TOPOLOGICAL DATA ANALYSIS IN ANALYZING THE SIMILARITY OF AIR POLLUTANTS' BEHAVIOR ACROSS AIR QUALITY MONITORING STATIONS IN MALAYSIA

MADUKPE VINE NWABUISI

UNIVERSITI SAINS MALAYSIA

2025

TOPOLOGICAL DATA ANALYSIS IN ANALYZING THE SIMILARITY OF AIR POLLUTANTS' BEHAVIOR ACROSS AIR QUALITY MONITORING STATIONS IN MALAYSIA

by

MADUKPE VINE NWABUISI

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

July 2025

ACKNOWLEDGEMENT

I am deeply grateful to my supervisor, Dr. Nur Fariha Syaqina Mohd Zulkepli, and my co-supervisor, Assoc. Prof. Dr. R. U. Gobithaasan, for their unwavering support, guidance, and encouragement throughout this research. I also sincerely thank Prof. Dr. Mohd Salmi Md Noorani for his insightful and timely feedback during the review phase, which significantly enriched this work. I want to express my heartfelt appreciation to the School of Mathematical Sciences, Universiti Sains Malaysia, for providing me with the opportunity to serve as a Graduate Research Assistant (GRA) under the Short-Term Research Grant (Project No: 304/PMATHS/6315649) and as a Graduate Teaching Assistant (GTA), which greatly supported my studies. Finally, I would like to thank the Department of Environment (DOE) Malaysia for granting access to the air quality data, a crucial component of this research.

TABLE OF CONTENTS

ACKNOWLEDGEMENT .... **ii**

TABLE OF CONTENTS .... **iii**

LIST OF TABLES .... **vii**

LIST OF FIGURES .... **viii**

LIST OF ABBREVIATIONS .... **xiv**

LIST OF SYMBOLS .... **xvi**

LIST OF APPENDICES .... **xvii**

ABSTRAK…….. .... **xviii**

ABSTRACT…... .... **xx**

CHAPTER 1 INTRODUCTION **1**

1.1 Research Background 1

1.2 Topological Data Analysis (TDA) 3

1.3 Problem Statements 6

1.4 Research Questions 6

1.5 Research Objectives 7

1.6 Scope of Study 7

1.7 Significance of Study 8

1.8 Thesis Outline 9

CHAPTER 2 LITERATURE REVIEW **10**

2.1 Introduction 10

2.2 Traditional HACA in Analysing Air Quality Data 10

2.3 Topological Data Analysis (TDA) in Environmental Studies 12

2.3.1 Conventional Mapper (CM) 15

2.3.2 Ball Mapper (BM) ... 16

2.4 Motivation of Study ... 17

CHAPTER 3 METHODOLOGY ... 19

3.1 Introduction ... 19

3.2 Datasets ... 20

3.3 Topological Data Analysis (TDA) ... 23

3.3.1 Conventional Mapper (CM) Algorithm ... 25

3.3.1(a) The Nerve Theorem ... 25

3.3.1(b) Reeb Graph ... 26

3.3.1(c) The Topological Version of CM ... 28

3.3.1(d) The Statistical Version of CM ... 29

3.3.4 Ball Mapper (BM) Algorithm ... 34

3.3.5 Graph Invariant ... 39

3.3.5(a) Estrada Index ... 40

3.3.5(b) Graph Density ... 41

3.3.5(c) Number of Cycle Basis ... 42

3.3.5(d) Fragmentation Index ... 43

3.3.5(e) Node Degree Distribution of Topological Graphs ... 44

3.4 Summary ... 45

CHAPTER 4 TOPOLOGICAL GRAPH REPRESENTATION OF AIR POLLUTANTS BEHAVIOR USING CONVENTIONAL MAPPER AND BALL MAPPER ALGORITHMS ... 47

4.1 Introduction ... 47

4.2 Simulation ... 48

4.2.1 Implementation of the CM Algorithm on the Simulated Dataset ... 50

4.2.2 Implementation of the BM Algorithm on the Simulated Dataset ... 52

4.3 PM_{10} Dataset for 2012 to 2015 and 2018 to 2020 56

4.3.1 CM Topological Graph for 2012 PM_{10} 57

4.3.2 CM Topological Graph for 2013 PM_{10} 63

4.3.3 CM Topological Graph for 2014 PM_{10} 67

4.3.4 CM Topological Graph for 2015 PM_{10} 70

4.3.5 CM Topological Graph for 2018 PM_{10} 73

4.3.6 CM Topological Graph for 2019 PM_{10} 75

4.3.7 CM Topological Graph for 2020 PM_{10} 78

4.3.8 Node Degree Distribution of Yearly PM_{10} Topological Graphs........ 80

4.3.9 Comparison between Quantitative and Qualitative Analysis of PM_{10} Behavior 82

4.4 Comparative Analysis of CM and BM in Analyzing CO, NO_2, $PM_{2.5}$, PM_{10}, O_3, and SO_2 Datasets for 2018 to 2020........ 84

4.4.1 CM and BM Topological Graphs of NO_2 86

4.4.1(a) Comparison between CM and BM for NO_2 88

4.4.2 CM and BM Topological Graph of PM_{10} 92

4.4.2(a) Comparison between BM and CM of PM_{10} 94

4.4.3 CM and BM Topological Graphs of $PM_{2.5}$ 96

4.4.4 CM and BM Topological Graph of CO 98

4.4.5 CM and BM Topological Graph of O_3........ 100

4.4.6 CM and BM Topological Graph of SO_2 102

4.5 Quantitative Analysis of CM and BM Topological Graphs for NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, and SO_2 Using Graph invariants........ 104

4.6 Comparison between TDA techniques and traditional HACA in Analyzing Air Pollutants Behavior........ 107

4.7 Summary........ 108

CHAPTER 5 DISCUSSION........ 110

5.1 Introduction.. 110

5.2 Discussion of Results.. 110

5.3 Significance of Quantitative Analysis of CM and BM Topological Graphs Using Graph Invariants.. 115

5.4 Advantages of Topological Clustering for Pollution Analysis and Control... 116

5.5 Study Limitations 118

CHAPTER 6 CONCLUSION AND FUTURE WORK................................ 119

6.1 Conclusion 119

6.2 Future Work.. 120

REFERENCES…….. 122

APPENDICES

LIST OF PUBLICATIONS

LIST OF TABLES

Page

Table 3.1 List of 60 air quality monitoring stations in Figure 3.2 21

Table 3.2 Air Pollution Index (API) 22

Table 3.3 Ambient Air Quality Guidelines (MAAQG) levels....................... 23

Table 3.4 Illustration of adjacency Table of Figure 3.12............................... 41

Table 3.5 Node degree of the topological graph in Figure 3.12 45

Table 4.1 The models used to simulate the stationary and non-stationary time series 49

Table 4.2 Descriptive statistics of PM_{10} daily average concentration for each year from 2012 to 2015 and 2018 to 2020.................................... 57

Table 4.3 Yearly comparisons of PM_{10} concentration mean differences with paired *t*-test 83

Table 4.4 Descriptive statistics of CO, NO_2, $PM_{2.5}$, PM_{10}, O_3, and SO_2 for 2018 to 2020 85

Table 4.5 Advantages and Disadvantages of CM and BM Algorithms....... 106

LIST OF FIGURES

Page

Figure 1.1 Overview of the topological data analysis Mapper process (Lum et al., 2013) 4

Figure 2.1 Datasaurus dozen (Gillespie et al., 2024) 13

Figure 2.2 Example of a topological graph 15

Figure 3.1 Research Methodology Flowchart 20

Figure 3.2 Location of 60 air quality monitoring stations across Malaysia. ... 21

Figure 3.3 Examples of simplices in different dimensions 24

Figure 3.4 Nerve construction from a point cloud and its associated simplicial complex 26

Figure 3.5 Reeb graph construction for a Torus using a height function 27

Figure 3.6 Conventional Mapper Flowchart 29

Figure 3.7 Flowchart of the PCA algorithm 31

Figure 3.8 HACA Flowchart 32

Figure 3.9 Toy P-shape point cloud used in illustrating CM algorithm implementation. 34

Figure 3.10 Ball Mapper Flowchart 35

Figure 3.11 Toy P-shape point cloud used in illustrating the BM algorithm implementation and its output 39

