TRAFFIC CONGESTION PATTERNS IDENTIFICATION USING GPS DATA ANALYTICS

DIO TONY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN ROBOTICS AND COMPUTATIONAL INTELLIGENCE SYSTEMS

SCHOOL OF ENGINEERING

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2024

KMITL-2024-EN-M-407-268

Thesis	Traffic Congestion Patterns Identification using GPS Data Analytics
Student	Mr. Dio Tony
Student ID.	63601042
Degree	Master of Engineering
Program	Robotics and Computational Intelligence Systems
Year	2024
Thesis Advisor	Assoc. Prof. Dr. Rathachai Chawuthai

ABSTRACT IN ENGLISH

Traffic congestion is a significant problem encountered worldwide in large urban areas. The issue intensifies during rush hours, leading to an expansion of congestion and deterioration of the traffic infrastructure. Each city has its own distinct traffic network, and the habits of its inhabitants influence its traffic flow. As such, a versatile method for identifying congestion patterns is needed. We suggested using Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) to identify the spread of traffic congestion through the analysis of congestion length distributions. Global Positioning System (GPS) trackers installed in taxis were utilized to identify traffic flow within Bangkok. The GPS probe data was divided into "prior" and "later" periods before undergoing clustering. Congestion hotspots identified from both periods were transformed into a congestion area, from which congestion lengths were derived based on three distinct approaches. Similarity comparisons of congestion length distribution were made against Longdo Traffic's list of the top 100 most congested roads in Bangkok. The results were promising on peak hours datasets, demonstrating the potential of HDBSCAN to significantly contribute to traffic management research.

ACKNOWLEDGEMENT

I express my deepest gratitude to Allah SWT for His blessings, which have enabled the successful completion of this research, conference, and thesis. His boundless mercy has guided me through this journey, fostering a supportive environment and conditions. I am immensely grateful for the unwavering support and guidance from my parents, friends, and advisor, which have been instrumental in my achievement.

I extend my heartfelt appreciation to Assoc. Prof. Dr. Rathachai Chawuthai, for his mentorship, compassion, and the invaluable learning opportunities provided throughout this academic endeavor as my advisor. Additionally, I am indebted to my university, King Mongkut's Institute of Technology Ladkrabang (KMITL) for fostering an environment conducive to personal and academic growth.

The completion of this thesis is a testament to the dedication, perseverance, and discipline cultivated during my master's degree studies in Thailand. I am profoundly grateful for the transformative experiences and the knowledge gained during this journey.

Dio Tony

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS AND SYMBOLS

1D	1 Dimensional
2D	2 Dimensional
3D	3 Dimensional
BMR	Bangkok Metropolitan Region
CCT	Causal Congestion Tree
CCTV	Close circuit television
d_{ij}	Distance between GT and synthesised data
DLT	Department of Land Transport
DOT	Department of Transportation
ETA	Estimation Time of Arrival
EXAT	Expressway Authority of Thailand
FHV	For Hire Vehicle
FHWA	Federal Highway Administration
f_{ij}	Mass of transported goods (congestion length)
GAN	Generative Adversarial Network
GPS	Global Positioning System
GT	Ground Truth
H_Max	HDBSCAN with maximum-based congestion length extraction
H_Mean	HDBSCAN_Mean-based congestion length extraction
H_Min	HDBSCAN with minimum-based congestion length extraction
HDBSCAN	Hierarchical Density Based Spatial Clustering of Applications with Noise
HDBSCAN(eom)	HDBSCAN with excess of mass as cluster stability measure
HDBSCAN(leaf)	HDBSCAN with `cluster_selection_method='leaf'`
IoT	Internet of Thing
iTIC	Intelligent Traffic Information Center Foundation
KL	Kullback-Leibler
km	Kilometre
km/h	Kilometer/hour
LH	Later Congestion Hotspots
LH_{avg}	Later Congestion Hotspots with coordinate based on average value

M	Average between ground truth and synthesised distribution
MB	Megabytes
NECTEC	National Electronics and National Information Center
NYC	New York City
OPTICS	Ordering Points To Identify Cluster Structure
P	Probability distribution of the ground truth
PH	Prior Congestion Hotspots
PH_{avg}	Prior Congestion Hotspots with coordinate based on average value
p_i	Length of congestion in distribution P
POI	Point Of Interest
Q	Probability distribution of the synthesised data
q_j.	Length of congestion in distribution Q
RSSI	Received Signal Strength Indicator
SVM	Support Vector Machines
TF1	Time Frame 1
TF2	Time Frame 2
TICM	Traffic Index Cloud Maps
TLC	Taxi and Limousine Commission
U.S.	United States
u_i	Location of data point in distribution P
V2I	Vehicle to Infrastructure
V2V	Vehicle to Vehicle
VBSCAN	Variable Density Based Spatial Clustering of Applications with Noise
v_j	Location of data point in distribution Q
X_i	Datapoint
X_{lat}	X-variable latitude
X_{lon}	X-variable longitude
X_{max}	Datapoint with the largest value within a dataset
X_{nm}	Normalised value against maximum value data
Y_{lat}	Y-variable latitude
Y_{lon}	Y-variable longitude

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND OF PATTERN IDENTIFICATION

Traffic congestion poses a significant challenge to sustainable urban development, impacting the economy, society, and the environment [1]. The periods of intensified traffic congestion, typically resulting in slow-moving or standstill traffic, are commonly referred to as “peak hours” and occur twice daily on weekdays. In Bangkok, the morning peak hours span from 6:00 to 9:00, while the evening peak hours extend from 16:30 to 19:30 [2]. In 2022, motorists in Bangkok experienced an additional of 93 hours of travel time for a 10 km commute due to peak hour traffic [3].

The primary cause of traffic congestion on urban roads and expressways often attributed to “traffic bottlenecks”. These bottlenecks are characterised by spatial discontinuities that result in a reduction of road capacity. [4]. Identifying and defining traffic bottlenecks is particularly challenging in urban traffic networks due to various complicating factors such as the complexity of road network topology, the contingency of congestion, and diverse travel behaviours [4].

Nowadays, traffic networks, their components, and their conditions can be easily monitored through navigation assistance applications such as Google Maps, or Waze, which make use of data from numerous resources. Typically, these applications represent the analysed data of traffic conditions in specific economic areas through traffic flow speed by using the colour-coded representation of road infrastructure. Given the availability of public and private dataset within the internet, it is possible to reconstruct the traffic conditions of a particular point of interest within a specific time to perform traffic monitoring and analysis efforts.

Identification of traffic congestion conditions involve numerous different procedures but commonly employs pattern recognition techniques. The decision in designing and employing the suitable pattern recognition techniques will determine the scalability of the algorithm (system as whole) in processing data from bigger area of interest, while maintaining acceptable, if not stable performance in synthesising the traffic congestion pattern.

1.2 PROBLEM DESCRIPTION

Each city has a distinct traffic network, and its traffic patterns are shaped by the behaviour of its residents. The traffic flow in traffic monitoring services commonly represented by colour code comprises of “green”, “yellow”, “red”, and “black”, indicating the range of traffic movement speed starting from free flow to standstill consecutively. Sample of traffic networks within Bangkok and their conditions are illustrated in Figure 1.

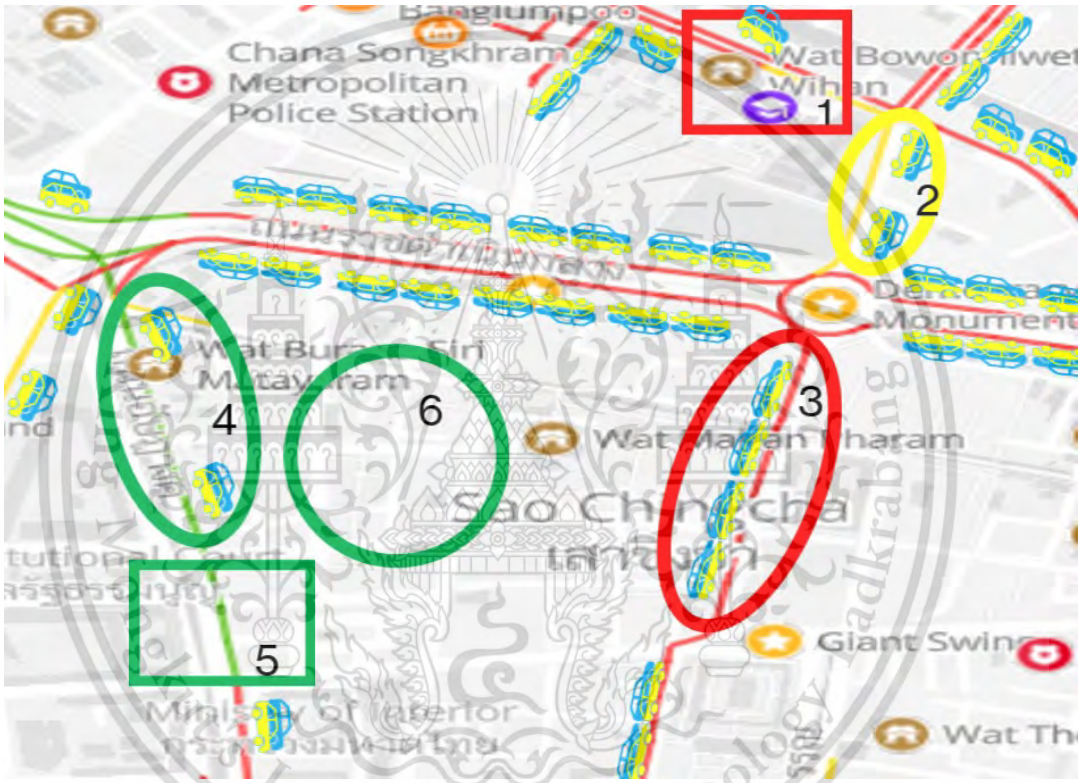


Figure 1. Several Factors Affecting Traffic Conditions Around Bangkok

Variation in traffic flow speeds illustrated within the traffic networks were the result of interactions among multiple traffic components, such as government infrastructure, religious facilities, and road capacity. These interactions are the cause for the variety in the observed traffic density ranging from standstill traffic (red coloured line) to free-flow traffic (green coloured line) within the illustrated traffic networks.

Often, research on traffic condition monitoring is performed at a junction level with 4 roads meeting. Such traffic model has limited capacity and insufficient when applied to the larger area of traffic network. Meanwhile, traffic congestion is the result of complex interaction between few junctions within a traffic network at the least level. The distinct traffic densities that are formed, resulted in unique traffic characteristics that are challenging to be represented with explicitly written algorithm.

1.3 RESEARCH OBJECTIVE

Development of flexible approach in monitoring over large traffic network is crucial for effective traffic monitoring. The objectives of this research are listed below:

1. Study the pattern of traffic congestion propagation through utilisation of unsupervised clustering technique.
2. Study the pattern of traffic congestion propagation through congestion length within congested area.
3. Study the pattern of traffic congestion propagation through utilisation of taxi GPS probe.

1.4 SCOPE OF STUDY

This study is focusing on the traffic congestion length, congested area, and congestion length density. The boundaries within this study are specified as below:

1. Research is conducted as Proof of concept (Technical Readiness Level: 0).
2. One technique will be used to unsupervised clustering technique.
3. Maximum of three congestion length extraction techniques will be proposed and compared.
4. Study will be performed to identify pattern of congestion within peak hours in one day only.
5. Does not concern on visualisation and comparison of congested segments.
6. All data are assumed to be produced by taxi operating on main roads.
7. Does not concern about computing resources.

8. Neglect errors from the usage of spherical and ellipsoidal earth model interchangeably.
9. Only uses spatial data for pattern generation.
10. Only uses taxi GPS probe data that is operating with passenger to ensure reliability of data used because such condition introduce more similar driving behaviour to private passenger cars.

1.5 EXPECTED CONTRIBUTION

The proposed approaches and methods are expected successfully provide length of congestion within a particular congestion area. Throughout the process, methods and techniques performed are expected to contribute towards the learning of:

1. Filtering process of false congestion data.
2. Utilisation of HDBSCAN and its role in pattern discovery.
3. Performance measurement techniques for pattern similarity.

CHAPTER 2
THEORY AND LITERATURE REVIEW

2.1 DATA

Data is seen as a key resource in the “knowledge economy”, a term which refers to economic system that is highly dependent on knowledge, and highly skilled labours [4]. Data is crucial in development of effective solutions and decision-making process. Generally, data can be defined as collection of natural or man-made (fabricated) events that are recorded in specific formats. In the scientific sectors, data can be further classified in to four different sources which are recorded in Table 1.

Table 1. Categorisation of Data in Scientific Sectors [5]

No	Data Categorisation	Source of data	Example of Application
1	Observation-based disciplines	Natural phenomena	Health Care, Geoscience, Meteorology, etc.
2	Experiment Results	Results from human-conducted experiments	Clinical trials, pH test, flow table test, etc.
3	Simulated data	Large-scale simulations	Digital Twins, etc.
4	Reference data	Highly curated datasets	Human genome, etc.

Data in its unprocessed (raw) form is not valuable because it cannot convey any information that will aid in knowledge development and wisdom for decision making. Data need to be processed, often with multiple techniques that are dependent upon the researchers’ knowledge and experience to reveal its value. The distinction of data’s worthiness within the different stages of value creation can be described using DIKW’s knowledge management model which is shown in Figure 2.

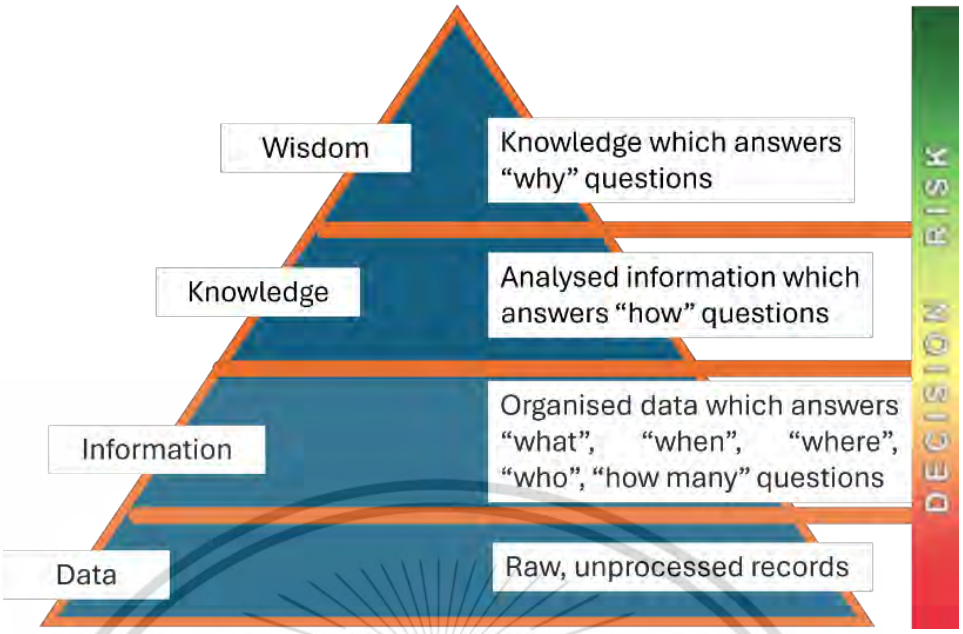


Figure 2. DIKW Knowledge Management Model and Its Value Creation during Decision Making Process

2.1.1 Data in Transportation Engineering and Urban Planning

The data format used within research mostly recorded in the form of textual records, sounds, images, and numerical scores [5]. In the field of traffic management, data is mainly derived from images and numerical scores. This raw data is then processed with multiple techniques, however, processing traffic data by reproducing techniques that works in other locations does not necessarily produce good result, because the behaviour of traffic is unique in different locations due to its traffic networks conditions and infrastructures.

One of few success cases of traffic management research by using image data was published by Li et al. [6] which has successfully developed supervised learning-based traffic congestion detection by image processing through extraction of histogram features on the images of city's Traffic Index Cloud Maps (TICM). Such method is claimed to achieve F1 score of 0.981 on Decision Tree based for a city-wide traffic network. However, at the time of writing, it is not possible to obtain similar data for both clean TICM images (without "Labels" in Google Maps) and details of traffic infrastructures (number of lanes, length of road, free flow speed and average speed of a particular road) within a city.

Various technologies have been implemented for traffic monitoring and data collection to improve transportation planning and traffic management. These technologies include loop detectors, Closed-Circuit Television (CCTV), Vehicle-to-Vehicle (V2V) communication, Vehicle-to-Infrastructure (V2I) communication, the Internet of Things (IoT), and Global Positioning System (GPS), among others. However widespread adaptation of those technologies has been hindered due to operational cost, occasionally leading to proprietary data and/or data commercialisation. On the other hand, the cost of GPS no longer a factor now [7], resulting in widespread deployment, particularly in navigation assistant applications. GPS' widespread deployment also gives rise to open dataset, which is a valuable resource in research field, thus facilitating the development of effective solutions in enhancing transportation engineering and urban planning.

2.1.2 Open GPS Data

Ground Positioning System (GPS) unit provides real-time spatial (geographic coordinates: latitude, longitude, and sometimes altitude) and temporal measurement (timestamp) data in the form of numerical values. GPS data is a valuable resource within the traffic monitoring system, providing real-time data which can be processed to indicate traffic conditions. Consequently, traffic congestion prediction and other alike traffic monitoring efforts can be produced by performing advance analysis on the presented traffic condition data, resulting in the development of effective traffic management solutions.

Monitoring over traffic conditions, especially traffic congestion condition is one of the most crucial elements in transportation engineering and urban planning field of study. Traffic congestion poses a significant hurdle to sustainable urban growth as it impacts the overall environment, especially the society and its economy [1]. Presently, Google Maps is considered as one of the most frequently used application for traffic monitoring and navigation assistance with, 10 billion downloads recorded solely on Google Play [8].

Google has been crowdsourcing congestion data from mobile phone's GPS (Google Maps) since 2009 [9]. Today, Google Maps has at least 97% accuracy on Estimation Time of Arrival (ETA) for its navigation assistant service. Such performance is powered by huge historical and real-time data, multiple services and technologies

including traffic condition monitoring and prediction algorithms [10]. Given the potential competitive advantage that could be gained with transportation related GPS data and its potential harm from privacy exploitation, public transportation GPS data are among the only few publicly available resources on GPS data.

Dataset on public transportation's GPS mainly made of bus and taxi. Taxi's GPS data offers better representation of traffic condition than bus's due to unrestricted navigation path on the road. Bus is a mode of transportations that focuses on mass transportation that is operating on a preplanned route, occasionally with a dedicated lane to reduce the effect of traffic congestion and maintain its performance in term of travel time. Further, taxi has greater quantity on the road due to its flexible scheduling and door-to-door service while partially supported by the demand for privacy, and personal space. There are two known providers for taxi GPS dataset which are given in the Table 2.

Table 2. List of Open Taxi GPS Datasets and Its Provider

No	Name	Details
1	New York City (NYC) Taxi and Limousine Commission (TLC) [11]	NYC TLC maintains datasets for various types of vehicles for hire, including for hire vehicles (e.g. limousines, etc), green taxis, and yellow taxis, among others. Around 1,000,000 trips recoded daily from 200,000 TLC licensees collectively [12].
2	Intelligent Traffic Information Center Foundation (iTIC) Thailand [13]	iTIC responsible for gathering and analysing public traffic data from both governmental (e.g. CCTV, etc.) and private sources including mobile phone probes, buses, taxis, and logistics data [14]. It hosts datasets for Thailand Location, Historical traffic incidents, Historical traffic information status of Thailand, and Historical raw data from mobile phone and vehicles and mobile probes in Thailand [15].

2.1.3 Vehicle Classification

The traffic network comprises of multiple infrastructures that are planned to facilitate the movement of goods and people such as road, bridge, highways, traffic signals, among others. The Department of Transportation (DOT) of New York City reported that their traffic network consists of street and highways with total length over 10 km [16]. In contrast, Bangkok's traffic network amounts to a total over 5 km long [17].

Traffic congestion is also contributed by poor infrastructure, which fails to provide sufficient service against the volume of vehicle. In 2023, a trip of 10 km drive took 21 minutes and 40 seconds in Bangkok while it took extra 170 seconds to finish similar trip in New York. Consequently, these cities were ranked 46th and 20th [18], respectively, in the global ranking of city with traffic congestion. The area that is commonly affected by traffic congestion within New York and Bangkok are illustrated in Figure 3 and Figure 4 respectively.

Figure 3. New York Typical Traffic [19]

Figure 4. Bangkok Typical Traffic [20]

Fundamental data on vehicle volume is needed to assist development of effective solutions for traffic challenges. One of such case is the usage of vehicle volume to design suitable pavement material. In the United States (U.S.), the Federal Highway Administration (FHWA) develops vehicle classification systems to organise the vehicle volume data in their data recording system to assist their pavement engineer in formulating materials that is durable and suitable for the operation of vehicles that are passing a particular section of the street [21].

Standardisation of vehicle classification system is an effort by governments through their transportation department in recording vehicle population which aims to achieve the following:

1. Uniformity: Standard ensures that the same categories and definitions are used across different regions, resulting in consistent data collection.
2. Mutual Understanding: The standardised vehicle classifications allows collaboration from agencies with different expertise.
3. Targeted Effort: Classification allows agencies to develop solutions that target specific challenges.

The vehicle classification standards is a continuously evolving regulation and exhibit variability in their implementation. In the U.S., the primary standard of vehicle classification is established by FHWA of the U.S. Department of Transportation, by utilising axle distance-based classification [21]. Individual states are allowed to devise their own standards guided by FHWA standards, which may include the factors such as gas emission, engine displacement, vehicle dimension, gross vehicle weight, among others. Conversely, Thailand employs a uniform standard developed by its Department of Land Transport (DLT). Vehicle classification standards of FHWA and DLT are shown in Appendix C and Appendix D, respectively.

2.1.4 Vehicle Population Publication

Frequency of vehicle population report publication are important to facilitate public and government agencies in formulating initiatives upon the changes of automotive demographic. Regular release of report provides foundational data for subsequent studies in formulating transportation and urban planning solutions. The FHWA releases their report annually [22] while the DLT release theirs on monthly basis [23] which are shown in Figure 5 and Figure 6, respectively.

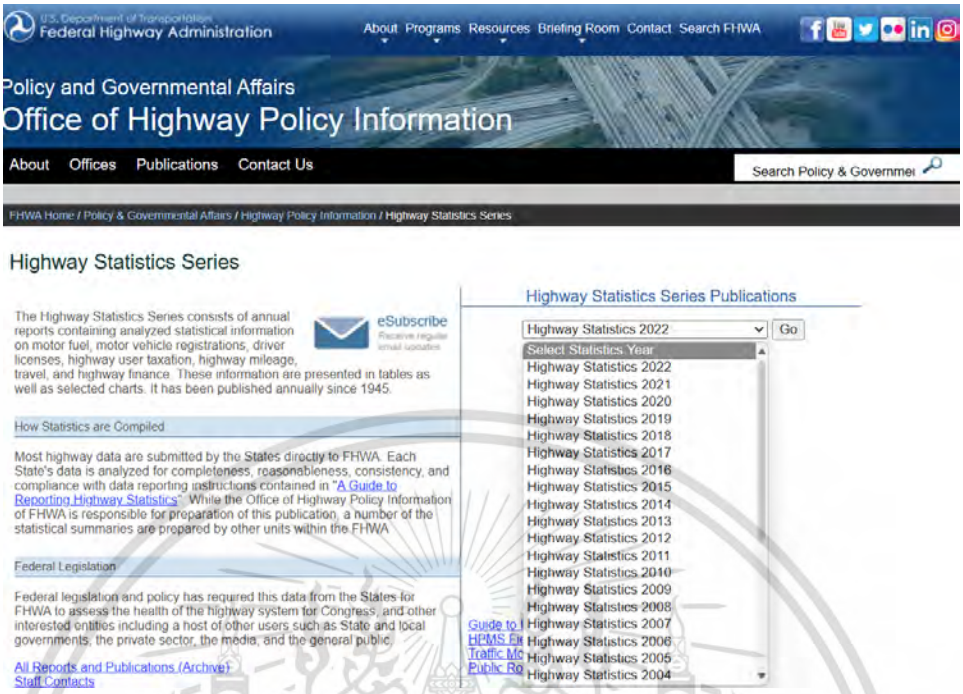


Figure 5. Publishing Frequency of Vehicle Population by FHWA [22]

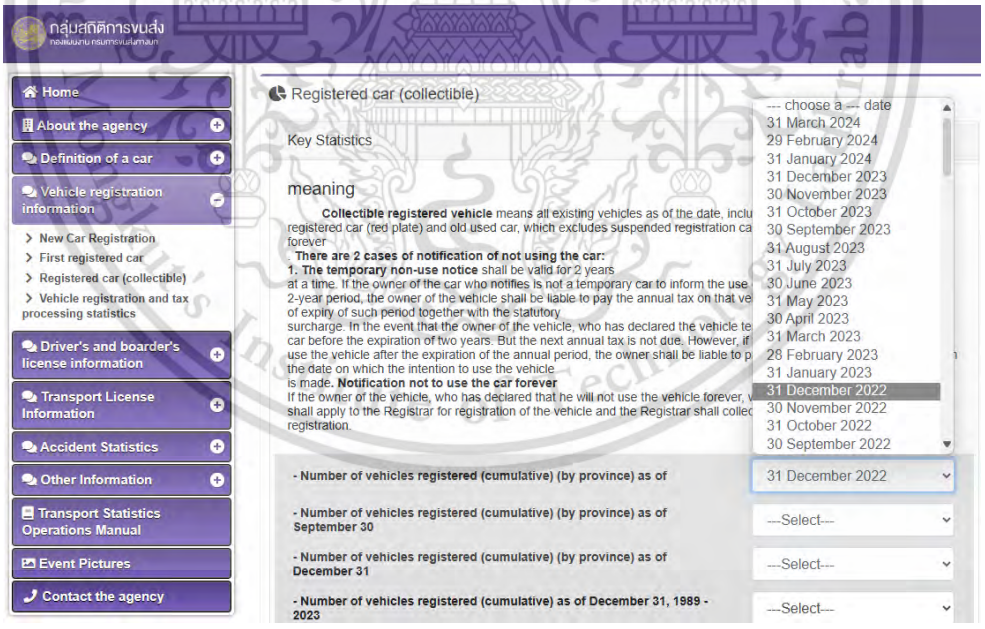


Figure 6. Publishing Frequency of Vehicle Population Report by DLT [23]

In the year 2022, New York reported vehicle population of 9,111,362 units while Bangkok was having 11,217,177 units of vehicle. The snapshot of vehicle population in New York and Bangkok are illustrated by Figure 7 and Figure 8, with complementary of vehicle population ratio shown by Figure 9 and Figure 10, respectively.

STATE	AUTOMOBILES			BUSES			TRUCKS			MOTORCYCLES			ALL MOTOR VEHICLES		
	PRIVATE AND COMMERCIAL (INCLUDING TAXICABS)	PUBLICLY OWNED	TOTAL	PRIVATE AND COMMERCIAL	PUBLICLY OWNED	TOTAL	PRIVATE AND COMMERCIAL	PUBLICLY OWNED	TOTAL	PRIVATE AND COMMERCIAL	PUBLICLY OWNED	TOTAL	PRIVATE AND COMMERCIAL	PUBLICLY OWNED	TOTAL
Alabama (2)	1,995,247	1,206	1,996,453	3,258	137	3,395	3,330,864	7,860	3,338,724	125,810	-	125,810	5,455,179	9,203	5,464,382
New Mexico	585,077	19,000	604,077	2,164	4,631	6,795	1,156,344	46,283	1,202,627	56,502	379	56,881	1,800,087	70,293	1,870,380
New York (3)	2,955,805	12,436	2,968,241	77,460	5,295	82,755	5,286,843	37,519	5,324,362	736,000	4	736,004	9,056,108	55,254	9,111,362

Figure 7. Vehicle Population in New York in 2022 [24]

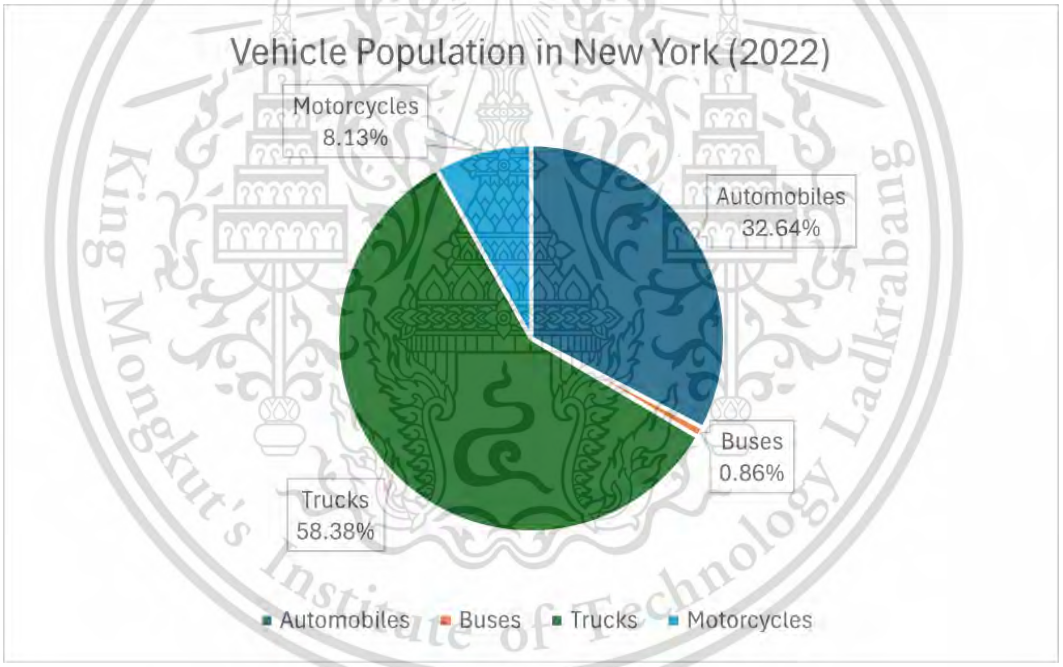


Figure 8. Ratio of Vehicle Population in New York in 2022

ประเภทรถ Type of Vehicle	ทั่วประเทศ Whole Kingdom	กรุงเทพฯ Bangkok
รวมทั้งสิ้น Grand Total	43,394,104	11,617,177
ก. รวมรถตามกฎหมายว่าด้วยรถยนต์ Total Vehicle under Motor Vehicle Act	42,035,133	11,423,635
รย. 1 รถยนต์นั่งส่วนบุคคลไม่เกิน 7 คน Sedan (Not more than 7 Pass.)	11,344,873	5,310,266
รย. 2 รถยนต์นั่งส่วนบุคคลเกิน 7 คน Microbus & Passenger Van	445,862	229,652
รย. 3 รถยนต์บรรทุกส่วนบุคคล Van & Pick Up	7,085,910	1,491,442
รย. 4 รถยนต์สามล้อส่วนบุคคล Motortricycle	1,362	743
รย. 5 รถยนต์รับจ้างระหว่างจังหวัด Interprovincial Taxi	-	-
รย. 6 รถยนต์รับจ้างบรรทุกคนโดยสารไม่เกิน 7 คน Urban Taxi	82,499	79,455
- บุคคลธรรมดา	27,189	26,313
- นิติบุคคล	54,762	52,596
- ไม่ระบุ	548	546
รย. 7 รถยนต์สี่ล้อเล็กรับจ้าง Fixed Route Taxi	2,247	1,730
รย. 8 รถยนต์รับจ้างสามล้อ Motortricycle Taxi (Tuk Tuk)	18,622	9,107
รย. 9 รถยนต์บริการธุรกิจ Hotel Taxi	3,610	450
รย.10 รถยนต์บริการทัศนาจร Tour Taxi	4,093	1,591
รย.11 รถยนต์บริการให้เช่า Car For Hire	77	52
รย.12 รถจักรยานยนต์ส่วนบุคคล Motorcycle	22,137,636	4,098,759
รย.13 รถแทรกเตอร์ Tractor	630,103	115,575

Figure 9. Vehicle Population in Thailand and Bangkok in 2022 [23]

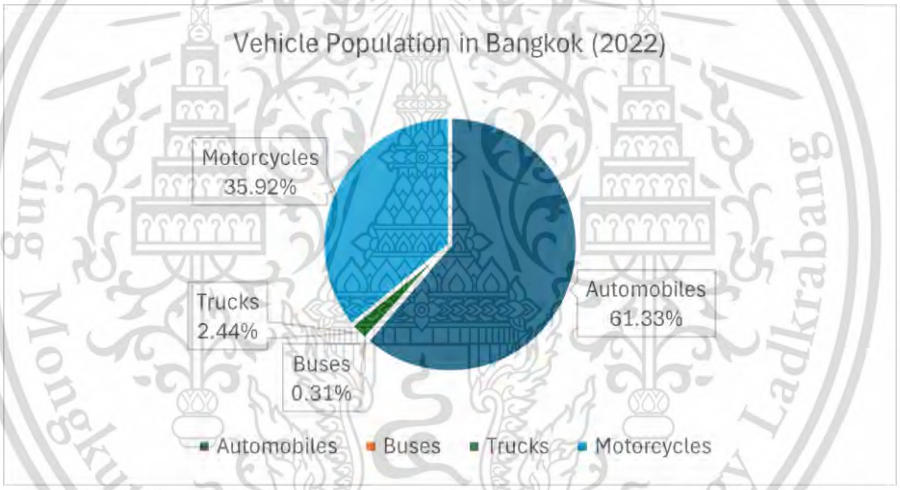


Figure 10. Ratio of Vehicle Population in Bangkok in 2022

Automobile categories used by the FHWA in their report refers to vehicle under class 2 and class 3 (All cars, cars with one-axle trailers, cars with two-axle trailers Pick-ups and vans, pick-ups and vans with one- and two- axle trailers) with total vehicle of 2,968,241 units [24]. Applying the same standard into DLT's classification will yield total vehicle of 7,124,133 units for Bangkok's automobile class (DLT's class: 1, 2, 3, 5, 6, 7, 8, 9, 10, and 11). Truck is defined as vehicle with two axles and six tyres, or vehicle at least three axles by New York's Department of Transport[25]. Truck population is higher than automobiles (passenger car) in New

York while in Bangkok, automobiles has the highest population among the vehicle populations.

