CHALMERS

7 Spatiotemporal Analysis of Traffic Data for

Identifying Congestions

Souptik Paul
Sangeeth John
Venkata Sai Dinesh Uddagiri
Madhumitha Venkatesan

Group 3

November 16, 2023

Contents

1 Abstract

This project, tackles the critical issue of traffic congestion in urban areas by employing innovative spatiotemporal data analysis techniques. Recognizing the multifaceted challenges posed by traffic congestion, we develop a model that seamlessly integrates spatial insights from distance-based clustering with temporal intricacies of vehicle counts. Leveraging the power of the Density-Based Spatial Clustering (DBSCAN) algorithm and Haversine's formula for calculating geographical distances, our approach transforms the complex road network into clusters of streets, capturing the inherent structure of traffic dynamics. The temporal dimension is then skillfully incorporated through a detailed analysis of vehicle counts, allowing us to discern peak congestion periods and identify temporal patterns indicative of varying traffic conditions. The model introduces a congestion threshold, based on quantiles obtained from the traffic data, facilitating the categorization of congestion levels. Applying this model to traffic data from the city of Aarhus, we successfully identify congestion patterns across different days of the week, providing valuable insights for transportation management and urban planning. The results showcase the model's accuracy in identifying congestion, with specific validation on a major highway confirming its authenticity. This project stands at the intersection of transportation engineering, data science, and urban planning, offering a holistic and predictive approach to address the complex challenges of traffic congestion in contemporary urban landscapes.

2 Introduction

In our rapidly urbanizing world, the escalating challenge of traffic congestion in cities has become a critical issue, affecting not only daily convenience but also causing economic losses, environmental degradation, and a decline in the overall quality of life for residents. To address this multifaceted problem, there is a growing need to employ data-driven approaches in transportation management and urban planning. This report outlines a project that sits at the intersection of transportation engineering, data science, and urban planning. It recognizes that an effective solution to traffic congestion requires a deep understanding of the underlying spatiotemporal data patterns. The project proposes an innovative spatial-temporal model that integrates distance-based clustering and temporal vehicle counts, providing a holistic representation of traffic congestion. Leveraging advanced techniques, including the application of Haversine's formula and Density-Based Spatial Clustering (DBSCAN) algorithm, the model aims to identify congestion patterns within urban environments. The report details the problem statement, analysis methods, and implementation steps, showcasing results obtained from real-world traffic data in the city of Aarhus. Through this comprehensive approach, the project seeks to contribute to the development of predictive models capable of capturing the dynamic interplay between spatial proximity and temporal patterns in traffic conditions.

3 Background

In our rapidly urbanizing world, the phenomenon of traffic congestion has reached critical proportions, plaguing cities with a myriad of challenges. This persistent issue arises from the complex interplay of various factors, including population growth, increased urbanization, and inadequate transportation infrastructure. As urban centers continue to expand, the consequences of traffic congestion extend far beyond mere inconvenience, encompassing economic losses, environmental degradation, and compromised quality of life for residents.

To address these multifaceted challenges, there is a growing imperative to employ data-driven approaches in the field of transportation management and urban planning. Spatiotemporal traffic data, which captures the movement of vehicles and their interaction with the urban environment over time, has emerged as a valuable resource. Analyzing such data can provide crucial insights into traffic patterns, identifying congestion hotspots, and inform decision-making processes aimed at optimizing transportation systems.

This project stands at the intersection of transportation engineering, data science, and urban planning. Its foundation lies in the recognition that the effective management of traffic congestion necessitates a deep understanding of the underlying data patterns. This project seeks to leverage advanced data clustering techniques and analysis methods to detect and analyze traffic congestion within urban environments.

In the upcoming section we define the problem of traffic congestion as a spatio-temporal analysis problem as well as discuss a few basis definitions with respect to spatio-temporal analysis.

4 Problem Statement and Analysis

4.1 Problem Statement

Urban areas face persistent challenges related to traffic congestion, leading to disruptions in transportation networks, increased travel times, economic losses, and environmental concerns. The objective of this project is to develop an innovative spatial-temporal model that seamlessly integrates spatial insights derived from distance-based clustering along with the temporal intricacies obtained from vehicle counts. The model aims to provide a holistic representation of traffic congestion as a multifaceted spatial-temporal clustering challenge. Similar work has done on this field by Toshniwal et.al. [1].

Before we can define our model for detecting traffic congestion, let us take a look at a few basic definitions related to this model.

4.2 Traffic Congestion

Traffic congestion [2] refers to a situation where the transportation network within a specific geographic region or locality experiences a significant reduction in the flow of traffic due to an overload of vehicles. This can occur on city streets, highways, or a combination of both. In such cases, the demand for transportation infrastructure exceeds its capacity to handle the volume of vehicles efficiently, leading to delays, slower speeds, and sometimes a complete halt in movement. For an illustrative example of traffic congestion, please refer to Figure 1.

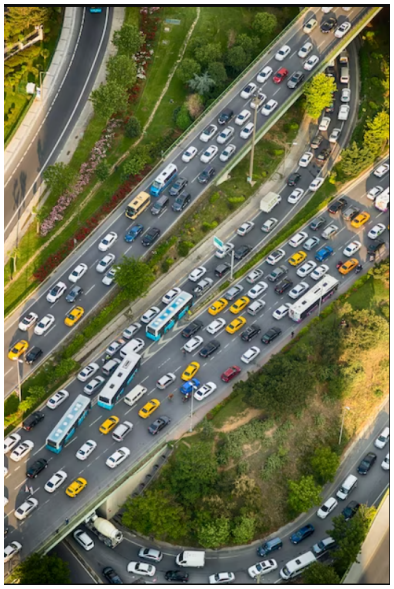

Figure 1: Traffic Congestion

Factors contributing to area-based traffic congestion may include a high population density, rapid urbanization, insufficient road capacity, inadequate public transportation options, inefficient traffic management, and a high number of vehicles for the available road space. The result is often increased travel times, frustration for commuters, economic losses, and environmental issues.

4.3 Spatiotemporal analysis

Spatiotemporal analysis refers to the examination and interpretation of data that varies both spatially and temporally. It involves the study of how phenomena change and interact not only across different locations but also over different points in time. This type of analysis is particularly

valuable in fields such as geography, environmental science, urban planning, epidemiology, and transportation, where understanding the evolving patterns of events or conditions is crucial.

4.4 Data Clustering

Data clustering using geographical distance [3] is a technique employed in data analysis to group spatially related data points based on their proximity in geographical space. This approach is particularly useful when dealing with datasets that have inherent spatial characteristics, such as geographical coordinates (latitude and longitude). The goal is to identify clusters of data points that are close to each other in the real-world physical space.

4.5 Our Approach

In our innovative approach to predicting traffic congestion, we intricately frame the challenge as a spatial-temporal analysis problem. Our objective is to find the congestion patterns for a particular area containing a number of streets. As traffic is constantly flowing, we presume that vehicles which are in a particular geographical area, will travel to a different street within the same geographical area, keeping the total number of vehicles for a particular area constant and thereby contributing to congestion. The foundational step involves the application of Haversine's formula to quantify the geographical distances between all streets, effectively transforming the intricate road network into a set of data points in a spatial coordinate system. Subsequently, we deploy the Density-Based Spatial Clustering (DBSCAN) algorithm to categorize streets into clusters, leveraging a minimum distance threshold to define the spatial cohesion within each cluster. This cluster formation inherently captures the inherent structure of traffic dynamics, encapsulating the idea that streets within a close geographical proximity are likely to share similar traffic characteristics. Each cluster represents an entire geographical location in the city of Aarhus, containing a number of streets as per the clustering algorithm. For an illustrative example of cluster that that represents a geographical area containing multiple streets can be seen in Figure 2.

