Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Application of clustering algorithms for Spatio-Temporal analysis of urban traffic data

Durga Toshniwal, Narayan Chaturvedi, Manoranjan Parida, Archit Garg, Chirag Choudhary, Yashpal Choudhary

Indian Institute of Technology, Roorkee 247667, India

Abstract

The large vehicle movement traffic datasets offer a lot of great opportunities for the evolution of new methodologies for the analysis of the transportation system. However, deriving relevant traffic patterns from such a vast amount of historical dataset is challenging. In this paper, several data mining techniques have been applied to obtain more understanding about urban traffic patterns by analyzing hourly and daily variation in urban traffic flow dataset. A model has been developed for the analysis of spatial and temporal patterns in urban traffic data. Model development involves the formulation of algorithms to be applied on the data and choice of various metrics to evaluate the clustering algorithm. Furthermore, these techniques have been applied to the traffic dataset of Aarhus, the second largest city of Denmark. Finally, results are analyzed to determine the various factors that affect the traffic flow patterns in an urban area.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Clustering; Temporal Variations; Spatial Analysis; Silhouette Coefficient; Dunn Index; Euclidean distance; Dynamic Time Warping (DTW); PAM; AGNES; Traffic Volume,

1. Introduction

In today's era, rapid urbanization not only brings a large physical infrastructure, but also brings a large amount of data infrastructures obtained from the sensor networks, government records databases, and social media data. The data combined from all these sources is a very powerful resource for making cities as the centre of innovation. Data mining provides us with many methods that are capable of knowledge extraction, clustering and classifying data, decision making and future prediction of values and status. Application of data mining in traffic management provides us with major benefits of traffic prediction and making the appropriate decision for it. This becomes more pragmatic as manual analysis of such huge data acquired and its frequent incoming makes it impossible.

* Corresponding author. Tel.: +91-9760316522
E-mail address: narayanchaturvedi@gmail.com

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

Due to economic, geographical, national and international developments, mobility has been increasing. However, this non-stop rise in transportation needs has adverse effects, such as air pollution and traffic congestion. In order to decrease this impact, various measures can be employed. For instance road pricing (toll taxes, etc), route guidance, intelligent traffic management system etc. It is therefore important to have knowledge about the traffic flow within the city to take important traffic management measures.

The process of traffic analysis can be done to obtain the detailed understanding about the functioning of traffic system along with its root phenomenon. Knowledge about the state of the traffic at various places and instances of times provides valuable information which may help estimate the location and times of bottlenecks in the city. It can also provide information regarding about the spare capacity of roads at different times and help to estimate the robustness of the network. Analysis of historical data, when combined with live data can be used to determine irregularities in traffic patterns, such as accidents. Such knowledge can be used to determine where and when improved traffic management is required. [Pascale et al., 2015] have worked on city centre London dataset collected through loop detectors for spatio-temporal analysis of city to find the similar areas over the space and time. [Lopez et al., 2017] used K-means and DBSCAN clustering methods to detect the urban traffic dynamics of Amsterdam city.

The objective of this study is to derive more knowledge about urban traffic patterns that may later be used for improving road and traffic management system of the urban area.

2. Literature Review

Researchers have mainly discussed traffic on highways. An obvious difference between traffic in cities and traffic on highways is the existence of multiple types of traffic modes on city roads network such as buses, trucks, cars, pedestrians, bicycles, while highways are mainly used by cars and heavy-duty vehicles such as buses and trucks. Another important characteristic is that city traffic network contains many intersections. Due to this, the traffic situation in cities shows many minor irregularities, in contrast to highways which generally show less irregularity. The city network managing a significant amount of local or short distance traffic also serves moderate large distance traffic to and from highways.

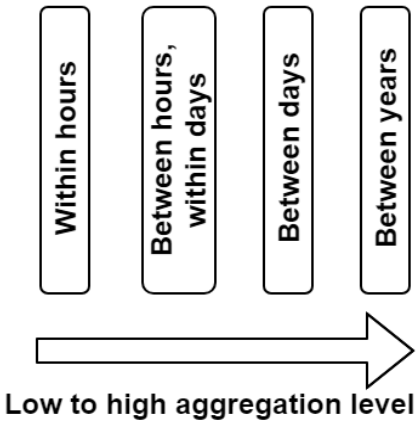

Fig.1. Common time scale for analysis of temporal variations

2.1. Temporal Variations

Time-based analysis of traffic data can be performed on different levels, varying from per minute variations to monthly variations. The variation dimensions for traffic volumes are:

Short-Term Variations: A distinction can be made on the basis of several factors like the day is a working day or not, what type of road is being analyzed, the difference in the type of vehicles, etc. An average working day shows a peak in the early morning, a peak in evening, and sometimes an off- time peak also. [Chrobok et al., 2004]

compared traffic on Dutch highways and observed that the evening peak is heavier than the morning peak. It also states that total delay is approximately the same for all weekdays, while morning peak and evening peak delays vary considerably between weekdays. Considering the traffic demand changes whole day, [Wang et al., 2016] examined morning peak, Noon off-peak and evening peak hours separately and identified factors with speed variation. According to the research done by [Chrobok et al., 2004], the traffic pattern of Mondays to Thursdays has a sharp AM peak and a higher and wider P.M. peak. According to [Stathopoulos and Karlaftis, 2001], there is no significant difference between different working days of the same week.

Studies on the influence of weather factors on traffic have given varying results. The effect of unseasonable weather on traffic has been investigated by [Al Hassan and Barker, 1999]. It was observed that the traffic increases in case of higher than expected sunshine hours or temperature whereas in case of lower than expected sunshine hours or temperature or higher than expected rainfall, the traffic volume decreases. Also, the stronger effect on weekends has been observed as compared to weekdays.

Long-Term Variations: Due to various economic, geographical, political and demographic factors, traffic volumes usually also vary over the years. [Festin, 1996] observed the traffic volumes in the USA between 1970 and 1995. The amount of traffic naturally increased during the period. According to his study, the period between 1985 and 1995 shows less amount of seasonal variation in traffic as compared to the period between 1970 and 1984. Weekly patterns appear to remain stable over time. Finally, he also found that weekday traffic is becoming increasingly concentrated between 5 AM – 6 PM. The afternoon peak is spreading faster into the middle of the day than as compared to the evening hours. Locally, land use and infrastructural developments determine long-term variations in traffic volumes.