Sedan is the dominant class of vehicle within the DLT's equivalent class of "automobile" in Bangkok. Thus, it is logical to assume that sedan (passenger cars) is among the main contributors of traffic congestion, leading to the urgency in investigating traffic congestion from the perspective of passenger car. Under the class 1 category of the DLT, sedan classification includes vehicle with 12 m in length.

2.1.5 Vehicle Sales Volume

Vehicles purchase directly related to the quantity of vehicle registration recorded within the vehicle population report. In December 2022, MarkLines Automotive Industry Portal, an online platform for automotive industry information data search platform, reported that vehicle of class 1 (DLT standard) is topping the sales chart in Thailand[26]. Coincidentally, the same models of vehicles also championing the sales records of the year 2022 in Thailand [27]. The details of the top selling vehicles in Thailand are given in Table 3, meanwhile the U.S.'s in Table 4.

Table 3. Top 10 Vehicle Sales and Their Volume in Thailand for Year 2022

No	Vehicle	Length (mm)	Width (mm)	Dimension Classification	Sales Number (unit) [27]
1	Isuzu D-Max	5265	1870	Pickup Truck	178,407
2	Toyota Hilux	5325	1855	Pickup Truck	142,578
3	Honda City	4574	1748	Sedan	45,791
4	Ford Ranger	5355	2180	Pickup Truck	28,689
5	Toyota Yaris	4425	1730	Sedan	28,157
6	Toyota Corolla	4630	1790	Sedan	27,414
7	Toyota Fortuner	4795	1855	SUV	26,869
8	Toyota Yaris Ativ	4425	1730	Sedan	26,394
9	Mitsubishi Triton	5305	1815	Pickup Truck	21,168
10	Isuzu MU-X	4850	1870	SUV	19,029

Table 4. Top 10 Vehicle Sales and Their Volume in the U.S. for Year 2022

No	Vehicle	Length (mm)	Width (mm)	Dimension Classification	Sales Number (unit) [28]
1	Ford F-Series	5,311	2,029	Pickup	653,957
2	Chevrolet Silverado	6,387	2,263	Pickup	520,936
3	Ram Pickup	5,890	2,062	Pickup	468,344
4	Toyota RAV4	4,625	1,854	SUV	366,741
5	Toyota Camry	4,895	1,839	Sedan	295,201
6	GMC Sierra	5,890	2,062	Pickup	241,521
7	Honda CR-V	4,628	1,856	SUV	238,155
8	Tesla Model Y	4,750	1,978	SUV	225,799
9	Jeep Grand Cherokee	4,915	1,969	SUV	223,344
10	Toyota Highlander	4,953	1,930	SUV	222,805

2.1.6 Rationale for the Data Selection

The usage of open GPS dataset is justified because GPS probe of (private) passenger vehicle is almost, if not, impossible to attain. Taxi datasets are chosen because taxis are capable of unrestricted road access, maximising the representation of traffic conditions. Further, Thailand is chosen as the focus of the research study because of the availability of benchmarking data, which was not readily available for New York from providers of navigation and traffic monitoring services such as Inrix, xmap, and TomTom traffic, probably hidden behind paywall.

Taxis are typically made with a capacity to accommodate 5-7 people. Taxi in Thailand falls under class 6 within the DLT vehicle classification with dimension not exceeding 2.5 m of width and 6 m of length., thus, it can be assumed to be equivalent to the criteria of passenger car in class 1 because the top selling vehicle models in Thailand have dimension that are within the criteria of class 6 vehicle.

The utilisation of the taxi dataset provided by the iTIC for research purpose as replacement of sedan (private passenger car) dataset due to unavailability, therefore, possible to be performed. Further, taxis are a preferred mode of public transport representing 17% of total public transport trips in the year 2017 in Bangkok [29], second only to public buses.

Moreover, the frequent publishing interval of the DLT of Thailand further strengthen this decision, as this research is intended to contribute towards the development of sustainable and timely solutions for traffic challenges, especially in the field of traffic condition monitoring.

Additionally, the dataset provided by the iTIC is also being implemented on Longdo Traffic, a local traffic monitoring service. This utilisation provides a more relevant context for performance benchmarking efforts. Finally, it as an effort of expressing gratitude and contributing towards the society for accommodating to the growth process of author in pursuing its tertiary study. The rationale supporting the determination of research direction is summarised in Table 5.

Table 5. Summarisation of Rationale Behind Choosing Data Source

No	Rationale	Thailand	United States
1	Open taxi GPS dataset	iTIC	TLC
2	Region of Interest	Bangkok	New York
3	Vehicle Classification Standard	DLT	FHWA, and NYC DOT
4	Passenger Vehicle Population	61%	32%
5	Vehicle Population Publishing Interval	Monthly	Annually
6	Focus of research	Congested traffic condition	
7	Benchmark Accessibility	Longdo Traffic	Not Found

2.2 LONGDO TRAFFIC AS GROUND TRUTH DATA FOR BENCHMARKING

Longdo Traffic is a navigation assistance and traffic monitoring service provider in Thailand. This service is developed by Metamedia Technonology Corporation Limited, a software development service company that was established in September 2005 [30]. Longdo traffic is a joint effort projects from numerous parties with different expertise ranging from telecommunication to traffic authority bodies which includes Forth Corporation, Expressway Authority of Thailand (EXAT), Samart Group, Royal Thai Police (Traffic Police Division), the National Electronics and National Information Center (NECTEC) of Thailand, and the iTIC [31]. The the project was launched to provide a web-based map service for traffic condition monitoring with Bangkok as the initial monitoring point in 2009 [31]. Longdo Traffic uses few different traffic monitoring resources including traffic cameras (CCTV) for live footage, and GPS unit within car navigation system to provide its traffic monitoring and navigation services to users while collecting extra information from users' report on traffic anomaly events that has not been updated [31].

Longdo Traffic uses the taxi GPS data as one of its data sources for historical and real-time data since the project initiation. Few of the known taxi GPS probe data sources are contributed by the iTIC and Oriscom [32]. In 2017, Ministry of Transport of Thailand launched 2 new taxi applications to boost public confidence in using taxi service, namely Taxi OK and Taxi VIP, whereby Oriscom was the contractor for of the GPS tracker [33] [34]. Initially, 40,000 unit of taxis were enrolled into the initiatives while setting up total target of 80,000 unit (72%) of taxis within Bangkok area to join by the end of 2018 [33]. The initiative provides favourable amount of traffic representation within Bangkok via GPS probe data and opportunities to conduct traffic monitoring research.

Today, Longdo Traffic incorporates auxiliary data such as rain radar, and pollution sensors into its map, enriching the offered features and improving its overall services. On micro level, colour coding the traffic network is a common technique used in traffic monitoring to indicate traffic condition on lane level. Meanwhile for over all traffic condition, Longdo Traffic the traffic conditions analysis score in the scale of 0 to 10 through "Longdo Traffic Index" feature (score of 10 represent traffic is standstill). Addition of more features can be done on top of currently used features

(additional layer) to be incorporated onto the current map view for further analysis such as shown by Figure 11.

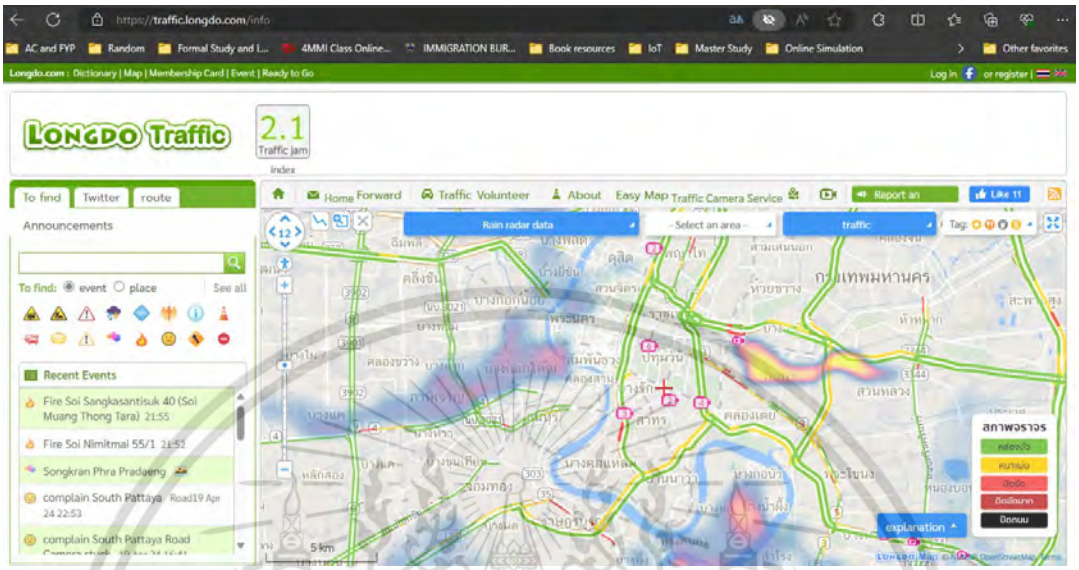


Figure 11. Longdo Traffic User Interface with Rain Heatmap [35]

Availability of comprehensive and inclusive datasets enable Longdo Traffic to perform meaningful analysis of traffic data and publish numerous findings such as traffic congestion pattern. Longdo Traffic mentioned that they have over 500,000 volunteers for live traffic monitoring through their mobile application, while typical active contributor is at about 31,000 – 53,000 users (approximately 6 - 10%, 42,000 on average) from the reported user number such as shown in Figure 12.

Figure 12. Active Volunteer on Moderate Traffic Index (Score ~ 5) [36]

Knowing traffic congestion pattern will allow public to plan their travel and allow authority to analyse the root cause of the congestion [37]. Longdo Traffic analysed the historical traffic data provided by the iTIC and identified the top 100 congested roads in Bangkok, presented on hourly basis [38]. The findings were published in their website, demonstrating the potential application of traffic monitoring data. Consequently, the identified top 100 congested road served as resource for benchmarking measurements and comparison in traffic congestion patterns recognition efforts within this research.

2.3 PRIOR RESEARCH ON TRAFFIC MONITORING EFFORTS AND ITS TECHNIQUES

Traffic monitoring research and its improvement efforts has been more important than ever to meet the challenges stemmed from urbanisation and the rapid development of vehicular transportation. Common challenge that is experienced by major cities worldwide is traffic congestion, that has been increasing in intensity and possess serious implications to the quality of their residence.

Traffic bottlenecks are a primary cause of congestion on urban roads and expressways, contributing to approximately 40% of traffic congestion incidents in the United States [39]. These bottlenecks are characterised by spatial discontinuities that reduce the capacity of road [40]. Identifying and addressing traffic bottlenecks is challenging due to various complicating factors such as unpredictable congestion patterns, topology of road network, and travel behaviour [40]. Successful identification and removal of these bottleneck can improve traffic congestion and bring network-wide improvements [39].

Previous research on traffic bottleneck identification has proposed various methodologies, including the construction of a causal tree based on graph theory [40]. Initially, a novel definition of traffic bottlenecks that considers both congestion level cost at the road segment itself and its contagion cost which denotes the effect of a congested segments when it propagates to another road segment. Additionally, an innovative technique combining, maximal spanning trees, Markov analysis, and graphical models has been proposed [39]. Further, simulation of their proposed method against “congested segment cost only” method on Taipei’s inductive loop-based traffic data was performed on SUMO software. Consequently, the simulation

outcomes showed that bottlenecks in the traffic network are not necessarily located at the most congested road.

Congestion propagation is another challenge and field of interest within the research of traffic monitoring. Determination of the impact given by a congested road (point) into the traffic network in term of magnitude and its after effect, often referred as causality relationship. This phenomenon to illustrate causality relationship with the well-known term of “butterfly effect” whereby flap of butterfly wing (congested road) could trigger a typhoon (network-wide congestion) elsewhere.

Congestion propagation, rather, traffic congestion is a spatio-temporal event which can be determined when adequate resources, mainly traffic data, is available. However, whenever data availability is an issue, researcher often concentrate their resources to collect and analyse data within the “peak hours”.

Peak hours is a term describing period of time in which traffic is at its worst condition throughout the day and night time, commonly caused by time dependant (periodic) event [41], such as starting and dismissal time of academic (school, and university), business (office, and market), social event (football match, weekly car free day), and religious (five prayers time for Muslim, Church service for Christian, and Incense offering and getting blessing for Buddhist), are among few to mention. It is logical to focus the limited resources towards the peak hours due to the abundance of vehicle volume, observation of traffic condition under stress, and formulation of critical traffic management strategy.

In the presence of traffic congestion, as per mentioned previously, there is probability of the other part within the traffic network to suffer the consequences from congestion in another place, especially when the road is connected or leading directly towards the congested road (area). However, discovering frequent pattern of congestion propagation in the traffic network could potentially enhance traffic management systems. Effective and timely traffic congestion prevention and clearing, and urban computing, are among few of the benefits of congestion propagation pattern identification to name [41].

Successful detection of traffic congestion within the network and causal interaction among them, challenges mentioned below need to be addressed [41]:

1. Heterogenous traffic patterns [41]: Each unique hours within the week, and each road segments often have distinct time-variant traffic patterns that are difficult to be represented by a single model of congestion detection across entire road network at different time periods.

2. Data sparseness and distribution skewness [41]: given the use of adequate quantity of sensors across a traffic network, the complexity of processing traffic data becoming unique and increasingly challenging as the traffic network expand due to different travel frequency within a particular road that may need to be weighted individually.

3. Causality among congestion [41]: given many congested roads (segments) could be determined, challenges in detecting the appearances, growths and/or transformations, and disappearances of congestion propagations by time.

Existence of the three main issues in the detection of traffic congestion propagation and its pattern is addressed by. Nguyen et al. [41] using Causal Congestion Tree (CCT), coupled with deployment of frequency discovery algorithm within the CCT that generates the most frequent sub-structures (subtree) as representative of recurrent propagation pattern in the data, and finally utilised Dynamic Bayesian Network for traffic congestion propagation modelling and causality probability estimation. Later, experiment was performed to observe the congestion propagation after period of 5 minutes with data with data obtained from road sensors installed across major Australian city. The dataset consists of 4 weeks' worth of data with 3:1 ratio on train-test data split. Consequently, the proposed method could achieve about 84% accuracy in predicting the congestion propagation, and claimed to be "general" enough for scalability and cross-field application (i.e., finding bottleneck in both internet traffic data, and water pipe data).

Causality (congestion correlation) among congested segments (road) was studied differently by Wang et al. [42] which related GPS data, and Point Of Interest (POI) around particular road network to determine the correlation pattern of congestion segments Initially, the proposed method started by extracting congestion information of each road segments, then mined the congestion relation between

each road segment pair through mining algorithm that was developed according to their congestion correlation definition, over dataset that is sourced from GPS trajectories of over 10,000 taxis, and road networks. Secondly, extraction of topological features and Point of Interest (POI) features from POI data and road networks. Then, dataset for training and testing are generated from the data obtained from the earlier process, to be fed into classifier. The proposed three-phased frameworks can produce stable classifier models that performed satisfactorily in term of performance assessments (precision and recall), and uncovering congestion patterns within road segments.

In the field of traffic congestion, much research come up with specific and unique definitions in defining traffic congestion components and its parameters. These components are then utilised within their traffic monitoring algorithms to determine the congestion condition of a traffic network. This phenomenon supports the notion that each unique area of study within any traffic networks possesses distinct behaviour, thus significant alteration, if not, completely new set up need to be developed for each area of study, impeding the scalability of the monitoring system and its adaptability towards evolving traffic components and their behaviour.

Based on limited literature review on prior research, it can be concluded that the most congested road segments may not be necessarily the bottleneck within the traffic network [39]. However, to achieve such level of analysis are not possible yet due multiple fundamental components of research are necessary, namely data from sensors (e.g. inductive loop, road network, POI, etc.) to determine the weightage/importance of a road in congestion correlation, possible need of permission from the authority, expertise in hardware engineering and transportation related knowledge, which are limited in possession within this research.

Acknowledging the gap of numerous resources within this research, especially in the knowledge of transportation field, the approaches designed within this research in addressing traffic congestion challenges are to minimise the “barrier to entry” and maximise possible contribution by utilising pattern recognition algorithm and minimal definition of traffic congestion components, while focusing on peak hour traffic condition.

2.4 PATTERN RECOGNITION

Pattern is observable characteristics of a subject of study (data) which is revealed from reliable frequent or widespread of incidents (event). Pattern can be seen either physically or it can be observed mathematically. Pattern recognition is a process of learning (i.e., data classification) using a machine learning algorithm to discover pattern (knowledge), which is illustrated by Figure 13.

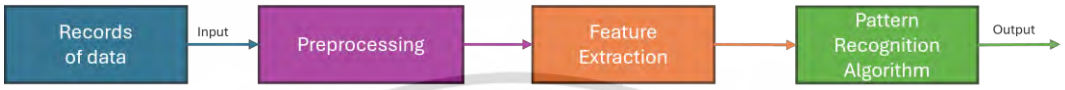


Figure 13. Pattern Recognition Process

The usage of pattern recognition becomes increasingly popular due to increase of data size (big data), and improvement in computing power. Pattern recognition empowers technologies such as speech recognition, and image recognition (e.g. automatic medical diagnosis), among others. Often, raw data is converted into a form that usable by machine during information extraction stage, so that regularity of patterns can be discovered by the machine learning algorithm on the successive stages.

Pattern recognition could be performed with at least 2 distinct methods, namely supervised learning, and unsupervised learning, among others. The difference among the mentioned methods lies on the nature of data used within the pattern recognition tasks. The dataset that has been processed by human (expert) in a way that the correct (known and/or desired) output for a particular input (event) within a specific (operational) condition has been obtained or predetermined, is known as labelled dataset. Data labelling requires the adding of label(s) of the context into the dataset that will help machine learning model in learning through process of context identification (characteristics, behaviours, classes) of the raw data by experts. Meanwhile, unlabelled dataset is piece of data that has not been tagged with label(s); information is yet to be discovered.

Supervised learning method uses labelled dataset. The algorithm used within supervised learning for pattern recognition task commonly termed as classification algorithms which learns the pattens within the given dataset with correct outputs to

try to generalise and predict the class of new examples (events) that it has not seen before. The outputs are made of discrete class labels such as spam or not spam, congested or free flow, etc. Availability of labelled inputs and outputs allow assessment of supervised learning model's performance and improvement on the learning process. Line classifiers, Support Vector Machines (SVM), decision tree, random forests, are among few to name on the example of supervised learning classification algorithm which is used to make predictions.

Unsupervised learning method takes in unlabelled data for inherent structure (knowledge) discovery without the intervention of human. The algorithm used within supervised learning for pattern recognition task commonly termed as clustering algorithms whereby similar characteristics or behaviours or experiences is being grouped together. Unsupervised learning models mainly used for data grouping, even though some may be modified for cluster prediction task too after fundamental clusters have been produced from an earlier learning task. However, often, there will be lack of transparency on how data is being clustered resulting in challenges in controlling accuracy of output [43]. Exclusive clustering, overlapping clustering, hierarchical clustering, and probabilistic clustering are among few of clustering types to name for unsupervised learning clustering algorithms category [44].

Given the limitation in expertise, unsupervised learning method through clustering algorithm is the more appropriate approach to be incorporated into the framework of congestion pattern recognition within this research because the use of unlabelled dataset of taxi GPS probe from the iTIC.

2.4.1 Unsupervised Clustering

Clustering algorithms groups unlabelled data into partitions (clusters) based on data differences or similarities (inherent structures), facilitating data exploration and analysis of patterns which helps in decision making. The data points within each identified clusters are more alike to each other, than the data points in the rest of the clusters.

Traffic congestion patterns identification involves the necessity of extracting congestion segments that is can also be seen as a hotspot because the congestion phenomenon often happens on that location, which gives rise to the term congestion hotspot. Application of unsupervised clustering algorithms for hotspot

mining application has been performed by numerous researchers in which clustering algorithms such as K-means, Density-based Spatial Clustering of Applications with Noise (DBSCAN), and Ordering Points To Identify Cluster Structure (OPTICS), are among few to name, had been proposed.

Refined K-Means++ algorithm was proposed by Wang et al. [45] to explore and master the actual supply and demands of unique hotspots at distinct time. They proposed a two-level subdivision concept and improved K-Means++ algorithm for the clustering of hotspot of taxi passengers. Grid unit was established for the area of interest, consisting of data grids from location and time data, whereby the suitable number of local regions for each data grid was determined comparing the result of by Aikake Information Criterion and Bayesian Information Criterion of the Gaussian Mixture Model of the data grid. Afterwards, the optimal number of k-cluster value was determined by the sum of the square distance errors for each local area [45]. K-Means++ clustering algorithm was then used to perform clustering within each area and determine the hotspot. The method was claimed to produce more accurate hotspot mining in comparison to direct clustering and DBSCAN clustering.

The approach of dividing region of interest into smaller local areas were performed to match the dynamic nature of taxi demand, resulting in different cluster numbers within each local area. Similar approach can be used within this research considering that spatio-temporal data of taxi GPS may also reveal the pattern of traffic congestion through hotspot mining technique. In term of clustering algorithm, density-based clustering algorithms are seen to be the more appropriate approach for congestion hotspot mining with taxi GPS probe to minimise assumption of normally distributed data, spherical cluster assumption, and eliminate necessity to predefine cluster quantity.

In density-based clustering, a cluster is developed by searching within the spaces of given input for high-density data regions that are separated by areas of low density, resulting in arbitrarily shaped clusters that are more robust against noise. The output clusters with dense data can be regarded as “hotspots”. These algorithms are claimed to be particularly suitable for nature-inspired spatial data [46], such as GPS coordinates in transportation [47].

The application of density-based clustering algorithm comes with limitations that have been discussed in paper authored by Wang et al. [45] in which flat cut is made based on similarity distance thresholds, among others. Alleviation method has been proposed through hierarchical method to choose suitable density threshold [46].

The most popular density-based approach for clustering is DBSCAN algorithm [48], [49]. Many algorithms are created as a modification or improvement to DBSCAN algorithm to overcome its apparent shortcomings that are dependent on two sensitive parameters, i.e., the minimum number of points in the neighbourhood (*MinPts*), and the radius of the neighbourhood (*eps*), such as VDBSCAN, OPTICS, among others [49]. HDBSCAN is another improved version of DBSCAN which capable of making variable density clustering, in which it matches the approach that was used by Wang et al. [45] in performing hotspot mining with additional benefits of arbitrary clusters shape.

2.4.1.1 HDBSCAN

Hierarchical Density Based Spatial Clustering Application of Application with Noise (HDBSCAN) is an extension to DBSCAN that construct a cluster hierarchical tree and extract flat cluster from the tree that meets the requirements of specific cluster stability measure [50][51], improving upon the flat labelling of clusters within dataset based on global density thresholds [50]. Taxi GPS dataset will have different density of data due to the mobile nature of the GPS unit that changes according to the trip, resulting in random data distribution, and sparsity of data that is dependent upon the quantity of operating taxis and popularity of specific location.

The demonstration of HDBSCAN capabilities in producing clusters of varying density has been proven in diverse field including air traffic management in [52], [53] and correlation of motion of protein in [54], among few to name, in which definition of distance matrix and its distance measurement techniques were varied to represent the data in the best way possible. Malzer et al. [51] specifically used HBDSCAN to cluster the remainder data that were declared as noise by DBSCAN clustering for GPS data points of ride pooling dataset. The approach could outperform HDBSCAN with excess of mass configuration as its cluster stability measure. However, the need of

domain expertise in deciding the suitable value of initial threshold remained as the main challenge, i.e., difficult to be determined [51].

2.4.1.2 Summary for Unsupervised Clustering

Based on limited literature review on prior research, it can be concluded that the data availability on each region may vary according to its popularity which had been seen from the methods used within refining of local area in [45], which support the notion of varying density in taxi GPS data. Therefore, the algorithm that is capable of variable density clustering, HDBSCAN is desirable to perform hotspot mining for congestion length pattern.

The input distance matrix for HDBSCAN clustering had been observed to be of varying techniques such as Euclidean distance[52], weighted Euclidean distance [53], and motion correlation matrix [54], in which the notion of developing more relevant distance measurement is encouraged to better describe the dataset in use. The insight obtained from Malzer et al. in [51] regarding usage of "excess of mass (eom)" as cluster stability measure in HDBSCAN clustering will be avoided since HDBSCAN(eom) had been shown to break up high density region which would have intuitively clustered into one or fewer clusters, in comparison to the micro-clusters produced [51]. Therefore, usage of leaf cluster stability measure will be used in HBSCAN clustering because it is intuitively easier to be controlled and understood.

2.5 DISTANCE MEASUREMENT

Taxi GPS coordinate consists of longitudes and latitude pairs in which the distance between two points will be derived from and fed into the distance matrix for HDBSCAN clustering. There are 2 model of earth which are flat (planar) and non-planar (ellipsoidal or spherical) model. The usage of distance calculation technique that mismatch the nature of data will result in measurement error that sometimes is not appropriate to be tolerated.

Planar distance techniques calculate distances in two dimensional (2D) Cartesian coordinate system, e.g. Euclidean distance, Manhattan distance, etc. Meanwhile, geodesic distance techniques calculate the distance across the curved surface of the world in a three dimensional (3D) spherical space [55]. The fact the earth is a 3D spherical object that is slightly flattened, i.e., ellipsoid, the usage of

planar calculation may be suitable only on certain case such as distance between equatorial cities, among others, in which the error is negligible for the distance between Singapore city and Nairobi [55]. Despite the scale of work which covers Bangkok Metropolitan Region (BMR) only, minimisation of error still being performed by applying non-planar distance measurement techniques i.e., Haversine formulas and Vincenty's formulae.

Haversine formula calculates distance between 2 points on a 3D spherical space in which the error introduced could be as high as 0.3% [56]. The distance between two points (pairwise distance) calculation of the Haversine formula given in (1).

$$D(x,y) = 2\arcsin\left[\sqrt{\sin^2\left(\frac{(x_{lat}-y_{lat})}{2}\right) + \cos(x_{lat})\cos(y_{lat})\sin^2\left(\frac{(x_{lon}-y_{lon})}{2}\right)}\right]. \quad (1)$$

Greater accuracy can be achieved with Vincenty's formulae i.e., up to 0.5 mm of error between two points [57]. However, Vincenty's formulae require more calculations compared to the simpler Haversine formula. Therefore, the usage of both will be utilised in a balance manner i.e the calculation of distance matrix for HDBSCAN will be performed with Haversine formula, while other more sensitive calculation may be performed with Vincenty's formulae.

2.6 SIMILARITY MEASUREMENT

Similarity measurements are techniques used in the process of quantifying similarity or dissimilarity between two objects, distribution, or data points. There are numerous techniques to perform similarity measurement, however it can be categorised into parametric and non-parametric approaches. While the nature of distribution of the taxi GPS dataset from the iTIC can be assumed to be non-parametric due to the 20 minutes sample size and the intuition that was developed from the insight obtained through limited literature review, the usage of parametric techniques will still be used to facilitate the explanation of similarities between the traffic congestion length distributions.

Parametric technique is used for dataset that is assumed to be normally distributed. Despite the contradicting nature of the dataset, characteristic, techniques such as measure of central tendency (mean) and standard deviation will help in describing the pattern of the congestion length distribution.

Suroso et al. [58] employed mean and standard deviation (SD) analysis to assess the similarity of Received Signal Strength Indicator (RSSI) values between Generative Adversarial Network (GAN) synthesised and an experimental dataset, focusing on the comparison of data distribution similarities.

Jensen-Shannon Divergence (JSD) is employed to quantify the similarity between the ground truth and synthetic data, as it was previously used by Xie et al. [59] for comparing biased and adjusted geodemographic distributions. JSD measurement is symmetrical, which means that no matter which equation is used as the ground truth between the two equations to be compared, the measurement result will be the same. JSD is derived from its non-symmetrical variant, the Kullback-Leibler (KL) divergence, which is given in (2).

$$\mathrm{KL}(\mathrm{P}||\mathrm{Q}) = \sum_{\mathrm{x} \in \mathrm{X}} \mathrm{P}(\mathrm{x}) \log_2 \left(\frac{\mathrm{P}(\mathrm{x})}{\mathrm{Q}(\mathrm{x})} \right) \tag{2}$$

. The average between the compared distribution is denoted by *M*, whereby $\mathrm{M} = (\mathrm{P} + \mathrm{Q})/2$. Subsequently, the JSD measurement between P and Q is calculated by (3).

$$\mathrm{JSD}(\mathrm{P}||\mathrm{Q}) = \frac{1}{2} \left(\sum_{\mathrm{x} \in \mathrm{X}} \mathrm{P}(\mathrm{x}) \log_2 \left(\frac{\mathrm{P}(\mathrm{x})}{\mathrm{M}(\mathrm{x})} \right) + \sum_{\mathrm{x} \in \mathrm{X}} \mathrm{P}(\mathrm{x}) \log_2 \left(\frac{\mathrm{Q}(\mathrm{x})}{\mathrm{M}(\mathrm{x})} \right) \right) \tag{3}$$

Earth Mover's Distance (EMD) is employed as a similarity measurement tool due to its prior successful application in quantifying movement model similarities for animals by Potts et al. [60]. The simple variant of EMD that utilised one-dimensional data is chosen. Let us assign *Q* as the representative for synthetic data distribution with data size of *m*, $\mathrm{Q} = \{(\mathrm{q_j}, \mathrm{v_j})_{\mathrm{j}=1}^{\mathrm{m}}\}$. Subsequently, *P* is used to as denote the distribution of the ground truth data with size *n*, $\mathrm{P} = \{(\mathrm{p_i}, \mathrm{u_i})_{\mathrm{i}=1}^{\mathrm{n}}\}$. The EMD

measurement between *P* and *Q* can be calculated using (4) by formulating them into a transportation problem such as shown in [61].

$$\mathrm{EMD}(\mathrm{P},\mathrm{Q}) = \min \sum_{\mathrm{F}=\{\mathrm{f_{ij}}\}} \frac{\sum_{\mathrm{i,j}} \mathrm{f_{ij}} \mathrm{d_{ij}}}{\sum_{\mathrm{i,j}} \mathrm{f_{ij}}} \tag{4}$$

In the transportation problem, elements of distribution *P* can be considered as "supplies" at location u_i while the elements of distribution *Q* can be taken as "demands" at location v_j, with the amount (weight) of supply, p_i and demand, q_j, . Consequently, EMD is defined as the minimum (normalised) work required for transportation of "supply" to fulfil the "demand", whereby d_{ij} denotes the distance between locations of supply, u_i and location of demand, v_j , while f_{ij} denotes the mass of transported goods from i^{th} supply in ground truth distribution to the j^{th}

Adapting the 1D calculation of EMD to our congestion length, *Q* is our synthesised data for a specific congestion length extraction approach, and *P* is our ground truth data. The EMD is calculating the minimum work required to resolve the "supply-demand" issue by using the distance between u_i and v_j with the congestion length of p_i and q_j.

Similarity measurement of non-parametric techniques i.e., EMD and JSD are seen to be suitable techniques to be utilised as the main indicator of performance measurement of the taxi GPS dataset. Meanwhile, application of parametric techniques i.e., mean and SD measurements are used to complement the analysis of the data distribution.