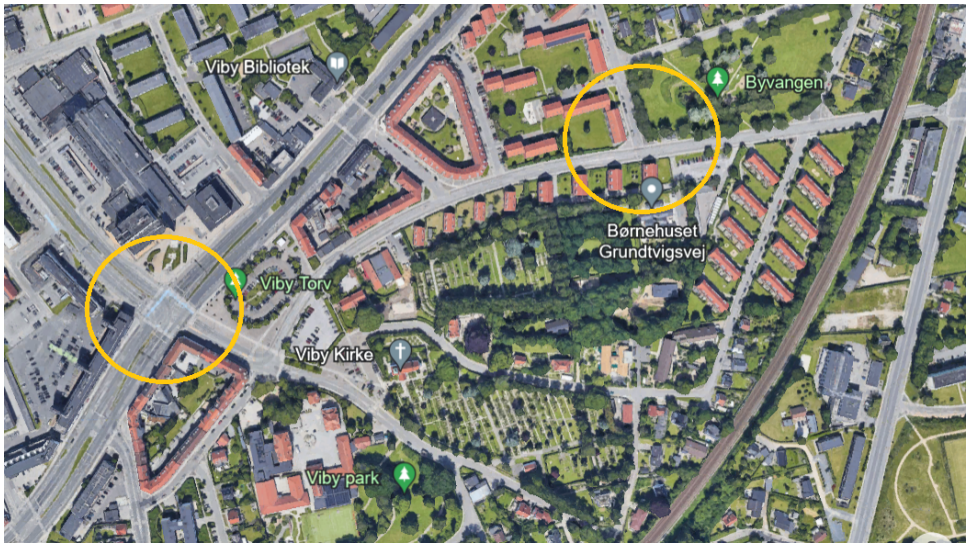

Figure 2: The each highlighted area(in Orange) shows a cluster that represents a geographical area containing multiple streets

The temporal dimension is then seamlessly integrated by conducting a comprehensive analysis of vehicle counts within each street across distinct time intervals, ranging from the early hours of the morning to midnight. This meticulous examination allows us to construct detailed time-series datasets for each street, unveiling the nuanced variations in traffic density throughout the day. It

is this temporal evolution of vehicular activity that enriches our model with the capability to discern peak congestion periods and identify the temporal patterns characteristic of varying traffic conditions. The final stage of our algorithm involves a judicious comparison of the total vehicle counts within each cluster (containing a number of streets) against a predetermined threshold. This comparison serves as the decisive factor in predicting congestion: if a cluster's total vehicle count surpasses the threshold, it is classified as heavy or moderately congested; conversely, if it falls below, the cluster is deemed un-congested. Since the clusters contain streets which are within a certain geographical area, if a cluster is congested, we can say that the geographical location of the cluster in facing traffic congestion. This threshold-based classification imparts adaptability to the model, allowing for customisation based on the unique characteristics of different urban environments.

In essence, our approach artfully combines the spatial insights derived from distance-based clustering with the temporal intricacies of vehicle counts, resulting in a holistic representation of traffic congestion as a multifaceted spatial-temporal clustering challenge. This nuanced perspective not only aligns with the intricate nature of real-world traffic phenomena but also lays the foundation for a predictive model capable of capturing the dynamic interplay between spatial proximity and temporal patterns in traffic conditions.

5 Implementation

5.1 Data Collection

For this project, we use traffic data [4] from the city of Aarhus. Aarhus, the second-largest city in Denmark with a population of approximately 361,544 residents as of 2023. Our project focuses on traffic data spanning from February 2014 to June 2014. The data collection process involves the strategic placement of sensors at two proximate locations on a street, where they diligently count the number of vehicles passing through during predefined time intervals. An example of a sensor placement with the sensor ID can be seen in figure 23. The data constitutes of 1 meta-data file and 449 sensor reading files. Each dataset entry corresponds to a specific road segment and includes essential information such as the timestamp, vehicle count, unique id for each reading, report id for the sensor and the latitude and longitude for the two points where the sensors are located. These counts were measured by individual detectors on various roads, on specific days, and during specified time periods.

5.2 Data Cleaning and Preprocessing

The traffic data needs to be processed so that it can be used for the mining of traffic patterns. For processing the data, it is merged into a single data frame. The data contains several fields such as timestamp(the time vehicles are recorded in the sensor), vehicle count(number vehicles that was recorded by the sensor), the latitude and longitude for the two points where the sensors are located. The data preprocessing method is essential as the data also contains additional fields such as average speed, average measured time extID(denotes the unique identifier assigned to each road). For this project, as it is essential for us to predict the congestion for a particular area, we only use a few particular fields from the data set. After the merging of the datasets, the final fields in the data frame are TIMESTAMP, vehicleCount, _id, REPORT_ID, POINT_1_LAT, POINT_1_LNG, POINT_2_LAT, POINT_2_LNG. Subsequently after the merging, we drop all the rows in the data frame, which has incomplete or missing data. After this, we drop all the rows which have vehicle count > 100. This is because some of the sensors used for the data recording process are of low quality, which abruptly record large traffic volumes and thereby affect the process of interpolation. Another preprocessing step is that the Timestamp(s), are converted into the datetime format. If there is no value obtained for 4 consecutive hours in a day, then the day is marked as invalid for processing. Finally, if the _id field, which denotes the unique id for each reading taken by the sensor has duplicates, they are removed.

5.3 Spatial Evaluation - Data Clustering

After the data has been preprocessed, we can now perform the spatial evaluation of this data. In the spatial evaluation, we group streets within a particular geographic distance to create the area for which congestion can be predicted. For calculating the geographic distance between two streets, we use Haversine's formula.

5.3.1 Haversine's Formula

Haversine's formula is a trigonometric equation widely employed in navigation and geospatial calculations to determine the great-circle distance between two points on the surface of a sphere. This formula is particularly useful in applications that involve measuring distances on the Earth, where the Earth is approximated as a sphere. Named after the Haversine function, which is central to its calculation, the formula takes into account the latitude and longitude coordinates of two locations and provides an accurate measurement of the shortest distance between them. A promising research using Haversine's formula, for calculating geographic distance has been performed by Prasetya et al. [5].

The Haversine formula is expressed as follows, where r is the radius of the Earth (6371 km). d is the distance between two points. ϕ_1 and ϕ_2 are the latitudes of the two points. λ_1 and λ_2 are the longitudes of the two points, respectively.

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\Phi_2 - \Phi_1}{2}\right) + \cos(\Phi_1)\cos(\Phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{1}$$

For example, let us obtain the distance between the following latitude and longitude values using Haversine's formula

$$\begin{aligned} \Phi_1 &= 56.2149789163549, \\ \Phi_2 &= 56.2085202681621, \\ \lambda_1 &= 10.139695703041, \\ \lambda_2 &= 10.16103838525. \end{aligned} \tag{2}$$

Now, let's calculate d:

$$\Delta\Phi = \Phi_2 - \Phi_1 = 56.2085202681621 - 56.2149789163549 \approx -0.0064586481928 \text{ radians} \tag{3}$$

$$\Delta\lambda = \lambda_2 - \lambda_1 = 10.16103838525 - 10.139695703041 \approx 0.021342682209 \text{ radians} \tag{4}$$

Now, we can substitute these values into the formula:

$$\begin{aligned} d = 2 \cdot 6371 \cdot \sin^{-1}\Bigg(\sqrt{\sin^2\left(\frac{-0.0064586481928}{2}\right) + \cos(56.2149789163549)\cos(56.2085202681621)} \\ \times \sin^2\left(\frac{0.021342682209}{2}\right)\Bigg) \\ \approx 1.10742318901 \text{ kilometers} \end{aligned} \tag{5}$$

So, the distance d between the two points is approximately 1.107 kilometers

5.3.2 Clustering using DBSCAN

Now that have obtained the means to calculate the geographic distance between two streets. We can now group the streets into the same cluster using DBSCAN algorithm.

Density-Based Spatial Clustering of Applications with Noise [6] (DBSCAN) is a clustering algorithm renowned for its ability to identify clusters of arbitrary shapes within a dataset. The algorithm operates based on the density of data points, distinguishing dense regions from sparser ones. DBSCAN defines clusters as continuous regions of high data point density, separated by areas of lower density. The reason we use DBSCAN over other clustering algorithms is its ability to discover clusters without assuming their shape in advance, making it particularly valuable in our case where clusters may exhibit complex and irregular patterns.

In our implementation, we cluster streets based on a minimum geographic distance. DBSCAN takes two parameters eps and min_samples. Eps represents the maximum distance, between two points for them to added to a particular cluster, for this case we have set eps = 0.5, which indicates that all streets within 500 metres of each other, will be added to the same cluster. On the other hand, min_samples represent the minimum number of points required to form a cluster. For our implementation, we have set min_samples to 2, which indicates that, at-least 2 streets are required to form a cluster. Then we create an instance of the DBSCAN model, with the defined eps and min_samples parameters. We also created a function for calculating the Haversine distance, which is passed to the DBSCAN model as a distance metric, allowing

DBSCAN to operate in a geospatial context. This provides us with a scatter plot, visualising the clusters formed from all the streets. The scatter plot showcasing DBSCAN model can be seen in Figure 3. After performing DBSCAN clustering, we get a number of clusters, denoted by cluster IDs, containing a number of streets.

