Accident Analysis and Prevention 50 (2013) 1034–1041

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Intelligent geocoding system to locate traffic crashes

Xiao Qin[a,*], Steven Parker[b,1], Yi Liu[c,2], Andrew J. Graettinger[d,3], Susie Forde[e,4]

[a] Department of Civil and Environmental Engineering, South Dakota State University, Brookings, SD 57007, United States
[b] Traffic Operations and Safety (TOPS) Laboratory, Department of Civil and Environmental Engineering, University of Wisconsin, Madison, WI 53706, United States
[c] Department of Computer Science, South Dakota State University, Brookings, SD 57007, United States
[d] University of Alabama, Civil, Construction, & Environmental Engineering, Box 870205, Tuscaloosa, AL 35487-0205, United States
[e] Data Management Section, Wisconsin Department of Transportation, Bureau of State Highway Programs, United States

ARTICLE INFO

Article history:
Received 10 February 2012
Received in revised form 19 July 2012
Accepted 5 August 2012

Keywords:
Accident locations
Geographic information systems (GIS)
Digital mapping
Crash map

ABSTRACT

State agencies continue to face many challenges associated with new federal crash safety and highway performance monitoring requirements that use data from multiple and disparate systems across different platforms and locations. On a national level, the federal government has a long-term vision for State Departments of Transportation (DOTs) to report state route and off-state route crash data in a single network. In general, crashes occurring on state-owned or state maintained highways are a priority at the Federal and State level; therefore, state-route crashes are being geocoded by state DOTs. On the other hand, crashes occurring on off-state highway system do not always get geocoded due to limited resources and techniques. Creating and maintaining a statewide crash geographic information systems (GIS) map with state route and non-state route crashes is a complicated and expensive task.

This study introduces an automatic crash mapping process, Crash-Mapping Automation Tool (C-MAT), where an algorithm translates location information from a police report crash record to a geospatial map and creates a pinpoint map for all crashes. The algorithm has approximate 83 percent mapping rate. An important application of this work is the ability to associate the mapped crash records to underlying business data, such as roadway inventory and traffic volumes. The integrated crash map is the foundation for effective and efficient crash analyzes to prevent highway crashes.

Published by Elsevier Ltd.

1. Introduction

The importance of having a comprehensive crash map has been highlighted as a critical component of safety data management in strategic highway safety plans by many state Departments of Transportation (DOTs). However, the status of having and maintaining an up-to-date crash map varies substantially state by state. Frequently reported obstacles and challenges include: the quality of the data collected from reported crashes, inaccurate crash location information, poor base maps, lack of an effective reference system, and the time-consuming process of manual map production. Moreover, with emerging federal requirements to expand the Highway Performance Monitoring System (HPMS) reporting, there is a need to review infrastructure deficiencies from a statewide perspective. State DOTs are becoming more engaged and involved in local highway programs and are providing more oversight to local projects. A geographic information system (GIS)-based site inventory will facilitate the programmatic transition that improves the overall transportation system design, management, and operational consistency.

Crash data, like any event data, can be converted to a point feature either based on an established linear reference system or an on-at roadway description. Once a geocoding mechanism is developed for crash data, the tool can be applied to other location data such as traffic control devices, bridge locations, etc., which offers a transition from a table-based site inventory to a map or GIS-based inventory.

In general, crashes occurring on state-owned or state maintained highways are a priority at the Federal and State level; therefore, state-route crashes are being geocoded by state DOTs. Crashes occurring on off-state highway system do not always get geocoded due to limited resources and techniques. It is acknowledged that creating and maintaining a comprehensive statewide crash GIS map with both state and non-state crashes is a complicated and expensive task.

A computer algorithm called Crash-Mapping Automation Tool (C-MAT) capable of handling both state highways and local road

* Corresponding author. Tel.: +1 605 688 6355.
E-mail addresses: Xiao.Qin@sdstate.edu (X. Qin), sparker@engr.wisc.edu (S. Parker), Yi.Liu@sdstate.edu (Y. Liu), andrewg@eng.ua.edu (A.J. Graettinger), susie.forde@dot.wi.gov (S. Forde).
[1] Tel.: +1 608 262 2591.
[2] Tel.: +1 605 688 5180.
[3] Tel.: +1 205 348 1707.
[4] Tel.: +1 608 266 7140.

0001-4575/$ – see front matter. Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.aap.2012.08.007

crashes is explained, a quality assurance/quality control (QA/QC) procedure is presented, and a flagging and debugging system for confidence of mapped events is described herein. The highway safety crash application developed treats all roads equally, irrespective of roadway ownership. This work addresses the key safety emphasis areas of: (1) improve data and decision support system, and (2) create a more effective decision processes/safety management system, both of which would benefit greatly with an automated crash mapping tool.

2. Background

For historical and business reasons, the Wisconsin Department of Transportation (WisDOT) maintains two separate data systems, one for state highways and one for local roads. State and local road Location Control Managers (LCMs) are tools used to maintain, display, and edit the state routes (12,000 miles) and non-state routs (100,000 miles).

The State Trunk Network (STN) and the Wisconsin Information System for Local Roads (WISLR) systems were developed and evolved fundamentally independent of each other to meet separate enterprise needs within WisDOT, although similarities and some common data exist between the systems. For example, both systems have the graphical entities of lines (links and chains) and points (nodes) and associated attribute tables that store individual feature data. STN contains virtually no local roads information. On the other hand, state routes are in WISLR, but are not used or maintained because all state route business data, reporting, and analysis currently use the STN system.

