Clustering the vulnerability of traffic accidents in Medan city with fuzzy c-means algorithm

K Syahputri, R M Sari, I Rizkya, U Tarigan, I Siregar and T A Farhan

Department of Industrial Engineering, Faculty of Engineering, Universitas Sumatera Utara

E-mail: khalida@usu.ac.id

Abstract. A traffic accident is one of the global problems that needs serious handling. Traffic accidents happen in various places at different times, this thing causes the difficulty to determine which area has traffic accidents-prone. Information about traffic accidents-prone areas is very needed for the people and regulators. That information can be a consideration for supervision or anticipation, especially for the police. Medan is the capital city of Sumatera Utara province, Indonesia. Medan is the third-largest city in Indonesia after Jakarta and Surabaya also the largest outside Java Island. Medan is one of the cities in Indonesia that has traffic accidents problems. 25.08% of all the traffic accidents in Sumatera Utara province in 2016 happened in Medan. 2018’s traffic accidents data acquired from the traffic unit in Medan Police Office shows a total of 947 traffic accidents that causes 148 died, 456 peoples heavily injured, and 616 people lightly injured. Then, this data is grouped based on the road they happened on and then the grouped data is clustered with the Fuzzy C-Means Algorithm based on the severity of the accidents, resulting in 5 roads being very vulnerable, 14 roads being quite vulnerable, and 224 roads being safe.

1. Introduction
One of the major problems in the world today is the rate of the traffic accidents and deaths on the roads. Each year, an estimated 1.24 million people died in a traffic accident and up to 20–50 million others injured, costing over US $500 billion worldwide. The cost of road traffic injuries is estimated to be between 1–2% gross national product in low-and-middle-income countries, which is over the US $100 billion a year. The current trends show that if urgent action is not taken, traffic accidents injuries could be the seventh leading cause of death by the year 2030, and ninety percent of these deaths occur in low and middle-income countries. Not to mention the emotional trauma of losing the loved ones, psychological impact on traffic accidents victims and permanent disability because of road traffic crashes. The causes of a traffic accidents are multi-factorial and involve the interaction of several pre-crash factors that include humans, vehicles and road environments. Many studies have been conducted to investigate and understand the factors that are contributing to a traffic accident to provide prevention [1].

Traffic accidents are a very common cause of admission to the hospital around the world, especially in developing countries. Disability, loss of days of work as a consequence of the traffic accidents lead to further socioeconomic deprivation in society. Traffic accidents lead to massive social and economic costs. Therefore, traffic safety is an important issue in the view of sustainable development. Road accidents can be the consequence of three different factors such as human, vehicle and environmental-related factors. Precautions regarding these three aspects are necessary to reduce the number of accidents in the long term. Nowadays, traffic safety is considered as the most studied subjects in the field of transportation engineering [2].

Nowadays, the issue of traffic accidents has become a crucial problem in the world. World Health Organization worked with partners including government and nongovernment organizations around the world to raise the profile of the preventability of traffic accidents. In many countries around the world,
there are many organizations in the world to maintain road safety for reducing the threat of the traffic accidents. On the other hand, researchers employed many techniques, especially statisticians techniques to identify the causes of the traffic road accidents using the historical traffic accidents dataset. The data miners explored various parameters or variables for the causes of the road accidents as well as behaviors of the drivers by using different data mining tools and techniques. So many researchers use to spend much of times searching for the best performing data mining algorithm for mining on the traffic road accidents dataset [3].

Data mining is a prominent technology to predict or do some analytics. Traffic accidents occur in different places around the world and the reason for accidents may vary. The main reasons for the deaths on road are traffic accidents, which is not following the traffic rules, overtaking in a wrong way, over speed, not following safety measures of road. Data mining is mainly used to identify the severity of accidents on roads. DMD (Data Mining Data Warehousing) has all the techniques to be used to predict or identify the severity of accidents on roads. DM is used to extract the semantic things over the data set that is a meaningful extract from the data available. The classification techniques like clustering, anomaly detection, clustering and classification rules are used for most of the DM operations on the road accidents [4]. Data mining techniques were chosen as a useful tool to find out the root cause of accidents [5].

