March, 2003

Georeferencing Historical FARS Accident Data: A Preliminary Report

Sungsoon Hwang, *DePaul University*
Jean-Claude Thill, *University of North Carolina at Charlotte*

Georeferencing Historical FARS Accident Data

A Preliminary Report

March 2003

Julie Hwang
Jean-Claude Thill

{shwang5, jcthill@buffalo.edu}

National Center for Geographic Information and Analysis
University at Buffalo

Amherst NY 14261

Contents

Figures

Tables

I. Introduction

This report documents procedures used to geo-reference[1] data in the FARS database (see the section II for the data description). The scope of the work reported here is limited to New York State and to the period from 1996 to 2001. Thus, it should be noted that running our program beyond this scope (other states or other time periods) may present unexpected errors. The expected audience of this report includes a general public interested in getting geo-referenced historical FARS data (i.e., shape file, arc/info coverage, SDTS) or in learning about implementing non-custom geo-referencing procedures in Geographic Information Systems (GIS).

The rest of this report contains two main parts. The first part (section II, III) describes the source data and reference data. The second part (section IV, V) presents how to geo-reference the source data against reference data. More specifically, section IV presents the basic ideas behind the proposed procedures for geo-referencing the FARS data as well as an overview of the procedures. Section V gives a more detailed description of the procedures. We demonstrate results in the Section VI. Finally, we conclude this report by suggesting possible program extensions for future work.

II. Source Data: FARS

1. Overview

A. FARS Data

The Fatality Analysis Reporting System (FARS) contains data on a census of fatal traffic crashes within the 50 States, the District of Columbia, and Puerto Rico. To be included in FARS, a crash must involve a motor vehicle traveling on a trafficway customarily open to the public and result in the death of a person (occupant of a vehicle or a non-occupant) within 30 days of the crash. FARS has been operational since 1975 and has collected information on over 989,451 motor vehicle fatalities. The system collects information on over 100 different coded data elements that characterize the crash, the vehicle, and the people involved. The specific data elements may be modified slightly at times, in response to users' needs and highway safety emphasis areas. All data elements are reported on four forms:

[1] *Geo-referencing* is defined as *establishing the relationship between page coordinates on a planar map and known real-world coordinates* (Kennedy, 2001).

- The Accident Form asks for information such as the time and location of the crash, the first harmful event, whether it is a hit-and-run crash, whether a school bus was involved, and the number of vehicles and people involved. (More details can be found in the section II-2.)
- The Vehicle and Driver Forms call for data on each crash-involved vehicle and driver. Data include the vehicle type, initial and principal impact points, most harmful event, and drivers' license status.
- The Person Form contains data on each person involved in the crash, including age, gender, role in the crash (driver, passenger, non-motorist), injury severity, and restraint use.

B. How Does FARS work?

NHTSA has a contract with an agency in each state to provide information on fatal crashes. In New York State, the agency in charge is the Department of Motor Vehicles. FARS analysts are state employees who extract crash information from various sources and put it in a standard format. Each FARS analyst attends a formal training program, and also receives on-the-job training. Data on fatal motor vehicle traffic crashes are gathered from the state's own source documents, and are coded on standard FARS forms. The analysts obtain the documents needed to complete the FARS forms, which generally include some or all of the following:

- Police Accident Reports (PARS);
- State vehicle registration files;
- State driver licensing files;
- State Highway Department data;
- Vital Statistics;
- Death certificates;
- Coroner/Medical examiner reports;
- Hospital medical records;
- Emergency medical service reports.

C. FARS Use

The mission of FARS is to make vehicle crash information accessible and useful so that traffic safety can be improved. FARS data are critical to

understanding the characteristics of the environment, trafficway, vehicles, and persons involved in the crash. This national census of motor vehicle fatalities is not available from any other source. FARS is NHTSA's primary database for highway safety analysis. FARS data can be used to answer many questions on the safety of vehicles, drivers, traffic situations, and roadways.

2. Data Fields for Geo-referencing

A. Accident Level Data Fields

Of the four forms composing the FARS, we use the accident level file only since our purpose is to pinpoint the location of accidents. This section describes data fields in the 1999 accident level file. There are slight variations in data elements for other years considered in this study.

