High accuracy crash mapping using fuzzy logic

Maria-Ioanna M. Imprialou *, Mohammed Quddus, David E. Pitfield

School of Civil and Building Engineering, Loughborough University, Loughborough LE11 3TU, United Kingdom

A R T I C L E   I N F O

Article history:
Received 25 October 2013
Received in revised form 5 February 2014
Accepted 3 March 2014

Keywords:
Crash location
Crash mapping
Fuzzy logic
Map-matching

A B S T R A C T

Accurate crash location data in crash databases can be shown to be essential for crash modelling, crash mapping, hazardous road segment identification and other studies that aim to decrease the number of crashes within a network area. In this paper a generic and high-accuracy automatic crash mapping method is developed and presented. The methodology is based on a transformed map-matching method for candidate road segment identification and on a fuzzy logic inference system for the final road segment selection. The method is implemented by employing all injury and fatal crashes that occurred during 2012 in the UK Strategic Road Network but can be transferred to other network/crash data. The accuracy of the developed method is estimated to be 98.9% (±1.1%) correct matches. The results of this method are compared to other less advanced crash mapping methods.

1. Introduction

In order to mitigate the impact of road traffic crashes, it is important that reported crashes contain key variables that include crash time and location, road, driver and vehicle characteristics and contributory factors related to drivers’ errors, defects in the vehicle(s) and the problems with the road environment. Crash data can therefore be considered as the primary safety performance data because an in-depth analysis of such databases may assist in designing effective countermeasures and formulating efficient crash prevention policies (Koike et al., 2000). Since road traffic crashes are predominantly spatial events, all the relevant information that describe the crashes may not be very useful if the locations where crashes took place on the network are not correctly specified. The identification of the crash location is very important as it contributes significantly to detailed crash analyses aiming to identify the contributory and causal factors including the reconstruction of the crash.

The accuracy of crash analyses in terms of modelling the severity and frequency of traffic crashes primarily depends on the availability of high quality and reliable safety performance data (Austin, 1995; Loo, 2006; Tarko et al., 2009). Crash data are often not reliable and therefore it is essential for it to be refined so as to enhance the overall quality. Spatial data are likely to be erroneous as space is multidimensional and that complicates the identification of a precise crash location, especially in complex road configurations such as junctions, roundabouts and flyovers (see Fig. 1). It is not known to what extent these errors are acceptable because the differences of road characteristics and the traffic conditions between two network points that are physically very close to each other can be important (e.g. the difference between a main motorway and its slip road) but there are indications that the estimated coefficients of a safety model differ significantly when the corrected crash
locations are used (Tegge and Ouyang, 2009). As a result, spatial crash data need to be enhanced through some special treatments in order to be confident about their quality.

Road traffic crash data are usually recorded by police officers who visit the crash scene just after a crash occurrence. Crash location is reported in several different ways: (1) by recoding spatial variables such as the road name and type, the proximity to junctions, (2) by assigning a crash at a point on a segment through the relative distances from some defined markers placed on the road and other physical facilities (i.e. a linear referencing system) and (3) by recording a single set of coordinates using a GPS or the national grid coordinate system. As crash databases usually do not include very precise geographic information though, it is reasonable to apply geocoding methods in order to estimate the exact crash location (Kam, 2003). Based on the fact that “each accident is a geographic event in the sense that is tied to a unique location defined in a given referencing framework” (Thill, 2000), a very common practice followed by many authorities around the world is recording the grid reference of the collision spot obtained by grid maps or Geographic Information Systems (GIS) so as to achieve a higher level of accuracy.

However, even when the crash location coordinates are recorded, it is still not guaranteed that the location of the crash can correctly be identified on a road network map. This is due to the error that these measurements may include and the inconsistency between the network databases that are used for crash mapping and the actual road network. For example, there are many simplified digital road maps in which roads are represented only by their centrelines and in some cases omitting features of the actual road geometry. As the majority of crash locations do not fall exactly on these centrelines, they have to be transferred and matched to the correct road segments. This is a quite challenging process, especially when they occur at areas with complex road networks where many road segments intersect, such as at junctions. Another important reason for the errors in recording crash location is that these crash data are primarily collected for administrative purposes rather than for scientific analyses (Loo, 2006). All the above, lead to the conclusion that when the identification of crash locations plays an important role in a traffic crash analysis study, a matching process of the crashes with the correct road segment would benefit the quality of the data and possibly the results.

