Multilevel Logistic Regression Modelling for Crash Mapping in Metropolitan Areas

Maria-Ioanna M. Imprialou
PhD Student
School of Civil and Building Engineering
Loughborough University
Loughborough LE11 3TU
United Kingdom
Tel: +44(0)1509 228545
E-mail : M.Imprialou@lboro.ac.uk

Professor Mohammed Quddus*
Professor of Intelligent Transport Systems
School of Civil and Building Engineering
Loughborough University
Loughborough LE11 3TU
United Kingdom
Tel: +44(0)1509 228545
E-mail : M.A.Quddus@lboro.ac.uk

Dr David E. Pitfield
Senior Lecturer
School of Civil and Building Engineering
Loughborough University
Loughborough LE11 3TU
United Kingdom
Tel: +44(0)1509223416
E-mail : D.E.Pitfield@lboro.ac.uk

* *Corresponding author*

Word Count: 6,222 words text +3 Figures*250+2 Tables*250= 7,472
Paper presented in the Transportation Research Board (TRB) 94rd Annual Meeting in Washington D.C., USA

Re-submission date: *1st of March 2015*

ABSTRACT

The spatial nature of traffic crashes makes crash locations one of the most important and informative attributes of crash databases. It is however very likely that recorded crash locations in terms of easting and northing coordinates, distances from junctions, addresses, road names and types are inaccurately reported. Improving the quality of crash locations therefore has the potential to enhance the accuracy of many spatial crash analyses. The determination of correct crash locations usually requires a combination of crash and network attributes with suitable crash mapping methods. Urban road networks are more sensitive to erroneous matches due to high road density and inherent complexity. This paper presents a novel crash mapping method suitable for urban and metropolitan areas that matched all the crashes that occurred in London from 2010-2012. The method is based on a hierarchical data structure of crashes (i.e. candidate road links are nested within vehicles and vehicles nested within crashes) and employs a multilevel logistic regression model to estimate the probability distribution of mapping a crash onto a set of candidate road links. The road link with the highest probability is considered to be the correct segment for mapping the crash. This is based on the two primary variables: (a) the distance between the crash location and a candidate segment and (b) the difference between the vehicle direction just before the collision and the link direction. Despite the fact that road names were not considered due to limited availability of this variable in the applied crash database, the developed method provides a 97.1% (±1%) accurate matches (N=1,000). The method was compared with two simpler, non-probabilistic crash mapping algorithms and the results were used to demonstrate the effect of crash location data quality on a crash risk analysis.

Keywords: crash location, crash mapping, metropolitan/urban networks, multilevel logistic regression

INTRODUCTION

Sustainable road safety programmes require constant enhancements to crash prevention policies and countermeasures. The precautionary measures should focus on the network areas where many crashes occur and the driving attitudes that are considered mostly responsible for crashes. Analysis of crash data aims to identify and explain the factors that lead to traffic crashes. The quality and reliability of the crash data that are used as an input for spatial crash analyses (e.g. identification of black spots, spatial crash modelling etc.) is closely related to the validity of their outcomes *(e.g.1–5)*. The spatial nature of crashes makes crash locations as one of the most important and informative attributes of crash databases *(5–7)* that at the same time are very likely to be inaccurately reported *(1–4, 7)*. Therefore, the refinement of the crash locations gives to the crash analyses the potential to improve in quality.

Traffic crashes create major problems to society as they are related to personal injuries, property damage and other disruptions such as traffic congestion and delays. Crashes in urban networks are more likely to involve unprotected road users such as pedestrians and pedal cyclists who are more vulnerable to serious or fatal injuries. Metropolitan areas are the ultimate form of the urban environment and they are characterised by high network density, large number of trips per day and road users with non-uniform driving attitudes who may also be unfamiliar with the environment (i.e. tourists, business visitors etc.). In these chaotic traffic conditions the explanation of the causal factors of crashes might be challenging; and that is why crash analyses of these areas are of particular interest. Crash locations are rarely reported very precisely as they are collected, for administrative rather than scientific purposes, by police officers who may not arrive immediately at the crash scene *(2)*. The complexity and density of the urban road network is likely to increase the chance of misreporting.

This paper presents a novel crash mapping method suitable for GIS-based crash data that is designed for London, one of the 20 largest metropolitan areas worldwide and the largest in Europe. The main challenges of this work are the high density and the complexity of the road network and the missing values for one of the most important information for identifying crash locations, the road names on which crashes occurred. The crash mapping algorithm developed in this work is based on a multilevel logistic regression model that employs the distance and direction differences as explanatory variables and it is the first wide-area network-level crash mapping attempt that does not include road names as supporting information.

This paper consists of: a review of the existing literature for crash mapping methodologies, a detailed description of the available data, a step-by- step explanation of the developed crash mapping method and the results of its implementation. Finally the main conclusions of this study are drawn.

LITERATURE REVIEW

Although accurate crash locations can be particularly useful for crash analyses the majority of existing studies proceed to the analysis of crash location data without reporting any sort of prior processing (e.g. *(6)*). The developed crash mapping approaches are closely related to the type of their locational input data. Crash locations are reported either using linear referencing, offset from intersections, addresses, or GIS coordinates.

The linear referencing method is a straightforward and relatively accurate method of crash location reporting. Studies that have available the indication of the closest milepost to the crash define this point of the network as the crash location *(8, 9)*. Although this approach demands minor processing, it is reporting mistakes insensitive. Moreover, its location error is equal to the half of the interval between two mileposts that can be as long as half mile.

Additionally, linear referencing can be used only on numbered roads and so it may not be applicable to urban networks.

