International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(3):1528-1545

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i03.04170

Mining Approach for Traffic Congestion Detection

Sekar Kidambi Raju[1], VVenkataraman[2*], V Rengarajan[3], G Sathiamoorthy[2], Ganesh Karthikeyan Varadarajan[1], Raj Anand Sundaramoorthy[1]

[1]School of Computing, SASTRA Deemed University, India, [2]Department of Mathematics, SASHE, SASTRA Deemed University, Thanjavur, India, [3]School of Management, SASTRA Deemed University, India. *Corresponding Author's Email: mathvvr@maths.sastra.edu

Abstract

To address the issue of traffic congestion, a new unsupervised incremental learning strategy has been proposed to identify and profile traffic congestion in metropolitan areas. The proposed model can effectively analyze and anticipate traffic situations in urban areas and improve traffic efficiency. Additionally, a clustering method based on Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points To Identify the Clustering Structure (OPTICS) has been suggested to cluster accident-prone regions. This method outperforms other algorithms based on synthetic and actual datasets. Furthermore, this research aims to determine and forecast the traffic flow of the road network using GPS data from floating cars. The traffic condition of the metropolitan road network is determined using an emerging hot spot analysis tool to look for diverse patterns of hot spot formation. Using the time series clustering approach, the road network is partitioned into groups with comparable spatiotemporal features. The three-time series forecasting models are also applied to estimate traffic operation status, and the proposed model outperforms the existing methods. Finally, this study proposes efficient and effective methods for managing traffic congestion in urban areas. These methods can help identify, assess, and forecast traffic congestion levels, crucial for improving traffic efficiency and ensuring public safety.

Keywords: Adjusted Information, Affinity Propagation, Clustering Structure, Density-Based, Hotspot Analysis, Spatial Clustering.

Introduction

In today's world, traffic congestion is one of the difficult scenarios faced in metropolitan cities. There is a direct correlation between the time people spend caught in traffic jams and the pollution in metropolitan areas. Traffic congestion is a major issue in many cities worldwide, leading to significant delays and negative impacts on health and safety. Traditional approaches to detecting congestion often rely on data from fixed sensors, which can be costly and have limited coverage. A new experimental approach using Particle Swarm Optimization (PSO) has been developed to address these limitations. PSO is a metaheuristic optimization algorithm that is effective in solving complex problems, and it has recently been applied to traffic congestion detection. Drivers lose 97 hours yearly due to traffic congestion (1). The increasing mobility of today's society has led to a rise in traffic accidents. Road traffic accidents are a public health and social issue because of the number of injuries and deaths they cause. To properly allocate resources for enhancing safety, adding value-added data to accident hotspots and analyzing their actions is essential. Using GIS and other value-added data forms, it is possible to acquire a better knowledge of the indicators of causal effects by finding road accident hotspots. GIS is a system for storing, retrieving, and analyzing geographic data (2). Recent academic research evaluates and uses initial data effectively to gain vital information (3). The theoretical and practical relevance of accurately analyzing Urban traffic congestion's spatiotemporal characteristics and fully exploring its complex operational regularity is significant. It has been over two decades since the Autoregressive Integrated Moving Average (ARIMA) was first used to forecast traffic flow issues (4). Spatial data management and analysis are essential for several applications. They are based on real-time applications, such as satellite imagery processing, X-ray crystallography, etc. It necessitates the use of automated knowledge discovery (5). The clustering technique is

(Received 21st February 2025; Accepted 12th July 2025; Published 30th July 2025)

developed based on density to organize geographical data better. The construction of spatial dataset algorithms must address the following fundamental concerns (6).

- There would be a need for more expertise in spatial data. Hence, the algorithm must work in an unsupervised way.
- No centroid or grids may be employed since the groupings will take on random forms.
- Databases are massive. Because clustering necessitates working with all of the available data, the algorithm should use processing approaches that are both quick and efficient.
- The proposed experimental approach using particle swarm optimization can significantly improve the accuracy of road traffic congestion detection modeling and analysis. This is because particle swarm optimization can efficiently search for the optimal combination of parameters to detect and analyze traffic congestion in real time accurately.
- The proposed approach can also improve road users' health and safety. Accurately detecting and analyzing traffic congestion can help reduce the risk of accidents caused by sudden stops or unexpected changes in traffic patterns.

The DBSCAN, a typical clustering technique, tries to mine datasets for potentially relevant information. Clustering analysis technology relies heavily on the DBSCAN algorithm. However, despite the DBSCAN algorithm's numerous benefits in clustering, it also has a few drawbacks. DBSCAN's minpts and eps will vary depending on the analyzed datasets. Traditional clustering methods may be divided into three categories, each with its own principles: partition-based and hierarchical (7). By reducing the total distances between points in a cluster, a partition-based approach is used to assign points to clusters while at the same time increasing the distances between the points in different clusters (8). The cluster count must be determined in advance to utilize this strategy.

Because of its excellent computing cost, the hierarchical clustering approach needs to be revised for substantial point clouds. Algorithms for detecting traffic jams have long relied on GPS data. The speed performance index (SPI) was developed using floating vehicle data to analyze traffic conditions. The journey duration is determined by analyzing the vehicle's GPS trajectory information (9). Traffic accidents and temporary traffic management are the most common causes of occasional congestion, but regular traffic flows are more likely to produce persistent congestion (10). Unlike occasional congestion, recurring congestion has particular laws for generating, spreading, and dispersing. The congestion is always generated and dissipated simultaneously on the same day, while the position and propagation direction are highly comparable in the spatial dimension. This research examined the spatial and temporal patterns of recurring workday congestion. The recurring congestion is controlled in the road network by continually monitoring its spatiotemporal pattern (11).

The novelty of this work is that they propose a broad approach to controlling traffic congestion in big cities using improved unsupervised learning schemes and clustering techniques. This work presents a strong method for clustering accident-prone areas to analyze and predict traffic events in metropolitan cities using DBSCAN and OPTICS. Besides, accurate determination and forecast of the traffic flow is possible due to GPS data of floating cars, which implements hot spot analysis and time series clustering. Using three-time series forecasting models takes the prediction to higher levels than previously achieved, with a significant boost from applying the models. Hence, this multiple perspective study develops and evaluates traffic congestion indicators and helps design the future tactics of increasing traffic efficiency and safety in urban areas.

The novelty of our method lies in a new blend of mining approaches, detection mechanisms, and type of data. In contrast with earlier studies that have a tendency to use static sensor readings and traditional clustering algorithms, our model combines Particle Swarm Optimization (PSO) with state-of-the-art density-based clustering algorithms such as DBSCAN and OPTICS to deal with non-conventional traffic patterns in an unsupervised fashion. In addition, we use the latest hotspot analysis and floating vehicle GPS data time-series clustering to model real-time spatiotemporal traffic flow dynamics. Not only does the integration improve the accuracy of congestion detection, but it also improves forecasting performance by utilizing several time-series models (CFF, ESF, and FBF), providing a more dynamic and predictive urban traffic congestion management system compared to past literature.

Incorporating the proposed model is also critical because the problem of traffic congestion in urban areas is complex, and the analysis involved needs to employ the best of different analytic and clustering methods. City traffic environments are dynamic and complex, with factors that differ in space and time and cannot be addressed without special techniques for assessment and prognosis. When using DBSCAN and OPTICS in conjunction, clustering the accident-prone areas yields a less streamlined yet more accurate method of detecting the critical hotspots that cause congestion. Also, the model can predict the traffic flow using GPS data and the emerging hot spot analysis in real-time, which is vital in timely interventions. However, this documented integrated approach of a consultant strengthens and optimizes traffic flow management approaches and traffic management systems to become efficient while reducing the danger associated with the urban transport system.

The GPS trajectory records gathered using floating vehicles serve as the main source of traffic information in the analysis. It is real-time, high-resolution spatiotemporal data. In order to maintain data reliability and prevent noise, there was a strong preprocessing pipeline was implemented. This involved the elimination of error trajectory points that are beyond the acceptable speed limits, rejection of duplicate or in-temporally incompatible records, and stationary points due to extended parking or idling. Invalid data out of the research scope were eliminated, and spatially contiguous records with equal time intervals were kept. Additionally, computations of average speeds and traffic performance measures were conducted upon rastering the road network into equal grids. Such strict cleaning and preprocessing ensured input data were representative, noise-free, and accurate, and a good representation of real traffic dynamics.

Experts and academics in traffic accident prediction employ clustering approaches extensively, allowing pedestrians to get fast but accurate and less costly safety delivery (12). Spatial clustering has been addressed in several types of research. Some of the most significant contributions to this field are analyzing some geographic locations with a higher incidence of road accidents. The riskiest regions on the road can be analyzed using the K-means algorithm (13). Several clustering techniques using internal and external measures are analyzed. Silhouette, Davies-Bouldin, and Calinski-Harabasz measures were employed to compare clusters with the researchers. Using these methods, a researcher may quickly choose the best cluster for analyzing data on traffic accidents (14).