The primary goal of merging STN and WISLR is to have a single system able to access, display, and analyze data. Although focused on crash data, the work presented herein can be leveraged for other business data. Because STN and WISLR are operational systems, it was requested that any merge technique minimize the impact on the systems, thereby creating the least disruption to the existing business practice.

In Wisconsin, state highway crashes are geocoded manually by first identifying crash locations from police reports and then assigning a location along a state highway or interstate. This process is referred to as reference point (RP) coding, which identifies a link and offsets associated with each highway crashes. The RP coding process generally lags several months behind crash report processing; therefore, the state highway crash map is not available until 6 months after the crashes are inventoried. Furthermore, no local road crashes were geocoded by WisDOT.

3. Literature review

Considering the value provided through a GIS crash map, researchers have attempted to develop, analyze, and disseminate crash-related geocoding procedures and digital maps to facilitate safety analysis and crash prevention. The successfully developed geocoding procedures include mile-post referencing system, GPS coordinates, address, and intersection and offset geocoding system (Dutta et al., 2007; Graettinger et al., 2001; Harkey, 1999; Kim et al., 1995; Miaou et al., 2005; Park et al., 2011). In a recent study, Zahran et al. developed computer algorithms to automatically process the geospatial road network data without any digitization (Zahran et al., 2011). In the crash report, the location is usually coded by using intersecting street or highway names, i.e., "on highway," "on street" or "at/from highway," "at/from street," and the offset distance and direction from the intersection. The intersection and offset geocoding method reads crash on-at location data from the crash reports, identifies the corresponding node in a map, and assigns latitude and longitude coordinates to the data. Seemingly straightforward, this method faces the challenges in the descriptive accuracy of the crash record, i.e., street name, direction, and offset distance, concurrent highway names, and alias (Dutta et al., 2007).

Geocoded crash data offer rapid visual of crash locations on a map, an extremely informational resource for researchers and engineers to identify crash spatial patterns. Using geocoded incident data, more sophisticated visual analytical tools can be employed for multiple visualizations (Wongsuphasawat et al., 2009). Moreover, crash data with coordinates can take full advantage of the development of geospatial statistics for in-depth inquiries of crash distributions, patterns, and causes. For example, spatial autocorrelation has been considered in the crash prediction models at the Traffic Analysis Zone (TAZ) level and corridor level with improved accuracy (Abdel-Aty and Wang, 2006; El-Basyouny and Sayed, 2009; Siddiquia et al., 2012). Kim and Yamashita analyzed spatial patterns of pedestrian crashes in Honolulu by using K-means clustering techniques (Kim and Yamashita, 2007). Thomas found several advantages in defining black zones using spatial autocorrelation and kernel methods on road segments (Thomas, 1996). Plug et al. analyzed the spatial structures of crashes using kernel density estimation and discovered significant differences in spatial-temporal patterns of single vehicle crashes for different causes (Plug et al., 2011). Recently, Khattak et al. explored the spatial distribution of the locations of secondary and nonsecondary incidents using kernel density analysis and they found that the positive correlation between the two did not necessarily exist (Khattak et al., 2010). Without geocoded crash records, the geospatial data analysis utilizing coordinate information is infeasible. Therefore, it is imperative to develop a robust geocoding methodology that is able to handle crashes regardless of where they occurred.

4. Data sources

To successfully geocode crash locations, two primary data sources are needed: the crash database of police reported crashes and geodatabase containing geo-spatial information and roadway attributes for all local roads in Wisconsin.

4.1. Crash database

Traffic crashes are, by statutory definition, "reportable" if someone is killed or injured, or if property damage exceeds a certain threshold. In Wisconsin, crash information is generally reported by a dispatched police officer via the Wisconsin MV4000 and is eventually archived in the WisDOT crash database. Key attributes from the MV4000 data used to locate a crash, along with field definitions, are provided below:

- DOCTNMBR or ACCDNMBR: a hard print number on the MV4000 used to uniquely identify a crash.
- RPNMBR: reference point number. A crash occurring on a state owned highway is assigned a reference point as a location reference.
- COUNTY: county in which a crash occurred.
- MUNICIPILITY: municipality in which a crash occurred.
- MUNITYPE: municipality type, such as city (C), village (V), or town (T), used to distinguish between municipalities having the same name but different types.
- ONHWY: name of the highway on which a crash occurred.
- ONSTR: name of the local street on which a crash occurred.
- ATHWY: name of the intersecting or nearby highway at/from which a crash occurred
- ATSTR: name of the intersecting or nearby street at/from which a crash occurred
- INTDIR: cardinal direction from the listed intersection

- INTDIS: distance from listed intersection location in hundredths of a mile
- ACCDLOC: type of location at which a crash occurs (public road intersection, public road non-intersection, parking lot, private property)

As attributes of a crash, these data elements are referred to as crash.attribute henceforth.

4.2. Wisconsin Information System for Local Roads (WISLR)

WISLR is a linear referencing system developed and maintained by WisDOT. The statewide local road network combines roadway data for local roads in Wisconsin with interactive mapping functionality. An On/At linear reference method is included in WISLR that facilitates the location of roadway attribute data such as crash location. This method takes advantage of the roadway street names for identifying locations along a defined route. The street name, as identified on the street sign, is used to express a given location and the data at that location. Because WISLR represents the statewide local road network and includes state highways for visual reference and continuous lines, WISLR was selected as the GIS platform for statewide crash mapping. More importantly, WISLR provides an opportunity to link physical roadway characteristic information to the crash reports for safety engineering analysis.

