Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2018, 5(8):75-82**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Clustering and Analysis of Headway and Speed Data of an Industrial Zone Traffic Flow

Oguzhan DOGAN[1*], Sancar AKBASAK[1], Onder M. TANRIYAPISI[1], Caglar KOSUN[2], Serhan OZDEMIR[1]

[1]Artificial Intelligence and Design Laboratory, Department of Mechanical Engineering, Izmir Institute of Technology, 35430, Urla İzmir, Turkey
[2]Department of City and Regional Planning, Ondokuz Mayıs University, 55100, İlkadım, Samsun, Turkey

Abstract Traffic flow data is in general a complex and a high-dimensional. Its analysis as well follows the same pattern. One of the straightforward methods to analyze a complex data such as traffic flow is to cluster the data, and has the probability distribution of it. In this research, three days data from an industrial zone have been considered. The authors have analyzed the data so as to categorize the traffic flow in normalized speeds and the head ways. The flow structure reveals that the normalized speed v\s headway distribution is similar throughout the week, as it is obvious in the speed his to grams .It is seen that in headway distribution, lognormal behavior is observed. A result of this analysis is the fact that on the third day, regardless of the vehicle speeds, tracking distance of the vehicles remained low, which were higher on the previous days.

Keywords K-means clustering, lognormal distribution, traffic flow

1. Introduction

Traffic flow has many characteristics and their analyses could be conducted with multifarious techniques. Some of those techniques may be ascribed to clustering. Cluster analyses could group a set of traffic flow variables according to their similarities or relations. In this study, K-means algorithm is utilized to analyze and categorize the traffic flow variables *i.e.* traffic speed and headway. Concisely, K-means algorithm is unsupervised and simply executed in which the centreoids are specified, all of the data points is assigned to their closest centroids considering Euclidean distance, and then the clusters are formed. Details are elaborated in the following section.

Should one examine K-means algorithm and traffic flow literature together, some of the related work could be listed as follows. For example, in the study [1] for the classification of the road types, ten most sensitive parameters e.g. vehicle compositions, hourly flows are selected, and K-means algorithm classifies the links into different groups providing that similar links are assigned to the same group. Roadside Pollution Monitoring (RPM) units are then classified by road type using K-means algorithm. One of the outcomes, for instance, indicates that there is a relationship between high number of heavy good vehicles on the roads and high level of pollution [1]. Besides, K-means algorithm is utilized in the framework of CO_2 emission reduction in the study [10]. In another study [2], the authors categorize the freeway traffic flow concerning traffic occupancy by employing K-means algorithm. The study assesses the relationship between different traffic states and safety performance on freeway traffic. In order to group the road accident locations K-means algorithm is implemented as well in the paper [9], and three categories i.e. high-frequency, moderate-frequency and low-frequency accident locations are specified considering frequency counts. The study [11] could be also representative for using K-means algorithm concerning pedestrian involved crashes.

In the other work, for Traffic Condition Recognition (TCR) system, K–means algorithm is used to separate the driving features into clusters. The clusters of the driving segments i.e. mean velocity-average accelerating,

average accelerating-idle time percentage are displayed. Hence, the authors present that the latter driving features becomes more proper for TCR [3]. For the prediction of travel time, the authors [4] proposed the modified K-means clustering approach and the details are explained in the paper. The authors also argue that the conventional K-means algorithm has some drawbacks, and they discuss the superiority of their clustering approach [4]. The authors of the work [5] refer to the advantages of the K-means clustering method, and K-means algorithm is implemented for two-intersection corridor and a small size of volume data to ascertain time-of-day breakpoints for traffic signal timing.

The paper [6] would also have to do with the examination of signal efficacy and K-means algorithm usage. K-means algorithm which is involved in two-level clustering is also utilized in the work related with evaluating aggressive driving behaviors in traffic [7].

2. Method

2.1. K-Means Clustering

In the statistical data analysis, identifying similar members of data in the same data set is called clustering [12]. Cluster analysis is revealed by Driver and Kroeber in the anthropology field in 1932 and so far developed in many fields [13]. As a clustering method, K-means clustering is commonly used [14]. Main purpose is to classify or to group the data based on similarities where K stands for the number of clusters. Automatic topic identification, image compression, density estimation, pattern recognition, pattern classification and vector quantization are the applicable areas for K-means clustering [15]. In this paper, K-means clustering is used analyze and categorize the traffic flow variables.

