

Analyzing Accident Prone Regions by Clustering

**Shuvashish Paul, Ashik Mostafa Alvi, Mahmudul Alam Nirjhor,
Shohanur Rahman, Adeeba Kashfee Orcho
and Rashedur M. Rahman**

Abstract Traffic accidents and injuries related to them have unfortunately become a daily incident for the people of Bangladesh and this is particularly true for people living in Dhaka City. This paper aims to identify the most hazardous regions for such incidents within the Dhaka Metropolitan Region as well as assess their influences. This research effort collects accident related data from the Accident Research Institute (ARI) at Bangladesh University of Engineering and Technology (BUET), Dhaka. This paper utilizes the k -means clustering and expectation maximization method to cluster related incidents together.

Keywords Traffic accidents • Clustering • Accident prone region

S. Paul · A.M. Alvi · M.A. Nirjhor · S. Rahman · A.K. Orcho · R.M. Rahman (✉)
Department of Electrical and Computer Engineering, North South University,
Plot – 15, Block – B, Bashundhara 1229, Dhaka, Bangladesh
e-mail: rashedur.rahman@northsouth.edu

S. Paul
e-mail: reisende@outlook.com

A.M. Alvi
e-mail: alvibdj@gmail.com

M.A. Nirjhor
e-mail: Inirjhor@gmail.com

S. Rahman
e-mail: shohan.nsu.cse@gmail.com

A.K. Orcho
e-mail: adeeba.kashfee@gmail.com

1 Introduction

Road accidents, formally known as traffic collisions refer to the incident where one vehicle collides with another vehicle, pedestrians, animals or any other object in the road. They can end up causing disability, disfigurement, loss of property and even loss of life in the most extreme cases. Several factors are known to influence such incidents—namely vehicle design, as well as several environmental, roadway and motorist characteristics. It is estimated that as many as 54 million people were affected globally by traffic collisions in the year of 2013 alone [1]. This ended up resulting in as many as 1.4 million fatalities [2]. These trends are continuing to surface in Bangladesh as well.

Since Bangladesh is a developing country, the penetration rate for car ownership is rather low. In fact, only about 3 people in every 1000 actually had access to a personal vehicle of their own in 2010 [3]. However, this is steadily rising as people’s purchasing power increases. This means that the rate of traffic accidents too is on the rise in Bangladesh [4]. Accidental injury research was created to tackle this problem and has been largely successful in identifying such trends on a global scale. Such efforts however are not common in Bangladesh and a large portion of incidents are actually never even properly reported. For every two traffic collisions resulting in a fatality in the United States, there are as many as 160 such deaths in Bangladesh [5]. It is no surprise that Bangladesh reportedly has one of the highest road fatality rates in the world [6]. To realize national trends and factors fueling this problem, Bangladesh University of Engineering and Technology (BUET) has established the Accident Research Institute (ARI) which aggregates accidental injury as well as severity data through a direct collaboration with the Dhaka Metropolitan Police (DMP) [7]. We have collected some pre-aggregated data from the ARI and analyzed it to discover relevant trends as well as Dhaka’s deadliest accident spots on a clustered level.

In brief, the major achievements of our work are: (i) Find and visualize the most accident prone areas within the Dhaka Metropolitan Region, (ii) Associate their impacts with local clusters built from the dataset, (iii) Build a cluster map signifying the most dangerous regions within Dhaka.

The remaining of the paper is arranged in 5 sections. Section 2 briefly discusses about the related work in this area. Section 3 showcases our research methodology in detail. Section 4 presents our research findings and Sect. 5 presents our final thoughts on the topic.

2 Related Works

Research related to the topic of traffic collisions and accidental injury is historically unfortunately rather uncommon in Bangladesh. In a recent study, it was found that nearly 22% of all reported traffic collisions in Bangladesh took place within Dhaka

Metropolitan City [8]. It was also found that large proportions of the main street network contained “blackspots,” areas of high relevance that were the largest contributors to traffic accidents.