2.2. Spatial Variations

Spatial analysis in urban traffic is done to identify areas with high traffic volume for a certain period of time. A spatial variation also provides the flow of traffic volume along different locations.

Spatial Variation of Traffic Volumes: Analysis of spatial variations in traffic data (without taking into considerations time variations) can be used to identify locations with high traffic volumes. [Zhao and Chung, 2001] and [Liu and Sharma, 2006] concluded that the expected traffic volumes on a road depend on its function. They developed a method for estimating average annual daily traffic (AADT) on roads for which no traffic volume counts were previously recorded.

Spatio-temporal traffic variations: The movement of a traffic through the network or the build-up of congestion are referred to as spatio-temporal traffic patterns. It can be used in traffic forecasting. [Stathopoulos and Karlaftis, 2003] declared that the routine daily flow observed was very similar for two detectors which were at a distance of less than 250 m from each other. The patterns observed were still quite similar even for larger distances.

Table 1. Attributes of a single record

Attribute name	Description
avgMeasuredTime	denotes number of seconds for which sensor records data
avgSpeed	denotes average speed of vehicles passing in the observed time duration
extID	denotes unique identifier assigned to each road
medianMeasuredTime	same as avgMeasuredTime
TIMESTAMP	denotes the start time measurement of traffic on a road
vehicleCount	number of vehicles passing between the two points of observation
_id	unique id assigned to each measurement of traffic
REPORT_ID	denotes unique id assigned to each place(road) of observation

3. Data Description and Preprocessing

In this study, traffic data provided by the city of Aarhus has been used to analyze traffic patterns in a metropolitan area. Aarhus is the second largest city (approximately 319,680 inhabitants as of 2003) in Denmark. Aarhus, the largest cultural city in Denmark lies on the northern latitude resulting variation in daylight hours between summers and winters. The data for the months of February 2014 to June 2014 is analyzed to discover patterns in traffic using several clustering techniques. The data has been collected by placing sensors at two nearby locations and counting the number of vehicles passing through them for a fixed duration of time. The data is available for 449 roads in CSV format. For each road, the record contains a traffic count measured by a certain detector on some road on a certain day for a certain time duration. Table 1 shows the attributes of the single record.

Besides the traffic volume record of every observation point, citypulse also provide meta information regarding the sensor locations used for measuring the traffic volume. This information is helpful in geographically locating the sensor on maps, and thereby creating a flow network of traffic volume. The metadata also contains useful information regarding the sensors such as a distance between two centers, duration of measurement, major roads etc. Each record of metadata contains information regarding the links/roads connecting the two sensors. A single record contains the following relevant variables as shown in Table 2:

Table 2. Variables of a single metadata record

Attribute name	Description
POINT_1_NAME	denotes an unique identifier of first observation point
POINT_2_NAME	denotes an unique identifier of second observation point
POINT_1_LAT, LNG	denotes latitude and longitude of first sensor location
POINT_2_LAT, LNG	denotes latitude and longitude of second sensor location
RECORD_ID	denotes unique identifier corresponding to the traffic record of street
DISTANCE	contains the distance between two sensor locations
DURATION	denotes duration of traffic measure

3.1. Data Preprocessing

The available traffic data is produced by the sensors located at different locations across the city. The number of vehicles given in the raw data is aggregated in 5 minutes, 15 minutes, 30 minutes or 1-hour intervals. The traffic data is processed so that it can be used for the mining of traffic patterns. Some traffic roads contain insufficient data and are cleaned. The various pre-processing techniques applied. The pre-processing is done in such a way that the final records formed, which are used for analysis contain all traffic counts for one street on a particular day. This pre-processed data is fed to data validation procedure in order to remove the invalid data. Various quality checks are evaluated and finally, all the processed records are aggregated, and the traffic data is linked to the type of day and month corresponding to the given date.

Preprocessing on individual streets: In the traffic data, we only require the traffic count at a certain interval of time for all the days. Therefore, we do not need information about the road (extID), the speed of cars (avgSpeed), measurement times (avgMeasuredTime and medianMeasuredTime) and other unique identifiers. We have reduced our data set so that it contains only relevant data fields used for mining patterns. Our records for a particular street now contain two fields: timestamp and vehicle Count.

Some of the records contain a deficient timestamp or period length. A deficient timestamp means that period of observation is shifted in time. In order to include such deficient timestamps or periods, vehicle Count has been linearly interpolated. These interpolated values are then used for various quality checks.
Data cleaning and transformation: In order to remove the records showing erroneous values, following quality checks have been employed.

1. Sometimes, low-quality sensors may have also recorded abruptly large traffic volumes. These traffic volumes largely affect the process of interpolation. Records having traffic count values greater than 100 are removed from the dataset.
2. If there is no value obtained for four consecutive hours in a day, then that day is marked invalid. If there are more than 20 percentage invalid days for a street, then that street is not considered for further analysis.

4. Modeling Approach

To meet the research goals, a model named CSTAT is proposed which is applicable to any traffic data set, though minor modifications may be required. The workflow of the proposed approach is shown in figure 2.