CHAPTER 3

METHODOLOGY

Unprocessed GPS data does not provide any useful information; therefore, it is organised by using clustering techniques. The knowledge on making a suitable definition of traffic congestion is the main challenge to the research within this field, therefore unsupervised clustering techniques was used to reduce bias (i.e., from human intervention) and reflect the dynamic of traffic behaviours. The definition of traffic congestion can be later refined into the desired definition of congestion by tuning the threshold of traffic congestion parameter. Analysis can be performed on the resulting congestion points by exploring patterns and causality observations between congestion points.

The primary objective of this study is to determine the pattern of congestion propagation through traffic congestion length. The research started by clustering taxi GPS data into clusters with congested traffic condition. Subsequently, distance measurement is made between two congestion hotspots within a congested area. Further, congestion propagation phenomenon is used as validation criteria for the obtained distance. Moreover, three distinct congestion length extraction approaches are proposed to process the validated distance into congestion length. Finally, similarity measurement is applied to gauge the performance of the proposed framework in generating congestion pattern against the ground truth data.

The proposed framework can be divided into three stages namely: input loading, data preprocessing, and finally, clustering and postprocessing. The overall approach to obtain the desired output is illustrated in Figure 14, and all the processes are executed with Google Colaboratory in a Python environment by utilising numerous libraries including Numpy [62], Pandas [63], are among few to name.

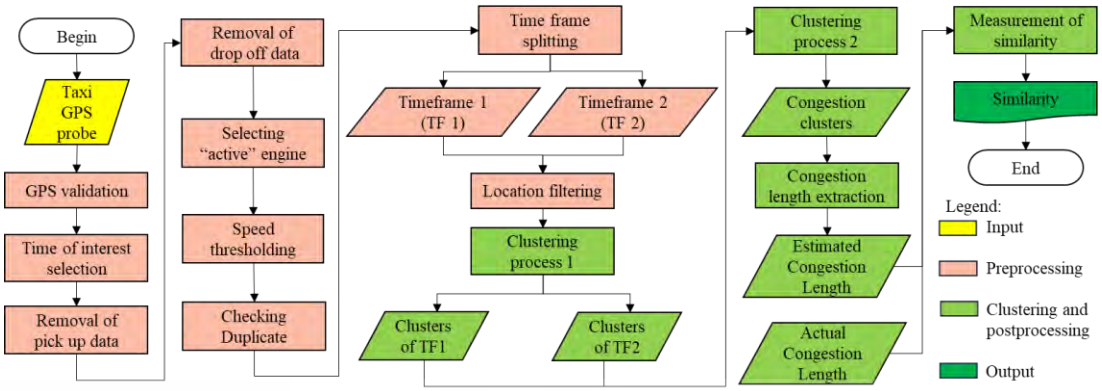


Figure 14. Overall Flowchart of Congestion Length Estimation with HDBSCAN from Taxi GPS Probe Data

The experiment began with loading taxi GPS transactions data as input. The input data was open data, licensed under Creative Commons Attribution 4.0, provided by Intelligent Traffic Information Center Foundation, iTIC Thailand [13]. The Taxi GPS dataset made of nine attributes in which these attributes will be selected according to the desired value given in Table 6.

Table 6. Taxi GPS Probe File Description and Its Desired Values

Field	Explanation	Data Type	Target Value
`VehicleID`	Uniquely assigned ID of the vehicle	str	All ID
`gpsvalid`	Indicator for good GPS entry	int	1
`lat`	Vehicle's latitude coordinate	float	Within Bangkok
`lon`	Vehicle's longitude coordinate		
`timestamp`	Date and time of records (GMT+7)	str	Within time of interest
`speed`	Travelling speed recorded in km/h	int	< 25 km/h
`heading`	Heading direction of vehicle (0-360) ° from North=0	int	-
`for_hire_light`	1 = light on - possibly no passenger, 0 = light off - possibly carrying passengers	int	0
`engine_acc`	1 (active, data collection every minute), 0 (inactive, data collection every 3 minutes)	int	1

The loaded input data was visualised into a map using Folium library [64], in which the resulting plot data indicated that the contributing taxis were operating all over Thailand, with largest density in Bangkok Metropolitan Region, such as illustrated in Figure 15.

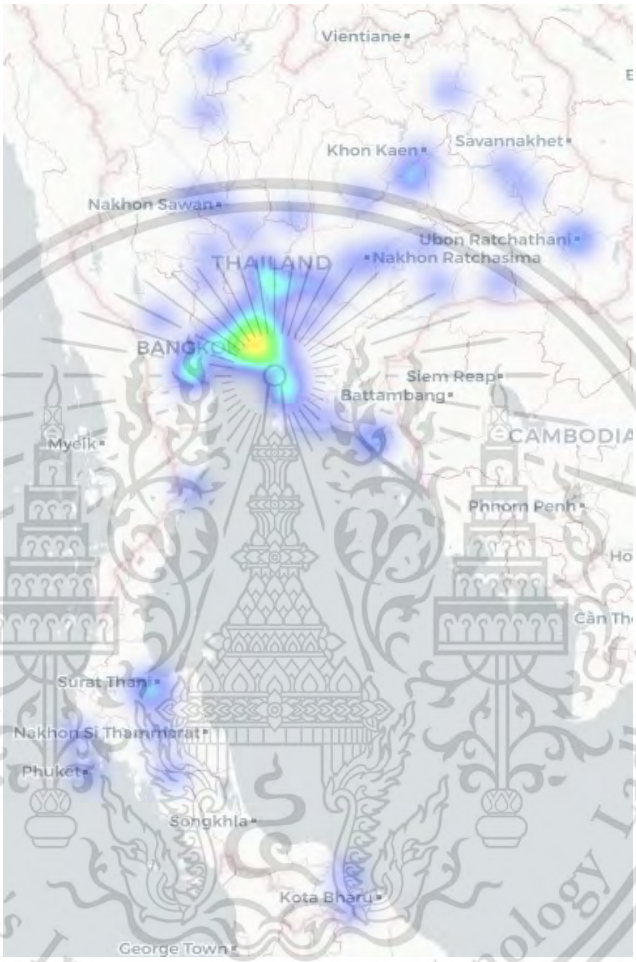


Figure 15. Density Distribution of Taxi GPS Data

3.1 GROUND TRUTH FOR TRAFFIC CONDITION DEFINITION

The presence of multiple attributes, varying data range, and possibly errors within input data require the establishment of set of rules and conditions to define the data validity, and to determine traffic congestion condition. Ground truth was required as a guide when determining traffic congestion condition. To the best of our knowledge within the limited literature review, the definition of congested flow speed is yet to be set, due to distinct characteristic of traffic network components for each area of interest.

3.1.1 Traffic Flow Speed

Our reference for traffic congestion flow speed in Thailand was taken from Longdo Traffic. It is a service provided by Metamedia Technonology Corporation Limited which has been publicly available since 2009 [31]. In Longdo Traffic, the traffic flow speed in urban road within Thailand was classified into four distinct colours. There are two different traffic flow speed classification based on the types of transportation infrastructure in Longdo Traffic which is given in Table 7 [37].

Table 7. Longdo Traffic's Colour Code Representation of Average Speed

Colour code	Road Type and Average Speed (km/h)	
	Main roads	Expressway
Black (Standstill)	< 10	< 20
Red (Heavily Congested)	< 15	< 30
Yellow (Slightly Congested)	< 25	< 60
Green (Free-Flow)	≥ 25	≥ 60

Here on, we assume that all data obtained from the iTIC on taxi GPS probe was produced by taxi operating on main road. Then, we proposed our own definition for the traffic condition categories based on the colour-to-average speed codes defined within Longdo Traffic for the main road infrastructure. Our definition only focusses on separation between congested traffic and free flowing traffic. The proposed definition is presented within Table 8.

Table 8. Longdo traffic's Speed Code

Average speed (km/h)	Colour code	Our definition
< 10	Black	Congested
< 15	Red	Congested
< 25	Yellow	Congested

3.1.2 Traffic Congestion Length

Data for Ground Truth (GT) was extracted manually from Longdo Traffic's Top 100 most congested roads in Bangkok in [38]. The ground truth data serves as reference and used to gauge the similarities of congestion pattern generated by HDBSCAN algorithm. The user interface of Longdo Traffic's Top 100 most congested roads in Bangkok is given by Figure 16.

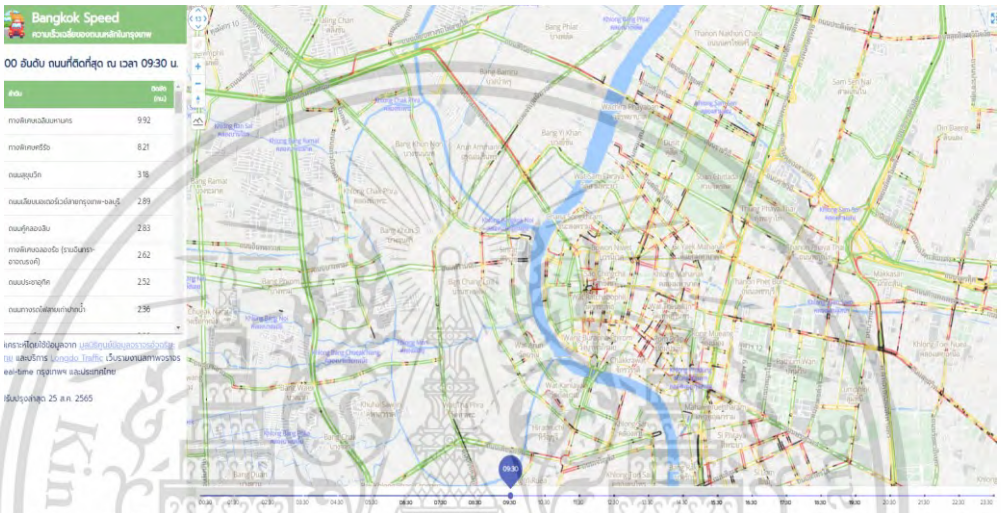


Figure 16. Web Page of Longdo Traffic's Top 100 Congested Road in Bangkok

The page hosts information of the typical traffic network outlook of Bangkok within a day represented with hourly granularity The website mentioned that the page was last updated on 25 August 2022 [38], in which we assumed that the last historical data used for the analysis of the published result is up to 24 August 2022. The data on most congested roads is given on the list located on the top left side of the window.

There ground truth data is extracted for the peak hours only which can be split into 2 sessions: morning peak and evening peak hours. In Bangkok, the morning peak runs from 6:00 to 9:00, whereas the evening peak hours runs from 16:30 to 19:30 [2]. The collection of ground truth data for different hours can be performed by adjusting the slider on lower right section of the webpage. All collected data of congestion length is measured in the unit of kilometers (km). The eight unique

datasets were obtained for the ground truth in which the details are summarised in Table 9.

Table 9. Summary of the Extracted Ground Truth

Ground truth data	Hour	Peak Period
G_Dataset 1	6:30	Morning
G_Dataset 2	7:30	Morning
G_Dataset 3	8:30	Morning
G_Dataset 4	9:30	Morning
G_Dataset 5	16:30	Evening
G_Dataset 6	17:30	Evening
G_Dataset 7	18:30	Evening
G_Dataset 8	19:30	Evening

3.2 DATA PREPROCESSING

Preprocessing stage was designed to refine the input dataset before its application in subsequent clustering procedures, aimed at eliminating extraneous data (noise) to enhance clustering efficacy and improve the accuracy of traffic congestion identification. The primary focus of preprocessing the taxi GPS data involved mitigating erroneous congestion indications. Notably, events like passenger pickup and drop-off could generate probe entries with speeds lower than the free-flow threshold.

Given the assumption made on section 3.1.2, whereby 25 August 2022 was the last used data by Longdo Traffic to produce our ground truth, we refer to another webpage within Longdo Traffic which showcases the day with most congested traffic of the year [65]. According to the website, November 2nd, 2021 (Tuesday) was identified as the weekday with the worst congestion in Bangkok throughout the year 2021. Thus, we applied our data preprocessing to the 2nd of November 2021 data obtained from November 2021 data of the iTIC [66].

The data of the iTIC taxi GPS for the date of 2nd of November 2021 was loaded into Pandas dataframe as input. Filtering is applied to the input to remove invalid entries that are not reflecting congested traffic condition. Summary of filtering method in preprocessing stages and their purposes are given in Table 10. The following subsections describe the filtering criteria to distinguish between valid and invalid data.

Table 10. Summary of Filters in Processing Stage

Sequence	Filtering Process	Description
1	GPS Validation	Only selects “gpsvalid” = 1
2	Time of interest selection	Select data of 20 minutes interval on each congested hours
3	Pickup event	Remove “congested” data resulted from passenger pickup
4	Drop-off event	Remove “congested” data resulted from passenger drop-off
5	Active engine	Operating and hired taxi only
6	Speed thresholding	Removal of “non-congested” entries
7	Duplicate data	Check for any duplicating data due to sensor error
8	Time frame splitting	Split data into 10 minutes interval, TF1, and TF2
9	Location filtering	Removal of non-Bangkok data entries

3.2.1 GPS Filtering

GPS signal is the earliest field to be filtered since it has validity flag which indicates the reliability of GPS coordinate produced by the GPS unit. Poor GPS signals are recorded with the value of “0” in the “`gpsvalid`” field within the dataframe.

3.2.2 Time of Interest

The loaded dataset consists of GPS data spanning for 24 hours. Therefore, it is necessary to segregate the period to be studied to match each peak time within the collected ground truth data. The smallest granularity of traffic observation is five minutes within Google Maps. However, changes in traffic conditions within five

minutes are often not observable, especially during peak hours due to standstill traffic. Thus, a period of ten minutes was chosen as the time of interest. Since traffic congestion may worsen or diminish from one time segment to the next, the time of interest accommodates two different time segments, totalling twenty minutes. The summary of time of interest for each dataset is given in Table 11.

Table 11. Summary for Time of Interest for All Peak Hours

The iTIC dataset Identifier	Peak time in GT	Time of interest
Dataset 1	6:30	6:10:00 to 6:29:59
Dataset 2	7:30	7:10:00 to 7:29:59
Dataset 3	8:30	8:10:00 to 8:29:59
Dataset 4	9:30	9:10:00 to 9:29:59
Dataset 5	16:30	16:10:00 to 16:29:59
Dataset 6	17:30	17:10:00 to 17:29:59
Dataset 7	18:30	18:10:00 to 18:29:59
Dataset 8	19:30	19:10:00 to 19:29:59

3.2.3 Passenger Pickup and Drop-off Events

Stoppage or stationary events were common event for taxi operators during passenger pickup and drop-off, resulting in entries with low speeds within these periods. The removal of these data was recommended because it could lead to erroneous congestion indications [7]. However, the data during this event could be used as valid data by making some assumptions.

Passenger pick-up events is a situation where a hired taxi obtained a target destination from passenger, resulting in the activation of taximeter to officially start the passenger's trip. The changes from “1” to “0” in the “`for_hire_light`” field signifies the passenger pick-up event.

In practice, the taximeter typically activates upon the passenger's seating in the taxi, even if the driver is unaware of the destination. Consequently, taxi drivers

tend to remain stationary until the destination is known. However, instances were observed where pickup events in the dataset exhibited non-zero values for the "speed" attribute, suggesting a delay within the GPS unit, as non-zero speeds imply that the taxi began moving immediately upon the taximeter's activation (changes from “1” to “0”) such as shown in Table 12.

Table 12. Snippet of Pickup Event

Entry	VehicleID	for_hire_light	speed
1	Taxi1	1	30
2	Taxi1	1	32
3	Taxi1	0	23
4	Taxi1	0	32

For analytical purposes, pickup events were assumed to have a duration of one minute, correlating to one GPS probe entry per “active” taxi. It was assumed that the GPS probe commenced recording immediately during the passenger pickup event. Data filtering commenced by grouping entries based on “VehicleID” and monitoring only those with both “0” and “1” values within the “for_hire_light” field.

In cases where the “speed” value dropped below the free-flow threshold upon the moment taxi get hired (“1” to “0”), either one of the specified measures below was taken to identify and eliminate erroneous congestion data entries:

1. Pickup event recorded on last entry such as shown in Table 13: If the “0” (pickup) was recorded at the end of the probe entry for a particular “VehicleID”, the “speed” was assessed from the “speed” before it is hired, assuming similar traffic conditions. Shall the speed in previous entries is greater than free flow, while the last entry speed showed “congested”, then the last entry speed is replaced with previous entry before it, resulting in 32 km/h as real speed for the case in Table 13, or,

2. Non-last entry pickup event, such as shown in Table 14: If the "`for_hire_light`" field recorded another "0" entry a after pickup event, then the "`speed`" was assessed from the "`speed`" of the succeeding entry, with assumption that traffic conditions towards the destination reflects traffic conditions better. Therefore, in the case shown in Table 14, the real speed at the second (2^{nd}) entry data is 27 km/h (not congested), which is against the earlier recorded speed of 5 km/h that reported a congested traffic condition, since the speed < 25 km/h (indicates congestion) while being hired.

Table 13. Snippet of Last Entry Pickup Event

Entry	VehicleID	for_hire_light	Speed	Validity	New Speed
1	Taxi1	1	30	-	30
2	Taxi1	1	32	-	32
3 (Last Entry)	Taxi1	0	23	Invalid	32

Table 14. Snippet of Non-Last Entry Pickup Event

Entry	VehicleID	for_hire_light	Speed	Validity	New Speed
1	Taxi1	1	30	-	30
2	Taxi1	0	5	Invalid	27
3 (Last Entry)	Taxi1	0	27	-	27

A drop-off event was presumed to happen for less than one-minute, in which taxi speed will decrease and possibly come to a standstill due to arrival at the destination and/or payment activity. This event could potentially generate erroneous congestion indications. Sample of drop-off event is illustrated in Table 15.

Table 15. Snippet of Drop-off Event

Entry	VehicleID	for_hire_light	Speed	Validity
1	Taxi1	0	30	Valid
2	Taxi1	1	32	-
3	Taxi1	0	23	Invalid
4	Taxi1	1	15	-

Drop-off entries were examined carefully and eliminated only if the taxi's "`speed`" value fell below the threshold of free-flow speed during the drop-off. The transition in the in the "`for_hire_light`" field from "0" to "1" indicated a drop-off occurrence. Therefore, for the first (1st) entry in Table 15, the data is valid while for the third (3rd) entry, the data is invalid and removed.

3.2.4 Travelling Speed, Engine Condition, and Vacancy Status

Analysis of pickup and drop-off events requires the usage of "`speed`", "`for_hire_light`", and "`engine`_acc" fields to determine valid entries. Afterwards, filtering for those three fields were performed simultaneously with condition such as below:

1. Speed filtering: to keep the data entries that were indicating congested traffic conditions in the traffic networks, i.e., "`speed`" below 25 km/h.
2. Engine condition filtering: keep entries with active engine, i.e., "`engine`_acc" = 1.
3. Vacancy status: behaviour of drivers varies greatly and unpredictable when there is no passenger within the vehicle. Driver might drive slower than the possible speed due to no urgency which will introduce false congestion condition. Only entries that possibly carrying passenger are kept, i.e., "`for_hire_light`" = 0.

3.2.5 Duplicate Data Entry

Duplicating data entry was checked to eliminate the possibility of entries being recorded multiple times by the GPS probe unit. The entries were grouped by "`lat`", "`lon`", "`timestamp`", and "`vehicleID`" fields. Following this, the uniqueness of

the remaining entries (so far) was confirmed by tallying the collected entries against the dimensions of the analysed data frame.

3.2.6 Timeframe Splitting

The time of interest in each peak hours' time was separated into two unique time frames, each spanning 10-minutes intervals. The split resulted in a "prior" time frame, TF1 (e.g. 16:10:00 to 16:19:59), and the "later" time frame, TF2 (e.g. 16:20:00 to 16:29:59). Clustering will be performed on each time frame, and supplementary techniques will be employed to delve deeper into their relationship.

3.2.7 Region of Interest

Preprocessing was finalised by verifying the location of GPS entry in which some entries had been seen to be recorded outside BMR by using mapping visualisation earlier. The desired area within the study, i.e. Bangkok Metropolitan Region, was set according to the boundaries outlined in Table 16.

Table 16. Boundaries of Study Area

Location	Coordinate
Top left boundary	14.03116, 10.23174
Top right boundary	14.03116, 100.91143
Lower left boundary	13.545157, 100.23174
Lower right boundary	13.545157, 100.91143

3.3 HDBSCAN FOR IDENTIFICATION OF CONGESTION

The HDBSCAN algorithm is deployed in Google Colaboratory, using the hdbscan library provided in [67]. Initially, a distance matrix was constructed to serve as an input to the clustering algorithm.

3.3.1 Distance Matrix and Distance Measurements

A distance matrix was generated for HDBSCAN clustering by computing the distance between every pair of GPS data points (pairwise distance). Given the Earth's nearly spherical shape, the haversine formula offers a reliable approximation of the

distance between two points on its surface, with an average error of less than 1% [56][68].

The Haversine formula needs the coordinates to be converted into unit of radians, hence the Numpy library was used to perform the conversion between coordinate in the unit of degree into radians, i.e., `numpy.radians()` function. The resulting measurement are fed into the Harversine formula implemented in Scikit-learn's library, i.e., `sklearn.metrics.pairwise.haversine_distances()` function [68].

3.3.2 HDBSCAN

The implementation of HDBSCAN in Python environment was contributed by McInnes et al. [69] which is accessible through `hdbscan.HDBSCAN()`function. This implementation of HDBSCAN necessitated a fundamental parameter: "`min_cluster_size`" to determine the identification of noise points [70], i.e., any cluster with data points lesser than the specified value of "`min_cluster_size`". Further, we included an optional argument "`metric=precomputed`" to inform HDBSCAN of the provided input's state, i.e., preprocessed distance matrix computed using Haversine formula, thereby eliminating the need for additional similarity calculations.

3.3.2.1 Congestion Hotspots Clustering

The optional argument "`cluster_selection_method=leaf`" was utilised to achieve more fine-grained and homogenous clusters [71]. Going forward, the term HDBSCAN(leaf) and HDBSCAN will be used interchangeably to refer to the same clustering method, with "`leaf`" as the argument for the parameter of "`cluster_selection_method` " within the HDBSCAN algorithm.

HDBSCAN clustering with "`leaf`" is preferred over "`eom`" (Excess Of Mass) due to the nature of the algorithm in which "`leaf`" clustering method will collect the leaf nodes of each subtrees as long as it meets the minimum datapoints required to be categorised as a cluster. Meanwhile "`eom`" will take the most stable/persistent clusters which is gauged based on the cluster "lifetime". Distance between datapoints is used as part of the persistency measurement, from which a Minimum Spanning Tree (MST) will be constructed with distance as the edges. As the threshold between datapoints is reduced, a cluster will either break down into few smaller clusters,

known as child clusters, or disappear if it no longer contains the minimum amount of datapoints due to reduction in distance threshold value, resulting in more compact clusters with arbitrary shapes. The summary of difference between behaviours of clustering methods of HDBSCAN(eom) and HDBSCAN(leaf) is illustrated in Figure 17.

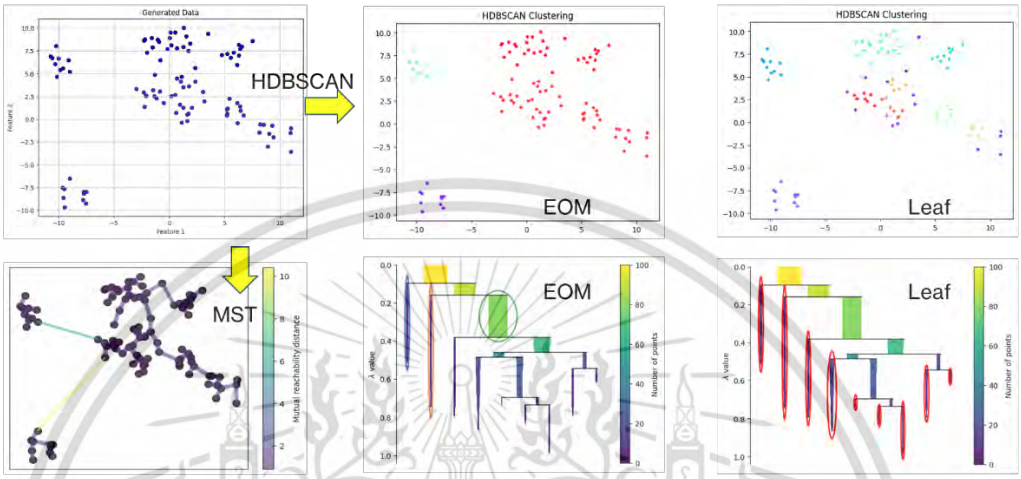


Figure 17. Comparison of Clustering Result between 'eom" and "leaf"

HDBSCAN(eom) produced 3 clusters while HDBSCAN(leaf) produced 9 clusters in Figure 17 because the clusters are more persistent at quantity of 3, which is not desired in traffic congestion. Meanwhile HDBSCAN(leaf) took all leaf clusters at subtrees as long as they have enough points to be categorised as a cluster.

The initial clustering procedure commenced with "`min_cluster_size=3`" to minimise the probability of fake congestion condition in case there are 2 taxis within the same location, e.g. traffic light. Any identified similarities during the clustering process generated clusters. This initial clustering used data from both TF1 and TF2, yielding congestion hotspots within both timeframes, subsequently referred to as "congestion hotspots".

A congestion hotspot cluster emerged when at least three data points were identified to belong to the same similarity group. The existence of numerous data points within a cluster will not be beneficial for the next clustering process because redundant data will be fed into the algorithm. Therefore, a representative point was developed to serve as a fresh datapoints for the clustering of congestion area.

Representative points were produced based on either one of the following logics:

1. Mode: the most frequently appeared data points location within a specific congestion hotspot will become the representative point. This location reflected the traffic the most since the number of data entry recorded was highest, or,
2. Mean: calculate the average of all data points within a particular congestion hotspot for each longitude and latitude recorded because this logic reflected the traffic condition the most.

3.3.2.2 Congestion Areas Clustering

The congestion hotspots from both outputs are utilised as input to the congestion area clustering process, another distance matrix (using Haversine formula) was developed. Congestion area clustering was started with "`min_cluster_size=2`" because precautionary step has been applied to congestion hotspots clustering process to reduce fake congestion condition, resulted in clusters labelled as a "congestion area" from here on.

3.4 CLUSTERING ALGORITHM EFFECTIVENESS

Effectiveness of HDBSCAN algorithm in the identification of congestion hotspot is performed by comparing it against K-Means++ algorithm which has been used in multiple hotspot mining research. There are about 505 intersections in Bangkok [72], which will be used as the cluster number, k, in the K-Means++ algorithm for Congestion Hotspot and Congested Area clustering. The implementation of K-Means++ algorithm is performed by using Scikit-learn's library, i.e., `sklearn.cluster.KMeans()` function [73]. The definition of valid clusters for K-Means++ is whenever a cluster consist of at least 3 datapoints, identical to the clustering argument given to HDBSCAN implementation on Congestion Hotspot clustering process.

The clustering will use one timeframe from one of the datasets and the performance will be measure by utilising Scikit-learn's library, i.e., `sklearn.metrics.silhouette_score()` function for Silhouette Scores [74], and `sklearn.metrics.davies_bouldin_score()` function for Davies Bouldin

Scores [75], whereby both of them have been widely used to gauge the performance of clustering effectiveness.

3.5 CONGESTION LENGTH EXTRACTION

Congested area comprised of congestion hotspots (data points) clustered by their similarity in distance. Congestion length is the outcome of pairwise distance calculation between all prior congestion hotspots (PH) and later congestion hotspots (LH) within a congested area. Initially, congestion hotspots within a specific congested area were sorted chronologically by timestamp, then they are grouped by the cluster of hotspots they were originated from. Subsequently, the pairwise distance (PH, LH) was calculated, yielding congestion length.

The pairwise distance calculation was performed using the ellipsoidal distance formulated by Thaddeus Vincenty which is accurate enough to produced error at most 0.5 mm [57]. The computational techniques for geodesic (ellipsoidal) distance was developed by Karney [76], and adopted into GeoPy library, in which was accessed through `geodesic().meters` function for resulting distance in the measurement unit of meters [77].

The resulting geodesic distance occasionally may produce excessively large distance between the data points, in which the distance value was still valid when it is compared against the maximum congestion length within the ground truth data, however turned into highly impractical, i.e., not when it is visualised due to the condition of the infrastructure presented within Bangkok, i.e., bridge. Therefore, a filter to validate the outcomes of geodesic distance was set by calculating the maximum distance that the vehicle could travel between the 2 data points given the average speed of 25 km/h and the difference in time from the timestamp.

The rationale behind the geodesic distance filter is derived from the condition of the data, which is GPS based with limited contributors, in which the entries within the dataset did not sufficiently describe the congestion condition in every road segment to support the notion of excessively large congestion area. Besides, the filter also kept the geodesic distances bounded, and focused around the epicentrum of the congested area.

On the contrary, there were clusters recoding data points of same coordinate with different timeframe, probably due to standstill traffic. Such condition resulted in

zero (0) value of distance measured during the geodesic distance calculation. Value of zero (0) distance is not recorded because it does not provide any congestion length value which will help in determining congestion length pattern. However, the congested area (cluster) remains valid shall it produced any valid distances. Each congestion areas might produce more than one pair of valid geodesic distances.

Finally, the valid entries of congestion length (geodesic distance) for each cluster are extracted into one single data point. There are three extraction techniques that are utilised to form its own congestion length pattern, which are detailed below and illustrated in Figure 18:

1. Extraction based on maximum distance (H_Max) approach extracts the largest distance value among valid congestion length within a cluster as representative of that cluster, then stores it as an element of H_Max distribution.
2. Extraction based on minimum distance (H_Min) approach extracts the lowest distance value among valid congestion length within a cluster as representative of that cluster, then stores it as an element of H_Min distribution.
3. Extraction based on average distance (H_Mean) approach extracts the average distance value among valid congestion length within a valid cluster as representative of that cluster, then stores it as an element of H_Mean distribution. The representative coordinates are produced by averaging the mean of all longitudes and cumulative longitudes both PH and LH, resulting in average prior hotspot, PH_{avg} , and average later hotspot, LH_{avg}. Then, the geodesic distance between the resulting pair (PH_{avg}, LH_{avg}) was calculated.

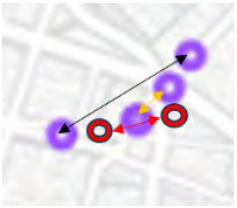


Figure 18. Congestion Length Extraction: Mean (Red), Min (Yellow), Max (Black)

3.6 SIMILARITY MEASUREMENTS

Similarity measurement techniques are employed to assess the success of our approach. Before application of similarity measurements techniques, the congestion length was scaled to ensure fair comparison between all pairs of congestion length extraction approaches against the ground truth.

3.6.1 Trimming Dataset for Comparison

Trimming was performed for other datasets that experienced inequality of data entries. The number of congestion area cluster produced may vary according to the amount of data entries and the distribution of the data itself within a particular dataset. Therefore, the amount of valid congestion clusters within each approach may exceed or fail to meet the number of entries recorded by the ground truth, i.e., one hundred entries. Excessive entries of valid clusters will be set to top one hundredth (100^{th}) entries only. Meanwhile, the GT dataset will be reduced to identical amount according to the number of resulting valid congestion clusters identified in respective dataset to facilitate fair comparison.

3.6.2 Normalisation – Max Scaling

Normalisation was performed to provide common scale, in which min-max scaling is one of the most used techniques. However, congestion length is seen to be not suitable to be scaled to the value of zero (0). Therefore, we use max scaling, X_{nm} to scale all congestion length distributions and confined them into range of value that is greater than zero to one ($0 < x \leq 1$), in which the observation of generated congestion length can be seen in equivalent scale and still maintaining the original distribution. The each datapoints, i.e., the obtained congestion length, X_i, is scaled against the greatest magnitude (maximum value, X_{max}) within each dataset, i.e., H_Mean, H_Min, H_Max, and the ground truth, facilitating fairer comparison of data pattern between the distinct frameworks. The calculation of max scaling is given in (5).

$$\boldsymbol{X_{nm} = \frac{X_i}{X_{max}}} \tag{5}$$

3.6.3 Measurement Techniques

Visualisation of all congestions lengths for all peak hours datasets are plotted for visual comparison by using `sns.violinplot()` function of seaborn library [78]. The interpretation of visualised distributions is then being assessed with similarity measurement techniques.

Similarity measurements for the mean and standard deviation (SD) are performed by utilising `pd.describe()` function of Pandas library which produces the said statistical information. Meanwhile, SciPy library are utilised to perform the calculation of EMD measurements and JSD measurement between two congestion length distributions using `scipy.stats.wasserstein_distance()` function and `scipy.stats.entropy()`, respectively.

