12th International Conference on Computer and Knowledge Engineering (ICCKE 2022), November 17-18, 2022, Ferdowsi University of Mashhad, Iran

City Intersection Clustering and Analysis Based on Traffic Time Series

Mohammad Aminazadeh
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
mohammad.aminazadeh98@gmail.com
m.aminazadeh@mehr.ui.ac.ir

Fakhroddin Noorbehbahani
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
noorbehbahani@eng.ui.ac.ir

***Abstract*— In the past few years, the number of private vehicles has risen intensely and will continue to grow in the future. Traffic management has become a significant concern as traffic issues directly affect the quality of life. Different methods have been implemented to track city traffic and find ways to reduce congestion, accidents, pollution, etc., but there is room for improvement. This study attempts to analyze Isfahan traffic through clustering traffic time series of intersections to find similar locations in the city. The result of this study would help traffic managers define appropriate policies and tactics for managing similar intersections effectively, which brings about time and cost saving for both citizens and the government.**

Keywords—time series clustering; city intersection clustering; time series visualization; traffic management

I. Introduction

Urbanization is increasing every year, and the number of vehicles traveling across cities is growing, leading to different issues in city traffic management. These issues include pollution, road congestion, accidents, and loss of time and money.

Numerous projects have been undertaken to enhance traffic management and reduce costs. So far, traffic volumes have been analyzed through road sensors or taxi trajectories (Pu et al., 2013). It is done to understand how traffic flow changes and how to enhance it via traffic management policies or by predicting traffic volume to reduce traffic jams and other related problems. It would be ineffective to apply the same traffic management policy to all locations in a city.

Isfahan traffic sensors are installed in 71 intersections across the city. To show the distribution of hubs that collect data, each hub is pinpointed on a map displayed in Fig. 1. It is obvious that hubs are well distributed across the whole city. This ensures that traffic values are not biased toward specific locations in the city.

To our best knowledge, no research has been conducted to find locations with similar traffic flow behaviors through traffic time series clustering. The primary motivation of this study is to partition city intersections based on the similarity of their traffic flows to detect similar/dissimilar locations in the city for managing traffic more productively. Our secondary purpose is to analyze traffic flow changes and find patterns for single intersections during 24 hours of a day through time series visualization. The research questions are listed below:

1. Which intersections across the city have similar/dissimilar traffic flow behaviors?
2. How does traffic flow change in 24 hours of the day concerning each day of the week?
3. What are the traffic flow trends in 24 hours of the day regarding each season of the year?
4. At which hours of the day do peaks and valleys happen in traffic flow?

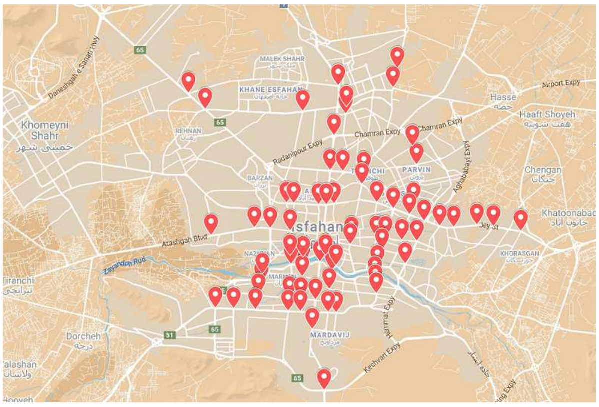

Fig. 1. Distribution of traffic sensors in Isfahan

The paper is organized as follows. Section II discusses related work, explaining advantages and disadvantages of other studies, and their differences from our study. In Section III, the methodology for analyzing traffic data is described. Section IV presents and discusses the results of the study. Section V concludes.

II. Related work

Recently, intelligent traffic management has received great attention. Lots of research has been done in fields such as traffic congestion and traffic pattern recognition, but most of the research done in city traffic management has addressed the issue of road congestion. Studies about intelligent traffic management could be categorized as follows:

- Traffic data visualization
- Traffic flow prediction
- Traffic pattern recognition and clustering

Chen et al. [1], concentrated mainly on traffic data visualization. They introduced the primary concepts of traffic visualization and its processes. These processes included data pre-processing, data visualization, and data visual analysis.

Hayashi et al. [2], also focused on traffic visualization, but additionally, they have tried to propose data summarization methods to be able to cover different-sized urban areas. They offered different methods and algorithms to reduce the volume of data and introduced two types of data summarization. They are road-network data summarization and transportation data summarization. The former focuses on reducing the complexity of city nodes and road networks. The latter is used to reduce the volume of transportation data in each area by removing unnecessary data and employing statistical methods to aggregate data.