Additionally, the researchers did not contrast clustering methods according to how quickly they ran. Run-time measurements may be helpful in the selection of an algorithm for non-experts. This work covers a wide range of topics that might benefit a novice researcher. K-means and DBSCAN clustering algorithms were compared for their performance (15). The researchers compared DBSCAN and k-means based on "run time" and accuracy. Even yet, when it came to showing which clustering method was superior based on the accuracy and complexity. The research compared the performance of several algorithms across various platforms (16). However, this research explores whether the method is superior for analyzing data from road accidents.

Diversity is an effective method for improving information retrieval (IR) (17). To account for the query-to-document lexical discrepancy, a probabilistic model was developed. The real-time and predictive data may be used to anticipate traffic congestion (18). They devised two assessment frameworks to test the proposed method's accuracy and computing efficiency. The natural and typical bus travel times are calculated using the vehicle's GPS trajectory information with the classification and congestion index (19). The road network's traffic status is exactly identified and rapidly disseminates crucial data that provides the logical flow of the driving paths to reduce traffic congestion. Researchers studying traffic congestion use the exact trajectory information.

The document retrieval system is developed using passage-based information with increased performance (20). A rating of passages generated in response to the query was used to evaluate the application of learning-to-rank-based document retrieval techniques. Based on the floating automobile data, a unique speed performance index (SPI) is developed to evaluate current traffic circumstances (21).

This work has several important advantages, such as the possibility of employing further and more advanced clustering concepts in connection with real-time data analysis to predict transportation congestion in large established urban areas effectively. Application of DBSCAN and OPTICS for identifying problematic zones in terms of raising the number of accidents increases the accuracy of congested area detection. Integrating the GPS data of floating cars will facilitate real-time traffic flow monitoring and prediction. Besides, the model's flexibility to changes in spatiotemporal traffic features allows for better response to traffic and subsequent control. Such combined strengths make it possible to produce a more general and anticipative approach to controlling urban traffic, which improves traffic flow, decreases congestion, and increases public safety.

The work has some shortcomings, but before listing them, it is worth stating that these limitations are typical of many similar studies. One vulnerability is that using DBSCAN, OPTICS, and real-time data implementation may be time-intensive and computationally extensive and may need to be more efficient in huge, dynamic city areas. Also, since the GPS information is collected only from float cars, there may be systematic biases if the information collected is not random or if the float cars' coverage is not random and comprehensive. Further, the approach may fail to consider such factors as weather or any other unexpected event that can cause a shift in traffic flow patterns. These constraints point right back to recommendations for enhancement to scale up while maintaining the representativeness of data and factoring resilience against disruptive events.

Methodology

The mining technique adopted in this work is based mostly on unsupervised clustering supported by metaheuristic optimization methods for optimizing clustering performance. Specifically, the method employs Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points To Identify the Clustering Structure (OPTICS) for identifying accident-prone areas and congestion clusters label-free. These are further enhanced with Particle Swarm Optimization (PSO) for enhancing cluster parameters towards achieving increased accuracy and responsiveness to a variety of traffic scenarios. The approach integrates time-series forecasting models and clustering models to identify spatiotemporal congestion patterns. The approach can thus accurately be termed as an unsupervised, multi-technique clustering framework for real-time urban traffic mining and prediction.

One of the most often utilized metaheuristic techniques for solving optimization issues is the Particle Swarm Optimization (PSO). Because PSO optimizes in a time-sensitive way, results from PSO may be retrieved in real time (22). This is a huge benefit. A multi-start PSO is used to accomplish density-based clustering, as illustrated in Figure 1. There are three primary steps to the proposed method: cluster thresholds, multi-start PSO for cluster node identification, and cluster construction.

This model collects and analyzes real-time data from traffic cameras and sensors to detect traffic congestion. The PSO algorithm is then applied to optimize traffic signal timings and control traffic flow, reducing congestion and minimizing the risk of accidents. The model also includes a predictive analytics component, which uses historical traffic data to predict congestion and recommend proactive measures to prevent it. The proposed model can improve traffic flow, reduce travel time, and enhance road safety, resulting in significant economic and social benefits.

In this experimental approach, utilizing Particle Swarm Optimization for road traffic congestion detection modeling and analysis is a promising solution that can enhance road health and safety by optimizing traffic signal timings and controlling traffic flow based on real-time and historical data. The overall framework of the research study is depicted in Figure 1.

Nodes are identified as belonging to a cluster based on the node's density in a particular region. The first step in the clustering process is to define the threshold level at which a point may be included in a cluster based on the data. The maxdist and minpts are critical features that must be considered while calculating the threshold. A node may only be included in an available cluster if it is within a certain distance (maxdist) of another node. Nodes must be surrounded by a minimum of minutes of nodes before they are considered part of the cluster (23). This characteristic is known as the direct density accessible for nodes n1 and n2. This cluster includes both of these sites.

This research is divided into three parts: identifying and forecasting urban traffic conditions and mining spatiotemporal patterns using various data sources. Many methods, such as processing trajectory data, are developed as a grid model, and road traffic performance indicators must be used to determine whether or not there is traffic congestion (24). Emerging hot spots and time series clustering are part of spatiotemporal pattern mining. The space-time cube and time-series approach predict traffic congestion on urban roadways.

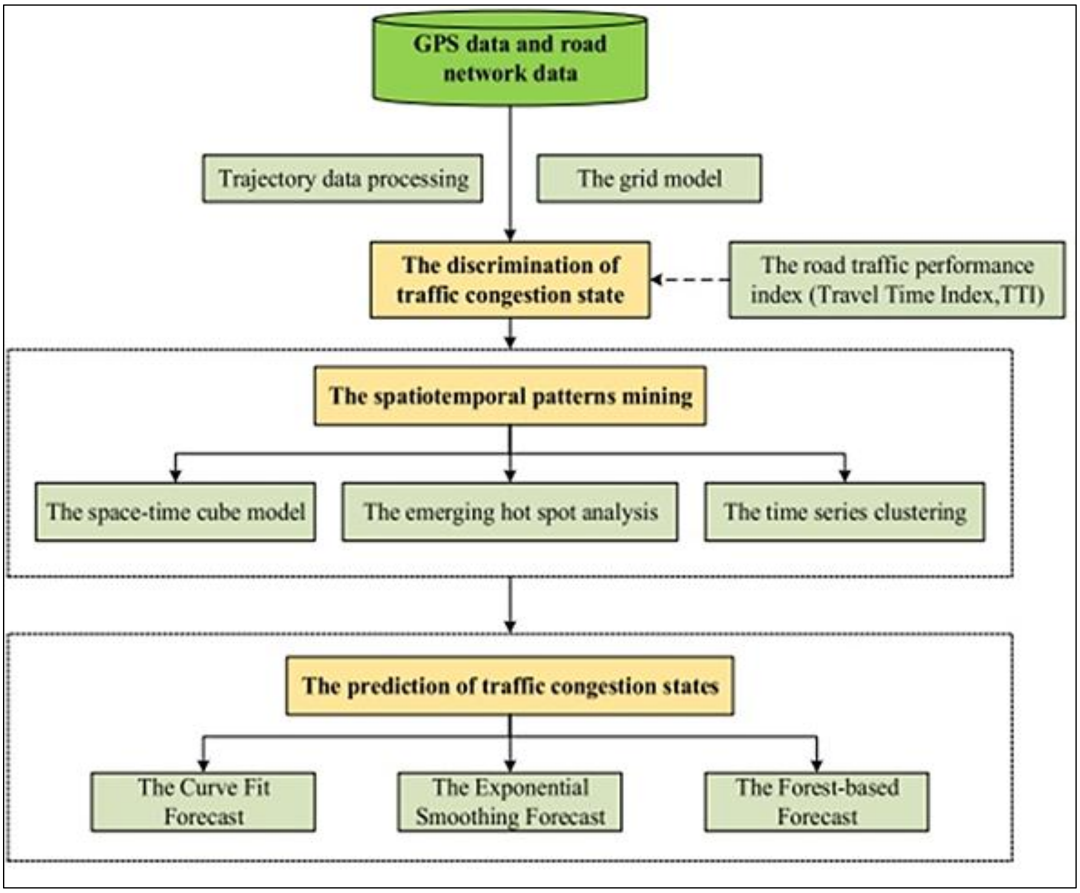
GPS data and road network data
Trajectory data processing
The grid model
The discrimination of traffic congestion state
The road traffic performance index (Travel Time Index,TTI)
The spatiotemporal patterns mining
The space-time cube model
The emerging hot spot analysis
The time series clustering
The prediction of traffic congestion states
The Curve Fit Forecast
The Exponential Smoothing Forecast
The Forest-based Forecast

Figure 1. Overview of The Framework for The Study

Trajectory Data Processing

Crucial phases in trajectory data processing include eliminating inaccurate and speed information. Figure 2 displays the data processing for trajectory data. The central processing units are listed below:

- Remove inaccurate data: The trajectories beyond the research's purview are deleted, and any trajectories that exceeded 100 km/h when the vehicle was in an abnormal state
- Precise invalid time-related data: A chronological sort of each vehicle's trajectory points was performed, and any points with multiple time stamps were purged. A trajectory point will be considered an isolated point and eliminated if the temporal separation between it and two neighboring trajectory points is more than 5 minutes (25).
- Clear invalid parking points: After 3 minutes, it was found that cars idling on the side of the road, waiting for passengers, were providing false data. Therefore, those trajectory points were eliminated.
- Compute the trajectory's average speed: The whole road network was broken down into equal-sized grids. After that, the latitude and longitude of each grid were rasterized, and the grids were assigned numbers. The final step is calculating each grid's path's average speed (26).
- Compute the traffic performance index for the road: The TTI for each grid was calculated based on the vehicle's speed.