CHAPTER 4

RESULT AND DISCUSSION

This section delves into the execution of the method, its results, and the discoveries made. The raw data was initially processed to remove any incorrect congestion scenarios. Subsequently, the clustering process yielded distinct clusters and noise points. This section is separated into two parts which detailed on the clustering result and the similarity measurements recorded for the proposed method.

The input file was a file of "`csv.out`" extension, that was obtained from the compressed files of November 2021 taxi GPS data in [66]. The size of the file for 2nd November 2021 data was approximately 150 MB, consisting of taxi GPS records spanning for approximately 24 hours. Entries of 1,983,657 rows were recorded from 3,258 unique taxis, with 9 columns wide in which the attributes data (values) were stored. Filtering for the time of interest was applied and eight unique datasets were obtained. For simplicity of explanation, a representative dataset will be used as example datasets among the peak hours. Meanwhile, the result of each dataset is presented within appendix E. The representative of dataset was selected by analysing the highest vehicle count before and after preprocessing. Figure 19 illustrates the outcomes for all eight distinct datasets.

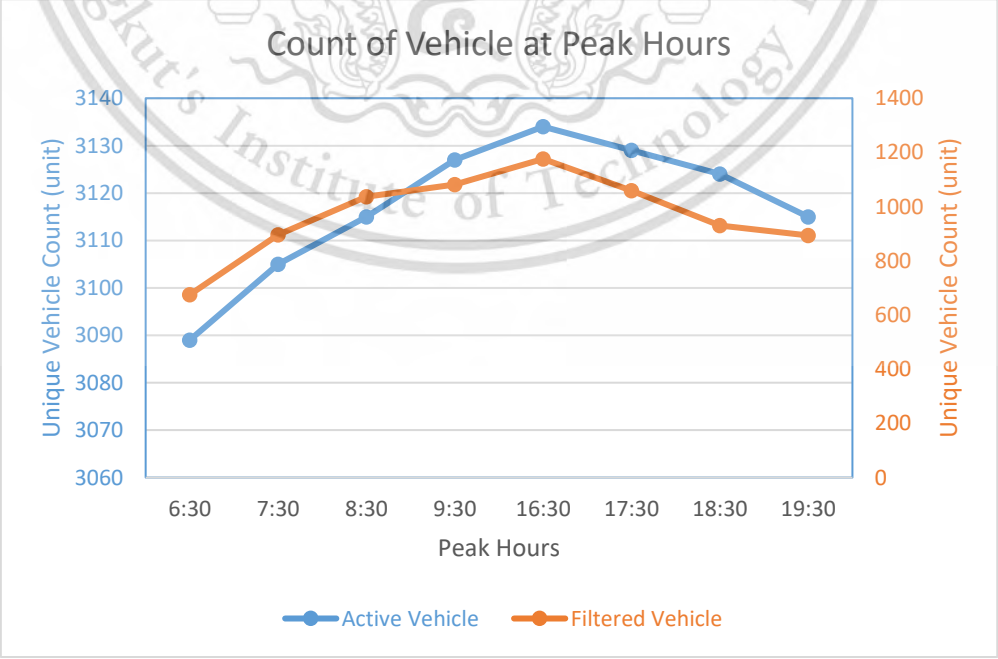


Figure 19. Vehicle Count Before and After Filtering for All Peak Hours' Datasets

The eight distinct datasets were having about approximately 3,117 unique vehicles count on average in each dataset. If the typical active contributor for Longdo Traffic's data recorded between 31,000 – 53,000 users, in which 42,000 users on average at any given time, the usage of the iTIC taxi GPS data covered approximately 7.42143% of the user contributor for in the total period of 20 minutes, i.e., cumulative span of TF1 and TF2 The details of the unique vehicle count for each dataset is illustrated summarised in Table 17.

Table 17. Summary of Unique Vehicle Count for All Datasets

Dataset	Measurement	Clustering Process			
		Raw	Filtered	TF1	TF2
1	Data count	26617	3405	1622	1783
	Vehicle ID count	3089	675	492	531
2	Data count	28983	5032	2397	2635
	Vehicle ID count	3105	896	672	724
3	Data count	30925	5816	2956	2860
	Vehicle ID count	3115	1037	805	815
4	Data count	31317	6233	3157	3076
	Vehicle ID count	3127	1082	842	853
5	Data count	30466	6850	3355	3495
	Vehicle ID count	3134	1176	933	927
6	Data count	31702	6608	3439	3169
	Vehicle ID count	3129	1059	866	808
7	Data count	30500	5956	2904	3052
	Vehicle ID count	3124	930	755	734
8	Data count	29573	5150	2546	2604
	Vehicle ID count	3115	894	710	709

4.1 PREPROCESSING RESULT

The “Dataset 5” which records the data at peak hour of time 16:30 is chosen as the representative for result explanation purpose within this section. The details of preprocessing outcomes for the representative dataset are recorded in Table 18.

Table 18. Preprocessing Outcomes of Dataset at Time 16:30 (“Dataset 5”)

Measurement	Data Frame			
	Raw	Filtered	TF1	TF2
Data count	30466	6850	3355	3495
Vehicle ID count	3134	1176	933	927

Initially, the dataset contains raw data of 30,466 entries from 3134 unique taxi units. After, data preprocessing, the dataset produced 6,850 data entries from 1176 unique taxi units which is then split into 2 different time frames, TF1 (16:10:00 to 16:19:59) and TF2 (16:20:00 to 16:29:59). Entries of 3355 data from 933 unique vehicles were hold by TF1, while TF2 was having 3945 data entries from 927 unique vehicles. The highest unique vehicle in the “Dataset 5” in comparison to others might had been contributed by ending of office hours.

4.1.1 Clustering Algorithm Effictiveness

The data of prior timeframe (TF1) from dataset 5 is used as input for clustering algorithms of HDBSCAN and K-Means++ to determine their effectiveness in clustering congestion hotspot. HBSCAN produced 355 clusters from TF1 while K-Means++ produces 390 clusters valid clusters out 505 intersections quantity in Bangkok that is assigned as the value of *k*-clusters numbers. The clustering output for both algorithms is given in Figure 20.

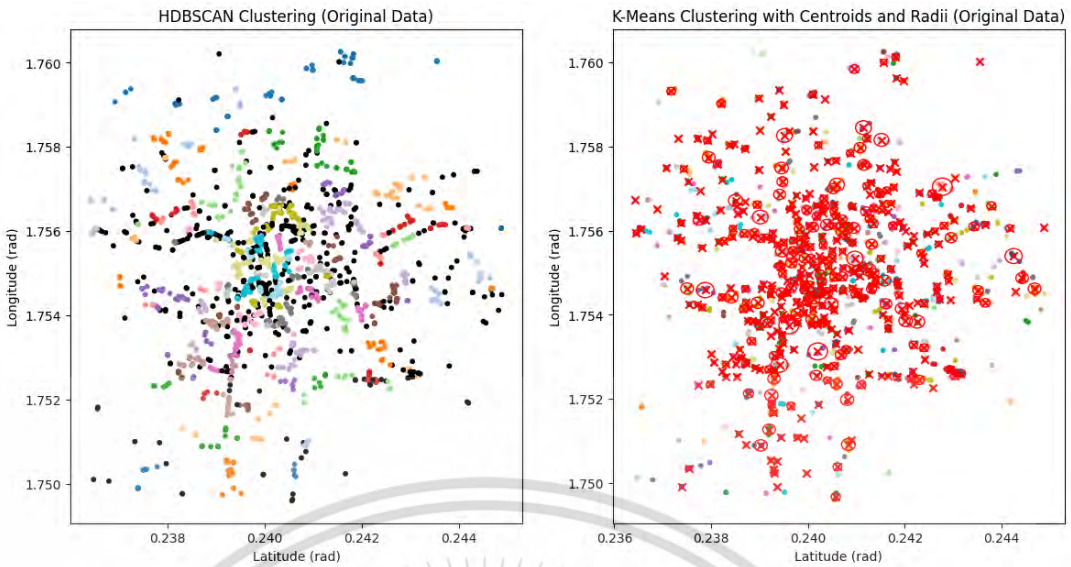


Figure 20. Clustering Outcomes with HDBSCAN and K-Means++

The noise points for both clusters in Figure 20 has not been removed whereby for HDBSCAN, it is marked with black coloured datapoints, while the valid clusters within K-Means++ is marked with circle for each cluster with "x" mark as its centroid locations. Visually, the noise points look almost identical which may indicate similar performance between the two algorithms. However, the K-Means derived algorithms such as K-Means++ are widely known to produce a cluster that made of having circular shape due to the spherical decision boundary that is made around the assigned centroid.

The result obtained by K-Means++ within this comparison is greatly influenced by the initial clusters value in which the quantity of intersections within Bangkok is used to produce optimal point of clusters that capture the traffic conditions from one road to another as they got connected in the intersection. On most occasion, the number of clusters will be statistically tested and iterated to find the best cluster quantity which will be inefficient on the computing resources. However, better performing algorithm is HBSCAN not only due to its ability in capturing arbitrary shape, also because its performance metrics results showed better performance which is summarised in Table 19.

Table 19. Result of Clustering Performance for HDBSCAN and K-Means++

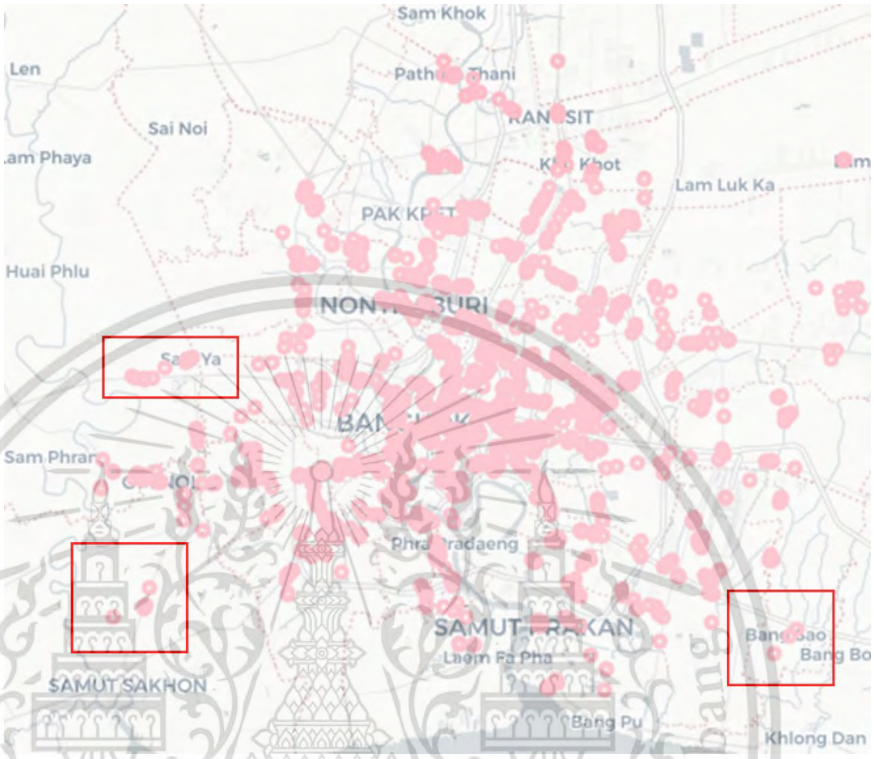
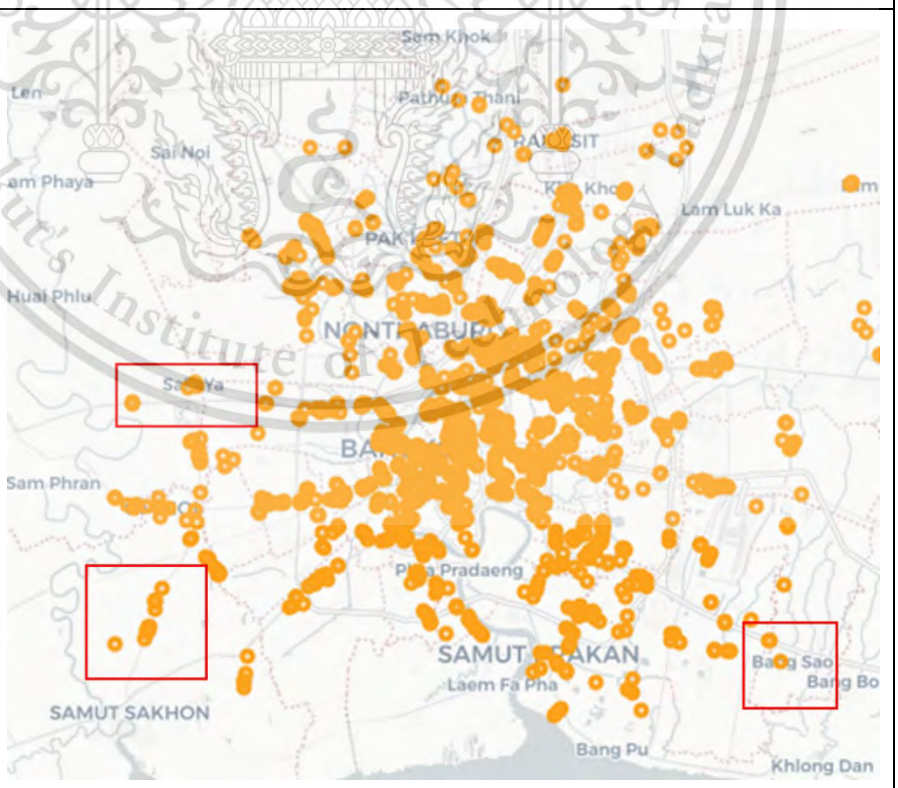
Clustering Algorithm	Assessment		
	Silhouette	Davies Bouldin	Cluster Output
HDBSCAN	0.72470	0.34752	355
K-Means++	0.69540	0.41625	390

The Silhouette Scores obtained by HDBSCAN indicated greater cohesion between datapoints within cluster in comparison with K-Means++, resulting in more compact clusters. On the other hand, the lower Davies Bouldin Scores obtained by HDBSCAN further strengthen the positive result obtained in Silhouette Score, indicating that HDBSCAN is more capable of producing well-separated clusters with tight datapoints within each cluster compared to K-Means++.

4.2 CLUSTERING RESULT: CONGESTION HOTSPOT CLUSTERS AND CONGESTION AREA CLUSTERS

Distance matrix with Haversine formula was developed to record the distance between two points in each timeframe. The distance matrix produced was fed into HDBSCAN algorithm as the starting point of the congestion hotspot clustering process with minimum cluster size of 3 datapoints and leaf clustering method, producing more fine-grained clusters, in which prior congestion hotspots of 355 clusters and later congestion hotspots of 364 clusters were identified. The illustration of identified prior congestion hotspot and later congestion hotspot are given in Table 20.

Table 20. Identified Prior and Later Congestion Hotspot at Time 16:30

Congestion Identifier	Visualisation
Prior Congestion Hotspots (16:10:00 to 16:19:59)	
Later Congestion Hotspots (16:20:00 to 16:29:59)	

The identified cluster of congestion hotspot from the given input of TF1 and TF2 produced many similar congestion areas in term of coordinate, especially within the Bangkok city and its vicinity. However, towards the ending of BMR, difference of congestion area in the prior timeframe and later timeframe can be seen through the red square marking that was given in the illustration of the congestion hotspots.

The congestion area clustering process was started by taking data points that were made from clusters of prior congestion hotspot and later congestion hotspot. The identified congestion hotspot clusters were converted into data points for. by assessing the appropriate representative point within the cluster and combined into one dataset. Then, clustering of congestion area was started by feeding the distance matrix of representative points into HDBSCAN algorithm with minimum cluster value of 2 datapoints and leaf clustering method.

Initially, the clustering process clumped all data points (congested hotspots) in a single large cluster i.e., the root of the cluster hierarchy. Then, a Minimum Spanning Tree (MST) of the mutual reachability distances between points was constructed, in which the edges in this MST represent the proximity between points based on density. Afterwards, the algorithm started to remove edges from the MST in descending order of weight (distance) value, i.e. removal by thresholding. The removal of an edge could have caused the cluster to split into two child clusters. The edge removal could also cause some points to get disconnected and fall out of the main clusters. These fallen points formed "child" clusters. Any “child” cluster that consisted of satapoints lesser than the defined value of “`min_cluster_size`” parameter was declared as noise points [71]. The details of congestion result output are given within Table 21.

Table 21. Clustering Process Outcomes on Time of Interest (16:30)

Measurement	Clustering Process		
	Prior hotspots (data points)	Later hotspots (data points)	Congestion Area (cluster)
Input count	3355	3495	719
Outlier count	496	469	169
Cluster count	355	364	122

There are 122 clusters of congested area identified in BMR through HDBSCAN at time 16:30. The illustration of identified congestion area clustering with HDBSCAN produced 122 clusters of congested area. The identified congestion areas around Bangkok are shown in Figure 21 whereby each cluster of congestion area is represented with same colour and located within the reasonable proximity. Consequently, the sample of input clusters that were affected during the identification of congestion area is given in Table 22.

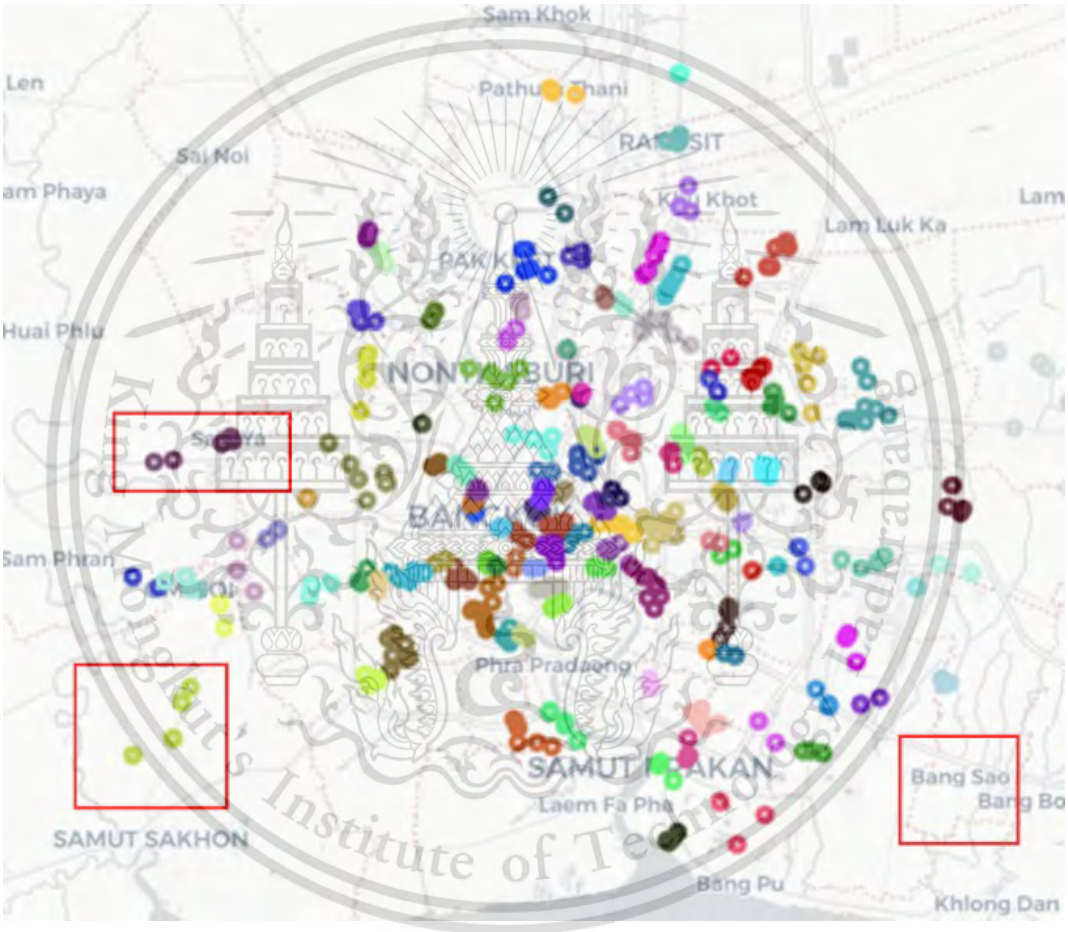
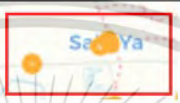


Figure 21. Identified Congestion Area in Bangkok with Its Associated Hotspots

Table 22. Sample of Input Clusters and Identified Congestion Area

No.	Input Clusters		Output Cluster
	Prior Hotspot	Later Hotspot	Congested Area
1			
2			
3			
4			
5			

The HDBSCAN algorithm did not necessarily produce congestion cluster whenever data points existed within prior and later timeframe such as seen in the Table 22 given previously. The application of “representative coordinate” caused the input clusters to be represented by either the “mode” or the “average” coordinate within its cluster. The concentration of data points in observation 1 within Table 22 was towards the top left of the red box while for the later hotspot, it was around the center of the box. Such distribution of data points could have resulted in the failure of meeting the criteria of congestion area in HDBSCAN clustering from distance perspective.

Demonstration of the adaptability of the HDBSCAN clustering algorithm was further strengthened by the successful clustering of densely packed area of data points within observation 5 from Table 22 into two distinct clusters and grouping sparsely distributed data in observation 4 within Table 22 into one cluster. Such feat demonstrated the scalability of the proposed framework. Consequently, the summary

of all identified congestion areas quantity for all datasets is illustrated in Figure 22 and given in Table 23.

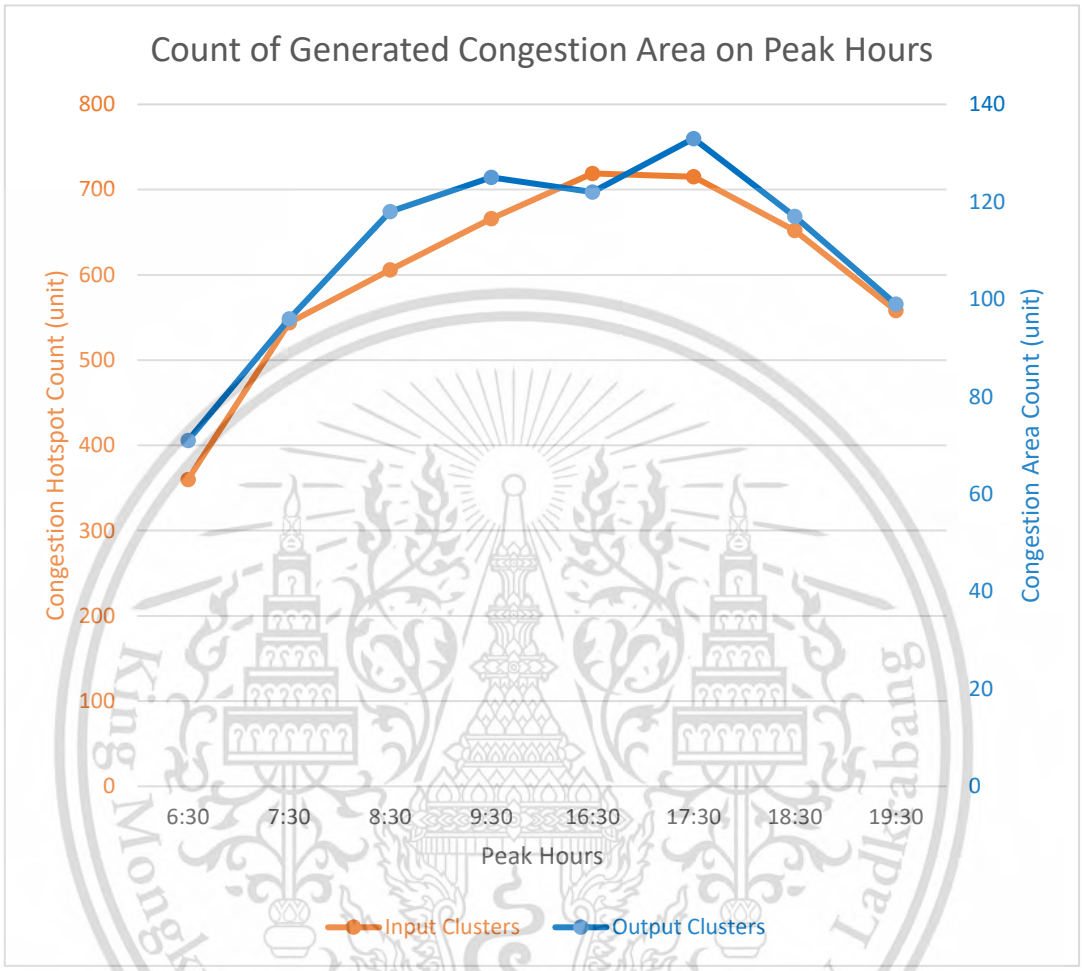


Figure 22. Summary of Generated Hotspots and Areas Count for All Datasets

Table 23. Summary of Input and Output for Congestion Area Clustering Process

Dataset	Measurement	Clustering Process		
		Prior hotspots (data point)	Later hotspots (data point)	Congestion Area (cluster)
1	Input count	1622	1783	360
	Outlier count	210	219	90
	Cluster count	171	189	71
2	Input count	2397	2633	544
	Outlier count	327	350	134
	Cluster count	258	286	96
3	Input count	2956	2860	606
	Outlier count	451	422	134
	Cluster count	305	301	118
4	Input count	3157	3076	666
	Outlier count	455	487	149
	Cluster count	343	323	125
5	Input count	3355	3495	719
	Outlier count	496	469	169
	Cluster count	355	364	122
6	Input count	3439	3169	715
	Outlier count	508	502	179
	Cluster count	389	326	133
7	Input count	2904	3052	652
	Outlier count	400	477	173
	Cluster count	322	330	117
8	Input count	2546	2604	558
	Outlier count	391	437	154
	Cluster count	278	280	99

4.3 SIMILARITY MEASUREMENTS

4.3.1 Valid Congestion Length and Congestion Area Clusters

A total of 122 congestion area clusters were identified in the dataset, each containing at least two data points. These data points, represented as pairwise hotspot distance (PH, LH), were utilised to calculate the congestion length within each cluster by using geodesic distance. The pairwise hotspot distance was calculated only when the hotspot pair were in (prior, later) orientation while others were filtered. This is due to the assumption that traffic congestion propagated from earlier time frame.

A filtering process was conducted to validate outcomes that complied with the maximum probable distance between the pair of data points. The filtering logic that was applied to the geodesic distance calculation process managed to identify and remove 10 invalid clusters. The removed clusters could be made of either one of reasons mentioned below:

1. The removed clusters are made of data points from the same time frame, e.g. prior hotspot only, or later hotspot only, or,
2. The removed clusters were made of data points that produced impractical geodesic distance value, either exceeding threshold or zero value.

The outcomes from geodesic distance calculation for the representative dataset produced 589 entries of valid distance data, which was recoded from 112 valid clusters for congestion length extraction process. The summary of the recorded valid entries, along with the quantity of input and output clusters for geodesic distance measurement in all datasets, is illustrated in Figure 23 and provided in Table 24.

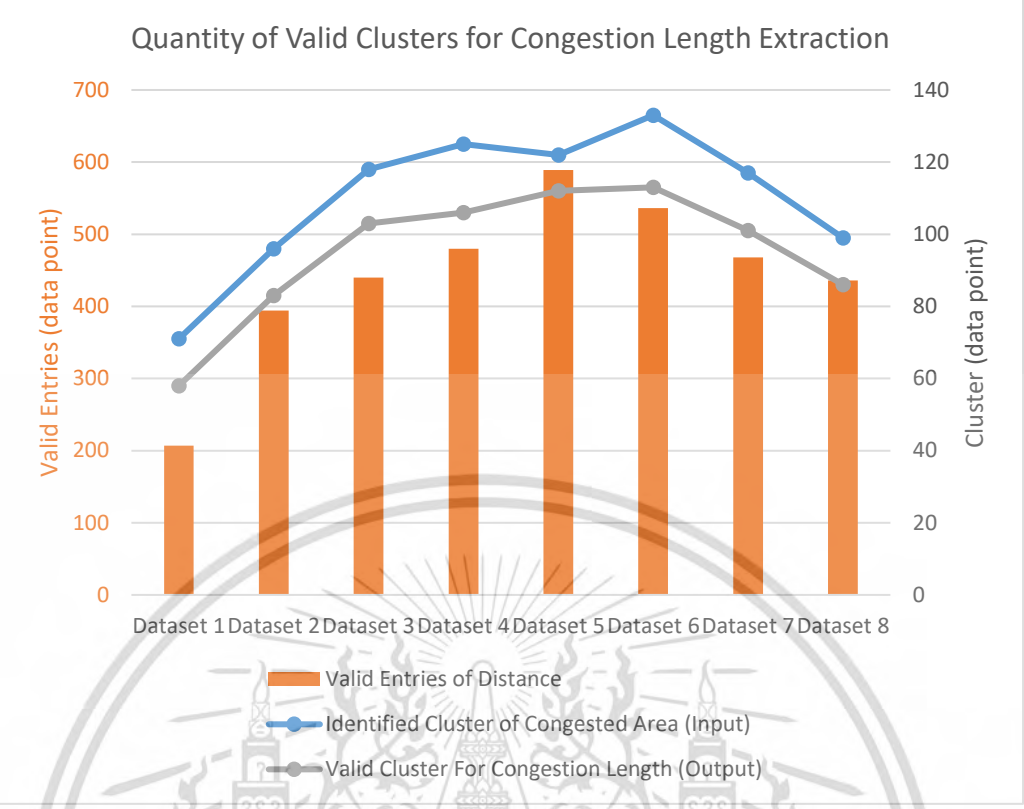


Figure 23. Summary of the Geodesic Distance Measurement in All Datasets

Table 24. Summary of the Geodesic Distance Measurement in All Datasets

Unique Identifier	Identified Cluster of Congested Area	Valid Entries of Distance	Valid Cluster For Congestion Length
Dataset 1	71	207	58
Dataset 2	96	394	83
Dataset 3	118	440	103
Dataset 4	125	480	106
Dataset 5	122	589	112
Dataset 6	133	536	113
Dataset 7	117	468	101
Dataset 8	99	436	86

The quantity of valid clusters of congested areas and its entries outcome however varied according to the amount of data entries and the distribution of the data itself within a particular dataset. Then, the congestion length extraction process for each H_Mean, H_Min, and H_Max approaches were performed by utilising the valid dataset of 589 data entries in which was grouped by its cluster label before the extraction, resulting in three distinct patterns of congestion length.

The number of valid clusters, i.e., 112 clusters in total, that were obtained from the representative dataset has exceeded the number of entries in the ground truth data. Therefore, the valid congested area was trimmed to the top 100 entries in descending order, i.e., to provide fair comparison between the ground truth and the test data. Meanwhile, in the dataset for traffic time of 6:30 (“Dataset 1”), the GT data was trimmed to the top 58 entries. The trimmed top 100 congestion length for all approaches and the ground truth at the time 16:30 (“Dataset 5”) were plotted and presented in Figure 24.

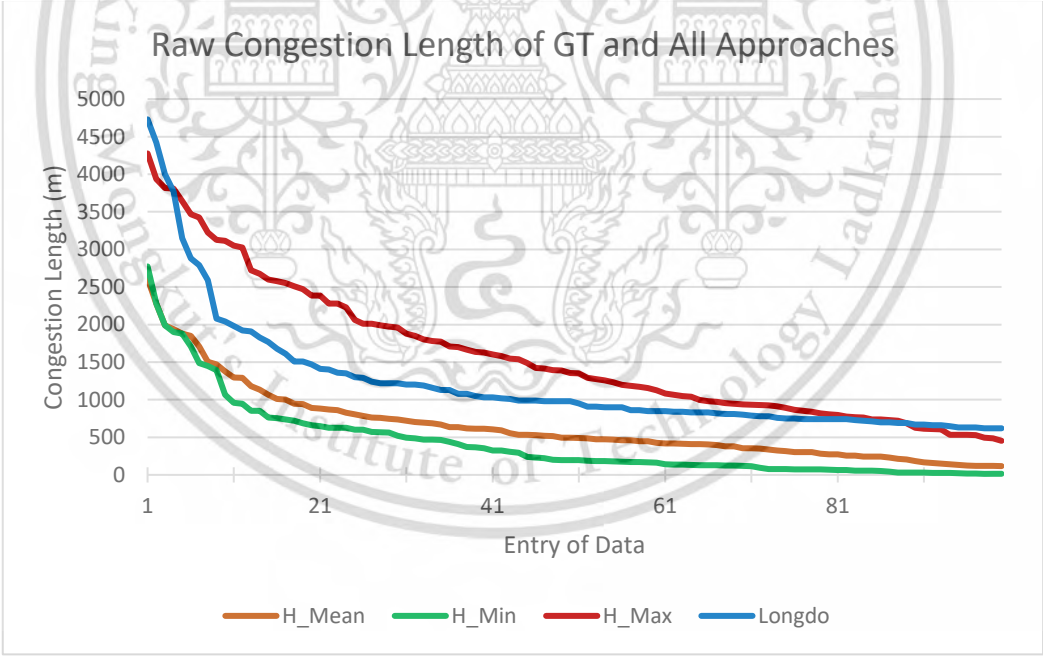


Figure 24. Extracted Congestion Length for All Approaches and Ground Truth

The plot showed that H_Mean approach and the H_Min approach were producing acceptable overall shape of the congestion length distribution, i.e., decaying exponentially at approximately similar rate, in which both approaches

produced congestion length values that were lower by approximately half compared to the GT in the range of 1st – 41st entry of the data. Meanwhile, the H_Max approach produced overshot measurement of from the range of approximately 5th within the data up to the 80th. The H_Mean approach produced the best decay rate while the mean was decaying faster and the H_Max approach had the slowest decay rate.

