Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

How Accurate are Locations in Malaysian Accident Data? Development of a Rectification Procedure based on Nested Filtered Search Technique

Ashar Ahmed[a,b*], Ahmad Farhan Mohd Sadullah[b], Ahmad Shukri Yahya[b], Mohammad Nishat Akhtar[c], Qummare Azam[c]

[a]*NED University of Engineering and Technology, Department of Urban and Infrastructure Engineering, Karachi 75270, Pakistan*
[b]*School of Civil Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia*
[c]*School of Aerospace Engineering, Universiti Sains Malaysia, 14300 Nibong Tebal, Penang, Malaysia*

Abstract

Errors in the accident data arise due to several individual and institutional shortcomings with inaccurate or ambiguous locations being the most serious amongst them. To overcome this issue a procedure has been devised in this study to rectify accident locations in the Malaysian accident data. The algorithm uses filtered search technique to match the accidents from the accident database with intersections from the field database. A total of 2809 accident records were screened for this purpose and 304 intersections were surveyed. The results showed that the algorithm is able to match up to 60% of the accidents to their respective locations depending upon the quality of the available records. Information related to only five parameters; major road width, landuse, lane marking, traffic control and geometry, are required for its implementation. Unlike geographic coordinates, these parameters are easily measurable and do not require any specialized equipment. The procedure is more relevant to low- and middle-income countries, where landmarks are much commonly used for location identification instead of geographic coordinates.

Keywords: traffic accident; crash location; matching algorithm; intersection; geometry

* Corresponding author. Tel. +92-333-353-9669.
E-mail address: ashar.ue17@gmail.com

1. Introduction

Road crashes are spatial events tied to a specific location(Imprialou, Quddus, and Pitfield 2014, Thill 2000). Errors in the location of accidents arise due to several individual and institutional shortcomings such as problems in data collection, management and processing (Tegge and Ouyang, 2009). Discrepancies in accident location can cause misidentification of black spots and hazardous zones and can make crash reconstruction impossible. They lead to detection of wrong parameters responsible for accident occurrence. These parameters are used in safety performance functions to obtain risk estimates which are then utilized by engineers, transportation planners and law enforcers to improve safety. But the estimates can only be accurate if the values provided belong to the site under investigation. It has been indicated in the literature that the estimated coefficients of a safety model vary considerably when rectified crash locations are used (Tegge and Ouyang 2009). Thus, the accuracy of crash location is extremely important to conduct all kinds of analysis from micro to macro scale.

Most countries rely on police to collect accident data (Miler, Todić, and Ševrović 2016), who have many duties to perform at the accident site and mistakes in recording are inevitable (Austin 1995). As a result location errors in accident data are prevalent in many countries around the world such as USA (Dutta et al. 2007, Qin et al. 2013), UK (Austin 1995, Imprialou, Quddus, and Pitfield 2014, Imprialou, Quddus, and Pitfield 2015), Australia (Howard, Young, and Ellis 1979), Saudi Arabia (Al-Ghamdi 2003), UAE (Hawas, Khan, and Maraqa 2012), Italy (Montella et al. 2013), China (Wang, Zhang, and Mao 2013), Canada (Burns et al. 2014), and Croatia (Miler, Todić, and Ševrović 2016).The extent of error ranges from 7% to 88%.

In the literature relevant to Malaysia, it has been reported that the Police data do not always refer to the exact location. Two studies pertinent to unsignalized intersections reported that the official reports do not contain the exact location, instead they give some indication of a typical pattern about the occurrence of the accident within the vicinity of a particular intersection (Abdul Manan and Várhelyi 2015, Abdul Manan 2014).But, no known initiative has been taken to establish a comprehensive procedure that can be used to rectify it. This argument became the initial motivation to conduct this study.

This paper aims to provide a rectification methodology to remove location errors from the Malaysian accident database. It uses a "Nested Filtered Search Algorithm" to match accidents with their respective locations. The novel features of this procedure are: (a) matching of accidents with crash sites without any prior GIS database backbone, (b) simple method that does not require any GPS-based geocoding which makes it extremely suitable for low and middle-income countries, (c) uses only five easily measurable parameters for location identification, (d) does not require highly skilled manpower for field survey, (e) more 'intelligent' as compared to other techniques available in the literature as it incorporates stepwise decision making process rather than matching coordinates only by using a 'distance measure'. The results indicate that the rectification procedure devised is capable of finding accident locations with accuracy depending upon the quality of raw data.

2. Literature Review

Most accident location rectification methods, matching techniques and validation procedures available in the literature rely on a matching algorithm. These techniques can be broadly classified into three categories. The first category comprises of those techniques which use name of road or other similar information such as road number, link ID, road class, district code, municipality and county for matching accidents to their respective locations. The second category of techniques utilizes GPS-coordinates to match the accidents to their corresponding locations. The third category encompass those techniques which use a combination of both, that is, names as well as coordinates in the matching process. A summary of studies pertinent to each category is presented in Table 1.

The name based techniques match the name of the road mentioned in the accident reporting form with the name of the road on which the accident actually occurred. Several other variables can be used in combination with name of road in the rectification process, depending upon the quality of available data. The difference between the actual and the reported name of roads or reported crash location and its distance from the segment/node on a road network where it has occurred is considered as the extent of error. A match between the information available in the accident report with the information on an official road map for the variables used in the matching process is considered as the output. The measured distance quantifies the accuracy of the matching technique. One of the earliest examples of this technique was the work done by Austin (1995), who matched the location of accidents in the UK accident reporting form, known as STATS 19, with the Ordinance Survey map. The accidents in the Wisconsin crash database were mapped with a higher accuracy by using roadway-route prefix, suffix, and name type along with the road names in the matching process (Dutta et al. 2007). Apart from the name variables, county ID was also used in combination by Tarko,Thomaz and Grant(2009), to assign accidents to their respective locations in Indiana, USA. A more improved methodology was used by Qin et al. (2013) in which the names of both major and intersecting roads/highways were utilized to pinpoint accidents on a state level map. Their Crash-Mapping Automation Tool (C-MAT) also used landuse and land jurisdiction variables.