Spatiotemporal Traffic Patterns and Space-Time Cube Model

ArcGIS Pro was utilized to mine spatiotemporal patterns in this investigation. By using time series analysis, the model creates the visual representations of spatiotemporal data. Hot and cool spots may be identified over time using new algorithms that use multidimensional data sets. The GPS data at one-hour intervals is analyzed using the hot spot analysis approach (27). The places are divided into separate groups in a space-time cube (STC), each having the same time-measure characteristics.

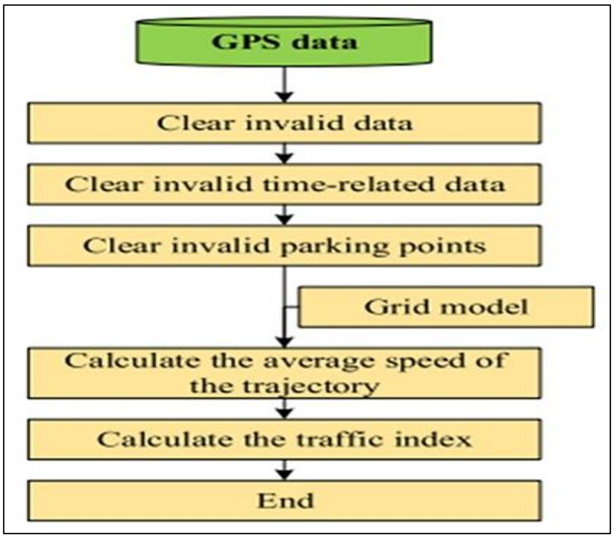
GPS data
Clear invalid data
Clear invalid time-related data
Clear invalid parking points
Grid model
Calculate the average speed of the trajectory
Calculate the traffic index
End

Figure 2: Trajectory Data Processing

The multidimensional cube of the raster layer was used to develop the multidimensional cubical structures. The STC will have the exact temporal and spatial resolution as a multidimensional raster. The raster cells dimension is converted from each space-time bin. The columns, rows, and time steps define the structure of the STC. To obtain the bin count in the STC, multiply the row and column counts by the number of time steps. Rows and columns define the geographical and temporal ranges of the cube. The cube of space and time is seen in Figure 3.

The suggestion acknowledges that the spatial representation of how effective the model is useful. To that end, congestion time-series plots and heatmaps are included in the study to illustrate the traffic congestion dynamics across locations and time periods. These are geo-referenced using ArcGIS Pro and the space-time cube (STC) technique, in which one can determine sustained, occurring, and transient hotspots. Time-series clustering outputs also illustrate that the congestion patterns change over time and provide better insight into both spatial density prediction and temporal change of metropolitan traffic. These visualization resources improve the interpretability and usefulness of the proposed model.

Hotspot Analysis

Watch out for hot and cold spots in the space-time cube that may be constant, sporadic, or oscillating. The Getis-Ord Gi statistic is applied, which considers every bin's information among its neighbors to calculate the degree of clustering. To determine which bins are included in a particular inquiry neighborhood, the method looks at neighborhood bins that fit under the defined conceptual link of Geographic distances. Bins must be included if they have been present in the same region for N neighboring time steps. Following that, two distinct analyses are performed: (1) the clustering strength of high and low values in adjacent bins was determined by examining the individual bins separately. These results are computed with Kendall’s measure using each bin's p-value, Z-score, and category in STC (28). Emergent hot spot analysis is organized in terms of structure, as shown in Figure 4.

Series Clustering

Series clustering identifies distinct network congestion where each cluster's constituents have comparable time measures. The kinds of congestion are grouped due to their shared historical background similarities. The image shows a grouping of time series. The location measures of time series inside and outside every group experience increasing levels of clustering when the space-time cube is partitioned into distinct groups. The correlation strategy groups the time series whose values fluctuate in lockstep. To evaluate the similarity of time series, this method uses statistical correlation that determines the differences between two time series. The positions of a cube were grouped according to the notion of similarity over time using the Partitioning Around Medoids (PAM) or K-medoids method. Determining how many clusters to utilize in a clustering procedure is challenging. An F test will determine the optimal number of clusters in this technique. When the cluster becomes closer to the time series, F values increase, showing that clustering is effective. Hostop classification and its significance are shown in Figure 5.

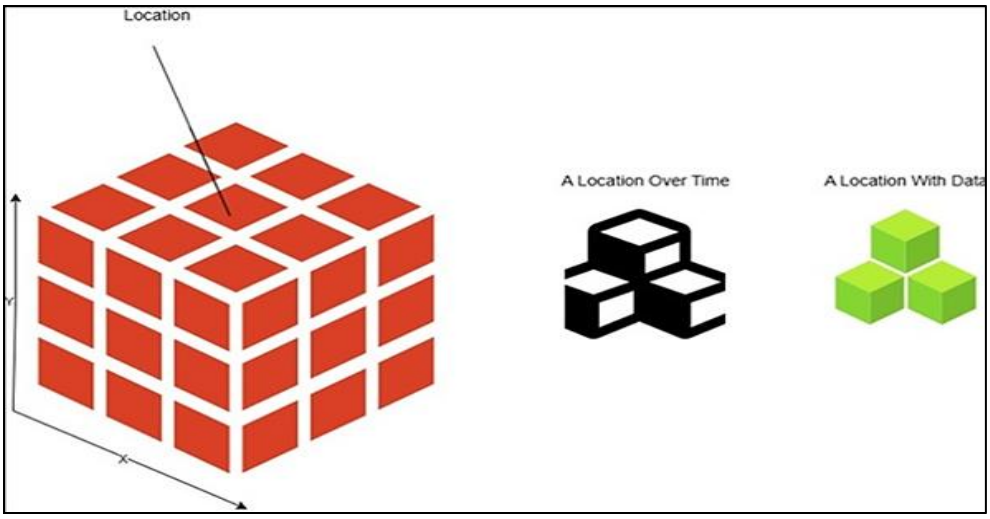

Figure 3: The Space-Time Cube

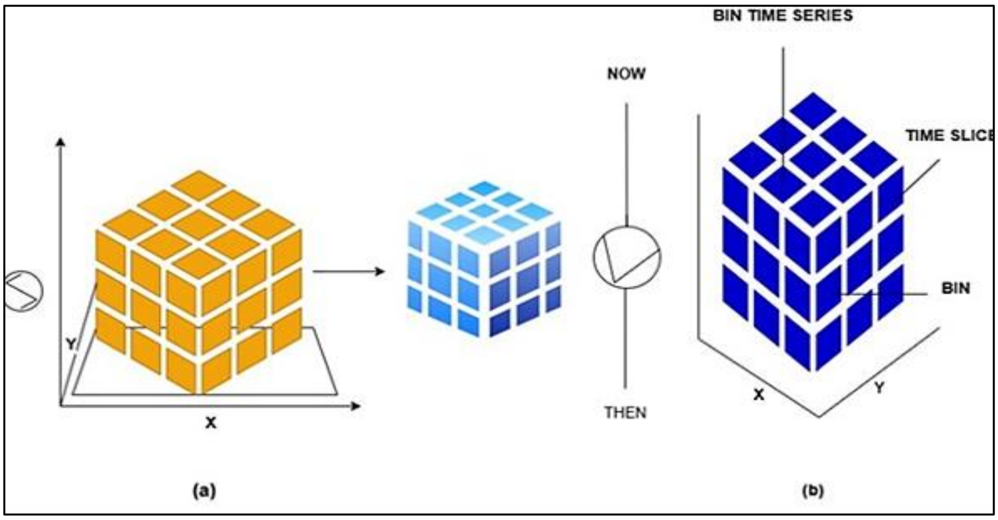

Figure 4: Hot Spot Trends

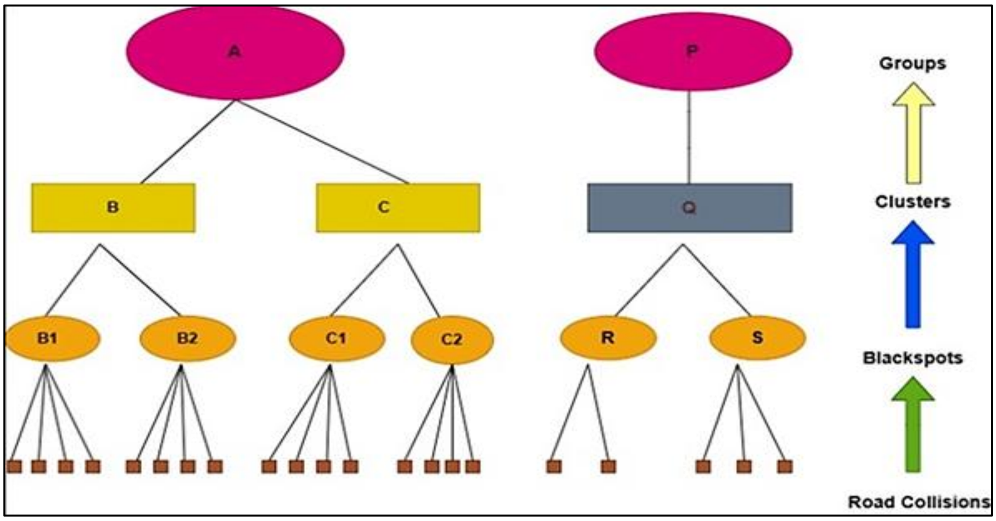

Figure 5: Hotspot Classification Method

Kernel Density Estimation

Many spatial tools have been created to better grasp how point patterns change geographically by estimating the Kernel densities (29). Kernel density estimation (KDE) has numerous benefits over statistical hotspots and clustering methods like K-means. For this strategy, the primary benefit is estimating the extent of an accident's danger. As a result of spatial dependence, there is an increased possibility of an accident near a designated cluster. By applying this density approach, it is feasible to create and homogenize an arbitrary geographic unit of analysis for the whole region, allowing for comparison and, eventually, classification (30).

