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.659037

Chazal, F., & Michel, B. (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.667963

Corbet, R., Fugacci, U., Kerber, M., Landi, C., & Wang, B. (2019). A kernel for multi-parameter persistent homology. *Computers & Graphics: X*, *2*, 100005. https://doi.org/10.1016/j.cagx.2019.100005

Edelsbrunner, H. (2022). *COMPUTATIONAL TOPOLOGY : an introduction.* Amer Mathematical Society.

El-Yaagoubi, A. B., Chung, M. K., & Ombao, H. (2023). Topological Data Analysis for Multivariate Time Series Data. *Entropy*, *25*(11), 1509. https://doi.org/10.3390/e25111509

Fatima, & Rahimi, A. (2024). A Review of Time-Series Forecasting Algorithms for Industrial Manufacturing Systems. *Machines*, *12*(6), 380–380. https://doi.org/10.3390/machines12060380

Garside, K., Henderson, R., Makarenko, I., & Masoller, C. (2019). Topological data analysis of high resolution diabetic retinopathy images. *PLoS ONE*, *14*(5), e0217413–e0217413. https://doi.org/10.1371/journal.pone.0217413

Gidea, M., Goldsmith, D., Katz, Y., Roldan, P., & Shmalo, Y. (2020). Topological recognition of critical transitions in time series of cryptocurrencies. *Physica A: Statistical Mechanics and Its Applications*, *548*, 123843. https://doi.org/10.1016/j.physa.2019.123843

Hashim, J. H., Adman, M. A., Hashim, Z., Mohd Radi, M. F., & Kwan, S. C. (2021). COVID-19 Epidemic in Malaysia: Epidemic Progression, Challenges, and Response. *Frontiers in Public Health*, *9*(560592). https://doi.org/10.3389/fpubh.2021.560592

Hodson, A., Pearce, J. M., Amlôt, R., & M Brooke Rogers. (2024). Delivering extreme event preparedness education in schools: A systematic review of educational preparedness resources available in England. *International Journal of Disaster Risk Reduction*, *100*, 104171–104171. https://doi.org/10.1016/j.ijdrr.2023.104171

Karim, W., Haque, A., Anis, Z., & Ulfy, M. A. (2020). The Movement Control Order (MCO) for COVID-19 Crisis and its Impact on Tourism and Hospitality Sector in Malaysia. *International Tourism and Hospitality Journal*, *3*(2). https://doi.org/10.37227/ithj-2020-02-09

Krakovská, A., Mezeiová, K., & BudáIová, H. (2015). Use of False Nearest Neighbours for Selecting Variables and Embedding Parameters for State Space Reconstruction. *Journal of Complex Systems*, *2015*, 1–12. https://doi.org/10.1155/2015/932750

Loong, Y. S., & Wan Amirah, W. U. (2022). The Malaysian Economy and COVID-19: Policies and Responses from January 2020 – April 2021. *United Nations Conference on Trade and Development, Geneva, Switzerlannd*.

Machado, J. A. T., & Lopes, A. M. (2020). Rare and extreme events: the case of COVID-19 pandemic. *Nonlinear Dynamics*, *100*(3), 2953–2972. https://doi.org/10.1007/s11071-020-05680-w

Munch, E. (2017). A User's Guide to Topological Data Analysis. *Journal of Learning Analytics*, *4*(2), 47–61. https://doi.org/10.18608/jla.2017.42.6

Nahid, A., Hesam, S., & Hormoz, S. (2018). Strategies for disaster risk reduction education: A systematic review. *Journal of Education and Health Promotion*, *7*(1), 98–98. https://doi.org/10.4103/jehp.jehp_31_18

Nur, Mohd, Fatimah Abdul Razak, Ismail, M., & Mohd Almie Alias. (2022). Hybridization of hierarchical clustering with persistent homology in assessing haze episodes between air quality monitoring stations. *Journal of Environmental Management*, *306*, 114434–114434. https://doi.org/10.1016/j.jenvman.2022.114434

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., & Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, *6*(1). https://doi.org/10.1140/epjds/s13688-017-0109-5

Peiman Alipour, S., Mohammad , N., & Djamel, K. (2019). The Potential of Data Analytics in Disaster Management. *Springer*, 335–348. https://doi.org/10.1007/978-3-030-03317-0_28

Phang, P., Ling, C. Y.-F., Liew, S.-H., Razak, F. A., & Benchawan Wiwatanapataphee. (2024). Nonlinear time series analysis of state-wise COVID-19 in Malaysia using wavelet and persistent homology. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-79002-0

Rai, A., Sharma, B. N., Salam Rabindrajit Luwang, Md. Nurujjaman, & Sushovan Majhi. (2024). Identifying extreme events in the stock market: A topological data analysis. *Chaos an Interdisciplinary Journal of Nonlinear Science*, *34*(10). https://doi.org/10.1063/5.0220424

Ranjai Baidya, & Lee, S.-W. (2024). Addressing the Non-Stationarity and Complexity of Time Series Data for Long-Term Forecasts. *Applied Sciences*, *14*(11), 4436–4436. https://doi.org/10.3390/app14114436

Ravishanker, N., & Chen, R. (2019). *Topological Data Analysis (TDA) for Time Series*. arxiv: 1909.10604.