Many WISLR database tables exist that are critical to the mapping algorithm. WISLR database tables relevant to crash location attributes, along with a brief description of each table, is provided below.

- RDWY_RTE: roadway route table is a unique list of road names within each municipality.
- ALT_RWRT_PRFX: alternate roadway route prefix table contains standard and alternate prefixes for road names.
- ALT_RWRT_NM: alternate roadway route name table contains standard and alternate spelling for common road names
- ALT_RWRT_TY: alternate roadway route type table contains standard and alternate types for road names.
- ALT_RWRT_SUFF: alternate roadway route suffix table contains standard and alternate suffixes for road names.
- VLD_ RWRT_PREX: standard roadway route prefix table contains standard prefixes for road names.
- VLD_ RWRT_NM: standard roadway route type table contains standard types for road names.
- VLD_ RWRT_SUFF: standard roadway route suffix table contains standard suffixes for road names.
- On-At: On-At table contains combination of road names that intersect each other.
- RDWY_LINK: roadway link table is a unique list of roadway links or segments.
- RDWY_RTE_LINK: roadway route link table establishes a relationship between routes and links.
- WISLR_PTY: WISLR party table contains municipality names and Ids.

The roadway route table contains a unique list of road names in WISLR separated into four parts: (directional) prefix, (road) name, (road) type, and (directional) suffix. For example, “E Washington Ave” would be separated into three fields and an empty fourth field: roadway-route-prefix = “E”, roadway-route-name = “Washington”, roadway-route-type = “Ave”, and roadway-route-suffix = “”. Each street name in this table is associated with a unique RDWY_RTE_ID.

In WISLR, intersections are identified with nodes listed in the On-At table. Each intersection is identified with a unique REF_SITE_ID. Two street names represented by the RDWY_RTE_ID (On- RDWY_RTE_ID and At- RDWY_RTE_ID) form a node as shown in Fig. 1. Note that the roadway route table is represented twice in Fig. 1 to illustrate the fact that a combination of the On-RDWY_RTE_ID and At- RDWY_RTE_ID is required to obtain the REF_SITE_ID.

Other tables in the WISLR database include alternative and standard tables. Alternate tables are used to standardize alternate spellings (aliases) that are part of a road name and capture different ways to record prefix and suffix information, such as “DRIVE” and “DR”. Standard tables contain the entire list of standard prefixes, types, and suffixes used in WISLR. The role of the Standard tables will be described in subsequent sections. Concurrent names for the same road, such as “US 18/US 151/Verona RD” are captured by the RDWY_RTE_LINK table.

5. Intersection offset geocoding methodologies

A crash can either be intersection-related or segment-related. Segment crash mapping is an extension of the intersection crash mapping, because a nearby intersection needs to be located first followed by the direction and distance calculations. Intersection identification and segment direction and distance calculations are described in the following sections.

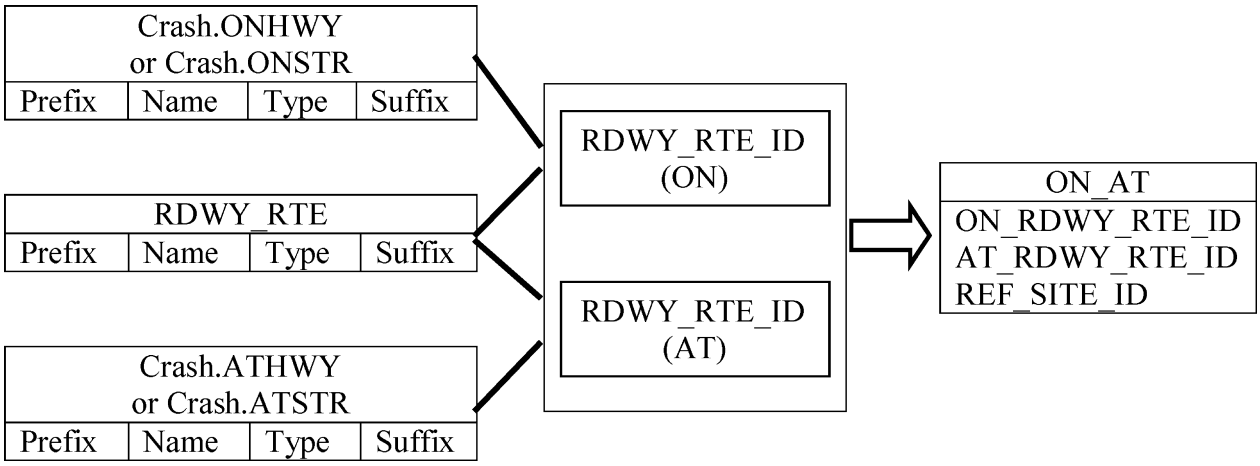


Fig. 1. WISLR relational tables used to identify a unique intersection.

5.1. Intersection level crash geocoding

Determining the intersection location for a crash record requires the algorithm to find the REF_SITE_ID in the On-At table following the logic illustrated in Fig. 1. The REF_SITE_ID is associated with a pair of RDWY_RTE_IDs corresponding to a crash.ONSTR or crash.ONHWY and a crash.ATSTR or crash.ATHWY. State highway names are predominately recorded as numbers, such as 94 for Interstate Highway 94, whereas the majority of street names are composed of letters, such as MAIN ST. Four combinations of On-At information are avaliable and used by the crash location algorithm. The preset combinations in order are: (crash.ONHWY, crash.ATHWY), (crash.ONHWY, crash.ATSTR), (crash.ONSTR, crash.ATHWY), and (crash.ONSTR, crash.ATSTR) with the priority levels from high to low based on the assumption that fewer mistakes will be made with numbers as compared to letters.