In order to convert crash locations that are reported using an offset from junctions to coordinates it is needed to combine the attributes of the crash and the network data. Qin et al. *(10)* and Dutta et al. *(11)* developed two algorithms for identifying crash locations in urban and sub-urban environment using "On-At tables" that demonstrate all the roads and their directions at each intersection. These algorithms are strongly dependent on the accuracy of the information included in the "On-At" tables and cannot be applied when junction information is missing. Qin et al. *(10)* report that overall 83% correct matches are achieved but this percentage is higher (89.7%) for local roads.

Address is an easily obtained spatial variable and consequently can be used for identification of crash locations. Burns et al. *(12)* tested some of the online geocoding APIs (Application Programming Interfaces) in terms of their capability to identify crash locations, when given the reported addresses. Google Maps API was found to give the highest matching rate (78%) although the accuracy of matching could not be quantified due to the ambiguity of the cases where the addresses include spelling or other mistakes. Another crash mapping method that is indirectly related to addresses is the method developed by Tarko et al. that attempted to link crash and network records *(3)*. The method was theoretically founded on probabilistic linking techniques used for matching hospital data and although it succeeded in matching all crashes with the correct roads, in some cases it matched one crash with multiple roads.

Crash locations that are reported with GIS coordinates when superimposed on a map of the road network rarely fall exactly on roads. In order to correct the locations additional crash related information should be employed. Some of the GIS-based studies use straightforward but simplistic approaches such as selection of the closest junction *(13)* or closest road section with road name filtering, *(2)* that may be effective for large datasets but not very precise. A similar approach includes the use of restrictive, pre-defined buffer zones along with some descriptive variables such as road name, class, speed limit and junction details *(1)*. Although these approaches have the benefit of a simple theoretic background and are easier to implement, are shown to produce significantly less accurate results than methods that use the vehicle directions *(7)*.

The use of the direction difference between the intended direction of the involved vehicles and a road segment was introduced by Wang et al. *(14)* who used a maximum weighted score to combine the distance and the angular difference to identify crash locations on the M25 motorway in the UK. A weighted score, although it gives slightly improved results than the simpler approaches (as estimated by Imprialou et al. *(7)*), is not suitable for dense and complex networks due to its strong dependency on vehicle direction. Direction difference was later used for two Artificial Intelligence-based crash mapping concepts *(4, 7)*. Deka & Quddus *(4)* developed an artificial neural network for matching crashes within the entire primary road network of the UK that considered the distance, vehicle direction, and the reported road name and type (accuracy level: 98.4%). Imprialou et al. *(7)* employed an empirically set fuzzy-logic inference system based on distance and direction combined with road name and type filters (accuracy level: 98.9%). One of the main shortcomings of these three methods is the expression of the direction of a crash by a single measure (i.e. the average of all the intended directions of all the involved vehicles) that can result in major information losses if the examined crashes include multiple vehicles.

In summary, although existing literature includes a variety of different crash mapping approaches, it seems that most of them are not applicable to dense urban networks. Moreover, none of the developed methods enables the identification of road crashes without taking into consideration the reported road name for each crash. Therefore, a new method should be developed that will be able to handle the complexity and the special characteristics of an urban road network such as inaccurate/missing road names in the crash records.

DATA DESCRIPTION

The data that are used for the implementation and validation of this crash mapping algorithm are extracted from the Integrated Transport Network of Great Britain (ITN) and the National Road Accident Database of the United Kingdom (termed as STATS 19).

a) Network Database-ITN

ITN is one of the layers of the digital Ordnance Survey MasterMap and represents in detail the centrelines and the direction of the entire road network of Great Britain with a system of links and nodes *(15)*. The part of the network that is considered for this study includes the City of London and Greater London (see FIGURE 1). The road network of this area is particularly dense and includes a variety of different road types; from motorways and A-roads to minor and private roads. The attributes of the network that are considered are:

- *Link Reference Number.*
- *Coordinates and Reference Number of start node*
- *Coordinates and Reference Number of end node*
- *Road name*
- *Road type*
- *Link length*

b) Crash Database-STATS 19

STATS 19 is the official Police crash database in the UK and consists of all injury crashes *(16)*. The database includes a number of different variables that describe the crash location, involved vehicles and casualties*(17)*. The main variables that may be of interest for this study are:

- *Accident Reference Number;*
- *Reported Location:* Easting and Northing obtained by the Ordnance Survey Grid map;
- *Vehicle movement compass point:* The intended direction of every involved vehicle reported using the four cardinal points and their intermediates (i.e. N, NE, E etc);
- *Road type*;
- *Junction Detail:* Type of junction (if any) within a radius of 20 meters from the crash location;
- *Road name;*
- *First point of Impact:* Each involved vehicle’s first point to come into contact with another vehicle (i.e. front, back, offside, nearside);

The crash database for this study consists of all the 72,710 reported crashes that occurred from 2010 to 2012 in the City of London and Greater London. TABLE 1 includes some descriptive information about the examined crashes. Not surprisingly for a dense urban network, over three quarters of the crashes are located less than 20 metres from junctions that are the most challenging road configuration in terms of crash mapping. Moreover, more than one quarter of the crashes occurred on minor urban roads where the road name is not reported. This is a significant limitation because road name has been proved to be a useful variable in previous crash mapping techniques and indicates that an alternative approach is required.