Commercialization of GPS technologies and their popularity in the first decade of 21st century marked the use of GPS based location coordinates for the identification of accident sites. This gave rise to coordinate based techniques developed by researchers for the validation and rectification of accident location. Missing information in the crash database such as crash coordinates is considered as the extent of error. The crash coordinates reported in the accident form falling within the zone defined for a road segment, node or intersection, are considered as the output. The measured distance between the reported and the matched coordinates quantifies the accuracy of the matching technique. Various methods such as the Lagrangian relaxation-based solution algorithm(Tegge and Ouyang 2009), weighted score estimation using perpendicular distance from the crash coordinates to the centerline of the carriageway and the angular difference between the direction of travel of the vehicle and the road(Wang, Quddus, and Ison 2009), Multilevel Logistic Regression modeling (Imprialou, Quddus, and Pitfield 2015), and application of Vehicle Black Box(Chung and Chang 2015), were used to allocate accidents on a road network.

The techniques which employed names or similar variables along with coordinates in the matching process were similar in nature to coordinate based matching techniques. The difference being that the name of road, highway or district provided an additional check to affirm if the accident belonged to the correct location. In the work done by Loo(2006), road name was used in the third step, after snapping the accidents to the nearest junction/road, to check whether the name existed in the road network database before proceeding further with the matching process. Similarly in the three step methodology developed by Imprialou, Quddus, and Pitfield (2014), road names were used in the first step, along with road type and coordinates, to filter the network segments.Miler, Todić, and Ševrović (2016) also developed a three step methodology in which road names were used in the second step to obtain the 'Jaro-Winkler' distance between the accident record street and the road network street.

In conclusion the techniques available in the literature used a 'distance measure' between the crash report and the crash site coordinates or crash report and the crash site characteristics to estimate their accuracies. However, no method is available that claims a complete match between the crash reported and the site where it occurred. This study fills this research gap by applying a simple procedure which is based upon the geometric and control characteristics of each crash site. A complete match between the reported and the actual crash site is said to be obtained when there exists no other intersection on the given road with the same characteristics as mentioned in the accident report.

Table 1. Accident location rectification/validation methods, description and output

Author (s)	Country	Extent of Error	Matching Technique	Variables Used	Output
Austin (1995)	UK	10%	NB	road class, road number, district, speed limit, pedestrian crossing facilities, junction control, junction detail, and carriageway type and markings	The procedure successfully identified the mistakes in official accident reports
Loo (2006)	China	27.5%	CB + NB	crash coordinates, grid references, road names, district board	The procedure estimated that there were 12.7% mistakes in road names and 9.7% mistakes in districts
Dutta et al. (2007)	USA	40.5%	NB	roadway-route prefix, suffix, name, type, and on-street and at-street names	The algorithm matched 79% of the crashes with 98% accuracy
Tarko, Thomaz and Grant (2009)	USA	-	NB	county ID, road name, road administration and road type	Only 20% of crashes had one road assigned while the rest had multiple roads assigned to one accident
Tegge and Ouyang (2009)	USA	-	CB	crash coordinates	55.6% of the crashes were either located on the borderline of multiple sites or

					do not coincide with the existing road network
Wang, Quddus, and Ison (2009)	UK	-	CB	crash coordinates	Location of 98% of accidents were rectified
Qin et al. (2013)	USA	28.3%	NB	names of major and intersecting highway/street, municipality and county, document ID, landuse, land jurisdiction, distance from the intersection location	10.38% and 23.93% of segment related and 10.17% and 14.06% of intersection related crashes, could not be mapped on the local roads and state routes respectively
Imprialou, Quddus, and Pitfield (2014)	UK	7.3%	CB + NB	crash location coordinates, road name type, and section label, vehicle's direction of travel	The algorithm matched accidents with an accuracy level of 98.9%
Burns et al. (2014)	Canada	50%	CB	crash coordinates	85% of traffic crash records were geo-coded
Chung and Chang (2015)	Korea	-	CB	crash coordinates, time of accident	Spatial difference of 84.5 meters and temporal difference of 29 minutes was found between police and Vehicle Black Box records
Imprialou, Quddus, and Pitfield (2015)	UK	26.6%	CB	crash location coordinates	The algorithm matched accidents with an accuracy level of 97.1%
Miler, Todić, and Ševrović (2016)	Croatia	33.5%	CB + NB	crash coordinates, street names	The algorithm validated over 66% of accidents

NB: Name Based, CB: Coordinate Based

3. Rectification Methodology

3.1 Study data and its limitations

Royal Malaysian Police (WHO 2015) is responsible for accident data collection in Malaysia. It is then transferred to MIROS (Malaysian Institute of Road Safety Research) (Hashim and Rahim 2009)which assist to maintain the crash database of the entire country. Six year crash data (2006-2011) pertinent to unsignalized intersections excluding roundabouts, for the state of Penang was acquired from MIROS for this study. This state is located on the northern west coast of Peninsular Malaysia. It has mixed landuse including residential, commercial, industrial, agricultural and educational. The traffic mix is highly heterogeneous comprising of various kinds of motorized and non-motorized vehicles such as bicycles, motorcycles, cars, vans, buses, small trucks, lorries and truck-trailers. The unique combination of different landuse and vehicles together creates an environment in which all types of accidents are observed. The accident reporting form was developed accordingly. For each crash recorded, there were ninety one attributes for which the data was required to be entered under the POL 27, which is the standard form typically used by the Royal Malaysian Police. A total of 78844 crashes were registered during the six year period for which the data was acquired. The details are shown in Table 2. Only 2809 crashes had complete information on all the variables used in this study. Therefore, datapoints with missing and incomplete information were filtered out to formulate the refined Accident Database (AD). The principal error in the AD was the location where the accident occurred. Among the locational variables only the name of the major road, its lane marking, area type and minor road's traffic control were available. Given the fact that there are several intersections along a particular road, there was an equal chance of an accident to occur at any one of them.

Table 2. Recorded and Missing Data Statistics

Year	Recorded Accidents	Complete Information	Missing %
2011	13663	285	97.91
2010	14866	334	97.75
2009	13988	582	95.84
2008	10167	543	94.66
2007	13167	748	94.32
2006	12993	317	97.56
Total	78844	2809	

The 2809 accidents, for which the information related to locational variables was available, were required to be further processed to segregate the ones which occurred on a particular road. Such screening was necessary because accident analysis requires site with a history of accidents, the common reference being at least four accidents at individual locations (Erdogan et al. 2008) or three or more serious injury accidents within three years (Elvik 2008). This led to the discovery of individual roads on which field survey was performed to gather data related to all intersections lying along their stretch.