Requiring a perfect match for a crash location returns extremely low match rates because a perfect match is very restrictive. Therefore, each crash record's crash.ONSTR, crash.ONHWY, crash.ATSTR and/or crash.ATHWY is parsed into prefix, name, type, and suffix components, and then matched with the Roadway-Route-Prefix, Roadway-Route-Name, Roadway-Route-Type and Roadway-Route-Suffix information in the WISLR RDWY_RTE table. The RDWY_RTE table is first filtered for the municipality where the crash occurred. Parsing is performed by splitting the crash location information field into multiple words, and then utilizing the WISLR tables to analyze each word to determine if that location component is a prefix, name, type, or suffix. The following assumptions are used by the parsing mechanism:

1. At least one word in the parsed data will be the street name field.
2. Only the first word can be tested to see if the word is a prefix.
3. The last word can be tested to see if the word is a suffix. If the last word is a suffix, the immediately preceding word can be tested to see if second to the last word is a type (if that does not violate assumption 1). If the last word is not a suffix, the last word can be tested to see if the word is a type.
4. If a word is not a prefix, a type, or a suffix, the word has to be used as the name field.

The parsing mechanism performs two levels of analysis with respect to the WISLR tables. Level 1 analysis attempts to parse *On-Street* and *At-Street* fields into the prefix, name, type, and suffix fields based on the contents of WISLR Standard tables. Level 2 uses the Alternate tables in WISLR in order to convert non-standard formats into standards ones during the parsing procedure. For instance, the alternate prefix 'North' could be standardized into 'N' in Level 2 parsing. Use of the Alternate tables is necessary since all street names are standardized in WISLR. The algorithm only parses using Alternate tables if the street is not found in WISLR after parsing with Standard tables.

This piecewise matching provides greater flexibility and intelligence when the information is incomplete. The match is conducted using a rigorous algorithm that considers spelling errors, roadway name aliases, and incomplete crash information. The primary challenge in developing the matching algorithm was the presence of incomplete street names in crash records. To handle this situation, five levels of matching, based on the available street name information, were established:

1. Name Matching: The Name field of the parsed crash field is matched to the *Roadway-Route-Name* field in the RDWY_RTE table. The additional Prefix, Type, and Suffix information is ignored.
2. Prefix-Name Matching: Both Prefix and Name fields of the parsed crash record are matched to WISLR. Suffix and Type information are ignored.
3. Name-Type Matching: Both Name and Type fields of the parsed crash record are matched to WISLR. Prefix and Suffix information are ignored.
4. Prefix-Name-Type Matching: Prefix, Name, and Type fields of the parsed crash record are matched to WISLR. Suffix information is ignored.
5. Prefix-Name-Type-Suffix Matching: Prefix, Name, Type, and Suffix information of the parsed crash record are matched to WISLR. This level takes into account all available street name information to find the WISLR *Roadway-Route-ID*.

Spelling errors in the name field of the parsed crash record are the most critical. If the name field cannot be matched, all match levels will be unsuccessful. A spell-check module was developed using the Damerau-Levenshtein distance algorithm (Damerau, 1964) to match a crash record street name(s), when a perfect match cannot be found, to the most similar street name(s) in the *Roadway-Route-Name* field in WISLR Roadway Route table.

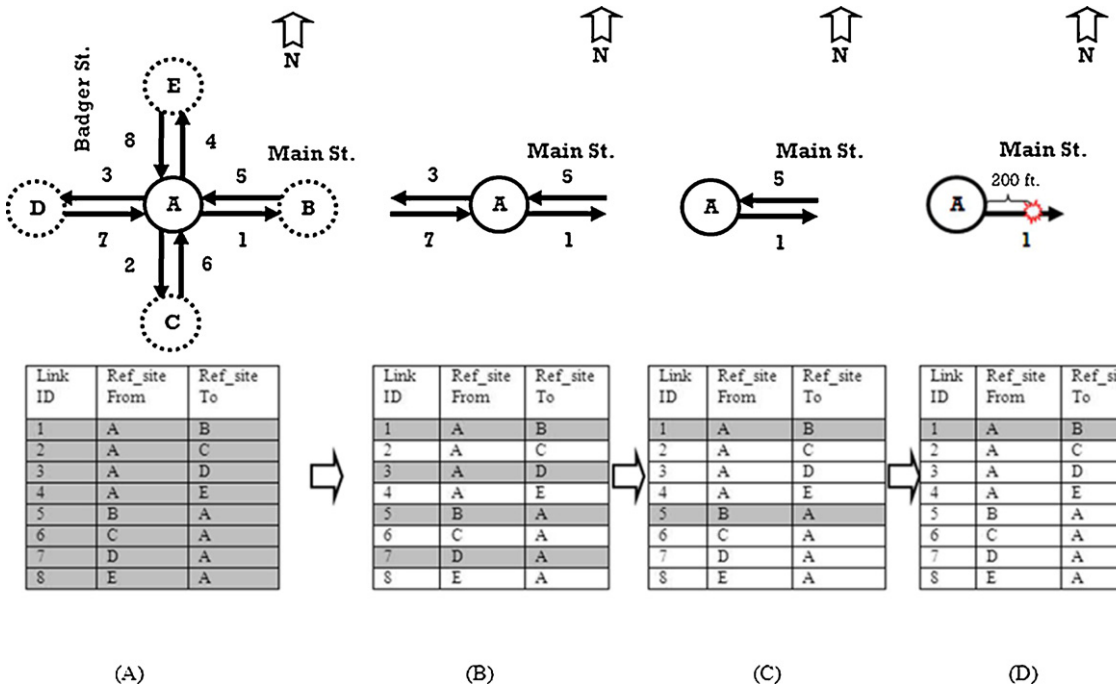
The implementation is designed in a stepwise fashion, starting with the most rigorous matching process and gradually relaxing the conditions until a successful match is found. In particular, the algorithm attempts to minimize any modifications of a street name in the crash record while at the same time the algorithm attempts to find a match at the highest match level.