The usage of different datasets between the ground truth and this research (iTIC taxi GPS probe) resulted in congestion lengths that were not possible to be compared directly for performance measurement. Therefore, the congestion length data is scaled against the maximum congestion length value extracted by each approach respectively. The snippet of the max-normalised data is given in Figure 25.

	Longdo	H_Mean	H_Min	H_Max
0	1.000000	1.000000	1.000000	1.000000
1	0.970509	0.653517	0.653517	0.980814
2	0.709115	0.578360	0.578360	0.913933
3	0.674263	0.572400	0.571961	0.857776
4	0.406166	0.571961	0.500898	0.842957

Figure 25. Snippet of Max-Normalised Extracted Congestion Lengths

4.3.2 Similarity Measurement Results

pAcknowledging that the dataset used to produce traffic congestion pattern by the GT and the research is different, usage of common scale (ratio) was seen to be the more suitable approach in comparing the similarity of congestion length instead of the directly measured value of congestion length. Max scaling technique facilitated the normalisation process. The normalised distribution of valid congestion length data that was extracted from HDBSCAN's congestion areas and Longdo Traffic's top 100 most congested roads in Bangkok length is illustrated in Figure 26.

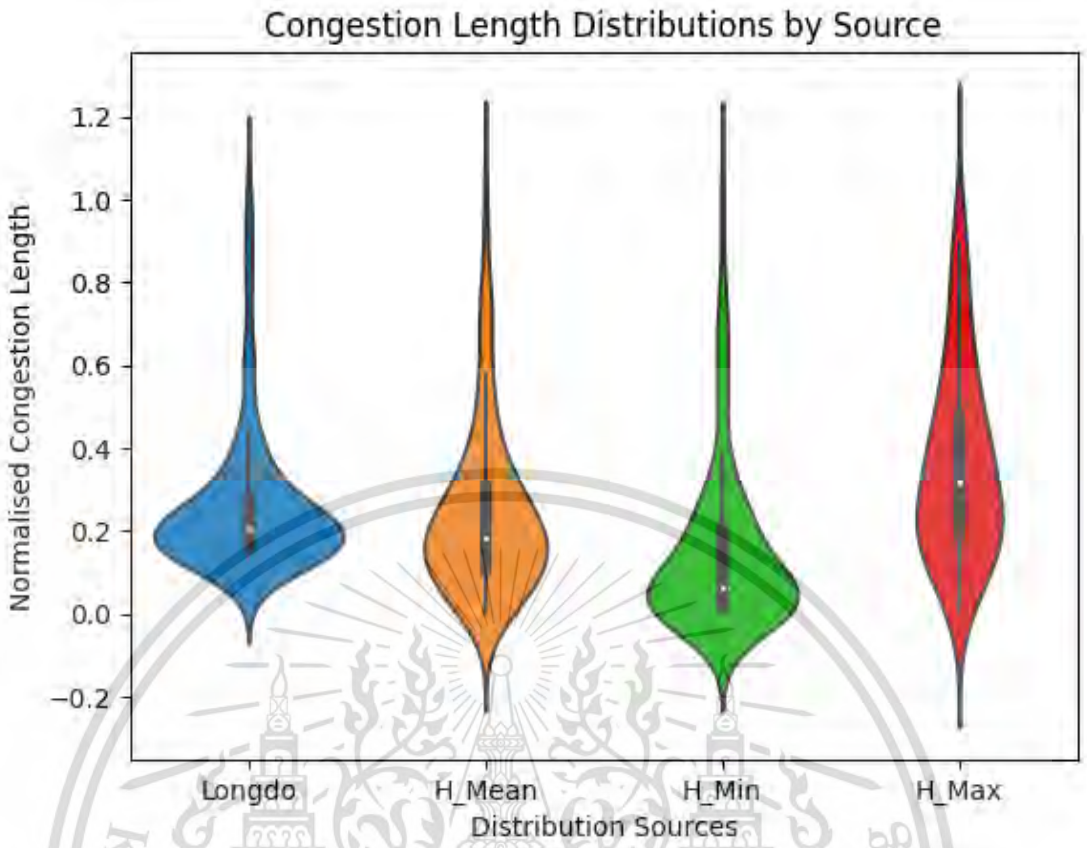


Figure 26. Congestion Length Distribution Based on Sources at 16:30

Visually inspecting the congestion pattern distribution of the experimental result and the GT that was illustrated with Violin Plot in Figure 26, three important observations can be made which are:

1. Median of the data that was represented by white dot within the box plot,
2. Violin plots' shape which represented data density,
3. Interquartile Range (IQR) of the data which describes the spread of the data, which was represented by the box plot.

The median value produced by the H_Mean approach was closest to Longdo's, with ending point of IQR that was also to Longdo's, in which those features indicated that considerable or even significant amount of congestion length values were shared between the GT and the H_Mean approach despite the data point density, in comparison to other approaches. Further, the H_Mean was having starting point of IQR closest to Longdo's IQR starting point. Lower IQR difference can be interpreted as better accuracy at predicting at which congestion length that

propagation of congestion most likely to happen. he density or concentration traffic congestion

Density of data distribution for Longdo Traffic, H_Mean, and H_min can be said to be having similar spread of frequency pattern. However, the H_Mean approach produced high frequency of congestion length in the region of data between the value of ~0.1 to 0.3, in which it was most alike to the data of GT, in comparison with other approaches. Despite, the H_Min approach produced more accurate accumulation of congestion length, resulting in more similar intensity of data point density which reflected through the height of frequency (shape) within the region below the value of ~0.1. Regardless, in term of congestion length value, H_Min approach produced many congestion length values that were erroneous (i.e., by at least 50%) when compared against Longdo's, which was reflected by the position of its median i.e., starting before the earliest data point within the GT's distribution.

The H_Max approach showed the least similarity in all observations i.e., median, IQR, and density of data distribution. The H_Max approach extracted distance of congestion length between two furthest points in a particular valid cluster, resulting in high magnitude congestion lengths that was reflected in an almost binomial distribution-like pattern that was slightly skewed to left (the lower 50% of data). The median of H_Max was located at a value whereby the IQR of the GT's distribution had ended, resulting in highly dissimilar density of data distribution pattern.

The similarity measurements between Longdo Traffic's top 100 most congested roads length and our HDBSCAN synthesised congestion lengths were conducted to support the observations that were derived from the visualised congestion propagation distributions. The summary of the utilised metrics i.e., measures of central tendency tools (mean), measure of dispersion (standard deviation, SD), JSD, and ESD were given in Table 25.

Table 25. Result of Similarity Measures of Congestion Length Distribution

SIMILARITY MEASUREMENT	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.25803	0.23916	0.14679	0.35779
Standard Deviation (SD)	0.16918	0.19909	0.19746	0.23245
Difference in mean against GT	0	0.01887	0.11124	0.09976
		7.3131 %	43.11127 %	38.66217 %
Difference in SD against GT	0	0.02991	0.02828	0.06327
		17.67939 %	16.71592 %	37.39804 %
JSD	0	0.02359	0.09818	0.02022
EMD	0	0.04677	0.11288	0.12052

The H_mean approach recorded difference in central tendency value of approximately 7% error with spread of data that was wider by approximately 18% against the GT. On the other hand, the H_Min approach scored the lowest standard deviation in all approaches against the GT at approximately 17%, i.e., lower by 1% compared to the H_Mean approach. However, the H_mean approach recorded the highest difference for the measure of central tendency against the GT at approximately 43%. The H_Max approach produced difference by approximately 38% in both measure of central tendency and the data spread from model, indicating a more widely distributed data compared to the GT.

The H_Mean approach obtained the best overall the lowest dissimilarity from the GT distribution in comparison to other approaches. Lower values in central tendency dissimilarity can be translated into lower error in each individual congestion length estimation, while lower SD error can be interpreted as better smaller data variation from the mean of the data. However, since all approaches are not normally

distributed (including the GT), the presented parametric statistical measurements were less relevant, i.e., complementary data to support observation made for the violin plots earlier.

Non-parametric statistical measurement are the more appropriate assessment tools to gauge the accuracy of the produced congestion length for non-normally distributed data since the distribution plot for all approaches and the GT were skewed to the left. The H_Mean approach scored the lowest EMD at 0.04677 in comparison to other approaches. The lowest EMD score recorded by the H_Mean approach can be interpreted as the approach that generated the congestion length with least error in each data points. Meanwhile, the H_Mean and H_Max recorded almost similar EMD scores at 0.11288 and 0.12052, respectively, indicated that they both accumulated almost similar total errors in congestion length which was justified by the median location of both distribution within the violin plot.

The Jensen-Shannon Divergence (JSD) score of the H_Max approach was the lowest among the approaches i.e., 0.02022 due to the distribution of the H_Max approach that stretched with considerably uniform data point density, resulting in low dissimilarity when the average of symmetrical divergence was measured. On the other hand, the H_Min approach scored the highest dissimilarity, i.e., 0.09818, due to the starting point of data distribution and the data concentration which was located before the starting of the GT's distribution, resulting in considerable difference around the region of data below the value of 0.2. Meanwhile, the H_Mean approach recorded JSD score of 0.02359 which is comparable to the H_Max approach. However, the usage of JSD score alone are not sufficient to measure the similarity of distributions because it had been shown that the H_Max was having the least similar data distribution from the violin plot. The use of ESD measurement is a must to validate the similarity result given by the JSD measurement because the ESD measurement indicates how much "work" need to be done to transform the compared distribution into the ground truth distribution.

4.3.3 Overall Result

The H_Mean approach was proven to be the best performing among others for the “Dataset 5”. However, the H_Min approach outperformed H_Mean in term distribution similarity whereby it generally has a sharper peak when inspected visually via violin plot, which is given in Figure 27.

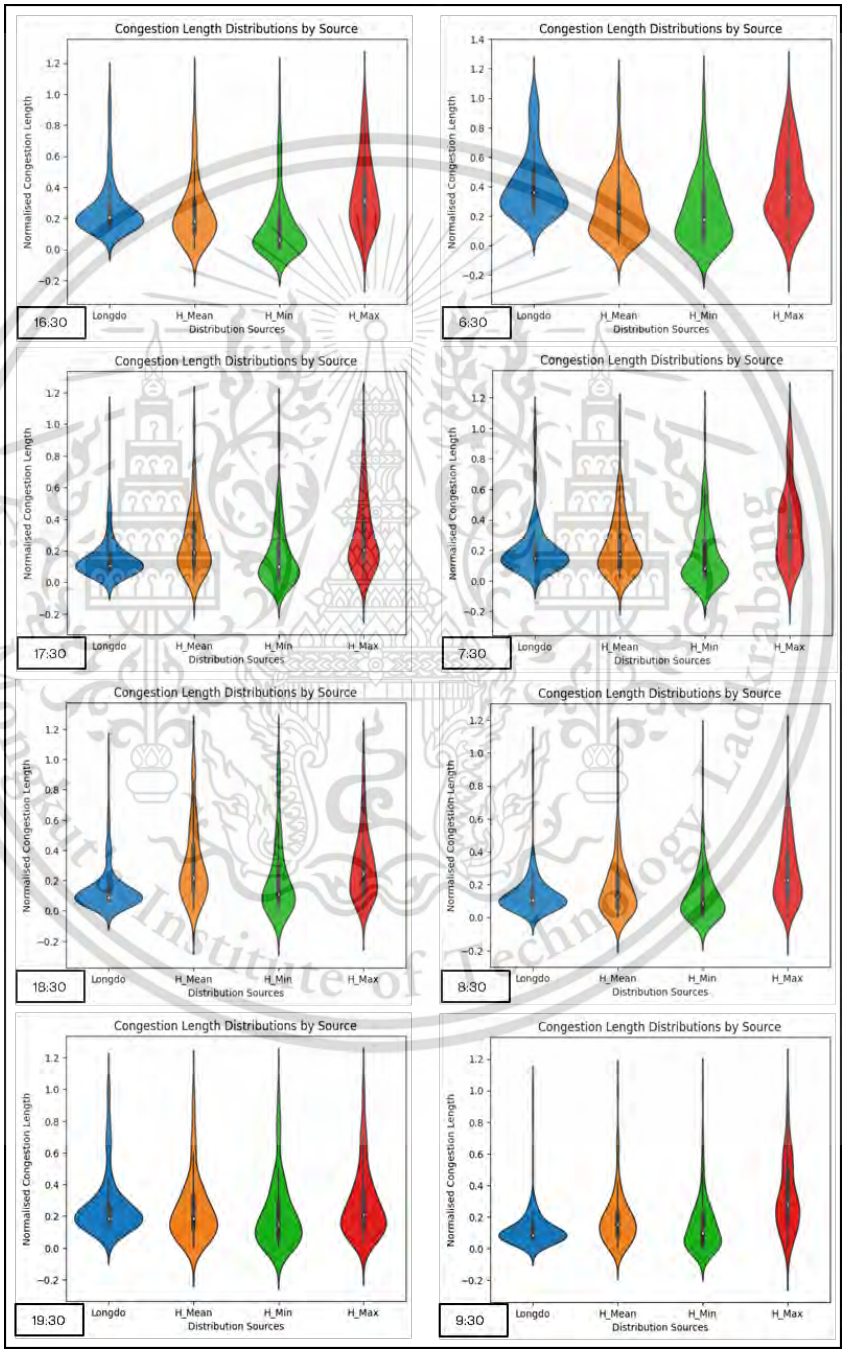


Figure 27. Violin Plot of All Approaches in All Datasets

Congestion in Bangkok is said to be started from 6:30 for the morning and 16:30 for the evening, meanwhile the congestion deescalates at 9:30 and 19:30 for the morning and evening, respectively. By analysing Figure 27, there are 5 points that can be shared:

1. At the starting of congestion (less intense) period, the whisker of box plot in GT extended beyond the value of 0.4, that is at time 6:30 and 16:30
2. As the congestion intensify, the whisker of box plot in GT only exists below the value of 0.4, that is at time 7:30, 8:30, 9:30,17:30, 18:30.
3. Looking at the trend of congestion length shown by GT from the time of 18:30 (intense congestion) to 19:30 (less intense congestion), we can deduce that the traffic condition at time 10:30 will have similar congestion trend to the time 19:30, just as how the transition of congestion has been shown from more intense to less intense period.
4. Congestion length is said to be less intense from the violin plot despite longer congestion length maybe due to the traffic management model that is used by Longdo Traffic, in which clustering might be used which caused the congestion length to be longer during less intense periods, while getting shorter during more intense period due to existence of more datapoints within the clusters.
5. Overall, H_Mean and H_Min captured similar distribution trend, however H_Min always has a shaper peak on the distribution pattern in comparison with H_Min. This may indicate that H_Min could have performed if the distance extracted can be slightly increase.
6. Poor performance of distribution during the time of 17:30 and 18:30 for both H_Mean and H_Min may have been contributed by driving behaviour of taxi whereby driver will try to take shortcut route while Longdo Traffic contributors stay on the main road.

The H_Mean approach recoded the least measurement error on congestion length through its EMD measurement score and the best similarity in distribution pattern of congestion length, which was reflected in its violin plot and JSD measurement score. Consequently, the H_Mean approach was found to be the best performing approach within all datasets, in which it recorded EMD score of 0.16551 in

the “Dataset 7”. The summary of the scores for EMD and JSD measurements of all approaches are given in Figure 28 and Figure 29, respectively.

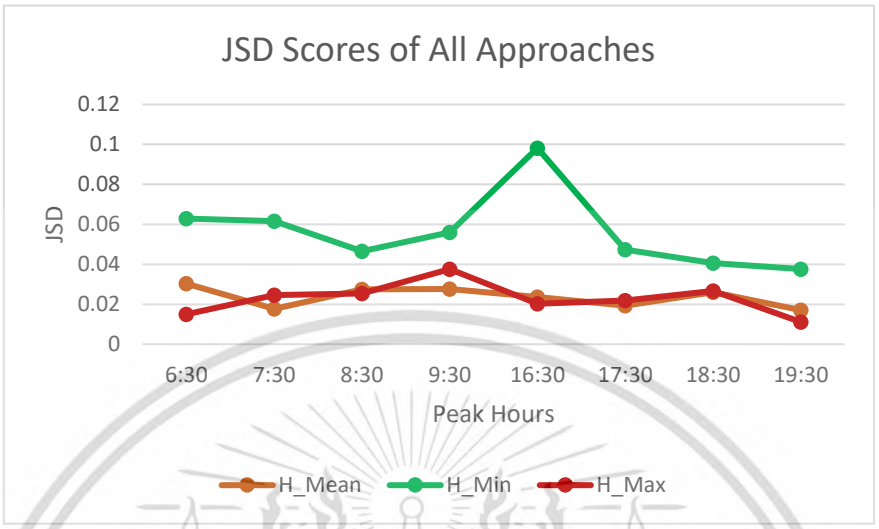


Figure 28. JSD Scores of All Approaches in All Datasets

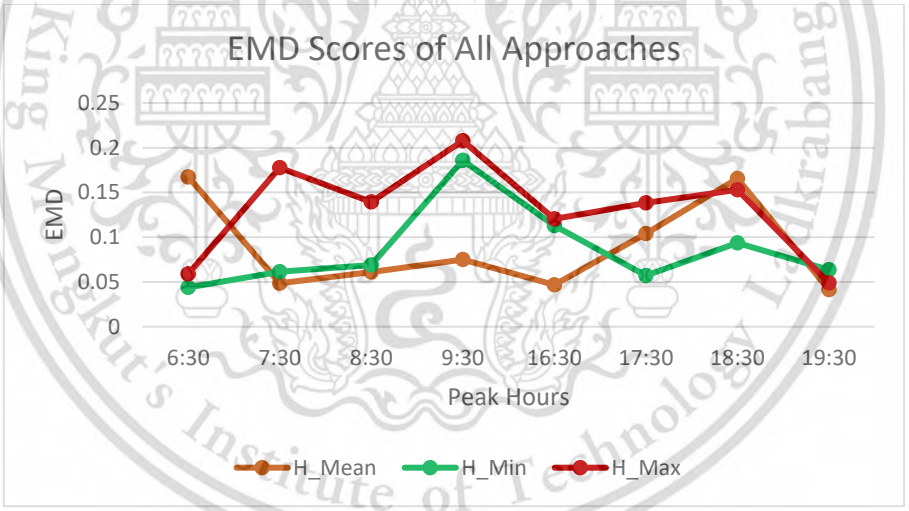


Figure 29. EMD Scores of All Approaches in All Datasets

Finally, the approach that scored the best on each EMD and JSD measurements, i.e., the lowest score compared to others for all peak hours’ dataset is illustrated Figure 30, while Figure 31 illustrated the worst performing approach on EMD and JSD measurement for all datasets.

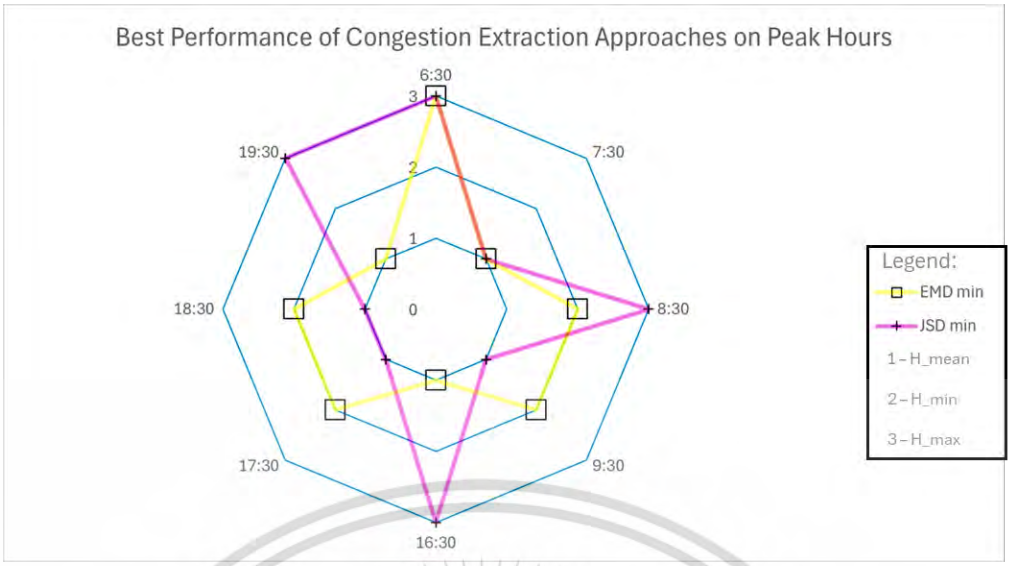


Figure 30. Best Performing Approaches in All Datasets

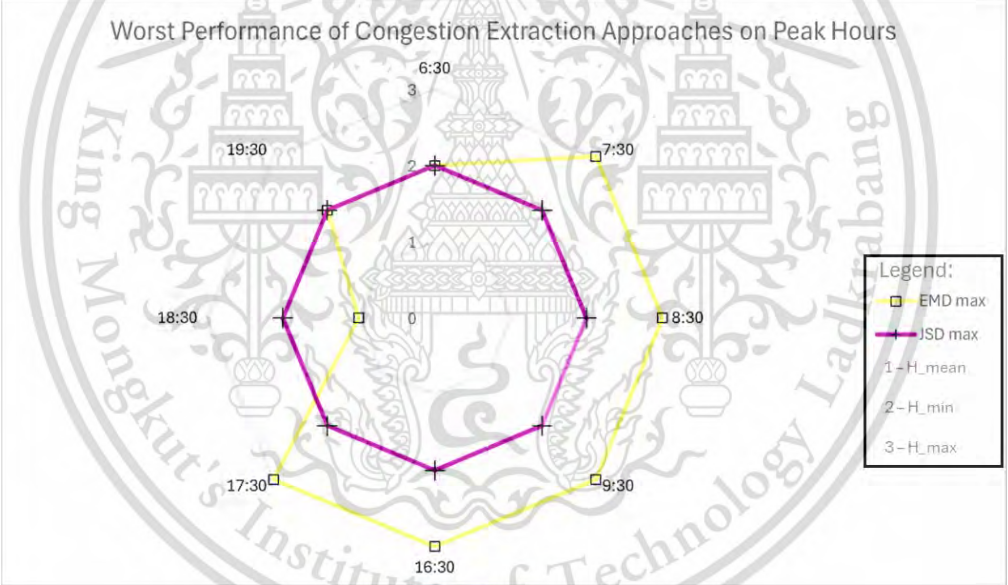


Figure 31. Worst Performing Approaches in All Datasets

4.3.4 Weekday and Weekday Congestion Pattern

Congestion pattern of Bangkok is further explored through its length by using HDBSCAN on the weekday and weekend. Two distinct dataset was prepared with 5 days of weekday data (1 week), and 4 days of weekend data (2 days of Saturday and 2 days of Sunday). The datasets were clustered with HDBSCAN and the congested length were extracted with the 3 approaches. We use Longdo Traffic's top 100

congested road list as our reference for weekday as Longdo Traffic stated that the dataset used in their top 100 congested road list in Bangkok were made up of weekday data only because weekend traffic has unique patterns [37]. The congestion length obtained were visualised via violin plot, then the traffic congestion length distributions for all congestion hours in the morning and evening of weekday and weekend are placed side by side according to its congestion hours which is given in Figure 32 and Figure 33, respectively.

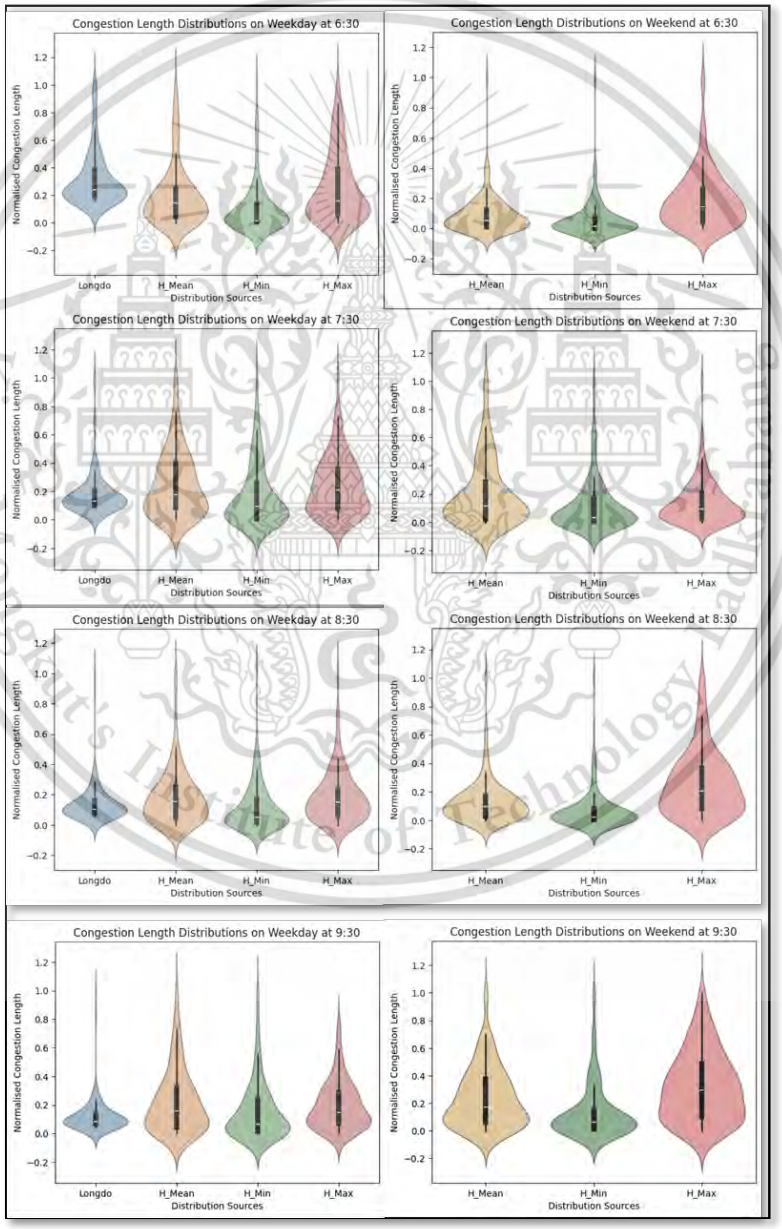


Figure 32. Morning Congestion Length Pattern on Weekday and Weekend in Bangkok

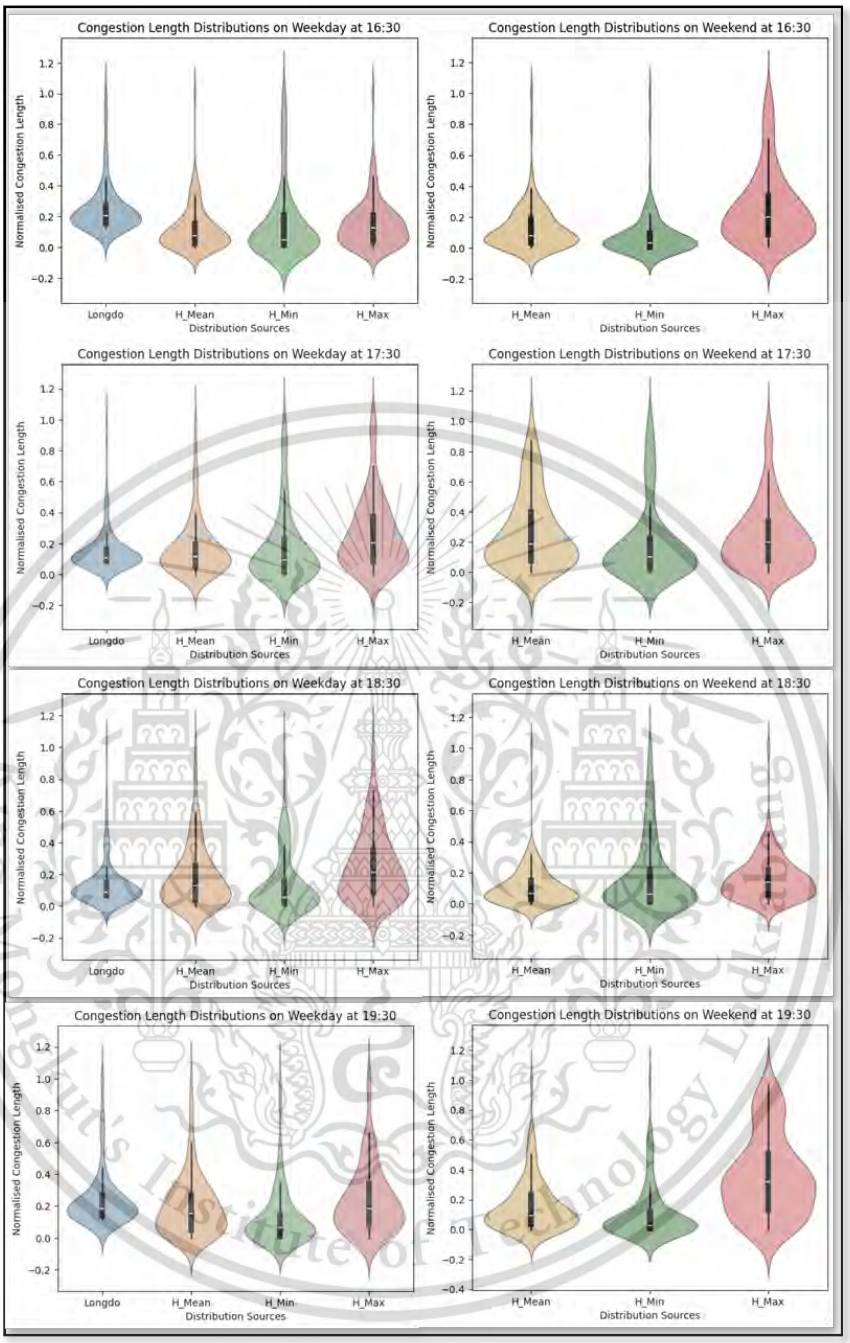


Figure 33. Evening Congestion Length Pattern on Weekday and Weekend in Bangkok

The performance of H_Min within the weekday dataset has shown improvement when more data is being used, resulting in the best similarity of distribution trend against GT with minimum EMD that is reflected from the span of its IQR. Looking at the performance shown during the weekend, we can assume that

H_Min and H_Mean are the better models to be used as reference for the congestion length pattern during weekday.

The data used by Longdo Traffic is taken from multiple sources as mentioned earlier (in section 2 and section 4.3.4) which might have caused the performance of the proposed methods to vary. As more data is being used for clustering, the driving behaviour of taxi drivers affects the characteristic of the dataset and clustering result. Meanwhile, Longdo Traffic takes in data from private vehicles which tend to stay on main road instead of shortcuts because of limited knowledge on Bangkok's traffic network and travel route. Regardless, H_Min showed promising performance.

Establishing the estimation of pattern between weekend and weekday may be challenging due to the unique nature that was mentioned by Longdo Traffic [37]. Regardless, an estimation to the weekend congestion length pattern was performed by analysing and matching the span of IQR and whisker of the box plot, then matching the weekend span onto similar weekday span for each congestion length extraction model. This method should allow us to get the estimate of the traffic congestion length pattern from the Longdo Traffic data. We can deduce by matching for each of the pattern as below:

1. The pattern shown by H_Min at weekend (7:30, 19:30) are similar to the pattern shown by H_Min at weekday (6:30) in which traffic is less intense.
2. The pattern shown by H_Min at weekend (9:30) is similar to the pattern shown by H_Min at the weekday (8:30, 19:30) in which the traffic is intense.
3. The pattern shown by H_Min at weekend (17:30) is similar to the pattern shown by H_Min at the weekday (16:30) in which traffic is less intense.
4. The pattern shown by H_Min at weekend (18:30) is similar to the pattern shown by H_Min at the weekday (9:30) in which traffic is intense.
5. The pattern shown by H_Mean at weekend (6:30,16:30) is similar to the pattern shown by H_Mean at the weekday (16:30) in which traffic is less intense
6. The pattern shown by H_Mean at the weekend (8:30) is similar to the pattern shown by H_Mean at the weekend (6:30), thus also similar to H_Mean at the weekday (16:30) in which traffic is less intense.