Figure 3.12 Toy topological graph G 40

Figure 3.13 Histogram of the node degree distribution of the topological graph in Figure 3.12 45

Figure 4.1 Example of a stationary simulated time series by AR (1) 49

Figure 4.2 Example of non-stationary simulated time series produced by ARIMA (1,1,1) 49

Figure 4.3 Topological graph of simulated time series with 5 covers and 20% overlaps produced from the CM application 50

Figure 4.4 Plot of simulated time series clustered in Nodes S1, S2, S6, S9, and S10 of simulated CM topological graph 51

Figure 4.5 Topological graphs of simulated time series by varying the parameters. (a) 5-Covers, 5% overlap, and the number of clusters is 2 (b) 5 Covers, 30% overlap and the number of clusters is 2 (c) 5 Covers, 50% overlap and the number of clusters is 2 (d) 5 Covers, 80% overlap and the number of clusters is 2 (e) 2 Covers, 30% overlap and the number of clusters is 4 (f) 4 Covers, 30% overlap and the number of clusters is 4 (g) 5 Covers, 30% overlap and the number of clusters is 4 and (h) 10 Covers, 30% overlap and the number of clusters is 4. 52

Figure 4.6 Topological graph of simulated time series generated using the Ball Mapper (BM) algorithm. 53

Figure 4.7 Simulated time series clustered within Nodes 0, 125, 131, and 139 of the Ball Mapper (BM) topological graph in Figure 4.6 with $\varepsilon = 0.04$ 54

Figure 4.8 BM topological graph of simulated time series with $\varepsilon = 0.003$ produced from the BM algorithm 55

Figure 4.9 BM topological graph of simulated time series with $\varepsilon = 0.001$ produced from the BM algorithm 55

Figure 4.10 Normalized Factor loading PC1 and partition of stations into 5 covers; Cover 0, 1, …, 4, with 30% overlap for 2012 PM_{10} 58

Figure 4.11 Graphical user interface displaying the topological graph generated using the Conventional Mapper (CM) algorithm through the KeplerMapper application. 59

Figure 4.12 2012 PM_{10} CM topological graph with 5 Covers and 30% overlaps 60

Figure 4.13 Time series plots of 2012 PM_{10} for monitoring stations clustered in Nodes A5, A3, A9, and A16 of the Conventional Mapper (CM) topological graph 61

Figure 14 Dynamic Time Warping (DTW) heatmap that measures the similarity between the time series of air quality monitoring stations clustered in Nodes A5, A2, A9, and A16 of the 2012 PM_{10} Conventional Mapper (CM) topological graph 62

Figure 4.15 Geographical locations of air quality monitoring stations clustered in the most connected Nodes (A1 through A5) of the 2012 topological graph 63

Figure 4.16 2013 PM_{10} CM topological graph with 5 Covers and 30% overlaps 64

Figure 4.17 Time series plots of 2013 PM_{10} for monitoring stations clustered in Nodes B1, B2, B11, and B8 of the CM topological graph. 65

Figure 4.18 DTW heatmap measuring the similarity between the time series of stations clustered in nodes B1, B2, B8, and B11 of the 2013 PM_{10} CM's topological graph 66

Figure 4.19 Geographical locations of air quality monitoring stations clustered in the most connected Nodes (B1 through B9) of the 2013 topological graph. 67

Figure 4.20 2014 PM_{10} CM topological graph with 5 Covers and 30% overlaps 68

Figure 4.21 Time series plots of 2014 PM_{10} for monitoring stations clustered in Nodes C7, C3, C1, and C12 of CM topological graph 68

Figure 4.22 DTW heatmap measuring the similarity between the time series of stations clustered in Nodes C1, C3, C7, and C11 of the 2014 PM_{10} CM's topological graph. 69

Figure 4.23 Geographical locations of air quality monitoring stations clustered in the most connected Nodes C1 through C6 of the 2014 topological graph. 70

Figure 4.24 2015 PM_{10} CM topological graph with 5 Covers and 30% overlaps for 71

Figure 4.25 Time series plots of 2015 PM_{10} for monitoring stations clustered in Nodes D1, D4, D7, and D15 of CM topological graph 71

Figure 4.26 DTW heatmap measuring the similarity between the time series of stations clustered in nodes D1, D4, D7, and D15 of the 2015 PM_{10} CM topological graph. 72

Figure 4.27 Geographical locations of air quality monitoring stations clustered in the most connected Nodes D1 through D12 of the 2015 topological graph. 73

Figure 4.28 2018 PM_{10} CM topological graph with 5 Covers and 30% overlaps for 74

Figure 4.29 Time series plots of 2018 PM_{10} for monitoring stations clustered in Nodes E4, E11, E8, and E13 of the topological graph 74

Figure 4.30 Geographical locations of air quality monitoring stations clustered in the most connected Nodes E1 through E4 of the 2018 topological graph. 75

Figure 4.31 2019 PM_{10} CM topological graph with 5 Covers and 30% overlaps 76

Figure 4.32 Time series of stations clustered in Nodes F7, F3, F1, and F16 of 2019 PM_{10} CM topological graph.. 76

Figure 4.33 DTW heatmap measuring the similarity between the time series of stations clustered in nodes F1, F2, and F3 of the 2019 PM_{10} CM topological graph. 77

Figure 4.34 Geographical locations of air quality monitoring stations clustered in the most connected Nodes F6 through F8 of the 2019 topological graph 78

Figure 4.35 2020 PM_{10} CM topological graph with 5 Covers and 30% overlaps for.. 79

Figure 4.36 Time series of stations clustered in Nodes G6, G4, G1, and G12 of 2020 PM_{10} CM topological graph.. 79

Figure 4.37 Geographical locations of air quality monitoring stations clustered in the most connected Nodes G1 through G10 of the 2020 topological graph. 80

Figure 4.38 Node degree distribution of 2012, 2013, 2014, 2015, 2018, 2019, and 2020 PM_{10} topological graphs 81

Figure 4.39 Spearman correlation heatmap showing the relationships between annual PM_{10} concentrations across all monitoring stations from 2012 to 2020 84

Figure 4.40 NO_2 CM topological graph with 7 Covers, 30% overlaps and 20 nodes 87

Figure 4.41 NO_2 BM topological graph with $\varepsilon = 0.078$ and 20 nodes 88

Figure 4.42 Time series of stations clustered in: (a) Node 12 of the CM graph, and (b) Node A1 of the BM graph 89

Figure 4.43 DTW heatmap measuring the similarities in the time series of randomly selected stations clustered in nodes of the CM and BM topological graph of NO_2. 90

Figure 4.44 Histogram comparing node degree distribution of BM with CM topological graphs of NO_2 91

Figure 4.45 PM_{10} CM topological graph with 4 Covers, 30% overlap, and 16 nodes 92

Figure 4.46 PM_{10} BM topological graph with $\varepsilon = 0.062$ and 16 nodes 93

Figure 4.47 Time series of the stations clustered in (a) Node B1 of the CM topological graph and (b) Node 36 of the BM's topological graph. 94

Figure 4.48 (a) DTW heatmap measuring the similarities in the time series of randomly selected stations clustered in Nodes N1, N5, and Node 36, 14 of the CM and BM topological graphs of PM_{10}, respectively. (b) Histogram comparing node degree distribution of CM with BM topological graphs. 96

Figure 4.49 $PM_{2.5}$ CM topological graph with 5 Covers, 20% overlap, and 20 nodes 97

Figure 4.50 $PM_{2.5}$ BM topological graph with $\varepsilon = 0.065$ and 19 nodes 97

Figure 4.51 Time series of the stations (a) clustered in Node H2 of the CM topological graph (b) clustered in Node 20 of the BM topological graph (c) Histogram comparing node degree distribution of BM with CM topological graphs 98

Figure 4.52 CO CM topological graph with 7 Covers and 40% overlap and 20 nodes 99

Figure 4.53 CO BM topological graph with $\varepsilon = 0.048$ and 20 nodes 99

Figure 4.54 Time series of the stations clustered in (a) Node M7 of the CM topological graph (b) Node 1 of the BM topological graph (c) Histogram comparing node degree distribution of BM with CM graphs of CO 100