For a street or highway name, if a match is not found, the match level value is reduced. If no match is found at the lowest match level, then the least possible amount of modification of street name is introduced, and again there is an attempt to find a match from the highest match level to the lowest match level. The process is repeated until one or more RDWY_RTE_ID is found or all match level and modification options are exhausted. In addition to the match levels set for the street names, the municipality specified in the crash data and the neighboring municipalities are considered in the matching process.

It is not unusual for a police officer to record the crash in a neighboring municipality, especially when a crash occurs on a jurisdictional boundary. Municipality issues are handled by separating the ON_AT table into the tables ON_AT_SAME_PTY and ON_AT_ADJACENCY. A match always starts with an on street and at street within the same municipality. When both (on) RDWY_RTE_ID and (at) RDWY_RTE_ID are available, a query is performed in the ON_AT_SAME_PTY table to find the REF_SITE_ID where the two roads intersect. If a REF_SITE_ID is found, the intersection level mapping is completed. If a REF_SITE_ID is not found in the officer recorded municipality, the matching process will try the street names in the neighboring municipalities and search for REF_SITE_IDs in the ON_AT_ADJACENCY table. Once a REF_SITE_ID is found for a crash, the first RDWY_LINK is selected from all the links connected to this REF_SITE_ID, with an offset of zero if the REF_SITE_ID is the From node, or the link length if the REF_SITE_ID is the To node. The reason for assigning a RDWY_LINK_ID and a RDWY_OFFSET is that the actual event geocoding is based on a linear referencing system where only link and offset information is used.

5.2. Segment-level crash geocoding

A precondition for a successful segment crash mapping is a successful intersection mapping. The appropriate link as well as the link offset or distance needs to be determined. A four-step process for geocoding a segment crash is illustrated in Fig. 2. A hypothetical crash occurred on Main St., 200 feet east of Badger St. Main St. and Badger St. intersect at node A, and there are eight links connected to node A, with four links on Main St. and four links on Badger St.



Link ID	Ref_site From	Ref_site To
1	A	B
2	A	C
3	A	D
4	A	E
5	B	A
6	C	A
7	D	A
8	E	A

Fig. 2. (A–D) Segment-level crash mapping used logic to identify an appropriate link.

The algorithm begins by identifying all links connected to the crash intersection. There are eight rows in the RDWY_LINK table that have node A as either a From or To node, as shown in Fig. 2A. Next, only the links of the roadway on which a crash occurred are kept; therefore, only four links are highlighted in Fig. 2B. This step is completed by limiting the candidate links to the ones corresponding to the RDWY_RTE on which a crash occurred, using the relationship established in the RDWY_RTE_LINK table. Next, as shown in Fig. 2C, the direction information from the MV4000 is used to determine which two links are on the appropriate side of an intersection. The RDWY_LINK table has been preprocessed to obtain the cardinal direction for each link. By matching crash.ONSTR or crash.ONHWY and crash.INTDIR to the links, the candidate links can be restricted to two. The last step shown in Fig. 2D is to decide on which link the crash should be placed. Since either link 1 or link 5 is potentially correct, the priority is given to the link with REF_SITE_FROM as the intersection. On one-way streets the REF_SITE_TO can be the only available link and is used.

Circumstances can become more complex if a crash is recorded further away from an intersection that the first link, i.e., crash.INTDIS is longer than the length of the link. In this case, the crash will traverse links in the crash.INTDIR until the crash.INTDIS is exhausted. The process is illustrated in Fig. 3 as a two-step process.

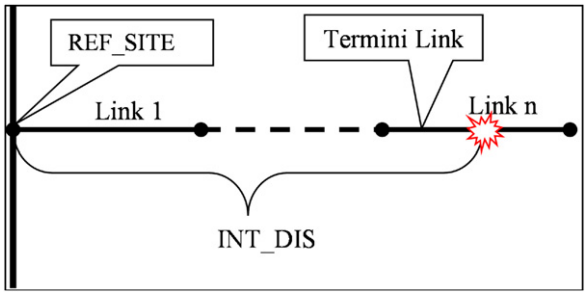


Fig. 3. Determining WISLR link ID and link offset when crash distance is longer than the link distance.

Step 1: Determine the termini link where the crash will be placed according to crash.INTDIS. In the example in Fig. 3, the termini link is Link n.
Step 2: Determine whether the termini link node is the FROM node or the TO node.
Step 2a: if the termini link node is the FROM node, the link id is link n and the link offset is defined as subtracting the sum of link lengths ($\sum_{i=1}^{n-1} L_i$) from crash.INTDIS.
Step 2b: if the termini link node is the TO node, the link id is link n and the link offset is defined as subtracting crash.INTDIS from the sum of link lengths ($\sum_{i=1}^{n} L_i$).