Another research effort pointed out these constraints, e.g., institutional weakness, lack of professional expertise, political support and proper policies to combat such problems in Bangladesh [9]. One researcher used Multinomial Logit Models (MNL) to analyze similar data but finally switched to Ordered Probit Regression since MNL could not provide ordering of severity levels [6]. They modeled injury severity using a 4 point Likert scale when preprocessing the accidental injury data.

Several researchers have recommended the use of k -means and its variant k -modes alongside Latent Class Clustering (LCC) in analyzing road accident data [10]. They also reported that LCC is computationally infeasible when the data contains a large number of categorical attributes. Other researchers also note that hierarchical clustering (such as Ward’s method or single linkage method) can also be beneficial when analyzing road safety and incident reports [11]. DBSCAN, which is a density based clustering algorithm, is also highlighted when dealing with spatial data (such as longitude and latitude) as being particularly good as they are not affected by outliers. Fuzzy clustering approach has previously been used in determining blackspots that lead to accidents [12]. The authors utilized Fuzzy C-Means (FCM) clustering which is an extension of the k -means algorithm to the fuzzy framework. It utilizes membership degrees instead of binary attributes that says “this point is a member of the cluster,” or “this point is not a member of the cluster.”

3 Methodology

The dataset(s) we acquired from the Accident Research Institute (ARI) at Bangladesh University of Engineering and Technology (BUET) were unfortunately only available as a Portable Document Format (PDF). It was also pre-aggregated by intersections/major city areas.

3.1 *The Dataset*

As the dataset was only available as PDF and as none of the statistical and data mining software suites allow PDF inputs, we had to have it manually converted back to the Comma Separated Values (CSV) format. Our dataset was pre-aggregated by major intersections within Dhaka City and included a range of X/Y coordinates defined atop an arbitrary axis that takes Dhaka’s GPO as the origin point and lays out other regions accordingly for a period of 10 years (2002–2012). An example value of those two attributes would be 304850–304879—which describes the range of the value of the X-coordinate for the area canonically known

Table 1 The dataset showcasing used attributes

Ro...	Tacc	Fatal	Ninjury	Pacc	Pfatal	Pninjury	Peracc	Xavg	Yavg
1	16	12	3	11	10	1	68.800	304,865	72,275
2	19	13	5	9	9	0	47.400	304,385	72,875
3	16	16	0	14	14	0	87.500	304,805	72,305

as Jatrabari. A small snapshot (of 3 records) showcasing the attributes of dataset (pre-analysis) can be seen in Table 1.

Table 1 showcases all of the attributes that we are taking into consideration when clustering. We will briefly describe them: (i) TACC stands for Total Accidents; it is the number of total incidents in that aggregation (data) point, (ii) Fatal stands for the amount of accidents that prove fatal for one or more involved people, (iii) NINJURY stands for normal injury, it describes the amount of non-fatal injuries sustained, (iv) PACC stands for pedestrian accident, it is the number of incidents where pedestrians are involved, (v) PFatal stands for pedestrian fatalities, the number of incidents that result in one or more pedestrian fatalities, (vi) PNINJURY stands for pedestrian normal injury, the number of incidents that result in non-fatal injury to pedestrians, (vii) PERPACC stands for “% of pedestrian accidents,” it simply signifies how many (out of total accidents for that location) incidents involved a pedestrian in it (viii) XAvg and YAvg roughly describe the aggregation point in a X–Y coordinate system *after* data preprocessing. That step is explained in detail in the next section. It is worth noting that several attributes were not considered as a part of the research (such as name of the general aggregation point and what type of road it is), we have chosen to ignore them from our analysis.

3.2 Data Preprocessing

Unfortunately, nominal clustering atop a range of values would not assist us in generating locally significant clusters to show the most hazardous areas within Dhaka. We had to get this converted into one value so we could represent this as a numeric attribute. Initially, we had chosen to simply take the starting part of the coordinate (“left of the hyphen”) and discard the rest. However, this proved to introduce large amounts of error since Dhaka is a relatively small place area wise.