Figure 4.55 O_3 CM topological graph with 5 Covers and 40% overlap and 20 nodes 101

Figure 4.56 O_3 BM topological graph with $\varepsilon = 0.06$ and 22 nodes................ 101

Figure 4.57 Time series of the stations clustered in (a) Node G1 of the CM topological graph (b) Node 2 of the BM topological graph (c) Histogram comparing node degree distribution of BM with CM 102

Figure 4.58 SO_2 CM topological graph with 7 Covers and 45% overlap, and 23 nodes.. 103

Figure 4.59 SO_2 BM topological graph with $\varepsilon = 0.014$ and 22 nodes........... 103

Figure 4.60 Time series of the stations clustered in (a) Node L7 of the CM topological graph (b) Node 37 of the BM topological graph (c) Histogram comparing node degree distribution of BM with CM graphs of SO_2.. 104

Figure 4.61 Graphs of the (a) density of BM and CM topological graphs (b) Estrada index (EE) of BM and CM topological graphs (c) Fragmentation index (FI) of BM and CM topological graphs (d) the number of cycle basis of BM and CM topological graphs.......... 105

Figure 4.62 NO_2 Dendrogram produced by traditional HACA 108

LIST OF ABBREVIATIONS

TDA	Topological Data Analysis
CM	Conventional Mapper
BM	Ball Mapper
HACA	Hierarchical Agglomerative Clustering Algorithm
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DTW	Dynamic Time Warping
DOE	Department of Environment
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation Projection
t-SNE	t-Distributed Stochastic Neighbor Embedding
O_3	Ozone
CO	Carbon monoxide
NO_2	Nitrogen dioxide
SO_2	Sulfur dioxide
PM_{10}	Particulate Matter less than 10 microns
$PM_{2.5}$	Particulate Matter less than 2.5 microns
AQI	Air Quality Index
API	Air Pollution Index
MAAQG	Malaysian Ambient Air Quality Guidelines
WHO	World Health Organization
MCO	Movement Control Order
AR	Autoregressive

MA	Moving Average
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
EE	Estrada Index
FI	Fragmentation Index
CC	Connected Components

LIST OF SYMBOLS

τ	Topology
X	Set
f	Continuous function
U	Open set or subset
α	Indexing set
$\mathcal{U}$	Cover or collection of open sets
$N(\mathcal{U})$	Nerve Complex
σ	Simplex
C_i	Cluster
$G(N,E)$	Topological graph G
N	Nodes of a topological graph
E	Edge of a topological graph
$B_\varepsilon(c_i)$	Ball centered at c_i
ε	Radius of ball
β_0	0^{th} Betti number
β_1	1^{st} Betti number
$\delta(n)$	Adjacency number
k	Node degree
$\mu g/m^3$	micrograms per cubic meter
ppm	Part Per Million

LIST OF APPENDICES

APPENDIX A	TABLE OF STATIONS CLUSTERED IN THE NODES OF THE CM TOPOLOGICAL GRAPHS

ANALISIS DATA BERTOPOLOGI DALAM MENGANALISIS KESERUPAAN KELAKUAN PENCEMAR UDARA ANTARA STESEN-STESEN PENGAWASAN KUALITI UDARA DI MALAYSIA

ABSTRAK

Pencemaran udara kini menjadi salah satu cabaran global utama yang berpunca daripada aktiviti perindustrian, pembinaan, pelepasan kenderaan dan pelbagai aktiviti manusia yang lain. Kajian ini menyelidik tingkah laku pencemar udara melalui pendekatan kualitatif dengan menggunakan teknik Analisis Data Bertopologi (ADB), iaitu Mapper konvensional (MK) dan variannya, Mapper Bebola (MB). Kaedah tersebut digunakan untuk menganalisis enam bahan pencemar udara utama (NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, dan SO_2) yang diambil dari 60 stesen pemantauan kualiti udara di seluruh Malaysia. Graf topologi yang dibina menggunakan algoritma MK mendedahkan kelompok kawasan dan stesen yang menunjukkan tingkah laku pencemar yang serupa. Hasil kajian menunjukkan persamaan dalam tingkah laku PM_{10} pada tahun 2013, 2015, dan 2020. Pendekatan ini berpotensi untuk menunjukkan visualisasi kelompok yang lebih jelas dan berintuitif, serta mencerminkan hubungan geografi berbanding kaedah pengelompokan kuantitatif tradisional seperti analisis kelompok aglomeratif berhierarki (AKAB). Struktur graf topologi ini turut dianalisis menggunakan metrik graf bagi membandingkan keberkesanan MK dan MB dalam menangkap corak pencemaran udara secara setempat dan menyeluruh. Walaupun kedua-dua kaedah ini memberikan maklumat yang bermakna dan visualisasi tingkah laku pencemar yang lebih baik, algoritma MB memberikan perwakilan yang lebih berkesan pada tingkah laku bahan pencemar gas yang kompleks. Dapatan ini terbukti melalui ketumpatan graf yang tinggi bagi graf topologi MB dengan masing-masing NO_2, O_3, CO, dan SO_2 menunjukkan nilai 15%, 10%, 32% dan 25%, manakala graf

topologi MK menghasilkan ketumpatan yang rendah, masing-masing menunjukkan 4%, 5%, 6% dan 5% untuk bahan-bahan pencemar tersebut. Secara keseluruhannya, hasil kajian ini mencadangkan pendekatan ini sebagai kaedah alternatif bagi mengenal pasti homogeniti antara stesen-stesen pengawasan kualiti udara dan seterusnya menyumbang kepada pengoptimuman pengawasan dan pengurusan kualiti udara.

TOPOLOGICAL DATA ANALYSIS IN ANALYZING THE SIMILARITY OF AIR POLLUTANTS' BEHAVIOR ACROSS AIR QUALITY MONITORING STATIONS IN MALAYSIA

ABSTRACT

Air pollution is increasingly recognized as a major global issue resulting from industrial processes, urban development, transportation emissions, and a wide range of human activities. This study investigates air pollutant behavior through a qualitative approach, utilizing topological data analysis (TDA) techniques, the conventional Mapper (CM), and its variant Ball Mapper (BM). It analyzes six major air pollutants (NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, and SO_2) collected from 60 air quality monitoring stations across Malaysia. The topological graphs generated using the CM algorithm revealed clusters of regions and stations exhibiting similar pollutant behavior. The results showed similar patterns in PM_{10} behavior across the years 2013, 2015, and 2020. This approach potentially offers a clearer and more intuitive visualization of cluster similarities, reflecting geographical relationships, compared to traditional quantitative clustering methods such as hierarchical agglomerative clustering analysis (HACA). The structure of the topological graphs is further examined using graph metrics to compare the effectiveness of the CM and BM algorithms in capturing local and global pollutant trends. While both methods provided valuable insights and better visualizations of pollutant behavior, the BM algorithm offered a more effective representation of the complex behavior of gaseous pollutants. This is evident in the higher graph densities of the BM topological graph for NO_2, O_3, CO, and SO_2, which are 15%, 10%, 32%, and 25%, respectively. In comparison, the CM topological graph of the same pollutants yields lower densities of 4%, 5%, 6% and 5%, respectively.

Overall, the findings suggest that this approach may serve as a viable alternative for identifying homogeneity in air quality monitoring stations, potentially contributing to the optimization of air quality monitoring and management.

CHAPTER 1

INTRODUCTION

1.1 Research Background

Air pollution has become a significant global challenge in recent decades, primarily due to rapid urbanization and industrialization, particularly in developing nations (Shi et al., 2020; Wu and Chang, 2020). The rising levels of air pollution and the challenges in managing it effectively are significant concerns (Zeng et al., 2019; Wang et al., 2020). As a developing nation, Malaysia also contends with increasing air pollution issues, which pose significant risks to public health and the environment (Usmani et al., 2020; Chin et al., 2019). Addressing this issue requires implementing effective mitigation strategies and conducting comprehensive analyses to better understand pollutant behavior through air quality monitoring and data analysis (Sofia et al., 2020). Air quality monitoring is essential for mitigating and managing the effects of air pollution within any ecosystem (Gulia et al., 2020).