3.2 Selection of parameters for location identification

Safety Performance Functions (SPFs) are developed on the basis of number of crashes occurring at a particular intersection. A SPF can't be established if there is no certainty about the crash location. Another limitation of such data is its inability to identify blackspots. Because there is no exact spot to pin the accidents along a road, there is no blackspot to conduct the detailed investigation. To solve the issue, physical parameters of intersections that do not change over the period of time were selected to identify the crash location. These parameters are listed in Table 3. Unlike *weather* such as rainy, cloudy or sunny; *volume* such as low, medium or high and *visibility* such as day-time, night-time, fog or mist; that keep changing and thus, can't be matched with the information recorded at the time of accident, the selected parameters can accurately point out the location of the accident provided that they have not been upgraded or undergone any physical change.

Table 3. List of Parameters

Parameter Name	Values
Road Width	0-9 m, 9.1-15 m, >15 m
Area Type	City, Town, Small Town, Rural
Lane Marking	Single, Double, One way, Divider, U-Turn, No marking
Control Type	Stop line/Stop sign, Yellow box, No control

The width of major road was divided into three categories that are 0 to 9 meters, 9.1 to 15 meters and greater than 15 meters, each representing a combination of one lane per direction, more than one lane in each direction and more than two lanes in each direction respectively. The area type where the intersection lies could be city, town, small town or rural. The major road's lane marking could be single line, double line, divider, U-turn or no marking at all. The traffic control type on the minor road could be stop-sign/stop-line, yellow box or no traffic control at all. Use of traffic control parameter, such as stop-sign, as a measure of intersection safety has been discussed in a previous study (Ahmed, Sadullah, and Yahya 2013). The effect of all the above four attributes on number of accidents have been studied thoroughly in another research (Ahmed, Sadullah, and Yahya 2014). Information on all the intersections for the above four parameters was collected through field survey.

3.3 Field survey

A Field survey of 11 roads was performed during May to July 2013. The map showing their location is presented in Fig. 1. The purpose of conducting this exercise was to collect all possible data pertinent to the geometric and physical attributes so that they can be matched with information given in the accident database. A complete match of attributes between the accident and field data authenticates the occurrence of a particular accident at a particular site. The equipment used were measuring wheel, safety vest, noting pad and pen. A single surveyor performed the survey during 9:00 a.m. to 6:00 p.m. by walking along the road taking readings of road width and noting information about traffic sign, lane marking and area type. The results of the survey are shown in Table 4. The data collected manually was fed into the computer using *MS Excel* to form the Field Database (FD). A total 304 intersections were surveyed out of which 120 were uncontrolled. That is, no sign post or pavement marking was present to control minor road traffic.

Table 4. Name of Road, Length of Road, and Number of Unsignalized Intersections

Name of Road	Length of Road (Km)	Number of Unsignalized Intersections
Jalan Bagan Ajam	0.75	6
Lorong Mak Mandin 5	1.2	6
Jalan Telaga Air	0.8	15
Jalan Sultan Azlan Shah	11.5	76
Jalan Dua Kepala Batas	1.1	11
Jalan Penaga P001	4.3	22
Jalan Jelutong	2.3	40
Jalan Bagan Lalang	2.0	14
Jalan Burma	3.6	40
Jalan Padang Bengali	4.7	24
Jalan Raja Uda	4.1	50
	36.35	304

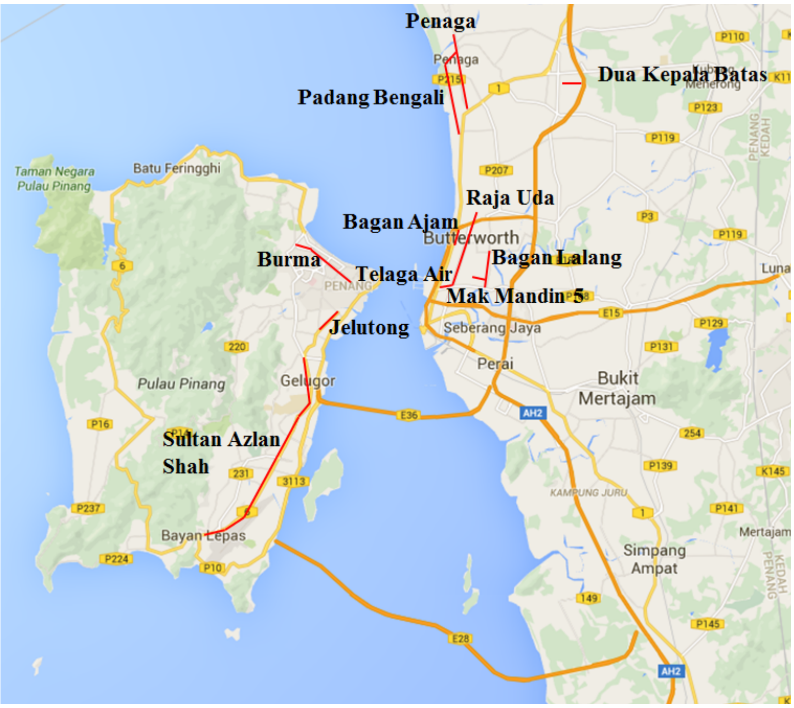


Fig. 1. Map of the 12 Roads Surveyed in the Penang State, Malaysia

3.4 Rectification procedure

The accident database and the field database contained more than one accident occurring and more than one intersection observed on each road respectively. This gave rise to the problem of 'Blind Crash Location'. It refers to the state that the user, is unable to locate which accident occurred at which intersection. Therefore, a computational technique known as "Nested Filtered Search Algorithm" was devised to envision the 'match' between an accident, stored in the accident database, and the intersection, stored in the field database, where it might have occurred thus, forming the rectified database in three steps as shown in Fig. 2. This algorithm screens accidents from accident database and intersections from field database with respect to the first parameter, which road width, listed in Table 3. After initial screening, datapoints are further filtered with respect to the second parameter and so on until each datapoint is sorted to the last parameter selected for analysis respectively. A complete match of parameter between a datapoint from the accident database and a datapoint from the field database authenticates the occurrence of a particular accident at a particular site. Accidents and intersections belonging to the same road are matched with each other. That is, the procedure is required to be run separately for each road.

