

Iilir Bejleri
Alberto Vazquez
Clara DiBella

TOOLS FOR INTEGRATING CRASH DATA INTO THE ARCGIS TRANSPORTATION DATA MODEL

Abstract: Palm Beach County Engineering and Public Works receives crash reports from the 33 law enforcement agencies within its jurisdiction and enters pertinent information into a non-spatial database. In order to capture, display, and analyze the spatial location of crashes and other transportation data, the Department has migrated to the ArcGIS Transportation Data Model (TDM). This paper presents the strategies used for implementing the TDM, describes tools developed in VBA for integrating crash data as events into the Model, and concludes with lessons learned from the experience in an enterprise setting.

Keywords: Transportation, Crash, GIS Application

A Brief History of Crash Data Collection in Palm Beach County

Introduction

In the 1970's, traffic accident records were collected and stored in filing cabinets at Palm Beach County's Highway Safety Department. Accident reports were filed under the law enforcement agency that wrote them. It was, as one would imagine, time consuming to hand search through crash reports to find the ones necessary for a study on a specific location.

At some point a mainframe computer arrived to make quicker work of locating crash reports and provide a computer printout of results. Around this time the Highway Safety Department was disbanded and the Accident Record Section was formed and relocated to the County's Cartographic Department.

Later on, the Accident Record Section was relocated to the Division of Emergency Services and then again to the Traffic Engineering Division where it resides to this day. The Cartographic Department would later become the Engineering Department's Geoprocessing (GIS) Section.

After the mainframe computer became the old cumbersome mainframe, a decision was made to move to software developed by the University of Florida for the small computer. The **Small Computer Accident Record System (SCARS)** and its subsequent iterations has been the database used by Palm Beach County for crash-related information since April of 1992.

Presently the Accident Records Section is responsible for collecting, coordinating, and analyzing all traffic accident reports for law enforcement agencies in Palm Beach County, including municipal police departments, the County Sheriff's office, and the Florida Highway Patrol. The Section currently processes over 50,000 crash reports per year and enters pertinent data into the SCARS database for over 38,000 that specifically pertain to public roadway crashes.

The Section also produces over 200 crash studies per year and generates an annual summary report of data collected. This annual report includes the number of raw accidents, accident rate by volume for select intersections, both intersection and non-intersection fatal crash locations, and other information like seat belt usage, driver age, motor vehicle type, crashes by month, day of week, hour of day, etc. Data not included in the database is the individuals name, address, race, insurance information, and the vehicle's identification number.

The current version of SCARS has been providing the 33 agencies within Palm Beach County and the Geoprocessing Section with a Dbase (.dbf) file of the database when requested. The program exports to this file format for selections, whether of accidents per jurisdiction, or type (like fatal, bike, or pedestrian) crashes, or time frame.

But the database is now old and needs to be improved and the decision was made to upgrade the database to Oracle. User needs were reevaluated and a redesign of the tables and relationships was put forth. Data entry user interfaces were also improved as part of the upgrade to a relational database. The Geoprocessing group thought it was a good time to design the new database with GIS analysis and mapping capabilities specifically within the newly created Palm Beach County's Transportation Data Model (PBC-TDM) framework.

One of the advantages of the existing crash data is that the (2) location fields per record can be geocoded allowing for intersection identification with an acceptable degree of accuracy once a fair amount of conditioning of the table (searching, replacing, parsing of fields, etc.) was accomplished. The Accident Records Section has been throughout the years entering a "node number" in the database for intersection locations.

It was thought that by having the ability to select spatially using the geocoded intersection node as defined by the Accident Records Section one could manage and automate the display of this data in a GIS environment using location referencing. It was further expected that for those non-intersection crashes that occurred, direction and distance offset from an intersection (node) captured in the database would represent measurement along a segment and would be better represented graphically in the model.

Problem Identification

GIS mapping of crash data prior to the creation of the TDM involved the time-consuming process of manually finding the appropriate intersection and placing a point and accident number in the appropriate field in an ArcView shapefile (.dbf) database. After the crash locations were identified, the accident number was used to join to the crash tables. For this reason, mapping of crash data was always conducted on a subset of the larger data for any given year, the subset usually being fatal, bike, and pedestrian crashes only.

Alternatively, attempts to geocode were met with poor results as street names used in the crash database to a large degree didn't match the official (postal standard) street names used by the County and reflected in their centerline GIS.