This environmental challenge stems from the release of harmful pollutants into the atmosphere, which harms both human health and the natural environment (WHO, 2024). Major pollutants affecting air quality include particulate matter (PM_{10} and $PM_{2.5}$), which can reduce visibility and penetrate deep into the lungs and bloodstream, causing respiratory and cardiovascular diseases (Areal et al., 2022; Maleki et al., 2022; Latif et al., 2018). Additionally, nitrogen dioxide (NO_2) and sulfur dioxide (SO_2) emissions, primarily from vehicle exhaust and industrial activities, contribute to the formation of acid rain and smog, thereby exacerbating air quality deterioration. Ground-level ozone (O_3), a secondary pollutant, exacerbates health issues such as asthma, while carbon monoxide (CO), primarily emitted from the combustion of fossil fuels, poses significant health risks (Peng et al., 2021). The accumulation of these pollutants harms human health, affects wildlife, damages vegetation, and contributes to climate change, highlighting the need for more analytical tools to monitor and investigate pollutant behaviors across designated air quality monitoring stations.

Investigating the behavior of pollutants between air quality monitoring stations is essential for enhancing the efficiency of air quality management. In Malaysia, the Department of Environment (DOE) monitors and generates large amounts of data by recording the daily average concentrations of six major air pollutants: NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, and SO_2 (DOE, 2024). However, the vast geographical area of Malaysia leads to varying air quality behavior, creating challenges in understanding spatial and temporal pollution trends for more accurate air quality forecasting and management (Naidin et al., 2024). Identifying regions with similar air pollution behaviors is crucial for accurately locating local emissions, reducing redundant air quality monitoring stations, and assisting decision makers and stakeholders in implementing effective pollution control measures to enhance air quality management (Stolz et al., 2020).

However, the volume of data generated from pollutant concentration measurements across numerous monitoring stations poses a substantial challenge for practical analysis. The complexity of the dataset requires advanced techniques to analyze and extract hidden information effectively. Traditional quantitative methods have proven helpful in this regard, each offering unique insights in the context of environmental monitoring. Time series analysis has been applied to monitor the trend of pollutants over time, enabling researchers to predict future trends of these pollutants (Yang et al., 2021). Descriptive statistics have been applied to summarize pollutant concentrations within specific areas, providing a clear overview of the dataset (Ibe et al., 2020). Also, correlation analysis reveals significant associations between pollutant concentration levels and their health impacts (Jalili et al., 2020). Regression analysis is used to predict pollutant concentrations based on various predictors, including weather conditions, traffic density, geographical location, seasonal variation, and industrial activities (Vasseur & Aznarte, 2021). In addition, hierarchical agglomerative clustering analysis (HACA), k-means, and DBSCAN clustering methods classify areas based on similar air quality patterns (Govender & Sivakumar, 2020; Zhang and Yang, 2022; Suris et al., 2022). These methods identify distinct groups of stations with similar pollutant behavior and potential sources, such as industrial activities and high-traffic areas, enabling the development of targeted policies for effective pollution control (Zhao et al., 2016; Chen et al., 2015). This approach also provides valuable insights into the quantitative characteristics of pollutants. However, traditional clustering methods are often lacking in qualitative aspects and in uncovering hidden insights inherent in the datasets. As a result, there is a need for more analytical methods capable of extracting this hidden qualitative information and providing a simpler representation of the dataset, such as topological data analysis (TDA) techniques.

1.2 Topological Data Analysis (TDA)

TDA is an emerging field of mathematics that focuses on uncovering hidden patterns in high-dimensional datasets, offering qualitative insights into their structure and visualization (Chazal & Michel, 2021). TDA employs persistent homology and the Mapper algorithm to extract the underlying "shape" of data, emphasizing the idea that data has shape and shape has meaning (Todd & Petrov, 2022). This thesis focuses on the Mapper algorithm illustrated in Figure 1.1. Mapper applies a filter function to partition point cloud data into overlapping subsets or covers. Within each partition, similar data points are clustered, and connections are drawn between clusters that share common data points. The resulting graph consisting of nodes and edges, captures the topological structure of the dataset by revealing patterns and relationships within the data. This graphical representation provides clearer and more interpretable visual insights into high-dimensional data.

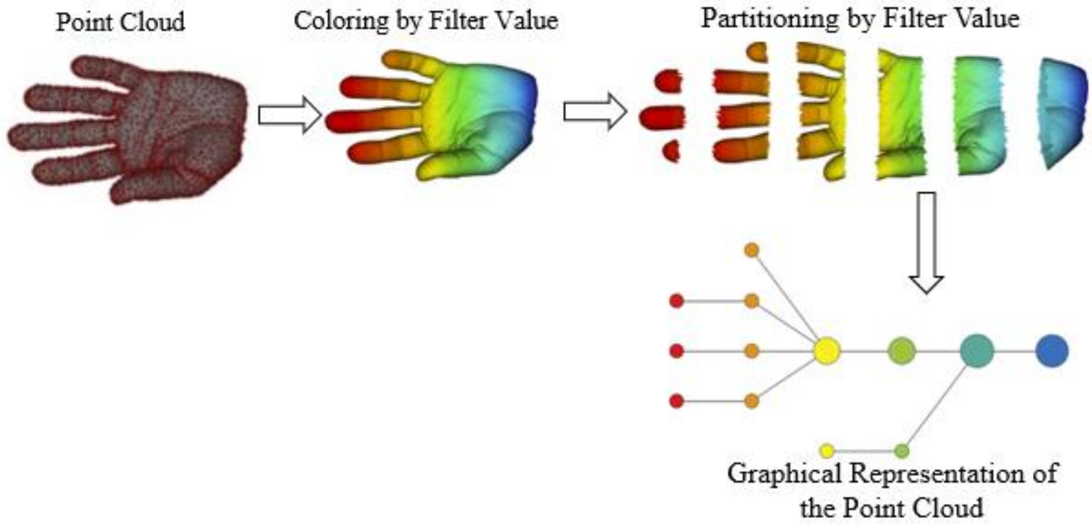

Figure 1.1 Overview of the topological data analysis Mapper process (Lum et al., 2013)

This emerging field has demonstrated a wide range of applications across various disciplines. In biology, Campbell et al. (2019) employed the TDA Mapper for feature extraction in pain recognition based on physiological signals, addressing the problem of inconsistent feature selection in emotion recognition systems. In neurology, Caputi et al. (2021) utilized persistent homology to differentiate between normal and abnormal brain states. In the area of flood detection, Musa et al. (2021) and Ohanuba et al. (2021) proposed using persistent homology as a preprocessing technique to enhance the early detection of critical transitions, particularly flooding events in rivers. Ofori Boateng et al. (2021) applied persistent homology to compare aerosol optical depth (AOD) maps across different spatial resolutions without requiring regridding. Within the financial sector, Ismail et al. (2022) introduced a novel method for detecting early indicators of financial crises using persistent homology. Yang et al. (2024) applied the technique to improve the robustness of the Light Gradient Boosting Machine (LightGBM) algorithm in image classification tasks.

These studies represent just a few of the many diverse applications of TDA across scientific and technological domains.

Although the TDA approach is increasingly applied across various fields, its potential remains underexplored in the study of air quality and pollution. A noteworthy study in this area was conducted by Zulkepli et al. (2022), who utilized TDA, specifically persistent homology, to extract topological features from air quality datasets for categorizing days predominantly affected by haze episodes. Persistent homology is a technique within TDA that focuses on tracking the birth and death of topological features in a high-dimensional dataset (Malott et al., 2020; Corbet et al., 2019). A clustering algorithm is often integrated with the topological information provided by persistent homology to produce a better cluster (Otter et al., 2017). Persistent homology output is visualized through persistence diagrams or barcodes. In contrast, Mapper generates a topological graph that visually represents relationships between clusters through nodes and edges.

This thesis uses the Mapper algorithm and its variant Ball Mapper to analyze six major air pollutants in Malaysia, with the former regarded as the conventional approach. The conventional Mapper introduced by Singh et al. (2007) acts as a method for clustering and visualizing datasets based on their topological structure. It employs multiple parameters to generate a simplified representation of the dataset as a network structure, referred to as a topological graph. Ball Mapper simplifies the process of the conventional Mapper by using a single parameter. This parameter is the radius of a closed ball that covers all data points, consequently producing a topological graph. This approach could enable a comparative exploration of both algorithms in capturing and visualizing the underlying patterns and relationships within air quality datasets.