The problem of traffic accidents vulnerabilities can be addressed by clustering. Clustering can be done with several methods, one of which is the fuzzy C-means algorithm. Fuzzy c-means (FCM) clustering is an unsupervised technique that has been successfully applied to feature analysis, clustering, and classifier designs in fields such as astronomy, geology, medical imaging, target recognition, and image segmentation [6]. In spatial analysis, a hotspot is generally defined as an area containing dense clusters of localized events. The identification of hotspots on the map is usually made by georeferencing as points the events happened in a certain period. One of the most known ways to detect hotspots is to use cluster techniques, which are an effective way of determining areas exhibiting high concentrations of localized events. Many clustering techniques were used for hotspot detection and event distribution and time-evolution in the spatial analysis [7].

Many previous studies have been conducted in clustering analysis of traffic accidents. Tortum (2015) researches in determining cities similar to each other in terms of health and traffic risk rates in traffic accidents happening in the provinces. In this study, provinces in Turkey were clustered using data about traffic accidents happening outside the cities in the year 2012. Both classical k-means and fuzzy c-means clustering techniques were used for clustering analysis. Clustering analysis of the provinces was conducted according to these risk rates that are determined as a health risk and traffic risk rates. It was observed that the fuzzy c-means technique usually produced more stable results. Also, the fuzzyc-means technique was observed to have been affected very much by exceptional data whereas the k-means technique was influenced very little [8].

2. Methodology
This research is conducted by researching every road section in the city of Medan. The object in the study was a traffic accident that occurred on every road of Medan. This study focuses on analyzing the vulnerability of traffic accidents that occur by performing a clusterization of vulnerabilities using fuzzy C-Means. On this research, data needed for the Fuzzy C-Means algorithm is the number of traffic accidents and the number of victims (Dead, Heavily Injured, Lightly Injured) on each road, according to the traffic unit of Medan Police Office, the order of the weight of traffic accidents victims condition from the highest to lowest are dead, heavily injured, and lightly injured. This method was developed by Dunn in 1973 and improved by Bezdek in 1981. The association to a cluster is verified by computing the reverse distance to the cluster center. The cluster centers verified by FCM depend on the geometric locations of the data points on the plane or space. In the FCM algorithm, an objective function, which should be minimized, is reflected as:

\[ F(U, V, m; X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (U_{ij})^m (|x_j - v_i|)^w \]  
(1)
where m is the fuzzy factor, k is the number of clusters, \( V = (v_1, v_2, \ldots, v_k)^T \) is a cluster centers vector containing the centers of the k clusters, n is the number of data points, \( X = (x_1, x_2, \ldots, x_n)^T \) is the data points vector, \( U = [u_{ij}]_{k \times n} \) is the membership matrix involving of the membership \( u_{ij} \) which shows the membership of \( x_j \) in the ith cluster, and \( \| . \| \) shows the Euclidean distance norm \( (\|Z\| = \sqrt{Z^T \cdot Z}) \). m is used to normalize and fuzzily the memberships the sum of which should be equal to 1. Minimization of \( F(U,V,m;X) \) is carried out through an iterative techniques such as alternating optimization (AO). When \( m > 1 \), an optimal solution that minimizes \( F(U,V,m;X) \) is found as:

\[
 u_{ij} = \left[ \sum_{p=1}^{k} \frac{||x_j - v_i||^2}{||x_j - v_p||^2} \right]^{\frac{m-1}{2}} 
\]

(2)

where \( 1 \leq i \leq k, 1 \leq j \leq n \), and the center of it’s cluster is achieved as:

\[
 V_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m} 
\]

(3)

After clustering the data, a validity index is used to express how well the data have been clustered [9]. The fuzzy clustering approach has previously been used in determining blackspots that lead to accidents. 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 say “this point is a member of the cluster,” or “this point is not a member of the cluster” [10].

3. Result and discussion

The clustering process of the Fuzzy C-Means Algorithm begins with determining the initial value of calculation that can be shown in Table 1 below.

| Number of Clusters (c) | 3          | The initial objective function (Po) | 0          |
|------------------------|------------|-------------------------------------|------------|
| Squared (w)            | 2          | Maximum iteration                   | 100        |
| Smallest error (ε)     | 0.001      | Initial iteration (t)               | 1          |

During 2018, it is recorded that there is 947 traffic accidents in Medan City, with 148 victims dead, 456 victims heavily injured, and 616 victims lightly injured. The number of the clusters that is going to be produced are \( C_1 \) (very vulnerable), \( C_2 \) (quite vulnerable), and \( C_3 \) (safe). Then, for each road, a random number raised for each cluster that the total of these three random numbers must be equal to 1 as an initial membership degree. The initial membership degree can be shown in Table 2 below.