The old SCARS database was developed with Dbase and was limited in functionality by today's standards. The database only allowed for a limited number of fields. Therefore, some information in the database tables were concatenated into one field. The amount of characters in each field was also limited, which caused large street names to be truncated. One could imagine the level of creativity that the user participated in to make a name fit within the allotted space.

Place names, generalized block addresses (ie: 5400 block of whatever street), and mile posts, were excluded from the crash table conditioning to achieve a geocoding match rate of 75% for existing node numbers. Street name truncation, issues with street type and directionals, along with data input errors were among the primary reasons 25% of the nodes didn't match. It is expected that further conditioning of the table and customization to the geocoding tool will increase the number of node numbers matched to intersections.

With the move of the centerline GIS data to the enterprise and the TDM, an alternate name table was created and the alternate name was specific for the road segment. Most of the misnamed roads in the crash table(s) were state roads within the County where the "SR" number was used instead of the official name. ESRI's out-of-the-box geocoding tool wants to convert SR to STHY, (US to USHY) etc. A prefix-type field in the Alternate name table was added to accommodate some of the problems with geocoding.

Goals and Objectives

The project as undertaken had several goals. To develop a methodology that achieves a high match rate to actual crash locations was one goal. The production of automated map products and the ability to visualize the complete or selected subsets of the dataset at any time is also of great importance.

Storing the crash data within the framework of the Palm Beach County Transportation Data Model, the facilitation of data management, and overall best database management practices represents an important goal. Also, the implicit -vs- explicit relationship to the centerline geometry that is afforded using location referencing lends to data entry personnel not needing to learn GIS to perform their job function. They simply work off an interface they are accustomed to using and processing for mapping occurs when requests are made from management.

Small improvements to the application interface, such as limiting the data entry persons' street name and alternate name choices in the database using the VB AutoComplete function, greatly reduces if not eliminates data entry errors. Also, the use of the measure values and directions taken from the crash tables utilizes existing data in an event table in which the graphic representation follows along the geometry of the route feature and represents a more accurate representation of the crash locations.

Methodology

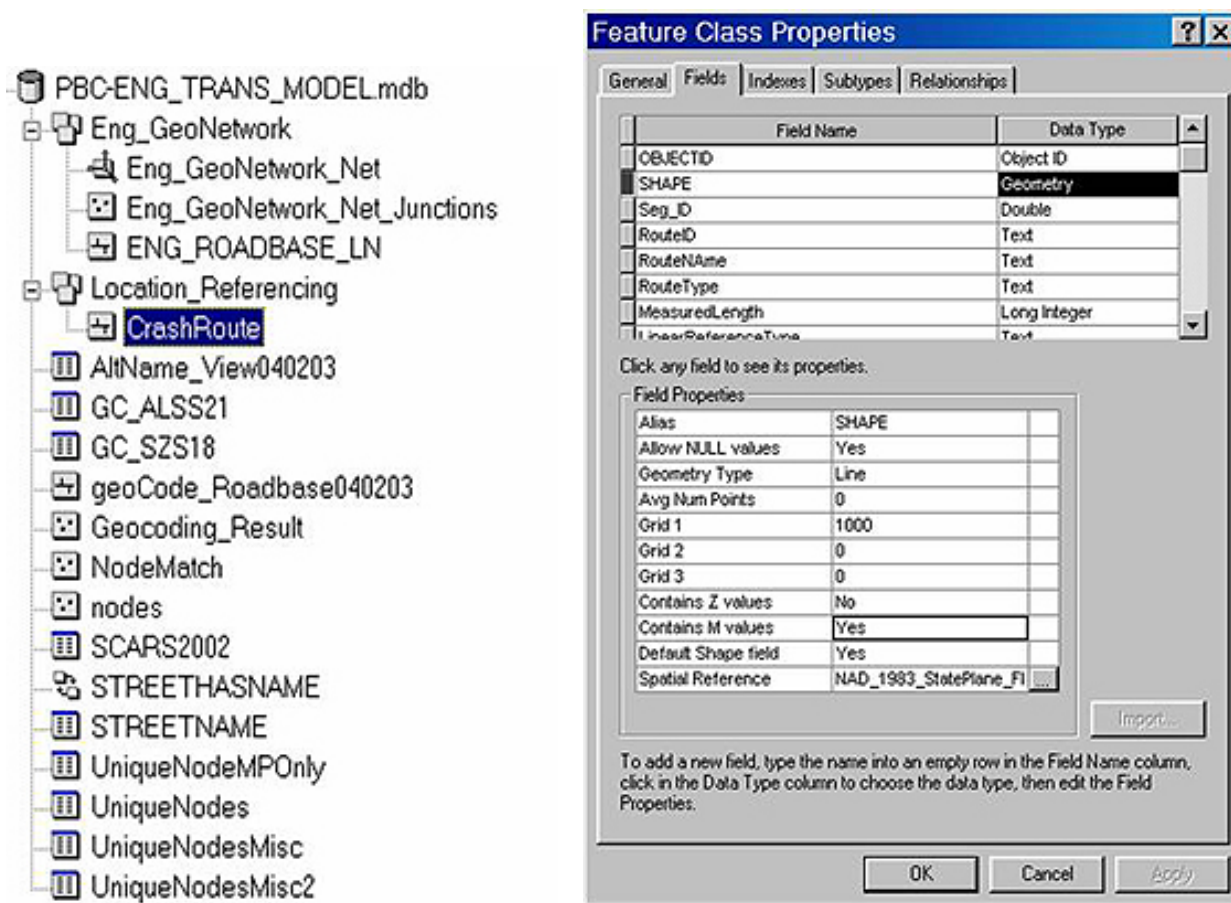
This section describes the methodology used to develop a custom tool using Visual Basic for Applications (VBA) for geocoding crash data and for integrating them into the Transportation Data Model. The methodology involves two major components: preparation of the required databases and creation of the event crash tables.

