Cluster analysis for road accidents investigation

M.N. Postorino & G.M.L. Sarnè
Faculty of Engineering - "Mediterranea" Reggio Calabria University, Italy

Abstract

Road accidents and particularly urban accidents are one of the most important negative impacts produced by travel, involving both users and non users of the transport system. The number of fatal injuries has increased in the last decades and their social cost has become more and more relevant requiring an in-depth study in order to resolve the problem. This paper proposes an analysis of accidents based on cluster techniques. Cluster techniques need a suitable database in order to group sets of "similar" objects and to identify the most relevant elements that represent the group. The identification of the relevant aspects common to different types of accidents is the first step for a conscious intervention in the transportation system, because the knowledge of the principal causes of accidents can help the analysts of the transportation systems both in the construction of mathematical relationships among accident and causes and can support the choice of suitable actions for reducing the number of accidents (mainly the fatal accidents).

1 Introduction

Road safety is an important aspect of urban and extra urban transportation systems, particularly due to the high social costs it involves. While different actions have been started for resolving the problem of the atmospheric pollution caused by the vehicles moving on the transportation networks as well as different efforts have been made to limit the environmental pollution at the end-of-life of the vehicles, in the safety field the situation is still very serious. The resources devoted to the road safety each year are largely smaller than the real needs; users spend more than 60 billion of ECU for accident refunds by means of the insurance companies, while the amount devoted to the prevention of accidents is
very lower. The situation is different from Country to Country, both for the differences in the regulations in force and for the different sensibility of users to the safety problem. To limit and reduce the number of accidents on the urban and extra urban roads, a first step is the increase of funds for improving the transportation system; in fact, if well designed, the improvement of the road network is one of the more efficient solutions for reducing the number and the severity of accidents in the time.

The social cost due to the accidents both in terms of medical and economical assistance weighs on the community and is really high. In 1990 the U.S.A. registered a total cost due to the accidents near to 137.5 billions of dollar per year. Similar analyses carried out in Europe valued the direct cost due to road accidents equal to 45 billions of ECU, with about 45000 fatal accidents registered on the European roads [12].

Different studies carried out in this field linked the risk of accidents, the percentage or the number of accidents, the number of fatal accidents (dependent variables) to different factors or explanatory variables (independent variables) such as: age, and/or sex of the driver, expert or inexpert drivers, speed, length of the network, use of the safety belts, meteorological conditions and so on [2], [3], [4], [5], [7].

Other kinds of studies concern the aggregate description of the accidents occurred in a region [1], [4], [10], [13], by using indicators for identifying the trend of some relevant variables (total number of accidents, number of fatal accidents, and so on, in a given location and in a given time period) and the black points.

However, this kind of analyses allows relating the event "accident" with one or more relevant variables or describing the trend in the space and in the time period considered, but it does not allow understanding which are the causes that produced the accident. In other words, these analyses cannot establish the degree of similarity among even apparently different events. On the other hand, the knowledge of the main causes producing the event "accident" is very important in order to decide if actions of information and user education are more relevant than actions on the road network or more generally on the network system [8], [6].

In this paper we propose the use of the cluster technique for grouping similar accidents, in order to identify some representative types for each group and then the representative accident of the category, following the approach proposed by Pas [11] for the classification of daily travel activity patterns. The analysis of the identified standard accidents allows better understanding the reasons that causes the beginning of the accident.

2 Main characteristics of urban accidents

As a general agreement, one considers that an "accident" has occurred when at least one vehicle moving along a road gets involved and moves away from the roadway; alternatively, it is considered that an accident has occurred when a vehicle collides with another vehicle, a person or an obstacle. If one or more
people involved in an accident die by thirty days, the accident is classified as "fatal".

Analysis of accidents can be performed by different points of view, because there are a very large number of factors that interact among them to produce the "accident event": human factors, technological factors and environmental factors. Studies already carried out attribute 10-11% of accidents to the non-observance of the safety distance, 15% to driver inattention, 2-3% to meteorological conditions and 52% to undefined causes.

Then, undefined factors are one of the most important causes of accidents, but the large class of factors collected in this set is not useful for analysis. On the other hand, the knowledge of the principal causes of accidents (not due to human factors) is a crucial aspect for the analysts of the transportation systems if some intervention have to be made for reducing the number of accidents (and mainly the fatal accidents).