Rhif, M., Ben Abbes, A., Farah, I., Martínez, B., & Sang, Y. (2019). Wavelet Transform Application for/in Non-Stationary Time-Series Analysis: A Review. *Applied Sciences*, *9*(7), 1345. https://doi.org/10.3390/app9071345

Sahimi, H. M. S., Mohd Daud, T. I., Chan, L. F., Shah, S. A., Rahman, F. H. A., & Nik Jaafar, N. R. (2021). Depression and Suicidal Ideation in a Sample of Malaysian

Healthcare Workers: A Preliminary Study During the COVID-19 Pandemic. *Frontiers in Psychiatry*, *12*. https://doi.org/10.3389/fpsyt.2021.658174

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, *461*(7260), 53–59. https://doi.org/10.1038/nature08227

Schindler, D. J., & Barahona, M. (2023). Persistent Homology of the Multiscale Clustering Filtration. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.04281

Shah, A. U. M. (2020). COVID-19 Outbreak in Malaysia: Actions Taken by the Malaysian Government. *International Journal of Infectious Diseases*, *97*, 108–116. https://doi.org/10.1016/j.ijid.2020.05.093

Siddiqui, S., Aarti Shikotra, Richardson, M., Doran, E., Choy, D. F., Bell, A., Austin, C. D., Eastham-Anderson, J., Hargadon, B., Arron, J. R., Wardlaw, A. J., Brightling, C. E., Heaney, L. G., & Bradding, P. (2018). Airway pathological heterogeneity in asthma: Visualization of disease microclusters using topological data analysis. *The Journal of Allergy and Clinical Immunology*, *142*(5), 1457–1468. https://doi.org/10.1016/j.jaci.2017.12.982

Sivakumar, B., & Deepthi, B. (2021). Complexity of COVID-19 Dynamics. *Entropy*, *24*(1), 50. https://doi.org/10.3390/e24010050

Syed Musa, S. M. S., Md Noorani, M. S., Abdul Razak, F., Ismail, M., Alias, M. A., & Hussain, S. I. (2021). Using persistent homology as preprocessing of early warning signals for critical transition in flood. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86739-5

Tan, E., Algar, S., Corrêa, D., Small, M., Stemler, T., & Walker, D. (2023). Selecting embedding delays: An overview of embedding techniques and a new method using persistent homology. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *33*(3), 032101. https://doi.org/10.1063/5.0137223

Uray, M., Giunti, B., Kerber, M., & Huber, S. (2024). Topological Data Analysis in smart manufacturing: State of the art and future directions. *Journal of Manufacturing Systems*, *76*, 75–91. https://doi.org/10.1016/j.jmsy.2024.07.006

Varrelman, T. J., Rader, B., Remmel, C., Tuli, G., Han, A. R., Astley, C. M., & Brownstein, J. S. (2024). Vaccine effectiveness against emerging COVID-19 variants using digital health data. *Communications Medicine*, *4*(1). https://doi.org/10.1038/s43856-024-00508-9

Wang, X., & Cheng, Z. (2020). Cross-sectional studies: Strengths, weaknesses, and recommendations. *Chest*, *158*(1), 65–71. NCBI. https://doi.org/10.1016/j.chest.2020.03.012

Wissel, C. (1984). A universal law of the characteristic return time near thresholds. *Oecologia*, *65*(1), 101–107. https://doi.org/10.1007/bf00384470

Yin, A. L., Guo, W. L., Sholle, E. T., Rajan, M., Alshak, M. N., Choi, J. J., Goyal, P., Jabri, A., Li, H. A., Pinheiro, L. C., Wehmeyer, G. T., Weiner, M., Safford, M. M., Campion, T. R., & Cole, C. L. (2022). Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. *International Journal of Medical Informatics*, *157*, 104622. https://doi.org/10.1016/j.ijmedinf.2021.104622

Yuhei, U., Junji, K., & Hideyuki, K. (2019). Topological Data Analysis and Its Application to Time-Series Data Analysis. *FUJITSU SCIENTIFIC & TECHNICAL JOURNAL*, *55*(2), 65–71.

Zakaria, S., Sulaiman, N. F. C., Roslan, U. A. M., Alias, A., & Malik, S. M. A. (2023). Impact of COVID-19 Pandemic on Malaysian Socio-Economics: Statistical-Dynamical Approach. *Journal of Mathematical Sciences and Informatics*, *3*(1). https://journal.umt.edu.my/index.php/jmsi/article/view/355/285