For the K-means clustering algorithm, following equation is used. For given x points in $\mathbb{R}$ and k is defined as cluster number. Aim is to construct a division S= (S_1, ..,S_k) and choose centroids C=(C_1, .., C_k) such that $\sum_{i=1}^{k} \sum_{x \in S_i} d(x, c_i)^2$ to be minimized. 'd' demonstrates the distance between x &c_i [16]. This equation is iteratively repeated so that at the end centroids of the clusters are found.

In the algorithm, number of clusters "k" is predetermined. Then, the centroids of the clusters are assigned arbitrarily. For every data point, the Euclidean distances between each points and cluster centroids are calculated so that the comparison could be done properly. Each point is assigned to the nearest cluster. The intention is to diminish the mean sum of distances to the closest centroid. So that the centroids could be changed according to ongoing calculations. If it is changed, the Euclidean distances are recalculated and same procedure is followed.

2.2. Log Normal Distribution

Likelihood of certain event may occur is called probability. Probability distribution is defined as the mathematical model shows probability presence of a variable out of population [17].Probability distributions are influential in data processing. The log normal distribution is seen inapplications such as failure and durability projections. The tendency of skewing through the limits of the distribution is observed in log normal distributions. Mathematical representations is shown as:

$$p(x) = \frac{1}{\pi(2\pi)^{1/2}} \exp[-\frac{1}{2} ln \frac{(x-x')^2}{\sigma^2}]$$

where $\sigma$is the standard deviation and x' is the normal distribution mean [18]. Generally, the decrease in life time of a product in time is presented with log normal distribution [17]. In log normal distribution, skewed distributions with low means and large variances are observed. The shape of this distribution is set by three different parameters. Those are σ shape parameter, m (median) scale parameter and μ location parameter [19].

3. Data Analysis and Results

In Figure 1, Figure 2 and Figure 3, K-means clusters of 3 different days of traffic data are shown. The system is 2-dimensional. Thus, plots cannot indicate the differences between the clusters of 3 days properly. On the other hand, in the second day one of the clusters penetrates into the other clusters. Further more on the third day, drivers closely follow each other nearly all of the speed ranges.

It can be easily seen that the speed and headway data show Gaussian and lognormal distribution, respectively. Headway data is more heavily jammed towards the origin as it can be seen in Figure 4, 5, 6. Therefore it shows

lognormal characteristic with average location parameter 0.6825 and average shape parameter 0.9393. The distribution of the speed values shows Gaussian characteristic. Combination of these two distributions matches up with clusters.