In [3], traffic congestion is visualized concerning traffic parameters in India. Parameters such as traffic density, traffic volume, headway, occupancy, and vehicle velocity are taken into the calculation to determine the traffic state. They have also done a time zone classification, which groups 24 hours of the day for extracting each group's traffic behavior and congestion. Zamani et al. have put effort into the same issue and grouped 24 hours of the day into five clusters [4].

Several studies have analyzed both spatial and temporal aspects of traffic data at the same time, which resulted in spatial-temporal visualizations and analysis. The spatial-temporal analysis is done in various forms. Zhang et al. [5], have tried to analyze the data using MAS (Multi-Agent System) to gather data from distributed data sources and analyze it through space and time.

In [6], the authors introduced a spatial-temporal visualization model for the first time. They displayed the spatial elements such as road networks on a 2D map on the x and y-axis and added the temporal element as the third dimension on top of it as the z-axis. Bonds showed the level of congestion with colors ranging from green to red. At the end of each time period, a bond is stacked on top of the roads to show past traffic condition for that period.

Picozzi et al., on the other hand, focused more on the spatial element only. They developed a traffic management system that calculates traffic and marks the roads' traffic state on its map in real-time [7]. Unfortunately, it is not applicable to show historical traffic changes on a 2D map.

In [8], a system named T-Watcher was proposed, which analyzed city traffic in three different layers. They have used taxi trajectory footprints to analyze city traffic. The layers mentioned include region, road, and vehicle fingerprints. Region fingerprint was the most abstract one, which showed the congestion level in a region within the city using a 2D map with a heat map coloring. The more congested a road is, the denser its color will be. The second layer was road fingerprint, which focused on the historical values of each route and tried to visualize them in more detail. It applied a circular visualization with different layers of rings to display the hours of the day. The congestion levels are shown by colors on each layer, and the central ring is divided into seven slices to indicate days of the week. The last layer was the vehicle fingerprint, which was directly resulted from the movement of each taxi in the city. This type of fingerprint combined the historical data of each vehicle with its live data to get a general and real-time estimation of the traffic situation. At this level, the speed, direction, and statistical information of each vehicle were gathered together.

To identify traffic congestion levels, several measures could be applied. For example, Po et al. employed the average speed of vehicles on the road as a value to measure congestion [9], and Wen et al. used Traffic Performance Index (TPI) instead, which is an index showing the whole area congestion intensity. They have considered a TPI between 0 to 2 as "No congestion", 2 to 4 as "Smooth congestion", 4 to 6 as "Slight congestion", 6 to 8 as "Moderate congestion", and any value above 8 as "Severe congestion" [10].

In most cases, the mean values related to city traffic are considered as a benchmark to find traffic anomalies. For example, the average speed of vehicles on a specific road is regarded its normal flow speed, and any situation showing a lower flow speed on that road is suspicious of being abnormal and could present congestion.

Several studies have been conducted on city traffic in big cities (for instance, Beijing), however, limited research has been undertaken regarding small cities, especially in developing countries. Additionally, road infrastructures in developing countries dramatically differs from those in undeveloped countries. Thus, general rules associated with investigations done in developed countries are less practical in emerging ones. One such rule of thumb is described in [9], implying that the existence of 100 vehicles in one kilometer of the road could be considered a signal of congestion. This pattern is not necessarily applicable to poor-built or narrow streets in developing countries.

This study has some similarities with some of the related works in case of traffic visualization, but the application of time series clustering to analyze traffic is novel.

III. METHOD

To extract valuable patterns from unstructured and dirty data, much effort must be put into pre-processing phase, to guarantee that input data is accurate. To conduct the present study, data has been received from the transportation department of Isfahan municipality. The data was noisy and unstructured, thus, cleaning and structuring it were critical before any analysis. Fig. 2 displays processes applied to analyze traffic data in succession.

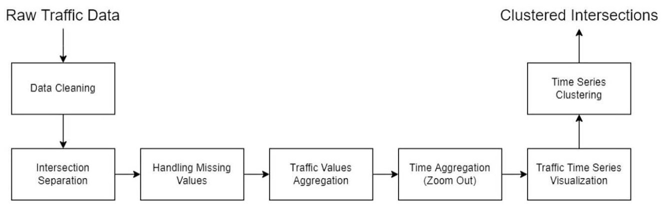

Fig. 2. Traffic analysis process

A. *Input Data*

Data is the most vital entity in any data analysis task. Thus, much attention must be given to the validity of input data.