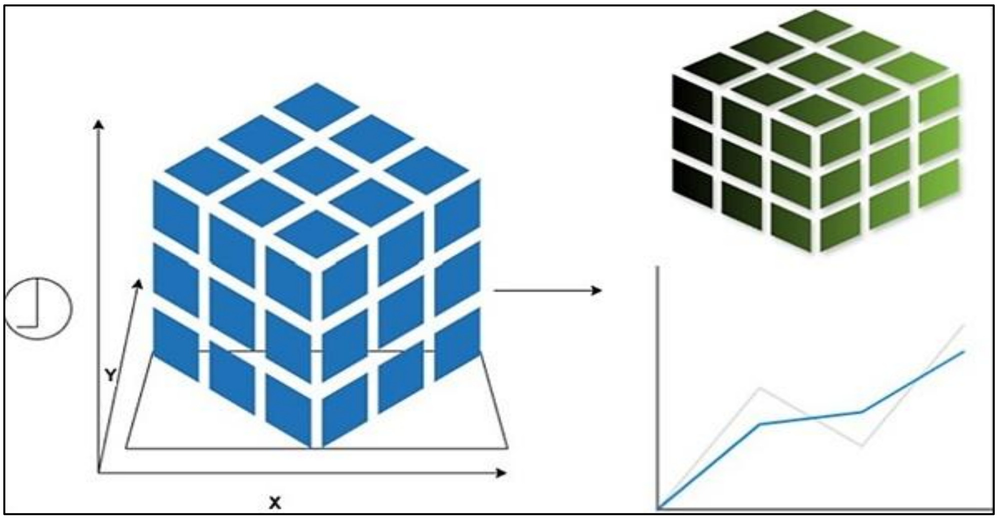

Figure 6: Time Series Clustering

In Kernel density estimation, each point is placed on the surface of a symmetrical polygon, the distance to a reference site is evaluated using a mathematical function, and the result for each polygon is summed. This technique is performed for each of the points on the compass. It is thus possible to estimate the density of the distribution of accident spots by adding the density estimates of each kernel. Time series clustering is presented in Figure 6.

$$f(x,y) = \frac{1}{nh^2}\sum_{j=1}^{n} \; K d_j C_h \qquad [1]$$

$$f(x) = \frac{1}{hn}\sum_{j=1}^{n} \; \frac{(x-x_j)}{h} K \qquad [2]$$

Assuming no outliers exist in the data, the density is estimated at the point (x, y) by multiplying the density estimate by the observation count (n) and the bandwidth h. To produce a smooth and continuous surface, kernel values are placed. The KDE approach is a method of determining bandwidth (the kernel) surrounding each place where the indication may be viewed. This applies a suitable function to the indicator's value at that particular moment. They add up all of these values, and even if there are no occurrences of the indicator variable, it yields a density estimate for the whole surface. There are two ways to determine density: the basic approach and the kernel method. Cell density values are computed by taking the number of features in the search area, considering the area's size, and then drawing a circle around each cell to determine how many cells there are. The neighborhood's radius influences the density map's appearance. If the circle's radius is enlarged, more feature points may be included, resulting in a more uniform density surface. The kernel approach uses a fixed number of cells to split the study area into sections. Instead of drawing a single neighborhood around each cell, a circular one is built around the accidental region instead of just one. A mathematical equation goes from 1 at every feature location to 0 at the boundary, as shown in Figure 7.

The particle velocity is updated using.

$$CapV_j = \{cr\frac{V_{max}}{\rho} - a\ if\ |V_j| < V_{min}\ V_j + r/c, otherwise \qquad [3]$$

The collision factor of a particle is deed using$cf\,cf_i = \frac{1}{1+e^{2-f(i)}}$ [4]

sigmoid operation $S(p_i)$ is given by

$$S(p_i) = 1/(1 + e^{-p_i}) \qquad [5]$$

Determine the updated velocities of particles using

$$V_i = wV_i + r_1\, c_1\, (C_{best} - C_i) + r_2\, c_2\, (C_{gbest} - C_i) \qquad [6]$$

The Department of Transportation and the Police collect this data, known as Stats19 data, which includes official accident information. To ensure the validity and reliability of the data collection, GPS information is recorded with the accident's position to a 10-meter precision. High-density grid cells (2290) were used to create the final surface. Many grid cells in densely populated city areas are clustered together, showing varying sizes. Both cell size and bandwidth (also called search radius) impact the KDE's performance. The bandwidth is critical for identifying the optimal density surface. The size of the hotspots may be influenced by the bandwidth used; in general, the more bandwidth used, the bigger the hotspots will be. Choosing the bandwidth and grid cell size for road accident density measurements is a guessing game due to the need for defined parameters. The search radius of 200 m is considered for this research since that's twice the size of our grid cell with the bandwidth at 200 m. Table 1 shows the environmental attributes and relative information in the relative information in context of the environment, referring to data or knowledge that is relevant or meaningful concerning a particular environmental attribute or issue.

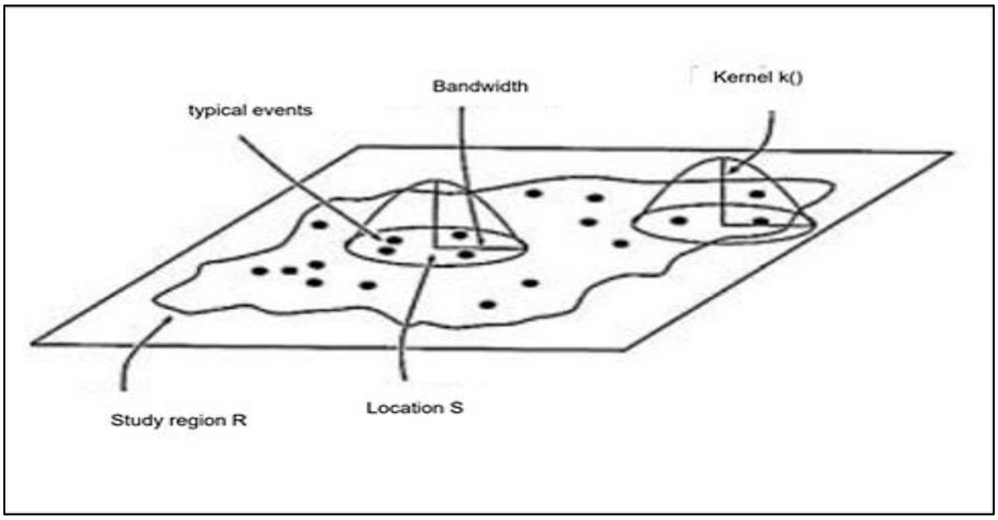

Figure 7: The Density Approach Based on the Quadratic Kernel Density

Table 1. Environmental Information

Attribute
Road Length Map
Length of Cycle Land
Pedestrian Crossings
City underground locations
Traffic lights
Bus stops
University & Schools
Speed cameras

Forecasting and Validation

Two models created using these methods are necessary to forecast the future of time series. Conversely, the prediction model is built to predict the time step's worth. A validation model is also utilized to confirm the predicted values. The time series data are fitted into the prediction model at each place in STC. Then, FRMSE is calculated for each time series using.

$$CapF_{RMSE} = \sqrt{\sum_{t=1}^{T} (c_i - r_i)^2/T} \quad [7]$$

T – time steps, c_t – curve value, and r_t – raw information at point t. The application of the forecast and validation models is shown in Figure 7. The method's original time series is evaluated using the FRMSE. The accuracy of the forecast model cannot be determined from this information. It typically fits the time series when extrapolating, but cannot produce reliable projections. The validation model offers an answer to this problem. A prediction model's ability to forecast future values for each time series is tested using the validation model. Each time series' last section is eliminated, and the forecasts from that data set are then used to build a prediction model. The difference between anticipated and raw values for time steps is expressed as the square root.

The Validation Root Mean Squared Error (VRMSE), a measure of the predictability utilized in the study, is computed. A validation model must be used if you wish to evaluate the precision of a prediction model, even when it isn't used for forecasting, to support model prediction where it is the prediction produced from time steps 1 through T-m, m, the count of time steps to be withheld for validation. Forecast and validation models are depicted in Figure 8.

