Information System and Smart City 2024, 4(1), 1737.
https://doi.org/10.59400/issc1737

Article

Data-driven insights: Unravelling traffic dynamics with k-means clustering and vehicle type differentiation

Anwar Mehmood Sohail[1,2], Khurram Shehzad Khattak[1,*], Zawar Hussain Khan[3]

[1] Department of Computer System Engineering, University of Engineering and Technology, Peshawar 25120, Pakistan

[2] Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia

[3] Department of Electrical and Computer Engineering, University of Victoria, Victoria V8W 2Y2, Canada

*** Corresponding author:** Khurram Shehzad Khattak, Khurram.s.khattak@gmail.com

CITATION

Sohail AM, Khattak KS, Khan ZH. Data-driven insights: Unravelling traffic dynamics with k-means clustering and vehicle type differentiation. Information System and Smart City. 2024; 4(1): 1737. https://doi.org/10.59400/issc1737

ARTICLE INFO

Received: 16 September 2024
Accepted: 22 October 2024
Available online: 21 November 2024

Abstract: Urban traffic poses persistent challenges, necessitating innovative approaches for effective traffic flow analysis and management. This research adopts a data-driven methodology, employing different algorithms such as K-Means clustering, multiple linear regression to analyse real-world traffic flow. The study utilizes road traffic data collected over seven days, spanning seven hours each day, comprising traffic count, vehicle speed, and categorization by vehicle type. Through rigorous data preprocessing and K-Means clustering, the research identifies distinct traffic clusters, revealing patterns beyond average counts and speeds. Notably, the differentiation of vehicle types within clusters provides nuanced insights into transport mode interactions. The findings contribute to the traffic flow analysis field and offer practical implications for informed urban traffic management strategies. Understanding traffic dynamics aids in developing effective congestion mitigation measures. The study concludes by highlighting potential areas for future research and improvements in optimizing traffic dynamics, emphasizing the importance of data-driven approaches in addressing urban traffic challenges.

Keywords: traffic flow analysis; k-means clustering; multi linear regression; machine learning; transportation

1. Introduction

Urban areas across the globe are facing an ever-growing dilemma posed by traffic congestion, a pervasive issue that not only impedes transportation efficiency but also poses significant challenges to environmental sustainability [1]. As cities expand and populations burgeon, the strain on transportation networks intensifies, exacerbating congestion woes and underscoring the urgent need for proactive measures. In light of these escalating challenges, it becomes increasingly imperative to prioritize effective traffic flow analysis as a cornerstone in the development of informed strategies aimed at alleviating congestion and enhancing urban mobility. By delving deeper into the intricacies of traffic patterns and dynamics, policymakers, urban planners, and transportation authorities can gain invaluable insights essential for the formulation and implementation of targeted interventions designed to mitigate congestion and foster sustainable urban development.

Urban traffic flow optimization present multifaceted challenges, necessitating advanced analytical methodologies [2]. Traditional methods often fall short in capturing the intricate patterns inherent in heterogeneous traffic, emphasizing the need for data-driven approaches, such as K-Means clustering [3]. Traditional traffic analysis has long relied on established methodologies such as traffic surveys, manual traffic counts, and statistical models [4]. These methods, while foundational, often face challenges in capturing the dynamic and complex nature of modern traffic patterns. Traditional methods exhibit limitations in scalability, accuracy, and real-time adaptability [4]. The reliance on historical data and manual data collection methods can lead to outdated insights, hindering the ability to address rapidly changing traffic dynamics. Recent advancements in technology have paved the way for data-driven approaches in traffic analysis. Leveraging real-time data from sources like sensors, GPS devices, and traffic cameras, researchers can attain a more granular understanding of traffic patterns [1,4].

More importantly with the advancement in computing power, artificial intelligence algorithms, as part of the broader data-driven paradigm, have found applications in identifying patterns within vast datasets. For example, previous studies have utilized clustering techniques to categorize traffic behaviour [2,3,5], contributing to a more nuanced understanding of traffic dynamics.

K-Means clustering, a popular unsupervised machine learning algorithm, partitions data into distinct clusters based on similarity. In the context of traffic flow analysis, K-Means can group similar traffic patterns, providing valuable insights into the heterogeneity of urban traffic. Successful applications of K-Means in traffic flow analysis showcase its ability to unveil hidden patterns and inform traffic management strategies. However, challenges such as sensitivity to initial conditions and the assumption of spherical clusters require careful consideration in implementation. Additionally, alongside K-Means clustering, we harness the potential of multiple linear regression techniques to delve deeper into traffic-related datasets, thereby augmenting our comprehension of traffic dynamics and facilitating more nuanced analyses.

This study endeavours to conduct a comprehensive analysis of traffic patterns over the span of a week, leveraging rich datasets encompassing traffic count, speed, and vehicle types. Through the application of K-Means clustering, our primary aim is to discern clusters exhibiting similar traffic behaviour, thereby enabling a nuanced comprehension of the multifaceted dynamics inherent within the urban road network. Additionally, we employ regression techniques to deepen our understanding of the impacts of various traffic flow variables. By harnessing data-driven insights, this research contributes to the enhancement of traffic management strategies, furnishing actionable information for optimizing road usage. The implications of our findings extend to policymakers, urban planners, and transportation authorities, offering potential avenues for improving traffic efficiency and mitigating congestion in urban settings.

The remainder of this work is structured as follows: Section 2 presents a comprehensive literature review. Section 3 delineates the methodology employed, while Section 4 provides detailed insights garnered from data analytics. Finally, Section 5 offers concluding remarks.

2. Literature review

This section provides an extensive review of past studies in urban traffic analysis and management, encompassing both traditional and data-driven methodologies. Through an examination of various approaches, including machine learning techniques such as K-Means clustering and regression analysis, our aim is to elucidate the progression of research in this domain. By synthesizing insights from a range of studies addressing diverse topics such as traffic flow pattern identification [2,5], congestion [3,6], traffic flow time slots [7,8], and traffic flow dynamics [9–11], we aim to discern key trends, methodologies, and existing gaps. This synthesis sets the foundation for our research methodology and underscores our contributions to this field.

In [2], the authors explore the utility of clustering analysis in enhancing transportation system management, operations, and modelling. They evaluate multiple clustering methods (k-means, k-prototypes, K-medoids, four variations of hierarchal method and combination of Principal Component Analysis (PCA) for mixed data) for discerning traffic patterns, highlighting their strengths and weaknesses. Stressing the importance of pattern recognition in transportation management, the study demonstrates how clustering algorithms can significantly contribute to this field. By aiming to bolster transportation system efficiency, the research utilizes clustering techniques to comprehensively analyse and model traffic behaviour. Through their investigation, the authors underscore the potential of clustering analysis in refining transportation modelling accuracy and decision-making processes, particularly emphasizing the effectiveness of K-prototypes and K-means with PCAs. Additionally, the paper offers practical recommendations for conducting and leveraging the outcomes of clustering analysis. In their study, [5] delve into clustering techniques to unveil traffic flow patterns in urban settings using extensive offline traffic data. Employing the K-means algorithm and incorporating temporal factors for pre-classification, the authors construct traffic flow vectors to identify distinct traffic patterns. Their methodology yields 11 clusters across workdays, weekends, and holidays, delineating patterns such as Morning-peaked, Evening-peaked, and non-peaked. Furthermore, the analysis of correlations between traffic patterns and geographical context enhances comprehension, suggesting promising applications in urban ITS.

In [3], the paper introduces an innovative K-means clustering method designed to identify urban hotspots by integrating spatial and temporal information, thus enhancing accuracy. Traditional K-means clustering encounters challenges with large datasets and high-dimensional data, prompting the development of a hybrid heuristic algorithm called FPSO-GAK to overcome these obstacles. Combining fuzzy system, particle swarm optimization, and genetic algorithm techniques, FPSO-GAK efficiently determines initial clustering centres for K-means. Experimental validation utilizing taxi GPS and multi-source datasets illustrates the superior performance of FPSO-GAK in accurately pinpointing urban hotspots compared to other clustering algorithms. Li et al. [6] undertake a comprehensive analysis of traffic congestion patterns in Beijing, employing a data-driven approach to tackle the growing congestion challenge in megacities. Leveraging extensive datasets on traffic flow and congestion, the study aims to uncover underlying patterns and variations contributing to congestion dynamics. The authors introduce an innovative weighted K-means clustering algorithm, which accounts for the varying importance of sampling points across different time segments, to identify representative congestion patterns accurately. Through the paired t-test method, they scrutinize the spatial and temporal variations of these patterns, drawing insights from a substantial dataset spanning six districts from 1 January 2017, to 31 December 2017. The findings underscore the temporal and spatial dependencies of congestion patterns, with notable impacts from automobile license plate restrictions. This research offers valuable insights for crafting targeted traffic optimization strategies to alleviate congestion and enhance overall traffic conditions in urban environments, particularly in cities like Beijing.