In this research, real data is analyzed that is produced by road traffic sensors. The dataset was collected from the transportation department of the Isfahan municipality. The traffic control system implemented in Isfahan is called SCATS, an Australian traffic control company that serves 28 countries including, Iran. Available data used in this study contains 37 months of traffic values, ranging from October 2018 to November 2021. The traffic data consist of 37 text files, and each file contains one-month records of all intersections. Each record contains traffic values of a sensor for 15 minutes.

B. *pre-processing*

The primary purpose of pre-processing is to clean data and bring structure to it. In the diagram displayed in Fig. 2, every task before time series visualization is considered pre-processing. As mentioned in [11], there exist five common issues with dirty data: out-of-date data, incomplete data, inaccurate data, duplicate data, and inconsistent data.

Regarding traffic data applied in this study, each line of text represented traffic data of several sensors located in the corresponding intersection. These values were separated by different spaces in the related text file. Hence, it was required to separate sensor values and remove the inconsistency. Next, any digit or letter, which did not relate to a practical value was marked as redundant and was removed from the lines that significantly reduced data size. Similar actions make it possible to store massive amounts of data on storage and could be used for historical data analysis or batch data processing. After that, incorrect (noise) values were detected using a statistical method. The values of sensors that were further than 5-sigma of the mean were considered outliers and replaced with the “NA” value.

As mentioned earlier, 71 intersections across the city collected traffic data. Unfortunately, data collected from each intersection were not aggregated in a single file. To analyze each intersection separately, firstly, records related to each intersection should be extracted and stored in a separate file.

Next, different types of missing values were handled. In one type, single or multiple sensors’ values were missing, but more than half of the values were available. In this category, missing values were filled with the mean of available values of other sensors located in the intersection. In the second type, several lines of data were entirely missing, and in the worst case, several months of data were wholly corrupted. Data related to these periods should be ignored and put focus on periods that have available data.

When all missing values are estimated, sensors’ values in each line are aggregated to result in a single number. This step could be done by selecting one statistical aggregation function such as summation, mean, or mode. In our case, the appropriate function was summation. The sum of values is added at the end of that line replacing single traffic values. This number presents sum of inflow and outflow volume of one intersection during 15 minutes. Intersections 1043, 1017 and 1075 were removed from analysis due to containing wide periods of corrupted data.

After that, each text file is converted to its related excel file. The dirty and unstructured text files are now replaced with clean and structured tables, each related to a separate intersection. These excel files are applied as inputs for the subsequent analytical phases. Table I, displays the specification of the final structured and cleaned traffic datasets.

TABLE I. TRAFFIC DATASET FEATURES

Feature	Type
Intersection ID	String
Gregorian Date	Date
Day of The Week	String
Time	Time
Traffic Sum	Integer

C. *Analysis*

1) Traffic Time Series Visualization: After preprocessing, there are 68 excel files corresponding to Isfahan intersections. In each file, traffic values from October 2018 to November 2021 are listed in 15-minute periods in order.

Hourly traffic pattern reveals traffic volume changes in 24 hours. To discover seasonal effects on hourly traffic patterns, hourly time series related to each season are displayed with a unique color. In case of observing distinct traffic behaviors on each day of the week, time series of various weekdays are produced and visualized separately. To create a single time series relating to a season or a day of the week, time aggregation is done first to convert records from a 15-minute scale to an hourly scale. To aggregate values, statistical operations such as mean, sum, minimum, maximum, mode, or count could be employed [12]. In this study, the sum of all

978-1-6654-7613-3/22/$31.00 ©2022 IEEE

traffic values relating to a single day was applied to aggregate 15-minute records to a new daily-scaled record. Next, traffic values relating to each hour of the day are averaged together. This will lead to 24 mean values, each indicating the mean value for a single hour of the day in a specific season or day of the week.

2) Intersections Clustering Based on Traffic Time Series: Three types of *k*-means clustering algorithms could be employed to cluster intersections' time series, including Euclidean, Dynamic Time Warping (DTW), and Soft DTW *k*-means. Euclidean *k*-means utilizes Euclidean distance to measure similarity between two time series. This distance measure is one of the most straightforward measures used in Euclidean spaces. Equation (1) shows how Euclidean distance is calculated, where q and p are two vectors in Euclidean space, q_i and p_i are i^{th} dimension of q and p, and n is the number of dimensions.

$$d(q,p)=\sqrt{\sum_{i=1}^{n}\left(q_i-p_i\right)^2} \tag{1}$$

The other two clustering algorithms apply DTW distance to measure the similarity between two points in space. DTW distance is helpful in situations where the comparison must be made without concerning time. Thus, the similarity between two temporal sequences that may vary in speed could be measured with DTW.