1.3 Problem Statements

Traditional clustering methods, such as hierarchical agglomerative clustering analysis (HACA), are commonly used to group air quality monitoring stations that exhibit similar pollutant behaviors. While these methods perform well in grouping data based on similarity, they are often limited in revealing topological relationships within the dataset. Additionally, the resulting clustering output can be complex to interpret, especially for high-dimensional datasets. This challenge can be addressed using TDA techniques such as the conventional Mapper algorithm.

The conventional Mapper algorithm requires multiple parameters to be set up to construct a topological graph. In contrast, its variant, the Ball Mapper algorithm, relies on a single-parameter greedy algorithm to produce a topological graph. This reduced parameter dependency of the conventional Mapper algorithm can enhance the construction of topological graphs that reflect pollutant dynamics. Despite this potential, conventional Mapper and Ball Mapper algorithms have been underexplored in air quality studies.

Finally, a gap remains in comparative analysis between traditional quantitative clustering and qualitative TDA-based methods, such as the conventional Mapper and Ball Mapper, in terms of simplified visualization of air pollutant behavior. A comparative evaluation using graphs can help determine which method better captures pollutant behavior and provides clearer visual insights.

1.4 Research Questions

This thesis addresses the following research questions:

1. To what extent does the conventional Mapper algorithm capture similarities in pollutant behavior and reveal geographical relationships between air quality monitoring stations?

2. How effective is the Ball Mapper algorithm through a topological graph in representing complex air pollutant dynamics based on the data shape?

3. How do HACA, conventional Mapper, and Ball Mapper differ in visually representing clusters of monitoring stations according to pollutant behavior?

1.5 Research Objectives

To address the research questions, the thesis focused on the following key objectives :

1. To construct topological graphs using the conventional Mapper algorithm that elucidate yearly PM_{10} behavior and highlight the geographical relationships among air quality monitoring stations, based on topological structure.

2. To construct air pollutants topological graphs using the Ball Mapper algorithm and assess its effectiveness in capturing pollutant complex dynamics.

3. To compare the performance and visualization capabilities of traditional HACA, conventional Mapper, and Ball Mapper algorithms in clustering air quality monitoring stations based on pollutant behavior using graph invariants.

1.6 Scope of Study

This thesis examines the behavior of air pollutants across six regions in Malaysia using a qualitative approach. It focuses on pollutant concentration data collected from air quality monitoring stations between 2012-2015 and 2018-2020. To guide this investigation, the study is structured around three main objectives. The first objective investigates the behavior of PM_{10} using data from 52 air quality monitoring stations, analyzed through conventional Mapper topological graphs. The most connected components (subgraphs) are visualized on a map of Malaysia to show how geographically dispersed stations exhibit similar pollution patterns based on the underlying data structure.

The second objective broadens the scope to include six pollutants: NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, and SO_2, using data from 60 monitoring stations between 2018 and 2020. Both the conventional Mapper and Ball Mapper algorithms are applied to generate topological graphs that capture the dynamic patterns of each pollutant based on the data shape.

Finally, the third objective focuses on a comparative analysis of the topological graphs generated by the conventional Mapper and Ball Mapper algorithms. Graph invariants, such as Graph Density, Fragmentation Index, Estrada Index, and Node Degree Distribution, are used to evaluate the structural characteristics of topological graphs. These metrics are used to assess which of the two algorithms, the Conventional Mapper or the Ball Mapper, better represents the behavior of each pollutant based on domain knowledge and literature. Additionally, the visualization and interpretability of topological graphs are compared with those generated by a traditional HACA dendrogram.

1.7 Significance of Study

This thesis aims to develop an efficient qualitative air quality analysis tool that can assist the Department of Environment (DOE) in identifying potential redundancies in monitoring stations and considering reallocations to more critical areas, thereby improving coverage and effectiveness. By leveraging topological structures, this study examines the potential for adopting a centralized approach to pollution control, based on stations clustered in a node and those in connected nodes.

TDA has the potential to provide valuable insights into evaluating the effectiveness of environmental policies, such as emission regulations, by analyzing changes in pollutant dynamics over time. By comparing topological graph structures before and after policy implementation, this approach may help identify whether pollutant behavior stabilizes, improves, or reverts to its previous pattern once the policies are no longer in effect. Additionally, this research offers an alternative perspective for environmental scientists and mathematicians interested in exploring new approaches to air quality research and management. Finally, it also provides qualitative insights that complement traditional statistical methods. Combining these approaches could help policymakers develop a more comprehensive understanding of air quality dynamics, contributing to more informed and evidence-based decision-making.

1.8 Thesis Outline

This thesis is structured into six comprehensive chapters, each contributing to a detailed exploration of air quality analysis through the traditional clustering method, HACA, and TDA techniques, namely, CM and BM algorithms. Chapter 2 offers an extensive literature review, focusing on the application of traditional HACA in air quality studies. It also explores the use of TDA techniques, including the CM and BM algorithms, across various scientific fields, providing a solid foundation for their application in this research. Chapter 3 introduces the air quality datasets used in the study, along with a detailed explanation of the methodologies and algorithms employed in the analysis. Chapter 4 presents the results from the analysis, providing interpretations that align with the research objectives. In Chapter 5, these findings are discussed in the context of the study's objectives, emphasizing their significance and addressing the limitations encountered during the study. The thesis concludes with Chapter 6, which synthesizes the key insights and offers recommendations for future research, paving the way for further exploration in the field of air quality management.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a literature review structured into four sections, providing a comprehensive overview of the methodologies employed in topological data analysis for air quality studies. Section 2.2 provides an overview of the hierarchical agglomerative clustering analysis (HACA) approach, summarizing its applications and contributions to air quality data analysis. Section 2.3 reviews the applications of TDA as a qualitative analysis approach across various fields, highlighting its potential as a valuable tool for analyzing air quality datasets. This is followed by a review of its tools, including CM and BM techniques. Additionally, Subsection 2.3.1 reviews studies that utilize the CM algorithm, highlighting its specific role and effectiveness in identifying patterns within a dataset. Finally, Section 2.3.2 explores the application and performance of the BM algorithm, highlighting its effectiveness in generating a simplified and enhanced visual representation of datasets. The chapter concludes in Section 2.4 with a discussion of the research motivations.

2.2 Traditional HACA in Analysing Air Quality Data

According to Govender and Sivakumar (2020), clustering is a data analysis technique used to explore and identify patterns in the underlying structure of a dataset by grouping data points with similar characteristics. Over the years, clustering techniques have been widely utilized in atmospheric science, with significant applications in climate and meteorological studies (Pampuch et al., 2023; Handhayani & Lewenusa, 2024). This technique groups data points into homogeneous clusters based on proximity measures derived from temporal information (Mehta et al., 2020). Numerous studies have demonstrated the effectiveness of cluster analysis in examining spatial patterns of air quality monitoring sites. For instance, Zhao et al. (2016) proposed a trajectory clustering method that effectively identifies hotspots within trajectory data, providing valuable insights for urban transportation planning and management.

Additionally, Vasudha and Rao (2023) applied HACA to classify air pollution monitoring stations in India into four distinct clusters based on average Air Quality Index (AQI) levels, highlighting the significant impact of topographical conditions on air pollution patterns. Alia et al. (2022) effectively classified air quality monitoring stations in Peninsular Malaysia based on the pollutant distribution using HACA. Zhang and Yang (2022) employed HACA to identify trends in $PM_{2.5}$ pollution and regional divisions across China, thereby aiding in the development of coordinated control measures for clean air policies. Stolz et al. (2020) optimized air quality monitoring stations in Mexico by identifying redundant stations. Xu et al. (2022) classified Chinese cities according to pollution levels, revealing the most and least polluted areas. Other significant studies include works by Lu et al. (2011), Zhang et al. (2016), Pires et al. (2008), Jorquera and Villalobos (2020), Rahman et al. (2022), Fang et al. (2022), and Suris et al. (2022).