Crash.INTDIS may contain multiple links but the key step is to determine the termini link and whether the termini link node is the From node or the To node, because the link offset is always measured from the From node. The output of the segment-level geocoding process is the link ID and link offset for each crash.

In order to match candidate RDWY_LINKs to a crash report crash.INTDIR field, it is necessary to determine the cardinal direction for each RDWY_LINK. Since RDWY_LINK cardinal directions are not included in WISLR, link directions were derived by considering a straight line connecting the RDWY_LINK start- and end-coordinates in the WISLR link shapefile. This process was run against each link in WISLR using an ESRI ArcObjects-based procedure and stored in a separate LINK_DIRECTION table. This method is not foolproof, i.e., links may have excessive curvature making the straight line approximation invalid.

6. Quality assurance and quality control (QA/QC)

To measure the success of the geocoding algorithm; accuracy, precision, and completeness were considered as the three most important performance metrics. Accuracy means the location of a geocoded crash has a high degree of conformity to the original crash location description. Accuracy was used to evaluate intersection-level mapping where only names are considered. Despite the fact that the original crash location description may contain imperfect information caused by spelling errors, missing prefix, suffix, or incorrect highway or street types, the mapped crash should be correct based on this information.

Precision means the geocoded position has a high degree of conformity with respect to the original crash location based on distance. Precision was used to evaluate segment-level crash mapping where the distance on the link is important. Though a crash may not be geocoded to the identical location of the original crash, the distance between the two should be shorter than a predetermined threshold. This specifically applies to urban areas having a high density network. This metric was used for crashes occurring on a link at a distance away from an intersection. Both accuracy and precision are microscopic metrics evaluating individual crashes.

Completeness is a macroscopic measure which is the percentage of crashes that can be geocoded with acceptable accuracy and precision. It is the ultimate measure of the power and robustness of the crash geocoding algorithm. State highway crashes and local crashes have respective QA/QC procedures because the benchmark sources are different. Local road crashes mapped with C-MAT were compared to Google maps output while state highway crashes mapped with C-MAT were compared to manually geocoded crashes which were transferred from STN to WISLR. Additional QA/QC was performed manually for a subset of crashes: 90 randomly sampled records per county or approximately 1% of total crashes for 2005–2009.

6.1. QA/QC for state highway crashes

Manually geocoded crashes can be used as ideal quality assurance measures given the quality control. To compare C-MAT output with manually geocoded crashes on the same platform, manually geocoded crashes were programmatically moved from STN to WISLR via the link method (Graettinger et al., 2009).

Agreement of crash locations are determined by comparing REF_SITE_ID for intersection crashes and LINK_ID for segment crashes. The results show 61,624 out of 70,046 algorithm mapped intersection-related state highway crashes matched moved manually coded crashes at 88% accuracy. However, the segment level match is relatively low; 53,515 C-MAT mapped crashes out of 87,675 total crashes or a 61% accuracy.

An examination of the unmapped crashes shows clusters at or near interchanges. Interstate interchanges are the most troublesome spots because of the complicated geometry and lack of unique names describing the interchange segments. Case studies reveal these complex interchanges are attributable to the low matching percentage of segment crashes.

6.2. QA/QC for local crashes

Unlike state highway crashes, crashes that occur on local streets do not have a benchmark dataset of manually geocoded crashes. Manual geocoding would be very time-consuming given the large number of local crashes. Finding a systematic approach to assess accuracy, based on an independent mapping source, was needed. An automatic approach was developed using Google API that identified the Universal Transverse Mercator (UTM) coordinates of a Google intersection based on the crash.ONSTR and crash.ATSTR information for each record. Algorithm mapped crashes were converted to UTM coordinates from a shapefile in ArcGIS. If the two locations are found to match, there is a high probability that the crash is mapped to the correct location. To measure the agreement between the crashes mapped by Google Maps and C-MAT, the Euclidean distance between the two locations was calculated from the coordinates. Due to the limitations of Google Maps, only intersection-level crashes could be compared.

The consistency between intersection locations on the WISLR map and Google source maps needs to be verified first because Google API uses Google source maps to locate crashes, same as C-MAT uses WISLR to locate crashes. To do this, intersection names in WISLR were imported into Google Maps to obtain the UTM coordinates for every intersection. The coordinates were then compared to the coordinates of WISLR nodes obtained from the WISLR shapefile. The comparison suggests a high agreement (more than 90%) was found between the Google source maps and WISLR map. After validating the source maps of both systems, crash location information was respectively mapped by Google API and C-MAT, and the results were compared. Where the results did not agree, the following procedure was employed.

1) Use coordinates found by Google Maps to retrieve address names, and analyze whether addresses match the original crash data.
2) Use C-MAT mapped coordinates in WISLR to retrieve address names, and analyze whether address match the original crash data.

This QA/QC procedure treats Google API as an "interpreter" of a crash location using its source maps. Input crash location descriptions are first translated into an X and Y coordinate, which may or may not be the correct location. One way to evaluate whether a Google translation is correct is to translate X and Y back to a text description and then verify that description with the original input. Similarly, C-MAT is another "interpreter." Checking the X and Y generated by C-MAT should return the same location description as the input.