To cluster traffic time series, time should be considered to calculate the similarity between intersections. Therefore, in this case, Euclidean *k*-means performs better than DTW and Soft DTW *k*-means. One limitation of Euclidean *k*-means is that input time series must be equally sized to calculate distance. To address this issue, more extended time series could be cut that results in loss of data, or the shorter ones could be manually extended. We chose to extend shorter time series to avoid data loss by adding zeros at the end of each time series to extend them to the length of the longest time series. Finally, after clustering and before plotting the time series of each cluster, zeros at the end of every time series were removed to enhance appearance of time series.

A significant parameter of the clustering algorithm is the number of clusters. Different methods exist to determine the number of clusters properly, which are discussed briefly. In research done by Makwana et al., six commonly employed ways to determine the optimum number of clusters have been introduced. Equation (2) is the "Rule of Thumb" method for calculating the number of clusters where k is the number of clusters estimated, and n is the number of objects [13].

$$k\approx\sqrt{\frac{n}{2}} \tag{2}$$

The second method, called "The Elbow Method" [13], plots the Within-Cluster Sum of Squares (WCSS) against the number of clusters to find the optimum number. WCSS is the sum of the squared distance between each point in the cluster and the centroid, which we attempt to minimize. Equation (3) shows how WCSS is calculated, where, k is the number of clusters and c_j is the cluster's centroid. WCSS for each number of clusters is calculated, and plotted against the number of clusters; an elbow-shaped graph is generated. The number showing the joint of the elbow shows the optimum number of clusters.

$$WCSS(k)=\sum_{j=1}^{k}\sum_{x_i\in\ cluster\ j}\|x_i-c_j\|^2 \tag{3}$$

Other methods are the "Information Criterion" approach, which seeks to maximize the likelihood of different types of models (different number of clusters) by adding parameters to the model, and the "Information Theoretic Approach" which, is based on a procedure related to distortion. The next method is "Using Silhouette", which utilizes various number of clusters besides checking and comparing within-cluster distance with between-cluster distance: more contrast shows a better fit. Final method is "Cross Validation", which is based chiefly on cluster stability. It splits data into several parts to employ some for clustering and others for validation. The idea behind stability is that good clustering algorithm should produce similar clustering results on data generated with specific distributions. [13].

In this study, the Elbow method is applied to determine number of clusters due to its excellent balance between simplicity and accuracy. To employ the elbow method, the Euclidean *k*-means algorithm is run iteratively. On the first iteration, the number of clusters is set to 2, and on each subsequent iteration, it is incremented by one. At the end of each iteration, after clustering is done, WCSS is calculated for that iteration. Finally, WCSS values are plotted for their corresponding number of clusters.

IV. RESULTS

A. Hourly Traffic Patterns

1) Hourly Time Series Separated by Seasons: Each intersection has a unique traffic flow behavior but due to massive amount of data, it is not possible to report patterns of all intersections separately. Hence, most insightful patterns will be presented only.

It has been discovered that, in general, traffic flows within hours of the day followed a relatively similar trend for most intersections. After midnight, traffic volume declined and reached the minimum at around 4 AM. After that, it rose and kept soaring until about 8 AM, which reached a relative maximum. Between 4 AM to 8 AM, traffic increased 21 times in value or 2100% for intersection 1002. From this point, different behaviors were discovered. Traffic in some intersections continued to rise until 1 PM, such as intersection 1002, which increased by 58%, but in other ones consolidated. However, no significant decline was observed from 8 AM to 1 PM. This behavior is called "The morning rush". Yin et al. have reported similar results for traffic flow patterns in 24 hours a day [14]. These results are shown in Fig. 3 and Fig. 5 where Q1, Q2, Q3, and Q4 denote, Winter, Spring, Summer, and Fall, respectively.

978-1-6654-7613-3/22/$31.00 ©2022 IEEE

In the afternoon, from 1 PM to 4 PM, some intersections, such as intersection 1002, showed a decline in traffic volume of around 33%, while for other intersections, such as 1018, shown in Fig. 4, it did not happen. The afternoon decline occurred in all seasons of the year, but we found out that in spring and summer, this drop in traffic volume was more noticeable due to the higher temperatures.

Next, traffic rose again from 4 PM and reached its maximum value between 6 PM to 9 PM. This rally is called “The evening rush”. However, the evening rush took place differently at intersections. For some of them, the maximum flow took place at 7 AM and started to go down after that. While for others, traffic went higher at this hour and reached the maximum at 9 PM to 10 PM.