In summary, it can be concluded that although precise crash locations are usually missing from the majority of crash analyses, assigning crashes onto the correct segments where they occurred is crucial if the purposes are:

- Locating hazardous segments within a network so as to design effective engineering countermeasures (e.g. altering road geometry) (Bíl et al., 2013; Karlaftis and Golias, 2002).
- Modelling segment-based traffic crashes with the aim of identifying the factors affecting crash frequency (e.g. curvature, gradient, traffic density and flow) and crash prediction (Wang et al., 2009).
- Spatial distribution of safety risk across the network. This can be employed for risk mapping that may lead to the introduction of targeted and specialised crash prevention measures (Steenberghen et al., 2004; Loo, 2009).

The primary objectives of this paper are therefore to highlight the importance and the challenges of crash mapping on road segments, especially in complex road networks and to develop a new generic and transferable crash mapping algorithm so as to map accurately traffic crashes onto road segments aiming to enhance the quality of crash data.

This paper is organised as follows: firstly, a brief literature review of prior research on crash mapping and an introduction to some map-matching techniques that can be applied to crash mapping are presented followed by the aim of this paper. Next, detailed descriptions of the data and the crash mapping method and the performance evaluation process are provided. Finally, the evaluation method and the results of the implementation of the crash mapping method are presented along with the main conclusions drawn from this study.
2. Literature review

2.1. Crash mapping methodologies

The number of crash mapping methodologies that can be found in the literature to date is relatively low as crash mapping is not a common area of research in road safety. The approaches vary according to the aim of each study and the data that are used in each case. Generally, crash mapping techniques use all types of spatial data that are included in crash records in order to increase the possibilities of accurate matching. Although Table 1 provides the key features of existing algorithms found in the literature, a brief discussion for each of the methods is presented in this section.

The earliest crash mapping methods are simplistic and are mainly based on the reported crash location and its distance from the adjacent roads. The matching road selection was done either by using buffer zones of predefined radius (Austin, 1995) or by allocating (henceforth: snapping) the crashes to the nearest intersections (Levine et al., 1995). A more refined approach of the latter method was applied by Loo (2006) where crashes were snapped to the closest segment or node of the network that had the same road name and type with the reported accidents. It is noteworthy that Loo (2006) found that 7.5–11.0% of the correct crash locations were not at the nearest segment or node of the road network.

More recent crash mapping algorithms tend to use additional variables from the crash reports and more sophisticated techniques. Wang et al. (2009) used a weighted score scheme based on the perpendicular distance of the crash location to the centreline of each candidate carriageway ($d_i$) and the angular difference between the direction of the involved vehicle and the road ($\Delta \theta_i$). These two measures were combined for the estimation of the weighting score ($W_{Si}$, see Eq. (1)) that was supposed to take the highest value for the most likely road link, that was therefore selected as the correct one.

$$W_{Si} = \frac{1}{d_i} + \cos(\Delta \theta_i), \; d_i \neq 0$$

Qin et al. (2013) developed an automatic crash mapping methodology (C-MAT) that converted the crash location data to specific points based on a linear referencing system and on an “On-At” table containing all the possible combinations of road intersections for each road. Crashes were assigned to a road segment that included the intersection – that was identified by specific points based on a linear referencing system and on an “On-At” table containing all the possible combinations of road name, direction and distance according to the crash reports.

A completely different approach was that of Tarko et al. (2009) who attempted to apply probabilistic linking techniques (that were used successfully for matching of various population records (Fellegi and Sunter, 1969)) for automatically