4.1. Pattern representation

Cluster analysis requires a pattern or a data point to be described by a number of attributes [Liao, 2005]. The CSTAT model suggests to describe the traffic pattern for a day as follows in figure 2:

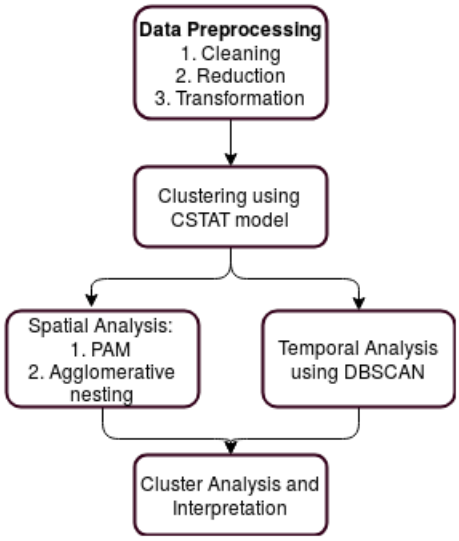

Fig.2. Workflow of the proposed approach

Profile (P): daily flow profile on street s at day d:

$P_{sd} = (q_{sd,1}, q_{sd,2}, q_{sd,3}, q_{sd,4}, ..., q_{sd,t})$

Where q is a traffic volume measurement indexed by street s on day d and a fixed time interval. The number of traffic measurements is dependent on the aggregation level chosen. The aggregation level or time interval for measurement chosen is such that short-term fluctuations in the data set are eliminated, which may have caused similar data points to be grouped into separate clusters.

4.2. Spatial Analysis

Two different algorithms for performing spatial analysis are used. These are Partitioning Around Medoids (PAM) and Agglomerative Nesting (AGNES). The first one is a partitioning clustering technique while the latter is a hierarchical one. Both of them use the Euclidean distance and the Dynamic Time Warping (DTW) distance metrics.

Partition Around Medoids (PAM): The aim of PAM [Kaufman and Rousseeuw, 1990] algorithm is to partition the data set of n objects into k clusters. This algorithm works on a dissimilarity matrix, with the goal of minimizing the total dissimilarity between the centroids of each cluster and its respective members. The algorithm uses the following function:

$$F(x) = \min \sum_{i}^{n} \sum_{i}^{n} d(i, j) z_{ij} \tag{1}$$

Where:

1. $\sum_{i=1,n} z_{ij} = 1,\ j = 1,2,...n$
2. $z_{ij} \le y_i, i, j = 1,2,..., n$
3. $\sum_{i=1,n} y_i = k, k = number\_of\_clusters$
4. $y_i, z_{ij} \in 0,1, i, j = 1,2,..., n$

Here F is the function to be minimized, $d(i, j)$ the dissimilarity measure between entities i and j and z_{ij} is a variable that is multiplied to ensure that only dissimilarity between elements from same clusters are considered in the function F. PAM works on two kinds of input data, the first one is the matrix that represents every input entity and the other is the dissimilarity matrix directly. In the analysis, the algorithm proceeds this way:

1. Choose k entities as initial medoids.
2. Calculate the dissimilarity matrix using a predetermined distance measure, such as Euclidean distance.
3. Assign every entity to the closest medoid.
4. For every cluster, search if any of the entities currently not a medoid lowers the average dissimilarity coefficient. If yes, select the entity that lowers this coefficient the most as the new medoid for the cluster.
5. If at least one medoid changes go to (3), else the algorithm terminates.

Agglomerative Nesting (AGNES): AGNES [Gowda and Krishna, 1978] is a type of hierarchical clustering technique which groups data objects into a tree of clusters. It begins by creating clusters composed of single data objects, and then iteratively, using some distance metric merges such clusters into larger ones. This second step is repeated several times until a single cluster is obtained.

Dynamic Time Warping (DTW): Dynamic time warping or DTW algorithm[Berndt and Clifford, 1994] is used for calculating the similarity between two temporal sequences that may differ in speed or time making it useful for time series analysis.

4.3. Temporal Analysis

For temporal analysis of streets, we used the DBSCAN algorithm for clustering the days of the week. We chose streets randomly from various spatial clusters and performed clustering for all the days in the dataset. The choice of DBSCAN algorithm was based on the observation that traffic volume patterns of days in adjacent weeks are very much similar to each other. DBSCAN, being a density-based algorithm is ideal for clustering data containing densely spaced points.

DBSCAN (Density Based Spatial Clustering for Applications with Noise): The algorithm as suggested by [Ester et al., 1996] classifies the elements of the database in two different types:

1. border points: located on/near the boundary of a cluster.
2. core points: located in the inner regions of a cluster.

The neighborhood of a point p is the set of points with a distance less than a threshold value eps. A point p is said to be directly density-reachable from another point q if it is present in the neighborhood of q and q's neighborhood size is greater than a fixed constant MinPts. Along with this, it defines a density-reachability criterion: a point p is called density reachable from q if a chain of points $p_1 \ldots p_n$, where $p_1 = q$, $p_n = p$, and p_{i+1} is directly density-reachable from p_i. These concepts are then used to define a cluster D:

1. For every q, p : if p belongs to D, and if q is density-reachable from p and p's neighborhood is greater than the MinPts threshold, then D contains q
2. For every q, p : if q, p belongs to D then there must exist a point t such that the points p and q are directly reachable from t

There may be some points which do not belong to any cluster. Such points are considered as noise points or, outliers. The algorithm randomly chooses a point p, marks it as visited, and computes its eps-neighbourhood. If it is a core point, then a new cluster is created, and the procedure is called for all points in p's neighbourhood. Otherwise, p is marked as a noise point. To find next cluster, algorithm chooses an unvisited point randomly, until all points have been visited.

4.4. Measuring effectiveness of clustering

The correctness of clusters formed can be measured on basis of two measures:

1. Cluster Cohesion: It measures how close objects are related to each other within a cluster. Cohesion can be measured by calculating the within-cluster sum of squares (WSS).
2. Cluster Separation: It measures how well-separated a cluster is from other clusters. Separation is measured by the distance between cluster sums of squares.

Silhouette Coefficient

Measuring [Aranganayagi and Thangavel, 2007] provides a method of interpretation and validation of clustering. The silhouette value takes into consideration both cluster cohesion as well as cluster separation. It ranges from [-1,1], where a high positive value indicates that the object is more similar to its own cluster compared to its nearest cluster. Assuming that the data has been clustered into K clusters. For every data item i, a(i) is the average distance of i from all other data points of the same cluster. b(i) is defined as the lowest average distance to any of the remaining clusters.

An s(i) value close to 1 shows that the data point is correctly clustered. An s(i) near zero means that the data point is on the border of two natural clusters to which it can belong. Thus, it shows how tightly grouped all the data points of the cluster are to each other.