| No. | Roads            | C1 (Very Vulnerable) | C2 (Quite Vulnerable) | C3 (Safe) |
|-----|------------------|-----------------------|-----------------------|-----------|
| 1   | Gatot Subroto    | 0.7                   | 0.2                   | 0.1       |
| 2   | Kapten Sumarsono | 0.6                   | 0.3                   | 0.1       |
| 3   | Jamin Ginting    | 0.7                   | 0.2                   | 0.1       |
| 4   | Sisingamangaraja | 0.4                   | 0.5                   | 0.1       |
| 5   | Gagak Hitam      | 0                     | 0                     | 1         |
| 239 | Turi              | 0.1                   | 0.3                   | 0.6       |
| 240 | Walikota         | 0.1                   | 0.3                   | 0.6       |
| 241 | Waringin         | 0                     | 0                     | 1         |
| 242 | Gunung Mahameru  | 0                     | 0                     | 1         |
| 243 | Asia Timur       | 0.2                   | 0.4                   | 0.4       |
After 42nd iteration, the number of objective function difference is 0.0007. This shows that the condition of the iteration must be stopped has been fulfilled (number of objective function difference is fewer than the number of smallest errors that is 0.001 on the initial value of calculation that can be shown in Table 2). The final membership degree of each roads can be shown in Table 3.

Table 3. Initial membership degree for each road.

| No. | Roads               | C1 (Very Vulnerable) | C2 (Quite Vulnerable) | C3 (Safe)   |
|-----|---------------------|----------------------|-----------------------|-------------|
| 1.  | Gatot Subroto       | 0.68372641           | 0.191223596           | 0.12505     |
| 2.  | Kapten Sumarsono    | 0.93721037           | 0.045282929           | 0.0175067   |
| 3.  | Jamin Ginting       | 0.99311768           | 0.004749704           | 0.00213262  |
| 4.  | Sisingamangaraja    | 0.88507145           | 0.084032983           | 0.03089557  |
| 5.  | GagakHitamstreet    | 0.52528654           | 0.373197114           | 0.10151634  |
| 239.| Turi                | 0.0004076            | 0.00339703            | 0.9961954   |
| 240.| Walikota            | 0.0004076            | 0.00339703            | 0.9961954   |
| 241.| Waringin            | 0.0004076            | 0.00339703            | 0.9961954   |
| 242.| Gunung Mahameru     | 7.374E-05            | 0.00064903            | 0.9992772   |
| 243.| Asia Timur          | 0.0011472            | 0.00097606            | 0.9897767   |

The biggest membership degree shows that the road is in that cluster. Take for example Gatot Subroto road that has the biggest membership degree on C1 (Very vulnerable) that is 0.68372641, compared to C2 and C3. Then, Gatot Subroto road is categorized as very vulnerable road.

Table 4. The initial value of the calculation.

| Cluster       | Number of Roads |
|---------------|-----------------|
| Very vulnerable| 5               |
| Quite vulnerable| 14             |
| Safe          | 224             |

Based on the table above, shows that 5 roads are very vulnerable, 14 roads are quite vulnerable and 224 roads are safe.

4. Conclusion
Medan is one of the cities that has high traffic accidents problems. From 947 traffic accidents that happened during 2018 in 243 roads clustered with Fuzzy C-Means Algorithm resulting on 5 roads being very vulnerable, 14 roads being quite vulnerable, and 224 roads being safe.

Acknowledgment
This research supported by Industrial Engineering Department, Faculty of Engineering, Universitas Sumatera Utara.

References
[1] C C Ihueze 2018 Accidents Analysis and Prevention 11
[2] F Ouni 2018 Modeling traffic accidents occurrence at Hazardous Road Locations: A case study in Tunisia 1
[3] S Hussain 2019 Information and Communication Technology for Intelligent Systems 106
[4] M Gupta 2018 IJECE 8
[5] L Sanmiquel 2018 Int. Journal of Environmental Research and Public Health 15
[6] K S Chuang 2006 Computerized Medical Imaging and Graphics 30
[7] F D Martino 2017 Fuzzy Sets and Systems
[8] A Tortum 2015 Iran J Public Health 44
[9] M J Rezaee 2018 Physic A 489
[10] S Paul Jurnal Farmasi dan Ilmu Kefarmasian Indonesia 4 (2), 60-67