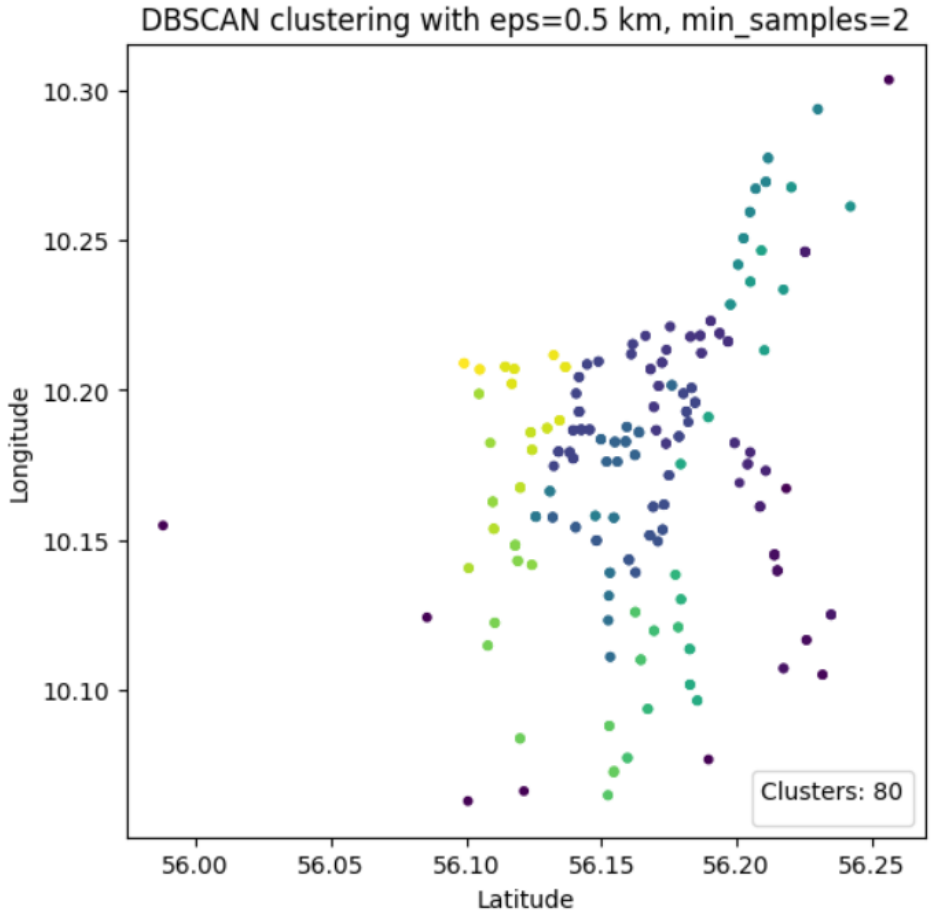

Figure 3: Density-Based Spatial Clustering with minimum geographic distance eps = 0.5 km

5.4 Temporal Evaluation - Time Series

After performing the spatial evaluation and obtaining the areas for which congestions can be identified, we now perform the temporal evaluation. In temporal evaluation, we obtain the time series graph, by plotting the number of vehicles observed on each sensor, at each hour for a particular day. In our data, the vehicles are observed at very short intervals of time, which makes it difficult to process. Now we assume that traffic congestion, does not happen instantaneously, rather it happen over period of time. Hence to make the processing easier and to identify congestion at a particular interval, we sum up the data leading to a particular hour. This gives us the total number of vehicles, observed in sensor for 1 hour time interval. This summation is then use to plot the time series graph, for 24 hours.This will be then be expanded to analysis the data for 7 days. The time series graph for one day can be seen in Figure 4.

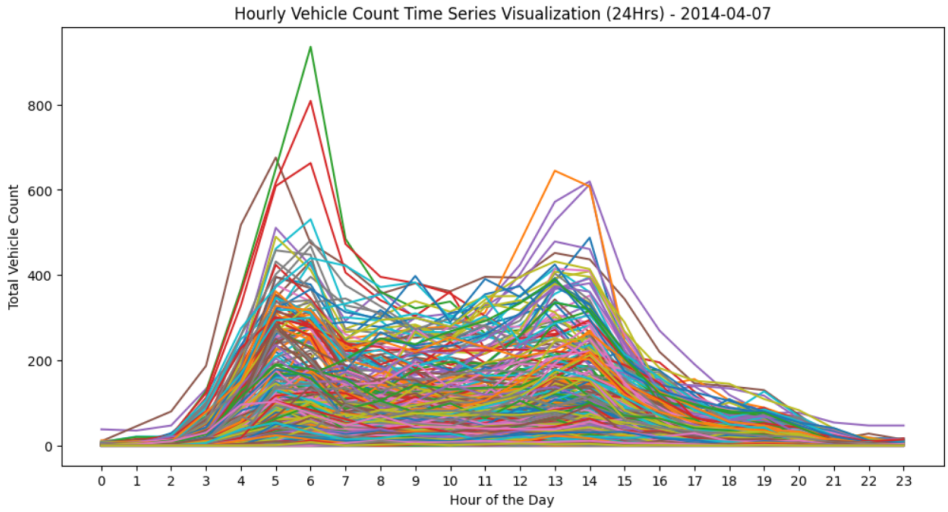

Figure 4: Time Series Plot for each street, Hour of the Day Vs Total Vehicle Count

5.5 Spatio - Temporal Evaluation

Now that we have performed the Spatial Evaluation and the Temporal Evaluation of the data. The Spatial Evaluation gives us a number of clusters with cluster IDs and corresponding report IDs(sensor IDs for each street). This data is then represented into a single data frame, which will be used for further processing. Similarly, the temporal evaluation gives us the date, hour in the which vehicles have been detected and the vehicle count observed for the particular hour. This data is also represented in a single data frame. Now that we have obtained the two dataframes for spatial evaluation and temporal evaluation, we now merge these two dataframes into a single data frame containing the date, cluster ID, hour for the particular day and vehicle count for that hour. This merged dataframe gives us the evaluated data for 7 days, containing the clusters or geographical areas where we can detect congestion and with the vehicles counts for those areas for particular days and hours. This merged dataframe can now be used, for identifying congestions in particular areas. The total vehicle count of a cluster at a particular hour can be calculated by summing up the vehicular counts for each street in the cluster for that particular hour. To quantify congestions, we need another factor called Congestion Threshold to identify, if an area is congested or not.

Congestion Threshold: For identifying traffic congestion in a particular area at a particular time, we define a congestion threshold. This threshold, indicates the minimum number of vehicles, that needs to present in the area(cluster), at a particular time to make the area congested. This threshold can depend on a number of factors such as number of streets, size of the streets, accidents, weather etc. However, as our data does not reflect factors such as accidents, weather and size of the streets, we use a quantile approach combined with the number of streets in a cluster, to calculate the threshold.

For calculating the congestion threshold, we use the quantile method on the dataframe. The quantiles specified are 0.50 (50th percentile, which is the median), 0.75 (75th percentile) and 1(100th Percentile - Maximum Value). These quantiles give us the values as 82.0, 246.0, 3551.0. Now that we have obtained the quantiles, we can categorize the congestion level of a specific cluster in the DataFrame based on the relationship between the total vehicle count, number of streets in a cluster, and predefined quantile values. This relation gives us the congestion categorisation by comparing the total vehicle count for a particular cluster at a specific hour, with the quantile multiplied with the number of streets in that cluster. The categorisation can be defined as follows:-

Table 1: Congestion categorisation in a cluster

Quantile	Threshold Calculation	Congestion Categorisation
82.0	[Total no. of vehicles in the particular cluster per hour] <= 82.0 x [No. of streets in the particular cluster]	No Congestion [GREEN Color]
246.0	[Total no. of vehicles in the particular cluster per hour] > 82.0 x [No. of streets in the particular cluster] and [Total no. of vehicles in the particular cluster per hour] <= 246.0 x [No. of streets in the particular cluster]	Moderate Congestion [ORANGE Color]
3551.0	[Total no. of vehicles in the particular cluster per hour] > 246.0 x [No. of streets in the particular cluster] and [Total no. of vehicles in the particular cluster per hour] <= 3551.0 x [No. of streets in the particular cluster]	Heavy Congestion [RED Color]

Now that we have obtained the relation for congestion categorization, the total vehicle counts in the dataframe for each hour and for each cluster, can be compared with the threshold to obtain the congestion congestion categorisation at each hour for every cluster. This is represented in the graph shown below, showcasing the cluster IDs and congestion categorisation for each hour for one day. The data represented in the graph is then represented on a map to visually see the congestion patterns at different hours across all the clusters. The Congestion Categorisation for each cluster IDS and an hour for one day can be observed in Figure 5 .