In Table 1, we also identify the data fields that may be useful for geo-referencing. The rightmost column of the table labels them. More particularly, RG indicates "relevant to geo-referencing" while RGBI indicates "relevant to geo-referencing, but incomplete." For instance, data field TRAFFICWAY IDENTIFIER is definitely used for geo-referencing purposes, whereas LATITUDE is not likely to be used because most values are left blank[2]. Data fields that are relevant to geo-referencing are marked as filled in the leftmost columns. More detailed description of the codes used to populate these fields will be given in the following sections, which is necessary to introduce our geo-referencing schemes.

[2] NHTSA is actively work with a consultant to provide geocodes (latitude, longitude) for the largest possible number of records for the current year. However, this geocoding effort faces the same obstacle as those described in this report for past years.

ACCIDENT LEVEL DATA FIELDS	TYPE	START	LENGTH	GEORef.
CASE STATE	N	1	2	RG
CITY	N	13	4	RG
COUNTY	N	17	3	RG
TRAFFICWAY IDENTIFIER	A/N	42	20	RG
MILEPOINT	N	62	5	RG
SPECIAL JURISDICTION	N	67	1	RGBI
RELATION TO JUNCTION	N	71	2	RG
LATITUDE	A/N	116	8	RGBI
LONGITUDE	A/N	124	9	RGBI

Table 1. Locational Fields in the FARS Database.

B. Coding Scheme for Fields Relevant to Geo-referencing

(i) State

Values = GSA state codes except for 43, Puerto Rico

This is the state in which the crash occurred. The state in which the vehicle(s) is (are) registered, REG_STAT, is found in the vehicle file.

(ii) City

Values =

Blanks	
0000	Not Applicable
0001-9996	Use GSA Geographical Codes
9997	Other
9999	Unknown

This is the locality where the incident occurred, as reported by the police officer. The locality in question can be an official city, a town, a village, or any other vernacular place name used and recognized locally. GSA geographical codes are somewhat stable. Occasionally one code will be divided into two codes in subsequent years. GSA geographical codes can be downloaded from the GSA Geographic Locator Codes web site at this URL: http://www.gsa.gov/glc.

(iii) County

Values = Blanks

000	Not Applicable
001-996	Use GSA Geographical Codes
997	Other
999	Unknown

GSA geographical codes are somewhat stable. Occasionally one code will be divided into two codes in subsequent years. GSA geographical codes can be downloaded from the GSA Geographic Locator Codes web site at this URL: http://www.gsa.gov/glc.

(iv) Trafficway Identifier, TWAY_ID

Values = Actual Posted Number, Assigned Number, or Common Name (if no posted or assigned number) except:

9999999999	Unknown

(v) Milepoint, MILEPT

Used in 1982 and subsequent years.

Values =

00000	None
Actual to Nearest 0.1 mile (Assumed decimal, e.g., 12345 = 1234.5)	
99999	Unknown

Five digits are always coded.

(vi) Relation to Junction, REL_JUNC

Values for 1991 and later =

01	NON-INTERCHANGE, Non-Junction
02	NON-INTERCHANGE, Intersection
03	NON-INTERCHANGE, Intersection Related
04	NON-INTERCHANGE, Driveway, Alley Access, etc.
05	NON-INTERCHANGE, Entrance/Exit Ramp Related
06	NON-INTERCHANGE, Rail Grade Crossing

07 NON-INTERCHANGE, In Crossover
09 NON-INTERCHANGE, Unknown
10 INTERCHANGE AREA, Intersection
11 INTERCHANGE AREA, Intersection Related
12 INTERCHANGE AREA, Driveway Access
13 INTERCHANGE AREA, Entrance/Exit Ramp Related
14 INTERCHANGE AREA, In Crossover
15 INTERCHANGE AREA, Other location in Interchange
19 INTERCHANGE AREA, Unknown
99 Unknown

Values for 1975 to 1990 =

1 Non-Junction
2 Intersection
3 Intersection Related
4 Interchange Area
5 Driveway, Alley, Access, Etc.
6 Entrance/Exit Ramp (Since 1978)
7 Rail Grade Crossing (Since 1979)
8 In Crossover (Since 1980)
9 Unknown

3. Getting Data

Most up-to-date data are downloadable directly from ftp://ftp.nhtsa.dot.gov/FARS/. Compact disks containing a more complete time series in ASCII format can be ordered from the Bureau of Transportation Statistics web site (http://www.bts.gov) free of charge. A complete user's guide is available at ftp://ftp.nhtsa.dot.gov/FARS/FARS-DOC/USERGUIDE-2002.pdf.