The predictive multilevel model that is described in the next section is built and validated using two independent representative samples (i.e. training and reference datasets respectively) extracted from the full crash dataset with the method of quota sampling. The training dataset consists of 700 crashes equally distributed all around the examined network in London during 2011. The reference dataset includes 1,000 crashes that occurred in 2012 (TABLE 1). All the crashes of these two sub-datasets were manually matched (vehicle by vehicle) to one of the candidate road links employing a range of quantitative and qualitative variables. The selection of the candidate links is based on the reported location, direction, road type, junction detail, first point of impact and road name (where available). In the absence of real reference data showing the road segments where crashes actually occurred, the manual crash mapping, although it is time inefficient, is considered to be the most reliable method *(12)*. Thus, the comparison of the results of the proposed method to the reference dataset is considered to capture the actual accuracy level.

METHOD: CRASH MAPPING USING MULTILEVEL LOGISTIC REGRESSION (CM-MLOGIT)

To achieve the maximum possible level of accuracy a crash mapping method should be designed to fit the special characteristics of the network and the crash data. As mentioned above, the examined road network is considerably dense and complex as it includes multiple different road types and classifications (from minor private roads to motorways). Moreover, the names of the roads where the crashes occurred are not always available (i.e. not available for all minor-road crashes that account for over the 26% of the crash database); increasing the difficulty of the task. In order to overcome the problems due to the above characteristics it is necessary to use a measure that in addition to the distance of a reported crash location to a road link (henceforth: *distance*) will enable the discrimination of a relevant to the crash road link from the non-relevant. The intended direction of each involved vehicle has been shown to have these qualities *(4, 7)* and consequently it is selected for this algorithm. In contrast to the previous approaches that used one single value of intended direction per crash (i.e. the average of the directions of all the involved vehicles), this algorithm considers a disaggregated measure of direction; the intended direction of each of the involved vehicle is considered separately so as to decrease the chance of mismatches.

The foundation of this approach lays on the examination of the crash moments before its occurrence. A crash is an unintentional meeting (collision) between a moving vehicle with other vehicles, road users or other obstacles and the crash location is the point of the network where this impact occurred. This point can be considered as the meeting point of the trajectories of the involved vehicles. Consequently, in order to find this unique crash location, firstly the road link, where each of the involved vehicles was travelling on just before the crash, should be found. This approach enables the identification of the locations of multiple-vehicle crashes that occurred on complex junctions without affecting negatively the mapping of simpler cases such as single vehicle crashes or crashes that included vehicles that were moving to the same direction. This section discusses the crash mapping process that is divided into three main procedures: a) Candidate road links identification, b) Matching probability estimation of candidate links and finally c) Identification of crash location.

a) Candidate road links identification

As the examined network covers a wide area and consists of over 300,000 road links the minimisation of the examined road links per crash is necessary for developing an efficient crash mapping algorithm. Thus, each crash is attempted to be matched with road links that fall into the area that is defined by an error circle with its centre at the reported crash location and a predefined radius (henceforth: *candidate links*). This method, that is extensively discussed in Imprialou et al. *(7)*, is an efficient way to narrow down the number of the examined road links without compromising the accuracy of the results.

The radius of the error circle is crucial for the successful implementation of this method as it should be sufficiently large for capturing the road links that are the most likely to be correct and simultaneously exclude most of the unnecessary links. The optimal value for the radius of the error circle was determined by the 99th percentile of the distance between the reported crash location and the manually selected road link in the training dataset. The 99th percentile of the distances is 27.46 m and the radius of the error circle was rounded up to 30 m. In order to identify candidate links for the extreme cases that did not include links within this area, the radius is expandable up to 230m with a step of 50m in each iteration.

Another measure to decrease the number of candidate links is the filtering of the links by road type. For crashes that are reported to occur on roundabouts or slip roads the final candidate links set consists of only roundabout or slip road links unless these road types do not appear in the error circle. In that case, all the links within the error circle are considered as candidate.

b) Matching probability estimation

As discussed, the primary goal is to identify the correct link on which a crash occurs. After identifying the set of candidate links, each link is assigned with a probability of being the link where each of the involved vehicles was travelling on just before the collision. At this stage both the distance (i.e. minimum distance from the reported crash location to a link) and the angular difference are considered. While the distance of a specific candidate link is the same for all the involved vehicles of a crash, the angular difference may differ. Thus, all the candidate links that were identified at the previous stage should be evaluated for their likelihood to be the matching links for each participating vehicle separately. In this way, crash data can be seen as a nested structure where candidate links are nested within vehicles and vehicles are then nested within crashes resulting in a three-level hierarchical dataset. FIGURE 2 represents graphically this hierarchical structure; every crash includes I (I $\geqslant$1) involved vehicles and each of these vehicles can be matched to J (J $\geqslant$1) candidate links.

The estimation of the probability of each link to be the correct link is based on a predictive three-level logistic regression model. Logistic regression is a probabilistic approach for modelling a binary response variable (here: a vehicle can be matched or cannot be matched to a link) in order to address a classification problem. The use of probability in classification problems is useful and logical when other labelling rules seem crude or ineffective due to the lack of a general criterion of classification *(18, 19)*. One of the main assumptions of traditional logistic regression is the independence of observations. Violation of this assumption may lead to biased standard errors, confidence intervals and significance tests that can result in erroneous conclusions*(20)*. In the examined case the candidate links cannot be considered as independent observations because the outcome of an observation can affect the outcome for other observations of the same candidate set. In fact, the structure of the crash data as it is aforementioned can be seen as hierarchical. Consequently, in order to avoid biases due to the independency assumption violation it is necessary to apply a hierarchical logistic regression model *(18, 21)*.