Muntean's study presents an innovative method for identifying critical traffic flow time slots through the integration of K-means clustering and decision trees [7]. The research aims to detect and analyse time periods marked by notable traffic flow fluctuations. Initially, K-means clustering categorizes traffic flow data into distinct groups based on similarity. Subsequently, decision trees are deployed to classify these groups and identify critical time slots based on predefined criteria. Through empirical evaluations and practical case studies, the study demonstrates the effectiveness of this approach in accurately pinpointing critical traffic flow time slots. [8] research investigates road junction time period analysis using the K-means algorithm. Their study aims to unveil meaningful patterns in traffic flow dynamics by applying K-means clustering to time periods at road junctions. Through this method, they aim to discern distinct time periods characterized by varying traffic conditions, providing insights for optimizing traffic signal timings and junction management strategies. By leveraging data mining techniques, particularly the K-means algorithm, the study contributes to addressing traffic congestion challenges. Empirical analysis illustrates the effectiveness of their proposed approach in accurately classifying traffic flow, highlighting differences between weekday and weekend traffic patterns.

Daniswara and Gunawan [9] introduce an innovative approach to simulating transport problems using a clustering velocity-density function. Their study aims to address transportation challenges through novel simulation techniques, employing clustering methods to discern distinct velocity-density patterns within the transport network, providing deeper insights into traffic dynamics. Through simulations and analysis, the authors showcase the effectiveness of their method in tackling transport issues. Utilizing K-Means clustering, they derive two clusters: jammed and light, each exhibiting different velocity-density functions derived from linear regression. Evaluation metrics such as RMSE and R-Squared are employed to gauge the accuracy of the method, yielding promising results. In their study [10], the authors investigate uninterrupted traffic flow dynamics using K-means clustering, focusing on a highway segment in Isfahan, Iran. Employing K-means clustering, they delineate three distinct traffic phases: free flow (F), synchronized (S), and wide-moving jam (J), consistent with the established three-phase theory of uninterrupted traffic flow. The research delves into transition speeds and densities among these phases, supplemented by Shannon entropy analysis for real-time traffic flow state monitoring. The integration of Shannon entropy analysis unveils temporal traffic flow dynamics, shedding light on transitions between free flow, synchronized, and wide-moving jam states. The study in [11] introduces a hybrid traffic state recognition model for urban expressways, amalgamating K-means clustering with AdaBoost-DS. By incorporating traffic flow, velocity, and occupancy parameters, the authors categorize expressway traffic states into four groups using the K-means clustering algorithm. The resulting traffic flow data then trains the AdaBoost-DS model for recognition purposes. Results from example verification and comparative analysis of urban expressway data showcase the method's effectiveness, achieving a recognition accuracy of 93.2%, surpassing BP neural network by 7.6%.

This section offers an encompassing review of prior endeavours in urban traffic analysis and management, underscoring the breadth of methodologies employed to address transportation complexities. Spanning from conventional techniques to state-of-the-art data-driven methodologies, researchers have explored diverse avenues to decipher and enhance traffic dynamics. In the continuum of this discourse, our study builds upon these established frameworks, striving to propel the understanding and optimization of urban traffic flow through a synergistic integration of clustering and regression analysis techniques. Leveraging K-Means clustering, our research unveils latent traffic flow patterns, shedding light on both the flow dynamics and the broader composition of traffic. Additionally, we employ regression methodologies to dissect the influence of vehicle types on traffic flow density and velocity, providing nuanced insights into the multifaceted dynamics of urban traffic.

3. Research methodology

Before data collection, careful consideration was given to the selection of data collection methods and equipment to ensure accurate and reliable data acquisition. The chosen sensor node configuration and data transmission protocols were designed to optimize both efficiency and cost-effectiveness.

3.1. Data collection

The road traffic dataset was acquired using a specialized sensor node designed for efficient and cost-effective data collection [1]. This sensor node consisted of a Raspberry Pi 4 (RPi), a low-cost Linux-based single-board computer, paired with a Pi camera v2 connected via the Camera Serial Interface (CSI) port. This setup facilitated the capture of high-resolution (1080p) roadside videos at 20 frames per second (fps). To ensure extended operation, a 10,000 mAh Xiaomi Mi Power Bank provided ample battery life. The sensor node transmitted traffic parameters to the cloud platform "ThingSpeak" using a Zong 4G Bolt + for reliable communication. Additionally, an Intel Movidius Compute Stick 2 enhanced computing capabilities, minimizing power consumption by offloading complex computations from the RPi.

Strategically positioned on University Road in Peshawar, Pakistan, located south of Islamia College and adjacent to the Bus Rapid Transport (BRT) station [1], the sensor node diligently captured data for seven consecutive days, commencing from Monday, 8 May 2023, to Sunday, 14 May 2023. Operating diligently for seven hours daily, from 9:00 AM to 4:00 PM, University Road was chosen for its bustling traffic volume and diverse vehicular composition, serving as a vital arterial route connecting major institutions. The sensor node efficiently transmitted a myriad of traffic parameters, including vehicle count, speed, direction, type, flow, peak hour factor, density, time headway, and distance headway, with updates streamed every 15 seconds to the ThingSpeak cloud platform.

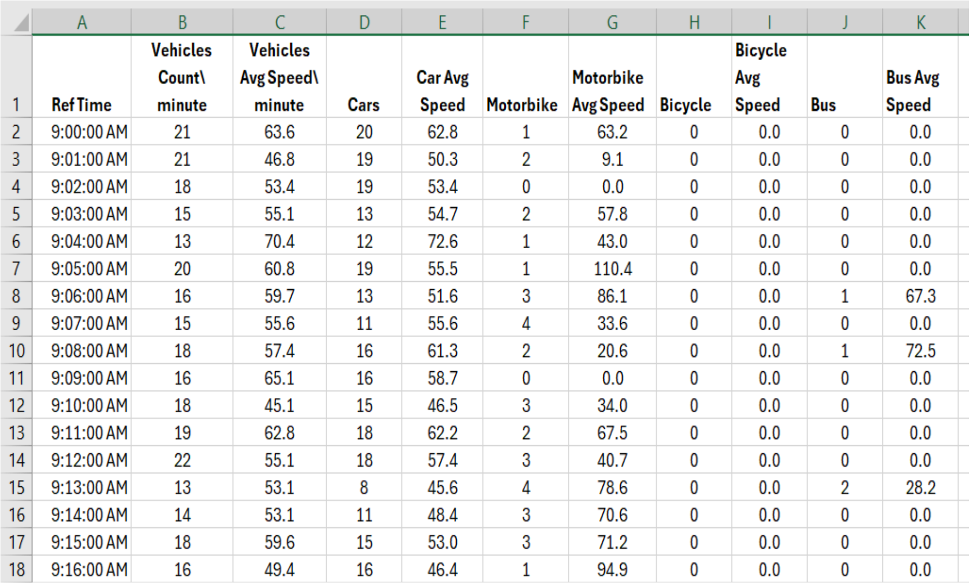
3.2. Data preprocessing

Accurate estimation of traffic parameters is essential for effective monitoring systems. Throughout the evaluation period, the sensor node operated flawlessly, capturing reliable traffic data detailed in **Table 1**. Evaluation primarily focused on vehicle count, speed, classification, and direction, with additional parameters like traffic flow, peak hour factor, density, and time/distance headway derived mathematically and streamed to the cloud platform. Manual testing conducted by the authors [1] reported accuracy rates of 79.8% for vehicle classification and 82.9% for speed estimation, representing the highest reported accuracy in similar literature. Vehicle count and speed were computed on a per-minute basis, along with separate calculations for each vehicle type.

Table 1. Breakdown of traffic flow on the observed road section.