<table>
<thead>
<tr>
<th>Author</th>
<th>Aim</th>
<th>Method</th>
<th>Main Variables</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin (1995)</td>
<td>Error identification/correction of accident reports</td>
<td>Buffer Zone</td>
<td>Location coordinates, road class, road number, district, speed limit, pedestrian crossing facilities, junction control, junction detail and carriageway types and markings</td>
<td>Simple, easy and fast to implement, especially in a GIS framework</td>
<td>Strict buffer zone limits, possible ambiguity of results for junction crashes</td>
</tr>
<tr>
<td>Levine et al. (1995)</td>
<td>Identification of spatial patterns in motor vehicle crashes</td>
<td>Crashes snapped to the closest junction</td>
<td>Crash location coordinates, road names</td>
<td>Very efficient for mapping large datasets</td>
<td>All crashes were allocated to junctions (53% inaccurate datasets)</td>
</tr>
<tr>
<td>Loo (2006)</td>
<td>Crash mapping and accident data correction.</td>
<td>Crashes snapped to the closest junction or road (road name and district check)</td>
<td>Crash location coordinates, road name, district board</td>
<td>Relatively reliable results</td>
<td>Junctions represented as a single point (lack of precision for the location of junction crashes), reported road types are not taken into account</td>
</tr>
<tr>
<td>Wang et al. (2009)</td>
<td>Spatial analysis of the relationship of crashes with congestion</td>
<td>Maximum weighted score</td>
<td>Perpendicular distance, Angular difference of the vehicle’s intended direction and the road direction</td>
<td>Straightforward, very efficient for selection between the two directions of a motorway</td>
<td>Not suitable for dense road networks, strong dependence on angular difference</td>
</tr>
<tr>
<td>Tarko et al. (2009)</td>
<td>Crash mapping</td>
<td>Probabilistic linking technique</td>
<td>Main and reference road names, county, junction type and other</td>
<td>Crash location coordinates not necessary</td>
<td>Identification of more than one segment matching</td>
</tr>
<tr>
<td>Qin et al. (2013)</td>
<td>Crash mapping</td>
<td>Conversion of linear reference system to a set of crash location coordinates</td>
<td>Road name (main and intersecting), road type, road direction, On/At tables</td>
<td>Reliable results (intersecting road information enhance the validity)</td>
<td>Strong dependence on the accuracy of on/at location information, no alternative selection way when junction information is inaccurate or missing</td>
</tr>
</tbody>
</table>
assigning crashes to roads by identifying the highest likelihood locations. The main drawback of this method was that it could not identify a unique road link per crash.

2.2. Crash mapping as an application of map-matching

Map-matching algorithms – like crash mapping algorithms – combine vehicle positioning data (i.e. x and y coordinates) with digital road network data aiming at identifying the correct segment in which a vehicle is travelling as well as the accurate position of the vehicle on the segment so as to track the routes of vehicles (e.g. Quddus et al., 2003, 2007). These two techniques are different from each other in two primary ways: (1) mapping a crash point onto a segment does not need data from other mapped crashes meaning that the process is quite independent; (2) crash mapping algorithms do not require to provide mapping results in real-time (i.e. a post-processing offline approach).

Despite their differences in function and purposes, some of the concepts used in the map-matching can be employed in crash mapping. For instance, two map-matching methods that could be used in crash mapping is the “point-to-point matching” and the “point-to-curve matching” (see more at: Bernstein and Kornhauser, 1996).

Crash mapping can also be aided by the initial mapping process of the probabilistic map-matching algorithm in which an elliptical error region (with semi-major equal to the error variances of GPS systems (Zhao, 1997)) around the position fix is placed to select the candidate segments (those whose nodes are within the ellipse) for the matching purpose. The correct segment is identified after the implementation of a number of criteria that differentiate between algorithms. Quddus et al. (2007) mention that in some cases the identification of candidate segments is not possible with the method of the error ellipse, and as a possible solution they suggest the formulation of an error circle instead. This change converts the identification of candidate segments to a circle-line intersection problem decreasing the complexity of calculations and the probability of errors.

2.3. Aim

Crash mapping methodologies that have been developed so far differ significantly from each other, in terms of methodological approach and the type and number of variables used, as they are strictly designed according to the individual characteristics of each study area and databases. One common point between them is that independently the accuracy of their results (that varied significantly) crash mapping methodologies suffer from lack of transferability.

This paper presents a generic methodology for automatic high-accuracy crash mapping that includes a combination of a map-matching methodology with fuzzy logic and demands limited manual intervention. The novel features of this work are: (a) the transformation of a map-matching concept to crash mapping, (b) the application of a Fuzzy Inference System for crash mapping, (c) the use of the network shape points for link disaggregation, (d) the development of an algorithm for efficient identification of candidate links within an error circle and (e) the development of a method for obtaining a representative sample of crashes from a road network. The fact that the elements that are used for the implementation of this methodology are some of the most common variables that crash databases include (e.g. reported road name, road type, reported crash location coordinates and vehicle’s intended direction) makes it transferable to different networks. Moreover, the use of multiple variables enhances the resistance of the method to erroneous data.

3. Data description

The data that were used in order to develop the new crash mapping methodology were obtained from: (a) the National UK Road Accident Database STATS 19 and (b) the Highways Agency Pavement Management System (HAPMS).