While traditional HACA approaches have been successful in analyzing air pollutant behavior, interpreting the relationships between clusters can be challenging. The HACA dendrogram can sometimes be difficult to interpret, particularly when working with high-dimensional datasets, where inter-cluster relationships may be unclear, as observed by Ezugwu et al. (2020), Ran et al. (2022), and Huang et al. (2021). These challenges associated with interpreting high-dimensional datasets highlight the need for alternative approaches, particularly qualitative methods that can offer clearer and more intuitive visualizations of clusters and their interrelationships.

To address limitations in traditional HACA, several studies have explored modified approaches. Particularly in air quality studies, Zulkepli et al. (2022) proposed a hybridization of HACA with persistent homology to assess haze-related air pollution in Southeast Asia. By replacing Euclidean distance with Wasserstein distance and incorporating persistent homology, the method captures deeper spatial structures. Results show improved clustering accuracy and more precise identification of patterns in elevated particulate matter levels compared to traditional HACA.

Finally, traditional HACA can be used as a clustering algorithm within the conventional Mapper framework to extract qualitative insights from complex datasets based on their shape. However, despite the effectiveness of this method it remains insufficiently applied in air quality research as evidenced by the existing literature. This highlights a contextual methodological gap in applying topological data analysis tools, particularly the Mapper algorithm, within the field of air quality research.

2.3 Topological Data Analysis (TDA) in Environmental Studies

TDA is an emerging field that leverages concepts from algebraic topology to analyze the shape and underlying structure of datasets (Carlsson, 2020). Its growing application across various studies provides fresh insights and innovative methods for data analysis (Corcoran & Jones, 2023). TDA, being a qualitative approach, highlights the topological features within datasets, revealing insights that traditional quantitative methods might overlook (Musa et al., 2021).

The ability of TDA to uncover hidden patterns and shapes in a dataset emphasizes the importance of qualitative approaches that focus on the topological and geometric structures within a dataset. Rather than relying solely on quantitative and numerical summaries, which do not capture the inherent shapes in a dataset. As shown in Figure 2.1, point clouds such as a dinosaur, circle, bullseye, wide lines, star, horizontal lines, scattered dots, and cross each have distinct shapes; yet, when analyzed using traditional quantitative methods, they yielded similar summary statistics. For example, each of the datasets has the same X Mean of 54.26, Y Mean of 47.83, X Standard Deviation of 16.76, Y Standard Deviation of 26.93, and a correlation of -0.06. This similarity in summary statistics, despite vastly different shapes, was observed by Matejka and Fitzmaurice (2017) and further examined by Gillespie et al. (2024).

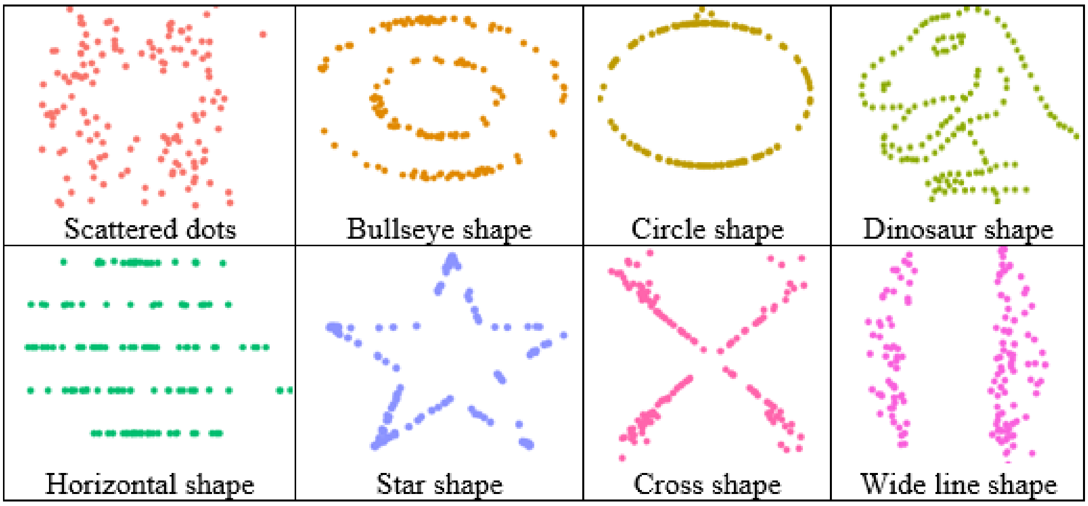

Figure 2.1 Datasaurus dozen (Gillespie et al., 2024)

TDA can capture complex patterns and relationships within a dataset that may be hidden by similar statistical values, enabling a deeper understanding of the dataset's structure. The TDA approach has proven to be effective in various fields of research. For instance, Chazal and Michel (2021) utilized TDA persistent homology to analyze high-dimensional biological data, including gene expression and protein interaction networks. Pereira et al. (2021) applied TDA Mapper to investigate the resilience of social networks. Hernández-Lemus et al. (2024) explored the application of TDA persistent homology in analyzing cardiovascular signals, highlighting its potential to improve understanding of cardiovascular diseases and their treatment. El-Yaagoubi et al. (2023) introduced TDA persistent homology concepts to a statistical audience, presenting an approach for analyzing multivariate time series data. Skaf and Laubenbacher (2022) reviewed emerging mathematical methods, including TDA, to enhance the ability of clinicians and researchers to analyze biomedical data. Finally, Uray et al. (2024) conducted a comprehensive survey on the applications of TDA in industrial production and manufacturing settings, as well as in other areas.

TDA has also been applied to various environmental science studies. For example, Islambekov et al. (2019) demonstrated that integrating persistent homology with existing change-point detection algorithms enhances the accuracy of identifying distributional regime shifts in environmental time series. Similarly, Ofori-Boateng et al. (2021) applied TDA persistent homology to compare aerosol optical depth (AOD) maps across varying spatial resolutions, revealing its strength in systematically evaluating spatial patterns without the need for resampling. Using AOD datasets from the Goddard Earth Observing System, this study highlights the applicability of TDA in Earth science studies. Additionally, Munch (2017) illustrated how TDA outputs, such as persistence diagrams and conventional Mapper topological graphs, enable quantification of shape and structure in data, making these techniques accessible to researchers in education, learning sciences, and various scientific domains. TDA persistent homology has been applied to the detection of early warning signals and floods (Musa et al., 2021; Ohanuba et al., 2021), as well as in the clustering of rainfall stations (Gobithaasan et al., 2022).

Recently, Zulkepli et al. (2022) utilized persistent homology to analyze air quality data, identifying the most and least polluted areas during haze events by extracting topological features. The study builds on earlier work by Zulkepli et al. (2020). Overall, the literature on TDA application in environmental studies remains limited despite its significant potential to address environmental challenges and handle high-volume data generated through environmental monitoring. TDA offers promising opportunities for deriving new insights through topological representations. This thesis aims to apply TDA tools, specifically the conventional Mapper and its variant the Ball Mapper algorithm to analyze air quality datasets.

2.3.1 Conventional Mapper (CM)

The CM algorithm proposed by Singh et al. (2007) has become a powerful tool in TDA for visualizing high-dimensional datasets. It transforms complex datasets into simplified topological graphs composed of nodes and edges, thereby creating a simpler representation of the dataset (Munch, 2017). As shown in Figure 2.2, each node represents a cluster of points with similar characteristics, while the edges connecting nodes indicate subtle relationships between points in the connected clusters. Node colors are assigned based on specific attributes or features of interest, enabling the visualization of additional insights into patterns within the dataset.

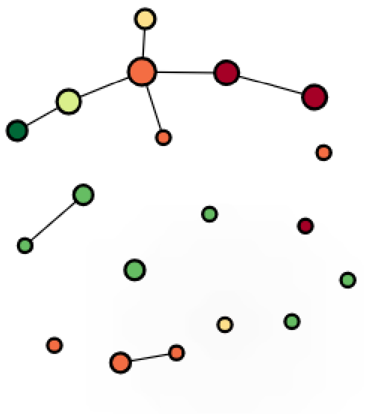

Figure 2.2 Example of a topological graph

The CM algorithm approach requires multiple parameters to be set up, including the filter function, cover, overlap percentage, and clustering algorithm, to produce topological graphs (Geniesse et al., 2019). CM has been successfully applied in various research areas, including the spread of coronavirus (Chen & Volić, 2021), medicine (Caputi et al., 2021), core gene expression (Palande et al., 2023), and mapping technological spaces (Escolar et al., 2020). These applications demonstrate the CM algorithm ability to uncover hidden patterns and inter-cluster relationships that are often limited in traditional methods, making it a promising tool for air quality analysis.