Day	Cars	Motorbikes	Buses	Bicycles	Total
Monday	8772	2012	60	250	11087
Tuesday	7361	1208	77	250	8899
Wednesday	7644	1729	53	195	9614
Thursday	5854	746	43	54	6696
Friday	7373	1621	175	297	9517
Saturday	7528	1592	103	369	9590
Sunday	8143	1496	61	390	10090
Total	52,675	10,404	572	1805	65,493
Percentage	81%	15.6%	2.5%	0.72%	

For this study, we meticulously computed the average vehicle count and speed on a per-minute basis. Additionally, we conducted separate calculations for the average count and speed per minute for each type of vehicle, as illustrated in **Figure 1**, to discern any nuanced variations across different vehicle categories.



	A	B	C	D	E	F	G	H	I	J	K
1	Ref Time	Vehicles Count\ minute	Vehicles Avg Speed\ minute	Cars	Car Avg Speed	Motorbike	Motorbike Avg Speed	Bicycle	Bicycle Avg Speed	Bus	Bus Avg Speed
2	9:00:00 AM	21	63.6	20	62.8	1	63.2	0	0.0	0	0.0
3	9:01:00 AM	21	46.8	19	50.3	2	9.1	0	0.0	0	0.0
4	9:02:00 AM	18	53.4	19	53.4	0	0.0	0	0.0	0	0.0
5	9:03:00 AM	15	55.1	13	54.7	2	57.8	0	0.0	0	0.0
6	9:04:00 AM	13	70.4	12	72.6	1	43.0	0	0.0	0	0.0
7	9:05:00 AM	20	60.8	19	55.5	1	110.4	0	0.0	0	0.0
8	9:06:00 AM	16	59.7	13	51.6	3	86.1	0	0.0	1	67.3
9	9:07:00 AM	15	55.6	11	55.6	4	33.6	0	0.0	0	0.0
10	9:08:00 AM	18	57.4	16	61.3	2	20.6	0	0.0	1	72.5
11	9:09:00 AM	16	65.1	16	58.7	0	0.0	0	0.0	0	0.0
12	9:10:00 AM	18	45.1	15	46.5	3	34.0	0	0.0	0	0.0
13	9:11:00 AM	19	62.8	18	62.2	2	67.5	0	0.0	0	0.0
14	9:12:00 AM	22	55.1	18	57.4	3	40.7	0	0.0	0	0.0
15	9:13:00 AM	13	53.1	8	45.6	4	78.6	0	0.0	2	28.2
16	9:14:00 AM	14	53.1	11	48.4	3	70.6	0	0.0	0	0.0
17	9:15:00 AM	18	59.6	15	53.0	3	71.2	0	0.0	0	0.0
18	9:16:00 AM	16	49.4	16	46.4	1	94.9	0	0.0	0	0.0

Figure 1. Processed traffic flow data.

3.3. Feature selection

Selecting relevant features for clustering involved considering various traffic parameters to capture underlying patterns in road traffic data effectively. Key features such as vehicle count, speed, and classification were chosen based on their significance in characterizing traffic flow dynamics and providing insights into traffic behaviour and patterns. Vehicle count and speed directly influence traffic congestion and road capacity utilization, while vehicle classification accounts for diverse vehicle compositions and directional variations in traffic flow. By focusing on these features, the clustering analysis aimed to capture the variability and complexity of road traffic patterns comprehensively.

3.4. K-means clustering

The K-Means clustering algorithm was utilized to group similar traffic patterns based on selected features. Operating iteratively, K-Means assigns data points to random cluster centroids and updates centroids based on the mean of data points assigned to each cluster until convergence. The optimal number of clusters (K) was determined using the silhouette coefficient method. K-Means clustering facilitated the identification of distinct traffic patterns, offering valuable insights into traffic flow dynamics such as peak traffic periods and congestion hotspots. Through this application, K-Means provided actionable information for informing traffic management strategies and enhancing urban mobility.