The general expression for a three-level logistic regression model is:

$$logit(p_{ijk}) = \log[p_{ijk}/(1-p_{ijk})] = \beta_0 + \boldsymbol{\beta_m x_{mijk}} + u_{0k} + u_{0jk} \qquad (1)$$

Where:

i, j and k: indices of Levels 1, 2 and 3 respectively;
y_{ij}: binary response variable;
$p_{ijk} = \Pr(y_{ij} = 1)$: probability the binary response variable belongs to the category represented by one;
$\boldsymbol{x_{mijk}}$: vector of the m^{th} explanatory variables;
β_0: intercept;
$\boldsymbol{\beta_m}$: vector of coefficients of the m$^{\text{th}}$ explanatory variables;
u_{0k}: random intercept of Level 3 ($u_{0k} \sim N(0, \sigma^2_{u_{0k}})$) and
u_{0jk}: random intercept of Level 2 ($u_{0jk} \sim N(0, \sigma^2_{u_{0jk}})$).

The predictive model is built using the training dataset that consists of 700 crashes that included 1,258 vehicles and 6,446 road links in total. The number of involved vehicles per crash varied from one to four and the candidate links per crash were from one to 31. The response variable is *score* that is equal to one (*score*=1) when the examined link is the unique "matching" link or equal to zero (*score*=0) for the rest of the links that are "not matching" with the examined involved vehicle of a crash. The explanatory variables are *distance* (i.e. minimum distance from the reported crash location to a link) and *angular difference* (i.e. minimum angular difference between the vehicle's intended direction and the link). The integration method that was used is mean and variance adaptive Gauss–Hermite quadrature *(21)*. The parameter estimation of the model is summarised in TABLE 2.

The tabulated results show that both the distance and the angular difference are statistically significant to the 95% level (i.e. p-values are less than 0.05) for the estimation of the probability of a candidate link to be the correct. As expected, an increase in both the explanatory variables decreases the probability of the examined link to be matching. More specifically, for every metre increase in the distance between the reported crash location and the examined link, the odds for it to be the matching decrease by 21.9% (i.e. distance odds ratio =exp(β_{Distance}) =0.781) and for every degree the angular distance increases, the odds drop by 4.9% (i.e. angular difference odds ratio =exp($\beta_{\text{Ang Difference}}$) =0.951). Moreover, from the Likelihood Ratio (LR) (i.e. LR test are used to compare the fit between two nested models) test it was found that the current model fits significantly better than the ordinary logistic regression model, justifying the superiority of multilevel modelling for the examined dataset. Equation (2) presents the combination of the expected values of the coefficients and the random effects that was used for prediction (according to the initial assumption random effects' expected values are equal to zero):

$$A = logit(P_{matching}) = 2.6154 - 0.2476\ (Distance) - 0.0505\ (AngDifference) \quad (2)$$

$$Score = P_{matching} = \frac{e^A}{1 + e^A} \quad (3)$$

Equation (3) was used for predicting the probability of each candidate link to be the link where the examined vehicle involved in the crash was travelling on just before the crash. For instance, a candidate link that is located 3m away from the reported crash location and its direction differs 11 degrees from the intended direction of an involved vehicle is 78.9% likely to be the matching link while the probability for another link with 2m distance and 70 degrees angular difference is only 19.6%. This probability (*score*) is used for ranking all the candidate links in terms of their goodness of matching with an involved vehicle and plays a primary role in the final crash location identification that is described in the following section.

Matching Link per Vehicle

The candidate link with the highest probability (i.e. score) is considered to be the correct link on which a vehicle was travelling on just before the crash. This link is termed as the matching link for this specific vehicle.

Although the links that ranked first in terms of score are almost always indeed the matching links, due to minor network digitisation problems there are a few cases where the matching link was found to be neighbour to the link with the highest score. This happens due to inaccuracy in the measurements of crash location and angular difference. In order to prevent this kind of mismatches an empirical rule was set: if the candidate link set includes links with distances smaller than the one of the link with the highest score and with angular differences up to 25 degrees bigger, then the link that has the highest score among them becomes the matching link.

c) Identification of crash location

The last step of this process is the identification of the final location on the selected link representing the most likely point of the network where the first impact occurred. In order to proceed to link-based locational or other statistical analyses of the crash data the crash location should be represented by a unique road link (i.e. *final matching link*) and a set of coordinates on that link representing the estimated point of first impact (i.e. *final crash location*). From the previously described process, every involved vehicle is matched with one road link. In order to select the final matching link of a crash, the matching links of all the involved vehicles must be taken into consideration. For that purpose it is useful to distinguish the two main categories of crashes in terms of the approach of the final selection they need.

1) Crashes including one matching link

This category includes all single-vehicle crashes and the crashes that behave like single-vehicle (i.e. crashes where all the vehicles are heading in the same direction and crashes where the vehicles are matched with the same link). The final matching link for these crashes is the link that was identified at the previous step (i.e. the link with the highest score) and the final crash location is the closest point from the reported crash location to this link.

2) Crashes including multiple matching links

This category includes crashes with vehicles heading in different directions. The majority of these crashes occurred on junctions. The rules of this process are described below:

i) If all the involved vehicles are matched with links that have one node (i.e intersection) in common, then this node is considered as the final crash location and the link that has the smallest distance from the reported crash location is the final matching link.

ii) If the involved vehicles are matched with links that do not all intersect, then the final matching link is the link with the highest score and the final crash location is the closest point from the reported crash location to this link.