The CM algorithm enhances traditional clustering techniques by offering a qualitative visualization of clustering results, representing the dataset as a topological graph. This approach provides a clearer understanding of complex datasets by simplifying them into nodes and edges, revealing patterns and relationships that may not be evident through conventional methods. This thesis aims to leverage the CM algorithm to provide a novel qualitative perspective on air pollutant data, enhance air quality monitoring management, and support decision-makers in formulating more effective pollution control strategies.

2.3.2 Ball Mapper (BM)

Recently, Dlotko (2019) introduced BM, a variant of the CM technique in TDA. BM, a more straightforward topological and geometric-based approach, captures the shape of the dataset by grouping neighborhood points using balls of equal radius through a greedy or Max-min algorithm without relying on a rigid clustering algorithm to produce a topological graph. Nodes are connected by edges when their respective balls intersect, indicating shared members. This approach yields a topological graph, similar to CM, where nodes represent clusters of similar data points (Dłotko et al., 2024). Unlike the CM algorithm which requires multiple parameters the BM algorithm uses a single parameter, the ball radius which is use as a cover in the algorithm. This improvement makes BM easier to implement while still providing a rich informational output.

Current studies that utilize the BM algorithm include Rudkin et al. (2023), which employed the BM algorithm to analyze regional voting behavior in the UK's Brexit referendum. Their study revealed that Leave votes were more geographically concentrated and homogeneous compared to “Remain” votes. The findings, consistent across different administrative levels, provided new insights into the regional economic and cultural factors influencing Brexit, demonstrating the effectiveness of BM in uncovering hidden spatial voting patterns. Dłotko et al. (2021) also employed the algorithm to visualize complex, non-monotonic relationships between key financial ratios and stock returns. It reveals hidden structures in asset pricing, illustrates how machine learning models capture return patterns, and identifies segments with significant abnormal returns, demonstrating the BM potential to reveal shape, connectivity, and spatial relationships within a dataset. However, the literature on BM is still emerging, and its full capabilities are yet to be extensively explored. The BM technique provides a clear and insightful approach to understanding complex datasets, effectively revealing inherent relationships and patterns within the data (Rudkin et al., 2023).

2.4 Motivation of Study

The motivation for this study lies in exploring the potential of TDA tools, notably the CM and BM algorithms, in environmental science, where they remain underutilized despite their demonstrated potential. Also, the TDA qualitative approach offers a unique perspective on complex data, highlighting structural and relational patterns that traditional quantitative methods often overlook. While persistent homology has shown success in environmental studies and recently in air quality analysis, there is limited research on CM and BM specifically within the context of air pollutant behavior. The CM and BM methodologies discussed in Chapter 3 enable a simplified yet informative visualization of high-dimensional datasets, which we believe will provide air quality researchers and managers with an alternative tool for monitoring and understanding pollutant patterns.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Figure 3.1 is the flowchart of the methodology employed in this thesis. The process begins with data acquisition and pre-processing, as presented in Section 3.2, followed by both qualitative and quantitative approaches. The TDA method, discussed in Section 3.3, comprises two techniques: the CM and BM algorithms. The first objective is addressed by applying the CM algorithm presented in Section 3.3.3 to construct topological graphs. These CM topological graphs aim to elucidate the yearly behavior of PM_{10} concentrations while capturing the underlying topological structure among air quality monitoring stations.

The second objective focuses on the application of the BM algorithm, a variant of CM, to construct alternative topological graphs of NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, and SO_2. By employing the BM algorithm discussed in Section 3.3.4, the study aims to provide a complementary lens through which to present and interpret air quality dynamics. The resulting BM topological graph is analyzed in the context of its structural and visual expressiveness.

In Objective Three, a comparative analysis of the CM and BM topological graphs constructed using the corresponding datasets will be implemented, utilizing the graph invariants discussed in Section 3.3.5. Also, the visualization of the topological graphs is compared against the HACA dendrogram. This comparative analysis evaluates the relative strengths and limitations of each method in presenting and interpreting the clustering behavior of air monitoring stations.

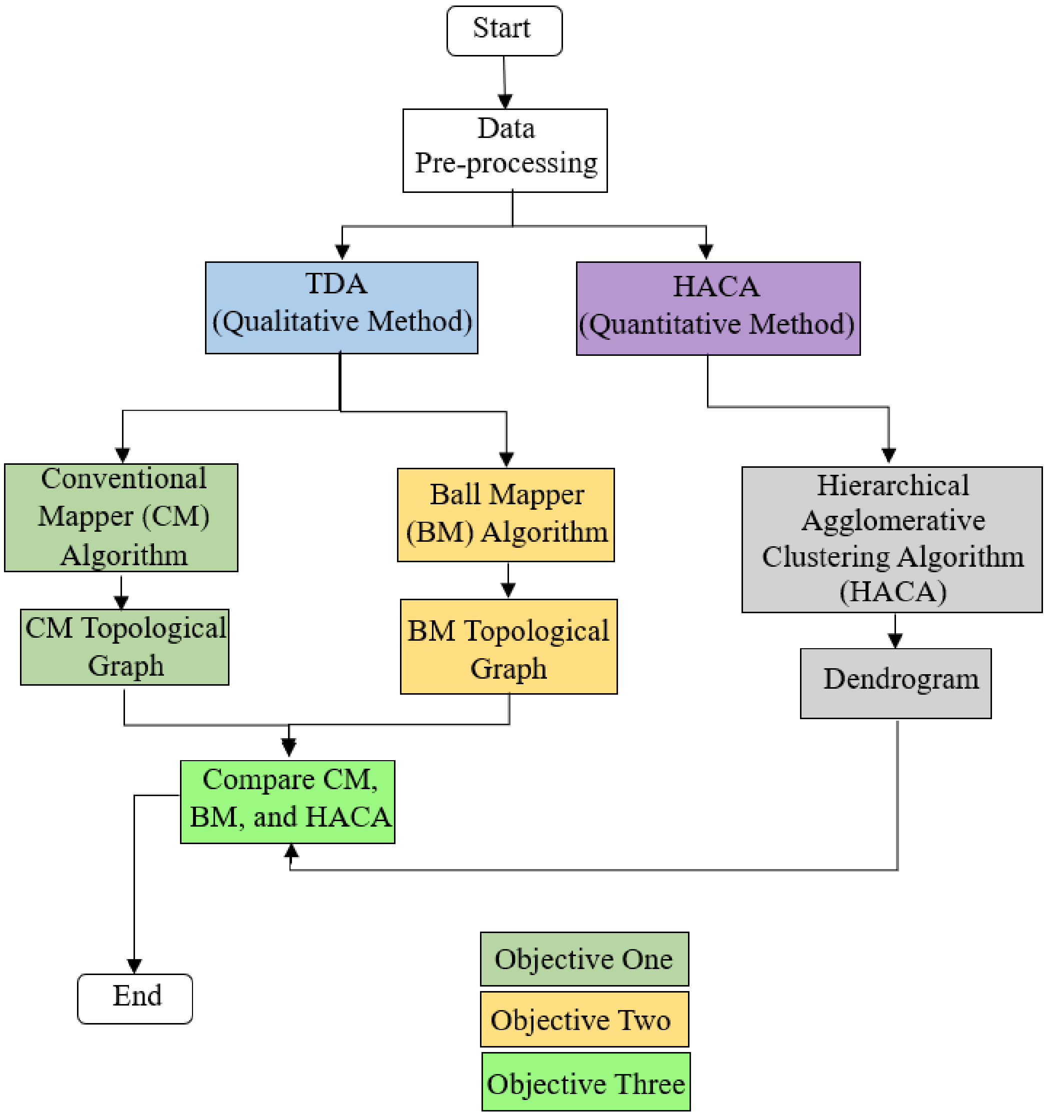

Figure 3.1 Research Methodology Flowchart

3.2 Datasets

The datasets used in this study are secondary data obtained from the Department of Environment (DOE) Malaysia (DOE, 2024). These data are collected from 60 air quality monitoring stations across the six regions, namely Northern, Central, Eastern, Southern, Sabah, and Sarawak in Malaysia. These stations are strategically located in the urban, suburban, rural, and industrial areas. Figure 3.2 presents a map of Malaysia, illustrating the locations of these stations, with different colors indicating the regional categorizations. The names of these stations, with their respective index number, are listed in Table 3.1. These categorizations serve as labels in the topological graph to investigate the interrelation of cluster similarities within the context of geographical relationships.