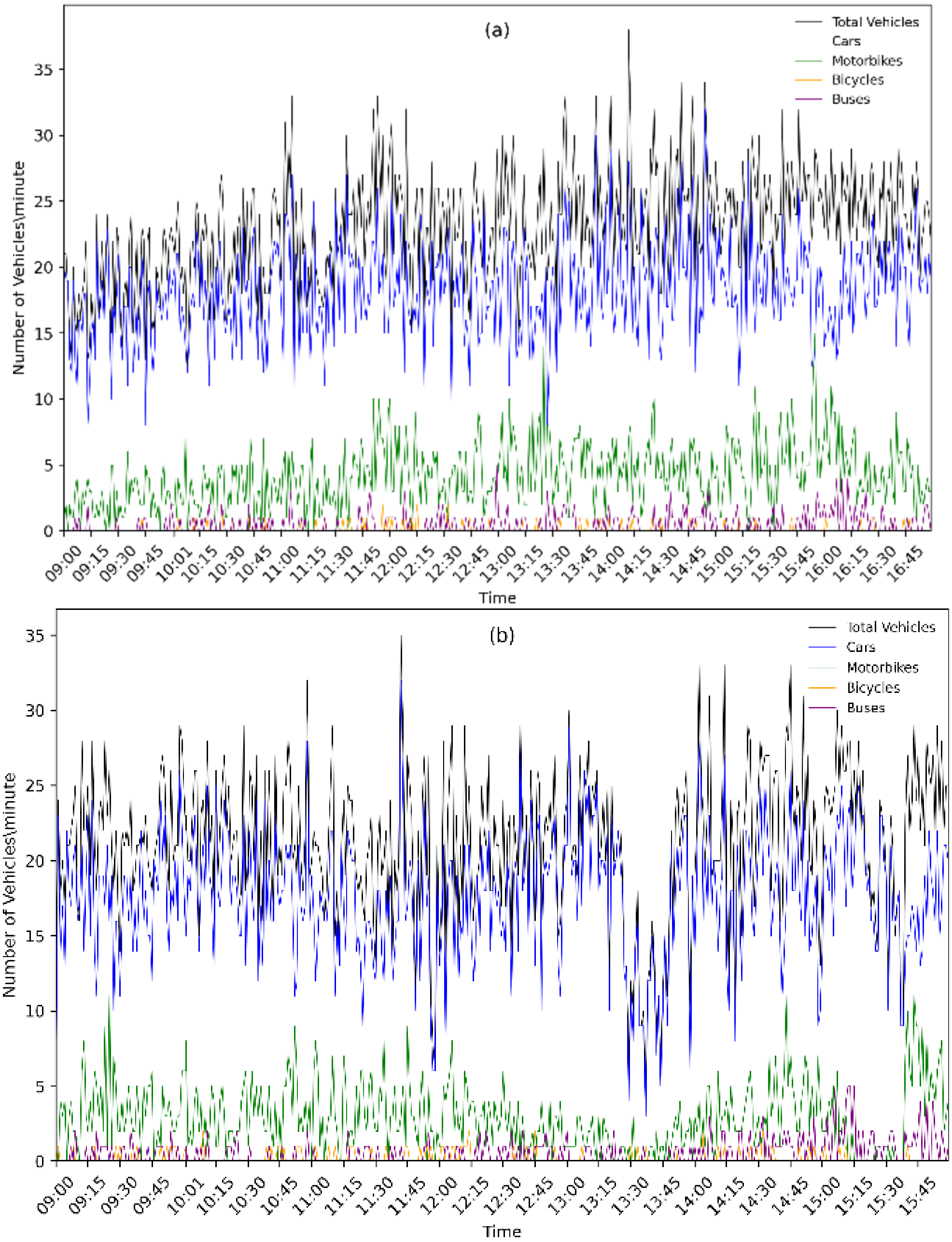
4. Results

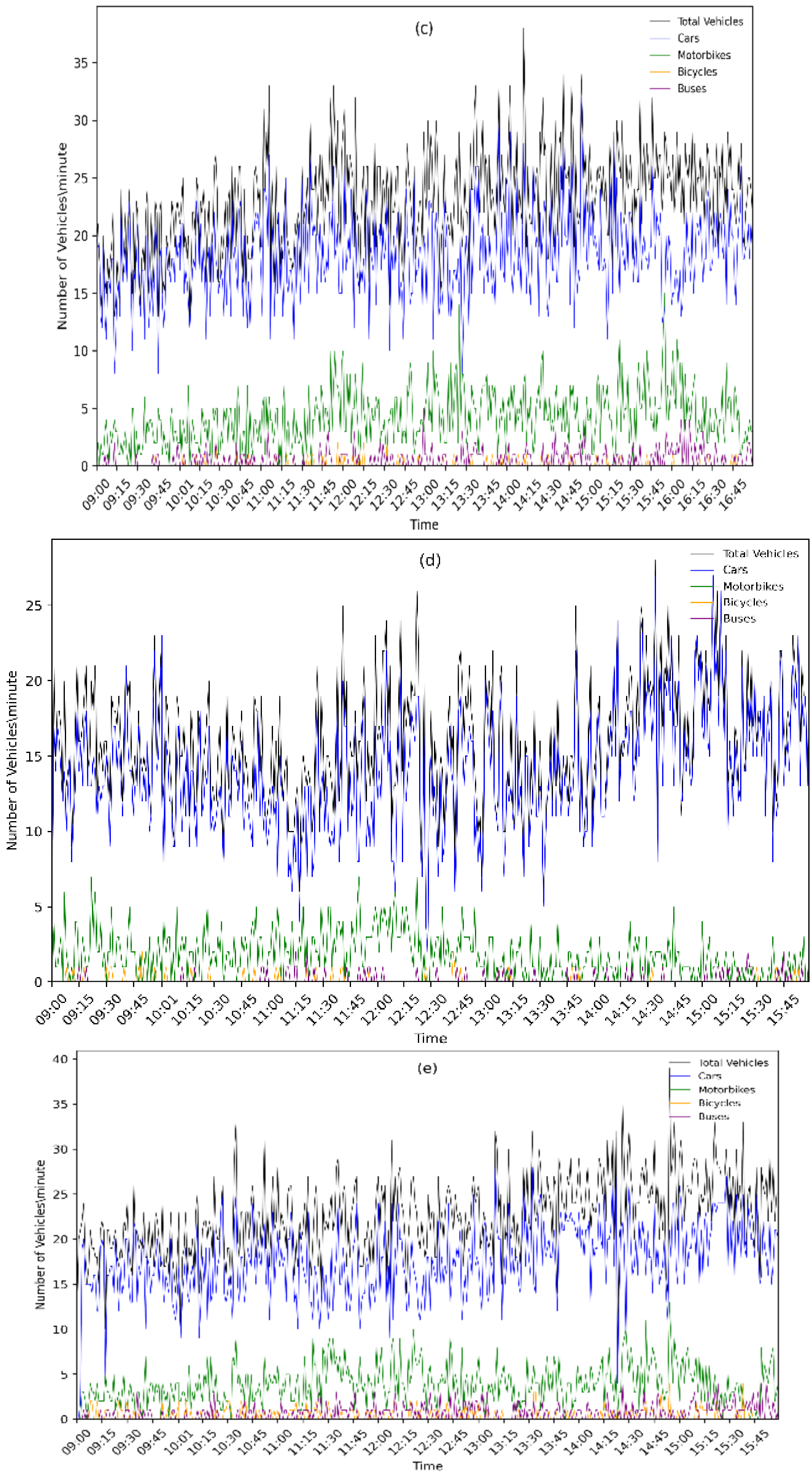
The following section presents the results obtained from the comprehensive analysis of traffic patterns observed over a week. The analysis includes descriptive statistics detailing the breakdown of traffic flow and an examination of the average speed of different vehicle types. These results offer valuable insights into the dynamics of urban traffic and provide a foundation for further interpretation and analysis.

4.1. Descriptive statistics

The traffic flow breakdown on the observed road segment over the one-week observation period is summarized in **Table 1**. Throughout the seven-hour period each day from 9:00 AM to 4:00 PM, an average of approximately 9,000 vehicles traversed the road, with a total count of 65,493 vehicles recorded for the week. Cars constituted the majority of the vehicular composition (81%), followed by motorcycles (15.6%), buses (2.5%), and bicycles (0.72%). Notably, Monday witnessed the highest traffic volume, while Thursday recorded the lowest, as illustrated in **Table 1**.

The breakdown of traffic flow is further elucidated in **Figure 2**. This figure demonstrates the fluctuation in traffic flow count per minute throughout the observation period, highlighting the predominance of cars and the sporadic presence of buses and bicycles. While motorcycles consistently contribute to the traffic flow, buses and bicycles exhibit varying patterns, with some time slots experiencing no presence of these vehicle types.


(a)
Total Vehicles
Cars
Motorbikes
Bicycles
Buses
Number of Vehicles\minute
0
5
10
15
20
25
30
35
Time
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45
16:00
16:15
16:30
16:45
(b)
Total Vehicles
Cars
Motorbikes
Bicycles
Buses
Number of Vehicles\minute
0
5
10
15
20
25
30
35
Time
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45


(c)
Total Vehicles
Cars
Motorbikes
Bicycles
Buses
Number of Vehicles\minute
0
5
10
15
20
25
30
35
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45
16:00
16:15
16:30
16:45
Time
(d)
Total Vehicles
Cars
Motorbikes
Bicycles
Buses
Number of Vehicles\minute
0
5
10
15
20
25
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45
Time
(e)
Total Vehicles
Cars
Motorbikes
Bicycles
Buses
Number of Vehicles\minute
0
5
10
15
20
25
30
35
40
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45
Time

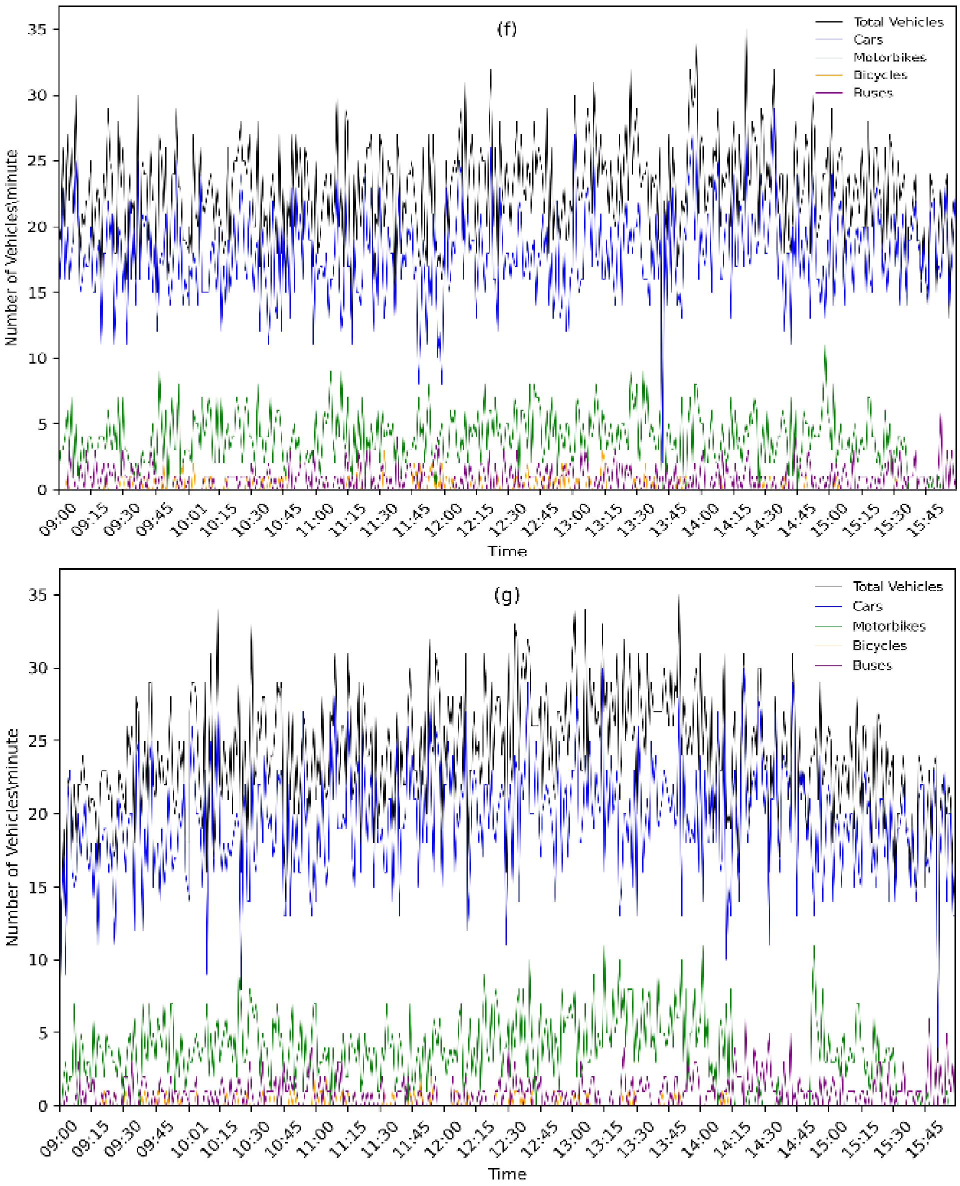
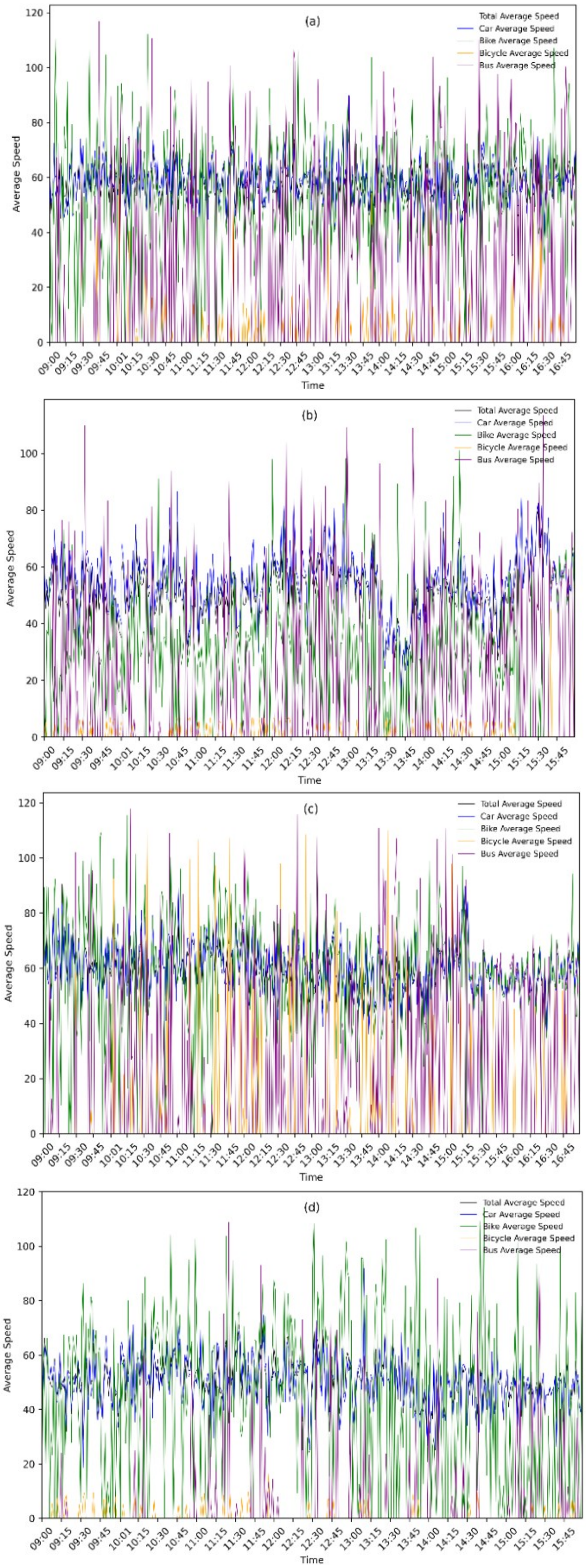