The knowledge of the factors that contribute to cause an accident is a fundamental aspect because that knowledge allows conceiving methods for improving the road safety.

Road accidents can be considered the consequence of the interactions among the users of the urban or extra urban transportation system and the environment in which they move. An accident is then the result of a sequence of actions and events, due to the interaction among the users and among the user and the system; the strong complexity of this interaction makes difficult to establish which factor could be the main cause of the accident and how little variations on the initial conditions could transform a slight accident in a fatal one.

Furthermore, a large set of factors could be considered for describing the accident, making more complex the classification. In order to reduce the number of possible causes, causal factors and contributory factors can be identified, with the following meaning:

- the causal factors are the deficiencies in the system and the manoeuvres that immediately precede the beginning of an accident, where the "deficiencies" of the system represent a condition of inefficiency that increases the risk of accident;
- the contributory factors represent the causes of some specific deficiencies and manoeuvres; then, each causal factor is linked to one or more contributory factors.

Studies on causal and contributory factors have been carried out in UK within a project of accident analysis developed by the Transport Research Laboratory [3], [4], [5]. The analysis carried out allowed linking the most usual pairs of causal and contributory factors in order to understand the dynamic of accidents.

Finally, an important aspect in the investigation of accidents is the standardisation of the data for verifying similarities among events occurred in different places. At European level, there is a project, called CARE [14], whose aim is the construction of a unified accident data base, at different level of information, for all the Countries of the European Community. Statistics and indicators at European level about the accidents are one of the outputs provided by the system.
3 Cluster analysis for grouping similar events

In order to group similar events in few homogeneous sets, an equal number of factors has to be defined for each event, specifically the accident. Let \( N \) be the total number of factors considered for each accident; an example of such factors is given in table 1. Naturally, the kind of factors considered in the table 1 can change to match the specific objective of the analyst in describing the accident. All the considered factors are qualitative and then a subsequent coding needs in order to make them quantitatively comparable.

Table 1 - An example of some main factors describing the accident

<table>
<thead>
<tr>
<th>human factors</th>
<th>age, sex, type of job, psychological conditions, physical conditions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>technological factors</td>
<td>type of vehicle, weight, age of the vehicle, number of kilometres already travelled</td>
</tr>
<tr>
<td>environmental factors</td>
<td>Meteorological conditions (sun, rain, fog, snow), pavement conditions, light conditions, locations, type of location (one way road, intersection, urban road, extra urban road, and so on);</td>
</tr>
</tbody>
</table>

The procedure presented for the classification of the accidents is briefly depicted in figure 1 and in the following the different steps are described more carefully.

**Step 1**

Step 1 produces the description of the accidents by means of the relevant factors previously identified; the result is a synthetic \( N \times M \) matrix, where \( N \) is the number of factors and \( M \) the number of accidents to be examined. This matrix is the input of the step 2.

**Step 2**

This step produces an analytical comparison among pairs of accidents, defined by their factors, in order to verify similarities or differences; the result of this step is a similarity (or distance) matrix, whose elements are a measure of the degree of similarity among pairs of accidents. The construction of the similarity matrix has been performed by using the following similarity index [11]:

\[
s_{ij} = \frac{\sum_{k=1}^{N} W_k s_k(x_{ik}, x_{jk})}{\sum_{k=1}^{N} W_k} \tag{1}
\]

where:

- \( s_{ij} \) is the similarity index between the accident \( i \) and the accident \( j \);
- \( x_{ik} \) is the observed value of the factor \( k \) for the accident \( i \);
$W_k$ the weight (if considered) of the factor $k$;  
$S_k$ a score function of the factor $k$  
$N$ the number of factors used for describing the accident $i$. 

Because of the factors $x_{ik}$ are defined with a code and they can have only few, predefined values, the score function used is simply:

$$S_k(x_{ik}, x_{jk}) = \begin{cases} 
1 & \text{if } x_{jk} = x_{ik} \\
0 & \text{otherwise}
\end{cases}$$

The result of the step 2 is a similarity matrix $S$, symmetric with element on the principal diagonal equal to 1 (each event $i$ is perfectly similar to itself); the generic element of this matrix describes the degree of similarity between the accident $i$ and the accident $j$, on the basis of the weighted average of the similarities among pairs of factors. The matrix $S$ is the starting point for the identification of similar objects.
Step 3

Step 3 produces the location of the accidents in the Euclidean real space $N$-dimension, being $N$ the number of factors.