Irrespective of being right or wrong, Google Maps usually provides a location. Without carefully reviewing the output, it is difficult to decide if the result is accurate. The Google API, which allows one to programmatically access Google Maps location information, included a mapping accuracy measure with nine levels indicating the mapping quality with zero being an unknown location. It was determined that only level 7, intersection level accuracy, and above had sufficient accuracy required for this study. The Google Maps accuracy levels are (Google Maps, 2011):

- Level 0: Unknown location.
- Level 1: Country level accuracy.
- Level 2: Region (state, province, prefecture, etc.) level accuracy.
- Level 3: Sub-region (county, municipality, etc.) level accuracy.
- Level 4: Town (city, village) level accuracy.
- Level 5: Post code (zip code) level accuracy.
- Level 6: Street level accuracy.
- Level 7: Intersection level accuracy.
- Level 8: Address level accuracy.
- Level 9: Premise (building name, property name, shopping center, etc.) level accuracy.

In addition to the Google Maps comparison, a manually verify sample of crashes were evaluated. A random sample of crashes was selected by taking 90 crashes per county (60 local road crashes and 30 highway crashes) for a total of 6480 crashes, or approximately 1% of all Wisconsin statewide reported crashes over the 2005–2009 timeframe. Each crash in the sample set is hand-inspected and assigned a flag indicating whether it is mapped correctly, mapped incorrectly, or not mapped. Crashes that are mapped incorrectly were further analyzed to determine the primary cause (e.g., problems with the source data, the WISLR network, or geocoding algorithm). The manual QA/QC results show that 5601 crashes or 86.44% are mapped correctly; 232 crashes or 3.58% are mapped incorrectly; and 647 crashes or 9.98% are not mapped. The manual QA/QC process produced results that are consistent with the Google API based analysis described above.

6.3. Debugging and flagging system

Though the results of crash mapping are encouraging, it is understood that manual intervention to improve accuracy will be required for some mapped crash. Unmapped crashes require manual processing, but inaccurately mapped crashes also require manual processing. Despite a relative high match percentage at the intersection level, it is not clear what mapped crashes require a manual review. To assist in identifying mapped crashes that should be manually reviewed, a QA/QC crash flag system was developed. The flag system: (1) provides confidence levels, (2) provides debugging details, (3) provides feedback about MV4000 source data, and (4) facilitates manual cleanup. These flags help calculate a confidence level which serves as an indicator of the quality of a mapped crash. The confidence level is a weighted average of nine flags, including match-level, match-parity, spelling-level, muni-flag, ref-site-flag, intdis-flag, intdir-flag, link-distance-flag, and link-direction-flag. The weights assigned to each flag vary by the importance related to map quality. Unfortunately, there is no easy way to determine the weights other than trial and error.

7. Results

C-MAT was able to map 498,976 crashes out of 603,267 (5-year Wisconsin crash data between 2005 and 2009) on the WISLR linework or 82.71 percent of the total crashes mapped. Note that approximate 2.75 percent of the crashes were filtered out in the mapping process due to incomplete location information, i.e., crashes without crash.ONSTR and crash.ONHWY or crashes without crash.ATSTR and crash.ATHWY. After excluding the cases missing critical location information, local and state highway crash mapping percentages are 89.70 and 79.73 percent, respectively. Table 1 shows the detailed summary statistics.

The QA/QC procedures and processes further developed will be implemented in subsequent years to produce a high quality, timely crash map.

For crash records where no location could be found, a manual review was performed on a sample basis to identify specific reasons contributing to the mapping failures. The manual review identified several issues listed below:

1) No valid On/At information
   The mapping algorithm currently implemented is based on complete and accuracy on/at location information. Crash data, without a pair of valid on/at location data, e.g., the state trunk highway crashes referenced on milepost or other linear referencing system, cannot be mapped with the current algorithm. Valid names are the Interstate, State, or US highways or their alternative names (local street names or alias) which are available in the WISLR tables.
2) No intersection in spite of available on/at information
   Crash report on/at location information does not always reference an intersection. This issue is frequently encountered at interchanges with under- or over-passes. Most interchange bridges are located in the state trunk network but the corresponding location information is not available in the WISLR tables. Mileposts are also used occasionally on the crash report to locate a crash.
3) Missing legs/approaches
   Some intersections have one or more approaches belonging to private roads such as the driveway of a trailer park, etc. In the WISLR tables, these locations with missing legs/approaches are not always stored as reference sites; therefore, cannot be displayed in the WISLR base map.
4) Complex interchanges

Table 1
Map outcome summary statistics.