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & if\ a(i) < b(i) \\ 0, & if\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & if\ a(i) > b(i) \end{cases} \tag{2}$$

Dunn Index

Dunn index (DI) is a measure for evaluating the quality of clusters produced. Defined as the ratio of the minimal inter-cluster distance and maximal intra-cluster distance, a high Dunn index indicates that the clusters formed are dense and well separated.

$$D = \frac{\min_{(1 \le i < j \le n)} d(i, j)}{\max_{(1 \le k \le n)} d'(k)} \tag{3}$$

5. Experiments and Results

5.1. Temporal Traffic Patterns

Daily flow profiles for different days are grouped using the DBSCAN algorithm for clustering. Data for working as well as non working days were grouped together for all roads for Feb-June months (105 days).

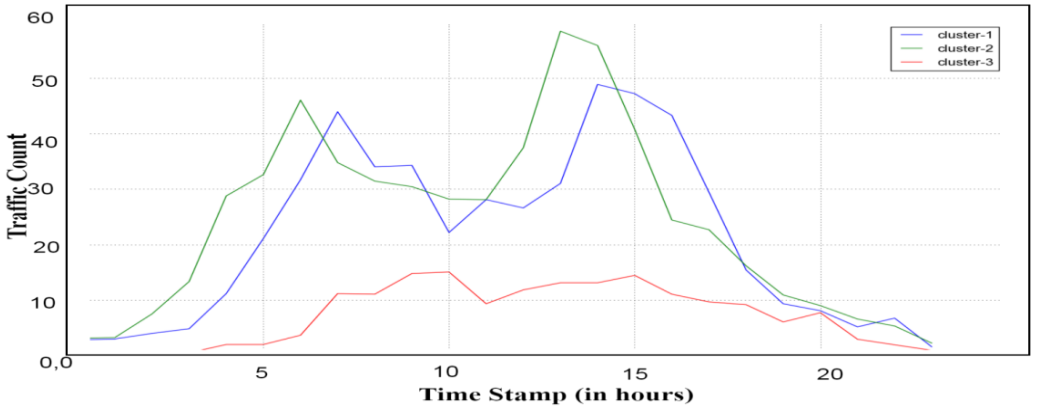

Fig.3. Temporal traffic pattern representing 3 clusters

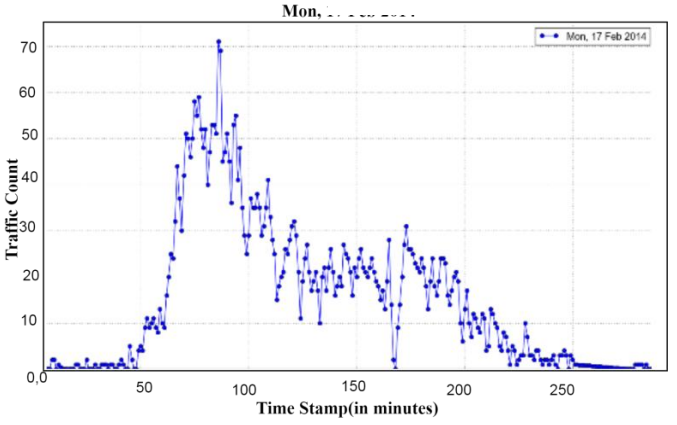

Fig.4. (a) 5 Minutes

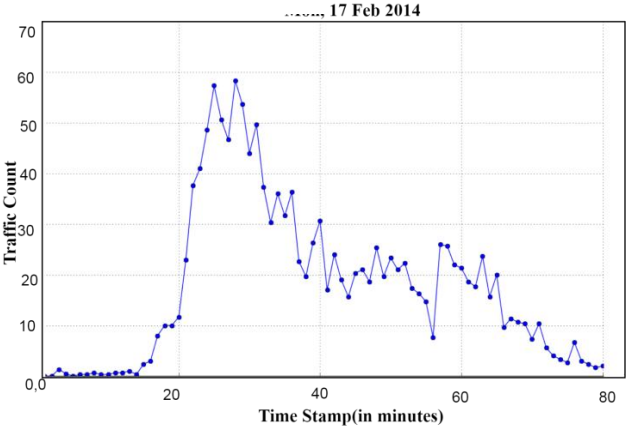

Fig.4. (b) 15 Minutes

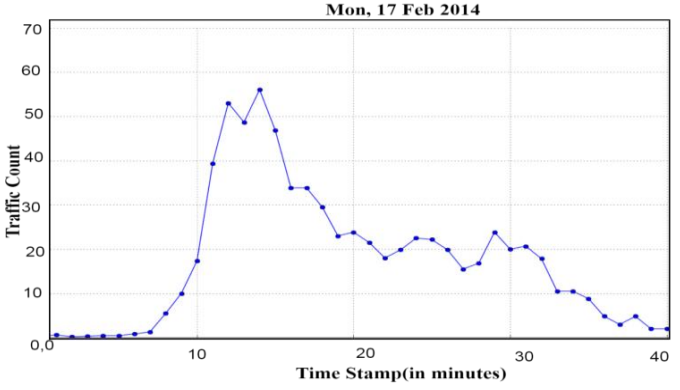

Fig.4. (c) 30 Minutes

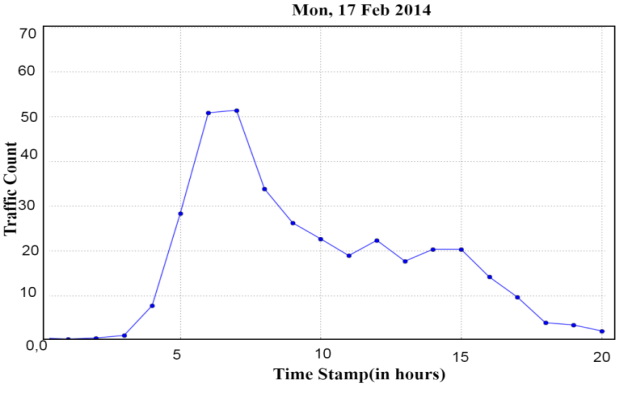

Fig.4. (d) 60 Minutes

Fig.4. Daily traffic plots for various aggregation levels

Clustering Analysis: Clustering is carried out for 320 streets (i.e number of streets for which enough data was available). The number of different clusters formed as well as the resulting traffic patterns differ by location. For the majority of streets, days are classified into mainly two clusters i.e weekdays or weekends. For some number of cases, all the days are clustered into 3 major groups. Detailed analysis for many such roads is carried out. In all cases, similar results are obtained. All working, as well as non-working days for a road, are clustered using DBSCAN and the resulting clusters formed are as shown in Figure 3. All the days for a given street get split into 3 clusters. The cluster depicted by redline consists of all the Saturdays and Sundays along with some weekdays. These weekdays are either public holidays or rainy days when the traffic on roads is drastically reduced as compared to their daily profile. The remaining two clusters contain a mixture of weekdays. A certain type of day, say Monday, is distributed over both the clusters. Mondays in the months of mid-February to the beginning of April are all present in cluster depicted by blue color, whereas the Mondays thereafter, that is from April 2nd week to the end of May, are present in cluster depicted by green color. Saturdays and Sundays remain in the same cluster. The shift of Mondays to a different cluster over time can be explained on the basis of seasonal changes. During this period, there is an increase in average temperature and decrease in