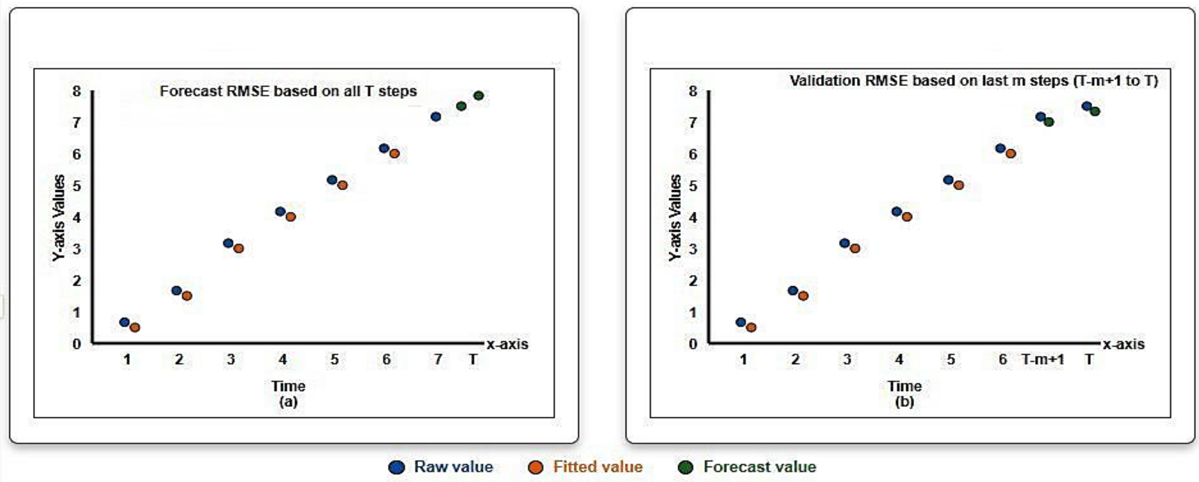

Figure 8: (A) The Forecast, **(B)** Validation Models

$$\text{Cap}V_{RMSE} = \sqrt{\sum_{t=T-m+1}^{T} (c_i - r_i)^2/m} \quad [8]$$

This paper covers all three facets—post-hoc trend analysis, real-time congestion detection, and prediction beforehand—to present an end-to-end system for the management of traffic congestion. Post-hoc analysis is conducted through hotspot mapping and trajectory clustering to detect patterns of congestion in the past. Real-time detection is accomplished through floating vehicles' GPS data that monitor real-time traffic status and feed into clustering algorithms (DBSCAN and OPTICS) for real-time congestion detection. In addition, sophisticated forecasting is obtained by applying time-series clustering and forecasting models (CFF, ESF, and FBF) to forecast future congestion levels based on spatiotemporal trends. It is a layered application that boosts responsiveness and accommodates short-term as well as long-term traffic management initiatives.

Results and Discussion

Reducing congestion on arterial highways is critical based on the network's congestion analyses. Building urban traffic microcirculation, increasing the density of the road network, and improving its capacity are essential. Pedestrian overpasses, safety islands, traffic signage, and line markings should all be highlighted as ways to keep roadways safe and convenient for everyone. Zebra crossings should also be redrawn. Shopping, catering, and entertainment establishments were found to be the most crowded areas during the evening rush hour. Side parking has become a prevalent problem in certain areas due to the need for more parking spaces nearby. This substantially affects road service levels and is critical to maximizing unused land usage and building more parking lots. To actively relieve parking issues, the shared parking policy is widely supported due to the various features of vehicle parking needs in various locations. The model increases public transportation use and service quality by optimizing the public transportation network. Spatiotemporal structures refer to patterns or configurations that evolve over space and time. These structures capture how certain phenomena or processes change and interact within a physical space across different periods in Figure 9. Time series clustering is a technique that groups similar time series data based on their patterns, trends, or other characteristics. By clustering time series data, we can identify groups of data that behave similarly over time, which helps simplify analysis, detect patterns, and make predictions in Figure 11. Clusters and hotspots refer to areas or groups where certain activities, events, or data points are concentrated. In the context of data analysis, a cluster is a group of similar items or occurrences that are close together either in a dataset or in a spatial area. A hotspot specifically refers to a region or location on the x-axis with a high concentration of activity or events on the y-axis, often indicating areas of special interest or concern in Figure 12.