We then decided to split each of these attributes into two portions—one that was in the left of the hyphen, and the other that was in the right. Afterwards, the rounded (up) mathematical average of the two numbers was chosen to represent each location in the X–Y plane.

$$x(\text{new}) = \text{ceil}\left(\frac{x(\text{left}) + x(\text{right})}{2}\right)$$

Several duplicate attributes had to be renamed to make the data importable in the data mining platforms that we put to use. The original dataset for example had two attributes with the same name of “Fatal.” As a part of preprocessing, we also filled in missing values using the global mean for that attribute or column. This data was then converted into the Attribute-Relation File Format through the use of a Java program that is available as a part of the Weka tool suite [13] under the namespace of “*weka.core.converters*”. The final dataset contains an aggregated report of 303 accidents out of which 189 (nearly 62%) proved fatal for one or more of the people involved in the accident. Preliminary analysis of the dataset is given: (i) Total Fatalities: 189 (62% of Total Accidents), (ii) Pedestrian Involvement 149 (49% of Total Accidents), (iii) Pedestrian Fatalities: 126 (42% of Total Accidents) (iv) Non-fatal/Normal Injury Count: 99 (33% of Total Accidents). The most common intersection count was Tee which appeared 24 times.

We also analyze the histograms for different attributes. For example, we can see that the majority of the values of TACC, which stands for Total Accidents, fall within the range of 0–24. We then apply data mining methods to form the clusters and attain aggregated results. These methods are: (i) K-means Clustering (with Euclidean distance) (ii) Expectation Maximization. In the next section, we describe *k*-means clustering in detail.

3.2.1 K-Means Clustering

K-means clustering is popular among all other clustering techniques used in data mining. The algorithm can be divided up into three steps: step (1) Given a list of records or examples, assign each record to the cluster that has the “nearest” distance. In this instance, we are using the Euclidean distance; step (2) Calculate the new mean of each cluster and create new centroids using them; step (3) Repeat from step-1 until the centroids no longer change

Usually, centroids are either randomly assigned from random records (which is the Forgy method), or a random cluster is assigned to each record and the update process is begun [14]. We are using the *k*-means implementations of Weka and Rapidminer in this paper. The main attribute(s) that we are clustering in this case are the X and Y coordinates of the X–Y plane system. They describe positional information within the grid, and we aggregate around this to derive the fatality and injury rates.

The attributes labelled SN (Serial Number), Name (Canonical Name of the Area), Type (Intersection Type), Xarea (X range of the Area) and Yarea (Y range of the Area) were ignored when generating the cluster (since they are nominal, and *k*-means cannot deal with nominal values directly.)

Now, before we could run the clustering algorithms, we had to figure out suitable values of *k* since both *k*-means and Expected Maximization (EM) need the user to tell them the number of clusters that we would like to form. We did some tests that measured the Sum of Squared Error (SSE) for $k = 2$, $k = 3$ and $k = 4$. Those results can be seen in the Table 2.

Table 2 SSE comparison for different values of K

KSSE	Distribution
17.70	C0: 48, C1: 6
15.83	C0: 47, C1: 5, C2: 2
12.7346	C0: 41, C1: 5, C2: 1, C3: 7

The distribution field in the above table describes how many records ended up in which cluster. C0: 47 for example states that 47 records were in cluster 0. SSE calculation was done using Weka.

Keeping in mind that Dhaka is a small city and partitioning it up in too many clusters will probably make the bigger picture impossible to see, we decided to try working with $k = 3$ and $k = 4$. $K = 5$ was ignored because it resulted in the creation of clusters with only one element in it. We ultimately decided to use $k = 3$ as it had an acceptably lower SSE value and did not result in the creation of single member clusters. We did the modelling in both Weka and Rapidminer (to be able to compare between multiple runs and implementations).

The Centroid Table(s) Table 3 containing the clustered values can be seen below.

Cluster 0 ends up with 87% of all records (47 aggregation points), Cluster 1 has only 9% (5 aggregation points) while Cluster 2 has 4% (2 aggregation points). Similarly, in iteration 2 using Rapidminer, the following Table 4 is discovered.