Estimates of the weekend traffic condition by matching congestion length distribution to Longdo Traffic's 100 top congested road list established the peak hours at weekend exist at the time 9:30 and 18:30 for morning and evening respectively. The summary of weekend traffic conditions estimates by referring to Longdo Traffic is given in Table 26.

Table 26 Summary Weekend Traffic Condition

Pattern of Reference at Weekend Time	Similar Pattern of Reference at Weekday Time	Traffic Condition Referring to Longdo Traffic
H Mean (6:30)	H_Mean (16:30)	Less Intense
H_Min (7:30)	H_Min (6:30)	Less Intense
H Mean (8:30)	H_Mean (16:30)	Less Intense
H_Min (9:30)	H_Min (8:30) H_Min (19:30)	Intense
H Mean (16:30)	H_Mean (16:30)	Less Intense
H_Min (17:30)	H_Min (16:30)	Less Intense
H_Min (18:30)	H_Min (9:30)	Intense
H_Min (19:30)	H_Min (6:30)	Less Intense

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Traffic congestion poses a significant global challenge, exerting adverse effects on social and economic progress. This study aims to assess the efficacy of HDBSCAN in generating traffic congestion patterns from taxi GPS probe data in Bangkok and evaluate its performance. Clustering effectiveness of HDBSCAN has been assessed to be better in comparison to K-Means++. The simulation of traffic congestion patterns was produced through unsupervised clustering with HDBSCAN(leaf) and taxi GPS data,. Three distinct distance extraction methods were devised to derive congestion length from valid congestion areas and their performance were assessed by utilising similarity measurement techniques namely mean, standard deviation, JSD, and EMD. On average, the iTIC taxi GPS data accounted for only 7.4214% of the total contributors compared to the reported ground truth dataset (Longdo Traffic).

The H_Mean approach was found to be the best overall congestion length extraction throughout all datasets within this study. Demonstration of the approaches by using the “Dataset 5” showed that H_Mean approach managed to produce score of 7.3131% on the mean measure and 17.67939% for standard deviation measure. Further, the H_Mean approach attained a score of 0.04677 in EMD measurement and exhibited a divergence of 0.02359 in JSD measurement, indicating a 97.641% similarity to the ground truth data in the “Dataset 5”.

While HDBSCAN(leaf) demonstrated favourable results on certain datasets, specifically the “Dataset 5”, “Dataset 7”, and “Dataset 2”, it also exhibited poor performance on other datasets. The suboptimal performance was evident from central tendency (mean) or standard deviation (SD) measurements exceeding 20% error. Furthermore, a performance degradation was indicated when the Earth Mover's Distance (EMD) value exceeds 0.1, suggesting that the generated congestion length density becomes irrelevant. These indicators of poor performance are further illustrated through the analysis of the violin plots, as detailed below:

1. Central Tendency (Mean): the top frequency of data distribution started slower or later than the GT’s plot.
2. Standard Deviation: the peak of top frequency was not as sharp as the GT’s due to greater spread of data.

3. Earth Mover's Distance: for measurement below 0.1, the IQR between GT's and the approach would have an overlap at least by 30%, in which the shape of distribution was generally the similar, but the plot started at either an earlier, or a later value that was significantly far from the GT's plot. Starting point.

On weekday and weekend data, Longdo Traffic top 100 congested roads in Bangkok list is only made by weekday data meanwhile iTIC Taxi GPS probe dataset consist of transactions for all days, in which only estimation of traffic conditions can be performed on the weekend data. H_Min shows better performance than H_Mean in term of trend distribution similarity, in which it was used to estimate the traffic condition in the weekend, along with H_Mean via visual analysis and deduction. The congestion length pattern produced has revealed dissimilarity between Longdo Traffic and iTIC Taxi GPS probe dataset due to weightage of driver behaviours. As the amount of data in iTIC dataset increased, the effect of taxi driver behaviours caused increment in the extracted congestion length value, which may be contributed by private vehicles that tend to stay on the main road, while taxi might take shortcut due to more intensive knowledge on the traffic networks and travel routes within Bangkok. Regardless, H_Min managed to exhibit promising performance.

This research provided understanding of congestion pattern generation from taxi GPS probe. While the application of the proposed method is yet to be seen, the utilisation of "representative coordinate" within the framework has the potential that could facilitate complementary traffic congestion analysis within a centralised traffic monitoring and management system with lesser data resources by tracing the data points involved in the identified congestion cluster. However, this feature was not properly demonstrated within this research. While the result of some datasets indicated poor performances, they also indicated that potential for significant improvement with the inclusion of more comprehensive data. Given the huge difference in active contributors' quantity for the data was used for traffic congestion propagation generation compared to the ground truth dataset, the outcomes are promising.

There are several limitations in this research that can be addressed, particularly given that this study serves as a proof of concept. Improvements can be made in areas such as data quantity, feature selection, and data filtering. Possible

enhancements and extensions to this study include, but are not limited to, the following:

1. A greater amount of data, whether from a longer timeframe or increased data quantity, could potentially enhance pattern generation and congestion propagation determination, given that Longdo's Traffic windows are separated hourly.
2. Combine other taxi probe data sources, such as data owned by TSquare Traffic Information Service which was reported in [79], to increase the number of unique taxi contributor.
3. Amendment of data filtering approach to increase valid data points with congestion condition, which in result may improve the performance of clustering of the HDBSCAN algorithm. However, filtering amendment must be performed with caution as it may cause the proposed method to lose its scalability due to the applied constrained on the input data.
4. The exploration of HDBSCAN(eom) algorithm usage could result in different pattern generation performance.
5. Hybrid data source to reduce data sparseness due to limited sample presented in the less travelled region (smaller road network), e.g. loop detector, bus GPS data, etc.
6. Utilisation of different distance measurement techniques can be applied when developing distance matrix such as Manhattan distance, since it captures the shape of road better than the usage of Haversine formula alone. A hybrid approach combining these methods may increase accuracy.
7. Post processing techniques to handle standstill datapoints could be introduced to discover new pattern of congestion propagation.
8. The "`heading`" field within the dataset may be utilised to provide traffic congestion estimation at more detailed such as at the lane level.

We believe this extension will offer valuable insights for knowledge expansion in the field of traffic management worldwide, particularly in Bangkok, Thailand.

.

REFERENCES

[1] M. A. Fattah, S. R. Morshed, and A. Al Kafy, "Insights into the socio-economic impacts of traffic congestion in the port and industrial areas of Chittagong city, Bangladesh," *Transportation Engineering*, vol. 9, Sep. 2022, doi: 10.1016/j.treng.2022.100122.

[2] "The average speed of personal cars in Bangkok (Morning peak 06:00 - 09:00)," Department of Land Traffic System Development. Accessed: Feb. 28, 2023. [Online]. Available: https://otp.gdcatalog.go.th/dataset/dataset_12_01/resource/ca5fbc90-70fa-48ed-a84c-fe0c4012b138

[3] "Traffic Index ranking," TomTom International BV. Accessed: Jul. 22, 2023. [Online]. Available: https://www.tomtom.com/traffic-index/ranking/

[4] G. Aparicio, T. Iturralde, and A. V. Rodríguez, "Developments in the knowledge-based economy research field: a bibliometric literature review," *Management Review Quarterly*, vol. 73, no. 1, pp. 317–352, 2023, doi: 10.1007/s11301-021-00241-w.

[5] "Data for science, technology and innovation: Definitions, scope and objectives," in *Enhanced Access to Publicly Funded Data for Science, Technology and Innovation*, Paris: OECD Publishing, 2020, p. 14. doi: 10.1787/947717bc-en.

[6] Y. Li and J. Xiao, "Traffic peak period detection using traffic index cloud maps," *Physica A: Statistical Mechanics and its Applications*, vol. 553, p. 124277, 2020, doi: https://doi.org/10.1016/j.physa.2020.124277.

[7] Z. Kan, L. Tang, M. P. Kwan, C. Ren, D. Liu, and Q. Li, "Traffic congestion analysis at the turn level using Taxis' GPS trajectory data," *Comput Environ Urban Syst*, vol. 74, pp. 229–243, Mar. 2019, doi: 10.1016/j.compenvurbsys.2018.11.007.

[8] "Google Maps - Apps on Google Play." Accessed: Apr. 23, 2024. [Online]. Available: https://play.google.com/store/apps/details?id=com.google.android.apps.maps

[9] "Official Google Blog: The bright side of sitting in traffic: Crowdsourcing road congestion data." Accessed: Apr. 20, 2024. [Online]. Available: https://googleblog.blogspot.com/2009/08/bright-side-of-sitting-in-traffic.html

[10] "Google Maps 101: How AI helps predict traffic and determine routes." Accessed: Apr. 20, 2024. [Online]. Available: https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/

[11] "TLC Trip Record Data - TLC." Accessed: Apr. 20, 2024. [Online]. Available: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[12] "About TLC - TLC." Accessed: Apr. 20, 2024. [Online]. Available: https://www.nyc.gov/site/tlc/about/about-tlc.page

[13] "Index of/opendata/probe-data," Intelligent Traffic Information Center Foundation. Accessed: Apr. 03, 2023. [Online]. Available: https://itic.longdo.com/opendata/probe-data/

[14] "What is iTIC foundation? | iTIC : Thai Intelligent Traffic Information Center Foundation." Accessed: Apr. 20, 2024. [Online]. Available: https://org.iticfoundation.org/node/36

[15] "iTIC Open Data Archives." Accessed: Apr. 20, 2024. [Online]. Available: https://itic.longdo.com/opendata/

[16] "NYC DOT - About NYC DOT." Accessed: Apr. 28, 2024. [Online]. Available: https://www.nyc.gov/html/dot/html/about/about.shtml

[17] "Bangkok Post - Clearing up congestion, one point at a time." Accessed: Apr. 28, 2024. [Online]. Available: https://www.bangkokpost.com/thailand/general/1586374/clearing-up-congestion-one-point-at-a-time

[18] "Traffic Index ranking | TomTom Traffic Index." Accessed: Apr. 28, 2024. [Online]. Available: https://www.tomtom.com/traffic-index/ranking/

[19] "New York - Google Maps." Accessed: Apr. 30, 2024. [Online]. Available: https://www.google.com/maps/place/New+York,+NY,+USA/@40.7466786,-73.9676767,11z/data=!4m6!3m5!1s0x89c24fa5d33f083b:0xc80b8f06e177fe62!8m2!3d40.7127753!4d-74.0059728!16zL20vMDJfMjg2!5m1!1e1?entry=ttu

[20] "Bangkok - Google Maps." Accessed: Apr. 30, 2024. [Online]. Available: https://www.google.com/maps/place/Bangkok/@13.7245449,100.5510417,11z/data=!4m6!3m5!1s0x311d6032280d61f3:0x10100b25de24820!8m2!3d13.7563309!4d100.5017651!16zL20vMGZuMmc!5m1!1e1?entry=ttu

[21] "Chapter 2. Introduction to Vehicle Classification - Verification, Refinement, and Applicability of Long-Term Pavement Performance Vehicle Classification Rules , November 2014 - FHWA-

HRT-13-091." Accessed: Apr. 27, 2024. [Online]. Available: https://www.fhwa.dot.gov/publications/research/infrastructure/pavements/ltpp/13091/002.cfm
[22] "Highway Statistics Series - Policy | Federal Highway Administration." Accessed: Apr. 27, 2024. [Online]. Available: https://www.fhwa.dot.gov/policyinformation/statistics.cfm
[23] "Definition of Cars," Transportation Statistics Group, Planning Division, Department of Land Transport. Accessed: Jul. 02, 2023. [Online]. Available: https://web.dlt.go.th/statistics/index.php
[24] "Table MV-1 - Highway Statistics 2022 - Policy | Federal Highway Administration." Accessed: Apr. 27, 2024. [Online]. Available: https://www.fhwa.dot.gov/policyinformation/statistics/2022/mv1.cfm#foot3
[25] "NYC DOT - Truck or Commercial Vehicle?" Accessed: Apr. 28, 2024. [Online]. Available: https://www.nyc.gov/html/dot/html/motorist/truckorcomm.shtml
[26] "Thailand - Flash report, Automotive sales volume, 2022 - MarkLines Automotive Industry Portal." Accessed: Apr. 27, 2024. [Online]. Available: https://www.marklines.com/en/statistics/flash_sales/automotive-sales-in-thailand-by-month-2022
[27] "Focus2move| Thailand Best Selling Cars - Top 100 in 2022." Accessed: Apr. 28, 2024. [Online]. Available: https://www.focus2move.com/thailand-best-selling-car-2022/
[28] "2022 US Vehicle Sales Figures By Model | GCBC." Accessed: Apr. 28, 2024. [Online]. Available: https://www.goodcarbadcar.net/2022-us-vehicle-sales-figures-by-model/
[29] S. Siangsuebchart, S. Ninsawat, A. Witayangkurn, and S. Pravinvongvuth, "Public transport gps probe and rail gate data for assessing the pattern of human mobility in the bangkok metropolitan region, thailand," *Sustainability (Switzerland)*, vol. 13, no. 4, pp. 1–28, Feb. 2021, doi: 10.3390/su13042178.
[30] "aboutus | metamedia technology." Accessed: Apr. 21, 2024. [Online]. Available: https://www.mm.co.th/aboutus?locale=th
[31] "Longdo.COM launches two new services, Longdo Traffic and Longdo Mobile |." Accessed: Apr. 21, 2024. [Online]. Available: https://www.mm.co.th/20090815-longdo-traffic-longdo-mobile
[32] "Longdo Traffic Bangkok and Thailand Traffic Information Report Latest Traffic Information." Accessed: Apr. 22, 2024. [Online]. Available: https://traffic.longdo.com/about
[33] "Bangkok Post - Taxi OK, Taxi VIP apps to launch on Nov 9." Accessed: Apr. 22, 2024. [Online]. Available: https://www.bangkokpost.com/thailand/general/1333595/taxi-ok-taxi-vip-apps-to-launch-on-nov-9
[34] "Taxi OK." Accessed: Apr. 22, 2024. [Online]. Available: https://www.oriscom.com/category.php?i=2&l=1&id=1&name=TAXI%20OK,TAXI%20OK
[35] "Longdo Traffic Bangkok and Thailand Traffic Information Report." Accessed: Apr. 21, 2024. [Online]. Available: https://traffic.longdo.com/info
[36] "Traffic volunteers." Accessed: Apr. 30, 2024. [Online]. Available: https://traffic.longdo.com/volunteer
[37] "Top 100 busiest roads in Bangkok on average over time - Longdo Map Blog." Accessed: Apr. 23, 2024. [Online]. Available: https://map-blog.longdo.com/top-100-bangkok-traffic-jam/
[38] "The Average Speeds of Main Roads in Bangkok," Longdo Traffic, Metamedia Technology. Accessed: Jul. 17, 2023. [Online]. Available: https://traffic.longdo.com/bkk-speed/
[39] C. Li, W. Yue, G. Mao, and Z. Xu, "Congestion Propagation Based Bottleneck Identification in Urban Road Networks," *IEEE Trans Veh Technol*, vol. 69, no. 5, pp. 4827–4841, May 2020, doi: 10.1109/TVT.2020.2973404.
[40] W. Yue, C. Li, and G. Mao, "Urban Traffic Bottleneck Identification Based on Congestion Propagation," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6. doi: 10.1109/ICC.2018.8422108.
[41] H. Nguyen, W. Liu, and F. Chen, "Discovering Congestion Propagation Patterns in Spatio-Temporal Traffic Data," *IEEE Trans Big Data*, vol. 3, no. 2, pp. 169–180, 2017, doi: 10.1109/TBDATA.2016.2587669.
[42] Y. Wang, J. Cao, W. Li, and T. Gu, "Mining Traffic Congestion Correlation between Road Segments on GPS Trajectories," in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2016, pp. 1–8. doi: 10.1109/SMARTCOMP.2016.7501704.
[43] "Supervised vs. Unsupervised Learning: What's the Difference? - IBM Blog." Accessed: May 04, 2024. [Online]. Available: https://www.ibm.com/blog/supervised-vs-unsupervised-learning/?utm_medium=OSocial&utm_source=Youtube&utm_content=WAIWW&utm_id=YTDescription-101-Supervised-vs-Unsupervised-blog-supervised-vs-unsupervised-learning

[44] "What Is Unsupervised Learning? | IBM." Accessed: May 04, 2024. [Online]. Available: https://www.ibm.com/topics/unsupervised-learning
[45] Y. Wang and J. Ren, "Taxi Passenger Hot Spot Mining Based on a Refined K-Means++ Algorithm," *IEEE Access*, vol. 9, pp. 66587–66598, 2021, doi: 10.1109/ACCESS.2021.3075682.
[46] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 1, no. 3, pp. 231–240, May 2011, doi: 10.1002/widm.30.
[47] C. Malzer and M. Baum, "A Hybrid Approach To Hierarchical Density-based Cluster Selection," Nov. 2019, doi: 10.1109/MFI49285.2020.9235263.
[48] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015, doi: 10.1007/s40745-015-0040-1.
[49] A. Starczewski, P. Goetzen, and M. J. Er, "A New Method for Automatic Determining of the DBSCAN Parameters," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 10, pp. 209–221, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:219168291
[50] D. and S. J. Campello Ricardo J. G. B. and Moulavi, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining*, V. S. and C. L. and M. H. and X. G. Pei Jian and Tseng, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172.
[51] C. Malzer and M. Baum, "A Hybrid Approach To Hierarchical Density-based Cluster Selection," *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 223–228, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:207794266
[52] L. Basora, J. Morio, and C. Mailhot, "A Trajectory Clustering Framework to Analyse Air Traffic Flows," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:209527952
[53] S. J. Corrado, T. G. Puranik, O. J. Pinon, and D. N. Mavris, "Trajectory Clustering within the Terminal Airspace Utilizing a Weighted Distance Function," *Proc West Mark Ed Assoc Conf*, vol. 59, no. 1, 2020, doi: 10.3390/proceedings2020059007.
[54] R. L. Melvin, J. Xiao, R. C. Godwin, K. S. Berenhaut, and F. R. Salsbury, "Visualizing correlated motion with HDBSCAN clustering," *Protein Science*, vol. 27, no. 1, pp. 62–75, Jan. 2018, doi: 10.1002/pro.3268.
[55] "Geodesic versus planar distance—Portal for ArcGIS | Documentation for ArcGIS Enterprise." Accessed: May 13, 2024. [Online]. Available: https://enterprise.arcgis.com/en/portal/latest/use/geodesic-versus-planar-distance.htm
[56] "Distance on an ellipsoid: Vincenty's Formulae - Esri Community." Accessed: May 07, 2024. [Online]. Available: https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-an-ellipsoid-vincenty-s-formulae/ba-p/902053
[57] T. Vincenty, "Direct and Inverse solutions of geodesics on the ellipsoid with application of nested equations," *Directorate of Overseas Surveys*, vol. XXIII, no. 176, pp. 88–93, 1975, Accessed: May 07, 2024. [Online]. Available: https://www.ngs.noaa.gov/PUBS_LIB/inverse.pdf
[58] D. J. Suroso, P. Cherntanomwong, and P. Sooraksa, "Synthesis of a Small Fingerprint Database through a Deep Generative Model for Indoor Localisation," *Elektronika ir Elektrotechnika*, vol. 29, no. 1, pp. 69–75, 2023, doi: 10.5755/J02.EIE.31905.
[59] Y. Xie, Y. Cheng, A. Agrawal, and A. Choudhary, "Estimating online user location distribution without GPS location," in *IEEE International Conference on Data Mining Workshops, ICDMW*, IEEE Computer Society, Jan. 2015, pp. 936–943. doi: 10.1109/ICDMW.2014.30.
[60] J. R. Potts, M. Auger-Méthé, K. Mokross, and M. A. Lewis, "A generalized residual technique for analysing complex movement models using earth mover's distance," *Methods Ecol Evol*, vol. 5, no. 10, pp. 1012–1022, Oct. 2014, doi: 10.1111/2041-210X.12253.
[61] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Trans Pattern Anal Mach Intell*, vol. 29, no. 5, pp. 840–853, May 2007, doi: 10.1109/TPAMI.2007.1058.
[62] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.
[63] W. McKinney, "Data Structures for Statistical Computing in Python," 2010, pp. 56–61. doi: 10.25080/Majora-92bf1922-00a.

[64] "Folium — Folium 0.16.1.dev65+g5086929b documentation." Accessed: May 12, 2024. [Online]. Available: https://python-visualization.github.io/folium/latest/
[65] "The Statistics of Traffic in Bangkok in 2021," Longdo Traffic, Metamedia Technology. Accessed: Apr. 10, 2023. [Online]. Available: https://traffic.longdo.com/statistics2021
[66] "Index of /opendata/traffic-status-feeds/2021." Accessed: May 15, 2024. [Online]. Available: https://itic.longdo.com/opendata/traffic-status-feeds/2021/
[67] "GitHub - scikit-learn-contrib/hdbscan: A high performance implementation of HDBSCAN clustering." Accessed: May 13, 2024. [Online]. Available: https://github.com/scikit-learn-contrib/hdbscan
[68] "sklearn.metrics.pairwise.haversine_distances — scikit-learn 1.4.2 documentation." Accessed: May 07, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.haversine_distances.html
[69] L. McInnes and J. Healy, "Accelerated Hierarchical Density Based Clustering," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, pp. 33–42. doi: 10.1109/ICDMW.2017.12.
[70] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.
[71] "How HDBSCAN Works — hdbscan 0.8.1 documentation." Accessed: May 07, 2024. [Online]. Available: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html
[72] "Artificial intelligence to run traffic lights at all Bangkok intersections - Bangkok News - Thailand News, Travel & Forum - ASEAN NOW." Accessed: Jun. 25, 2024. [Online]. Available: https://aseannow.com/topic/1078510-artificial-intelligence-to-run-traffic-lights-at-all-bangkok-intersections/#google_vignette
[73] "KMeans — scikit-learn 1.5.0 documentation." Accessed: Jun. 25, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
[74] "silhouette_score — scikit-learn 1.5.0 documentation." Accessed: Jun. 26, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
[75] "davies_bouldin_score — scikit-learn 1.5.0 documentation." Accessed: Jun. 26, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html
[76] C. F. F. Karney, "Algorithms for geodesics," *J Geod*, vol. 87, no. 1, pp. 43–55, Jan. 2013, doi: 10.1007/s00190-012-0578-z.
[77] "Welcome to GeoPy's documentation! — GeoPy 2.4.1 documentation." Accessed: May 07, 2024. [Online]. Available: https://geopy.readthedocs.io/en/stable/
[78] M. Waskom, "seaborn: statistical data visualization," *J Open Source Softw*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.
[79] A. Peungnumsai, A. Witayangkurn, M. Nagai, and H. Miyazaki, "A Taxi Zoning Analysis Using Large-Scale Probe Data: A Case Study for Metropolitan Bangkok," *The Review of Socionetwork Strategies*, vol. 12, no. 1, pp. 21–45, 2018, doi: 10.1007/s12626-018-0019-4.

AUTHOR BIOGRAPHY

Name	Mr. Dio Tony
Date of Birth	January 8, 1994, in Medan
Address:	No. 13, Jalan Panglima Awang 35/126A, Alam Impian, 40470, Seksyen 35, Shah Alam, Selangor, Malaysia.

Educational Background:

2016:	Bachelor of Engineering (Hons) Electronic and Electrical Engineering (First Class Honors), King Mongkut's Institute of Technology Ladkrabang.
2024:	Master of Engineering in Robotics and Computational Intelligence Systems, King Mongkut's Institute of Technology Ladkrabang.

Work Experience and Research Achievements:

2023	Best Paper Award at iSAI-NLP-AIoT 2023 Conference.
2022 – 2024	Part time role and functions at KMITL
2017 - 2020	Operation Executive at PT. N.H.F. Auto Supplies

APPENDIX A

PUBLICATION

Assessing HDBSCAN Implementation for Traffic Congestion Pattern Estimation in Bangkok with Taxi GPS Probe

Dio Tony
Department of Robotics and AI,
School of Engineering,
King Mongkut's Institute of Technology Ladkrabang,
Bangkok, Thailand
63601042@kmitl.ac.th

Rathachai Chawuthai*
Department of Computer Engineering,
School of Engineering,
King Mongkut's Institute of Technology Ladkrabang,
Bangkok, Thailand
rathachai.ch@kmitl.ac.th

2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) | 979-8-3503-7121-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/iSAI-NLP60301.2023.10354829

***Abstract*—Traffic congestion is a major issue that is experienced globally in metropolitan cities. The phenomenon becomes more serious during peak hour as congestion increases and degrades the traffic networks. Each city possesses a unique traffic network, and the behaviour of its residents affects its traffic patterns. Therefore, a flexible congestion pattern identification approach is desirable. We proposed the employment of Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) to estimate traffic congestion propagation patterns through congestion length distribution. Global positioning System (GPS) probe of taxis were utilised to represent traffic pattern within Bangkok. The dataset was preprocessed into two successive timeframes, namely "later" timeframe and "prior" timeframe before being clustered. The identified congestion hotspots from both timeframes were transformed into a congestion area from which congestion lengths were extracted. Similarity measurements on congestion lengths distribution were conducted against Longdo Traffic's top 100 most congested roads list in Bangkok, showed encouraging results across all tests, with more than 90% similarity in one of the measurements, which indicated that HDBSCAN was feasible to make a key contribution to traffic management research.**

Keywords— traffic congestion pattern, taxi GPS probe, HDBSCAN, similarity measurement techniques

I. INTRODUCTION

Traffic congestion is one of the major barriers to sustainable urban development: it affects economy, society, and the general environment [1]. The period during which traffic congestion intensifies, typically resulting in slow to standstill traffic, is commonly known as "peak hour" and occurs twice daily on weekdays. In Bangkok, morning peak time starts from 6:00 to 9:00, while the evening peak hours starts from 16:30 to 19:30 [2]. In 2022, Bangkok's motorists wasted an additional 93 hours in peak hour congestion [3].

The root cause of traffic congestion in urban roads or expressways is the "traffic bottleneck". This phenomenon is characterised by spatial discontinuity which leads to a reduction of road capacity [4]. Traffic bottlenecks are difficult to define and identify, especially in urban traffic networks due to multiple complicating factors, including road network topology, congestion contingency, and travel behaviour [4].

Prior research that focused on congestion propagation identification have proposed method such as the construction of a causal tree accompanied based on graph theory [4] to identify traffic bottleneck. Nguyen et al. [5] employed slightly different approach by implementing Dynamic Bayesian Network to identify the congestion propagation patterns.

Numerous technologies have been utilised for traffic monitoring and data collection including Close Circuit Television, Loop Detector, Internet of Things, Vehicle-to-Vehicle, Vehicle-to-Infrastructure, and Global Positioning System (GPS), among others. Application of these technologies aims to enhance of transportation planning and traffic management. However widespread adaptation has been hindered due to their operational cost. On the other hand, the cost of GPS no longer a factor [6], resulting in widespread deployment, particularly in navigation assistant application. The prevalence of GPS usage has turned it into valuable data source for the planning and management of traffic and transportation.

Taxis are a preferred mode of public transport, offering flexible route selection for a trip. In 2017, taxi usage represented 17% of total public transport trips in Bangkok [7], second only to public buses. Taxis in Thailand fall under "category 6" within the vehicle category criteria set by Department of Land Transport with dimensions of 2.5 meters by a length less than 6 meters and capacity of 7 passengers [8], which is equivalent to private passenger car in "category 1".

Registered private passenger cars made up 23.59% of total registered vehicle within Thailand in 2017, positioning them as the second largest category after motorcycle (55.31%) [9]. Therefore, investigating traffic congestion from the perspective of passenger car is crucial. This task can be achieved through the utilisation of taxi GPS probe data, as demonstrated by Wang et al. [10] in determining correlation of traffic congestion between road segments. Additionally, Buapang et al. [11] used taxi GPS data with a spectral graph neural network to perform traffic prediction.

Clustering algorithms groups data into partitions based on data similarity, facilitating data exploration and analysis of patterns within data. In density-based clustering, a cluster is developed by searching within the spaces of given input for high density data region which is separated by less dense region [12]. Such algorithms are particularly suitable for spatial data [13], such as GPS coordinates in transportation [12]. Region with dense data also referred as hotspot, is the outcome of clustering algorithm. Clustering algorithms such as K-Means, Density-based Spatial Clustering of Applications with Noise (DBSCAN), and Ordering Points To Identify Cluster Structure (OPTICS), has been proposed previously for hotpot mining application. Their application and limitations have been discussed by Wang et al. [14] which are sensitive to outliers, and make a flat cut based on similarity distance threshold, among others.

* Corresponding author

Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) [15] [16] is a density-based clustering algorithm with a hierarchical base, which selects clusters from multiple levels within the final dendrogram rather than making a flat cut. HDBSCAN clusters by using a cluster stability metric and "mutual reachability distance" [15] [16] to determine the meaning of "high density" for a given dataset [17].

The primary focus of this project was to assess feasibility of HDBSCAN in estimating traffic congestion patterns in Bangkok based on Taxi GPS probe data. The application of HDBSCAN algorithms to cluster taxi GPS data into congestion hotspots explores the feasibility of this approach, with potential benefits extending to various applications, including traffic prediction, pattern generation, among others. The study centred on analysing data distribution of traffic congestion length within the identified hotspots.

This paper is organised as follows. Section I introduces the problem and discusses previous works. Section II details the methods used in the experiment including data preparation and algorithm setting. Section III shows and discusses some outcomes of the proposed method. Section IV concludes with a summary and future works.

II. Methodology

Obtaining length of congestion was the main objective of this study. It was extracted from the distance between two congestion hotspots within a congested area. There were three main steps needed to be performed to produce congested areas namely: input loading, data preprocessing, and finally, clustering and postprocessing. The overall approach to get the desired output is illustrated in Fig. 1, and all the processes were executed with Google Colaboratory in a Python environment.

The experiment began with loading taxi GPS transaction as input data. The input data was open data, licensed under Creative Commons Attribution 4.0, provided by Intelligent Traffic Information Center Foundation (iTIC) Thailand [18]. The input data had nine attributes and their desired values within this study are given in Table I.

Visualised input data indicates that the taxi installed with a GPS Probe unit was operating mainly in Bangkok Metropolitan Region (BMR) but not limited to it, such as shown in Fig. 2.

Traffic conditions need to be defined to determine the valid data, and ground truth need to be established. Our reference for traffic condition category is taken from Longdo Traffic which classifies the traffic condition in urban road within Thailand into four distinct colours. We proposed our own definition for the traffic categories which is presented within Table II.

TABLE I.
TAXI GPS PROBE FILE DESCRIPTION AND ITS DESIRED VALUE.

Attribute	Description	Data Type	Desired Value
VehicleID	Unique vehicle ID	string	All VehicleID
gpsvalid	enough satellite for GPS fix	float	1
lat	GPS location up to 5 decimal places	float	Within Bangkok boundary
lon	GPS location up to 5 decimal places	float	Within Bangkok boundary
timestamp	GPS timestamp (GMT+7)	string	Within time of interest
speed	km/h	integer	< 25/h
heading	vehicle heading direction (0-360) ° from North=0	integer	-
for_hire_llight	1 = light on - possibly no passenger, 0 = light off - possibly carrying passengers	integer	0
engine_acc	1 (active, the data will be collected every minute), 0 (inactive, data collected every 3 minutes)	integer	1

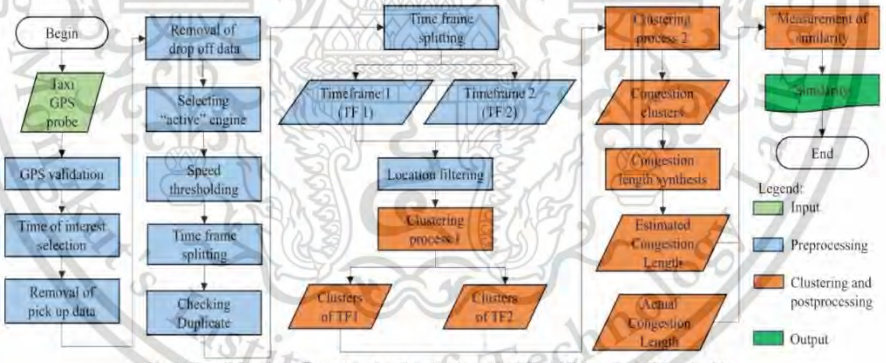


Fig. 1. Overall flowchart of congestion length estimation with HDBSCAN from Taxi GPS probe data.