The relationship among the objects, specified by the similarity matrix, can be represented in a real space [11] and in order to achieve this goal the similarity matrix has to be positive semidefinite. Gower [15], [16] showed that a matrix of indices of similarity is positive semidefinite if all data of the set of observations are considered. Particularly, each object has to be valued and compared to all the others.

Because of each object (specifically the accident) is defined by the same number of factors and there are no reasons for eliminating an object from the process of valuation and comparison, it can be argued that the all data in the observed set are considered and then the similarity matrix is positive semidefinite. In this way, the matrix can be represented in the real Euclidean space: each accident can be represented by a point and the distances among the different point reproduce the relationship expressed by the similarity matrix.

The result of this step is then a matrix of spatial coordinates that allows identifying each accident as a point in the $N$-dimension space.

The clustering algorithm used in the literature can directly use the similarity matrix for identifying similar object, but if we want to identify the centroid for each group and the distance among each element of the group and the centroid the step 3 cannot be omitted because this kind of information cannot be directly deduced by the similarity matrix.

Step 4

In this step similar accidents are grouped together, on the basis of their location in the Euclidean real space and then it provides a clustering of similar accidents on the basis of the similarity relationships defined at the previous step two.

Each clustering process involves loss of information; in fact the object are grouped on the basis of their similarity but because of the similarity cannot be identified with a perfect equality, the clustering reduces the specific characteristics of each object, specially if the cluster is identified by only one representative element. On the other hand, if we increase the number of cluster for maintaining the peculiarity of the objects, this number will tend to coincide with the number of objects.

The loss of information produced by the clustering process, particularly when each cluster is represented by only one centroid, can be defined in terms of sum of the squares of the errors. More specifically, if $N$ is the dimension of the considered space, referred to the number of factors, $G$ the number of groups and $N_g$ the number of observations in the group $g$, the following expression holds:

$$WSS_G = \sum_{g=1}^{G} \sum_{i=1}^{N} \sum_{n=1}^{N_g} (X_{gin} - \bar{X}_{gn})^2$$

(2)

where:
WSSG is the sum of the squares of the errors for the G groups; 
Xgn is the location along the dimension n of the observation i of the group g; 
Xgn is the average location along the dimension n of the observations of the group g.

Among the different techniques of cluster analysis, the hierarchical agglomerative can be used. Particularly, the algorithm proposed by Ward [17] is one of the most known for resolving the problem. It minimizes the sum of the squares of the errors provided by eqn (2) at each step of the clustering procedure. Furthermore, the algorithm of Ward implicitly assumes that the relationships among the objects, specified by the similarity matrix, can be represented in a real space. At the end of the clustering process, a set of groups is obtained, each one of them includes accidents similar among them.

**Step 5**
The last step analyzes the results obtained at the previous step and identifies the accident representative of each cluster. Because of the aim of the clustering is the identification of homogeneous groups and then the discussion about the characteristics of the objects in the cluster, the final step is the identification of relevant elements that can represent the group. A barycentre (centroid) can be defined in each cluster, as representative of the group; however, the centroid not necessarily has to coincide with an element of the cluster. Furthermore, for each element of the cluster the distance from the centroid can be calculated. Particularly, beginning from the centroid a specific distance can be defined that identifies a boundary: all the elements whose distance from the centroid is less then the boundary distance can represent the group. The following relationship can be defined [11]:

\[ d_{ig} \leq d_g + a\sigma_g \]  

where:
- \(d_{ig}\) is the distance among the centroid and the \(i\)th element of the cluster \(g\);
- \(d_g\) is the distance among the centroid and the element nearer to the centroid of the cluster \(g\);
- \(\sigma_g\) the standard deviation of the distance among the elements and the centroid of the cluster \(g\);
- \(a\) an experimental parameter.

This relationship allows considering more than one representative element of the cluster, based on the not-homogeneity degree of the cluster measured by \(\sigma_g\) that estimates the dispersion into the group.