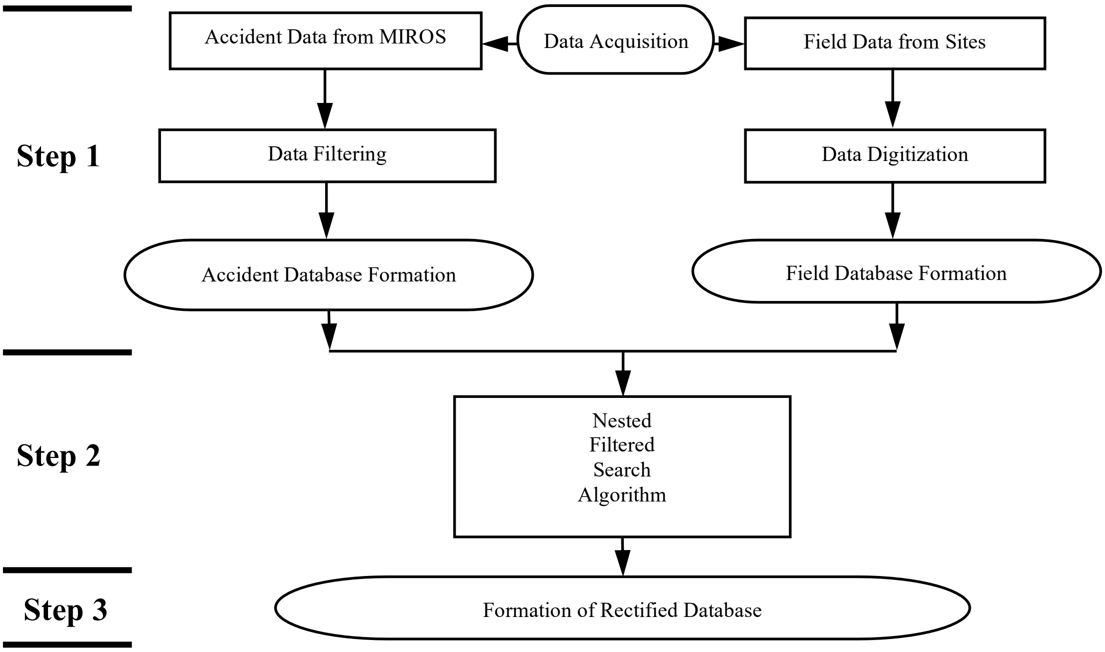


Fig. 2. Procedure of Rectified Database Formation

3.4.1 Algorithm

The search algorithm designed uses nested filters to solve the problem statement. Two files are maintained in order to store object data. The first file maintains a set of query while the second file is the target file. Pre-processing is done on the objects of the first file, so that the one which satisfies a specific match property with the target file can be listed out very effectively. Therefore, the computation of data objects appears as a two-fold step (search data, report data).

The idea of filter nesting searching requires discussion on the feasibility approach. In order to make this search algorithm applicable, it is essential for the problem statement to have an exhaustive enlisting of the object data residing in the query file which matches with the object data residing in the target file. As per literature (Chazelle 1986), if S is a finite set of object data to be filtered, Q be a domain query and P be a predicate defined for every single pair in S x Q, then the equation defined to preprocess could be computed efficiently as shown in Equation (1):

$$g: Q \rightarrow 2^s; [g(q) = \{v \in S | P(v, q) is\ true\}] \quad (1)$$

Here g is a function defined, so that by reporting g it is meant that each object is only once in the set *g (q)*. The Pseudocode for the designed rectification technique is presented in Fig. 3. The code was written in C++ using Microsoft's Visual C++ 2010. The program required input in the form of data files imported from MS Excel for processing and then delivers the output which was exported back to *MS Excel* for further analysis. The input for the query file is the accident data and the input for the target file is the field data. Fig. 3 shows the Pseudocode of the designed algorithm.

```
Query_Match (object data set) {
N: Number of object data rows in query file
arrayA[ ]: Object data set array residing in query file (Road Width)
arrayB[ ]: Object data set array residing in target file (Road Width)
AT  : Area Type parameter
LM  : Lane Marking Parameter
TC  : Traffic control parameter
Count 1: Common Counter for query/target file
for loop (int j=0; j<N; j++)
        if (range division[arrayA] € range division[arrayB])
                Start matching arrayA[AT, LM, TC] && arrayB[AT, LM, TC]
                        if (match) then print hit
                                else print no match
End for loop
}
```

Fig. 3. Pseudocode for the designed rectification technique

For the designed algorithm, *arrayA[ ],* and *arrayB[ ]* comprised of road width data and was considered to be the primary parameter or the root parameter. On the basis of road width data, the area type (AT), lane marking (LM) and traffic control (TC) parameters were filtered out. Each iteration performed by for-loop reduces the object data by one unit. Fig. 4 depicts a tree model for the designed algorithm. The function of range filter is to filter-out the road width data into 3 segments of multiple ranges i.e. *R1, R2* and *R3*. If there is a match between the road width data of the query file and the road width data of the target file, then the corresponding data associated with each road width data (query and target) are compared and the final output is generated.

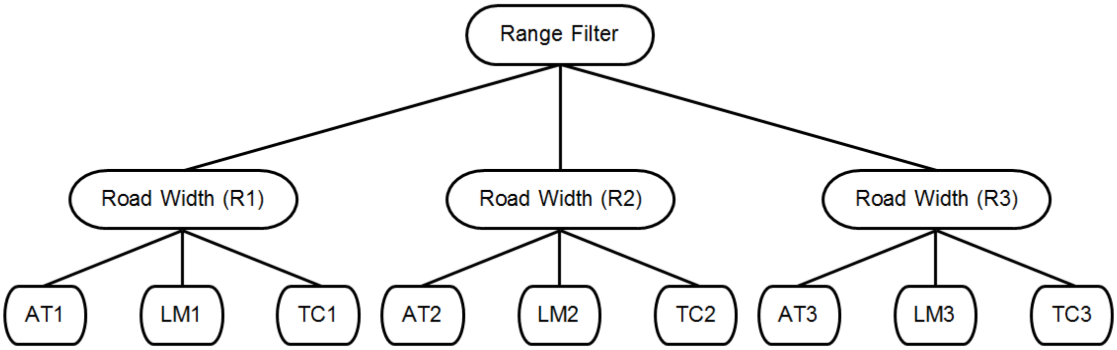


Fig. 4. Tree model of the designed algorithm

It is known that various programs do not execute for a long period of time without repeating instruction. As per the foundation study (Havlak 1997), a loop is considered to be a chunk of code which has got repeated executions without the implementation or repetition of any surrounding code. For the designed search algorithm, series of nested loops are maintained, where master loop (ML) comprises of range filter. The nesting of loops inside master loop with header '*h*' are supposed to be the exterior most loops with respect to the tree model node set (ML-h).