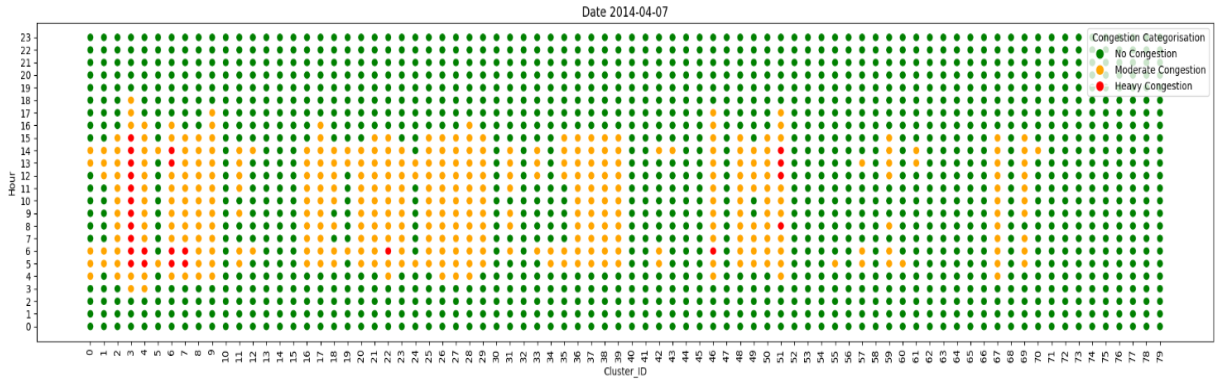

Figure 5: The Congestion Categorisation for each cluster IDS and an hour for one day

6 Results

Our traffic congestion detection model was tested for a week starting from 6th April 2014 to 12th April 2014. A week or consecutive seven days was chosen so that we could analysis congestion patterns across different days of the namely weekdays(Monday-Friday) and weekends(Saturday,Sunday). As per the model, first we use DBSCAN, with Haversine's formula to obtain the clusters for all available streets. The clusters obtained after the DBSCAN clustering model can be seen in Figure 3.

Now that we have obtained the clusters, the time series graphs can be obtained for each day of the week for the specified week. The time series graphs can be seen below.