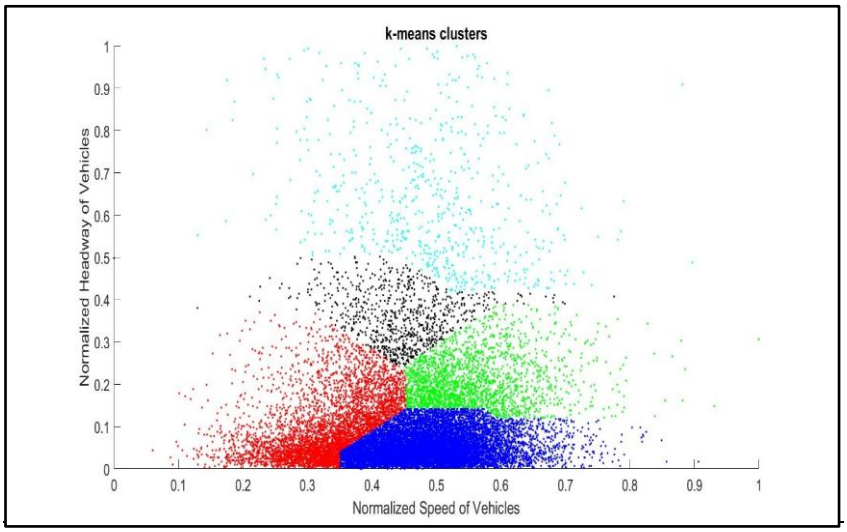

Figure 1: Clustering of Day 1

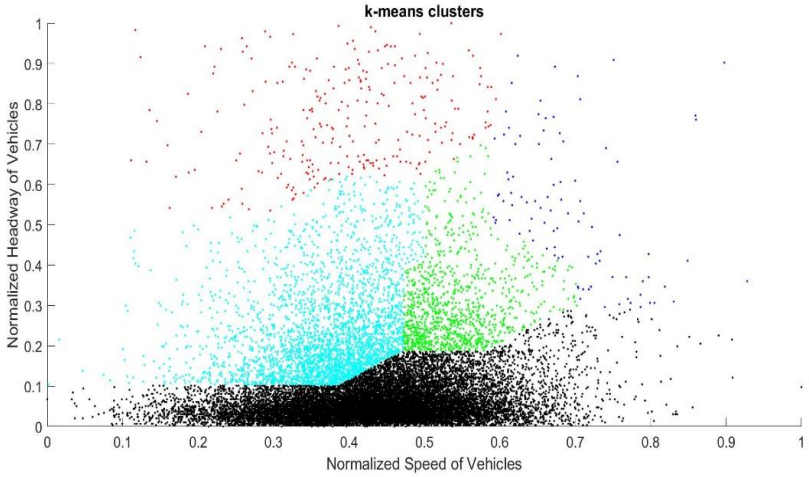

Figure 2: Clustering of Day 2

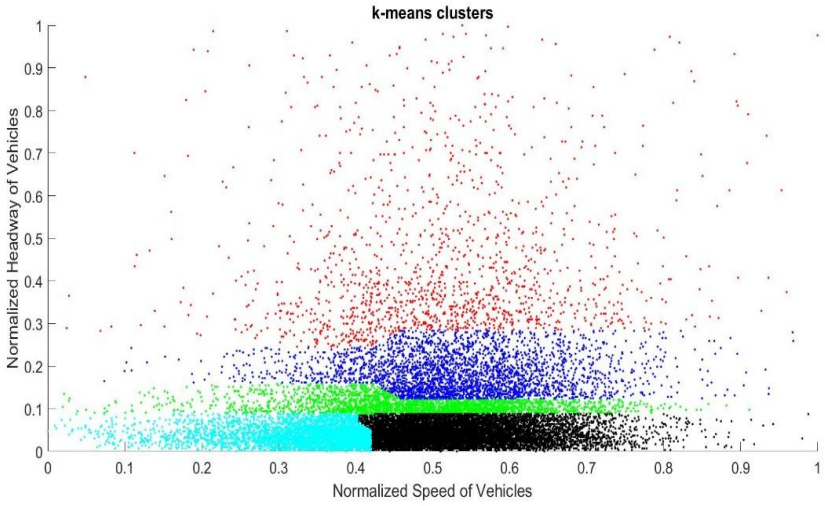

Figure 3: Clustering of Day 3

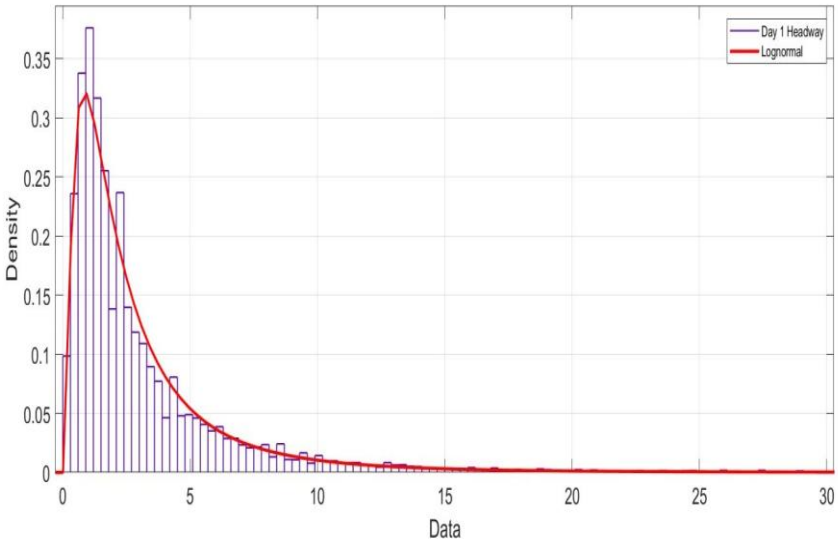

Figure 4: Lognormal Distribution Fitting of Headway Distribution of Day 1

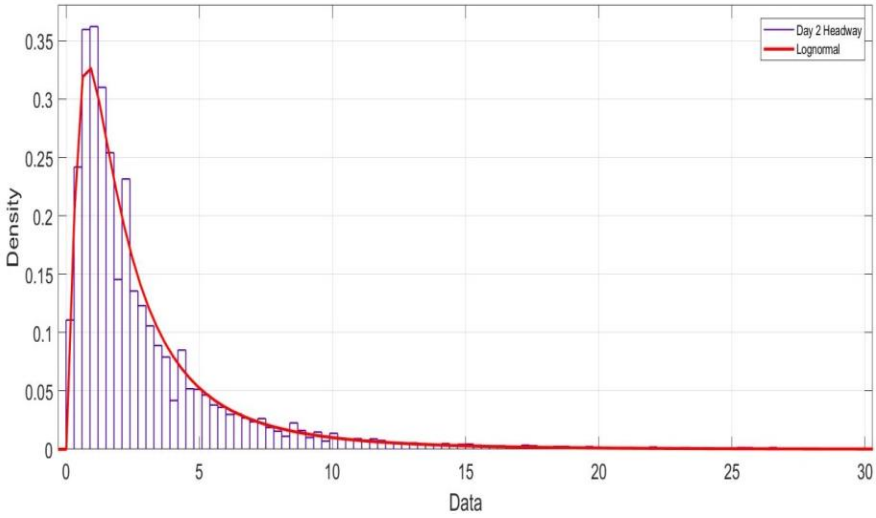

Figure 5: Lognormal Distribution Fitting of Headway Distribution of Day 2

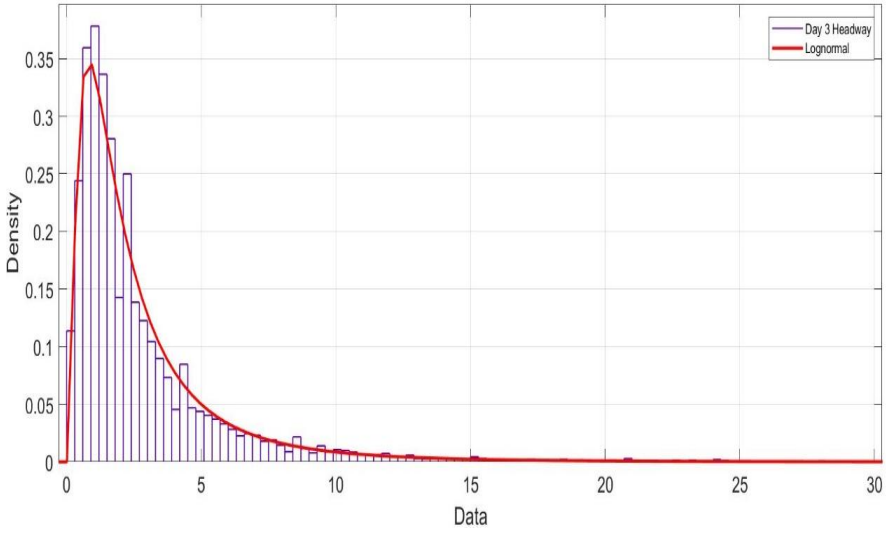

Figure 6: Log normal Distribution Fitting of Headway Distribution of Day 3

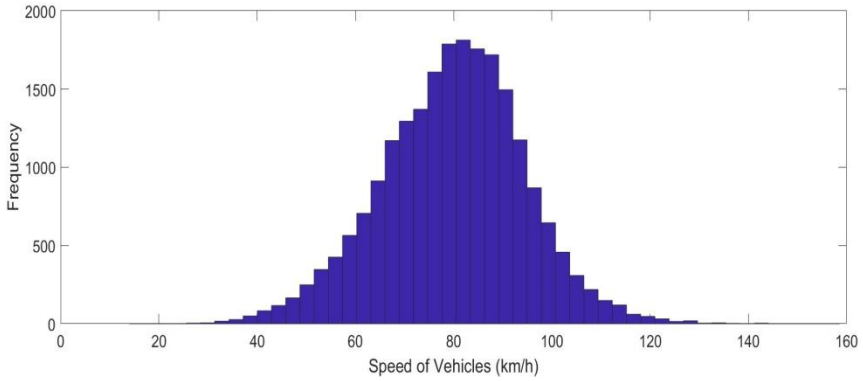