humidity levels. People, with change in weather, tend to change their working hours. This shift in clusters shows the shift in morning traffic peak from around 7 AM to nearly 6 AM in the morning. This clearly shows a change in the general lifestyle of people with change in weather. In order to estimate the epsilon in DBSCAN algorithm, we have used K-Nearest Neighbour Plot (K-NN plot) with K being equal to number of MinPts used in algorithm. As depicted in the Figure 5, we have chosen epsilon close to 19 meters as there is the large increment in epsilon thereafter. The number of MinPts are chosen is such a way that it maximizes the silhouette coefficient.

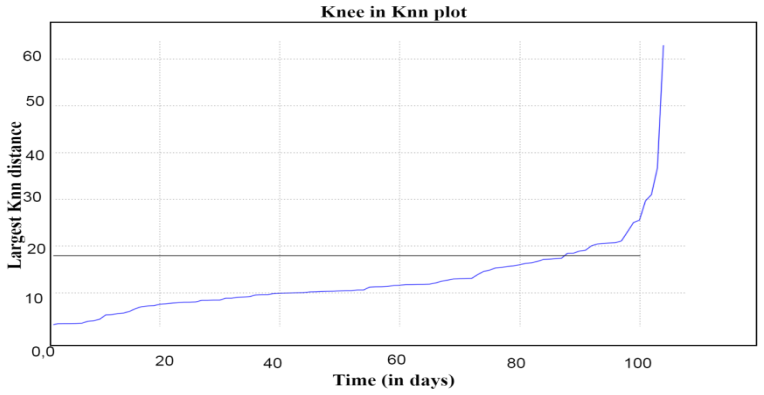

Fig.5. KNN plot for choosing epsilon in DBSCAN algorithm

5.2. Spatial Traffic Patterns

The working day profiles for roads remaining after preprocessing were clustered using various clustering algorithms as follows:

1. Partition Around Medoids (partitional clustering).
2. Agglomerative Nesting (hierarchical clustering).

We have used both Euclidean distance and Dynamic Time Warping (DTW) distance as two distance measures for calculating dissimilarity between data points. For calculating the quality of clustering, we have used the following measures:

1. Dunn Index (DI)
2. Average Silhouette Coefficient (S)

The quality measure for each of the clustering algorithm is compared along with different distance metric, and then appropriate resultant clustering is chosen.

Clustering Analysis: This section contains a comparison of two clustering algorithms, with two different distance metric. Also, in order to define the number of clusters to be used in the algorithm, we plotted the values of Dunn Index (DI) and Silhouette Coefficient (S) for various values of K and chose the values of K for which maximum values of these measures were obtained.

Partition Around Medoids (k medoids partition algorithm)

- For PAM using Euclidean distance, the maximum value of Silhouette coefficient is obtained for K = 5, while for Dunn Index, maxima occur at K = 6 as shown in Figure 6.
- For PAM using DTW distance, the maximum value of both Silhouette coefficient and Dunn Index occur at K = 6 as shown in Figure 7.

Agglomerative Nesting (Hierarchical clustering)

- For AGNES using Euclidean distance, the maximum value of Silhouette coefficient is obtained for K = 4, while for Dunn Index, maxima occur at K = 7 as shown in Figure 8.
- For AGNES using DTW distance, the maximum value of both Silhouette coefficient and Dunn Index occurs at K = 6 as shown in Figure 9.

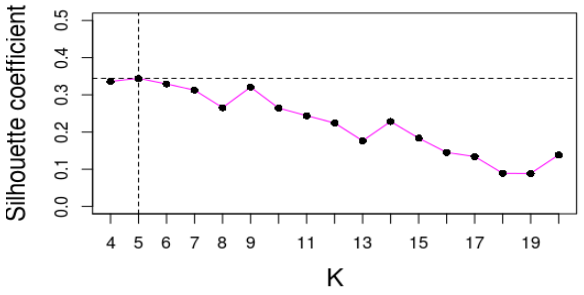

Fig.6. (a) Silhouette coefficient vs no. of clusters

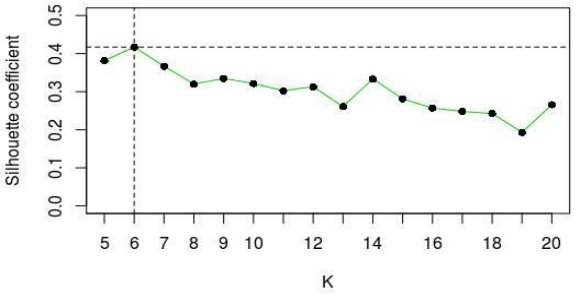

Fig.7. (a) Silhouette coefficient vs no. of clusters (K)