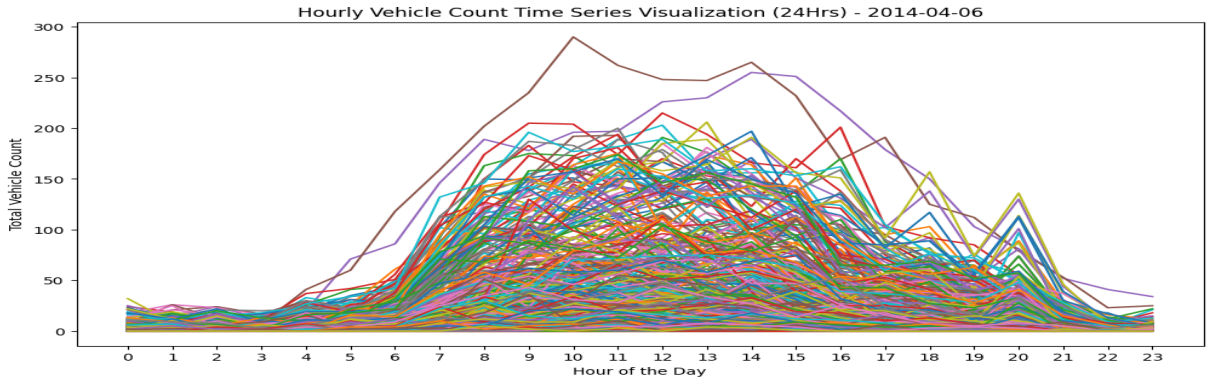

Figure 6: Time Series Graph visualizing vehicle counts for 06-04-2014 Sunday

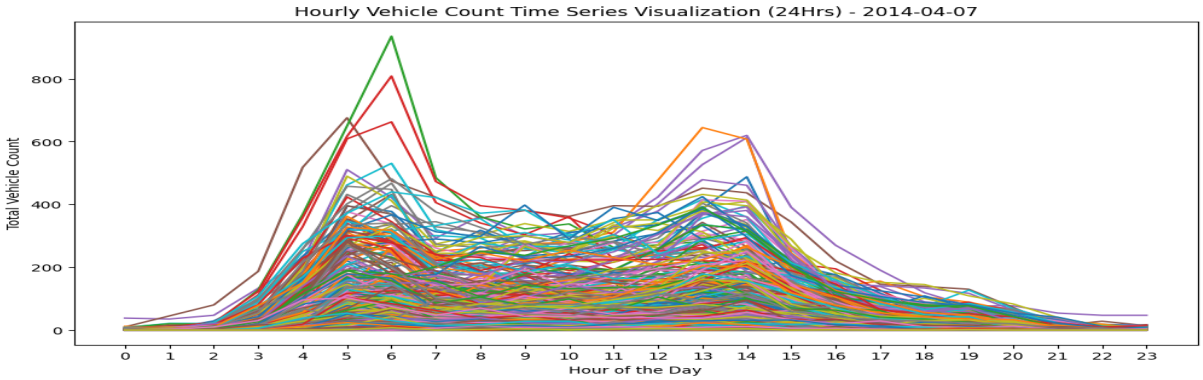

Figure 7: Time Series Graph visualizing vehicle counts for 07-04-2014 Monday

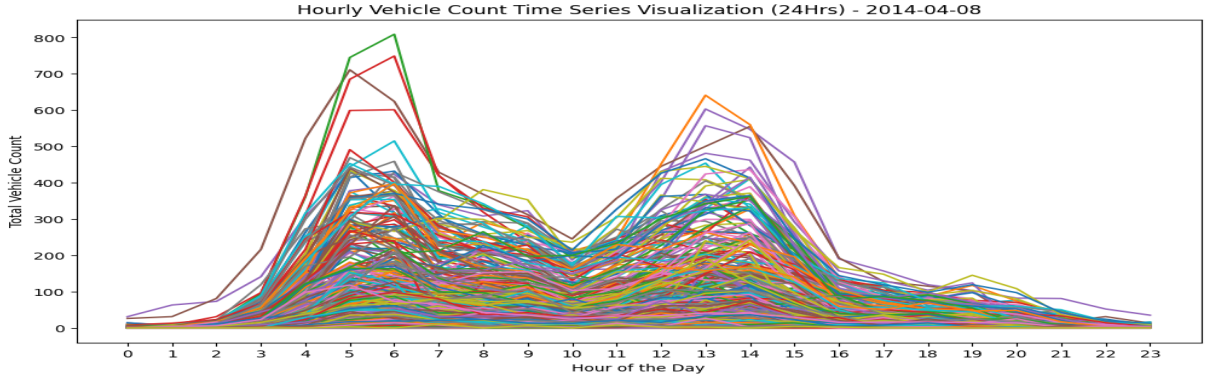

Figure 8: Time Series Graph visualizing vehicle counts for 08-04-2014 Tuesday

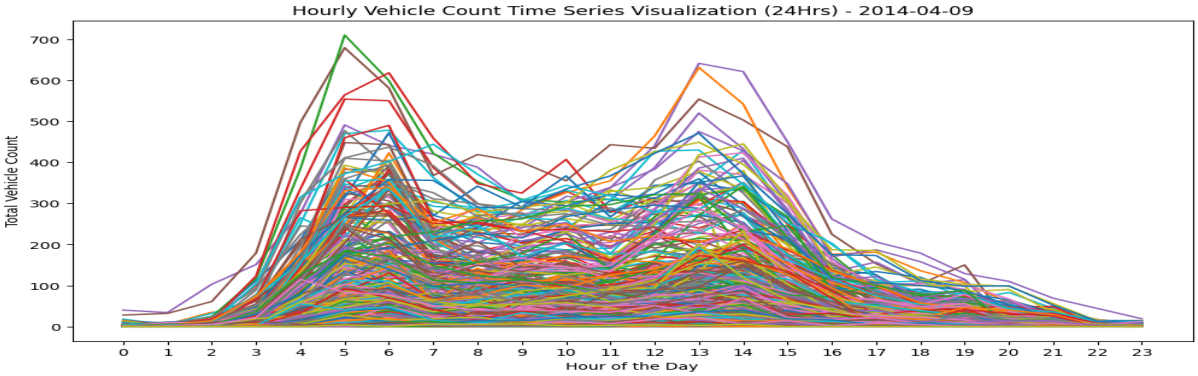

Figure 9: Time Series Graph visualizing vehicle counts for 09-04-2014 Wednesday

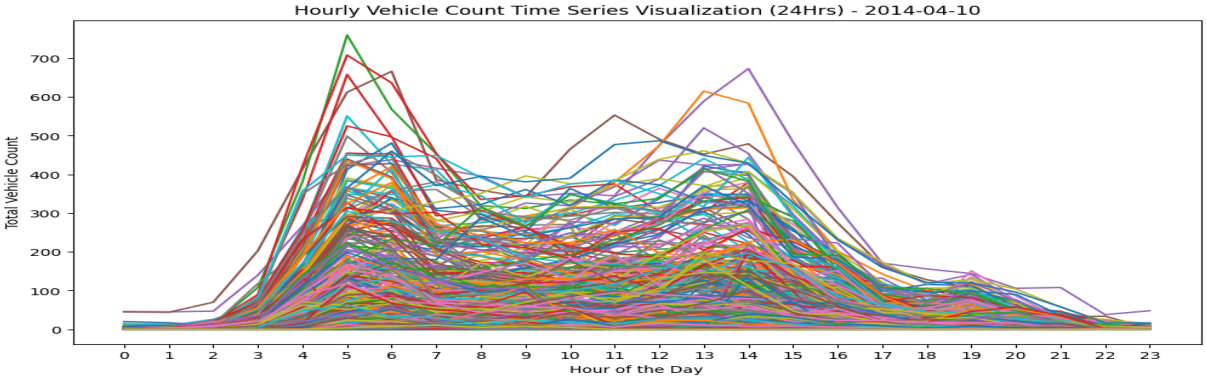

Figure 10: Time Series Graph visualizing vehicle counts for 10-04-2014 Thursday

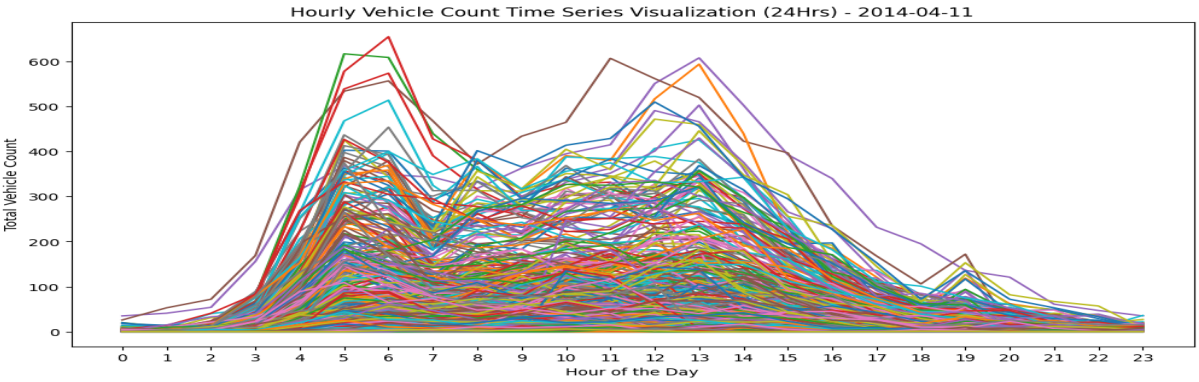

Figure 11: Time Series Graph visualizing vehicle counts for 11-04-2014 Friday

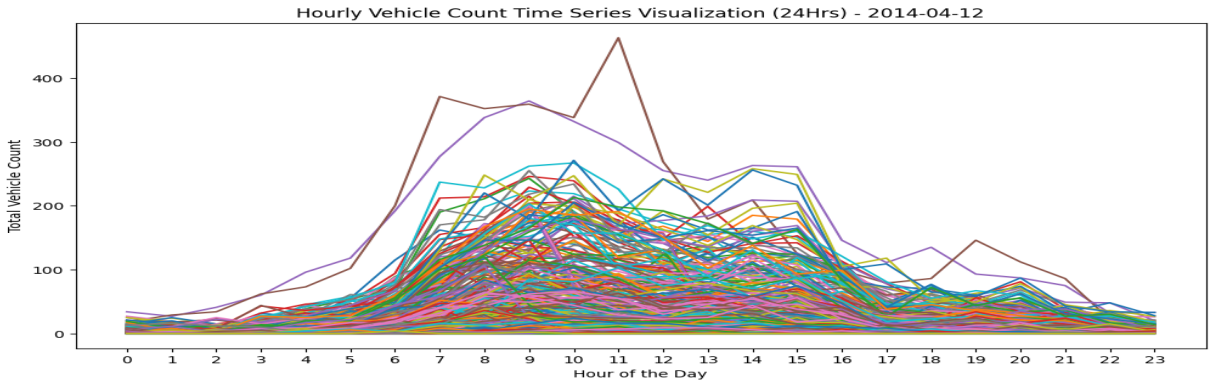

Figure 12: Time Series Graph visualizing vehicle counts for 12-04-2014 Saturday

Once the time series graphs and clusters have been obtained, the total count of each cluster for every hour of a day can be compared with the threshold conditions and plotted. The graphs below showcase the congestion pattern for each cluster ID at every hour for each day of the week.

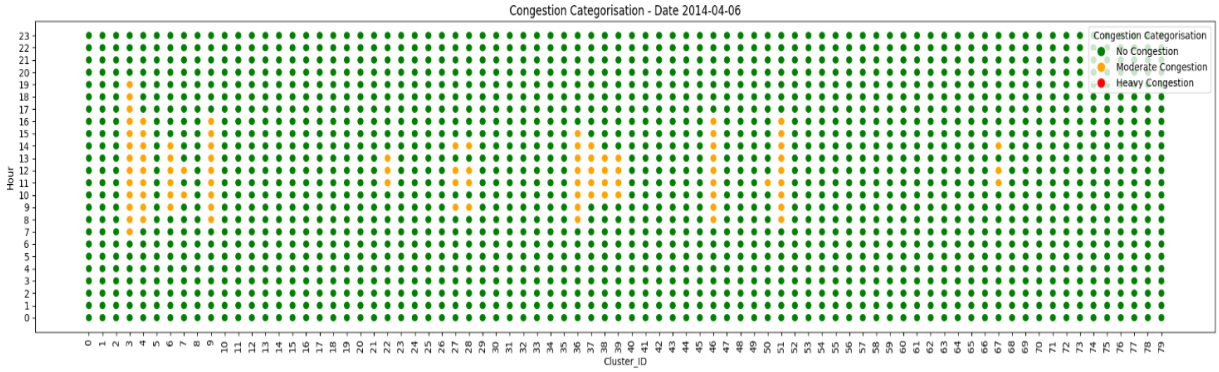

Figure 13: Congestion Categorisation for at each hour for all clusters on 06-04-2014 Sunday

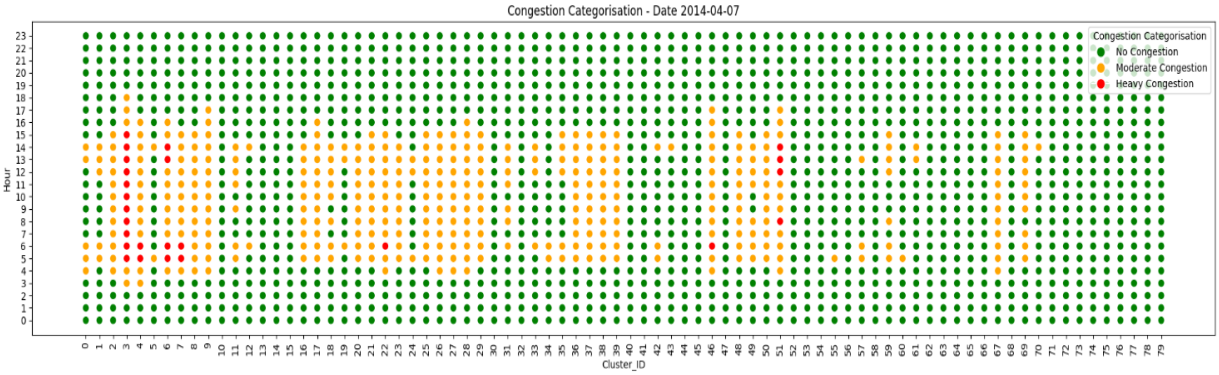

Figure 14: Congestion Categorisation for at each hour for all clusters on 07-04-2014 Monday