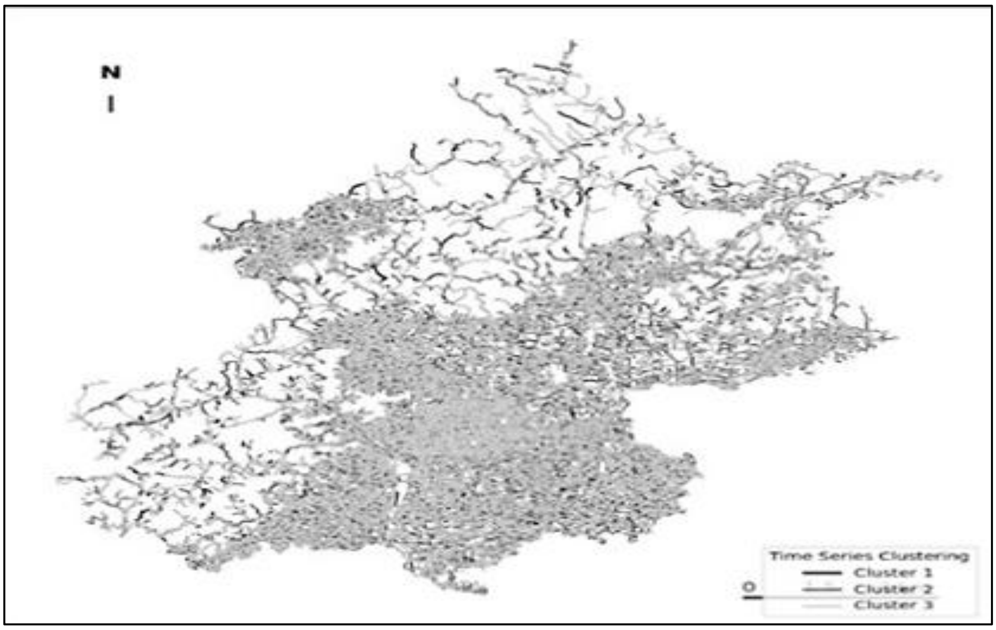

Figure 9: Spatiotemporal Structures

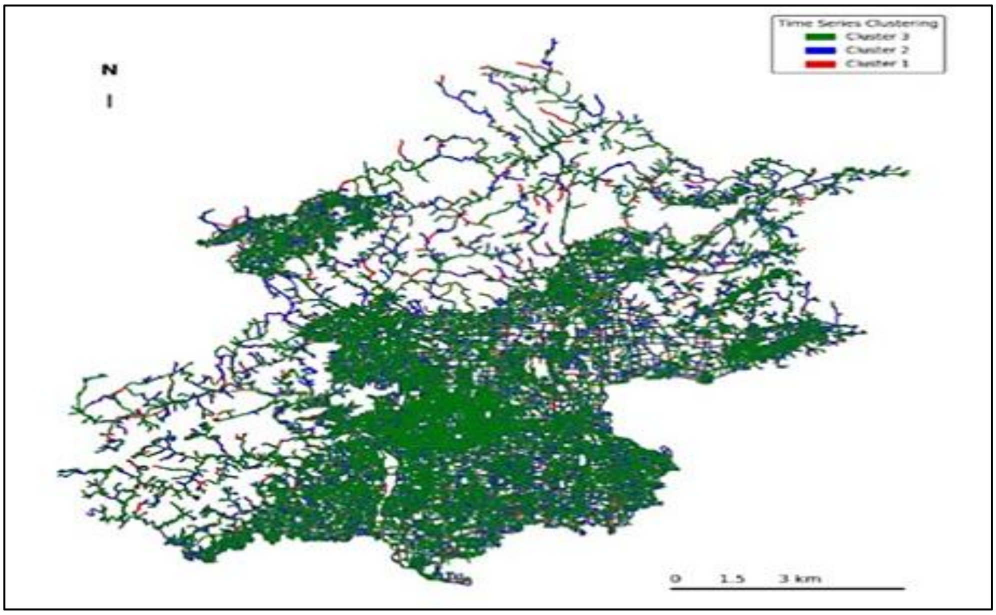

Figure 10: The Cluster-based Time Series

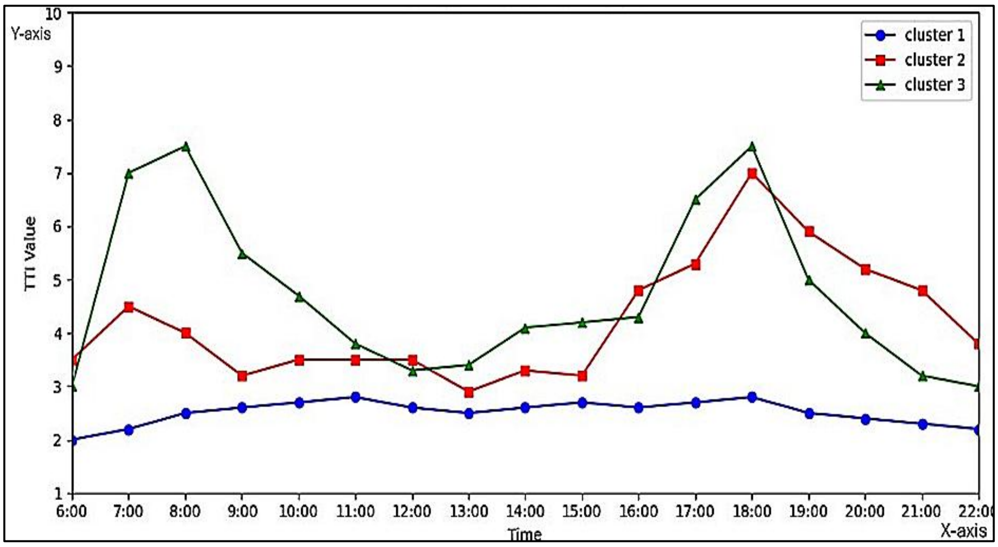

Figure 11: The Result of Time Series Clustering

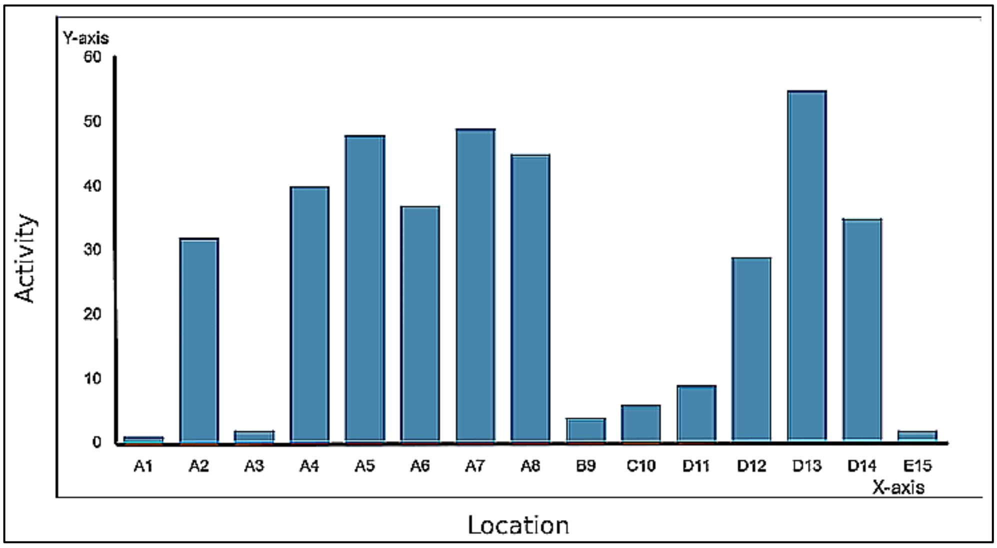

Figure 12: Clusters and Hotspots

There are intriguing patterns throughout time and geography due to the different cluster types. Mainly, the a large gap between clusters that incorporate walkers and bicyclists and those that don't; pedestrian and bicycle clusters predominate in the Centre of the city, whereas vehicle-only clusters are more common on the city's busier, arterial highways. This approach generates a database of various sizes and densities of hotspots and the collisions that occur inside the hotspot's limits. The database structure indicates that individual collisions are not examined. Still, the collection of collisions that are close to each other implies that there is some common or connecting causal element. At the same time, each site may be deemed distinct with characteristics – number of pedestrians, number of cyclists, and frequency of accidents in different climates. This similarity may construct a "like for like" comparison between hotspots. Table 2 represents the cluster characteristics, groups of related or similar entities that share common characteristics or attributes. These characteristics can vary depending on the type of cluster being considered.

Table 2. Cluster Characteristics

High			Low		
Variable	**Index**	**Variance**	**Variable**	**Index**	**Variance**
Accidents	499	81.52	Casualities*4	9	5.45
Number Cells	485	81.33	Rain with wind	3	6.27
Other wallets	153	10.62	Vehicles*4	2	3.7
Severity * Fatal	137	7.66	Snowing	0	2.3
Vehicles * 1	124	55.92	Vehicles*5	0	11.07
Unknown	95	13.07	Casualities*5	0	6.74
Tube Stations	93	2.8	Casualities*6	0	3.22

Table 3. Comparison of Clustering Methods

K-means (10)	**Mini K-Means (9)**	**DBSCAN (13)**	**Optics (12)**
8394 (1)	8521 (4)	5310 (3)	5310 (3)
6634 (2)	6691 (0)	2831 (1)	2831 (1)
5312 (4)	5317 (2)	1745 (6)	1745 (6)
3093 (5)	3215 (5)	1539 (5)	1559 (5)

Table 3 compares the outcomes of the clustering techniques used for this study. A look at the most excellent accident-prone locations packed together using various clustering algorithms reveals a noticeable variation in the information quantity in each significant cluster that may confuse the users. Various comparisons are performed in this research to discover the best clustering technique for data on traffic accidents. These algorithms' efficacy is evaluated using two criteria in this section: execution time and internal cluster validation metrics.

It is possible to determine the clustering quality using the Silhouette coefficient. As well as demonstrating how specific or well-separated an item is from another cluster, it also displays how closely an object relates to the other objects. Objects with a silhouette score greater than 1 are both well-matched inside their cluster and poorly matched with clusters on either side. On the other hand, a lower number indicates that the item is a poor match inside its cluster and has some similarities with clusters to its immediate north and south. When an item is shown this way, it indicates how distinct or far off it is from other clustered objects (cohesion) (separation). From 1/+1, which signifies that the item is matched well in its cluster and badly matched to its neighbors, the silhouette score ranges from 1/-1. Low values imply objects that are both unsuitable for their cluster and have some similarities with those in the clusters to which they belong. DBSCAN returns the maximum result of 0.751, as seen in Fig. 8. Two distances make up a silhouette score: a medium and a far one.

$$s = \frac{(b-a)}{((a,b))} \qquad [9]$$

As you can see, a is the average distance among each point in one cluster, whereas b is the distance between the next nearest cluster's other points and sample, defined by d – d-inter-cluster distance. The clustering is predicted if a and b are equal.

The value of suitable visualization to facilitate better model validation and real-world applicability for traffic authorities and urban planners. To this end, the study places emphasis on visual attributes such as spatiotemporal heatmaps, hotspot categorization, space-time cube plots, and time-series cluster plots in order to represent intricate traffic patterns effectively visually. These visualizations not only confirm the result of clustering and prediction but also demystify technical results into intuitive observations for non-technical users. In the future years, these developments can be realized to integrate dashboard-based interfaces or GIS-enhanced visualization tools in an effort to enable real-time monitoring and decision-making within metropolitan traffic networks.

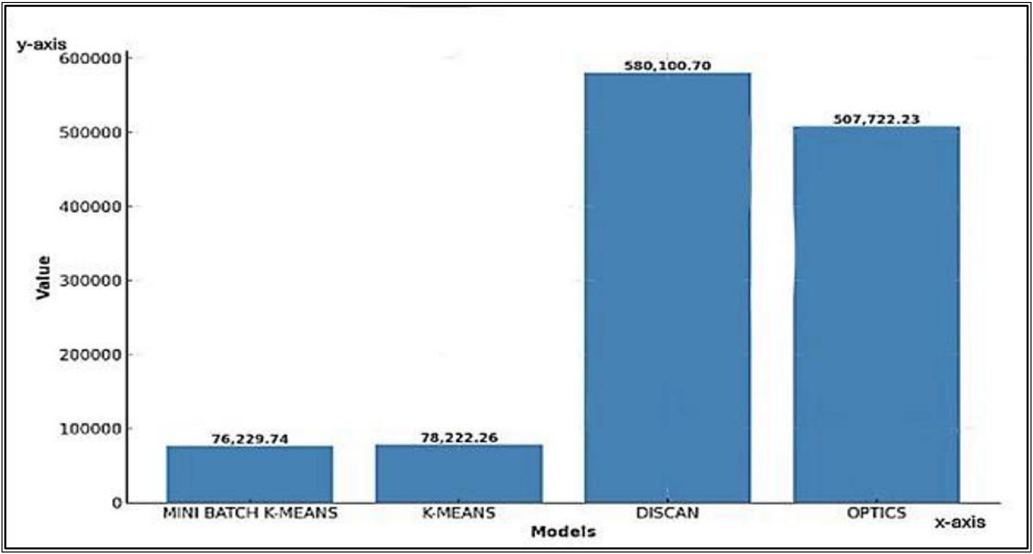

Figure 13: According to the Silhouette Coefficient, A Comparison of Clustering Techniques

In other words, a value of one indicates that the two clusters are touching, a value more than one implies that the clusters overlap, and a value lower than one denotes a respectable distance between the clusters. Put another way, we're searching for small values that show how well two clusters can be separated. The Silhouette score and Davies-Bouldin Index distinguish between each data point's centroid and the allocated cluster center point.