Figure 2. Traffic flow count per minute on the under observed road section from 9:00–16:00 on **(a)** Monday; **(b)** Tuesday; **(c)** Wednesday; **(d)** Thursday; **(e)** Friday; **(f)** Saturday; **(g)** Sunday.

Another critical aspect of traffic flow analysis is the speed of different vehicle types and their impact on overall traffic flow speed. **Figure 3** presents the breakdown of average speed for various vehicle types on different days. The average speed of cars closely aligns with the overall traffic flow speed, indicating their significant contribution to traffic dynamics. Similarly, motorcycles generally adhere to the traffic flow speed, with occasional deviations observed in certain time slots.

Notably, buses exhibit spikes in average speed, attributed to specialized vans navigating aggressively to maximize passenger pickup. This observation underscores the influence of vehicle behaviour on traffic flow dynamics and highlights the impact of unique vehicle types on urban traffic patterns.


(a)
Total Average Speed
Car Average Speed
Bike Average Speed
Bicycle Average Speed
Bus Average Speed
Average Speed
0
20
40
60
80
100
120
Time
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45
16:00
16:15
16:30
16:45
(b)
Total Average Speed
Car Average Speed
Bike Average Speed
Bicycle Average Speed
Bus Average Speed
Average Speed
0
20
40
60
80
100
Time
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45
(c)
Total Average Speed
Car Average Speed
Bike Average Speed
Bicycle Average Speed
Bus Average Speed
Average Speed
0
20
40
60
80
100
120
Time
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45
16:00
16:15
16:30
16:45
(d)
Total Average Speed
Car Average Speed
Bike Average Speed
Bicycle Average Speed
Bus Average Speed
Average Speed
0
20
40
60
80
100
120
Time
09:00
09:15
09:30
09:45
10:01
10:15
10:30
10:45
11:00
11:15
11:30
11:45
12:00
12:15
12:30
12:45
13:00
13:15
13:30
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30
15:45

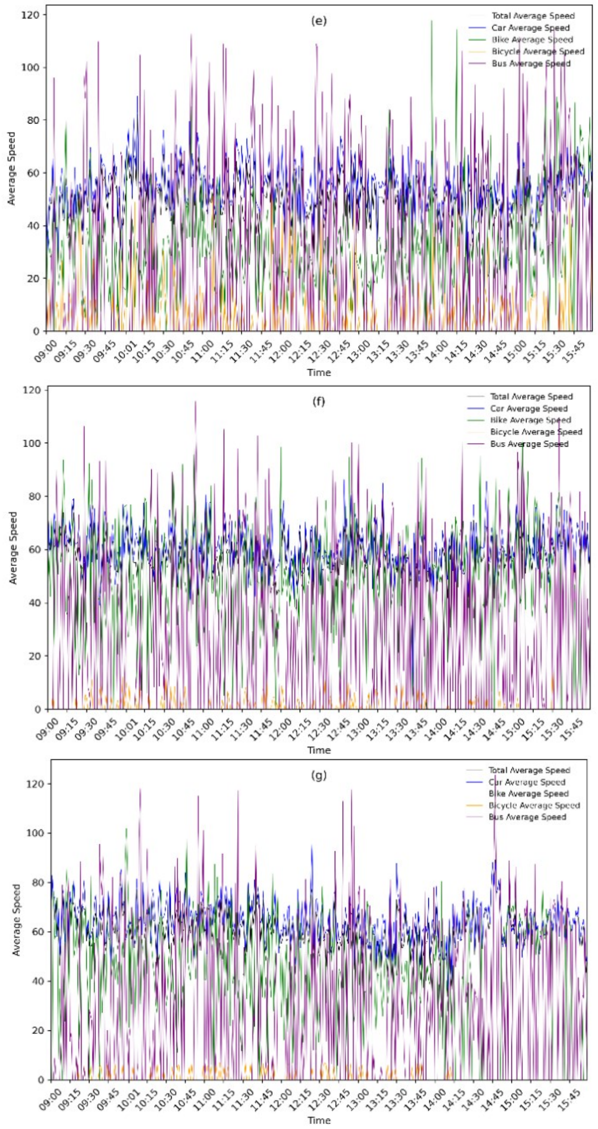


Figure 3. Average traffic flow speed on the under observed road section from 9:00AM–4:00PM on **(a)** Monday; **(b)** Tuesday; **(c)** Wednesday; **(d)** Thursday; **(e)** Friday; **(f)** Saturday; **(g)** Sunday.

These descriptive statistics and analyses provide a comprehensive overview of traffic dynamics on the under-observed road section, offering insights into vehicular composition, traffic volume, and speed variations. Such insights are instrumental in informing traffic management strategies and optimizing urban mobility to alleviate congestion and enhance overall transportation efficiency.

4.2. K-means clustering

The application of the K-Means clustering algorithm to the pre-processed traffic dataset aimed to uncover distinct traffic patterns and consolidate similar data points. Initially, the dataset underwent preprocessing to standardize variables like traffic flow speed, vehicle type, and traffic density, ensuring equitable representation of features during clustering. The K-Means algorithm iteratively processed the standardized data, with the number of clusters (K) determined based on optimal values. Each data point was then assigned to the nearest cluster centroid using the Euclidean distance metric, with centroids continually updated until convergence was achieved. This iterative process yielded clusters representing cohesive groups of traffic data with shared characteristics, offering insights into underlying dataset patterns.