Data Preparation

There are three main data files required for the mapping of crashes as events using location referencing: The routes (streets) database, the spatial intersections Node layer and crash data tables.

Creating the route file

The route file, named CrashRoute was developed as a separate feature dataset inside the ArcGIS Transportation Data Model. It is important to remember when creating a route feature, to enable M values. ArcGIS stores x,y,m in the feature geometry. Enabling the M value, is what makes location referencing possible. When creating a route, under field properties, the default SHAPE field needs to be changed to yes. The feature class stores the measured value in the field called SHAPE. Documentation on how to create a route feature class is well documented in the ArcGIS Transportation Data Model literature. The computer screen image captures below (Figure 1) demonstrates this process with Palm Beach County data.



Shape	OBJECTID	Seg_ID	RouteID	RouteName	RouteType	SHAPE_Leng	STREET
Polyline M	1	9		DAKOTA	2	151.102783	DAKOTA CT
Polyline M	2	10		5TH		250.930944	5TH AVE S
Polyline M	3	12		TECUMSEH		424.981963	TECUMSEH DR
Polyline M	4	16		BEARS CLUB		832.340738	BEARS CLUB DR
Polyline M	5	39		170TH		622.501518	170TH RD N
Polyline M	6	67		SPUR CLOSE		264.128349	SPUR CLOSE
Polyline M	7	73		COLLIN	2	265.709597	COLLIN DR
Polyline M	8	79		CORAL LAKES		345.824175	CORAL LAKES DR
Polyline M	9	98		SUMMIT RUN		851.626652	SUMMIT RUN CIR
Polyline M	10	118		FLORIDA	2	874.325113	FLORIDA BLVD
Polyline M	11	122		CHESTNUT HILL	2	251.798808	CHESTNUT HILL RD
Polyline M	12	150		M&RCH		241.253752	M&RCH CIR

Figure 1. CrashRoute table

Creation of the spatial Node layer

In the existing SCARS application a node number is assigned when inputting the crash report information into the database for the first time (see Figure 2). The existing nodes are based on the intersection of streets in which crashes occurred or are measured from. The location description fields (LOC_DESC1 and LOC_DESC2) describe this in the database. The SCARS crash tables were imported into an MS Access database and LOC_DESC1 and LOC_DESC2 were concatenated. Geocoding was then performed based on the intersections and the existing node geometries were captured into the GIS.