We end up with a similar cluster membership breakdown where Cluster 0 has 38 aggregation points, Cluster 1 has only 1 aggregation point and Cluster 2 has 15 aggregation points. These tables show the values of each centroid that is at the center of each cluster generated. For example, using Weka, the average value of total accidents for Cluster 0 is 4.10. For cluster 1, this is 18.6, the highest among all the clusters. Cluster 2 is in the middle with 8.5. It is important to note that k -means has a chance of differing from implementation to implementation as the results entirely depend on how the initial centroids are chosen. The average of sum of squared errors in three clusters is 15.82 for k -means clustering algorithm.

3.2.2 Expectation Maximization Clustering

Expectation Maximization, commonly known as EM refers to a recurrent process of visualizing the maximum likelihood estimates of parameters that belong to some statistical model. More information on EM could be found elsewhere [15, 16]. We will be using the EM implementation of the (Weka) in this paper. We chose to create 3 clusters with maximum iterations set to 100 and minimum standard deviation set to 1.0E-6. Table 5 depicts the result derived for EM algorithm. The values in the brackets represent the standard deviation across the mean value as derived by the EM algorithm while the values outside the brackets are the avg. mean across that cluster. For example, the average count of Total Accidents in Cluster 0 is 5.8688 ± 2.23 , while in Cluster 1 it is 17.1636 ± 4.18 and in Cluster 2 it is 1.4823 ± 1.15 . Other values are distributed similarly. The cluster membership

Table 3 Centroid values for each cluster’s attributes

Cluster	Attribute	Toolkit	Value
Cluster_0	Total Accidents	Weka	4.1064
Cluster_0	Fatal Accidents	Weka	2.3404
Cluster_0	Normal Injury	Weka	1.5957
Cluster_0	Ped. Accident.	Weka	1.6383
Cluster_0	Ped. Fatality	Weka	1.3404
Cluster_0	Ped. Normal Injury	Weka	0.383
Cluster_0	% of Ped. Acc.	Weka	37.3085
Cluster_0	X-coord	Weka	304493.5106
Cluster_0	Y-coord	Weka	72881.7021
Cluster_1	Total Accidents	Weka	18.6
Cluster_1	Fatal Accidents	Weka	13.6
Cluster_1	Normal Injury	Weka	3.6
Cluster_1	Ped. Accident.	Weka	11.8
Cluster_1	Ped. Fatality	Weka	11.2
Cluster_1	Ped. Normal Injury	Weka	0.8
Cluster_1	% of Ped. Acc.	Weka	63.78
Cluster_1	X-coord	Weka	304607
Cluster_1	Y-coord	Weka	72975
Cluster_2	Total Accidents	Weka	8.5
Cluster_2	Fatal Accidents	Weka	5.5
Cluster_2	Normal Injury	Weka	3
Cluster_2	Ped. Accident.	Weka	6.5
Cluster_2	Ped. Fatality	Weka	3.5
Cluster_2	Ped. Normal Injury	Weka	3
Cluster_2	% of Ped. Acc.	Weka	75.7
Cluster_2	X-coord	Weka	304415
Cluster_2	Y-coord	Weka	72777.5

Table 4 Rapidminer clustering around significant attributes

Attribute	Cluster_0	Cluster_1	Cluster_2
TACC	5.211	8	6.467
Fatal	3.105	6	4.333
NINJURY	1.632	6	2.067
PACC	2.474	4	3.400
PFatal	2.026	4	3
PNINJURY	0.526	0	0.533
PERPACC	37.784	50	49.200
Xacg	304520	304595	304447
Yavg	72626.711	75515	73369.333