For almost every intersection, the traffic volume was higher in spring and summer than in fall and winter. An interesting observation was the difference in late-night traffic drops from 8 PM to 12 AM between these seasons. As shown in Fig. 5, late night drops happened sooner and more intense in fall and winter, which is due to the cold weather at late night hours in these two seasons.

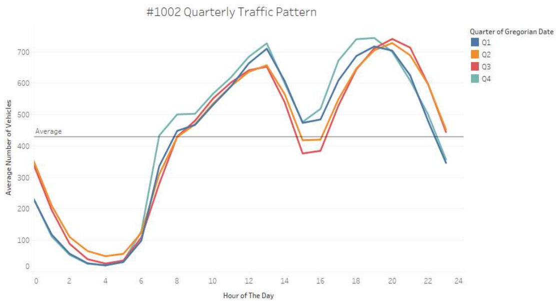

Fig. 3. Hourly time series of intersection 1002 separated by quarters/seasons of the year.

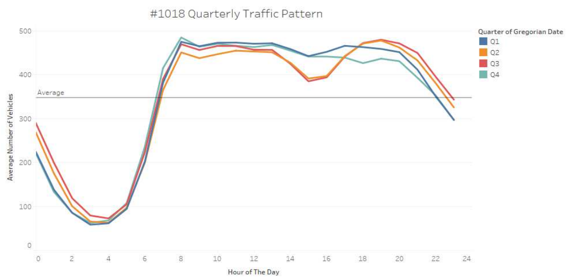

Fig. 4. Hourly time series of intersection 1018 separated by quarters/seasons of the year.

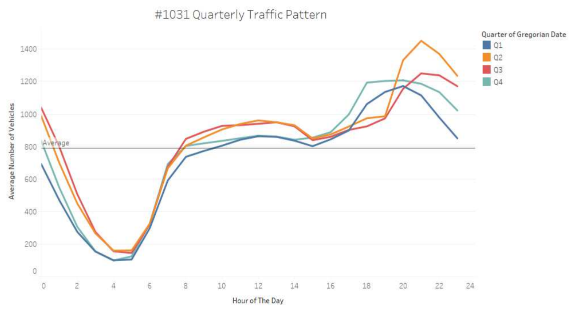

Fig. 5. Hourly time series of intersection 1031 separated by quarters/seasons of the year.

2) *Hourly Time Series Separated by Days of the Week*: In Iran, working days are from Saturday to Thursday, and Friday is a holiday. The results of visualizing time series related to each day of the week, denoted that working days’ time series were almost entirely similar. At the same time, the holiday showed different and unique behavior. As depicted in Fig. 6 to 8, generally, Friday had the minor traffic volume among days of the week and experienced morning and evening rush with a much more gentle slope.

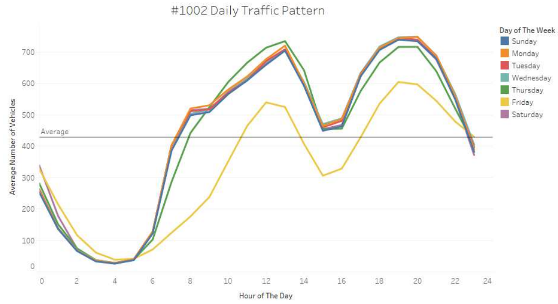

Fig. 6. Hourly traffic time series of intersection 1002 separated by days of the week.

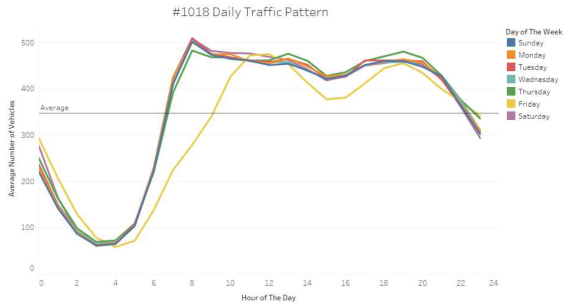

Fig. 7. Hourly traffic time series of intersection 1018 separated by days of the week.

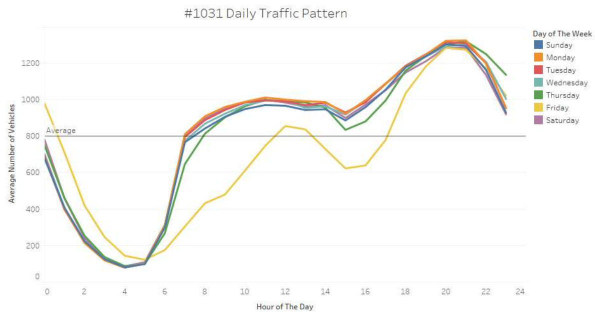

Fig. 8. Hourly traffic time series of intersection 1031 separated by days of the week.