Accident report (Esc) - EXIT F1-Save Record #XXXXX

DATE [12/05/02] TIME [15:20] ACC NUMBER [3241026] ROUTE ON [3259]
 PED/INJ/KILL[/ /0] NODE NUMBER [58561] ROUTE AT [3271]
 EMS NO [3241026] Crash [0] # of Persons [2] Pstd Spd [0]
 LOC DESC [LYONS RD][SR 806] First Aid [0]
 Taken to [] Trnsp by []

	Act	Spd	Ins	Age	Al/Drg	Res	Sex	Inj	S Equip	Eject
Dru#1	[3]	[0]	[Y]	[24]	[0]	[0]	[2]	[0]	[0]	[0]
Dru#2	[3]	[0]	[Y]	[36]	[0]	[0]	[1]	[0]	[0]	[0]

	Age	Loc	Inj	S Equ	Eject
Pas#1	[0]	[0]	[0]	[0]	[0]
Pas#2	[0]	[0]	[0]	[0]	[0]
Pas#3	[0]	[0]	[0]	[0]	[0]
Pas#4	[0]	[0]	[0]	[0]	[0]
Pas#5	[0]	[0]	[0]	[0]	[0]
Pas#6	[0]	[0]	[0]	[0]	[0]

Figure 2. Data entry screen for old SCARS database, note Node Number and Location Description fields

As mentioned earlier a fair amount of conditioning of the existing crash data was performed to improve the results of geocoding node locations. This effort included search and replace of State Road and US Highway numbers to match the primary names as they exist in the official street name table. The elimination of obvious errors, the segregation of place names, generalized block numbers, mile posts, were set aside to be handled later. The idea being to validate the methodology on what would be the majority of data as expressed in the database. It is expected that further work will need to be performed to street names in the old/existing crash data. Continued enhancements to the Alternate name table will further improve the output representation of crash data as they are described in the crash reports and subsequent entry into the crash database. The creation of a place name table from the County's Landmark GIS will also help in this process.

It should be noted that the geocoding of the nodes is a one-time effort. This is in contrast to the geocoding of crashes every time a desired map output is required. Future crash data will be associated to a complete set of nodes as updated in the regular update cycle of the road centerline GIS. As intersections are created, so will new nodes. Future crashes will have a near 100% match rate against the nodes and mapping will be automated thus allowing the GIS Analyst (and management) to spend more time performing analysis and making decisions on countermeasures with the data rather than simply mapping it and capturing statistics.