Table 5 Clustered attribute values according to EM

Attribute	Tool	Cluster_0	Cluster_1	Cluster_2
Total accidents	Weka	5.8688 (2.23)	17.1636 (4.18)	1.4823 (1.15)
Fatal accidents	Weka	3.34 (1.94)	12.50 (3.41)	0.86 (0.94)
Normal injury	Weka	2.20 (1.97)	3.50 (1.80)	0.72 (1.21)
Ped. accidents	Weka	2.53 (1.08)	11.16 (3.33)	0.41 (0.49)
Ped. fatalities	Weka	1.95 (0.96)	10.16 (3.81)	0.41 (0.50)
Ped. normal injuries	Weka	0.72 (0.74)	1.17 (1.07)	0 (0)
% of Ped. accidents	Weka	46.03 (16.63)	66.48 (15.81)	25.38 (36.66)
X-coord	Weka	304501.47 (156.97)	304525.02 (253.14)	304492.85 (240.13)
Y-coord	Weka	72872.81 (584.56)	72991.41 (628.94)	72874.23 (435.07)

distribution included 26 (48%) aggregation points within Cluster 0, 6 (11%) aggregation points within Cluster 1 and 22 (41%) aggregation points within Cluster 2. The log likelihood, which is a statistical attribute for measuring the quality of a model, was -25.27 for this estimation. Lower values are preferable because it indicates maximization of the EM likelihood function. Table 5 depicts the results.

4 Our Findings

If we take the standard deviation values from EM into account, it appears that both k -means and EM seem to produce similar results. To illustrate this, we manually attach labels to the dataset aggregation points as follows—total accident counts greater than 15 is labelled ‘highly dangerous,’ counts between 5 and 14 is labelled ‘moderately dangerous,’ and counts less than 5 are labelled ‘moderately safe.’ We then compare the results of the two clustering algorithms to find the best one. Those results are illustrated in the validation table, i.e., Table 6. It compares the two clustering algorithms based on their accuracy.

As we can see from the Table 6, EM results in better accuracy. We believe this is due to suboptimal centroid formation on the k -means implementations (since it is randomized on every run)—EM does not have to deal with this problem. Generally speaking, k -means can be thought of as a variant of EM with an assumption that it will result in the creation of spherical clusters. It does this by “hard assigning” data points to specific clusters when it converges, while EM only “soft assigns,” as in it is open to entertaining the possibility of a certain point belonging to any centroid.

After going through the results generated, we can come up with the following findings. To make this easier to visualize, we have plotted the clusters atop a map of the Dhaka City in Fig. 1.

Table 6 Cluster validation between methods

Method	Correctly classified records	Accuracy (%)
Expectation Maximization (EM)	48	88.89
K-means	41	75.93