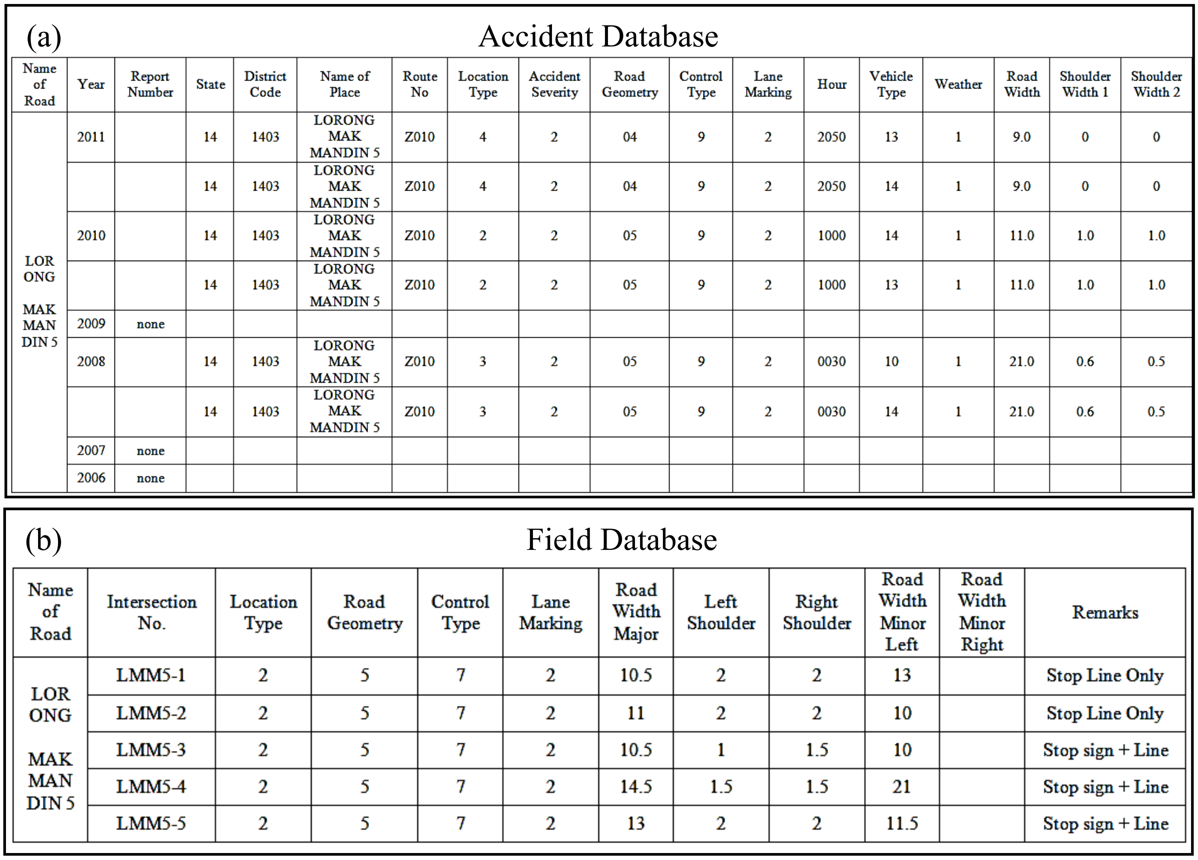
There are three master loops in the algorithm each with a different header '*h*'. For the first master loop, the header '*h*' is road width range 1 (*R1*) which is zero to nine meters. For the second master loop, the header '*h*' is road width range 2 (*R2*) which is 9.1 to 15 meters and for the third master loop, the header '*h*' is road width range 3 (*R3*) which is greater than 15 meters. Moreover, Fig. 4 also demonstrates the loop-nesting relation with a tree where each parent node contains exactly the node of the corresponding loop.

3.4.2 Introduction of intersection geometry as additional filtering parameter

The algorithm matched a group of intersections with a group of accidents as explained earlier. The output achieved the objective, which was, to remove the randomness in the choice of crash location, through identification of the accidents that occurred at a certain set of intersections out of all the intersections that lie on a particular road. In order to further improve the results, a fifth parameter was introduced which was intersection geometry. This significantly improved the matching process and resulted into an absolute one to one match of intersection and accidents. For example, if there were four accidents that were matched with a group of two intersections which included both T-junction as well as cross-junction, then further filtration with respect to number of legs, that is, geometry will result into the identification of a particular accident or multiple accidents that occurred on a particular intersection. This part of the rectification exercise was performed manually which is similar to the methodology used by the transportation agencies of Indiana, USA, where human control is required when the automated process is unable to identify a unique location for a particular accident as mentioned in Tarko, Thomaz and Grant (2009).

4. Rectification Results

The results of the rectification process carried out in this study are divided into two segments. The first segment is the output of the three step rectification procedure and the second segment is the final one to one match achieved after introducing geometry into the rectification process. In the first step of the three step rectification procedure two distinct databases were formed, as shown in Fig. 5. The raw accident data was filtered for missing and fictitious values. As a results an accident database was formed which contained 262 accidents from 11 roads. In parallel the data related to geometric and physical attributes was collected through field survey. They were digitized in a tabulated form to obtain a field database, as shown in Fig. 5.The field database contained information related to 304 intersections from 11 roads. In the second step the Nested Filtered Search Algorithm was run for each road at a time. For example, for the road named "Burma", also called Jalan Burma, the program first asked the user to input the accident data file for the said road. Then the user was asked to input the field data file for the said road. The program reads both data files and after processing it gave output in a tabular format which was then exported to MS Excel for further analysis. Thus, the program was run 11 times to obtain the rectified database.