PALM BEACH COUNTY TRANSPORTATION DATA MODEL

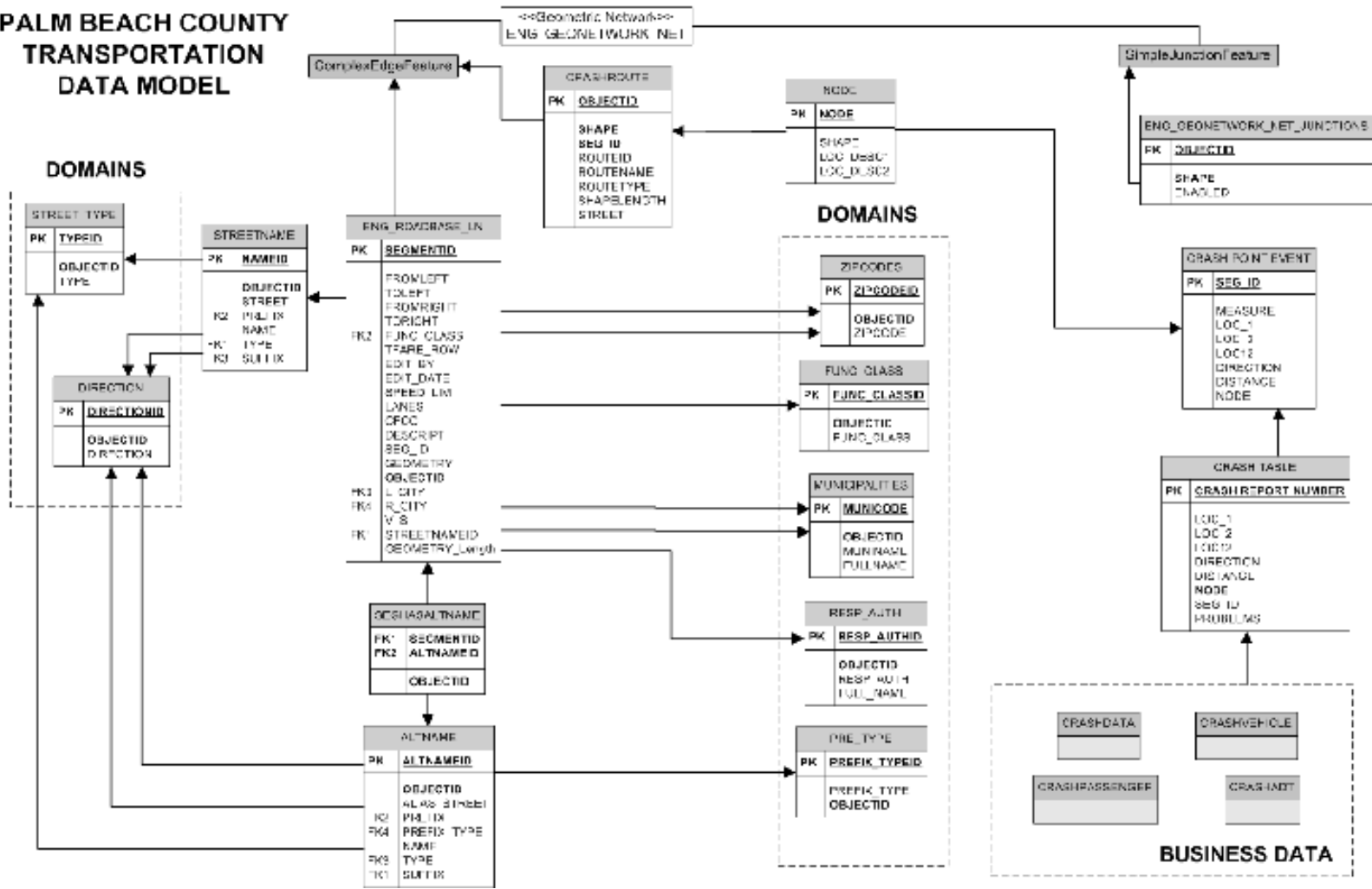


Figure 3. The Palm Beach County Transportation Data Model

Creating the Crash Event Table

As stated earlier the initial effort to match crashes using ESRI's standard geocoding engine produced very poor results due to inconsistencies between street names in the address crash field and the corresponding street names in the route layer. After reviewing alternative options it was decided to explore the possibility of mapping crashes as events along a route system. This method is different from the regular address geocoding and is possible because of the existence of the 'node' for each crash. The 'node' is a number that uniquely identifies each intersection in Palm Beach County. Since most of the crash addresses are reported either at intersections or on a street at a given distance from the intersection, it is logical to use the intersection 'node' number as the starting point for geocoding of crash data.

To map a point event on a route two variables are needed: the route ID on which the event happened and the distance of the crash location from the beginning of the route segment. The distance of the crash from a given intersection is available from the crash report. The unknown variable remains the ID of the route or the street where the accident happened.

The name of the street is part of the crash address but its relation with the street database is not established. Querying the street database for a given street name would typically produce many records since there are many street segments that have the same name. The other reason for not using the attribute queries is the discrepancy of names between crash reports and street database explained above. However, fortunately the crash tables in Palm Beach County do contain the 'node' number that identifies the streets at the intersection related to the crash location. The node information is essential for an alternate method of geocoding presented below:

Initially, the 'node' number for each crash is captured from the crash data. Then the node layer is queried to identify the spatial location of the node at hand. Next, a spatial query (point to line intersection) between the node point and the streets layer is performed to select only the streets intersecting the node.