B. *Intersection Clustering*

1) *Optimum Number of Clusters:* As mentioned earlier, in this research, the elbow method was employed to find the optimum number of clusters. Euclidean *k*-means algorithm is run several times, and WCSS values are calculated for every iteration and plotted for the corresponding number of clusters (see Fig. 9).

978-1-6654-7613-3/22/$31.00 ©2022 IEEE

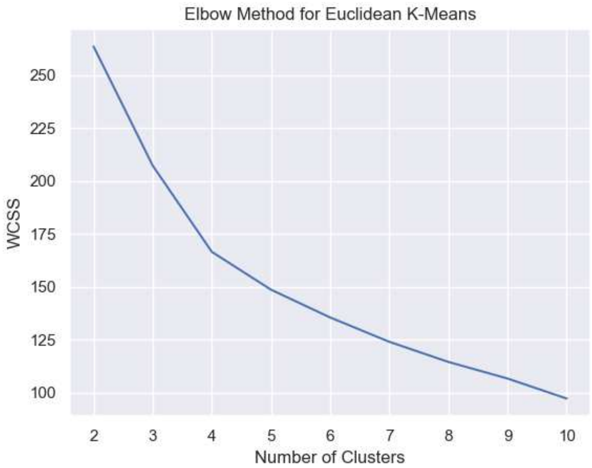

Fig. 9. Elbow method result for Euclidean *k*-means

WCSS dropped sharply and continued to fall noticeably until 5 clusters. However, in the case of the increasing clusters of more than 5, WCSS does not change significantly. It could be concluded that 5 is the optimum number of clusters for Euclidean *k*-means clustering in this study, hence intersections are grouped into 5 clusters.

2) Clustering Results: Intersections were divided into 5 clusters. Table II displays the distribution of intersections into clusters. To get deeper insights into clustering results, intersections belonging to each cluster are shown on the map with same colors. As shown in Fig. 10, It is observable that intersections, which are in line or connected, show similar traffic flow behaviors. However, when distributed intersections belong to the same cluster, hidden patterns are revealed about similar intersections. Each cluster's time series is plotted and displayed in Fig. 11 to 15.

Cluster 1 contains intersections on the main vertical route, connecting northern and the southern regions of the city. This cluster includes intersections that support travels mainly from the southern region to the northern region of the city and vice versa.

Cluster 3 contains intersections next to or on the southern side of the river, where primary business and entertainment centers are located. Some of the main pathways on the northern side of the city are also in this cluster. Additionally, we know that most business and entertainment centers are located in the southern part of the river. Thus, it makes sense to conclude that intersections enclosed in cluster 3, relate directly or indirectly to primary business and entertainment centers.

Clusters 2 and 4, contain intersections that cover the borders of the city and are located on main routes connecting the city to surrounding towns. Therefore, these intersections are pathways between Isfahan and other towns and are employed mainly for out-of-city transportation.

Cluster 5 contains 4 intersections, which 3 out of them are located next to the river which shows that this cluster contains data about riverside transportation. Each cluster's transportation functionality is summarized in Table III.

TABLE II. INTERSECTION CLUSTERS AND THEIR ASSOCIATED COLORS ON THE MAP

Cluster	Color on Map	Intersection IDs
1	Green	1001-1016-1035-1071-1080-1084-1090-1092
2	Black	1019-1037-1059-1065-1067-1072-1074-1083-1088-1115-1119-1502
3	Blue	1010-1014-1021-1029-1031-1033-1034-1039-1042-1044-1047-1053-1057-1068-1069-1077-1078-1079-1081-1082-1086-1093-1109-1112-1113-1117-1501-1504
4	Yellow	1002-1012-1018-1028-1038-1040-1041-1048-1051-1054-1064-1066-1070-1085-1095-1503
5	Purple	1032-1036-1094-1116

TABLE III. INTERSECTION CLUSTERS AND THEIR TRANSPORTATION FUNCTIONALITIES

Cluster	Functionality
1	North-South Connection/Transportation
2	Out-of-City Transportation
3	Main Business/Entertainment Centers
4	Out-of-City Transportation
5	Riverside Transportation