(a) Accident Database

Name of Road	Year	Report Number	State	District Code	Name of Place	Route No	Location Type	Accident Severity	Road Geometry	Control Type	Lane Marking	Hour	Vehicle Type	Weather	Road Width	Shoulder Width 1	Shoulder Width 2
LOR ONG MAK MAN DIN 5	2011		14	1403	LORONG MAK MANDIN 5	Z010	4	2	04	9	2	2050	13	1	9.0	0	0
			14	1403	LORONG MAK MANDIN 5	Z010	4	2	04	9	2	2050	14	1	9.0	0	0
	2010		14	1403	LORONG MAK MANDIN 5	Z010	2	2	05	9	2	1000	14	1	11.0	1.0	1.0
			14	1403	LORONG MAK MANDIN 5	Z010	2	2	05	9	2	1000	13	1	11.0	1.0	1.0
	2009	none															
	2008		14	1403	LORONG MAK MANDIN 5	Z010	3	2	05	9	2	0030	10	1	21.0	0.6	0.5
			14	1403	LORONG MAK MANDIN 5	Z010	3	2	05	9	2	0030	14	1	21.0	0.6	0.5
	2007	none															
	2006	none															

(b) Field Database

Name of Road	Intersection No.	Location Type	Road Geometry	Control Type	Lane Marking	Road Width Major	Left Shoulder	Right Shoulder	Road Width Minor Left	Road Width Minor Right	Remarks
LOR ONG MAK MAN DIN 5	LMM5-1	2	5	7	2	10.5	2	2	13		Stop Line Only
	LMM5-2	2	5	7	2	11	2	2	10		Stop Line Only
	LMM5-3	2	5	7	2	10.5	1	1.5	10		Stop sign + Line
	LMM5-4	2	5	7	2	14.5	1.5	1.5	21		Stop sign + Line
	LMM5-5	2	5	7	2	13	2	2	11.5		Stop sign + Line

Fig. 5. (a) Sample accident database; (b) Sample field database

The algorithm applies four filters to reach the decision if the accident belonged to the intersection on the given road. The first filter is the road width range, which is 0-9 m, 9.1-15 m, and greater than 15 m. After the application of the range filter, the accidents and intersections were then filtered with regards to area type, lane marking and traffic control respectively. Hence, the final output, which is the rectified database, contain an accident or group of accidents matched to a potential intersection or group of intersections which possess the same area type, lane marking and traffic control falling into a particular road width range. Since the identification process was based on geometric and physical parameters for each data point in both the databases, therefore, percentage of match was dependent upon these parameters correctly measured by the police and noted in the crash record. Incorrect information such as road width greater than or less than the one measured on site, lane marking, area type and traffic control different than the one observed in the field, renders a data point as noise resulting into its screening. This limits the possibility of identification of location of all the crashes occurring on a given road. The algorithm was able to match the location of 55 accidents on their respective roads as shown in Table 5. The output is similar to the result obtained by Tarko, Thomaz and Grant (2009) who found that when accident records are matched with road inventory records, multiple road segments are linked to a single accident.

Table 5. Number of matched accidents by NFSA

Name of Road	Number of Accidents in Accident Database (1)	Number of Matched Accidents (2)	% Match =(Column 2 ÷ Column 1) X 100
Jalan Raja Uda	59	10	16.95
Jalan Bagan Ajam	37	8	21.62
Lorong Mak Mandin 5	6	2	33.33
Jalan Bagan Lalang	36	4	11.11
Jalan Burma	22	4	18.18
Jalan Jelotong	31	5	16.13
Jalan Dua Kepala Batas	16	8	50
Jalan Telaga Air	11	2	18.18
Jalan Padang Bengali	7	2	28.57
Jalan Penaga P001	15	8	60
Jalan Sultan Azlan Shah	22	2	9.09

In order to remove the ambiguity in the location of accidents due to multiple intersections being matched with multiple accidents, an additional filtering parameter was introduced in the rectification process which was intersection geometry. Since the intersection geometry is based on the number of legs, therefore the chance of its wrong identification by the officer is very less. This additional parameter proved to be extremely beneficial in converging the accident or group of accidents, distributed over a group of intersections, to a particular intersection where it/they had occurred. As a result, a total of 24 accidents were identified on 16 intersections. The details of which are presented in Table 6.

Table 6. Site name and the number of accidents matched as a result of the rectification process

Site Name	Site Number	Matched Number of Accidents
Jalan Bagan Lalang Int 2	1	1
Jalan Bagan Ajam	2	4
Jalan Bagan Lalang Int 1	3	1
Jalan Telaga Air	4	1
Jalan Raja Uda Int 2	5	1
Lorong Mk Mandin 5	6	1
Jalan Dua Kepala Batas	7	4
Jalan Penaga Int 2	8	1
Jalan Raja Uda Int 1	9	2
Jalan Penaga Int 1	10	1
Jalan Padang Bengali	11	1
Jalan Jelutong Int 1	12	1
Jalan Raja Uda Int 3	13	1
Jalan Jelutong Int 2	14	2
Jalan Penaga Int 2	15	1
Jalan Burma	16	1

5. Discussion and Conclusion

The typical concept to have two databases, an accident database and a road network database as shown in Fig. 6, matched with each other was common in the methods available in the literature (Deka and Quddus, 2014; Dutta et al., 2007; Imprialou, Quddus, and Pitfield 2014; Qin et al., 2013; Tegge and Ouyang, 2009; Wang, Quddus, and Ison 2009). They worked on the principle of estimating the difference between the locations of accidents mentioned in the accident database with the location on the road network they were reported to have been occurred. A distance measure was used in the estimation process which can not work without the use of coordinates of the accident and its respective roadway facility. Moreover a GIS or road inventory/network map was also a prerequisite in the matching process. In this study all records in the accident data did not have coordinates which constrained the use of existing matching processes. As a consequence the matching algorithm devised and its output, differed from the algorithms and techniques presented in previous researches. One such example is the work done by Wang, Quddus, and Ison (2009) in which a weighted score was calculated with the help of crash location coordinates to estimates the correctness of the segment assigned. They claimed that their technique was able to rectify 98% of accidents. Although an improved technique was utilized by Qin et al. (2013) in which additional variables such as document ID, names of the municipality, landuse and land jurisdiction, were used but their algorithm was able to map only 83% of all crashes. Comparatively the maximum numbers of accidents matched by NFSA was 60%. Given the constraints in the input data, it is still higher than the output of the weighted-based algorithm, which was 56%, used for performance evaluation of the ANN-based algorithm formulated by Deka and Quddus (2014).