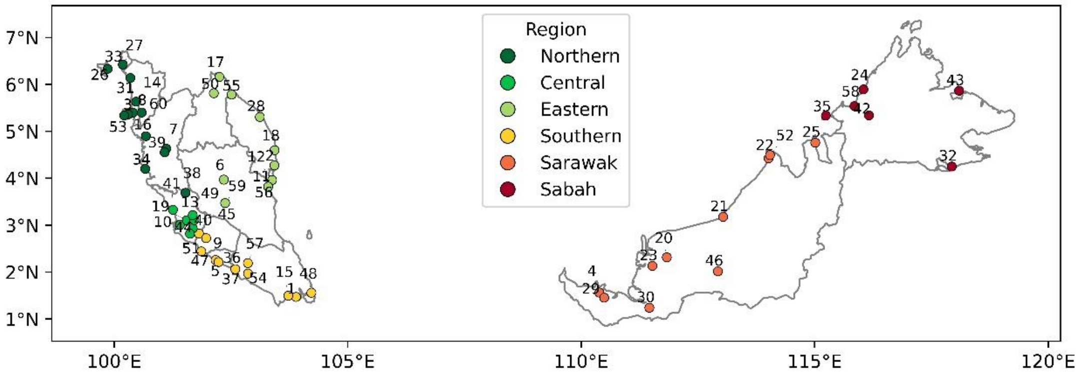

Figure 3.2 Location of 60 air quality monitoring stations across Malaysia.

Table 3.1 List of 60 air quality monitoring stations in Figure 3.2

No.	Station	No.	Station	No.	Station
1	Pasir_Gudang	21	Bintulu	41	Kuala_Selangor
2	Kemaman	22	Miri	42	Keningau
3	Perai	23	Sarikei	43	Sandakan
4	Kuching	24	Kota_Kinabalu	44	Putrajaya
5	Bukit_Rambai	25	Limbang	45	Cheras
6	Jerantut	26	Langkawi	46	Kapit
7	Tasek	27	Kangar	47	Port_Dickson
8	Seberang_Jaya	28	Kuala_Terengganu	48	Kota_Tinggi
9	Nilai	29	Samarahan	49	Batu_Muda
10	Klang	30	Sri_Aman	50	Tanah_Merah
11	Kuantan	31	USM	51	Banting
12	Balok_Baru	32	Tawau	52	ILP_Miri
13	Petaling_Jaya	33	Alor_Setar	53	Balik Pulau
14	Sungai_Petani	34	Seri_Manjung	54	Batu Pahat
15	Larkin	35	Labuan	55	Besut
16	Taiping	36	Bandaraya_Melaka	56	Indera
17	Kota_Bharu	37	Muar	57	Segamat
18	Paka	38	Tanjung_Malim	58	Kimanis
19	Shah_Alam	39	Pegoh	59	Temerloh
20	Sibu	40	Seremban	60	Kulim

The DOE monitors six major air pollutants nitrogen dioxide (NO_2), particulate matter with diameters less than 10 micrometers (PM_{10}), particulate matter with diameters less than 2.5 micrometers ($PM_{2.5}$), ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2) across 60 designated air quality monitoring stations, as listed in

Table 3.1. This thesis examined the daily average of PM10 datasets from selected years (2012, 2013, 2014, 2015, 2018, 2019, and 2020). However, the 2016 and 2017 datasets were excluded from the analysis because they had over 20% missing data. Also, the datasets containing average daily concentrations for all six pollutants (NO_2, PM_{10}, $PM_{2.5}$, O_3, CO, and SO_2) from 2018 to 2020 are utilized.

In Malaysia, the DOE reports air quality using the Air Pollution Index (API), a measure that communicates air quality levels to the public. The API summarizes the concentrations of various pollutants into a single number, providing a clear and straightforward way to understand air pollution levels and their potential health effects. Based on the air quality status supplied by the DOE in the Environmental Quality Report 2020 (DOE, 2024), the API status is categorized into five classes, as shown in Table 3.2.

Table 3.2 Air Pollution Index (API)

API	Air Quality Status
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very unhealthy
>300	Hazardous

Based on the API status, Unhealthy days occur when pollutant concentrations significantly exceed the limits set by the Malaysian Ambient Air Quality Guidelines (MAAQG). These guidelines establish specific thresholds for the daily average concentrations of various pollutants, providing a benchmark to assess the air quality on public health. As shown in Table 3.3, pollutant levels that exceed the benchmark may pose potential health risks to the public.

Table 3.3 Ambient Air Quality Guidelines (MAAQG) levels

Pollutant	Unhealthy Threshold level
PM_{10}	100 µg/m³ (24-hour average)
$PM_{2.5}$	35 µg/m³ (24-hour average)
SO_2	0.03 ppm (24-hour average)
NO_2	0.05 ppm (24-hour average)
CO	8.75 ppm (8-hour average)
O_3	0.05 ppm (8-hour average)

Based on the air quality dataset described in this section, this thesis employs TDA techniques to explore the underlying shape and patterns within the dataset qualitatively. The following section introduces the fundamental concepts of topological data analysis.

3.3 Topological Data Analysis (TDA)

TDA is a framework that applies principles from algebraic topology to study the shape of data. It emphasizes the properties of spaces and shapes that are invariant under continuous transformations such as stretching or bending. To formally introduce the basic concepts in TDA, it begins with the definition of a topology (Definition 3.1) and topological space (Definition 3.2), which provides the mathematical framework for studying shape and structure.

Definition 3.1 *Topology τ is the mathematical study of properties of spaces and shapes that remain unchanged under continuous transformations, such as stretching, bending, and twisting, while excluding alterations like tearing or gluing* (John, 2020). This provides the foundational language for analyzing geometric and spatial relationships in TDA. With the notion of topology established, a topological space is formally defined as follows:

Definition 3.2 *A topological space (X,τ) is a set X equipped with a topology τ, where τ is a collection of subsets of X, called open sets that satisfy the following axioms* (Sakai, 2020):

1. $\varnothing \in \tau$ and $X \in \tau$

2. τ is closed under arbitrary union: For any index set A if $\{U_\alpha\}_{\alpha \in A} \subseteq \tau$, then: $\bigcup_{\alpha \in A} U_\alpha \in \tau$.

3. τ is closed under finite intersection: if $U_1, U_2, ..., U_n$ are sets in τ, then: $\bigcap_i^n U_i \in \tau$.

The basic building blocks in TDA are simplices, the geometric objects that generalize points, line segments, and triangles to arbitrary dimensions, with definitions provided below:

Definition 3.3 *a simplex is the simplest possible shape in a given dimension, serving as the foundation of a simplicial complex* (Barrabés & Mateu-Figueras, 2006). For instance, a 0-simplex represents a point, a 1-simplex an edge, and a 2-simplex a filled triangle, as illustrated in Figure 3.3.

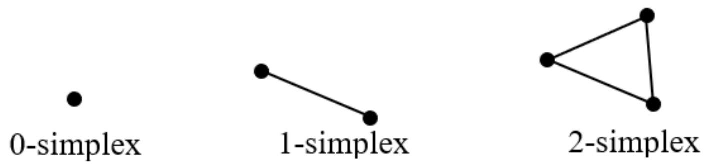

Figure 3.3 Examples of simplices in different dimensions

Topological summaries of high-dimensional datasets are constructed using various tools in TDA that reveal the underlying structure or “shape” of the data. Among these tools, the Mapper algorithm is commonly used to transform high-dimensional data into simplified graphical representations. This method enables visual interpretation of patterns, relationships, and clusters within the dataset. The following section introduces the conventional Mapper algorithm and provides a brief discussion on its topological and statistical versions.