Fig. 10. Clustered intersections displayed by colors

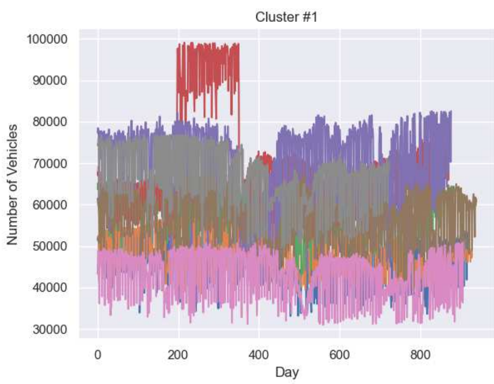

Fig. 11. Cluster 1 time series

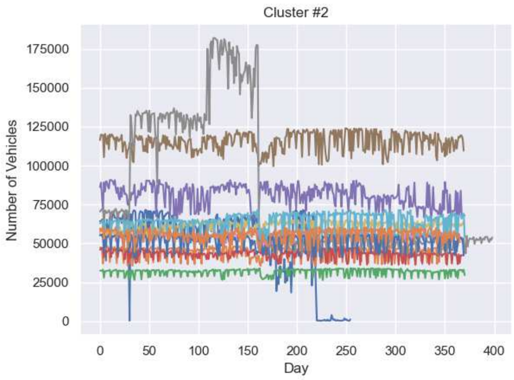

Fig. 12. Cluster 2 time series

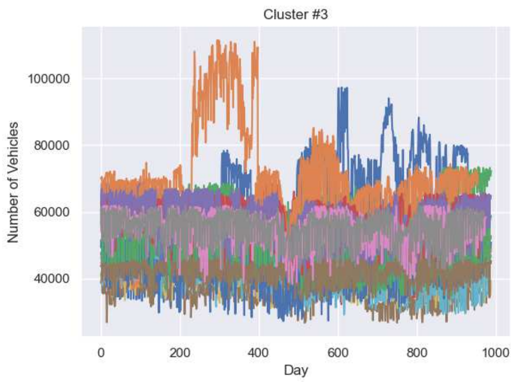

Fig. 13. Cluster 3 time series

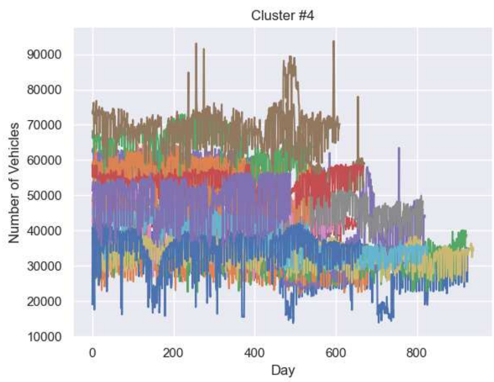

Fig. 14. Cluster 4 time series

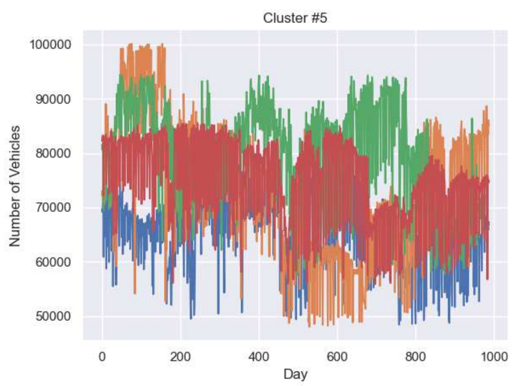

Fig. 15. Cluster 5 time series

V. Conclusion and Future Work

As the population grows in big cities, traffic management becomes a crucial concern. In this study, 37 months of Isfahan traffic data were analyzed to extract valuable patterns for better traffic management.

After preprocessing and cleaning traffic data, traffic time series were clustered to find out which intersections behave the same and analyze the reasons behind those patterns. Moreover, traffic time series data related to each intersection were visualized to observe trends.

Clustering intersections having similar traffic behaviors, makes it possible to apply particular traffic policies for each cluster to control and enhance the traffic flow. This would help optimize traffic management policies, which leads to the reduction of costs, time loss, traffic jams, pollution, and accidents. For instance, intersections located around main business and entertainment centers must be targeted with different traffic control policies than intersections that support out-of-city transportation.

For example, to employ a transportation policy for controlling COVID-19, cluster 3 should be targeted to transportation limitations because it relates to entertainment or business centers that attract people the most. Any transportation limitation on intersections in clusters 2 and 4 could have a massive economic loss due to being the main pathways applied for freight transport to nearby cities. Additionally, the results could be used for better banner advertising. Banners should be installed at intersections having the highest transportation and at times of their traffic peak volume (i.e. intersections belonging to clusters 1 and 3).