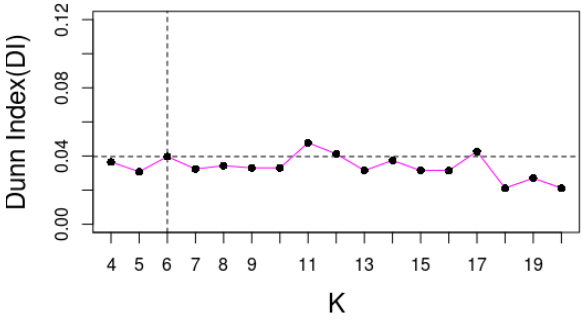

Fig.6. (b) Dunn index vs no. of clusters (K)

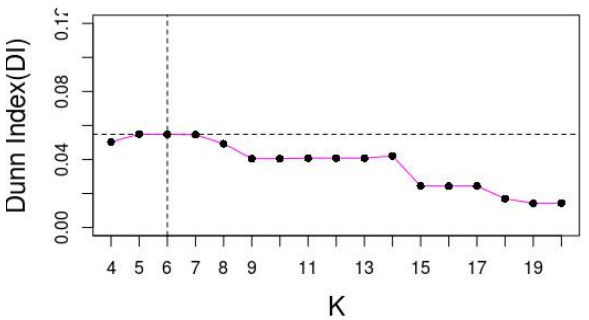

Fig.7. (b) Dunn index vs no. of clusters (K)

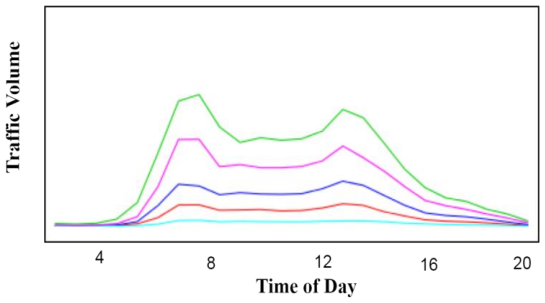

Fig.6. (c) Clusters mean with K=5

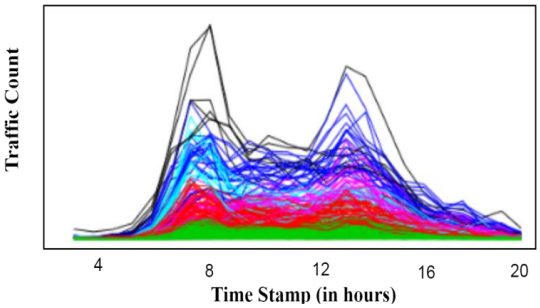

Fig.7. (c) Clusters mean with K=6

Fig.6. Clustering using PAM algorithm and Euclidean distance metric

Fig.7. Clustering using PAM algorithm and DTW distance metric

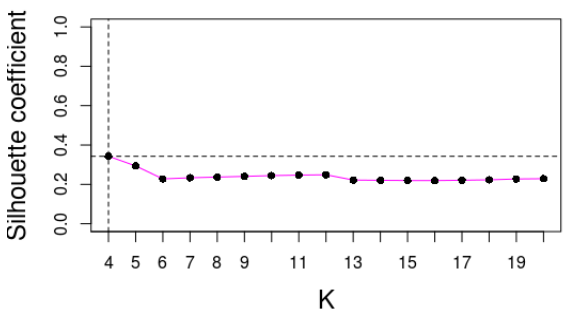

Fig.8. (a) Silhouette coefficient vs no. of clusters

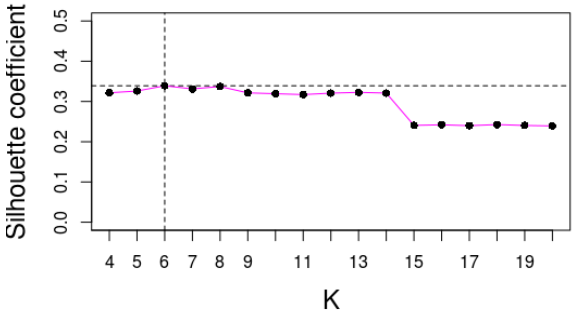

Fig.9. (a) Silhouette coefficient vs no. of clusters

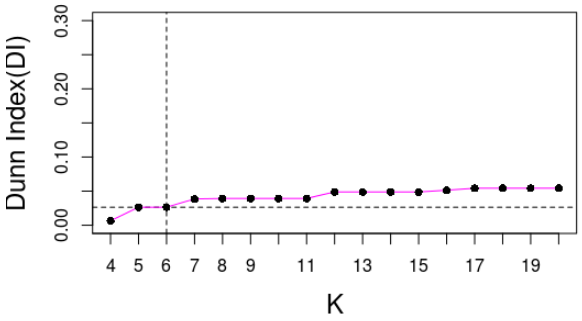

Fig.8. (b) Dunn index vs no. of clusters (K)

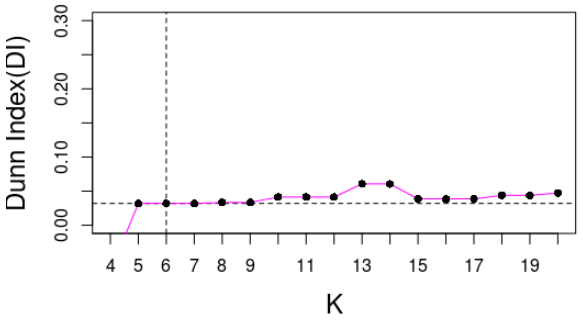

Fig.9. (b) Dunn index vs no. of clusters (K)