Fig. 2. Heat map of taxi GPS data.

TABLE II. LONGDO TRAFFIC'S SPEED CODE.

Average speed (km/h)	Colour Code	Our definition
< 10 km	Black	Congested
< 15 km	Red	Congested
< 25 km	Yellow	Congested
≥ 25 km	Green	Free Flow

Finally, reference for actual congestion length was extracted from Longdo Traffic's top 100 most congested roads in Bangkok in [19]. The reference was used to gauge similarities against HDBSCAN's generated congestion length patterns. The following subsections discuss the other required processes to acquire the desired values in detail.

A. Data Preprocessing

Preprocessing step cleaned the input data for usage in the following clustering process, by filtering unwanted data (noise), resulting in better clustering performance, and more accurate traffic congestion detection. Addressing false congestion condition was a major task within data preprocessing of Taxi GPS. Events such as passenger pickup and drop-off might introduced probe entry with "speed" values lower than free flow speed. In this section, we describe the criteria of filtering condition to distinguish between valid and invalid data entries. The dataset used within this study was collected from 2nd November 2021 (Tuesday) as it was declared as the most congested weekday in 2021 for Bangkok by Longo Traffic website [20].

Initially, poor GPS signals were removed to improve congestion hotspot detection. Poor GPS events were recoded with a "0" in the "gpsvalid" field. Next, time of interest was chosen. In Thailand, schools and government offices hours generally begin at 8:00 to 8:30.

Congestion propagation could be identified when two different time frames existed. Morning rush hour was chosen with 7:30 as the time of interest, because of a high possibility of congestion hotspot formation due to the operation of government facilities.

Stoppage or stationary was a common behaviour of taxi operators during picking up and dropping off passengers which resulted in a slow speed being recorded within this period, possibly leading to false congestion conditions. Thus, its removal was important [6]. However, we tried to use the pickup and drop-off data by making some assumptions.

Passenger pickup events were defined as situation where an occupied taxi has established a destination and started the taximeter to signify the beginning of a passenger's trip. This event was indicated by changes from "1" to "0" in "for_hire_light" attribute. Generally, taximeter started when a customer was being seated within the taxi, even without the driver knowing the destination. Regardless, taxi drivers will most likely remain stationary when the destination has not yet been acquired, leading to our assumption that delay was presence within the system since some pickup event entries within the dataset were found to have nonzero "speed" attribute, signifying that the taxi had started to move when the taximeter was started.

Pickup events were assumed to be one minute long in duration, which is equivalent to one GPS probe entry for a particular taxi ID. We assumed that GPS probe unit started its recoding at the instant of the passenger pickup event. The filtering process started by grouping data based on "VehicleID" and monitoring only those "VehicleID" that had both "0" and "1" within its "for_hire_light" attribute. If the "speed" value fell below the free flow speed as the taxi announced its hiring, either of the following approaches were used to prevent and remove false congestion condition data entry:

1) *Pickup event recorded on last entry:* If the "0" (pickup) was recorded at the end of the probe entry for a particular "VehicleID", the "speed" was assessed from the "speed" before it is hired, assuming similar traffic conditions, or,

2) *Non-last entry pickup event:* If the "for_hire_light" field recorded another "0" entry after a pickup event, then the "speed" was assessed from the "speed" of the succeeding entry, with assumption that traffic conditions towards the destination reflects traffic conditions better.

Drop-off event is associated with reduction of vehicle speed (or even stationary) upon arrival and payment activity. Drop-off is assumed to be 1 minute event. Drop-off is signified by the changes from "0" to "1" within "for_hire_light" field which might result in false congestion condition. We monitored and removed the drop-off entry only if the taxi speed falls under free flow speed.

Speed filtering afterward removed all data entries that were not reflecting congested conditions within the traffic network with threshold value of 25 km/h. Next, the congestion time of interest were split into two distinct time frames with intervals of 10 minutes, resulting in a "prior" time frame, TF1 (7:10:00 to 7:19:59) and a "later" time frame, TF2 (7:20:00 to 7:29:59).

Furthermore, duplicate data entry checking was conducted to anticipate the occurrence of error from the GPS probe unit. We grouped all data entries based on "lat", "lon", "timestamp", and "VehicleID" attributes. Subsequently, the uniqueness of each data entry was verified by the count of entries obtained against the dimension of the analysed data frame. Finally, boundaries in Table III were used to extract data within Bangkok city only.

TABLE III. BOUNDARIES OF STUDY AREA.

Location	Coordinate
Top left	14.03116, 10.23174
Top right	14.03116, 100.91143
Lower left	13.545157, 100.23174
Lower right	13.545157, 100.91143

B. HDBSCAN for identification of congestion hotspot

We implemented HDBSCAN algorithm in Google Colaboratory by employing the library provided by Scikit-learn [21]. Initially, a distance matrix was developed to serve as an input which stored the distance value for all pairs of GPS data. Distance between a pair of GPS data, $D(x,y)$ was calculated by using Haversine formula given in (1).

$$D(x,y) = 2\arcsin\left[\sqrt{\sin^2\left(\frac{(x_{lat}-y_{lat})}{2}\right) + \cos(x_{lat})\cos(y_{lat})\sin^2\left(\frac{(x_{lon}-y_{lon})}{2}\right)}\right] \quad (1)$$

HDBSCAN [22] required one basic parameter: "min_cluster_size" to establish "noise points" (any cluster smaller than the specified minimum_cluster_size) identification. Additionally, we passed an optional argument "metric=precomputed" so that the HDBSCAN recognised the state of the provided input which had been processed and did not require further similarity calculation. Moreover, we passed another optional argument "cluster_selection_method=leaf". Subsequently, the first clustering process was started with "min_cluster_size=3".

Any similarities identified during clustering produced clusters. The initial clustering process was conducted on both TF1 and TF2 data, resulted in congestion hotspots within both timeframes that were termed as "congestion hotspots" from here on. By using congestion hotspots from both outputs as input, another distance matrix for a second clustering was developed. This second clustering process was started with "min_cluster_size=2" and resulted in clusters termed as "congestion area" from here on.

A congested area consisted of congestion hotspots that were clustered by distance similarity. Congestion length was obtained through pairwise distance calculation between all congestion hotspots. Initially, congestion hotspots within a particular congestion area were grouped based on timeframe and sorted chronologically, then pairwise distance was calculated using the Haversine formula. Finally, we implemented three different statistical techniques to extract the distance from each congested area, namely average distance (mean), maximum distance (max), and minimum distance (min).

Similarity measurements evaluated the success of our approach. Similarity observation by using statistical mean and standard deviation was implemented as used by Suroso et al. [23] as a similarity gauge on received signal strength indicator (RSSI) between a GAN synthesised dataset against an experiment-based dataset.

Moreover, Jensen-Shannon Divergence (JSD) similarity measurement was utilised to measure the similarity of synthesised and ground truth since it was implemented on the similarity measurement for biased and adjusted data of geodemographic distribution by Xie et al. [24]. JSD is symmetrical derivation of Kullback–Leibler (KL) Divergence shown in (2). Let P represents the probability for ground truth and Q for synthesised data, and M for the average of the compared distributions, $M = (P + Q)/2$. JSD between the two probability is given by (3).

$$KL(P||Q) = \sum_{x \in X} P(x)\log_2\left(\frac{P(x)}{Q(x)}\right) \quad (2)$$

$$JSD(P||Q) = \frac{1}{2}\left(\sum_{x \in X} P(x)\log_2\left(\frac{P(x)}{M(x)}\right) + \sum_{x \in X} P(x)\log_2\left(\frac{Q(x)}{M(x)}\right)\right) \quad (3)$$

Additionally, Earth Mover's Distance (EMD) similarity measurement was used for similarity measurement due to its implementation in measuring similarity in movement models for animals by Potts et al. [25]. In our simple 1D data, let P be our ground truth, $P = \{(p_i, u_i)_{i=1}^{m}\}$ and Q be our synthesised data, $Q = \{(q_j, v_j)_{j=1}^{n}\}$ with size m, n, respectively, the EMD between them can be modelled as a solution to the transportation problem [26] in (4).

$$EMD(P, Q) = min \sum_{F=\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \quad (4)$$

In the transportation problem, elements in P can be treated as "supplies" that are located at u_i and elements in Q as "demands" located at v_j, with the amount (weight) of supply and demand indicated by p_i and q_j, respectively. Then EMD can be defined as the minimum (normalised) work required for resolving the supply-demand transport [26] with distance between position u_i and position v_j, denoted by d_{ij} and the mass of transported goods from ith supply to the jth denoted by f_{ij}. We calculated EMD similarity using the SciPy library [27].

III. RESULT AND DISCUSSION

Implementation of our approach and its outcome, along with our findings are discussed in this section. Raw input data was preprocessed to eliminate the false congestion condition. Afterwards, clustering produced clusters and noise points. A summary of preprocessing and clustering outcomes is recorded in Table IV and Table V, respectively.

TABLE IV. PREPROCESSING OUTCOMES ON TIME OF INTEREST (7:30).

Measurement	Data Frame			
	Raw	*Filtered*	*TF1*	*TF2*
Data count	28986	5036	2403	2633
Vehicle ID count	3105	896	672	724
Within BKK	-	4219	1988	2231

TABLE V. CLUSTERING PROCESS OUTCOMES.

Measurement	Clustering Process		
	Prior hotspots (data points)	*Later hotspots (data points)*	*Congestion Area (cluster)*
Input count	1988	2231	542
Outlier count	332	350	148
Cluster count	256	286	84

Fig. 3 shows a sample of identified congestion areas and its associated hotspots by HDBSCAN. The HDBSCAN clustering algorithm generated the grouping based on minimum cluster size threshold ("min_cluster_size=2"), which retained a cluster if it has at least member of two points.

Fig 3. Congestion area and its associated hotspots.

Initially, the clusters started as a single large clump, then the edges between hotspots were disconnected from largest to shortest distance by thresholding the edge weight, resulting in more clusters with fewer points. Child clusters that had fewer points than the minimum cluster threshold were declared as noise (points falling out of clusters). The distribution of congestion length data that was extracted from HDBSCAN's congestion area and Longdo Traffic's top 100 most congested roads in Bangkok length is illustrated in Fig. 4.

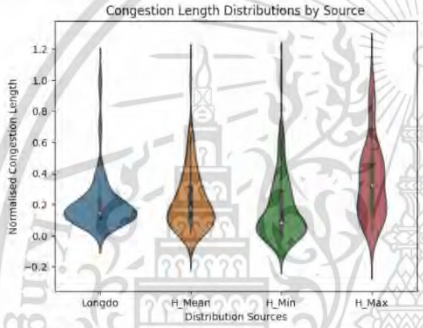


Fig. 4. Congestion length distribution based on sources.

Visually, congestion length distributions from Longdo Traffic, H_mean, and H_min had a similar spread, while H_Max showed the least similarity. H_Max extracted distance of congestion length between two furthest point in a particular cluster, resulting in an almost binomial distribution-like (high uniformity) pattern on its congestion length count. Meanwhile, referring to Longdo's distribution, the count of smaller congestion length dominated at least 60% of the data. Also, H_Min displayed higher visual similarity than H_Mean against Longdo.

The similarity measurements between Longdo Traffic's top 100 congested roads length and our HDBSCAN synthesized congestion lengths are summarized in Table VI. The H_Mean approach had the lowest dissimilarity scores by mean, standard deviation, JSD, and ESD measurements, indicated that H_Mean provided better estimation on the distribution of congestion length against other distance extraction approaches.

TABLE VI.
RESULT OF SIMILARITY MEASURES OF CONGESTION LENGTH DISTRIBUTION.

SIMILARITY MEASUREMENT	Source of Congestion Length			
	Longdo Traffic (GT)	*HDBSCAN Mean (H_Mean)*	*HDSCAN Min (H_Min)*	*HDBSCAN Max (H_Max)*
Mean	0.204551	0.222811	0.173721	0.354097
Standard Deviation (SD)	0.166766	0.183817	0.197197	0.242535
Difference in mean against GT	0	0.018260	0.030830	0.149546
Difference in SD against GT	0	0.017051	0.030431	0.075769
JSD	0	0.019335	0.061386	0.022062
EMD	0	0.057655	0.073289	0.158685

Statistical similarity measurement simply indicated that H_Mean approach had the smallest deviation on the distribution of data, about 8.93% and 10.22% for mean and standard deviation from the GT, respectively. The JSD measured the sum of information lost when probability distribution M was used to approximate distribution P and distribution Q. JSD measurement was bounded from zero to one, in which all HDBSCAN achieved less than 7% loss across all HDBSCAN variants in this study. Interestingly, H_Max scored a lower JSD value than H_Min as the result of an almost uniform event distribution in H_Max's distribution, when it was compared against H_Min's.

The EMD measured the amount of work needed to transform distribution P into distribution Q, to quantify the similarity between distributions. EMD score was not bounded, ranging from zero to infinity. The result in Table VI indicated that H_Mean had the smallest difference against GT. Interestingly, H_Min had better similarity score than H_Max against GT. EMD reflected the difference between distribution better as it considers the distance between event in distributions, magnifying the existence of outlier and error between the compared distributions.

Max scaling was used to post-process the congestion length before applying similarity measurement since the magnitude extracted was smaller than GT's. The difference may be contributed by the amount of (input) data available. Different distance measurement techniques for distance matrix could be applied, such as Manhattan distances since they capture the shape of road better than simple subtraction and Euclidian distances. Additionally, effective filtering was needed to ensure that a minimum number of false congestion conditions were included in the input. While the quantity, type, and sources of data used by Longdo Traffic in producing their top 100 congested roads list was not declared, Siangsuebchart et al. [7] mentioned that current contributors to the taxi GPS probe data represented approximately 4 % of all registered taxis in Bangkok, which confirmed the significance of taxi and its equivalent vehicle category in portraying the traffic congestion pattern.

The contribution of this study, in addition to the estimation of congestion length, is proposing an approach for the enhancement of taxi GPS data to better reflect traffic conditions by eliminating false congestion conditions within the dataset. Effort to minimize over filtering concern mentioned by Kan et al. [6] were made by introducing some heuristic techniques, increasing the amount of data available and thus improving its reliability.

IV. CONCLUSIONS AND RECOMMENDATIONS

Traffic congestion is a significant issue globally that affects the advance of social and economic growth negatively. We assessed HDBSCAN ability to generate traffic congestion patterns from taxi GPS probe data in Bangkok and measured its performance. This study used HDBSCAN and taxi GPS data to simulate traffic congestion patterns through clustering. Three different distance extraction methods were applied to extract congestion length from identified congestion areas. The performance of these distance extraction methods was evaluated by using similarity measurement techniques: statistical mean, standard deviation, JSD, and EMD. The mean-based distance extraction method, H_Mean is the best performing approach, scoring 8.93% and 10.22% on the mean and standard deviation measures, respectively. Moreover, H_Mean achieved score of 0.057655 on EMD measurement and displayed 0.019335 of divergence in JSD measurement, translating to 98.07% similarity to the ground truth data.

The extension to this study will involve the use of additional data to enhance congestion detection accuracy, and the "heading" field within the data to provide traffic congestion estimation at the lane level. We believe this extension will offer valuable insights for traffic management in Bangkok and worldwide.

REFERENCES

[1] M. A. Fattah, S. R. Morshed, and A. Al Kafy, "Insights into the socio-economic impacts of traffic congestion in the port and industrial areas of Chittagong city, Bangladesh," Transportation Engineering, vol. 9, Sep. 2022, doi: 10.1016/j.treng.2022.100122.

[2] "The average speed of personal cars in Bangkok (Morning peak 06:00 - 09:00)," Department of Land Traffic System Development. Accessed: Feb. 28, 2023. [Online]. Available: https://otp.gdcatalog.go.th/dataset/dataset_12_01/resource/ca5fbc90-70fa-48ed-a84e-fe0c4012b138

[3] "Traffic Index ranking," TomTom International BV. Accessed: Jul. 22, 2023. [Online]. Available: https://www.tomtom.com/traffic-index/ranking/

[4] W. Yue, C. Li, and G. Mao, "Urban Traffic Bottleneck Identification Based on Congestion Propagation," in 2018 IEEE International Conference on Communications (ICC), 2018, pp. 1–6. doi: 10.1109/ICC.2018.8422108.

[5] H. Nguyen, W. Liu, and F. Chen, "Discovering Congestion Propagation Patterns in Spatio-Temporal Traffic Data," IEEE Trans Big Data, vol. 3, no. 2, pp. 169–180, 2017, doi: 10.1109/TBDATA.2016.2587669.

[6] Z. Kan, L. Tang, M. P. Kwan, C. Ren, D. Liu, and Q. Li, "Traffic congestion analysis at the turn level using Taxis' GPS trajectory data," Comput Environ Urban Syst, vol. 74, pp. 229–243, Mar. 2019, doi: 10.1016/j.compenvurbsys.2018.11.007.

[7] S. Siangsuebchart, S. Ninsawat, A. Witayangkurn, and S. Pravinvongvuth, "Public transport gps probe and rail gate data for assessing the pattern of human mobility in the bangkok metropolitan region, thailand," Sustainability (Switzerland), vol. 13, no. 4, pp. 1–28, Feb. 2021, doi: 10.3390/su13042178.

[8] "Definition of Cars," Transportation Statistics Group, Planning Division, Department of Land Transport. Accessed: Jul. 02, 2023. [Online]. Available: https://web.dlt.go.th/statistics/index.php

[9] "TRANSPORT STATISTICS REPORT IN 2018," Thailand, Mar. 2019. Accessed: Jul. 03, 2023. [Online]. Available: https://web.dlt.go.th/statistics/load_file_select.php?tmp=7647.1500001202085&data_file=1188

[10] Y. Wang, J. Cao, W. Li, and T. Gu, "Mining Traffic Congestion Correlation between Road Segments on GPS Trajectories," in 2016 IEEE International Conference on Smart Computing (SMARTCOMP), 2016, pp. 1–8. doi: 10.1109/SMARTCOMP.2016.7501704.

[11] S. Buapang and V. Muangsin, "Traffic Prediction With a Spectral Graph Neural Network," in 2022 7th International Conference on Business and Industrial Research (ICBIR), 2022, pp. 341–346. doi: 10.1109/ICBIR54589.2022.9786482.

[12] C. Malzer and M. Baum, "A Hybrid Approach To Hierarchical Density-based Cluster Selection," Nov. 2019, doi: 10.1109/MFI49285.2020.9235263.

[13] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," Wiley Interdiscip Rev Data Min Knowl Discov, vol. 1, no. 3, pp. 231–240, May 2011, doi: 10.1002/widm.30.

[14] Y. Wang and J. Ren, "Taxi Passenger Hot Spot Mining Based on a Refined K-Means++ Algorithm," IEEE Access, vol. 9, pp. 66587–66598, 2021, doi: 10.1109/ACCESS.2021.3075682.

[15] R. Campello, D. Moulavi, and J. Sander, Density-Based Clustering Based on Hierarchical Density Estimates, vol. 7819. 2013. doi: 10.1007/978-3-642-37456-2_14.

[16] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection," ACM Trans. Knowl. Discov. Data, vol. 10, no. 1, Jul. 2015, doi: 10.1145/2733381.

[17] R. L. Melvin, J. Xiao, R. C. Godwin, K. S. Berenhaut, and F. R. Salsbury, "Visualizing correlated motion with HDBSCAN clustering," Protein Science, vol. 27, no. 1, pp. 62–75, Jan. 2018, doi: 10.1002/pro.3268.

[18] "Index of/opendata/probe-data," Intelligent Traffic Information Center Foundation. Accessed: Apr. 03, 2023. [Online]. Available: https://itic.longdo.com/opendata/probe-data/

[19] "The Average Speeds of Main Roads in Bangkok," Longdo Traffic, Metamedia Technology. Accessed: Jul. 17, 2023. [Online]. Available: https://traffic.longdo.com/bkk-speed/

[20] "The Statistics of Traffic in Bangkok in 2021," Longdo Traffic, Metamedia Technology. Accessed: Apr. 10, 2023. [Online]. Available: https://traffic.longdo.com/statistics2021

[21] F. Pedregosa FABIANPEDREGOSA et al., "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011, [Online]. Available: http://scikit-learn.sourceforge.net.

[22] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," The Journal of Open Source Software, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.

[23] D. J. Suroso, P. Cherntanomwong, and P. Sooraksa, "Synthesis of a Small Fingerprint Database through a Deep Generative Model for Indoor Localisation," Elektronika ir Elektrotechnika, vol. 29, no. 1, pp. 69–75, 2023, doi: 10.5755/J02.EIE.31905.

[24] Y. Xie, Y. Cheng, A. Agrawal, and A. Choudhary, "Estimating online user location distribution without GPS location," in IEEE International Conference on Data Mining Workshops, ICDMW, IEEE Computer Society, Jan. 2015, pp. 936–943. doi: 10.1109/ICDMW.2014.30.

[25] J. R. Potts, M. Auger-Méthé, K. Mokross, and M. A. Lewis, "A generalized residual technique for analysing complex movement models using earth mover's distance," Methods Ecol Evol, vol. 5, no. 10, pp. 1012–1022, Oct. 2014, doi: 10.1111/2041-210X.12253.

[26] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," IEEE Trans Pattern Anal Mach Intell, vol. 29, no. 5, pp. 840–853, May 2007, doi: 10.1109/TPAMI.2007.1058.

[27] P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," Nat Methods, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

APPENDIX B

CONFERENCE'S AWARD AND PROOF OF PARTICIPATION

iSAI-NLP-AIoT 2023

The 18th International Conference on Artificial Intelligence and Natural Language Processing
and The International Conference on Artificial Intelligence and Internet of Things
27 - 29 November 2023 | The Sukosol Hotel, Bangkok, Thailand

The Best Paper Award

The Session of Data Analytics and Machine Learning

Presented to

Dio Tony and Rathachai Chawuthai

for the paper

"Assessing HDBSCAN Implementation for Traffic Congestion
Pattern Estimation in Bangkok with Taxi GPS Probe"

(Assoc. Prof. Dr. Sommai Piysa-Art)
President, RMUTT

(Dr.Thepchai Supnithi)
Vice President, Artificial Intelligence Association of Thailand

iSAI-NLP-AIoT 2023

The 18th International Conference on Artificial Intelligence and Natural Language Processing
and The International Conference on Artificial Intelligence and Internet of Things
27 - 29 November 2023 | The Sukosol Hotel, Bangkok, Thailand

Present this honor certificate to

Dio Tony and Rathachai Chawuthai

for your kind participation as an oral presenter on titled

"Assessing HDBSCAN Implementation for Traffic Congestion Pattern
Estimation in Bangkok with Taxi GPS" Probe"

(Assoc. Prof. Dr. Sommai Piysa-Art)
President, RMUTT

(Dr.Thepchai Supnithi)
Vice President, Artificial Intelligence Association of Thailand

APPENDIX C

FHWA (U.S.) VEHICLE CLASSIFICATION CODE

Table C. FHWA vehicle classification and its definitions [21]

Class Group	Class Definition	Class Includes
1	Motorcycles	Motorcycles
2	Passenger Cars	All cars, cars with one-axle trailers, cars with two-axle trailers
3	Other Two-Axle Four-Tire Single-Unit Vehicles	Pick-ups and vans, pick-ups and vans with one- and two- axle trailers
4	Buses	Two- and three-axle buses
5	Two-Axle, Six-Tire, Single-Unit Trucks	Two-axle trucks
6	Three-Axle Single-Unit Trucks	Three-axle trucks, three-axle tractors without trailers
7	Four or More Axle Single-Unit Trucks	Four-, five-, six- and seven-axle single-unit trucks
8	Four or Fewer Axle Single-Trailer Trucks	Two-axle trucks pulling one- and two-axle trailers, two-axle tractors pulling one- and two-axle trailers, three-axle tractors pulling one-axle trailers
9	Five-Axle Single-Trailer Trucks	Two-axle tractors pulling three-axle trailers, three-axle tractors pulling two-axle trailers, three-axle trucks pulling two-axle trailers

(continued) Table C. FHWA vehicle classification and its definitions [21]

10	Six or More Axle Single-Trailer Trucks	Multiple configurations
11	Five or Fewer Axle Multi-Trailer Trucks	Multiple configurations
12	Six-Axle Multi-Trailer Trucks	Multiple configurations
13	Seven or More Axle Multi-Trailer Trucks	Multiple configurations
14	Unused	----
15	Unclassified Vehicle	Multiple configurations

APPENDIX D

DLT (THAILAND) VECHICLE CLASSIFICATION CODE

Table D. DLT vehicle classification and its definition [23]

Class Group	Class Definition	Class Includes
1	Personal sedan car (at most 7 passengers)	Car with at most 2.5 m wide and 12 m long
2	Personal microbus & passenger van (more than 7 passengers)	Car with at most 2.5 m wide and 12 m long, and the length of the body measured from the centre of rear axle to the back of the car must not exceed 2/3 of the length measured from the centre of front axle to the centre of rear axle
3	Personal van & pick up	Car used for non personal transportation according to the law of DLT with at most 2.5 m wide and 12 m long, and the length of the body measured from the centre of rear axle to the back of the car must not exceed 3/5 of the length measured from the centre of front axle to the centre of rear axle
4	Personal three-wheeled vehicle	Car with at most 1.5 m wide and 4 m long, with at most 550 cc of combined cylinder capacity of the engine

(continued) Table D. DLT vehicle classification and its definition [23]

5	Interprovincial taxi	Two-door sedan with no less than four doors, with weight at least 1 ton. The width at most 2.5 m and the length at most 6 m. At least 1500 cc of the engine combined cylinder capacity
6	Urban taxi (at most 7 passengers)	Two-door sedan. The width at most 2.5 m and the length at most 6 m. at least 4 doors with no central lock system installed. The windshield must be clear, transparent, and free of any modifications, except for signs or documents as prescribed by law or materials for blocking or filtering sunlight at the front windshield as specified by the Department of Land Transport. The engine must have At least 1000 cc of combined cylinder capacity. For taxis (TAXI – METER) registered since 17 April 1992 onwards (except for taxis that the owner registers in place of taxis registered before 17 April 1992), it must be a two-door sedan or a two-door sedan with a cargo area inside the vehicle (two-door van) that is factory-made. The vehicle width at most 2.5 m, the length at most 6 m, it should at least four doors, and the engine should have a total cylinder capacity of at least 1,500 cc
7	Small four-wheeled taxi	A two-part car with at least two doors. The width is at most 1.5 m, the length is at 4 m, and the engine capacity is at most 800 cc

(continued) Table D. DLT vehicle classification and its definition [23]

8	Three-wheeled taxi (Tuk Tuk)	Compact vehicle, with 2 rows or 2 sections of seats. The width at most 1.5 m and the length at most 4 m, with engine total capacity of at most 550 cc.
9	Hotel taxi	Passenger or rental car. Same as class 5 in term technical specification. This is a vehicle used to transport passengers between airports, seaports, transport stations, or train stations and hotels, residences, passenger offices, or the offices of the business service provider.
10	Tour taxi	Passenger or rental car. Same as class 5 in term technical specification.
11	Car for hire	Passenger or rental car. Same as class 5 in term technical specification.
12	Personal motorcycle	Vehicle that runs on engine or electric power and has at most two wheels. At most one additional wheel, it shall include a when side trailer is installed. The motorcycle has at most 1.1 m of width and the length at most 2.5 m. The side trailer has at most 1.1 m of width and the length at most 1.75 m. The distance between the rear wheel of motorcycle and the trailer's wheel is at most 1.5 m

(continued) Table D. DLT vehicle classification and its definition [23]

13	Tractor	Vehicle that has wheels or belts and an engine for movement. Used for digging, scooping, pushing, or pulling, etc., or a vehicle for towing which is not used for personal transportation under the law on land transport must have a width of at most 4.40 m and a length of at most 16.20 m.
14	Road roller	Vehicle used for compacting materials on the ground and has a self-propelled engine or uses another vehicle to tow. Must have at most 3.50 m in width and 8 m in length.
15	Farm Vehicle	Vehicle for agricultural use, using an engine that is not specifically used for vehicles. It must be a vehicle with three or four wheels, at most: 1.6 ton in weight, 2 m in width, 6 m in length, and 1200 cc in total engine cylinder capacity
16	Automobile trailer	Vehicle that moves by being towed. At most: 2.5 m in width and 12 m in length

(continued) Table D. DLT vehicle classification and its definition [23]

17	Public (taxi) motorcycle	Motorcycle used for carrying passengers (for hire). At most: 1.1 m in width, 2.5 m in length, 2 m in height, and 125 cc of engine cylinder capacity. If it is powered by an electric motor power rating between 250 W and 4 kW, and able to drive at least 45 km/h continuously for at least 30 minutes. Does not include motorcycles with side trailers and bicycles with engines.
18	Ride-hailing car	Same as class 1 in term of technical specification, however usage age does not exceed 9 years from the date of first registration. Registered class is changed to taxi type carrying at most seven passengers through an electronic system

APPENDIX E

PERFORMANCE OF ALL CONGESTION EXTRACTION APPROACH ON ALL DATASETS

Table E1. Summary of output for peak hour at time 6:30

<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>
Table E1-1. Preprocessing Outcomes
<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Data Frame</th></tr>
<tr><th>Raw</th><th>Filtered</th><th>TF1</th><th>TF2</th></tr>
<tr><td>Data count</td><td>26617</td><td>3405</td><td>1622</td><td>1783</td></tr>
<tr><td>Vehicle ID count</td><td>3089</td><td>675</td><td>492</td><td>531</td></tr>
</table>
</td></tr>
</table>

(continued) Table E1. Summary of output for peak hour at time 6:30

Clustering and post processing	**Table E1-2. Clustering Outcomes**

Measurement	Clustering Process			
	Prior hotspots (data points)	Later hotspots (data points)	Congestion Area (cluster)	Valid Congestion Area (cluster)
Input count	1622	1783	360	71
Outlier count	210	219	90	13
Cluster count	171	189	71	58

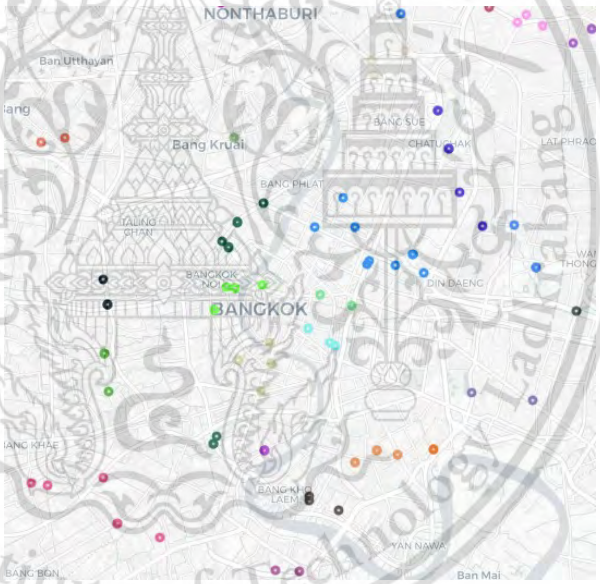


Figure E1-1. Identified Congestion Area in Bangkok with Its Associated Hotspots

(continued) Table E1. Summary of output for peak hour at time 6:30

Congestion length distribution similarity

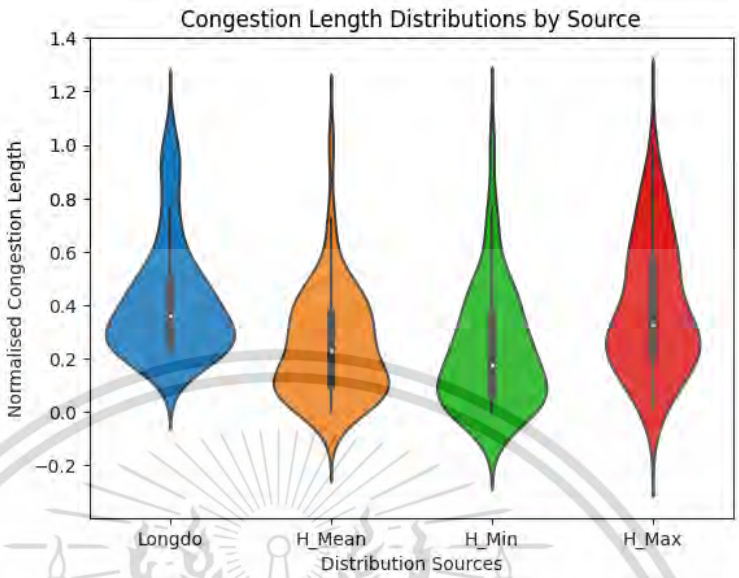


Figure E1-2. Congestion Length Distribution Based on Sources

Table E1-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.41665	0.24923	0.23092	0.3895
Standard Deviation (SD)	0.21158	0.19718	0.21798	0.24015
Difference in mean against GT	0	0.16742	0.18573	0.02715
Difference in SD against GT	0	0.0144	0.0064	0.02857
JSD	0	0.0303	0.0629	0.01496
EMD	0	0.16742	0.18573	0.05905