III. Reference Data

1. Data Lists

Geo-referencing procedures require data in reference to which source data (here FARS accident records) can be positioned. We describe here the reference data that we used for geo-referencing accident records of the FARS database. The

selection of reference data is based on two primary considerations. First, we consider how relevant/compatible the data is to the source data, FARS accident records. Second, we compromise between data accessibility and data quality. For instance, TIGER/LINE roads are chosen over other road network data that may be more accurate because TIGER is in the public domain.

Related FARS Field	Organization	Layer Name	Data Source (URL or product) / Metadata Documentation
Tway_id	DOT-FHWA	NHPN	http://www.fhwa.dot.gov/planning/nhpn
			http://www.fhwa.dot.gov/planning/nhpn/docs/metadata.html
	US Census Bureau	TIGER	http://arcdata.esri.com/data/tiger2000/tiger_download.cfm
			http://www.census.gov/geo/www/tiger/rd_2ktiger/tlrdmeta.txt
County	NY State GIS clearinghouse	*Nyshore*	https://www.nysgis.state.ny.us/s_dot/data/state/nyshore.zip (user login required)
			http://www.nysgis.state.ny.us/gis3/data/dot.nybndry.html
City	NY State GIS clearinghouse	*Nybndry*	https://www.nysgis.state.ny.us/s_dot/data/state/nybndry.zip (user login required)
			http://www.nysgis.state.ny.us/gis3/data/dot.nyshore.html
	Caliper	*Ccplacec*	Academic TransCAD® version 3.5d (program CD)
			N/A
	GSA	GSA code table	http://www.gsa.gov/attachments/GSA_PUBLICATIONS/extpub/glcout_1.zip
			N/A

Table 2. Sources of Reference Data

2. Preprocessing

Some preprocessing is necessary to put together the reference data from multiple sources. Major tasks include transforming different coordinate systems into a single unified one, creating a field for joining data tables, and finally assembling the spatial data. In principle, the needs for preprocessing arise from the differences between source data and reference data particularly in the context of a geo-referencing task.

A. Conversion between different coordinate systems

The information about spatial reference (i.e., coordinate system, projection) can be found in each metadata specified in Table 2. We use the geographic coordinate system in NAD83 (North American Datum of 1983) as a single unified coordinate system. Since the data from the NYS GIS clearinghouse are in the UTM system, we convert these data into those in the targeted coordinate system.

In addition, datum difference needs to be configured. *Ccplacec* (census place centroid) from TransCAD® version 3.5d is in NAD27 (North American Datum of 1927) while other dataset is in NAD83. So we convert the datum of *Ccplacec* into NAD83.

B. Relating GSA codes to Geographic layers

To match the GSA geographic codes populated in city and county fields (see Section II-2-B) with geographic data (viz. county/city related reference data such as *nybndry*, *nyshore*, *ccplacec*) that do not contain a GSA code, we need to create a key field for joining. From the example shown in Figure 1 below, we can infer city code 0010 (for geographic data) from the depicted relations. As a result, we can write the 0010 to the new field, say city_gsa in geographic data, which is used to relate a non-spatial FARS locality to a locality recorded in the geographic data.