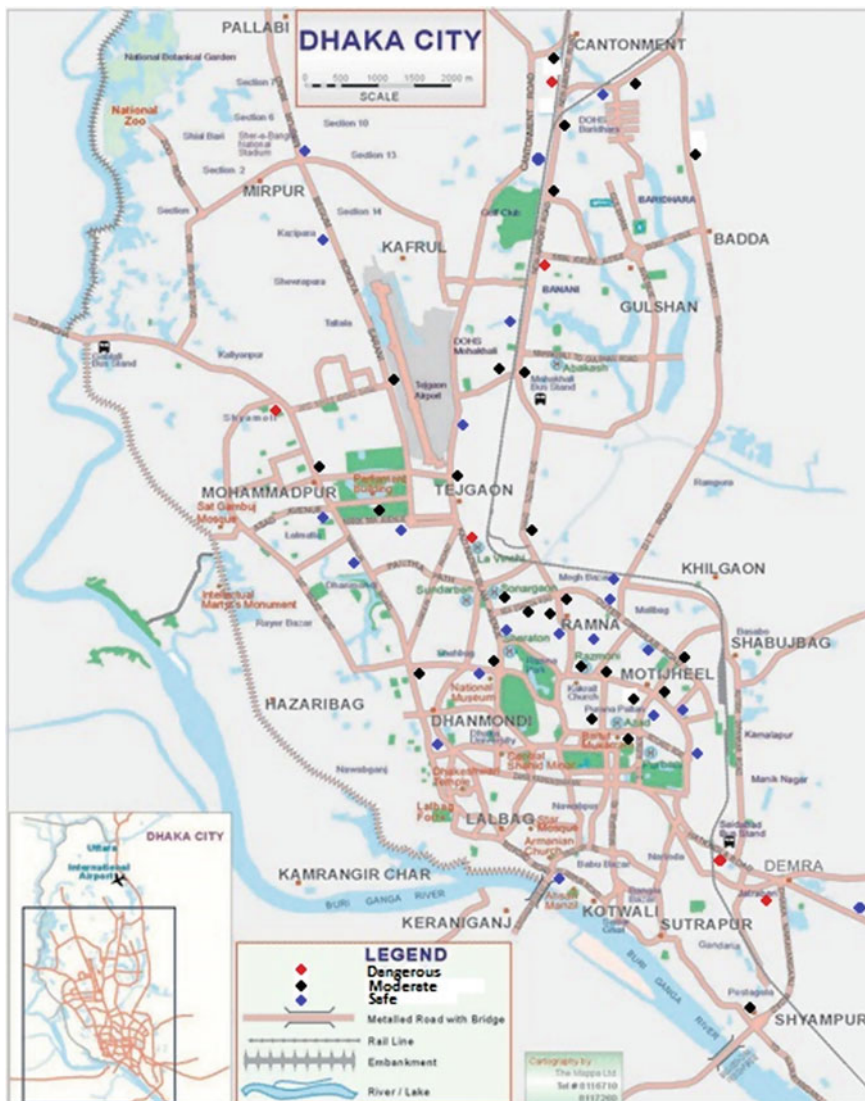


Fig. 1 Cluster map

Cluster 0 proves to be the safest with only around 4 or 5 total accidents reported in the entire dataset when using k -means. EM however finds another set with a much lower accident count—around 1.48 ± 1.15 within Cluster 2. Fatal accidents among those are even lower (around 0.86 ± 0.94), normal injury rate is around 0.72 ± 1.21 . On average, pedestrians are involved in 0.41 ± 0.49 incidents out of which 0.41 ± 0.50 incidents result in fatalities. No pedestrians suffer a normal or non-fatal injury in this cluster. The centroid of this region is located at $X = 304492.85$ and $Y = 72874.23$ and encompasses regions such as Zia Colony Cantonment Gate, Osmani Uddyan, Tongi Diversion Road, Mohakhali and so on.

Cluster 1 proves to be the deadliest using both analytic methods with an average of around 17.16 ± 4.18 accidents out of which nearly 12.50 ± 3.41 prove fatal. Normal injury count is the highest here at 3.50 ± 1.80 , and pedestrians are involved in nearly 67% of these incidents. The pedestrian accident count is 11.16 ± 3.33 out of which nearly 10.16 ± 3.81 prove to be fatal. Pedestrians suffer normal or non-fatal injuries in 1.14 ± 1.07 cases. The centroid of this region is located at $X = 304525.02$ and $Y = 72991.41$ and it encompasses regions such as Jatrabari, Farmgate, Saidabad, Jasim Uddin Crossing, etc.

Cluster 2 (while using k -means) and Cluster 0 (while using EM) provides a moderate accident prone cluster where total accidents are around 5.87 ± 2.23 where around 3.34 ± 1.94 of them prove to be fatal. Nearly 2.53 ± 1.08 out of these cases involve pedestrians out of which 1.95 ± 0.96 end in fatalities in average. The amount of accidents that result in non-fatal injuries to the pedestrians is around 0.72 ± 0.74 . The centroid of this region is located at $X = 304501.47$ and $Y = 72872.81$ and encompasses regions such as Shapla Chattar, Progoti Sarani (Badda), Shahbag, Bijoy Sarani, etc.

5 Conclusion

From the above findings, it can be concluded that several areas of the Dhaka Metropolitan City are indeed a lot more dangerous than other areas. The primary purpose of this research paper is to locate those areas so measures can be taken to deal with the problems. We are glad to report that our findings match up with those found by other researches. Kamruzzaman et al.'s work on a similar subject matter produced similar results [6]. Traditionally, aggregation has been used with geospatial as well as accidental research data in this arena to recognize areas of high importance. We have put data-mining techniques to use in this paper to do the same and can conclude that such techniques work well for this purpose.

Data scarcity has posed a problem for our research effort since day one. While it is known that Bangladesh is a very accident prone country [1, 2], the reporting/recording of information relating to these incidents is often entirely manual and forgotten. We want to give suggestions to higher authorities for better management of accident related data. More extensive research initiative could be taken by other researchers on this computerized rich data set.