To identify the optimal number of clusters, several techniques were explored, including the elbow method and silhouette analysis. These methods assessed clustering quality, balancing intra-cluster similarity and inter-cluster dissimilarity to determine the most suitable cluster count. The selection rationale considered dataset properties, computational efficiency, and cluster interpretability. Notably, the silhouette coefficient consistently indicated an optimal two-cluster solution across all days of the week, with values ranging from 0.41 to 0.44 from Monday to Sunday. This uniformity underscores the stability and reliability of the chosen cluster configuration, facilitating robust interpretation and analysis of traffic patterns.

4.3. Cluster analysis

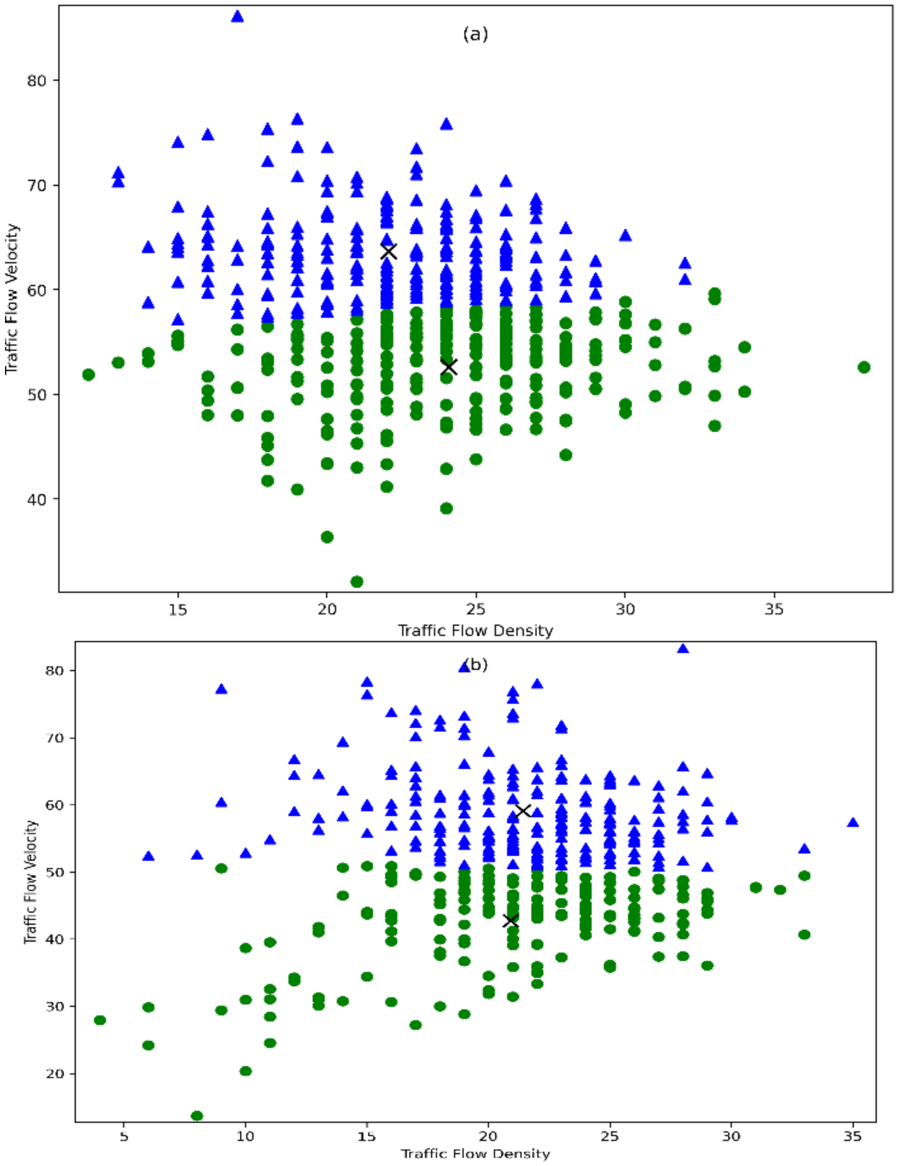
The application of the K-Means clustering algorithm yielded distinct clusters, each representing unique traffic patterns within the dataset. These clusters group together traffic data points with similar characteristics, offering insights into the underlying dynamics of traffic flow. For instance, the range of traffic flow velocities varies across different days, reflecting fluctuations in traffic density and velocity as can be seen in **Figure 4**. Notably, weekdays exhibit greater variability in traffic flow characteristics compared to weekends, where traffic conditions are more consistent.

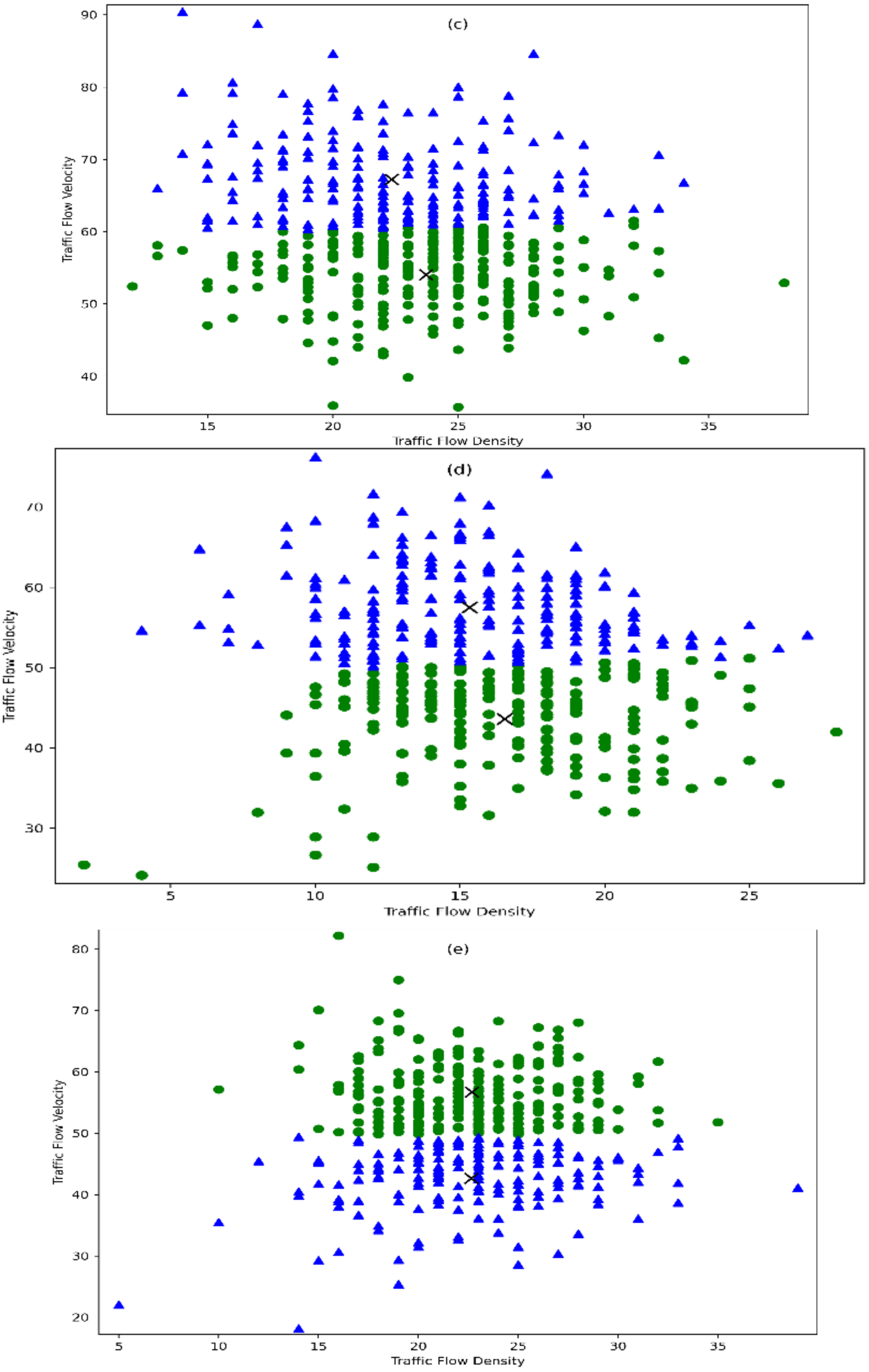
In **Figure 4**, the variability in traffic flow velocity across different days is apparent. For instance, on Saturday, the traffic flow velocity ranges between 40-80 km/h, while the slowest speeds are observed on Friday, with a range of 30-70 km/h. Additionally, the spatial distribution of data points on the charts varies depending on the day, influenced by the volume of on-road vehicles. For instance, on Tuesday and Wednesday, significant fluctuations in both per-minute traffic flow velocity and density are observed, indicating dynamic traffic patterns. Conversely, weekends exhibit more uniform traffic flow characteristics, resulting in data points clustering closely together within a small circumference on the charts.