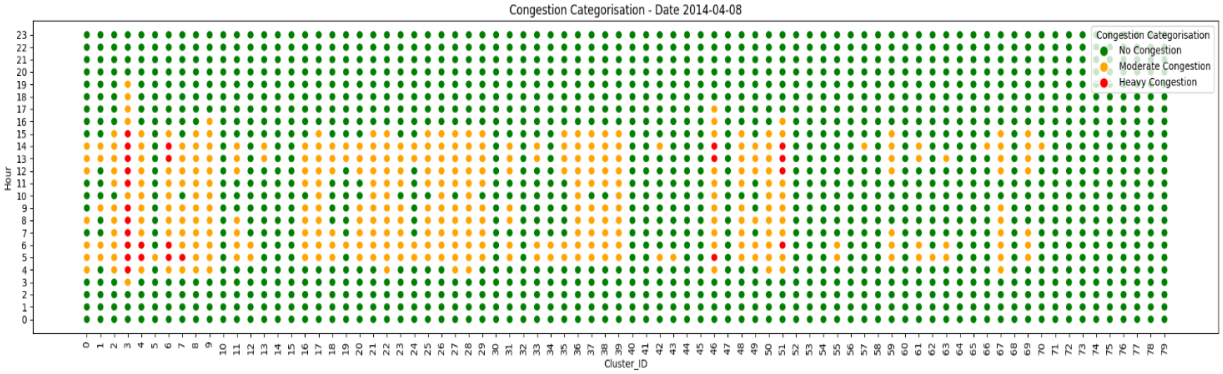

Figure 15: Congestion Categorisation for at each hour for all clusters on 08-04-2014 Tuesday

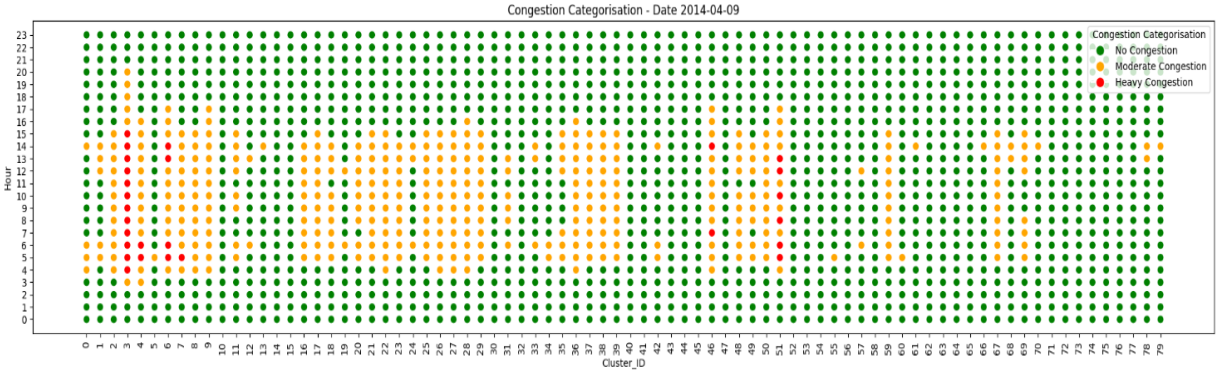

Figure 16: Congestion Categorisation for at each hour for all clusters on 09-04-2014 Wednesday

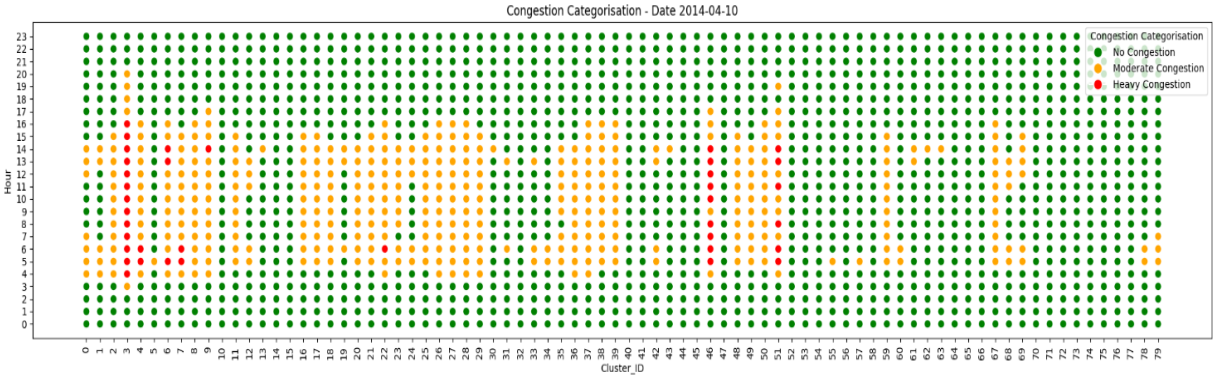

Figure 17: Congestion Categorisation for at each hour for all clusters on 10-04-2014 Thursday

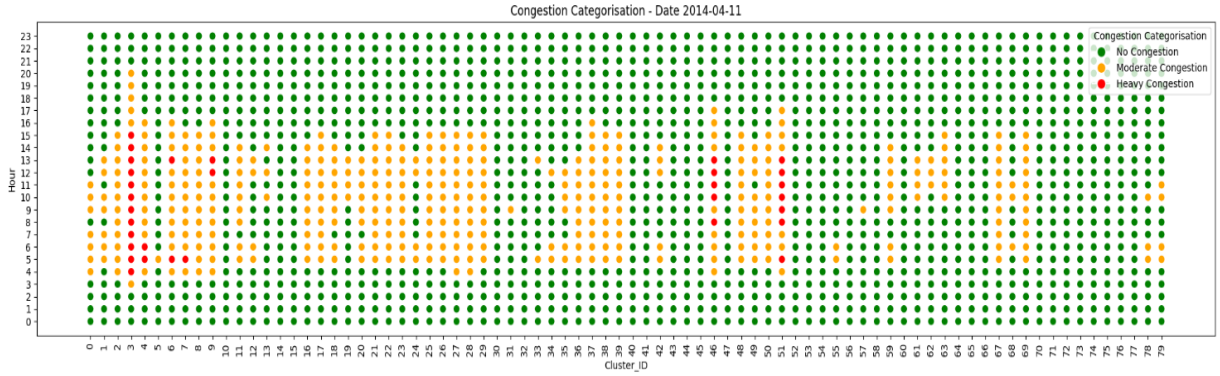

Figure 18: Congestion Categorisation for at each hour for all clusters on 11-04-2014 Friday

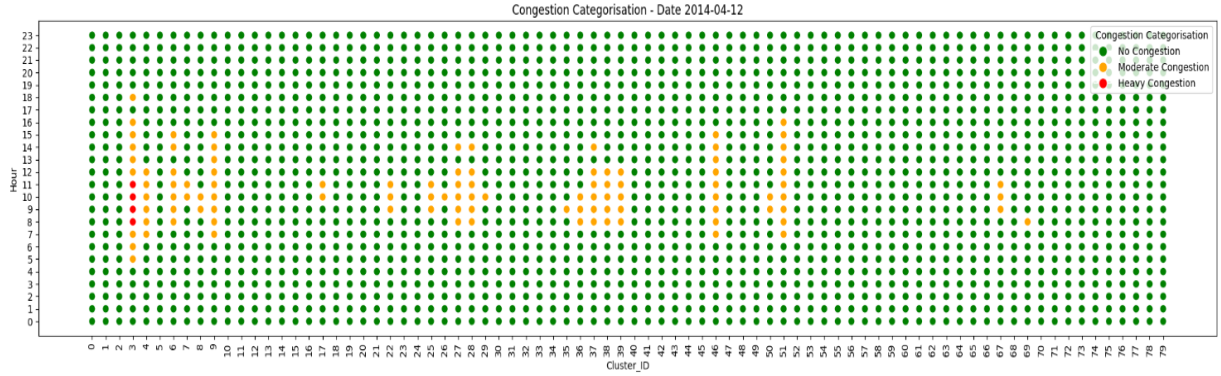

Figure 19: Congestion Categorisation for at each hour for all clusters on 12-04-2014 Saturday

Now that we have obtained the congestion patterns for all the clusters at each hour for 7 days. They can be plotted, on a map, to visualize the actual cluster formation and congestion categorization. Figures 20-22 show three clusters with cluster ID = 11 for day = 12-04-2014 at hour = 1300, cluster ID = 16 for day = 11-04-2014 at hour = 1300, and cluster ID = 51 for day = 11-04-2014 at hour = 1200 having the congestion patterns as shown.