References

1. Global Burden of Disease Study 2013, Collaborators: Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 301 Acute and Chronic Diseases and Injuries in 188 Countries, 1990–2013: A Systematic Analysis for the Global Burden of Disease Study 2013, 20 Apr 2016. doi:[10.1016/S0140-6736\(15\)60692-4](https://doi.org/10.1016/S0140-6736(15)60692-4)
2. Global, Regional, and National Age–Sex Specific All-Cause and Cause-Specific Mortality for 240 Causes of Death, 1990–2013: A systematic analysis for the global burden of disease study 2013. *The Lancet* **385**(9963), 117–171 (2015). doi:[10.1016/s0140-6736\(14\)61682-2](https://doi.org/10.1016/s0140-6736(14)61682-2)
3. The World Bank. Motor Vehicles (per 1,000 People): The World Bank. 2013. Archived from the Original <http://data.worldbank.org/indicator/IS.VEH.NVEH.P3> (2014). Accessed 20 Apr 2016
4. Momin, S.M.: Traffic accidents: are the drivers solely to blame? *The Independent*. <http://www.theindependentbd.com/printversion/details/10945> (2015). Accessed 20 Apr 2016
5. Walsh, D.: Can an app really make Bangladesh’s notoriously deadly roads safer? *The Guardian*. <http://www.theguardian.com/world/2015/mar/13/bangladesh-deadly-road-accidents-criticalink-app-emergency-services> (2015). Accessed 20 Apr 2016
6. Kamruzzaman, M., Haque, M.M., Ahmed, B., Yasmin, T.: Analysis of traffic injury severity in a mega city of a developing country. Paper presented at the 4th Road Safety International Conference held in Sydney, Australia, 4–5 Mar 2013
7. Road Safety Facts, 1st edn. Accident Research Institute (ARI), Dhaka (2013). Web, 20 Apr 2016
8. Hoque, M.S., Mahmud, S.M., Kawsar, C., Siddiqui, A.: Road Safety in Bangladesh and Some Recent Advances, 1st edn. Accident Research Centre (ARC), Dhaka (2012). Web, 20 Apr 2016
9. Mahmud, S.M., Hoque, M.S.: Proceedings of the 4th Annual Paper Meet and 1st Civil Engineering Congress, 22–24 Dec 2011, Dhaka, Bangladesh. <https://www.k4health.org/sites/default/files/30bfullpaperroadsafetyresearch.pdf>. Accessed 20 Apr 2016
10. Kumar, S., Toshniwal, D.: A data mining framework to analyze road accident data. *J. Big Data*. doi:[10.1186/s40537-015-0035-y](https://doi.org/10.1186/s40537-015-0035-y). Accessed 20 Apr 2016
11. Doğru, N., Subaşı, A.: Comparison of clustering techniques for traffic accident detection. *Turk. J. Electr. Eng. Comput. Sci.* **23**, 2124–2137 (2015). doi:[10.3906/elk-1304-234](https://doi.org/10.3906/elk-1304-234)
12. Sazi, Y.: Fuzzy clustering approach for accident black spot centers determination. In: *Fuzzy Logic—Emerging Technologies and Applications* (2012). doi:[10.5772/35521](https://doi.org/10.5772/35521)
13. Waikato University: Converting CSV to ARFF. WEKA. <https://weka.wikispaces.com/ConvertingCSVtoARFF>. Accessed 20 Apr 2016
14. Hamerly, G., Elkan, C.: Alternatives to the K-means algorithm that find better clusterings. In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management—CIKM ‘02* (2002). doi:[10.1145/584792.584890](https://doi.org/10.1145/584792.584890)
15. Chen, Y., Gupta, M.R.: EM demystified: an expectation-maximization tutorial. University of Washington, 20 Apr 2016
16. Hartley, H.O.: Maximum likelihood estimation from incomplete data. *Biometrics* **14**(2), 174 (1958). doi:[10.2307/2527783](https://doi.org/10.2307/2527783)