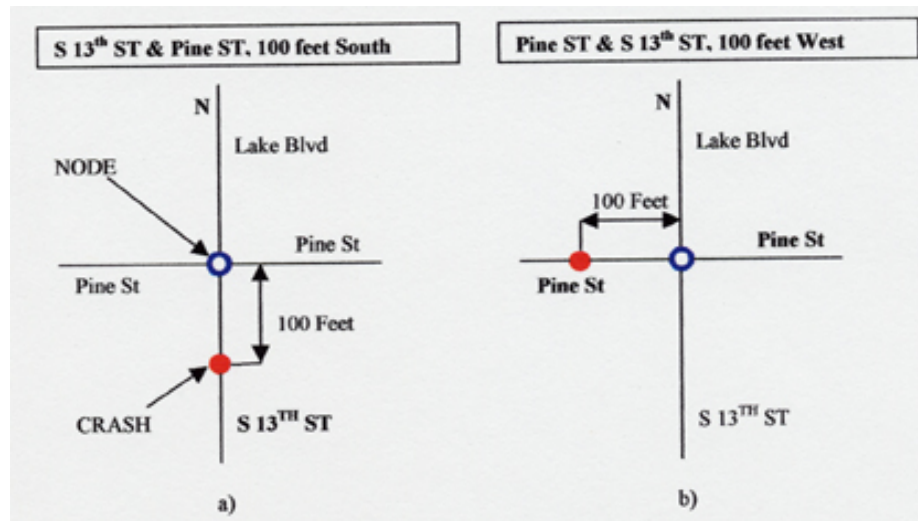


Figure 4. Determining the correct segment

In the example illustrated in Figure 4 all four streets intersecting the node are selected: Pine St (two segments), Lake Blvd and S 13th St. The next step is to select the street associated with the crash. Each street name is compared in turn with the name of the first street in the crash address. In the example S 13th St. is compared to all four streets. In the simplest case the comparison produces only one match i.e. S 13th St. (Figure 4a). However there may be more complex case. In the case of the second address on the right Pine ST and S 13th St. the comparison produces two segments with the same name (Pine St). An additional step is required to identify the correct segment. Here the offset direction is crucial. All the candidate segments are compared by their angle value and the one closer to the offset direction is selected.

In the example the western segment of the Pine St is selected since the address specifies the crash to be located West of S 13th St. (Figure 4b). In cases when the offset distance is zero either one of the candidates can be selected.

The second item required for the crash event table is the 'measure', i.e. the distance of the crash point from the beginning of the segment. By a quick examination it appears that the measurement should simply be equal to the offset distance from the intersection. However that's not always the case because in a linear referencing system the distances are measured from the beginning of the segment. Each street segment stores two nodes named 'From-node', indicating the beginning of the segment and the 'To-node' indicating the end of the segment (Figure 5). When the 'From-node' happens to be at the intersection, then the measure should equal the distance (Figure 5a).

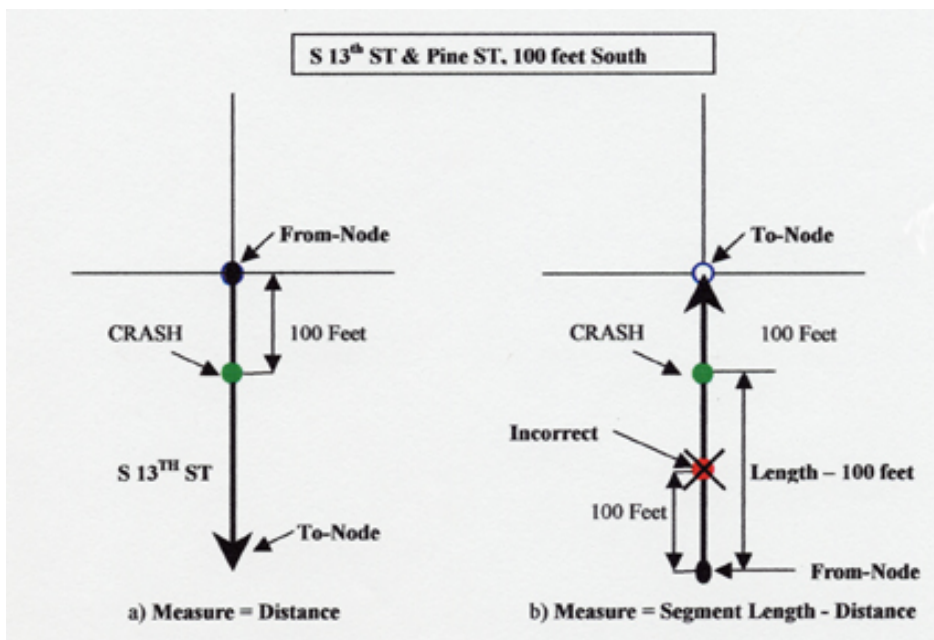


Figure 5. Determining the measure

However, if the From-node is not located at the intersection, using the distance as the measure would produce an incorrect result (the red dot in Figure 5b). The correct measure should be determined by subtracting the distance from the entire length of the segment (Figure 5b). In cases when the offset distance is zero and the To-Node is at the intersection the measure results equal to the length of the segment ($\text{Length} - 0 = \text{Length}$)

Displaying Crash Events

The above algorithm is coded as a series of routines in VBA and is executed by a button in the ArcMap toolbar. The program determines the route ID and the measure for each crash record and writes them in the crash table. The crashes are then displayed on the map as route events using the route streets database and the crash event table. Figure 6 shows the route key field (seg_id) and the measure field (Measure) added and populated in the crash table.