Figure 7: Vehicle Speed Distribution of Day 1

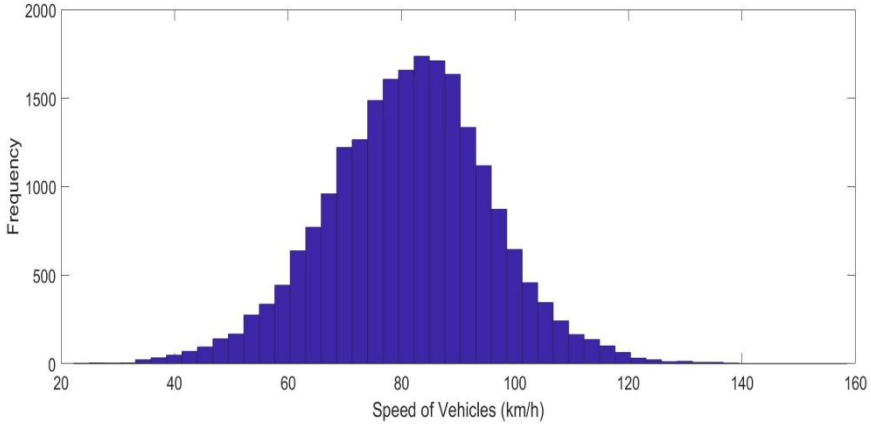

Figure 8: Vehicle Speed Distribution of Day 2

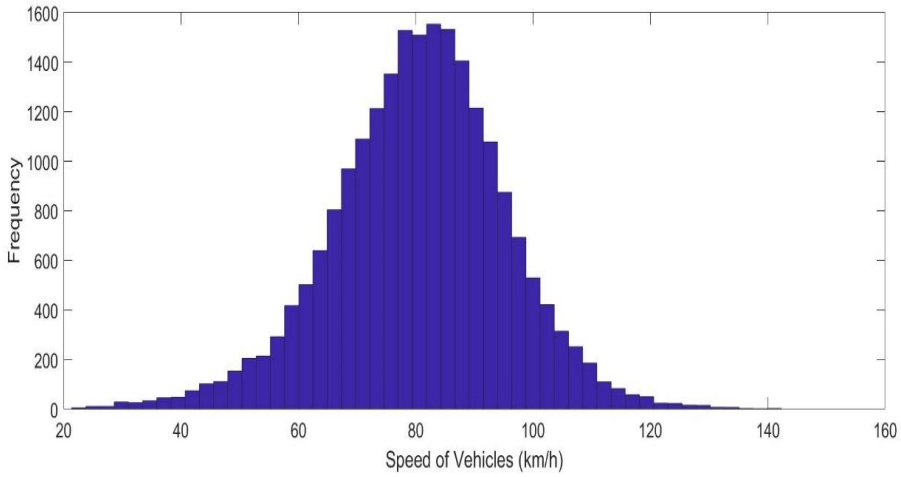

Figure 9: Vehicle Speed Distribution of Day 3

The speed distributions of each day are extracted. As it is seen in Figure 7, 8, 9, the distributions are approximately Gaussian. When inspecting day to day, the distribution characteristics of Day 1 and Day 2 have aproximity, which are symmetrical to the greatest extent. The histogram of Day 3 is somewhat different from the other days, and it is a bit left-skewed. The kurtosis and skewness values in Table 1 also support this inference. The kurtosis and skewness values of Day 1 and Day 2 are closer to 3 and 0, respectively.

Table 1: Skewness and Kurtosis values per day

	Day 1	Day 2	Day 3
Skewness	-0.0823	-0.0522	-0.2037
Kurtosis	3.3854	3.4974	3.6773

Table 2: Average velocity and headway values per day

	Day 1	Day 2	Day 3
Average Velocity (km/h)	79.89	81.2302	80.5672
Average headway (sec.)	3.2431	3.1742	2.9454

The average speeds of each day are found quite similar (Table 2), but with respect to average headway values, Day 1 and Day 2 are much closer to each other.