When comparing clustering techniques using the Silhouette Coefficient, the method that yields the highest average Silhouette score across all data points is generally considered to have produced the most effective clustering. This comparison helps select the most appropriate clustering technique for a given dataset in Figure 13.

Davies-Bouldin Index is defined based on the "within cluster" and "between-cluster" distances ratio.

$$Dij = \quad \sum_{i=1}^{k} D_{ij} \qquad [10]$$

$Open($Between clusters 0, 1, 2, 3, and 4, D_{ij} values are calculated, taking the sum of the values from those Clusters. Clusters 1, 2, 3, and 4 will be treated similarly. Then, the average of the highest values in the dataset is taken. A measure of how close the i and j clusters are together is called the "within-to-between cluster distance ratio," or D_{ij}.

$$D_{ij} = \frac{d_i + d_j}{d_{ij}} \qquad [11]$$

In this case, di and d_j are the average distances between the centroid and each data point in cluster i. A distance of d_{ij} separates the two clusters' centroids. In this case, D_{ij} is the cluster similarity index, estimated by multiplying the standard deviations by the difference between the center velocities and dividing that result by the sum of the squares. Di and die are both tiny, which shows that the ij clusters are incomparable. For example, a value of one indicates that the two clusters are nearby; a value of more than one indicates that the overlapping clusters; and a value of < 1 defines that the clusters are well separated. In other words, tiny numbers are considered to indicate how effectively two clusters are separated in Figure 14.

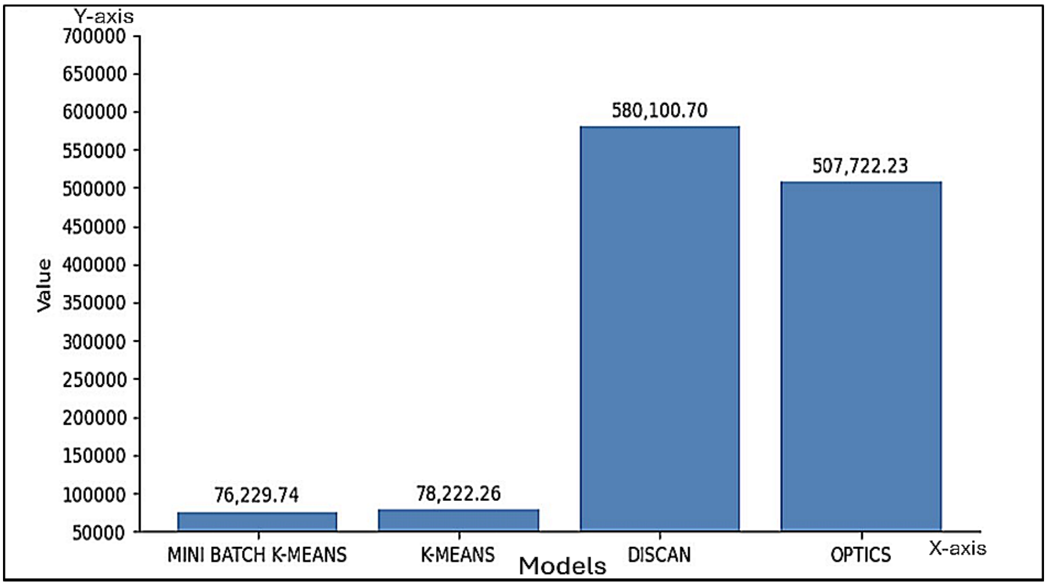

Figure 14: Comparison Based on Davies-Bouldin Index

Each cluster's "inter-cluster dispersion" and "intra-cluster dispersion" are added together to arrive at the Variance Ratio Criterion (VRC). The "better" the algorithm performs, the "higher" the index's score might be. There is no specified range for this value. The higher the score, the more distinct and dense the clusters are, which is consistent with the cluster concept.

Calinski-Harabasz Index: The index, sometimes called the VRC, is calculated as the total dispersion inside and across each cluster. The "higher" the index's score, the "better" the algorithm performs. This value's range is not given. The clusters get clearer and denser as the score increases, which is compatible with the idea of a cluster. OPTICS returns the highest number, 597,722, as seen in Figure 10. The global centroid, as well as the relative centroids for each of the k clusters. Intercluster SSB and intracluster dispersion (SSw) (38).

$$SS_w = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|^2 \qquad [12]$$

$$TSS - SS_w \qquad [13]$$

Where k – clusters, and N – data points. The total number of squares in a dataset equals the sum of the square distances between the dataset's centroid and each data point. For each data point, Ci stands for the cluster it belongs to; mi is the cluster's centroid, and ||xmi || is their distance. This is followed by the Calinski-Harabasz Index (CHI) and then the ratio of dispersion inside and across clusters:

$$CH_I = \frac{SS_B}{SS_W} \frac{(N-k)}{(k-1)} \qquad [14]$$

Total data points are represented by N. Indexes such as the Calinski-Harabasz and the Silhouette measure the efficiency of a group by comparing pairwise differences in the distances inside and across clusters, respectively.

According to the Silhouette Metric and Davies-Bouldin Index, it has been concluded that DBSCAN is a superior option. The CHI is much higher for convex clusters than for alternative cluster definitions, such as clusters gathered from DBSCAN. K-means uses the closest mean to split the space, creating convex sections. As long as a line segment joining two locations is inside the set, it's called convex. Using DBSCAN and OPTICS, a maximum cluster radius is set. The algorithm will group points accessible from one another. However, a non-convex cluster is possible. A higher Calinski-Harabasz Index indicates better-defined and more distinct clusters, making it a useful tool for comparing different clustering methods or determining the optimal number of clusters in a dataset in Figure 15. It's now obvious that DBSCAN creates clusters faster than OPTICS has shown in Table 4.

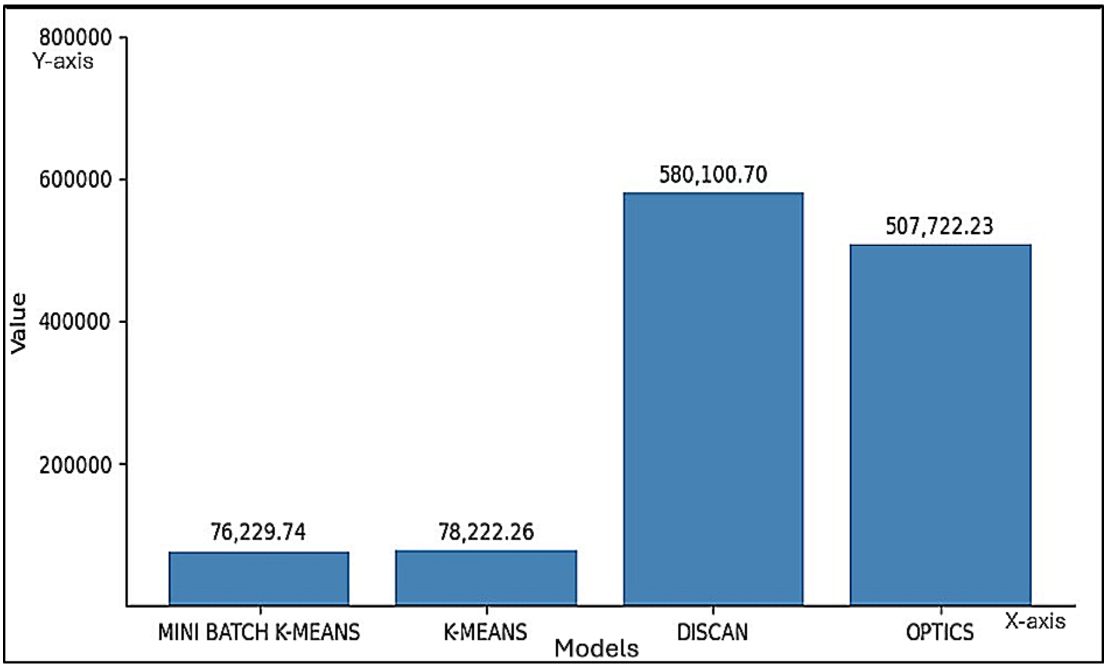

Figure 15: Comparison Based on CHI

Table 4. Clustering Times

Total No of Test Cases.	Time Taken to cluster (sec)			
	K-means	Mini Batch K-means	Optics	DBSCAN
8700	0.05	0.097	11.349	1.1126
20554	0.251	0.1	41.675	4.951
32407	0.348	0.11	81.656	7.899

Real-World Deployment and Smart City Integration

The mining technique adopted in this work is based mostly on unsupervised clustering supported by metaheuristic optimization methods for optimizing clustering performance. Specifically, the method employs Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points To Identify the Clustering Structure (OPTICS) for identifying accident-prone areas and congestion clusters label-free. These are further enhanced with Particle Swarm Optimization (PSO) for enhancing cluster parameters towards achieving increased accuracy and responsiveness to a variety of traffic scenarios. The approach integrates time-series forecasting models and clustering models to identify spatiotemporal congestion patterns. The approach can thus accurately be termed as an unsupervised, multi-technique clustering framework for real-time urban traffic mining and prediction.