RESULTS

Method Evaluation

The matching accuracy of CM-MLOGIT is estimated by comparing its output with the 1,000 manually matched crashes of the reference dataset discussed above. If the identified road link by CM-MLOGIT for each crash is different from the manually selected link then this is considered as a mismatch. The confidence interval of the result (*d*) that is estimated based on categorical data sample equation *(22)* is approximately one percent.

$$d = sqrt(\frac{Zp(1-p)}{N_S}) \tag{4}$$

Where d: acceptable margin of error, Z: Z-value (here: 1.96 for the 95% confidence level), p: percentage of expected error, N_s: sample size (1,000 crashes).

CM-MLOGIT is a method that requires a certain amount of time for training the data and fitting the three-level logit model. In order to test whether CM-MLOGIT could be substituted by non-probabilistic, less time-demanding methods, the reference dataset results were compared to the results of:

a) A minimum distance based algorithm that selects the closest road link to the reported location (DMIN) as suggested by Levine et al. *(13)*;
b) A minimum angular distance based algorithm that selects the link that has the most similar direction with at least one of the involved vehicles (AMIN). This algorithm is tested for the first time in this paper. [1]

The three methods were applied for the full three-year crash database (2010-2012) in order to develop three indicative crash rate maps of the entire study area. The risk level of each road link is quantified using the ratio of total crashes per unit of length. Using crash rates is not the strongest method for risk estimation *(23)* but the scope of this paper is not the identification of high risk areas, but the comparison of crash mapping methods. Consequently, the maps should be seen more as a representation of the differences between the methods than of the actual hazard-proneness of the network links.

Accuracy Estimation

The reference dataset was employed to estimate the accuracy of the three methods discussed above**.** The percentage of accurate matches is 76.6%, 55.1% and 97.1% (±1%) for DMIN, AMIN and CM-MLOGIT respectively. Not surprisingly, the two simplistic methods are significantly less accurate than CM-MLOGIT. The accuracy of DMIN is slightly lower than that of the simple distance-based mapping algorithm (81.6%) for sub-urban and rural roads (7) and the higher density of the examined network is the most probable reason for this difference. A minimum angular difference algorithm has not been tested before and it is found that based on the direction differences alone, only half of the crashes can be matched to the correct links. Despite that, the results of an algorithm that uses an optimal combination of angular difference and distance can be very precise. It is clear that both these variables are very useful for GIS-based crash mapping independently of the density of the network and the availability of other supporting information (i.e. road name). CM-MLOGIT is the first network-level crash mapping algorithm that does not include road name filtering at any of its stages. The accuracy level that is reached, taking into account the absence of a dominant variable for all the existing crash mapping approaches, is satisfying.

The mismatches of CM-MLOGIT are closely related to the complexity and density of the urban network; they are either due to the configuration of some links or the large distance of the reported location from the most probable actual location. More specifically, the complexity of the links' shape cannot be accurately represented using a pair of nodes and in some extreme cases this leads to erroneous estimation of the distance or the angular

[1] The candidate links for both the methods were the same with those of CM-MLOGIT.

difference (or both) and consequently of the matching score. An attempt to address this problem by dividing road links into straight smaller segments indicated by their shape points was not successful though. The candidate links set was replaced by a candidate segments set and the rest of the matching process remained the same. Although the candidate segments capture the real configuration of the road, the accuracy level was 95.6%. This slight increase in mismatches is caused by minor inherent digitisation imperfections of the network map. The accuracy of DMIN and AMIN after the links' segmentation reached 77.4% and 56.3% respectively, but this increase is not significant enough to change the conclusions about the matching capability of either of the algorithms. The second type of mismatched crashes includes cases where the correct link was located further from the reported location than a parallel neighbouring link. This kind of mismatches could be avoided if more variables of the crash reports were included in the algorithm (such as junction detail or first point of impact or road name where available). However, improvement of the overall results is not guaranteed as the crash databases include some erroneously reported information *(1)* and an increase of the number of the considered variables would increase the chance of mismatches due to misreporting.

FIGURE 3 depicts the rate maps that correspond to each of the three crash mapping methods (a, b and c for DMIN, AMIN and CM-MLOGIT respectively). Crash rate (crashes per kilometre) is represented with descending order with the red, orange, blue and green zones. Comparing the three maps it can be seen that the risk pattern for the same area appears to be different depending on which crash mapping method is implemented (e.g. junction A). The most realistic representation belongs to the map produced based on CM-MLOGIT because the crash locations used are by far more precise and the quality of the input is proportional to the quality of the output. Accordingly, this outcome can be extended to other crash analyses; the accuracy of a crash mapping method affects the results of the subsequent crash analyses and consequently their conclusions (e.g. decisions for prevention policies and countermeasures).

CONCLUSION

Allocation of traffic crashes to the precise locations where the first impact occurred is directly related to the accuracy of spatial crash analyses. A review of literature highlighted that crash mapping in metropolitan and urban areas is rather challenging because of high complex land-use patterns and road density of the road network and due to incorrect/missing records of road names on which the crashes occur.

This paper developed a new probabilistic crash mapping algorithm (termed as CM-MLOGIT) that has the capability of mapping traffic crashes in dense urban networks. The two unique features of the algorithm are: (1) separate matching of all individual vehicles involved in a crash for identifying the final matching link for the crash and (2) allowing hierarchical nested structure of data (i.e. links are nested within vehicles where vehicles are nested within crashes) for developing the relationship between distance, angular difference and the goodness of matching (i.e. a three-level logistic regression). When the algorithm applied to 1,000 crashes in London, the accuracy level of CM-MLOGIT was found to be 97.1% revealing that the vehicle by vehicle examination of crashes and the use of an optimal combination of angular difference and distance through the use of a multilevel logistic regression can counterbalance the exclusion of road names in the mapping process. Moreover, CM-MLOGIT outperforms two non-probabilistic crash mapping methods (i.e. DMIN and AMIN) that are based on the minimum distance and angular difference respectively. A preliminary crash risk analysis showed that the examined algorithms provide different results, meaning that the accuracy of identified crash locations can affect the outcomes of spatial crash analyses.