In fact, eqn (3) allows depicting an ideal circle whose centre is on the centroid and whose diameter is equal to the expression at the second hand of the eqn (3). All the elements that are into the circle can represent the cluster. If \(\sigma_g\) is very large (i.e., the cluster is formed by elements not homogeneous among them) the diameter of the circle increases and the circle is larger. If \(\sigma_g\) is
very low (i.e., the cluster is formed by elements homogeneous among them) the diameter decreases and the circle is smaller, collapsing into the centroid. In this way, even the elements farther from the centroid can be represented, and the information they contain is preserved.

4 An example

In this section a test example is presented by using a set of accidents data collected in the medium size city of Reggio Calabria, in the south of Italy. The use of only a test example is due to the impossibility of having all the data in a short time, because a large part of this data are reported in a widely descriptive format while the procedure described needs encoded data. In this stage, 7 significant groups are identified (table 2). Groups with less than three elements are not considered because they are too specific and they cannot identify a standard event; furthermore, the number of accidents in each group is decreasing from the top to the bottom in table 2, by implicitly defining more frequent and less frequent standard event.

Table 2 – Test example: the identified standard accidents

<table>
<thead>
<tr>
<th>cluster</th>
<th>Description of the standard event representative of the cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Accident occurred at the intersection, with faint light, no fog, no rain, in normal pavement conditions; driver less than 35, inattentive; age of the car less than five.</td>
</tr>
<tr>
<td>2</td>
<td>Accidents occurred along a two-way road, with faint light, rainy weather, normal pavement conditions; driver less than 35, high speed driving; age of the car less than five.</td>
</tr>
<tr>
<td>3</td>
<td>Accident occurred along a two-way road, good light conditions, no fog, no rain, bad pavement conditions; driver less than 35, inability to control the car; age of the car more than five and less than ten.</td>
</tr>
<tr>
<td>4</td>
<td>Accident occurred along a two-way road, good light conditions, rainy weather, normal pavement conditions; driver less than 35, inability to control the car; age of the car more than five and less than ten.</td>
</tr>
<tr>
<td>5</td>
<td>Accident occurred at the intersection, with good light conditions, no fog, no rain, in normal pavement conditions; driver more than 35 and less than 50, inattentive; age of the car more than five and less than ten.</td>
</tr>
<tr>
<td>6</td>
<td>Accident occurred along a one-way road, good light conditions, no fog, no rain, bad pavement conditions; driver less than 35, high speed; age of the car more than five.</td>
</tr>
<tr>
<td>7</td>
<td>Accidents occurred along a two-way road, with faint light, no fog, no rain, normal pavement conditions; driver less than 35, high speed driving; age of the car less than five.</td>
</tr>
</tbody>
</table>

As showed in table 2, the accidents identified as representative are
characterized by different combination of the considered factors; particularly, in this first stage only a subset of the factor previously indicate in table 1 are here considered. A complete standardisation of the accidents in the medium size sample city considered cannot be extracted because of the analyses are still in progress. However, even if still incomplete, the analyses carried out with the procedure described shows that accidents can be standardised although the apparently large number of cases. When the most relevant standard events are identified, they can be analyzed in order to verify which factors are determinant for the beginning of the accident. In other words, the identification of the standard events is the starting point for linking the accident with the relevant factors and for verifying quantitatively the weight they have in explaining the origin of the accident.

5 Conclusions

In this paper a procedure for the evaluation of the accidents has been described, in order to obtain a clustering of similar accidents on the basis of their characteristics. The procedure is formed by different interrelated steps that allow identifying homogeneous groups of accidents. The clustering of accidents in homogeneous groups allows separating the elements that are not similar, but, at the same time, the individual characteristics of each event is maintained in the group. Further developments can be expected by defining a different score function (that in this work is a simply binary one) and above all by constructing an analytical function that links the main identified factors for each representative event to the generic accident, in order to verify the impact produced by modifying one or more of the values of the factors. This last aspect is already object of the search in progress.

References

[6] Brouwer M Road safety information system: Key information supporting traffic safety policy in the Netherlands, 7th International Conference on Traffic Safety on Two Continents, 1997


[16] Gower J. C. A general coefficient of similarity and some of its properties, Biometrics, 27, 1971