Table 2. Breakdown of cluster statistics.

Day	Cluster 0			Cluster 1		
	Traffic count	Centroid traffic count	Centroid traffic velocity	Traffic count	Centroid traffic count	Centroid traffic velocity
Monday	48	24	52	48	22	63
Tuesday	42	20	42	42	21	59
Wednesday	48	23	54	48	22	67
Thursday	42	16	43	42	15	57
Friday	42	22	56	42	22	56
Saturday	42	23	53	42	21	66
Sunday	42	25	55	42	22	68

The clustering analysis reveals distinct traffic flow states across different days of the week, highlighting the dynamic nature of urban traffic. Each cluster's centroid encapsulates the average characteristics of its data points. Typically, Cluster 0 signifies lower vehicle counts and slower average speeds, indicative of congested or pre-congestion states, as shown in **Table 2**. Conversely, Cluster 1 is associated with higher vehicle counts and faster average speeds, suggesting free-flowing traffic conditions. Analysis of the centroids from Monday to Friday shows that Cluster 0 tends to have lower vehicle counts and slower speeds during peak traffic hours, signaling potential congestion. In contrast, Cluster 1 generally exhibits higher vehicle counts and faster speeds, often corresponding to off-peak hours or less congested periods. Interestingly, on weekends (Saturday and Sunday), both clusters' centroids show higher vehicle counts and faster speeds compared to weekdays, possibly reflecting reduced congestion and smoother traffic flow due to decreased commuter traffic. These findings from the clustering analysis provide valuable insights into the temporal variations of traffic flow states, enhancing our understanding and management of urban traffic dynamics.




(c)
Traffic Flow Velocity
90
80
70
60
50
40
15
20
25
30
35
Traffic Flow Density
(d)
Traffic Flow Velocity
70
60
50
40
30
5
10
15
20
25
Traffic Flow Density
(e)
Traffic Flow Velocity
80
70
60
50
40
30
20
5
10
15
20
25
30
35
Traffic Flow Density

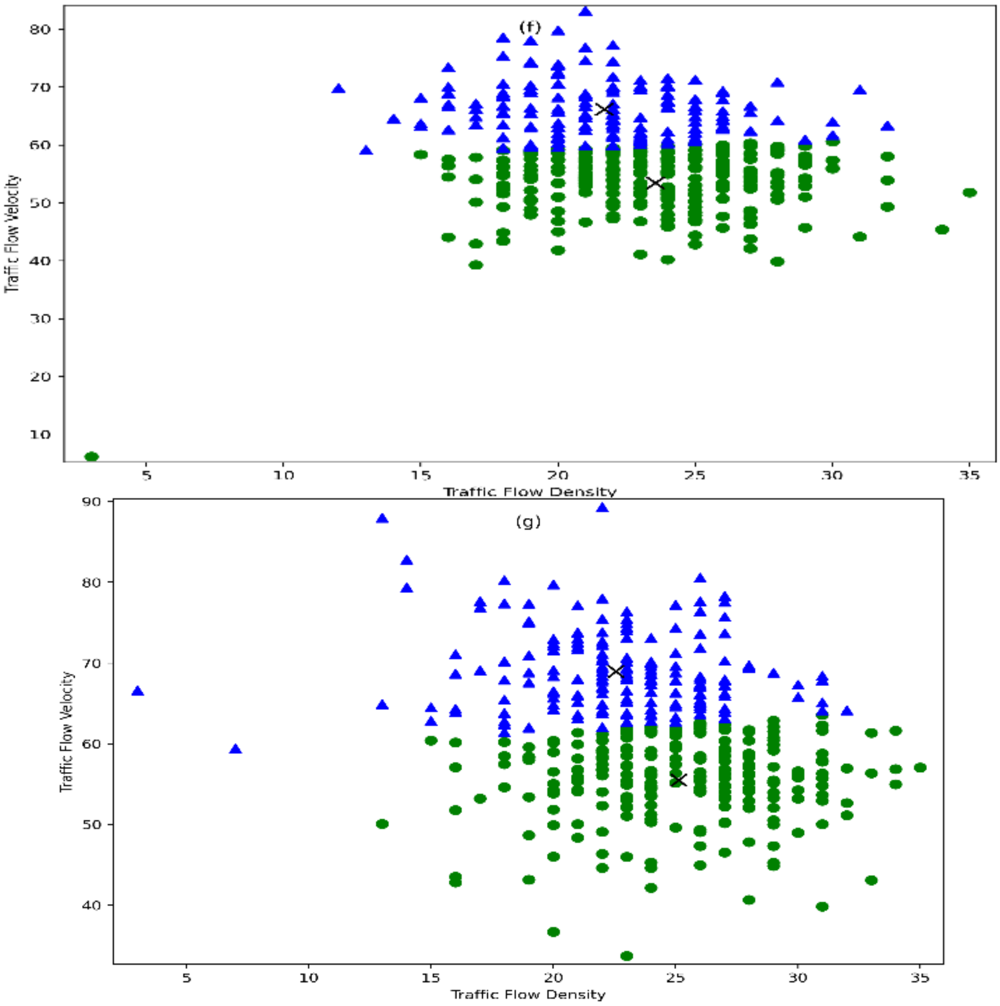


Figure 4. Traffic flow clusters on the under observed road section from 9:00 AM–4:00 PM on **(a)** Monday; **(b)** Tuesday; **(c)** Wednesday; **(d)** Thursday; **(e)** Friday; **(f)** Saturday; **(g)** Sunday.

The data presents the centroids of vehicle count, average speed, and the average number of vehicles for each day of the week, serving as key indicators for delineating traffic flow states. By analysing these centroids, we can identify prevailing traffic conditions and categorize them into distinct flow states, ranging from free-flowing to congested. Variations in vehicle count and average speed heavily influence this classification; higher centroid values for vehicle count coupled with lower average speeds may indicate congestion or slower-moving traffic, while lower vehicle counts and higher average speeds imply smoother, free-flowing conditions. This thorough examination of traffic flow patterns across different days provides insights into the variability of traffic flow states over the week, thereby facilitating informed decision-making for traffic management and optimization strategies.

4.4. Multiple linear regression

In this study, multiple linear regression analysis was employed to investigate the relationship between the average traffic flow speed per minute and the individual average speeds for cars, motorbikes, buses, and bicycles per minute. The regression equations derived for each day of the week are presented below:

Where $x1,\ x2,\ x3,\ x4$ represent the average speed of cars, motorbikes, buses, and bicycles, respectively.

$$Average\ Traffic\ Flow\ Speed\ (Monday) = 0.798x1 + 0.012x2 + 0.103x3 - 0.056x4$$

$$Average\ Traffic\ Flow\ Speed\ (Tuesday) = 0.891x1 + 0.019x2 + 0.036x3 - 0.113x4$$

$$Average\ Traffic\ Flow\ Speed\ (Wednesday) = 0.842x1 + 0.008x2 + 0.098x3 - 0.008x4$$

$$Average\ Traffic\ Flow\ Speed\ (Thursday) = 0.863x1 - 0.006x2 + 0.058x3 - 0.369x4$$

$$Average\ Traffic\ Flow\ Speed\ (Friday) = 0.815x1 + 0.028x2 + 0.103x3 - 0.048x4$$

$$Average\ Traffic\ Flow\ Speed\ (Saturday) = 0.828x1 + 0.014x2 + 0.080x3 - 0.438x4$$

$$Average\ Traffic\ Flow\ Speed\ (Sunday) = 0.856x1 + 0.021x2 + 0.038x3 - 0.596x4$$

The coefficients associated with each predictor variable provide insights into their respective impact on the average traffic flow speed. A positive coefficient indicates a positive relationship, signifying that an increase in the average speed of a particular vehicle type leads to a corresponding increase in the average traffic flow speed. Conversely, a negative coefficient suggests an inverse relationship, implying that an increase in the average speed of a specific vehicle type results in a decrease in the average traffic flow speed.