3.1. UK crash records database – STATS 19

STATS 19 is a set of data collected by the Police in the UK (Department for Transport, 2011). Reportable crashes are those that include at least one injured casualty. The crash database for this study comprises all 10,520 STATS 19 reports of crashes that occurred during 2012 in the network of motorways and trunk roads of England that are managed by the Highways Agency, the so-called Strategic Road Network (SRN), and the most important variables of the database that were used were:

- **Accident Reference Number**: A unique seven-digit sequence that is used to distinguish road crashes.
- **Location**: A pair of six-digit coordinates (Eastings and Northings) obtained by the Ordnance Survey Grid map.
- **1st Road class**: The class of the road where the crash occurred. In the SRN Road Class can be either motorway (M) or main single carriageways (A).
- **1st Road number**: The number that corresponds to the road where the crash occurred.
- **Road type**: Roundabout, one way street, dual carriageway, single carriageway, slip road or unknown.
- **Speed limit**: The posted speed limit on the road where the crash occurred.
- **Junction detail**: This variable expresses the proximity of the crash location to a junction. If the crash took place on a road section that is located less than 20 m from a junction, then the junction type should be reported (e.g. Roundabout, slip road, junction with more than four arms etc.).
- **2nd Road class**: The class of the intersecting road (if any).
- **2nd Road number**: The number that corresponds to the intersecting road.
- **Vehicle movement compass point**: The intended direction of the vehicle just before the incident measured with compass and reported using the four cardinal points and their intermediates (N, NE, E, SE, etc.)

As the reporting method for the Vehicle Movement Compass Point is restrictive, the accuracy of the vehicle direction measurements is limited with the possible deviation from the actual intended direction of the vehicle reaching up to 22.5°. For example, if the actual vehicle direction is 112.5° then in STATS 19 the Vehicle Movement Compass Point will be either E (90°) or SE (135°) that both have 22.5° difference from the actual intended direction.

3.2. Road network database – HAPMS

HAPMS (Highways Agency Pavement Management System) is a computer-based model of the SRN (scale 1:2500) that is represented by a section referencing system that divides the network into sections with consistent road characteristics (road type, name, number of lanes etc.) and specified starting and ending points. The network is represented by a system of segments and nodes and each section is assigned a unique name (i.e. section label) (HA, 2009). The maximum horizontal error of a digital map with map scale 1:2500, like the HAPMS network data, is 2.84 m (Quddus, 2006) indicating that the base map employed in the analysis possesses good accuracy.

In order to decrease the length of individual road sections so as to improve the road direction accuracy and consequently to increase the probability of mapping each crash to the accurate location, HAPMS polylines are divided into smaller segments defined by shape-points. Shape-point coordinates by section label were extracted in a GIS environment and a new set of nodes and segments was developed using the method proposed by Quddus (2006). In this way, the network that initially comprised of 20,734 road sections with average length 743.77 m, was divided into 211,247 piece-wise straight segments with average length: 72.31 m. Fig. 2(a) represents two road sections of the HAPMS network on the A500 main carriageway (3400A500/126 and 3400A500/226). The start and the end nodes of these sections are displayed by the rhombuses. Fig. 2(b) represents these two sections as they are divided by their shape points into 18 and 19 smaller road segments respectively. The boundaries of each segment are illustrated by the numbered points (0 to 1, 1 to 2, and so forth).

The features of the final network database were:

- Coordinates of starting node of the segment.
- Coordinates of ending node of the segment.
- Road name.
- Road type.
- Section label.

![Fig. 2. Disaggregation of the network data into a new network database based on the shape points (a) road sections and (b) disaggregated road sections.](image-url)
4. Fuzzy logic based crash mapping (FLCM)

The objective here is to identify the correct section of the road for each of the 10,520 crashes in the STATS 19 database. Once a road section is identified then the next task is to determine the location (i.e. x- and y-coordinates) of the crash on the road section. For this purpose, a three step process is developed (Fig. 3):

1. The network segments are firstly filtered in respect to their adjacency to the reported crash location and their road name and type so as to form a set of candidate segments.
2. Each of the candidate segments is then evaluated for its “goodness of matching” with the reported location and vehicle movement direction just before the crash using a fuzzy logic (FL) Inference System.
3. Each crash is allocated to a suitable point on the selected segment that is considered to be the actual location of the first impact.

4.1. Candidate segment identification

As the crash coordinates that are recorded in the STATS 19 reports rarely fall exactly onto a specific road section, a combination of checks between the common variables within the two databases (i.e. crash and network) are carried out in order to identify the road section that has the highest likelihood to be the correct one in which the crash occurred. Since the network map is represented by 211,247 piece-wise straight segments, it is essential to apply an initial selection process where only the most relevant segments to the crash location will be kept for further processing and for the final matching (hence termed as candidate segments).