Thus, clustering performance, spatiotemporal behavior, and forecast output have been checked and verified with literature results. For example, DBSCAN and OPTICS usage have been compared against standard clustering algorithms such as K-means set in the literature (13–16). Likewise, our forecasting for time series is consistent with recent predictive traffic modeling literature (17–21). These analogies are to provide context for authenticating the strength and originality of our approach and placing it in the general intellectual and practical debate of traffic congestion analysis.

Conclusion

To get around the drawbacks of many earlier techniques for grouping accident-prone sites, this work suggests OPTICS and DBSCAN. The density-based cluster strategies are more adept at finding geographical clusters and handling outliers than other methods. The simulation on a real-world dataset supports the same. This research is targeted to assess and forecast the current level of traffic operations on roads. The research focused on weekday traffic congestion and used the emerging hotspot analysis to search for fresh, escalating, recurrent, and sporadic hotspot patterns every hour. The network's traffic status is split into three groups using a time-series clustering method. The time series prediction models are applied - CFF, ESF, and FBF. The forecast and VRMSE are measured for the model validation. As a result of this analysis, the following spatiotemporal characteristics of network congestion were identified: Intersections and significant urban arteries are the most often congested sites. Commuter routes are congested both in the day and night, with every rush period showing the most congestion. The lowest average FRMSE/RMSE figures clearly show that the CFF exactly anticipates the traffic state. However, this study has several drawbacks. More data might be incorporated into the model as a first step to improve accuracy. Future studies could look towards leveraging GPS data from vehicles to address the limitations of taxi GPS. These ML strategies are applied in this research to anticipate operational traffic conditions; more strategies still need an inquiry. Future studies should include a range of methods for traffic prediction. This research only applied TTI throughout various periods to create predictions. In the future, new evolutionary and hybrid methods will be applied further to reduce computing time and to increase the measures of performance.

Future Work

Using spatial data expertise has become increasingly crucial in developing unsupervised algorithms that cluster massive databases quickly and efficiently. The random groupings require innovative processing approaches to detect and analyze road traffic congestion in real time accurately. The proposed experimental approach using particle swarm optimization offers a promising solution to improving the accuracy of traffic congestion modeling and analysis. Furthermore, it can significantly enhance the health and safety of road users by reducing the risk of accidents caused by sudden stops or unexpected changes in traffic patterns. Overall, these advancements in traffic congestion detection and analysis will ultimately lead to more efficient and safer road networks for everyone.

Abbreviations

ARIM: Autoregressive Integrated Moving Average, CHI: Calinski-Harabasz Index, DBSCAN: Density-Based, Spatial Clustering of Applications with Noise, FRMSE: Forecast Root Mean Squared Error, GIS: Geographic Information System, GPS: Global Positioning System, KDE: Kernel Density Estimation, OPTICS: Ordering Points To Identify the Clustering Structure, PSO: Particle Swarm Optimization, SPI: Speed Performance Index, STC: Space-Time Cube, TTI: Traffic Performance Index, VRC: Variance Ratio Criterion, VRMSE: Validation Root Mean Squared Error.

Acknowledgment

None.

Author Contributions

Sekar Kidambi Raju: Original draft writer, V Venkataraman: Methodology, V Rengarajan: Data curation, G Sathiamoorthy: Validation, Ganesh Karthikeyan Varadarajan: Software, Raj Anand Sundaramoorthy: Visualization.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Approval

Ethical approval was not required for this study, as it did not involve human participants or animal subjects.

Funding

None.

References

1. Zheng Y, Li Y, Own CM, Meng Z, Gao M. Real-time prediction and navigation on traffic congestion model with equilibrium Markov chain. International journal of distributed sensor networks. 2018 Apr;14(4):1550.
2. Huang Z, Xia J, Li F, Li Z, Li Q. A peak traffic congestion prediction method based on bus driving time. Entropy. 2019 Jul;21(7):709. doi:10.3390/e21070709.
3. Lauf S, Haase D, Kleinschmit B. The effects of growth, shrinkage, population aging and preference shifts on urban development—A spatial scenario analysis of Berlin, Germany. Land Use Policy. 2016 Mar;52:240–54.
4. Chen H, Chen H, Zhou R, Liu Z, Sun X. Exploring the mechanism of crashes with autonomous vehicles using machine learning. Math Probl Eng. 2021 Feb;2021:1 10.
5. Zou Y, Zhu T, Xie Y, Li L, Chen Y. Examining the impact of adverse weather on travel time reliability of urban corridors in Shanghai. J Adv Transp. 2020 Dec;2020:1–11.

6. Shimazaki H, Shinomoto S. A method for selecting the bin size of a time histogram. Neural Comput. 2007 Jun;19(6):1503–27.
7. Mashfiq Rizvee M, Amiruzzaman M, Islam MR. Data mining and visualization to understand accident-
8. Prasannakumar V, Vijith H, Charutha R, Geetha N. Spatiotemporal clustering of road accidents: GIS-based analysis and assessment. Procedia Soc Behav Sci. 2011;21:317–25.
9. Sisodia D, Singh L, Sisodia S, Saxena K. Clustering techniques: a brief survey of different clustering algorithms. International Journal of Latest Trends in Engineering and Technology (IJLTET). 2012 Sep;1(3):82-87.
10. Liu K, Shang Y, Ouyang Q, Widanage WD. A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery. IEEE Transactions on Industrial Electronics. 2020 Mar 18;68(4):3170-80.
11. Xu X, Liu X, Xu Z, Dai F, Zhang X, Qi L. Trust-oriented IoT service placement for smart cities in edge computing. IEEE Internet Things J. 2019;7:4084–91.
12. Zhang Y, Yin C, Lu Z, Yan D, Qiu M, Tang Q. Recurrent Tensor Factorization for time-aware service recommendation. Appl Soft Comput. 2019;85:105762.
13. Zhao Z, Chen W, Wu X, Chen PC, Liu J. LSTM network: A deep learning approach for short-term traffic forecast. IET Intell Transp Syst. 2017;11:68–75.
14. Yang T, Jiang Z, Shang Y, Norouzi M. Systematic review on next-generation web-based software architecture clustering models. Comput Commun. 2021;167:63–74.
15. Mukherjee A, Goswami P, Yang L, Yan Z, Daneshmand M. Dynamic clustering method based on power demand and information volume for intelligent and green IoT. Comput Commun. 2020;152:119–25.
16. Hua L, Jing S, Yi X, Wang H, Xin H. A new hybrid method based on partitioning-based DBSCAN and ant clustering. Expert Syst Appl. 2011;38:9373–81.
17. Yan C, Wei X, Liu X, Liu Z, Guo J, Li Z, Lu Y, He X. A new method for real-time evaluation of urban traffic congestion: A case study in Xi'an, China. Geocarto Int. 2020 Jul;35(10):1033–48.
18. Chen H, Chen H, Liu Z, Sun X, Zhou R. Analysis of factors affecting the severity of automated vehicle crashes using XGBoost model combining POI data. J Adv Transp. 2020 Nov;2020:1–12.
19. Yu X, Xiong S, He Y, Wong WE, Zhao Y. Research on campus traffic congestion detection using BP neural network and Markov model. J Inf Secur Appl. 2016 Dec;31:54–60.
prone areas. In Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020. Singapore: Springer Singapore. 2021 May 18:143-154. https://doi.org/10.48550/arXiv.2103.09062
20. Ma C, Zhou J, Xu X, Xu J. Evolution regularity mining and gating control method of urban recurrent traffic congestion: A literature review. J Adv Transp. 2020 Jan;2020:1–13.
21. Yang X, Luo S, Gao K, Qiao T, Chen X. Application of data science technologies in intelligent prediction of traffic congestion. J Adv Transp. 2019 Apr;2019:1–14.
22. Du M, Ding S, Jia H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. Knowl-Based Syst. 2016;99:135–45.
23. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Inkdd 1996 Aug 2;96(34):226-231.
24. Cai L, Meng J, Stroe DI, Peng J, Luo G, Teodorescu R. Multiobjective optimization of data-driven model for lithium-ion battery SOH estimation with short-term feature. IEEE Transactions on Power Electronics. 2020 Apr 16;35(11):11855-64.
25. Liu H, Pang L, Li F, Guo Z. Hough Transform and Clustering for a 3-D Building Reconstruction with Tomographic SAR Point Clouds. Sensors. 2019;19:5378.
26. Zhao J, Kigen KK, Wang M. Modeling the operation of vehicles at signalized intersections with special width approach lanes based on field data. IET Intell Transp Syst. 2020;14:1565–72.
27. Gao F, Zhang Q, Han Z, Yang Y. Evolution test by improved GA with application to performance limit evaluation of APPS. IET Intell Transp Syst. 2021;15:754–64.
28. Joshan Athanasius J, Vasuhi S, Vaidehi V, Shiny Christobel J, JerartJulus L. Adaptive density-based data mining technique for detection of abnormalities in traffic video surveillance. J Intell Fuzzy Syst. 2020;39:3737–47.
29. Bryant A, Cios K. RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates. IEEE Trans Knowl Data Eng. 2018 Jun;30(6):1109–21.
30. Xie P, Li T, Liu J, Du S, Yang X, Zhang J. Urban flow prediction from spatiotemporal data using machine learning: a survey. Inf Fusion. 2020;59:1–12.