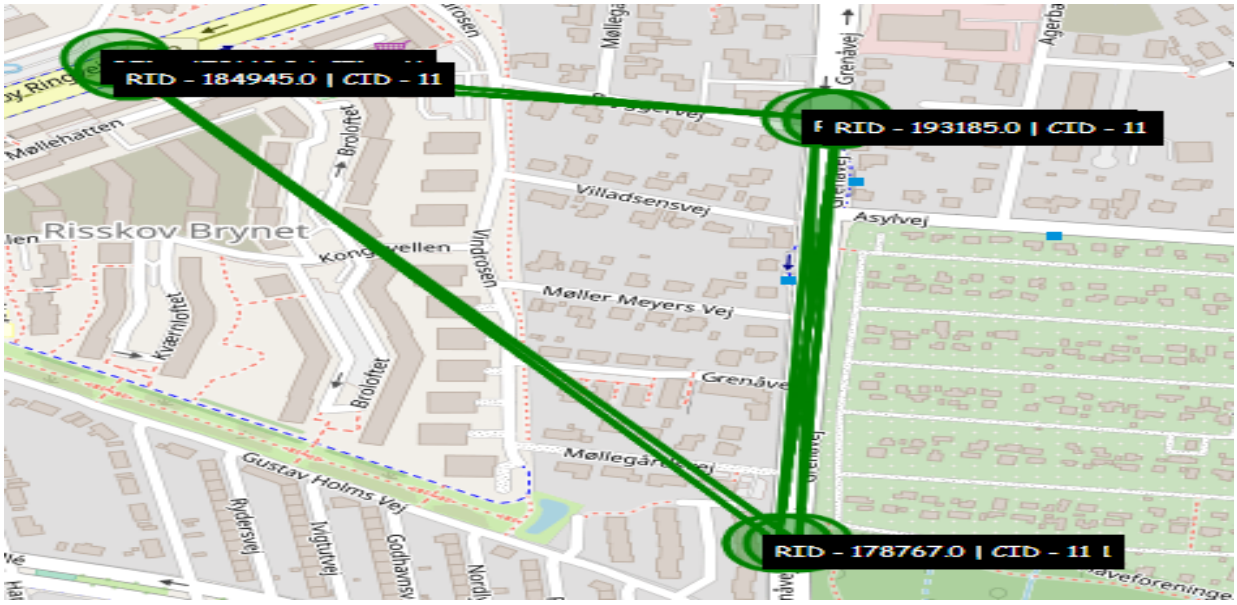

Figure 20: Congestion Categorisation for at hour = 1300 for clusters ID = 11 on 12-04-2014 Saturday

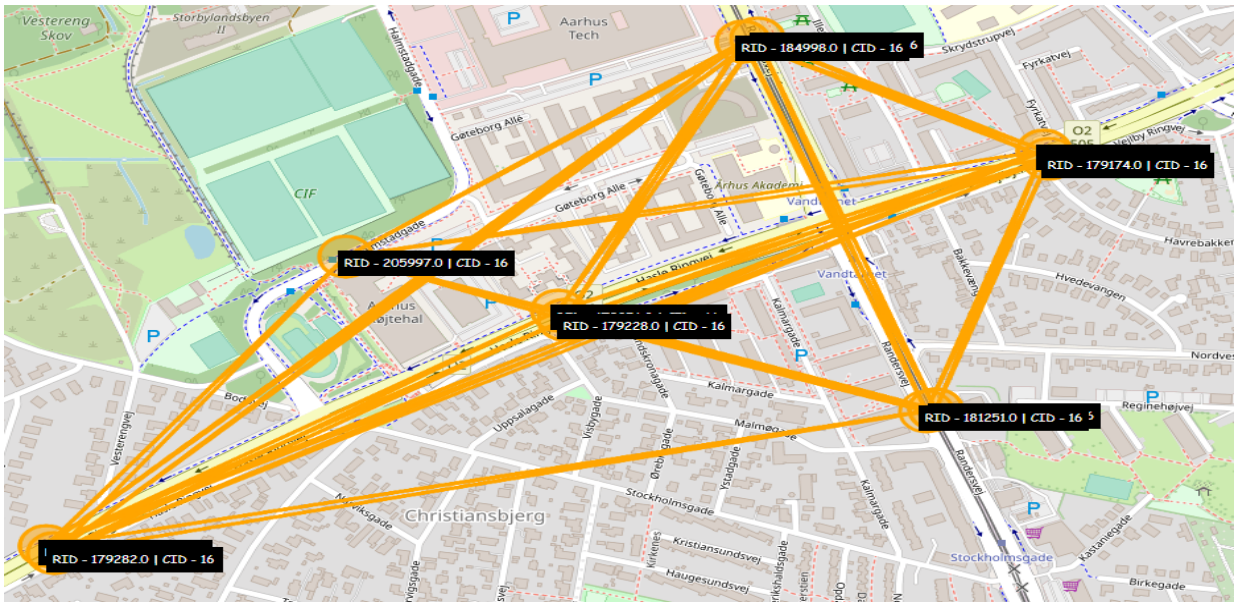

Figure 21: Congestion Categorisation for at hour = 1300 for clusters ID = 16 on 11-04-2014 Friday

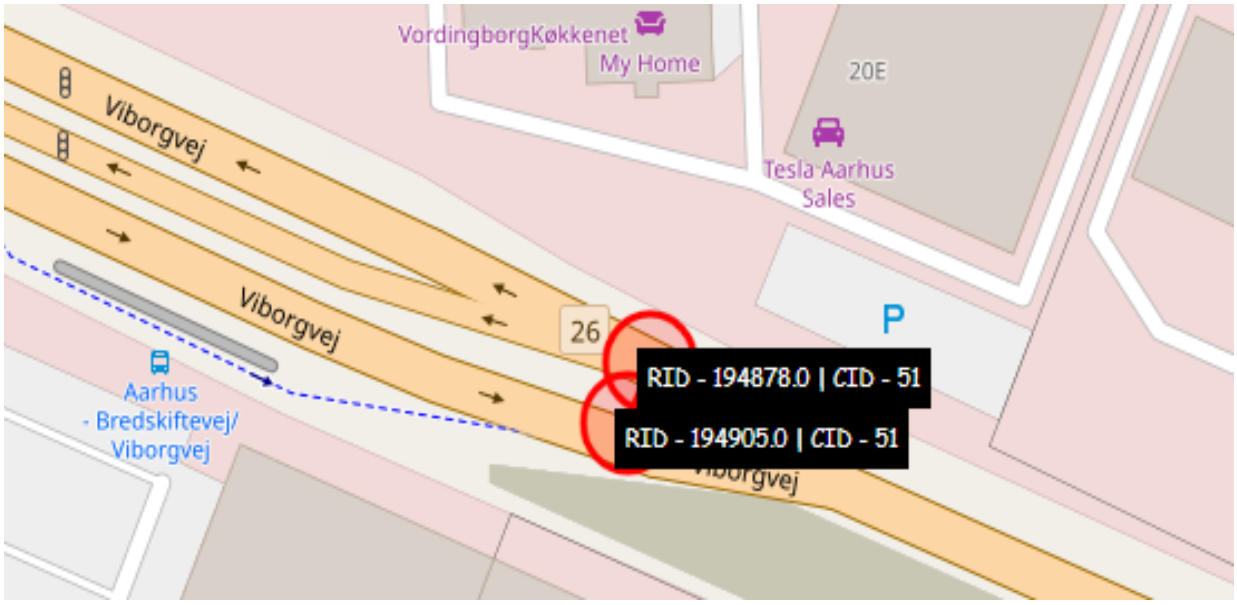

Figure 22: Congestion Categorisation for at hour = 1200 for clusters ID = 51 on 7-04-2014 Monday

7 Discussions

7.1 Project Outcome

Once the congestion patterns of each cluster at every hour for 7 days are obtained in Figures 13-19, a number of inferences can be made. Firstly, for 06-04-2014, which is a Sunday, we can see that no heavy congestion was observed across any of the clusters. This observation makes sense as traffic flow, on weekends is comparatively lesser as compared to weekdays, as most people have to travel to work, on weekdays. Similarly, we can also see that on 12-04-2014, Saturday, the number of heavy and moderate congestions obtained were very few, further supporting the previously mentioned hypothesis. Now for all the days between and including, 07-04-2014 to 11-04-2014, which are weekdays, we can see similar traffic congestion patterns are obtained. For all the weekdays, moderate to heavy traffic congestion is observed for most of the clusters, between the hours of 04:00 to 20:00. This further makes sense as most professional activities are performed between these hours.