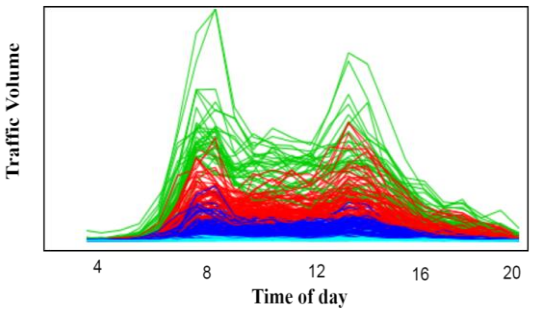

Fig.8. (c) Clusters mean with K=4

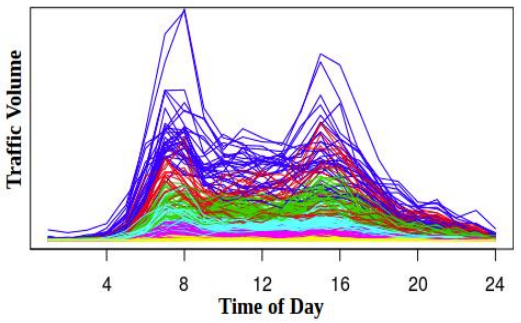

Fig.9. (c) Clusters mean with K=6

Fig.8. Clustering using AGNES algorithm and Euclidean distance metric

Fig.9. Clustering using AGNES algorithm and DTW distance metric

PAM versus AGNES

From the above analysis, we can observe that PAM algorithm with DTW distance measure gives better values of average Silhouette coefficient and Dunn Index. Thus, PAM with DTW distance measure is used for analyzing the Aarhus traffic dataset using Google maps.

Clusters Interpretation: Each of these cluster shown in figure 6(c) and figure 7(c) have been examined separately and similarity between every street belonging to same clusters have been explained in this section.

- Red cluster: This cluster contains roads which have a constant low/medium amount of traffic throughout the day. These are the local roads of everyday use alongside which markets, bus stands, train stations, etc are located.
- Green cluster: The cluster depicted by green color contains roads which are mostly present in residential areas. These are also the streets connecting nearby places, due to which traffic density on these roads is

low for all periods, except for some slight peaks in morning's time. For example, in the region shown in figure 12, all the roads present in a residential area fall within this category.

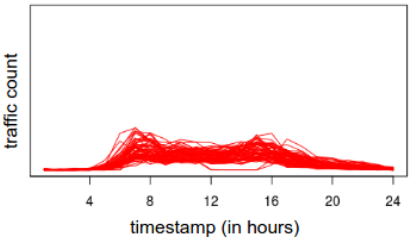

Fig.10. (a) Red Cluster

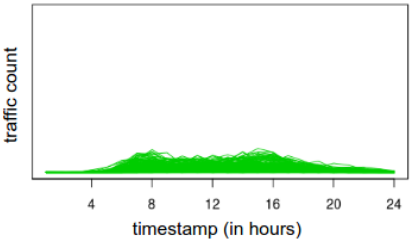

Fig.10. (b) Green Cluster

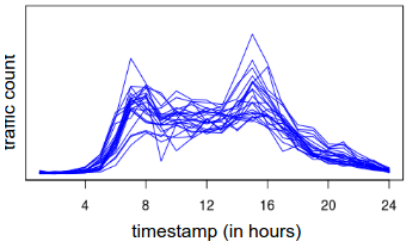

Fig.10. (c) Blue Cluster

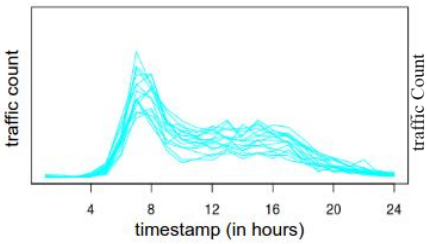

Fig.10. (d) Cyan Cluster

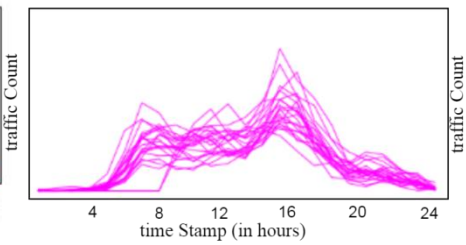

Fig.10. (e) Pink Cluster

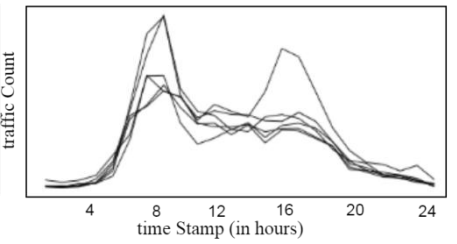

Fig.10. (f) Black Cluster

Fig.10. Density of different clusters

- Pink and Cyan cluster: This cluster contains all such links which have a peak in the afternoon, around 4 PM with no AM peak. These roads are used by traffic returning home from schools or other activities during the afternoon. For example many major museums and libraries open from 11 AM- 5 PM and are most visited during this period of the day. Cluster cyan, on the other hand, shows a morning peak, at around 7-8 AM, rather than an evening peak. It partly consists of links with traffic in the opposite direction of links in the pink cluster.
- Black cluster: This cluster shows heavy traffic during the early morning and medium traffic during the afternoon hours. As roads are present on the outskirts of the city, they may be used [Chrobok et al., 2004] for early transportation, recreational purposes, for instance.