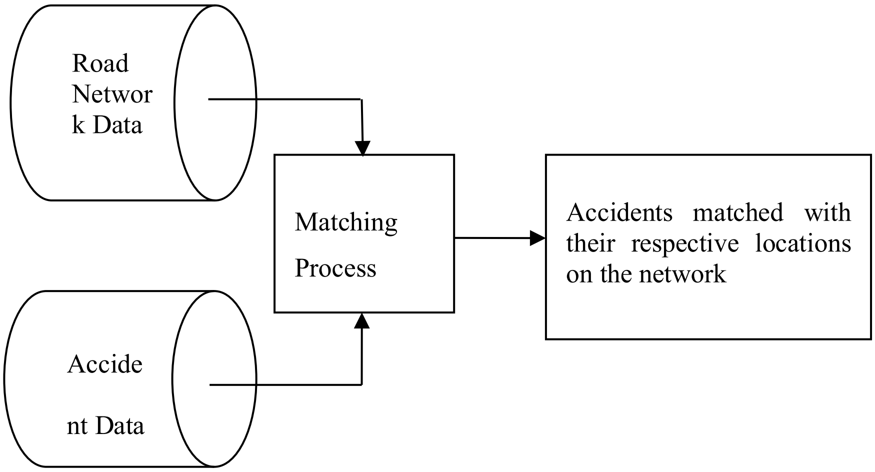

Road Networ k Data
Matching Process
Accidents matched with their respective locations on the network
Accide nt Data

Fig. 6. Typical rectification process

A Lagrangian based heuristic algorithm was used by Tegge and Ouyang (2009) to optimize the number of accidents that occurred on a particular segment. This can also be computed using parallel computing but the technique is suitable for large data sets. In this study the quality of raw data limited the number of utilizable records that can be given as input to the computer program. Therefore, it was not necessary to introduce such sophistications into the algorithm. In the work done by Imprialou, Quddus, and Pitfield (2014) the starting and ending node coordinates were used in the matching process. They claimed that the mean distance between the reported crash and the matched point given on a segment was 8.53 meter. The match point assigned was considered to be the actual location of the initial impact. In this study the actual location of the accident was accepted only when there existed 100% match between the landuse, lane marking and traffic control parameters given the major road width mentioned in the crash report falls within the range filter. Moreover, it has been mentioned in another research (Deka and Quddus, 2014) that accurate identification of accident location is not possible using only one parameter. Therefore, a combination of parameters is required as utilized in this study.

The use of geometry increases the precision with which a location is identified but it limits the number of usable accident records. This result is similar to Steenberghen et al. (2004) who stated that the utilization of relative references, such as milestone, in a matching problem where co-ordinates are missing, increases the accuracy but reduces the number of accidents that could be located. Unlike expressways where mileposts or milestones are readily available and can be easily used to estimate the location of an accident, local roads do not have such markers. Since most of the unsignalized intersections and access points lie on local roads, therefore; major road width, landuse, lane marking, traffic control and number of intersection legs were used as relative references to find the location of an accident. Furthermore when the relative references or the parameters used for matching are not available, the accident could not be located at all. This argument is supported by a study related to the state of Indiana, USA; in which it was stated that 50% of accidents that could not be mapped were on local roads (Tarko, Thomaz and Grant 2009).

There were three reasons for the less number of sites identified. One was the excessive number of blank fields in the raw data obtained from MIROS. The other was the incorrect or impractical data entries such as widths of roads being zero or one meter only. The third was the difference in names of roads and use of ‘shortcuts’ in noting them by different police officers. This error is identical to the one reported by Loo (2006) who stated in his study that the Hong Kong Police accident database had 12.7% mistakes related to road names. Similar to the above, spelling and typographic errors have also been observed in the databases used in the development of traffic safety GIS for Honolulu, Hawaii and have been reported by Kim and Levine (1996) which proves that having such errors in the Malaysian accident database is not an anomaly. These errors resulted into data entries being filtered out which reduced the total number of records that could be utilized in the rectification procedure. Furthermore, the quality of raw accident data is primarily dependent upon how accurately and how much the accident reporting form (ARF) is filled. As compared to accident reporting forms of other low and middle-income countries like Bangladesh in which the total number of items is 67 and India in which the total number of items is 44 only, the Malaysian accident reporting form, known as ‘POL 27’, is much lengthier and contains 91 items with several attributes. In high-income and developed countries like USA the total number of items is even less. In some states, such as New York, the total number of items are as less as 30. The total number of items in the UK’s accident reporting form, known as ‘STATS 19’, is 69 (Austin, 1995). In contrast to the examples of various accident reporting forms given, the verboseness of POL 27 is obvious, which is a four page official document divided into 14 sections. This is the main cause of the less percentage fill of its items and incomplete filling of the entire form that results into blank fields, impractical entries and use of substandard language by the Police officer incharge of investigation. Ultimately the data from such forms becomes part of the MIROS database and creates an erroneous pool of raw data which limits the total number of accidents whose location could be identified.

Acknowledgements

The authors would like to acknowledge the Ministry of Higher Education (MOHE) Malaysia's FRGS grant No. 203/PAWAM6071257 for the financial support that helped conduct this research. Royal Malaysian Police and MIROS are acknowledged for providing data for this study. In addition to this, the authors would also like to acknowledge the Institute of Post Graduate Unit, Universiti Sains Malaysia for the USM Fellowship support.

References

Abdul Manan, Muhammad Marizwan. 2014. "Motorcycles entering from access points and merging with traffic on primary roads in Malaysia: Behavioral and road environment influence on the occurrence of traffic conflicts." *Accident Analysis & Prevention* no. 70:301-313. doi: http://dx.doi.org/10.1016/j.aap.2014.04.009.

Abdul Manan, Muhammad Marizwan, and András Várhelyi. 2015. "Motorcyclists' road safety related behavior at access points on primary roads in Malaysia – A case study." *Safety Science* no. 77:80-94. doi: http://dx.doi.org/10.1016/j.ssci.2015.03.012.

Ahmed, A., A. F. M. Sadullah, and A. S. Yahya. 2013. Contemporary developments in the safety analysis of unsignalized intersections. In *Proceedings of the International Conference on Advances in Structural, Civil and Environmental Engineering – SCEE 2013*. Kuala Lumpur, Malaysia: Institute of Research Engineers and Doctors.

Ahmed, Ashar, Ahmad Farhan Mohd Sadullah, and Ahmad shukri Yahya. 2014. "Accident Analysis Using Count Data for Unsignalized Intersections in Malaysia." *Procedia Engineering* no. 77:45-52. doi: http://dx.doi.org/10.1016/j.proeng.2014.07.005.

Al-Ghamdi, Ali S. 2003. "Analysis of traffic accidents at urban intersections in Riyadh." *Accident Analysis & Prevention* no. 35 (5):717-724. doi: http://dx.doi.org/10.1016/S0001-4575(02)00050-7.

Austin, Kevin. 1995. "The identification of mistakes in road accident records: Part 1, locational variables." *Accident Analysis & Prevention* no. 27 (2):261-276. doi: http://dx.doi.org/10.1016/0001-4575(94)00065-T.