One limitation of the current study is that data were limited to 37 months and were full of missing values. Those records that were entirely missing could not be analyzed. In this research, data visualization focused only on time, in future works, spatial aspect could be considered while analyzing traffic flows.

To obtain better results in the future, intersections connections forming traffic corridors could be considered. Hence, traffic would be analyzed through these corridors rather than a single intersection. Inputs and outputs of intersections could be connected to form a directional graph, and patterns would be extracted from traffic flow changes between the output of one vertex and inputs of adjacent vertices. This methodology could be employed to find the intensity of effects on one intersection's traffic flows by applying traffic control policies in another intersection. Another application of the results obtained from this methodology is for predicting one intersection's traffic flow change by having another adjacent or nonadjacent intersection's traffic flow.

Acknowledgment

We would like to thank the transportation department of the Isfahan municipality for sharing data and their valuable supports.

978-1-6654-7613-3/22/$31.00 ©2022 IEEE

REFERENCES

[1] W. Chen, F. Guo, and F. Y. Wang, “A Survey of Traffic Data Visualization,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 2970–2984, 2015, doi: 10.1109/TITS.2015.2436897.

[2] N. Hayashi and Y. Shiraishi, “A method for summarizing and visualizing city traffic data,” *2014 7th Int. Conf. Mob. Comput. Ubiquitous Networking, ICMU 2014*, pp. 165–170, 2014, doi: 10.1109/ICMU.2014.6799089.

[3] T. Tsuboi, “Traffic congestion visualization by traffic parameters in India,” *Int. Conf. Innov. Comput. Commun.*, pp. 371–380, 2020.

[4] Z. Zamani, M. Pourmand, and M. H. Saraee, *Application of Data Mining in Traffic Management: Case of City of Isfahan*. .

[5] H. Zhang, Y. Zhang, Z. Li, and D. Hu, “Spatial – Temporal Traffic Data Analysis Based on Global Data Management Using MAS,” *Ieee Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 267–275, 2004.

[6] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, “Stacking-based visualization of trajectory attribute data,” *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2565–2574, 2012, doi: 10.1109/TVCG.2012.265.

[7] M. Picozzi, N. Verdezoto, M. Pouke, J. Vatjus-Anttila, and A. Quigley, “Traffic visualization: Applying information visualization techniques to enhance traffic planning,” *GRAPP 2013 IVAPP 2013 - Proc. Int. Conf. Comput. Graph. Theory Appl. Int. Conf. Inf. Vis. Theory Appl.*, pp. 554–557, 2013, doi: 10.5220/0004291605540557.

[8] J. Pu, S. Liu, Y. Ding, H. Qu, and L. Ni, “T-watcher: A new visual analytic system for effective traffic surveillance,” *Proc. - IEEE Int. Conf. Mob. Data Manag.*, vol. 1, pp. 127–136, 2013, doi: 10.1109/MDM.2013.23.

[9] L. Po, F. Rollo, C. Bachechi, and A. Corni, “From sensors data to urban traffic flow analysis,” *5th IEEE Int. Smart Cities Conf. ISC2 2019*, pp. 478–485, 2019, doi: 10.1109/ISC246665.2019.9071639.

[10] H. Wen, J. Sun, and X. Zhang, “Study on Traffic Congestion Patterns of Large City in China Taking Beijing as an Example,” *Procedia - Soc. Behav. Sci.*, vol. 138, no. 0, pp. 482–491, 2014, doi: 10.1016/j.sbspro.2014.07.227.

[11] M. Nheng, “Five common dirty data issues and how your business can avoid them,” 2020. .

[12] “Time Aggregation,” *IBM Documentation*, 2005. https://www.ibm.com/docs/en/tnpm/1.4.5?topic=series-time-aggregation (accessed Dec. 03, 2021).

[13] P. Makwana, T. M. Kodinariya, and P. R. Makwana, “Review on Determining of Cluster in K-means Clustering Review on determining number of Cluster in K-Means Clustering,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, 2013, [Online]. Available: https://www.researchgate.net/publication/313554124.

[14] J. Yin, D. Zhou, and Q. Q. Xie, “A clustering algorithm for time series data,” *Parallel Distrib. Comput. Appl. Technol. PDCAT Proc.*, pp. 119–122, 2006, doi: 10.1109/PDCAT.2006.1.

978-1-6654-7613-3/22/$31.00 ©2022 IEEE