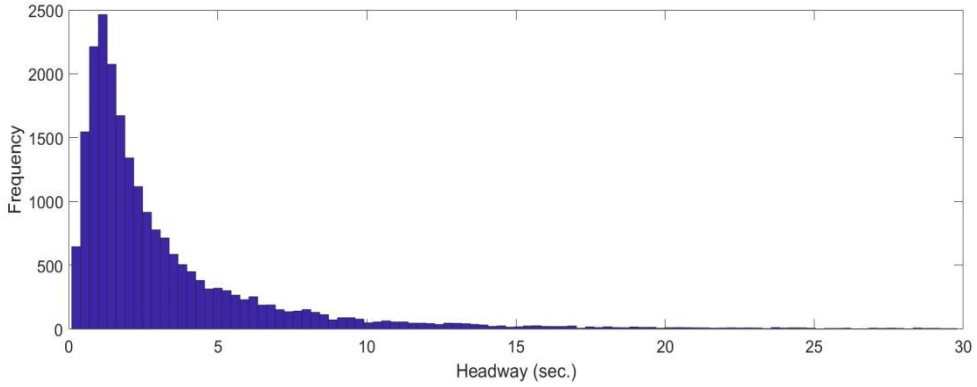

Figure 10: Headway Distribution of Day 1

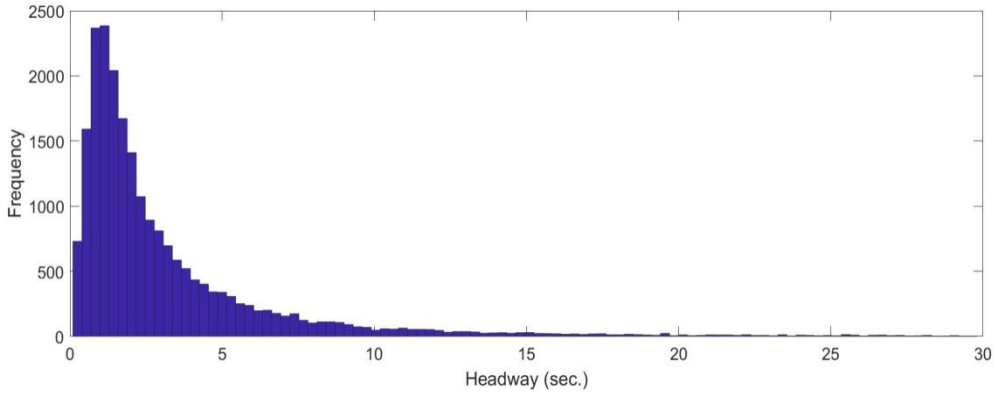

Figure 11: Headway Distribution of Day 2

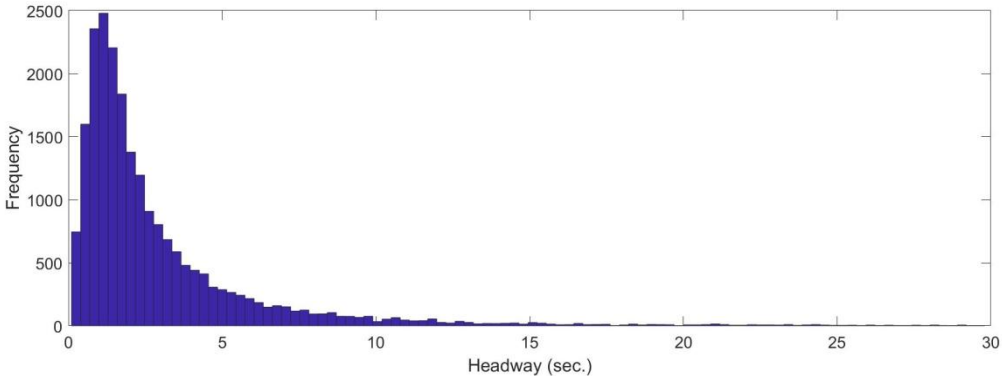

Figure 12: Headway Distribution of Day 3

Headway distributions of each day are analyzed, and to illustrate, the distributions of Day 1, Day 2 and Day 3 are plotted on Figure 10, 11, 12 respectively. As it is seen on these figures, the headway data could fit lognormal-like distribution.

4. Conclusions

This paper has dealt with a three day analysis, where a three separate traffic data were considered. The industrial zone that is considered is not only an intersection for lightweight vehicles but the main artery for intermediate and heavy cargo trucks to all major locations in Turkey. This zone, hence, provides an interesting insight into similar mixed traffic locations, where all classes of motor vehicles could be observed. As such, speed variations also range in wider scales, so that no speed class could be attributed to a certain vehicle class, light or heavy. The general characteristics on this road segment shows that the structure of the traffic is more or less stable on varying days, ignoring minor changes. In the possible case that data may hide inner paradigms that might be opaque to the viewer, K-means classification has been employed. The code is also composed by the authors, to provide certain options on the analysis. The zone is roughly comprised of 5 vehicle types, so clustering is made into 5 segments.