REFERENCES

1. Austin K (1995) The Identification of Mistakes in Road Accident Records: Part 1, Locational Variables. Accident analysis and prevention 27:261–276.

2. Loo BPY (2006) Validating crash locations for quantitative spatial analysis : A GIS-based approach. Accident analysis and prevention 38:879–886.

3. Tarko AP, Thomaz J, Grant D (2009) Probabilistic Determination of Crash Locations in a Road Network with Imperfect Data. Transportation Research Record: Journal of the Transportation Research Board 2102:76–84.

4. Deka L, Quddus M (2014) Network-level accident-mapping: Distance based pattern matching using artificial neural network. Accident; analysis and prevention 65:105–13.

5. Tegge R, Ouyang Y (2009) Correcting erroneous crash locations in transportation safety analysis. Accident analysis and prevention 41:202–209.

6. Koike H, Morimoto A, Hanzawa Y, Shiraishi N (2000) Development of Hazard Map Using GIS to Reduce Traffic Accidents. Computing in Civil Engineering 217–224.

7. Imprialou M-IM, Quddus M, Pitfield DE (2014) High accuracy crash mapping using fuzzy logic. Transportation Research Part C: Emerging Technologies 42:107–120.

8. Geurts K, Wets G, Brijs T, et al. (2006) Ranking and selecting dangerous crash locations: correcting for the number of passengers and Bayesian ranking plots. Journal of safety research 37:83–91.

9. Monsere CM, Bertini RL, Bosa PG (2006) Comparison of Identification and Ranking Methodologies for Speed-related Crash Locations. Oregon Department of Transportation Reseach Unit, Washington DC

10. Qin X, Parker S, Liu Y, et al. (2013) Intelligent geocoding system to locate traffic crashes. Accident analysis and prevention 50:1034–41.

11. Dutta A, Parker S, Qin X, et al. (2007) A System for digitising Winsconsin crash location information. 86th Annual Meeting of the Transportation Research Board

12. Burns S, Mechanics A, Building ME, et al. (2013) An Accessible and Practical Geocoding Method for Traffic Collision Record Mapping : A Quebec Case Study. Transportation Research Record

13. Levine N, Kim KE, Nitz LH (1995) Spatial Analysis of Honolulu Motor Vehicle Crashes: I, Spatial Patterns. Accident analysis and prevention 27:663–674.

14. Wang C, Quddus M a, Ison SG (2009) Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. Accident analysis and prevention 41:798–808.

15. Syrvey O (2009) OS MasterMap Integrated Transport Network Layer User guide. UK, Southampton

16. Department for Transport (2011) STATS 20 - Instructions for the Completion of Road Accident Reports from non-CRASH Sources.

17. Department for Transport (2011) STATS19 road accident injury statistics – report form. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230590/stats19.pdf.

18. Guo, Guang Zhao H (2000) Multilevel Modelling for Binary Data. Annual Review of Sociology 26:441–62.

19. Khan HR, Shaw JEH (2011) Multilevel Logistic Regression Analysis Applied to Binary Contraceptive Prevalence Data. Journal of Data Science 9:93–110.

20. Vanlaar W (2005) Multilevel modeling in traffic safety research: two empirical examples illustrating the consequences of ignoring hierarchies. Traffic injury prevention 6:311–6.

21. Stata (2013) STATA Multilevel mixed-effects reference manual. College Station, Texas, US

22. Bartlett JE, Kotrlik JW, Higgins CC (2001) Organizational Research : Determining Appropriate Sample Size in Survey Research. Information Technology, Learning and Performance Journal 19:43–50.

23. Cheng W, Washington SP (2005) Experimental evaluation of hotspot identification methods. Accident; analysis and prevention 37:870–81.

LIST OF TABLES

LIST OF FIGURES

TABLE 1: Summary of crash reports 2010-2012 and the training and reference sub-datasets.

		Year			Sub-Datasets	
		2010	**2011**	**2012**	**Training**	**Reference**
Total crashes		24,145	24,468	24,097	700	1000
Percentage (%)	Fatalities	0.51	0.63	0.56	1.00	0.70
	Serious injuries	10.88	10.42	11.59	10.00	11.80
	Slight injuries	88.61	88.94	87.85	89.00	87.50
	Roundabouts	4.35	3.96	3.98	4.57	4.20
	Slip Roads	0.62	0.58	0.59	0.86	0.80
	Junctions	75.44	76.81	77.79	77.14	77.30
	Missing Road Name	27.23	26.14	26.39	25.00	27.40
Min number of vehicles		1	1	1	1	1
Max number of vehicles		8	11	8	4	5

TABLE 2: Estimated parameters, standard errors and confidence intervals of the hierarchical logit model.