Candidate segment identification follows the method outlined in Quddus et al. (2007) in which an error circle is formed around the reported crash point. The radius of the circle is subject to the quality of the network and crash location data. This can be empirically derived from a sample.

Road segments that fall within the error circle or segments that physically intersect with the error circle are considered to be the candidate segments. This can be termed as a classical circle-line intersection problem in which a segment can be considered as a candidate segment when one of the two following conditions is satisfied: (a) both of the nodes of a segment fall within the circle, or (b) there exists at least one intersecting point between a segment and the circumference of the circle. In other words a road segment \((N_1N_2)\) (Fig. 4) is candidate for a crash with reported location \(O(x, y)\) (error circle \((O, r)\)) when:

\[
(\overrightarrow{OP}) \leq r
\]

Unless:

\[
\overrightarrow{v_{1m} \cdot v_{2n}} > 0 \quad \text{for } n = 1, 2;
\]

and

\[
|\overrightarrow{v_{mn}}| > (l_1l_2) \quad \text{for } m = 1 \text{ or } m = 2 \text{ and } n = 1, 2.
\]

where \(\overrightarrow{v_{mn}} = \overrightarrow{l_nN_m}\) for \(m = 1, 2 \text{ and } n = 1, 2\).

Since road configuration close to junctions can be very complex, the number of candidate segments with different characteristics is high. To avoid to the extent possible the number of mismatches the candidate segments are filtered based on their road name and road type. Segments that have different road name and road type from that reported in the crash database are excluded from the next step of the process. However, due to some inconsistency that appears to exist between the databases, this filtering process can be proven to be very restrictive where no match is found. In these cases, the filtering rules are relaxed and the segments that remain for further evaluation are those that have either the same road name or road type with the crash report.

This process is briefly displayed at Fig. 5 in which the crash location in the STATS19 database is denoted by round symbol A. An error circle of radius 100 m is then formed around the crash point, A (Fig. 5(b)). The initial set of candidate segments are
shown in Fig. 5(c). According to STATS19 crash database, the crash occurred on A5013 with ‘main carriageway’ road type. Based on this information, Fig. 5(d) shows the final set of candidate segments on A5103.

The radius of the error circle that is used in this research is set to 100 m based on empirical observations and the characteristics of the network data. From some initial manual matching of crashes (using the STATS19 crash database of 2011) it was found that the perpendicular distance of the reported crash location to the matching segment was usually considerably less than 50 m. As a consequence, and taking also into consideration the segment lengths’ descriptive statistics, an error circle with 100 m radius includes a reliable number of candidate segments, eliminating both the probabilities of false alarm and missed detection.

If no segments are found within the 100 m circle then the radius is gradually expanded up to 200 m with a 50 m step and the filtering rule changes allowing the inclusion of segments with either the same road name or the road type as depicted in the flow chart (see Fig. 6). However, if still no candidate segments can be identified, the crash remains unmatched and is flagged in order to be manually matched.

4.2. Candidate segment evaluation and selection of the correct segment

For the final selection of the matching segment, the “goodness of matching” of each of the filtered candidate segments with the crash information is evaluated by applying a Mamdani Fuzzy Inference System (FIS) using the fuzzy logic design application included in Matlab 13. In this study it would not be easy, and possibly not successful, to set some pre-defined thresholds that would indicate which segment among the candidates is the most likely for the crash to occur. Fuzzy logic is a technique that is used when reasoning is not determined by exact rules but from different levels of truth (from completely true to completely false) and for interpretation of linguistic terms (such as “short” distance). Additionally, fuzzy logic is suitable for analyses that include data likely to be imprecise (MathWorks, 1999). The two input variables of the FIS are: the distance ($D$) from the crash point to a candidate segment and the angular difference ($\Delta \theta$) of the intended vehicle direction and the direction of a candidate link. $D$ is defined as the perpendicular distance from the reported location point to the candidate link, if the reported location’s projection point is between the two nodes of the link; otherwise it is the minimum of the distances between the crash location and the link’s nodes.

After selecting the set of input and output variables, the remaining two major components of a FIS are: (1) fine-tuned membership functions and (2) fuzzy rules (MathWorks, 1999). Since membership functions can take on different shapes and forms, one of the challenging aspects of designing a FIS is the identification of membership functions in terms of their shapes and fine-tuning. In this study, an empirical analysis was conducted in which most of the commonly used membership functions (e.g. triangular, trapezoidal, sigmoidal and Gaussian) were explored. The results indicated that the performance of the FIS does not significantly vary by the shape of membership functions. Triangular and trapezoidal functions however offer slightly better accuracy (about 1% higher). Therefore these are considered in the FIS.