Since the speed v\s headway data distribution is pretty much homogeneous through the days, the variations in K-means may safely be ignored, and instead certain regions may be put in focus. One such attempt discloses a driving structure on the third day in such a way that no matter what the vehicle speed, drivers opted to drive close to the traffic ahead, compromising the safety. Also it is clear that the distribution of the speed data reflect a nearly normal distribution, with a minor deviation from the Gaussianity except for the third day, where the deviation is rather obvious.. In addition, it is observed that, the headway data behaves as lognormal distribution.

References

[1]. Chen, H., Namdeo, A., & Bell, M. (2008). Classification of road traffic and roadside pollution concentrations for assessment of personal exposure Environmental modelling & Software, 23(3), 282-287.

[2]. Xu, C., Liu, P., Wang, W., & Li, Z. (2012). Evaluation of the impacts of traffic states on crash risks on freeways. Accident Analysis & Prevention, 47, 162-171.

[3]. Montazeri-Gh, M., & Fotouhi, A. (2011). Traffic condition recognition using the K-means clustering method. ScientiaIranica, 18(4), 930-937.

[4]. Nath, R. P. D., Lee, H. J., Chowdhury, N. K., & Chang, J. W. (2010, September). Modified K-means clustering for travel time prediction based on historical traffic data. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 511-521). Springer, Berlin, Heidelberg.

[5]. Wang, X., Cottrell, W., & Mu, S. (2005, September) Using K-means clustering to identify time-of-day break points for traffic signal timing plans. In Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE (pp. 586-591). IEEE.

[6]. Datesh, J., Scherer, W. T., & Smith, B. L. (2011, June). Using K-means clustering to improve traffic signal efficacy in an IntelliDrive SM environment. In Integrated and Sustainable Transportation System (FISTS), 2011 IEEE Forum on (pp. 122-127). IEEE.

[7]. Lee, J., & Jang, K. (2017)A framework for evaluating aggressive driving behaviors based on in-vehicle driving records. Transportation Research Part F: Traffic Psychology and Behaviour. (Available online 13 December 2017, in press, corrected proof).

[8]. Chen, Z., &Xiong, R. (2017). Driving cycle development for electric vehicle application using principal component analysis and K-means cluster: with the case of Shenyang, China. Energy Procedia, 142, 2264-2269.

[9]. Kumar, S., &Toshniwal, D. (2016). A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24(1), 62-72.

[10]. Velázquez-Martínez, J. C., Fransoo, J. C., Blanco, E. E., & Valenzuela-Ocaña, K. B. (2016). A new statistical method of assigning vehicles to delivery areas for CO2 emissions reduction. Transportation Research Part D: Transport and Environment, 43, 133-144.

[11]. Kim, K., &Yamashita, E. Y. (2007). Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. Journal of Advanced Transportation, 41(1), 69-89.

[12]. MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999.
[13]. Bailey, Ken (1994). Numerical Taxonomy and Cluster Analysis. Typologies and Taxonomies. p. 34. ISBN 9780803952591.
[14]. Hartigan J.A. and Wong M.A. (1979). A K-means Clustering Algorithm. Applied Statistics, p. 28:108.
[15]. Xindong W., Kumar V., J. Quinlan R., Ghosh J., Yang Q., Motoda H., Geoffrey J. McLachlan, Angus F. M. Ng, Liu B., Philip S. Yu, Zhou Z., Steinbach M., Hand D.J., and Steinberg D. (2008). Top 10 algorithms in data mining. Knowl. Inf. Syst., 14(1):1–37, 2008.
[16]. Bhowmick A. (2009). A theoretical analysis of Lloyd's algorithm for k-means clustering, unpublished Ph. D. dissertation, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur
[17]. D. Montgomery, (2009),Introduction to Statistical Quality Control, John Wiley & Sons, Jefferson City, 6th edition, 63-165
[18]. S. Figliola, D. Beasley, (2006), Theory and Design for Mechanical Measurement, John Wiley & Sons, United States of America, 4th edition, 109-141
[19]. Kalecki, M.(April, 1945) On the Gibrat Distribution. Econometrica. Journal of the Economic Society. Vol. 13, No. 2, 161-170