		Uniquely mapped	Duplicates	Not mapped	Total	
Total state wide	Local roads	258,114 (82.59% of state wide crashes on local roads)	22,217 (7.11% of state wide crashes on local roads)	32,182 (10.30% of state wide crashes on local road)	312,513 (53.26% of total crashes are on local roads)	586,724[a] (total mappable crashes)
		280,311 (89.70% of the local road crashes mapped)				
	State routes	162,672 (59.32% of state wide crashes on state routes)	55,973 (20.41% of state wide crashes on state routes)	55,566 (20.26% of state wide crashes on state routes)	274,211 (46.74% of total crashes are on state routes)	
		218,645 (79.73% of state route crashes mapped)				
	Total	420,786 (71.72% of state wide crashes mapped uniquely)	78,190 (13.33% of state wide crashes mapped to duplicate locations)	87,748 (14.96% of state wide crashes did not map)		
		498,976 (85.04% of the mappable crashes or 82.71% of the total crashes[a] mapped)				
Segment related crashes	Local roads	156,499 (81.67% of segment crashes on local roads)	15,230 (7.95% of segment crashes on local roads)	19,883 (10.38% of segment crashes on local roads)	191,612 (52.64% of segment crashes are on local roads)	363,971 (62.03% of total crashes are segment related)
	State routes	94,819 (55.01% of segment crashes on state routes)	36,293 (21.06% of segment crashes on state routes)	41,247 (23.93% of segment crashes on state routes)	172,359 (47.36% of segment crashes are on state routes)	
Intersection related crashes	Local roads	101,615 (84.05% of intersection crashes on local roads)	6987 (5.78% of intersection crashes on local roads)	12,299 (10.17% of intersection crashes on local roads)	120,901 (54.28% of intersection crashes are on local roads)	222,753 (37.97% of total crashes are intersection related)
	State routes	67,853 (66.62% of intersection crashes on state routes)	19,680 (19.32% of intersection crashes on state routes)	14,319 (14.06% of intersection crashes on state routes)	101,852 (45.72% of intersection crashes on state routes)	

[a] Of the 603,267 crashes in Wisconsin between 2005 and 2009, only 586,724 had complete location information reported.

Interchanges are more complicated than intersections because of ramps, ramp intersections, and over- or underpasses. The interchange layout can be depicted by dozens of links and nodes representing various vehicular movements and changes in geometric characteristics, whereas the interchange name may be as simple as a pair of on/at locations. Based on on/at information only, it is difficult to assign a crash to an appropriate link and offset.

Some of these issues may be resolved by supplementing other data such as bridge location information or other DOT maintained business data.

8. Conclusions

The goal of this study was to develop and apply the Crash-Mapping Automation Tool (C-MAT) based on the intersection offset geocoding method to produce an accurate, precise, and complete single Wisconsin crash map. The research introduced in the paper identified issues in the crash mapping process, tested alternatives to improve the tool, developed and validated promising solutions to the issues, and established rigorous QA/QC procedures. The current version produced a mapping percentage of approximately 83 based on the outcome of 603,267 crashes (5-year Wisconsin crash data between 2005 and 2009). Approximate 3 percent of the crashes were filtered out in the mapping process due to incomplete location information.

Different QA/QC approaches were employed to evaluate local and state highway crashes, respectively. For local crashes, crashes mapped by Google Maps were used as a reference; for state highway crashes, WisDOT manually geocoded crashes were used as a reference. Though relatively high agreement was found in various mapping tools, the disparity was investigated.

The confidence level of crash locations developed in this study provides useful feedback to the quality of mapping. This quantitative measurement however needs to be thoroughly assessed to ensure a truthful reflection of the potential problems associated with a mapped crash.

References

Abdel-Aty, M., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. Transportation Research Record 1953, 98–111.

Damerau, F.J., 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM 7 (3), 171–176.

Dutta, A., Parker, S., Qin, X., Qiu, Z.J., Noyce, D.A., 2007. A system for digitizing Wisconsin crash location information. Transportation Research Record 2019, 256–264.

El-Basyouny, K., Sayed, T., 2009. Urban arterial accident prediction models with spatial effects. Transportation Research Record 2102, 27–33.

Google Maps, 2011. Google Maps JavaScript API V2 References, http://code.google.com/apis/maps/documentation/javascript/v2/reference.html (accessed June 2011).

Graettinger, A.J., Rushing, T., McFadden, J., 2001. Evaluation of inexpensive GPS units to collect crash location data. Transportation Research Record 1746, 94–101.

Graettinger, A.J., Qin, X., Spear, G., Parker, S., Forde, S., 2009. Combining state route and local road linear referencing system information. Transportation Research Record 2121, 152–159.

Harkey, D.L., 1999. Evaluation of truck crashes using a GIS-based crash referencing and analysis system. Transportation Research Record 1686, 13–21.

Khattak, A.J., Wang, X., Zhang, H., 2010. Spatial analysis and modeling of traffic incidents for proactive incident management and strategic planning. Transportation Research Record 2178, 128–137.

Kim, K., Levine, N., Nitz, L., 1995. Development of a prototype traffic safety geographic information system. Transportation Research Record 1477, 41–47.

Kim, K.E., Yamashita, E.Y., 2007. Using K-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. Journal of Advanced Transportation 41 (1), 681–689.

Miaou, S., Tandon, R., Song, J.J., 2005. Providing personalized traffic safety information to the public: using web-based geographical information system technologies. Report SWUTC/05/167424-1. Texas Transportation Institute, College Station.

Park, S.H., Bigham, J.M., Kho, S.Y., Kang, S., Kim, D.K., 2011. Geocoding vehicle collisions on Korean expressway based on postmile referencing. KSCE Journal of Civil Engineering 15 (8), 1435–1441.

Plug, C., Xia, J., Caulfield, C., 2011. Spatial and temporal visualization techniques for crash analysis. Accident Analysis and Prevention 43 (6), 1937–1946.

Siddiquia, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis and Prevention 45, 382–391.

Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidents. Accident Analysis and Prevention 28 (2), 251–264.

Wongsuphasawat, K., Pack, M., Filippova, D., VanDaniker, M., Olea, A., 2009. Visual analytics for transportation incident data sets. Transportation Research Record 2138, 135–145.

Zahran, E.M., Bennett, L.D., Smith, M.J., 2011. A GIS-based approach to automating the collection of geospatial road network data. Traffic Engineering and Control 52 (7), 295–299.