Membership functions are fine-tuned based on empirical observations and the nature of the data that are used for calculating $D$ and $\Delta \theta$ respectively. The threshold values of each membership function are determined empirically after matching manually 200 crashes obtained from an independent database; the STATS19 crash reports of 2011. The exploratory manual matching process was essential for understanding the range of the two input variables and consequently for the determination of the number of the membership functions for each of the input variables as well as their thresholds. The angular difference ($\Delta \theta$) includes two membership functions (i.e. “small” and “large”) that represent how acceptable an angular difference is taking into account that for a correct segment it rarely and only slightly exceeds the measurement error (22.5°). The distance ($D$) includes three membership functions (i.e. “short”, “medium”, “long”) expressing the level of the relative distance of the candidate segment to the reported crash location. The fuzzy rules represent all the possible
Fig. 7 shows the fine-tuned input and output membership functions. The following six rules are applied to the FIS:

- If \( D \) is “short” and \( \Delta \theta \) is “small”, then (Matching Score is “very good”).
- If \( D \) is “medium” and \( \Delta \theta \) is “small”, then (Matching Score is “good”).
- If \( D \) is “long” and \( \Delta \theta \) is “small”, then (Matching Score is “moderate”).
- If \( D \) is “short” and \( \Delta \theta \) is “large”, then (Matching Score is “moderate”).
- If \( D \) is “medium” and \( \Delta \theta \) is “large”, then (Matching Score is “bad”).
- If \( D \) is “long” and \( \Delta \theta \) is “large”, then (Matching Score is “very bad”).

Fig. 5. Candidate segment identification process in four steps: (a) reported crash location, (b) error circle formation, (c) potential candidate segments and (d) final set of candidate segments.
Fig. 6. The sequence of checks for each road segment of the network for the candidate segment set formulation: (a) Initial error circle radius \( r = 100 \text{ m} \) and (b) expanded error circle radius \( 100 \text{ m} < r < 200 \text{ m} \).

Fig. 7. FIS input (a and b) and output (c) membership functions.
For the defuzzification of the rules the method of centroid was applied and the output variable was a matching score that ranged from 0 to 100. The candidate segment with the highest “Matching Score” is considered to be the most likely to be the segment where the examined crash should be located (Correct Link). In the case of more than one segments having equal “Matching Scores” the segment with the smaller $D$ from the crash location is selected.

4.3. Crash location identification

After the identification of the correct segment the exact location of the crash on the selected segment is determined. The actual location is a point on the correct segment that is either the perpendicular projection of the reported crash location point on the correct segment if that falls between the two nodes of the segment or the closest of the two nodes. Following, the distance between the reported and the estimated location is calculated (henceforth: Distance).

Aiming to ensure the results’ accuracy to the maximum level, the three-step procedure described above was applied three times separately for: all the crashes, crashes that were reported to occur on roundabouts and crashes that were reported to occur on slip roads. By separating roundabout and slip road crashes and adding some additional steps where appropriate, mismatches are avoided to the maximum level. As it is mentioned at Section 4.1, from the initial manual checks it was observed that Distance was at the majority of the cases lower than 50 m. That is why is when FLCM selects a segment that is placed 50 m or more from the reported location a manual check on the accuracy of this result is useful.

- All crashes
  Network Database: HAPMS
  Additional steps:
  (1) If $Distance \geq 50$ m, the crash location is manually checked and changed if necessary.
  (2) If Candidate Segments = 0, the crash location is manually determined.
- Roundabout/Slip Road crashes
  Network Database HAPMS (Roundabouts/Slip Roads only)
  Additional step:
  (3) If $Distance \geq 50$ m or Candidate Segments = 0, the crash location is substituted with the respective correct segment estimated by All crashes.