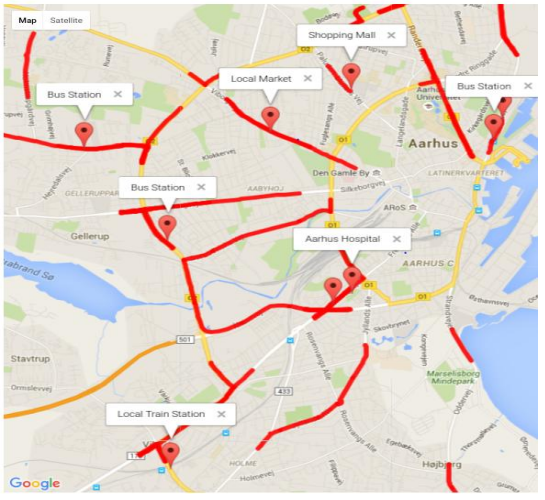

Fig.11. Map showing streets corresponding to Red cluster

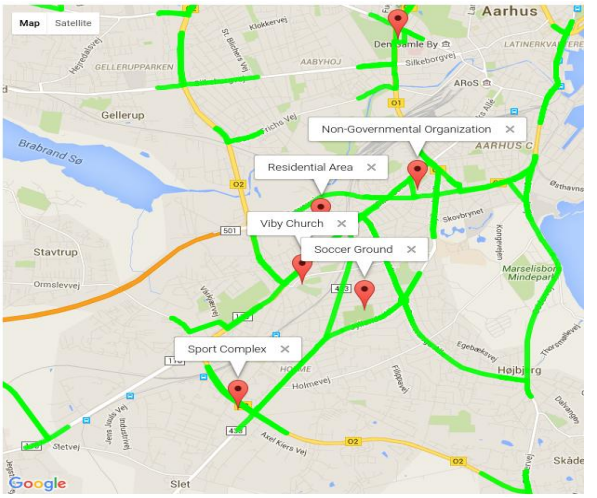

Fig.12. Map showing streets corresponding to Green cluster

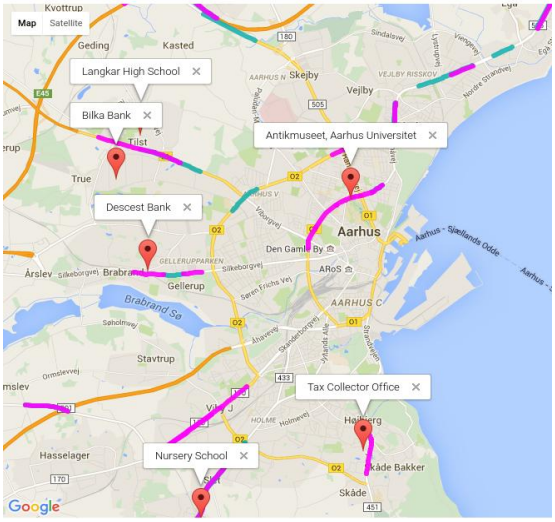

Fig.13. Map showing streets corresponding to Cyan and Pink cluster

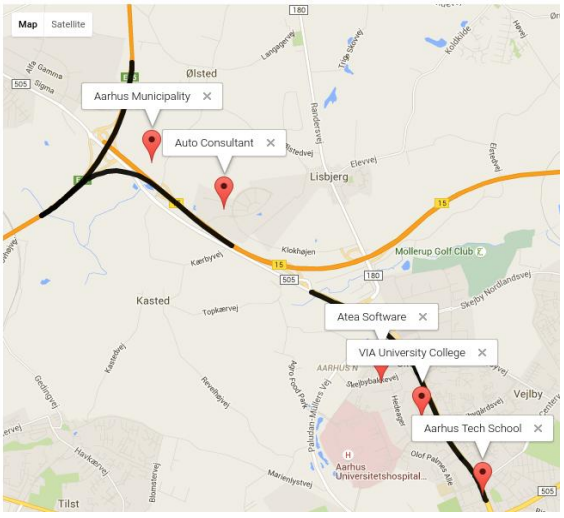

Fig.14. Map showing streets corresponding to black cluster

6. Conclusions

Due to the implementation of various information centers of urban traffic, more traffic data is now available. In order to mine patterns from the urban traffic data, various clustering algorithms such as PAM, AGNES, and DBSCAN have been used. We have seen that different clustering methods provide different insight into the traffic data. The various clusters formed are satisfactorily explained on the basis of seasonal and demo-graphical factors.

Limitations and Future Work: The data-set used stretched across a period of only 4 months. As a result, traffic patterns due to seasonal changes could not be analyzed. So, it is recommended to carry out the same analysis when more data is available. Second, long-term variations in traffic patterns of Aarhus couldn't be analyzed because data spanning over only a few months were considered. Traffic patterns may be expected to change on the basis of social-economic and geographical developments, infrastructural changes and/or general demographic, in the long run. It could thus be fascinating to analyze the effect of these local changes on traffic patterns in the city. Finally, only traffic data of Aarhus city were studied here. It would be insightful to compare the traffic flow patterns for various cities. By doing this, we can observe the effect of social, economic and cultural factors in urban areas having similar geography and populations.

References

Al Hassan, Y., & Barker, D. J. (1999). The impact of unseasonable or extreme weather on traffic activity within Lothian region, Scotland. Journal of Transport Geography, 7(3), 209-213.

Aranganayagi, S., & Thangavel, K. (2007, December). Clustering categorical data using silhouette coefficient as a relocating measure. In Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on (Vol. 2, pp. 13-17). IEEE.

Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In KDD workshop (Vol. 10, No. 16, pp. 359-370).

Chrobok, R., Kaumann, O., Wahle, J., & Schreckenberg, M. (2004). Different methods of traffic forecast based on real data. European Journal of Operational Research, 155(3), 558-568.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

Festin, S. M. (1996). Summary of national and regional travel trends: 1970-1995 (No. FHWA-PL-96-021). United States. Federal Highway Administration.

Gowda, K. C., & Krishna, G. (1978). Agglomerative clustering using the concept of mutual nearest neighbourhood. Pattern recognition, 10(2), 105-112.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

Liu, Z., & Sharma, S. (2006). Statistical investigations of statutory holiday effects on traffic volumes. Transportation research record, 1945(1), 40-48.

Stathopoulos, A., & Karlaftis, M. (2001). Temporal and spatial variations of real-time traffic data in urban areas. Transportation Research Record: Journal of the Transportation Research Board, (1768), 135-140.

Stathopoulos, A., & Karlaftis, M. G. (2003). A multivariate state space approach for urban traffic flow modeling and prediction. Transportation Research Part C: Emerging Technologies, 11(2), 121-135.

Wang, X., Fan, T., Li, W., Yu, R., Bullock, D., Wu, B., & Tremont, P. (2016). Speed variation during peak and off-peak hours on urban arterials in Shanghai. Transportation Research Part C: Emerging Technologies, 67, 84-94.

Zhao, F., & Chung, S. (2001). Contributing factors of annual average daily traffic in a Florida county: exploration with geographic information system and regression models. Transportation Research Record: Journal of the Transportation Research Board, (1769), 113-122.

Pascale, A., Mavroeidis, D., & Lam, H. T. (2015). Spatiotemporal clustering of urban networks: Real case scenario in London. Transportation Research Record: Journal of the Transportation Research Board, (2491), 81-89.

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N., & Lint, H. (2017). Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. Scientific Reports, 7(1), 14029.