Attributes of crashes_test4						
	SEG ID	DIRECTION	DISTANCE	NODE LOC1	NODE LOC2	MEASURE
	5665		0	JOHN F KENNEDY D	CONGRESS AVE	1744.544299
▶	21084		0	N COUNTRY CL	MILITARY TRL	490.149695
	26069		0	SW 4 ST	SW AVENUE E	0
	9229		0	SW 5 ST	SW AVENUE B PL	216.156794
	56159	W	100	SW 12 ST	SW AVENUE E	100
	46147		0	SW 12 ST	SW AVENUE E	281.762485
	3855	E	50	SW 8 ST	SW AVENUE E	605.623438
	46147		0	SW 12 ST	SW AVENUE E	281.762485

Record: [Navigation icons] 2 [Navigation icons] Show: [All] [Selected] Records (0 out of 1570 Selected.)

Figure 6. Sample crash table with event fields

Results and Discussion

The custom crash mapping tool was used to map a table of 1570 crashes. It resulted in 100% match rate (Figure 7).

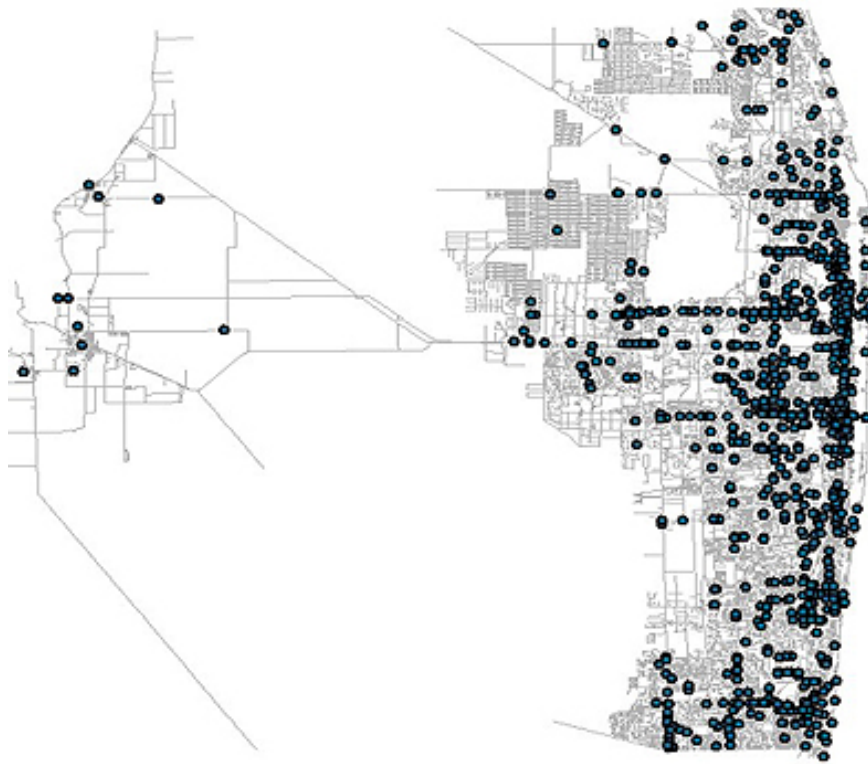


Figure 7. Crashes displayed as route events

Although all the records matched, the matching accuracy is not the same for all of them. Below is a table that shows the break down of matches based on their accuracy.

OID	Status	Cnt	Status	Percent
0	Matched: At Intersection - Invalid offset direction	119		7.6
1	Matched: At Intersection - Offset distance greater then line length	126		8
2	Matched: OK	312		19.9
3	Matched: OK - No name matched but seq_id found by offset direction	567		36.1
4	Matched: OK at Intersection - No name matched but has no offset	446		28.4

Record: [Navigation icons] 3 [Navigation icons] Show: [All] [Selected] Records (0 out of 5 Selected.)