5. Method evaluation

In order to quantify the accuracy level of the FLCM a sample of the examined cases were matched manually to road sections, and the results were compared to the respective sections that are identified by the developed method. For the manual crash mapping additional variables from the STATS 19 were used such as speed limit, junction details, 2nd road class and number. The inclusion of the additional variables to the manual checks and the fact that during a manual crash mapping process it is possible to treat each case individually when necessary (e.g. crashes near complex configurations), makes the results of this process highly reliable. Therefore, manually mapped crashes form a suitable reference set for the examined algorithm. The sample was obtained by dividing the entire road network into 70 exhaustive and mutually exclusive clusters of equal areas and selecting crashes randomly with the quota sampling approach (Fig. 8). The size of the sample is 716 cases and it was estimated using the equation of sample size for categorical data (see more (Bartlett et al., 2001)) as follows:

\[
n_0 = \frac{Z^2 \ast p \ast (1 - p)}{d^2}
\]

\[
N_s = \frac{n_0}{1 + (n_0/N_p)}
\]

where $Z$: Z value (here: 1.96 for 95% confidence level), $p$: percentage of expected error (here: 2.5%), $d$: acceptable margin of error of the estimated proportion (here: 1.1%), $N_p$: population size (here: 10,520) and $N_s$: sample size.

The proportions of reported road types (roundabout, slip road and main carriageway) that appeared in the sample were similar to those of the entire database as it can be seen in Table 2.

Apart from the estimation of the accuracy of the FLCM method it is useful to examine the results of the proposed method to the results of existing crash mapping methods that are applicable to the study area. The three methods that fulfilled this criterion were:

- Crash Mapping Method 1 (CM1) – based on Levine et al. (1995): Crashes are snapped to the closest segments of the network.
- Crash Mapping Method 2 (CM2) – based on Loo (2006): Crashes are snapped to the closest segments of the network that has the same road names and types as the reported crash.
- Crash Mapping Method 3 (CM3) – based on Wang et al. (2009):
Crashes are snapped to the segment that has the highest Weighting Score \((W_{Si})\) that is calculated using Eq. (1), where \(d_i\) is the distance of the reports crash location from the road segment \(D\) and \(D_{h_i}\) the angular difference between the intended vehicle direction and the link’s direction \(\Delta \theta_i\).

6. Results

FLCM was implemented to the 10,520 crashes discussed earlier. It has been found that the time required to process 10,520 crashes is 230 min (by using a laptop PC with 4 GB RAM and 3.4 GHz processing speed). This suggests that the developed method can process 46 crashes per minute. The accuracy of the four algorithms (CM1, CM2, CM3, FLCM) was then evaluated using the reference 716 crashes discussed in the previous section; the accuracy that was estimated for each algorithm was the percentage of reference cases that were assigned to the same road segment with the one that was selected manually. The percentage of accuracy was estimated at the 95% confidence level with a confidence interval of ±1.1%. All the results can be found at Table 3. The total accuracy levels were found to be 81.6%, 87.7%, 85.0% and 98.9% for the CM1, CM2, CM3 and FLCM algorithms respectively. The percentages of accuracy for the three main reported road types (roundabouts, slip roads and main carriageways respectively) were also estimated in order to identify possible weaknesses of the method in the identification of specific categories of road crashes. It was revealed that CM1 and CM3 face particular difficulty in identifying the correct location for crashes that occurred on roundabouts and that CM2 has the same problem for main carriageway crashes. In contrast, FLCM gives accurate results for roundabout and slip road crashes whereas, for main carriageway crashes the
results are less accurate. The mismatches on main carriageways are mainly due to the errors of the reported road name and type of the crash database. The mean Distance for each of the methods was calculated and found to be: 4.43 m for CM1, 6.47 m for CM2, 14.54 m for CM3 and 8.53 m for FLCM. It is clear that the FLCM method with error just above to 1% gives the most reliable matching results among the examined methods. An interesting outcome is that the correct segment is not always the closest to the reported crash location that is in line with the conclusion of Loo (2009). An additional result to that was that even the closest segment that has the same road name and road type with the examined crash can be erroneous. This highlights the importance of the intended vehicle direction as a variable for a crash mapping algorithm. However, from the results of the CM3 method – that considers the vehicle’s intended direction – it is revealed that the inclusion of this variable in an inflexible formula does not guarantee the accuracy of the results. This supports the selection of fuzzy inference systems for crash location identification that provide a flexible framework that adapts to the reported data of each case individually. For the FLCM method, the 99th percentile of the Distance was found to be 56.8 m and the 98.5th percentile 49.9 m confirming the validity of the selection of the 50 m threshold boundary. In other words, the fact that 98.5% of the cases have Distance less than 49.9 m justifies the need for a manual check in order to confirm the accuracy of the segment selection (as it is described at the Additional Steps 1 and 3 at page 14) when the Distance from the selected segment is over 50 m.