Burns, Shaun, Luis Miranda-Moreno, Joshua Stipancic, Nicolas Saunier, and Karim Ismail. 2014. "Accessible and Practical Geocoding Method for Traffic Collision Record Mapping." *Transportation Research Record: Journal of the Transportation Research Board* no. 2460:39-46. doi: doi:10.3141/2460-05.

Chazelle, Bernard. 1986. "Filtering search: A new approach to query-answering." *SIAM Journal on Computing* no. 15 (3):703-724.

Chung, Younshik, and IlJoon Chang. 2015. "How accurate is accident data in road safety research? An application of vehicle black box data regarding pedestrian-to-taxi accidents in Korea." *Accident Analysis & Prevention* no. 84:1-8. doi: http://dx.doi.org/10.1016/j.aap.2015.08.001.

Deka, Lipika, and Mohammed Quddus. 2014. "Network-level accident-mapping: Distance based pattern matching using artificial neural network." *Accident Analysis & Prevention* no. 65:105-113. doi: http://dx.doi.org/10.1016/j.aap.2013.12.001.

Dutta, Arup, Steven Parker, Xiao Qin, Zhijun Qiu, and David Noyce. 2007. "System for Digitizing Information on Wisconsin's Crash Locations." *Transportation Research Record: Journal of the Transportation Research Board* (2019):256-264.

Elvik, Rune. 2008. "A survey of operational definitions of hazardous road locations in some European countries." *Accident Analysis & Prevention* no. 40 (6):1830-1835. doi: http://dx.doi.org/10.1016/j.aap.2008.08.001.

Erdogan, Saffet, Ibrahim Yilmaz, Tamer Baybura, and Mevlut Gullu. 2008. "Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar." *Accident Analysis & Prevention* no. 40 (1):174-181. doi: http://dx.doi.org/10.1016/j.aap.2007.05.004.

Hashim, H. H., and S. A. S. M Rahim. 2009. The construction of road accident analysis and database system in Malaysia. Paper read at 14th IRTAD Conference.

Havlak, Paul. 1997. "Nesting of reducible and irreducible loops." *ACM Transactions on Programming Languages and Systems (TOPLAS)* no. 19 (4):557-567.

Hawas, Yaser E., Md Bayzid Khan, and Munjed A. Maraqa. 2012. "Analysis of the Effectiveness of the Road-Crash Database in the United Arab Emirates." *Journal of Transportation Safety & Security* no. 4 (3):225-243. doi: 10.1080/19439962.2012.659417.

Howard, B. V., M. F. Young, and J. P. Ellis. 1979. Appraisal of the existing traffic accident data collection and recording system - South Australia. In *CR6*. Canberra: Dept. of Transport, Office of Road Safety.

Imprialou, Maria-Ioanna M, Mohammed Quddus, and David E Pitfield. 2015. "Multilevel logistic regression modeling for crash mapping in metropolitan areas." *Transportation Research Record: Journal of the Transportation Research Board* (2514):39-47. doi: 10.3141/2514-05.

Imprialou, Maria-Ioanna M., Mohammed Quddus, and David E. Pitfield. 2014. "High accuracy crash mapping using fuzzy logic." *Transportation Research Part C: Emerging Technologies* no. 42:107-120. doi: http://dx.doi.org/10.1016/j.trc.2014.03.002.

Kim, Karl, and Ned Levine. 1996. "Using GIS to improve highway safety." *Computers, Environment and Urban Systems* no. 20 (4–5):289-302. doi: http://dx.doi.org/10.1016/S0198-9715(96)00022-1.

Loo, Becky P. Y. 2006. "Validating crash locations for quantitative spatial analysis: A GIS-based approach." *Accident Analysis & Prevention* no. 38 (5):879-886. doi: http://dx.doi.org/10.1016/j.aap.2006.02.012.

Miler, Mario, Filip Todić, and Marko Ševrović. 2016. "Extracting accurate location information from a highly inaccurate traffic accident dataset: A methodology based on a string matching technique." *Transportation Research Part C: Emerging Technologies* no. 68:185-193. doi: http://dx.doi.org/10.1016/j.trc.2016.04.003.

Montella, Alfonso, David Andreassen, Andrew Tarko, Shane Turner, Filomena Mauriello, Lella Imbriani, and Mario Romero. 2013. "Crash Databases in Australasia, the European Union, and the United States." *Transportation Research Record: Journal of the Transportation Research Board* no. 2386:128-136. doi: doi:10.3141/2386-15.

Qin, Xiao, Steven Parker, Yi Liu, Andrew J. Graettinger, and Susie Forde. 2013. "Intelligent geocoding system to locate traffic crashes." *Accident Analysis & Prevention* no. 50:1034-1041. doi: http://dx.doi.org/10.1016/j.aap.2012.08.007.

Steenberghen, T., T. Dufays, I. Thomas, and B. Flahaut. 2004. "Intra-urban location and clustering of road accidents using GIS: a Belgian example." *International Journal of Geographical Information Science* no. 18 (2):169-181. doi: 10.1080/13658810310001629619.

Tarko, Andrew, Jose Thomaz, and Darion Grant. 2009. "Probabilistic Determination of Crash Locations in a Road Network with Imperfect Data." *Transportation Research Record: Journal of the Transportation Research Board* no. 2102:76-84. doi: doi:10.3141/2102-10.

Tegge, Robert, and Yanfeng Ouyang. 2009. "Correcting erroneous crash locations in transportation safety analysis." *Accident Analysis & Prevention* no. 41 (1):202-209. doi: http://dx.doi.org/10.1016/j.aap.2008.10.013.

Thill, Jean-Claude. 2000. "Geographic information systems for transportation in perspective." *Transportation Research Part C: Emerging Technologies* no. 8 (1–6):3-12. doi: http://dx.doi.org/10.1016/S0968-090X(00)00029-2.

Wang, Chao, Mohammed A. Quddus, and Stephen G. Ison. 2009. "Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England." *Accident Analysis & Prevention* no. 41 (4):798-808. doi: http://dx.doi.org/10.1016/j.aap.2009.04.002.

Wang, Yonggang, Chunbo Zhang, and Chengyuan Mao. 2013. "Fatal motor vehicle crashes on road segments in Harbin, China: combining rates into contributory factors." *Transport* no. 28 (2):117-129. doi: 10.3846/16484142.2013.794372.

WHO. 2015. Global status report on road safety 2015. In *Global status report on road safety*. Geneva 27: World Health Organization.