Now we take a look at Cluster 3. Cluster 3 contains all the streets, which are part of the Østjyske Motorvej(East Jutland Motorway). Østjyske Motorvej is a motorway from the new Lillebæltsbro and on over the Vejlefjordbroen to Søften northwest of Aarhus. It is one of the busiest highways in Denmark, with approximately 65000 vehicles traveling on this highway per day [7]. As per our model, we see that cluster 3 is heavy to moderately congested for all days of the week, except for Sunday, between the hours of 03:00 to 20:00. This observation is very crucial as it confirms the accuracy and authenticity of our model.

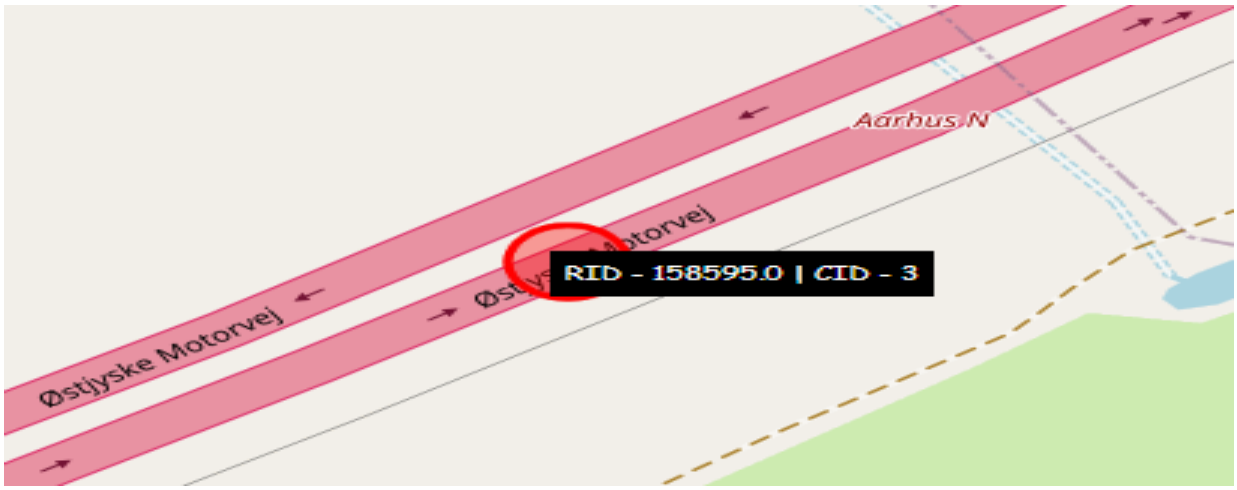

Figure 23: Congestion Categorisation for at hour = 1200 for clusters ID = 3 on 7-04-2014 Monday

7.2 Limitations

While the described project on traffic congestion detection through spatiotemporal analysis and data clustering demonstrates a comprehensive and innovative approach, it is important to acknowledge certain limitations. One notable limitation lies in the assumption that congestion is solely influenced by the number of vehicles and their spatial distribution, neglecting external factors such as accidents, weather conditions, or road maintenance, which can significantly impact traffic flow. The model's reliance on the quantile method for determining congestion thresholds may oversimplify the dynamic nature of congestion, as different clusters may have unique characteristics that cannot be fully captured by a uniform threshold. Additionally, the project's focus on a specific city, Aarhus, might limit the generalizability of the model to diverse urban environments with distinct traffic patterns and infrastructure. The accuracy of the

congestion predictions heavily relies on the quality and coverage of the traffic data, and any gaps or inaccuracies in the dataset could affect the reliability of the results. Moreover, the temporal aspect of the model, while capturing daily variations, may not fully address longer-term trends or seasonal influences on traffic congestion. In summary, while the project provides valuable insights, considering these limitations is essential for a nuanced understanding of its applicability and potential refinements for future implementations.

7.3 Future Work

For future iterations of this traffic congestion detection project, several avenues of improvement and expansion could be explored. First and foremost, incorporating additional factors such as weather conditions, road incidents, and special events would enhance the model's accuracy and reliability. Introducing real-time data streaming could also contribute to the dynamic nature of congestion prediction, allowing for prompt responses to changing traffic conditions. The project could benefit from scalability enhancements to handle larger datasets and accommodate different cities with diverse urban structures. Integrating advanced visualization techniques and geographical information systems (GIS) could provide more comprehensive insights into congestion patterns, aiding urban planners and policymakers in making informed decisions. Additionally, considering long-term trends and seasonality in traffic patterns could contribute to a more holistic understanding of congestion dynamics. Collaborations with local transportation authorities and the integration of smart city technologies could provide access to richer datasets and facilitate the implementation of real-world solutions. Lastly, continuous validation and refinement of the model against ground truth data and feedback from city planners and traffic management experts would be crucial for ensuring its practical utility and effectiveness in addressing the challenges posed by urban traffic congestion.

8 Conclusions

In conclusion, our project aimed to address the critical issue of traffic congestion in urban environments by developing an innovative spatiotemporal model for congestion identification. Leveraging advanced data clustering techniques and temporal analysis, our model successfully identified congestion patterns across various clusters in the city of Aarhus over a week. The results revealed a logical correlation between congestion and weekdays, with significant congestion observed during typical professional activity hours from 04:00 to 20:00. The accuracy of our model was further validated through the examination of Cluster 3, which encompasses the Østjyske Motorvej, a major Danish highway. The model accurately depicted the highway's congestion patterns, consistent with its status as one of the busiest in Denmark, with approximately 65,000 daily vehicles. Our approach, combining distance-based clustering and temporal analysis, provided a comprehensive understanding of the dynamic nature of traffic congestion. The introduction of a congestion threshold based on quantile values and the number of streets in a cluster added adaptability to the model. Overall, our model demonstrates the potential for effective traffic management and urban planning through data-driven methodologies, contributing to the broader discourse on addressing congestion for improved economic productivity, environmental sustainability, and enhanced quality of life in urban areas.

9 Appendix

The entire code for the Traffic Identification model can be found at this GitHub Repository.

References

[1] D. Toshniwal, N. Chaturvedi, M. Parida, A. Garg, C. Choudhary, and Y. Choudhary. "Application of clustering algorithms for spatio-temporal analysis of urban traffic data". In: *Transportation Research Procedia* 48 (2020), pp. 1046–1059.

[2] M. Ashifuddin Mondal and Z. Rehena. "Intelligent traffic congestion classification system using artificial neural network". In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 110–116.

[3] D. Birant and A. Kut. "ST-DBSCAN: An algorithm for clustering spatial–temporal data". In: *Data & knowledge engineering* 60.1 (2007), pp. 208–221.

[4] *Smart City Denmark Traffic Dataset*. `https://www.kaggle.com/code/tengrihan/smart-city-denmark-traffic-dataset/input`.

[5] D. A. Prasetya, P. T. Nguyen, R. Faizullin, I. Iswanto, and E. F. Armay. "Resolving the shortest path problem using the haversine algorithm". In: *Journal of critical reviews* 7.1 (2020), pp. 62–64.

[6] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady. "DBSCAN: Past, present and future". In: *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. IEEE. 2014, pp. 232–238.

[7] *Østjyske Motorvej*. URL: `https://da.wikipedia.org/wiki/%C3%98stjyske_Motorvej/`.