From the 10,520 cases that were matched with the FLCM method, there were 266 (2.5%) that needed manual check according to the additional steps 1 and 2. 14 of them were crashes that could not be matched with any segment of the database due to simultaneous road name and road type mismatch with all the potential candidate segments. From the 266 cases that were checked manually, 107 (1%) needed manual correction. From the entire database, there were overall 206 (2%) cases of crashes that were matched to road segments that had different road names than the reported, and 557 (5.3%) segments that had different road types than reported. After the 107 manual corrections, there were 36 (0.3%) cases where either the road name or type of the selected road segment was the same with those referred on the crash report. This situation indicates the existence of some inconsistency between the network and the crash database that is mostly responsible for the estimated error of the developed method.

Figs. 9 and 10 represent graphically the crash locations when they are superimposed on the digital road network; before crash-mapping (a) and after the implementation of the CM1(b), CM2 (c), CM3(d) and FLCM(e) algorithms respectively. It can be easily noticed that the majority of the reported crash locations (Figs. 9a and 10a) do not fall exactly onto a road segment and some of them are placed between two or more road sections. The crash locations indicated by the four methods (Figs. 9b–e and 10b–e) have both similarities and differences. The locations of CM1 and CM2 (Figs. 9b, c and 10b, c) are very similar to each other, as it was expected, but they are quite different from the FLCM (Figs. 9e and 10e) locations as the crashes are not necessarily assigned to the closest road segment, but the one that has the highest Matching Score. CM3 method locations are almost identical to those indicated by FLCM on the main carriageway crashes (Fig. 9d) however the locations for the roundabout crashes (Fig. 10d) are very different as CM3 was found to have only 52.0% of accuracy for roundabouts.

![Fig. 9. Crash locations at a segment of the M4 motorway according to: (a) STATS19 crash reports (unmapped), (b) CM1, (c) CM2, (d) CM3 and (e) FLCM.](image-url)
7. Conclusions and future work

Crash databases rarely include accurate information concerning crash locations. However, the quality of spatial data is important for studies that focus on crashes and their relationships with geometric or road traffic characteristics. Consequently, refined crash location data can potentially contribute to the quality and reliability of the results of studies such as segment-based crash modelling, risk mapping and identification of hazardous road segments within a road network. Prior crash mapping methodologies mainly focus on the specific characteristics of their study areas and data and are characterised by limited transferability, some methodological shortcomings and in some cases by a high percentage of error.

The crash mapping method developed in this paper (i.e. FLCM) is a generic, high accuracy, automatic method of crash location identification that demands limited manual intervention. The method is based on a transformed map-matching candidate segment selection technique and a fuzzy inference system that was formed in respect to the characteristics of the study area (the Strategic Road Network of the UK) and the limitations of the crash database information (STATS19 reports). Although the presented thresholds were selected specifically for these crash and network databases, the overall method is applicable to any other road network after the determination of appropriate threshold values without any further
modifications providing that the two main components, location coordinates and vehicle intended direction, are available. Moreover, this method has the potential to be implemented in a GPS framework so as to obtain real-time crash coordinates. The results after the implementation of this method to all the reported crashes that occurred during 2012 within the study area showed 98.9% accurate road section selection (confidence level 95%, confidence interval ±1.1%). From the comparison of this method with three other prior crash mapping methods (CM1, CM2, CM3) it was shown that the closest road segment (independently of whether it has the same road name and type with the reported crash) is not always the correct one, highlighting that crash mapping is a much more challenging process than it probably seems and that successful crash mapping demands more input variables than the reported crash location coordinates and reported road name and type. The research presented in this paper is significant as the improved crash locations identified by the algorithm were used by the UK Highways Agency for risk mapping (in which crash risk was estimated from the observed crash data normalised by traffic flow) across their entire road network.

Although there are indications for the importance of crash locations on road safety analyses this statement is not clearly justified yet and in fact most of research does not take into consideration precise crash locations. As a consequence, the next step we consider as appropriate for this study is the examination of the impact of accurate crash locations on crash analyses such as network-level crash prediction models aiming to directly access the usefulness of data-refining methods like the one we propose here.

Acknowledgements

This paper is based on a component of a project commissioned by the UK Highways Agency. The authors would like to express their special thanks to Stuart Lovatt and Louisa Cliffe (both from the Highways Agency) and Dr. Ramesh Perera and Chris Page (both from URS) for their support during the course of this work. The opinions in this paper are those of the authors and do not necessarily reflect those of the UK Highways Agency or URS. The authors remain solely responsible for any errors or omissions. We would also like to thank Prof. Lin Zhao and his research team at the College of Automation at Harbin Engineering University, China for hosting the lead author as an exchange student for a month during which she coded a part of the crash mapping algorithm in Matlab.

References