Table E2. Summary of output for peak hour at time 7:30

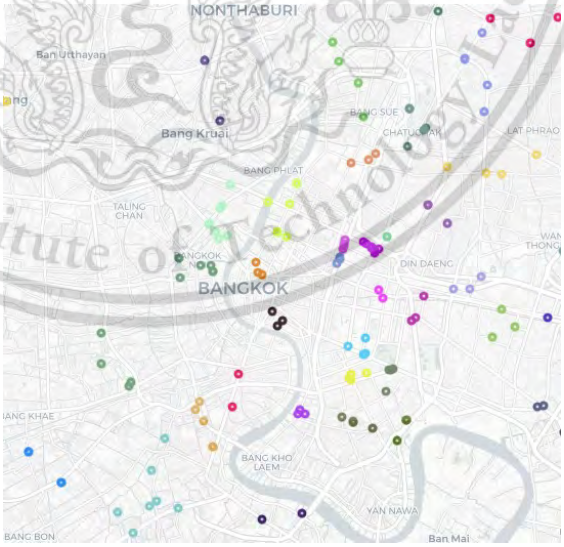
<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>
Table E2-1. Preprocessing Outcomes
<table>
<tr><td rowspan="2">Measurement</td><td colspan="4">Data Frame</td></tr>
<tr><td>Raw</td><td>Filtered</td><td>TF1</td><td>TF2</td></tr>
<tr><td>Data count</td><td>28983</td><td>5032</td><td>2397</td><td>2635</td></tr>
<tr><td>Vehicle ID count</td><td>3105</td><td>896</td><td>672</td><td>724</td></tr>
</table>
.
</td></tr>
<tr><td>Clustering and post processing</td><td>
Table E2-2. Clustering Outcomes
<table>
<tr><td rowspan="2">Measurement</td><td colspan="4">Clustering Process</td></tr>
<tr><td>Prior hotspots (data points)</td><td>Later hotspots (data points)</td><td>Congestion Area (cluster)</td><td>Valid Congestion Area (cluster)</td></tr>
<tr><td>Input count</td><td>2397</td><td>2633</td><td>544</td><td>96</td></tr>
<tr><td>Outlier count</td><td>327</td><td>350</td><td>134</td><td>13</td></tr>
<tr><td>Cluster count</td><td>258</td><td>286</td><td>96</td><td>83</td></tr>
</table>


Figure E2-1. Identified Congestion Area in Bangkok with Its Associated Hotspots
</td></tr>
</table>

(continued) Table E2. Summary of output for peak hour at time 7:30

Congestion length distribution similarity

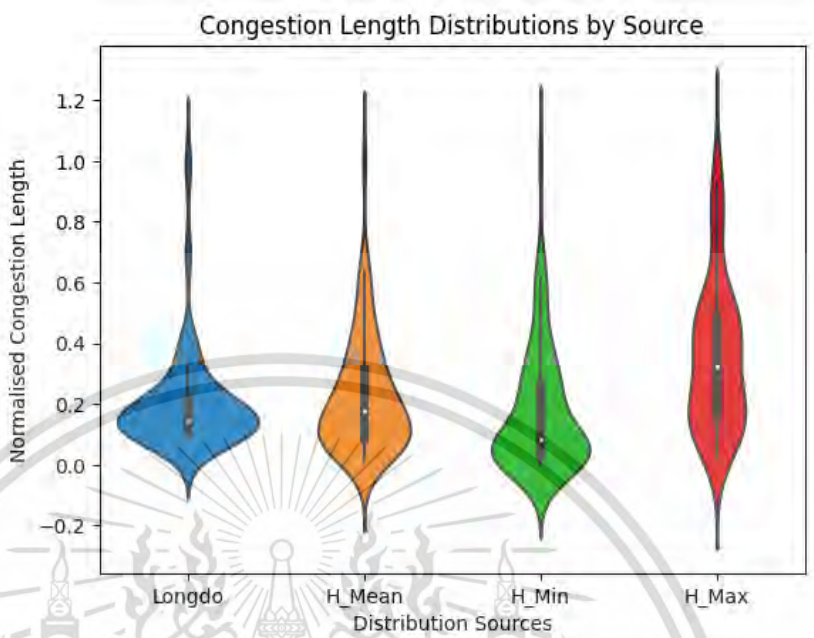


Figure E2-2. Congestion Length Distribution Based on Sources

Table E2-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.20589	0.2075	0.15363	0.375
Standard Deviation (SD)	0.16733	0.17549	0.18283	0.24597
Difference in mean against GT	0	0.00162	0.05225	0.16912
Difference in SD against GT	0	0.00816	0.0155	0.07864
JSD	0	0.01773	0.06162	0.02466
EMD	0	0.0486	0.06878	0.17779

Table E3. Summary of output for peak hour at time 8:30

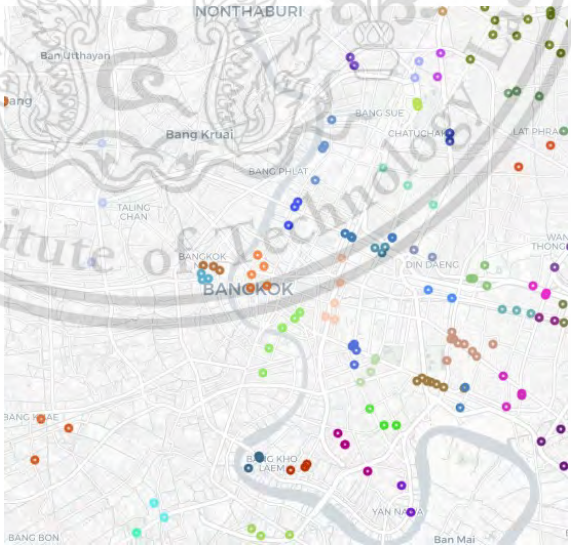
<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>
Table E3-1. Preprocessing Outcomes
<table>
<tr><td rowspan="2">Measurement</td><td colspan="4">Data Frame</td></tr>
<tr><td>Raw</td><td>Filtered</td><td>TF1</td><td>TF2</td></tr>
<tr><td>Data count</td><td>30925</td><td>5816</td><td>2956</td><td>2860</td></tr>
<tr><td>Vehicle ID count</td><td>3115</td><td>1037</td><td>805</td><td>815</td></tr>
</table>
.
</td></tr>
<tr><td>Clustering and post processing</td><td>
Table E3-2. Clustering Outcomes
<table>
<tr><td rowspan="2">Measurement</td><td colspan="4">Clustering Process</td></tr>
<tr><td>Prior hotspots (data points)</td><td>Later hotspots (data points)</td><td>Congestion Area (cluster)</td><td>Valid Congestion Area (cluster)</td></tr>
<tr><td>Input count</td><td>2956</td><td>2860</td><td>606</td><td>118</td></tr>
<tr><td>Outlier count</td><td>451</td><td>422</td><td>134</td><td>15</td></tr>
<tr><td>Cluster count</td><td>305</td><td>301</td><td>118</td><td>103</td></tr>
</table>


Figure E3-1. Identified Congestion Area in Bangkok with Its Associated Hotspots
</td></tr>
</table>

(continued) Table E3. Summary of output for peak hour at time 8:30

Congestion length distribution similarity

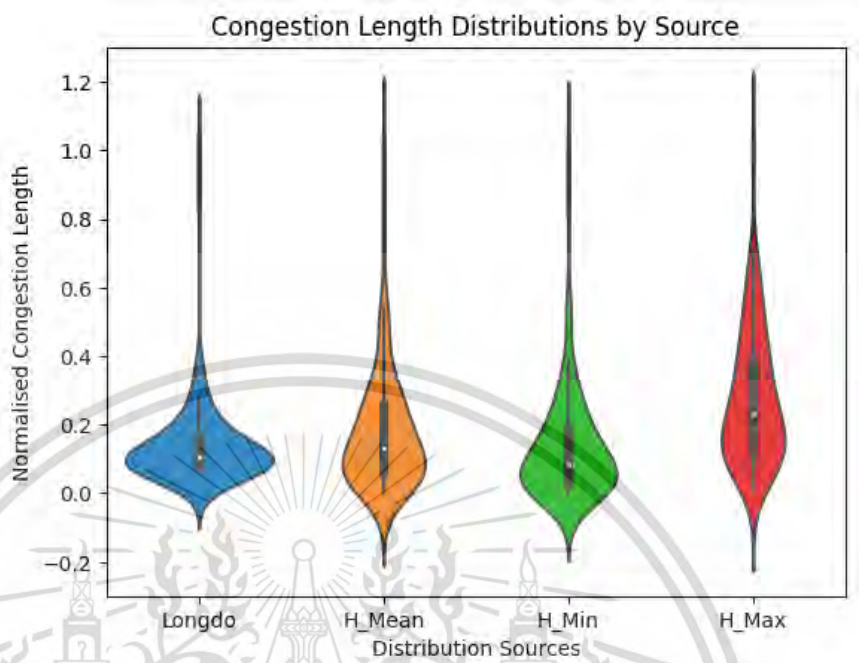


Figure E3-2. Congestion Length Distribution Based on Sources

Table E3-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.14206	0.18131	0.13705	0.27339
Standard Deviation (SD)	0.13396	0.17778	0.16603	0.1906
Difference in mean against GT	0	0.03925	0.00501	0.13133
Difference in SD against GT	0	0.04382	0.03207	0.05664
JSD	0	0.02745	0.0466	0.0254
EMD	0	0.06107	0.04388	0.13922

Table E4. Summary of output for peak hour at time 9:30

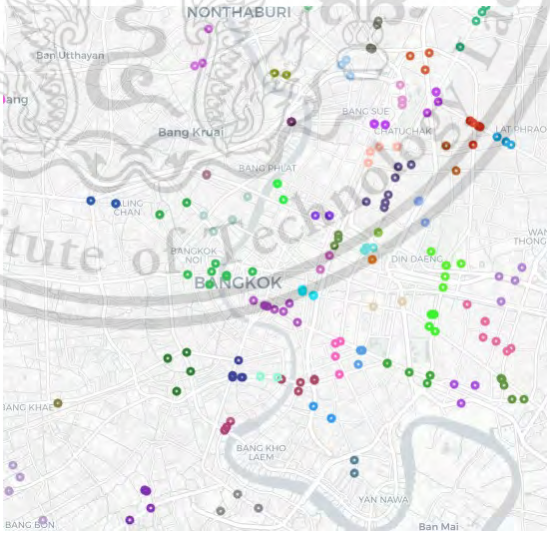
<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>
Table E4-1. Preprocessing Outcomes
<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Data Frame</th></tr>
<tr><th>Raw</th><th>Filtered</th><th>TF1</th><th>TF2</th></tr>
<tr><td>Data count</td><td>31317</td><td>6233</td><td>3157</td><td>3076</td></tr>
<tr><td>Vehicle ID count</td><td>3127</td><td>1082</td><td>842</td><td>853</td></tr>
</table>
.
</td></tr>
<tr><td>Clustering and post processing</td><td>
Table E4-2. Clustering Outcomes
<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Clustering Process</th></tr>
<tr><th>Prior hotspots (data points)</th><th>Later hotspots (data points)</th><th>Congestion Area (cluster)</th><th>Valid Congestion Area (cluster)</th></tr>
<tr><td>Input count</td><td>3157</td><td>3076</td><td>666</td><td>125</td></tr>
<tr><td>Outlier count</td><td>455</td><td>487</td><td>149</td><td>19</td></tr>
<tr><td>Cluster count</td><td>343</td><td>323</td><td>125</td><td>106</td></tr>
</table>


Figure E4-1. Identified Congestion Area in Bangkok with Its Associated Hotspots
</td></tr>
</table>

(continued) Table E4. Summary of output for peak hour at time 9:30

Congestion length distribution similarity

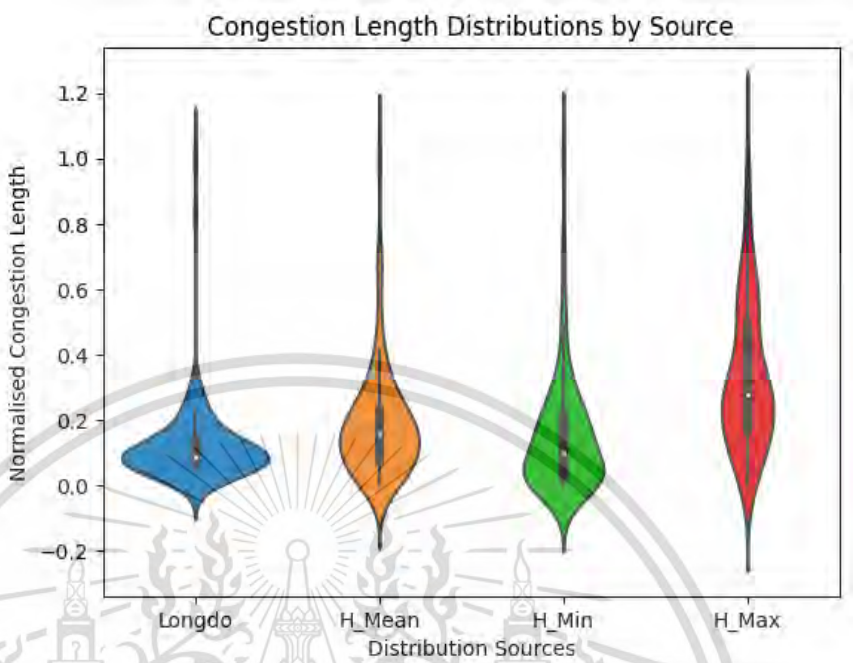


Figure E4-2. Congestion Length Distribution Based on Sources

Table E4-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.12335	0.18433	0.14638	0.32642
Standard Deviation (SD)	0.12932	0.16308	0.17078	0.21974
Difference in mean against GT	0	0.06098	0.02303	0.20307
Difference in SD against GT	0	0.03376	0.04146	0.09042
JSD	0	0.02763	0.05605	0.03758
EMD	0	0.07497	0.06151	0.20777

Table E5. Summary of output for peak hour at time 16:30

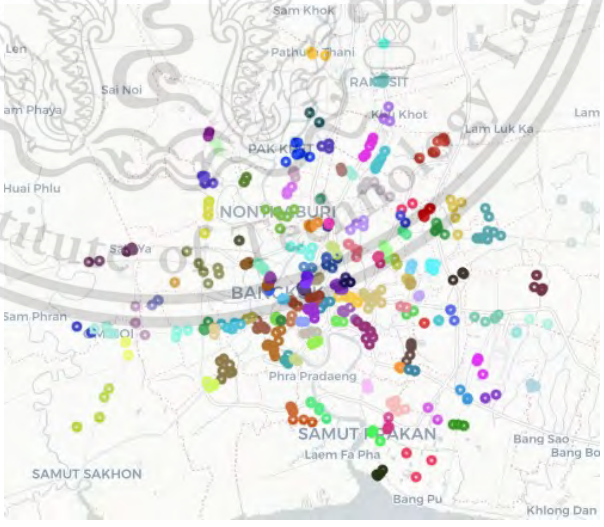
<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>
Table E5-1. Preprocessing Outcomes
<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Data Frame</th></tr>
<tr><th>Raw</th><th>Filtered</th><th>TF1</th><th>TF2</th></tr>
<tr><td>Data count</td><td>30466</td><td>6850</td><td>3355</td><td>3495</td></tr>
<tr><td>Vehicle ID count</td><td>3134</td><td>1176</td><td>933</td><td>927</td></tr>
</table>
.
</td></tr>
<tr><td>Clustering and post processing</td><td>
Table E5-2. Clustering Outcomes
<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Clustering Process</th></tr>
<tr><th>Prior hotspots (data points)</th><th>Later hotspots (data points)</th><th>Congestion Area (cluster)</th><th>Valid Congestion Area (cluster)</th></tr>
<tr><td>Input count</td><td>3355</td><td>3495</td><td>719</td><td>122</td></tr>
<tr><td>Outlier count</td><td>496</td><td>469</td><td>169</td><td>10</td></tr>
<tr><td>Cluster count</td><td>355</td><td>364</td><td>122</td><td>112</td></tr>
</table>


Figure E5-1. Identified Congestion Area in Bangkok with Its Associated Hotspots
</td></tr>
</table>

(continued) Table E5. Summary of output for peak hour at time 16:30

Congestion length distribution similarity

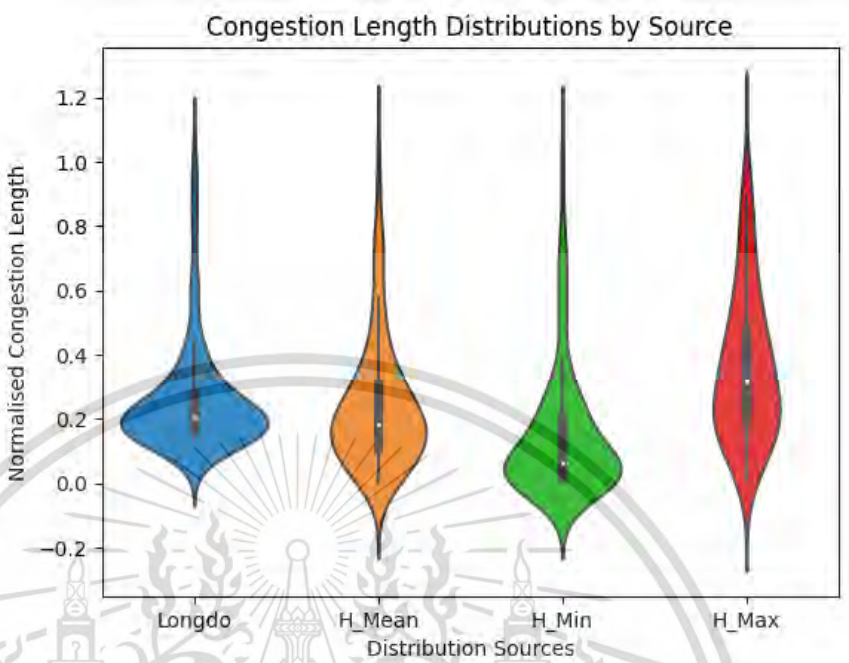


Figure E5-2. Congestion Length Distribution Based on Sources

Table E5-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.25803	0.23916	0.14679	0.35779
Standard Deviation (SD)	0.16918	0.19909	0.19746	0.23245
Difference in mean against GT	0	0.01887	0.11124	0.09976
Difference in SD against GT	0	0.02991	0.02828	0.06327
JSD	0	0.02359	0.09818	0.02022
EMD	0	0.04677	0.11288	0.12052

Table E6. Summary of output for peak hour at time 17:30

<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>

Table E6-1. Preprocessing Outcomes

<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Data Frame</th></tr>
<tr><th>Raw</th><th>Filtered</th><th>TF1</th><th>TF2</th></tr>
<tr><td>Data count</td><td>31702</td><td>6608</td><td>3439</td><td>3169</td></tr>
<tr><td>Vehicle ID count</td><td>3129</td><td>1059</td><td>866</td><td>808</td></tr>
</table>

.

</td></tr>
<tr><td>Clustering and post processing</td><td>

Table E6-2. Clustering Outcomes

<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Clustering Process</th></tr>
<tr><th>Prior hotspots (data points)</th><th>Later hotspots (data points)</th><th>Congestion Area (cluster)</th><th>Valid Congestion Area (cluster)</th></tr>
<tr><td>Input count</td><td>3439</td><td>3169</td><td>715</td><td>133</td></tr>
<tr><td>Outlier count</td><td>508</td><td>502</td><td>179</td><td>20</td></tr>
<tr><td>Cluster count</td><td>389</td><td>326</td><td>133</td><td>133</td></tr>
</table>

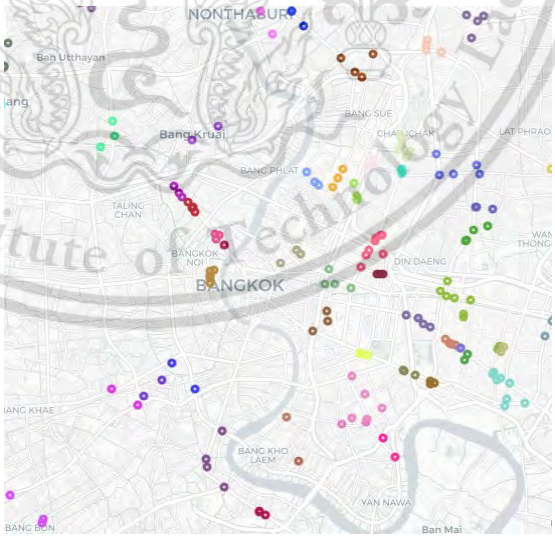


Figure E6-1. Identified Congestion Area in Bangkok with Its Associated Hotspots

</td></tr>
</table>

(continued) Table E6. Summary of output for peak hour at time 17:30

Congestion length distribution similarity

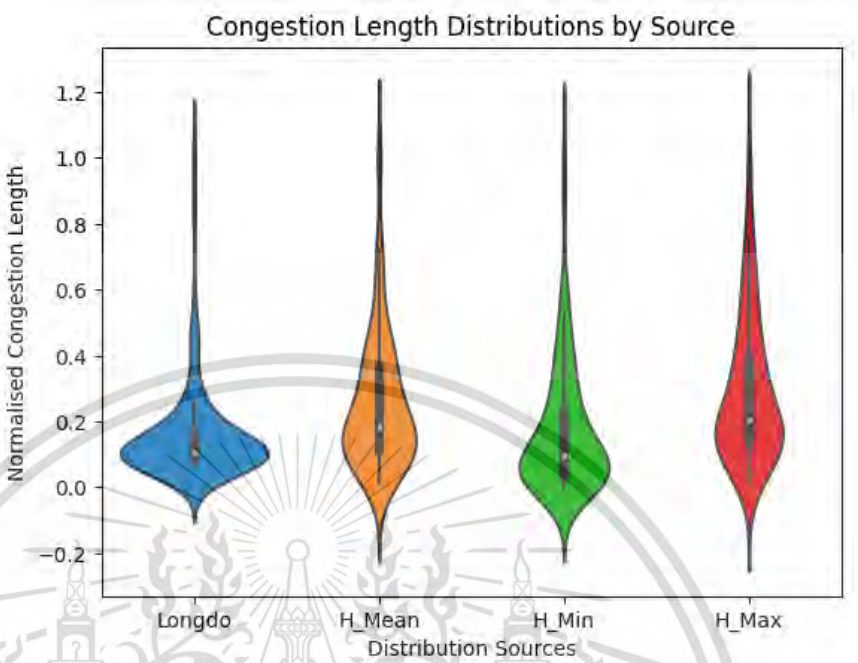


Figure E6-2. Congestion Length Distribution Based on Sources

Table E6-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.15676	0.25216	0.16768	0.28971
Standard Deviation (SD)	0.1469	0.20083	0.19006	0.21775
Difference in mean against GT	0	0.0954	0.01092	0.13295
Difference in SD against GT	0	0.05393	0.04316	0.07085
JSD	0	0.01934	0.04736	0.02188
EMD	0	0.10396	0.05699	0.13848

Table E7. Summary of output for peak hour at time 18:30

<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>

Table E7-1. Preprocessing Outcomes

<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Data Frame</th></tr>
<tr><th>Raw</th><th>Filtered</th><th>TF1</th><th>TF2</th></tr>
<tr><td>Data count</td><td>30500</td><td>5956</td><td>2904</td><td>3052</td></tr>
<tr><td>Vehicle ID count</td><td>3124</td><td>930</td><td>755</td><td>734</td></tr>
</table>

.

</td></tr>
<tr><td>Clustering and post processing</td><td>

Table E7-2. Clustering Outcomes

<table>
<tr><th rowspan="2">Measurement</th><th colspan="4">Clustering Process</th></tr>
<tr><th>Prior hotspots (data points)</th><th>Later hotspots (data points)</th><th>Congestion Area (cluster)</th><th>Valid Congestion Area (cluster)</th></tr>
<tr><td>Input count</td><td>2904</td><td>3052</td><td>652</td><td>117</td></tr>
<tr><td>Outlier count</td><td>400</td><td>477</td><td>173</td><td>16</td></tr>
<tr><td>Cluster count</td><td>322</td><td>330</td><td>117</td><td>101</td></tr>
</table>

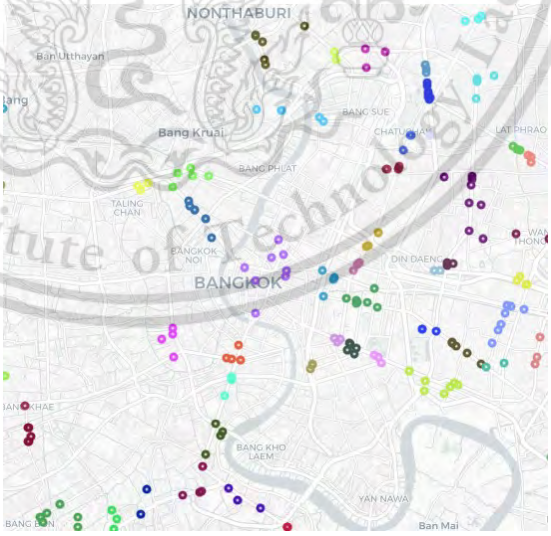


Figure E7-1. Identified Congestion Area in Bangkok with Its Associated Hotspots

</td></tr>
</table>

(continued) Table E7. Summary of output for peak hour at time 18:30

Congestion length distribution similarity

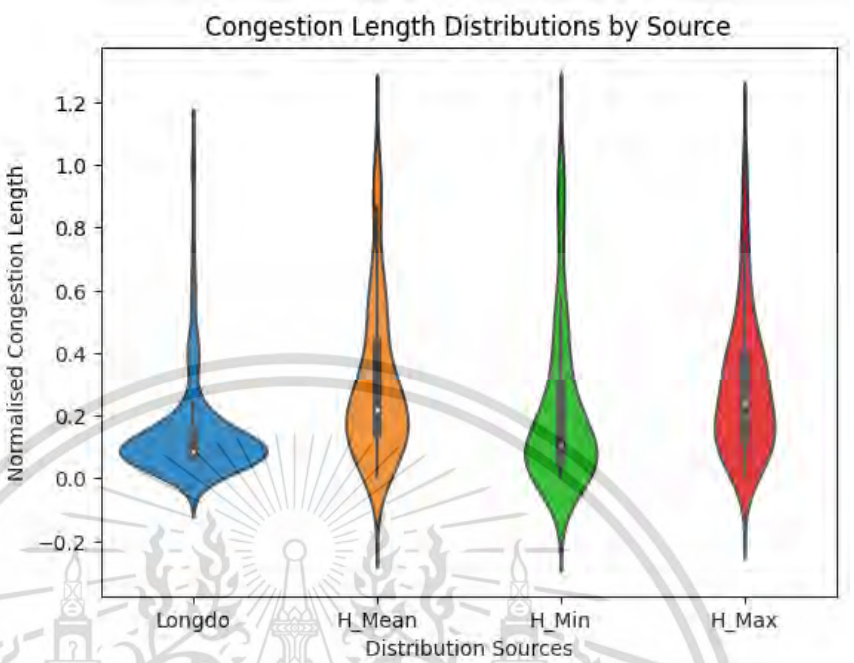


Figure E7-2. Congestion Length Distribution Based on Sources

Table E7-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.14366	0.30392	0.21434	0.292
Standard Deviation (SD)	0.1494	0.24302	0.24761	0.21914
Difference in mean against GT	0	0.16026	0.07068	0.14834
Difference in SD against GT	0	0.09362	0.09821	0.06974
JSD	0	0.02606	0.04066	0.02672
EMD	0	0.16551	0.09366	0.15293

Table E8. Summary of output for peak hour at time 19:30

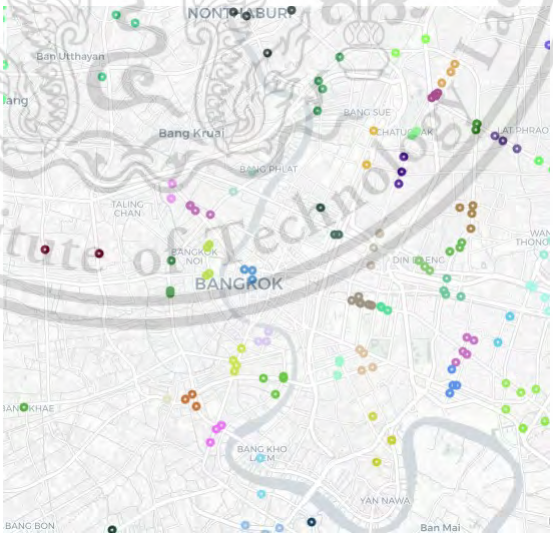
<table>
<tr><th>Stage</th><th>Result</th></tr>
<tr><td>Preprocessing</td><td>
Table E8-1. Preprocessing Outcomes
<table>
<tr><td rowspan="2">Measurement</td><td colspan="4">Data Frame</td></tr>
<tr><td>Raw</td><td>Filtered</td><td>TF1</td><td>TF2</td></tr>
<tr><td>Data count</td><td>29573</td><td>5150</td><td>2546</td><td>2604</td></tr>
<tr><td>Vehicle ID count</td><td>3115</td><td>894</td><td>710</td><td>709</td></tr>
</table>
.
</td></tr>
<tr><td>Clustering and post processing</td><td>
Table E8-2. Clustering Outcomes
<table>
<tr><td rowspan="2">Measurement</td><td colspan="4">Clustering Process</td></tr>
<tr><td>Prior hotspots (data points)</td><td>Later hotspots (data points)</td><td>Congestion Area (cluster)</td><td>Valid Congestion Area (cluster)</td></tr>
<tr><td>Input count</td><td>2546</td><td>2604</td><td>558</td><td>99</td></tr>
<tr><td>Outlier count</td><td>391</td><td>437</td><td>154</td><td>13</td></tr>
<tr><td>Cluster count</td><td>278</td><td>280</td><td>99</td><td>86</td></tr>
</table>


Figure E8-1. Identified Congestion Area in Bangkok with Its Associated Hotspots
</td></tr>
</table>

(continued) Table E8. Summary of output for peak hour at time 19:30

Congestion length distribution similarity

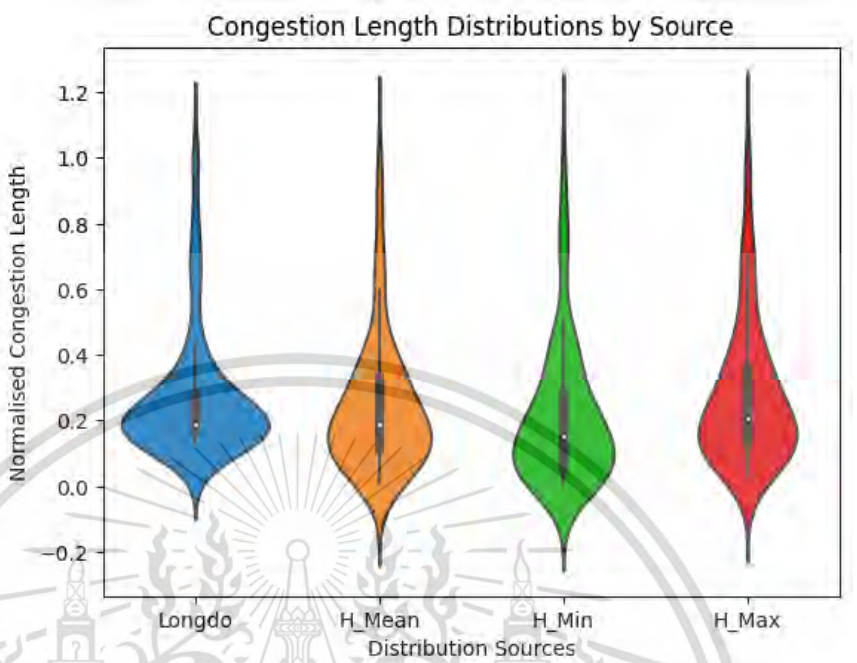


Figure E8-2. Congestion Length Distribution Based on Sources

Table E8-3. Similarity Scores of Congestion Length Distributions

Similarity measurement	Source of Congestion Length			
	Longdo Traffic (GT)	HDBSCAN Mean (H_Mean)	HDBSCAN Min (H_Min)	HDBSCAN Max (H_Max)
Mean	0.26537	0.24113	0.21119	0.27315
Standard Deviation (SD)	0.18653	0.20139	0.21077	0.21159
Difference in mean against GT	0	0.02424	0.05418	0.00778
Difference in SD against GT	0	0.01486	0.02424	0.02506
JSD	0	0.01706	0.03762	0.01114
EMD	0	0.04137	0.06352	0.04886