These findings offer valuable insights into the factors influencing traffic flow dynamics, thereby informing urban transportation planning and management strategies. For instance, the regression results reveal an inverse relationship between bicycles and the overall average traffic flow speed. Notably, while motorbikes significantly outnumber buses, the impact of bus speed on average traffic flow speed far surpasses that of motorbikes, as evidenced by their respective coefficients in the regression equations. This discrepancy underscores the aggressive driving behaviour of buses, where they consistently travel at speeds notably higher than the average traffic flow speed, as depicted in **Figure 3**.

4.5. Discussion

The application of K-Means clustering in urban traffic analysis marks a significant evolution from traditional methodologies that predominantly rely on simpler statistical techniques or manual observations. Traditional traffic analysis methods, such as regression models or time-series analysis, often focus on aggregate measures like vehicle count or average speed, which can overlook more intricate patterns. In contrast, K-Means clustering facilitates a data-driven approach that uncovers hidden patterns by segmenting traffic flow data into distinct clusters based on underlying characteristics. This shift not only provides a deeper understanding of traffic dynamics but also introduces a level of granularity and precision that traditional methods often lack.

When comparing this study to prior research, there are both similarities and important distinctions. For instance, studies like [2] and [5] also employed K-Means clustering to identify traffic patterns, with the former evaluating different clustering techniques and the latter focusing on offline traffic data to classify workday and weekend patterns. In our research, we extend the clustering approach by not only identifying peak and off-peak traffic flow conditions but also incorporating vehicle types and their impact on traffic density and velocity. Unlike the work by [5], which yielded a fixed number of clusters across different days, our analysis reveals dynamic traffic clusters that vary over weekdays and weekends, reflecting more nuanced flow dynamics. Moreover, studies like [6] and [7] introduced hybrid clustering methods to account for temporal variations, whereas our work focuses on the pure K-Means approach while acknowledging its limitations, particularly the sensitivity to outliers and the assumption of spherical clusters.

In line with [3] and [6], who introduced enhanced K-Means algorithms to handle large datasets and spatial-temporal information, our study also demonstrates that K-Means can effectively classify traffic flow states, especially when preprocessing techniques like outlier removal and parameter tuning are applied. However, while these studies have incorporated advanced variations like fuzzy logic and hybrid algorithms, our focus remains on the practical application of traditional K-Means clustering to offer a simpler yet insightful framework for understanding urban traffic dynamics.

4.6. Practical implications

The findings of this study carry significant practical implications for traffic management strategies in urban areas. By leveraging the insights gained from K-Means clustering, transportation authorities can develop targeted interventions to alleviate congestion, optimize road usage, and improve overall traffic flow efficiency. For example, the identification of congestion hotspots and peak traffic periods can inform the deployment of traffic control measures such as signal timing adjustments or lane management strategies [12,13]. Additionally, the clustering analysis can aid in the design of more efficient public transportation routes and the allocation of resources for infrastructure improvements.

4.7. Recommendations for implementation

To effectively translate the findings of this study into actionable strategies, it is imperative for transportation authorities to adopt an integrated approach that combines data analytics with stakeholder collaboration and community engagement. Establishing partnerships with local governments, transportation agencies, and community organizations can facilitate the implementation of targeted traffic management initiatives and ensure their acceptance and effectiveness. Furthermore, investing in data collection infrastructure and analytics capabilities will be crucial for sustaining ongoing monitoring and evaluation of traffic management interventions. Lastly, continuous refinement and adaptation of strategies based on real-time data feedback and evolving traffic patterns will be essential for maintaining the effectiveness and relevance of traffic management efforts in dynamic urban environments.

5. Conclusion

This research delved into the intricate dynamics of urban traffic flow through a multi-faceted analysis encompassing descriptive statistics, cluster analysis, and multiple linear regression. The findings unearthed distinct traffic patterns across different days of the week, shedding light on the temporal variability of traffic flow states. Through K-Means clustering, we identified heterogeneous clusters representing diverse traffic conditions, offering valuable insights into congestion dynamics and traffic behaviour variations. Additionally, the multiple linear regression analysis elucidated the nuanced relationship between vehicle types and average traffic flow speed, highlighting the differential impact of various vehicles on overall traffic dynamics.

Our study significantly contributes to the field of traffic flow analysis by providing comprehensive insights into urban traffic dynamics. By synthesizing data-driven methodologies and techniques, we offer a deeper understanding of traffic patterns and their underlying determinants. The identification of distinct traffic states and their temporal variations facilitates informed decision-making for urban transportation planning and management, thereby contributing to the optimization of traffic flow efficiency and congestion alleviation strategies.

Despite the valuable insights gleaned from this study, certain limitations warrant acknowledgment. The analysis primarily focused on a specific urban road segment, limiting the generalizability of the findings to broader traffic contexts. Future research could explore additional road segments and urban areas to enhance the robustness and applicability of the findings. Moreover, while our study delved into the relationship between vehicle types and traffic flow speed, further investigations could incorporate additional factors such as weather conditions, road infrastructure, and driver behaviour for a more comprehensive understanding of traffic dynamics. Additionally, the integration of real-time data streams and advanced machine learning algorithms could enhance the predictive capabilities of traffic flow models, paving the way for more adaptive and responsive traffic management systems.

Author contributions: Conceptualization, AMS and KSK; methodology, KSK; software, KSK; validation, AMS, KSK and ZHK; formal analysis, KSK; investigation, AMS; resources, KSK; data curation, KSK; writing—original draft preparation, KSK; writing—review and editing, AMS; visualization, AMS; supervision, KSK; project administration, KSK; funding acquisition, KSK. All authors have read and agreed to the published version of the manuscript.

Data availability statement: Data available on request from the corresponding author upon reasonable request.

Conflict of interest: The authors declare no conflict of interest.

References

1. Khan, Khattak, Khan et al. Edge computing for effective and efficient traffic characterization. Sensors. 2023; 23(23): 9385.
2. Saha, Tariq, Hadi et al. Pattern recognition using clustering analysis to support transportation system management, operations, and modeling. Journal of Advanced Transportation. 2019; 2019(1): 1628417.
3. Li, Zhou, Gu et al. A novel K-means clustering method for locating urban hotspots based on hybrid heuristic initialization. Applied Sciences. 2022; 12(16): 8047.
4. Khattak, Minallah, Khan et al. Sensing technologies for traffic flow characterization: From heterogeneous traffic perspective. Journal of Applied Engineering Science. 2022; 20(1): 29–40.
5. Song R, Yang H. Clustering and understanding traffic flow patterns of large scale urban roads. In: Proceedings of the 2021

International Conference on Control, Automation and Information Sciences (ICCAIS); 2021.

6. Li X, Gui J, Liu J. Data-driven traffic congestion patterns analysis: A case of Beijing. Journal of Ambient Intelligence and Humanized Computing. 2023; 14(7): 9035–9048.
7. Muntean MV. Identifying critical traffic flow time slots using k-means and decision trees. In: Proceedings of the 2020 IEEE 10th International Conference on Intelligent Systems (IS); 2020.
8. Chu HC, Wang CK. Using K-means algorithm for the road junction time period analysis. In: Proceedings of the 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST). 2017.
9. Daniswara FA, Gunawan P. Simulation of transport problem with clustering velocity-density function. In: Proceedings of the 2020 8th International Conference on Information and Communication Technology (ICoICT). 2020.
10. Esfahani RK, Shahbazi F, Akbarzadeh M. Three-phase classification of an uninterrupted traffic flow: a k-means clustering study. Transportmetrica B: transport dynamics; 2018.
11. Gao, Chen, Chen Z et al. Traffic state recognition of urban expressway based on K-means clustering and AdaBoost-DS. In: Proceedings of the Sixth International Conference on Traffic Engineering and Transportation System (ICTETS 2022). 2023.
12. Zeb, Khattak, Rehmat et al. HetroTraffSim: A macroscopic heterogeneous traffic flow simulator for road bottlenecks. Future Transportation. 2023; 3(1): 368–383.
13. Khattak KS, Khan ZH. Evaluation and Challenges of IoT Simulators for Intelligent Transportation System Applications. Science. Engineering and Technology. 2024; 4(1): 94–111.