	Coefficient	Std. Error	p-value	95% Confidence intervals	
				min	max
Intercept	2.615373	0.1687828	0.000	2.284565	2.946181
Distance	-0.2476173	0.0136814	0.000	-0.2744324	-0.2208023
Angular Difference	-0.0505259	0.0022673	0.000	-0.0549698	-0.0460819
$\sigma(u_{0k})$	0.5716379	0.182914		0.3053174	1.070263
$\sigma(u_{0jk})$	6.20E-28	9.61E-21		-	-

Log-Likelihood (Three-level Logit Model) = -1759.3957

Log-Likelihood (Simple Logit Model) = -1770.9883

LR=23.1752>$\chi^2_{.01,\,2}$

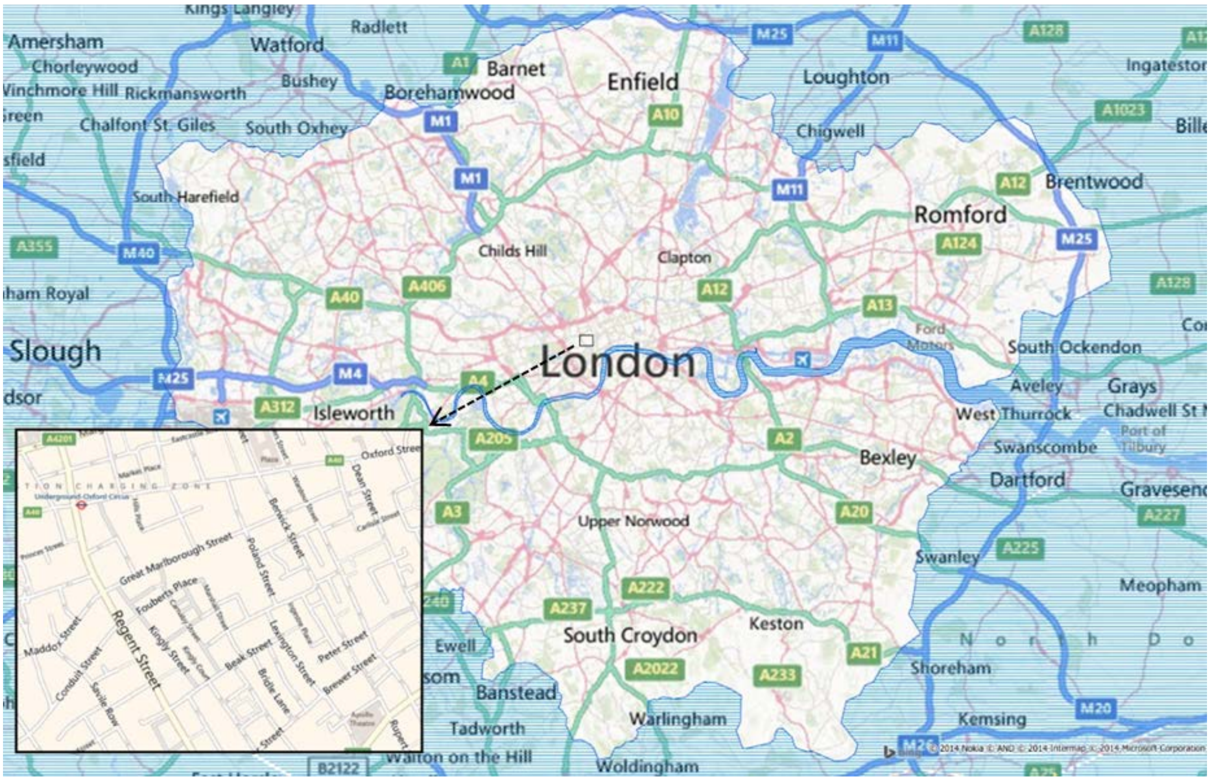

FIGURE 1: City of London and Greater London boundary. (Left: A zoomed detailed portion of the road network showing the inherent complexity in mapping crashes onto correct segments) *Source:* Bing^{TM}Maps