Figure 8. Break down of matched crashes by accuracy level

Group One - 100 percent accuracy. This group consists of three subgroups:

- Crash records that have correct street names and correct definition of the offset distance and direction in the crash report. About 20% of crashes fall in this category (row OID 2).
- Crash records that have discrepancies in street names with the street database but don't have an offset distance. As such they match at the intersection. About 28% of the crashes fall in this category (row OID 4)
- Crash records that have discrepancies in street names with the street database but do have an offset distance. Since the street names don't match the correct street segment is detected by using the offset direction. About 36% of the crashes fall in this category (row OID 3)

Altogether the records with 100 percent accuracy rate consist of about 85% of the crash records.

Group Two – Less than 100 percent accuracy. This group consists of two subgroups:

- a. Crash records that have correct street names and correct definition of the offset direction but the offset distance is greater than the total length of the street segment where the crash happened. This is a result of an incorrect estimation of the law enforcement officer that collected the crash address. These addresses are matched at the intersection. There is a spatial inaccuracy of the size of the offset distance that varies from crash to crash. However since the offset distance itself is an estimation (at times incorrect), by placing the crash at the intersection provides enough spatial accuracy useful for overall spatial crash mapping and analysis. About 8% of crashes fall in this category (row OID 1).
- b. Crash records that have invalid offset direction. E.g. the offset is specified incorrectly to be at a direction where there is no street. These crashes are matched at the intersection as well. About 7% of the crashes fall in this category (row OID 0)

Altogether the records with less than 100 percent accuracy rate consist of about 15% of the crash records. These results show an excellent matching rate with an excellent matching accuracy. Typically in standard address matching using geocoding the match rate is about 75% (to be optimistic) but only a part of this percentage is with 100% accuracy. In the case of the crash data in Palm Beach County the results of matching using address geocoding were at much lower rates.

It should be noted that the tool developed for crash matching will produce results similar to those described above only for the past crash data that contain a number of deficiencies due to the limitation of previous data entry and storage system. The new crash data entry system engages various checks to minimize errors and thus increases accuracy. It's expected that all the new crashes entered in the system in the future should match 100% with an accuracy of 100%.

There were a number of difficulties encountered during the process of the development of the custom crash mapping tool. Most of them are related to the preparation of the data required for crash mapping. Difficulties include the creation of the spatial layer node using standard geocoding techniques when most of the street names didn't match the corresponding street names in the street database. A good lesson learned in this process is the importance of having accurate alternate street names. Alternate street names are a must for any address geocoding process. Last, probably the biggest lesson learned from this experience is the need for coordination with the other departments responsible for the crash address data entry and storage. Integration of GIS street data with the data entry system can reduce at a minimum most of the problems encountered during this process.

Conclusion

Judging based on the achieved results it can be concluded that the custom tool developed to map crashes in Palm Beach County overcomes by a large margin the limitations of the standard address geocoding and produces a very high rate of successful matches. The time spent for the creation of the spatial node layer is well justified by the high rate of matches and the minimization of time for mapping the rest of crashes for the past 10 years and all future crashes. The automation of the process saves time and effort and produces high quality results. Storing the crashes implicitly as events in the County Transportation Data Model has the benefits of better data coordination and management and provides integrated access to data query and analysis in the County enterprise setting.

References

Ensley, David. Programmer/Consultant, SCARS/CARS database application
Orlando, Florida

ESRI and the Transportation Data Model (UNETRANS) websites

<http://support.esri.com/index.cfm?fa=downloads.dataModels.filteredGateway&dmid=14>

<http://www.ncgia.ucsb.edu/vital/unetrans/>

Frank, Robert. Accident Records Section Manager,
Palm Beach County Engineering and Public Works, West Palm Beach, Florida

Steiner, Ruth. Schneider, Richard. Bejleri, Ilir. Wright, Scott
Department of Urban and Regional Planning, University of Florida
GIS Mapping of Pedestrian and Bicycle Crashes, Final Report, Sept. 2002

Author Information

Ilir Bejleri, Ph.D.

Assistant Professor

Department of Urban and Regional Planning, University of Florida

Email: ilirbeu@yahoo.com

Alberto Vazquez

GIS Supervisor

Palm Beach County Engineering and Public Works

Email: avazquez@co.palm-beach.fl.us

Clara DiBella

GIS Support Specialist

Palm Beach County Engineering and Public Works

Email: cdibella@co.palm-beach.fl.us