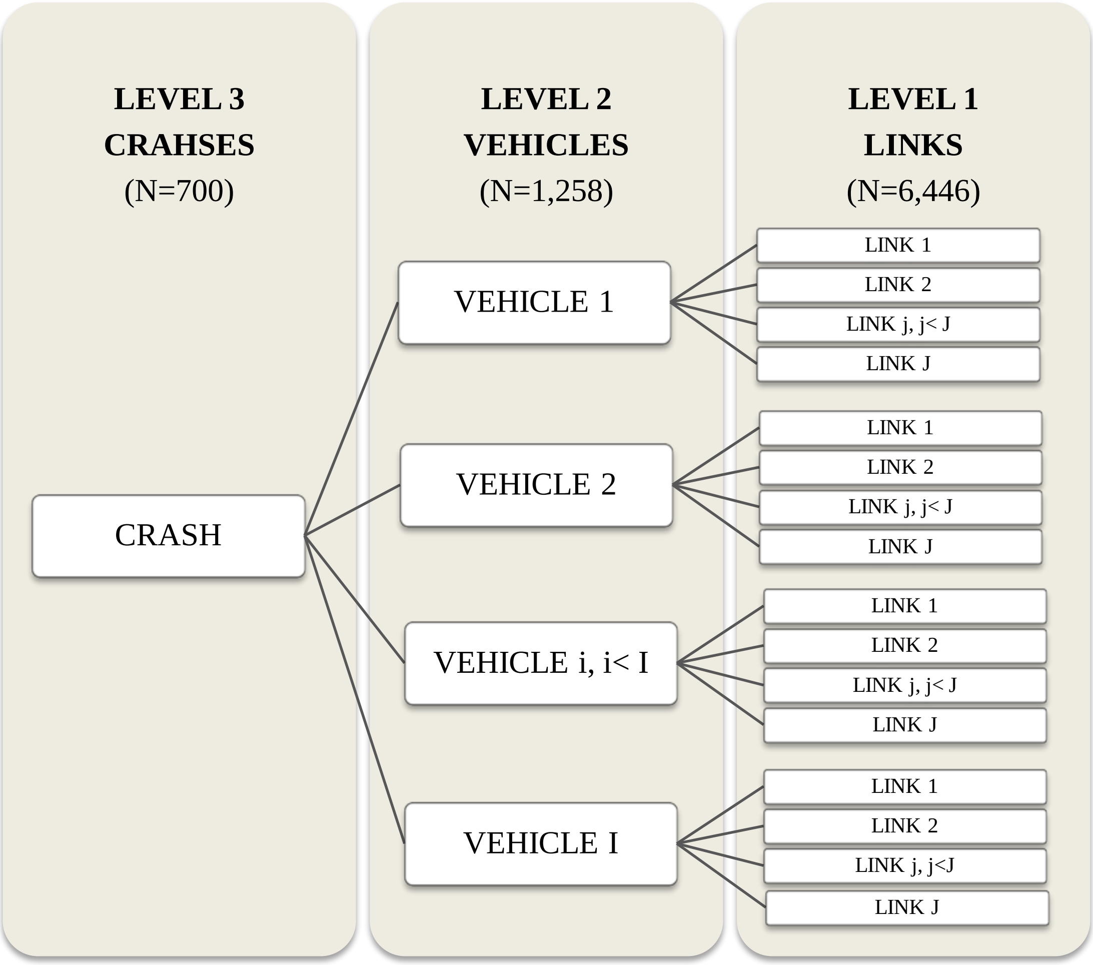

FIGURE 2 : Graphical representation of the hierarchical structure of a crash with I involved vehicles and J candidate links.

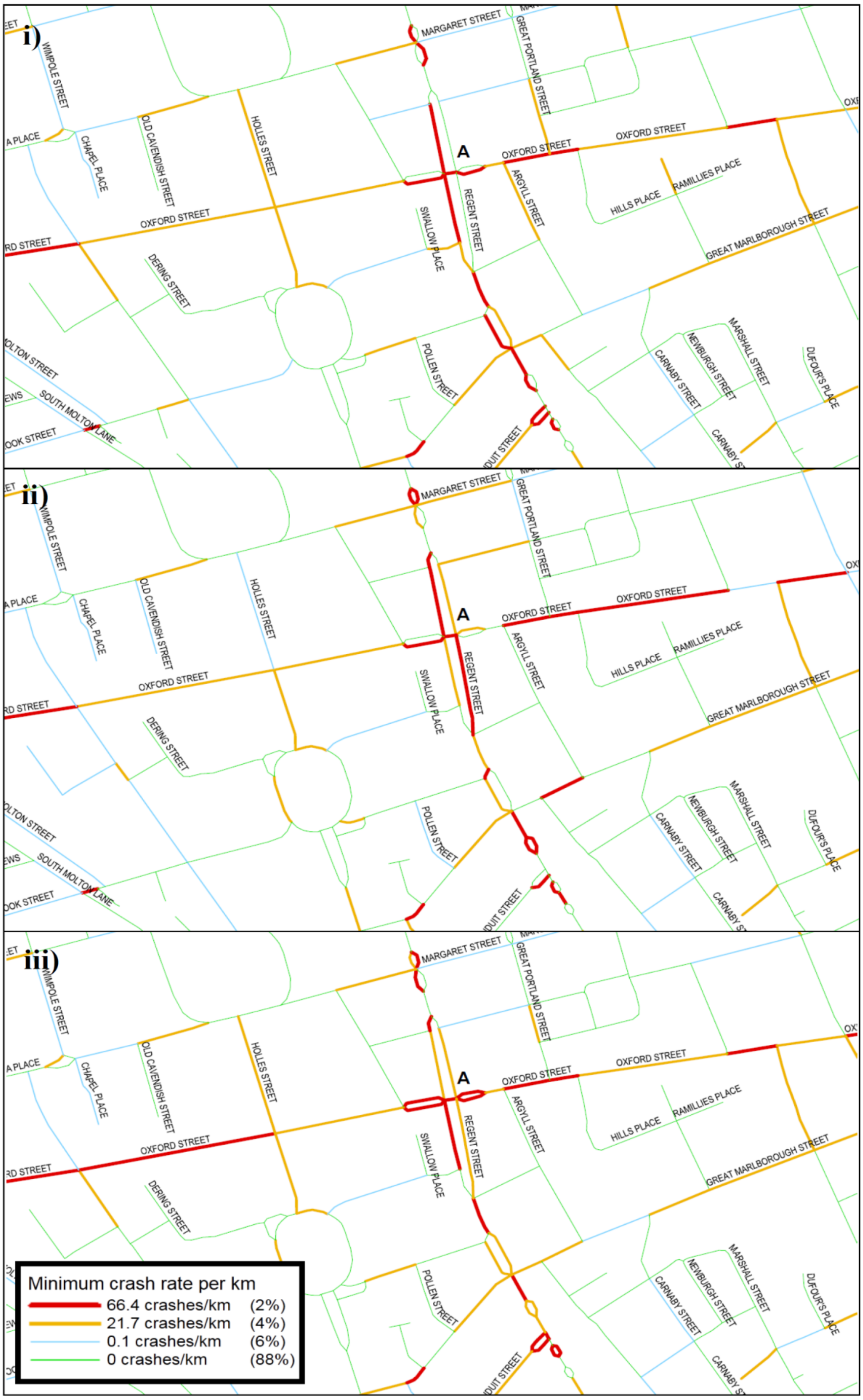

FIGURE 3: Crash rate maps of a part of the study area based on the results of: i) DMIN, ii) AMIN and iii) CM-MLOGIT.