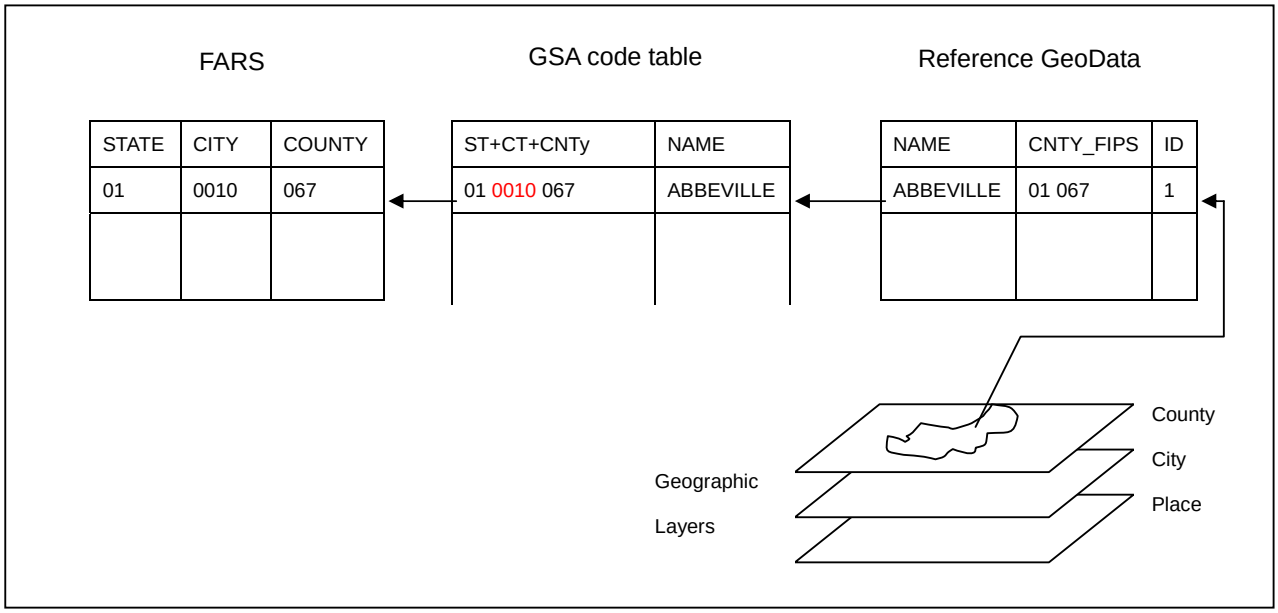


Figure 1. Joining Spatial and Non-spatial Data.

C. Merging Town and City

The data field CITY in FARS accidents is loosely defined as a general place name, but it does not distinguish between different types of zones. One of the reference data, *Nybndry* consists of multiple regions such as county, town, city, village, airport, and park. For example, the Town of Amherst and the City of Buffalo are coded in the same field CITY without regard for the fact that these two entities belong to different classes of places (towns and cities, respectively), whereas the reference data *Nybndry* distinguish them. It may be necessary to preprocess *Nybndry* in order to match two datasets (source attribute CITY and reference *Nybndry*). Consequently, two polygon databases are merged, namely town and city, since it is more efficient to relate the FARS locality to one geographic layer than to two separate layers. This is facilitated by the fact that the two boundaries, town and city, never overlap. For convenience's sake, the merged polygon layer is called PLACE_PL; the other CITY-related point layer, *Ccplaces*, is renamed PLACE_PT.

D. Extracting County Polygons

The COUNTY geographic layer is extracted from *Nybndry*. To summarize, the reference data, *Nybndry* are now divided into following separate layers: COUNTY and PLACE_PL. *Ccplacec* is renamed PLACE_PT. We are going to use these names as aliases of the reference data instead of the original name of data from now on. FARS field names and reference data names are marked in capital letter to contrast.

IV. Geo-referencing Procedures

1. Overview

Geo-referencing the FARS accident records is a matching problem. We attempt to match the source data (non-spatial) to the reference data (spatial) in order to identify the coordinates of the location of an accident. Unfortunately, it is not a custom geocoding[3] (or address matching) procedure as suggested by the attributes relevant to geo-referencing such as TWAY_ID and CITY. More specifically, essential information usually required for an address matching, such as house number and zip code, is missing in the case of local streets. Moreover, CITY is not

[3] It is defined as *The process of identifying the coordinates of a location given its address*. For example, an address can be matched against a TIGER/Line street network to determine the location of a home. Also referred to as address geocoding (Kennedy, 2001).

equivalent to the usual postal areas used for the purpose of mailing. (eg. John James Audubon Pkwy, Buffalo, Erie County, NY instead of 400 John James Audubon Pkwy, Amherst, NY 14228). The case of highways is little more straightforward in terms of completeness of the information given. They are usually coded as highway name (TWAY_ID) with milepoint (MILEPT), which enables the geo-referencing in the specific location unlike the case of local streets.

Given the intrinsic differences between the coding schemes to highways and to local streets, the matching problem is divided into two types. One is to use a linear referencing system[4] (LRS) for highways (Linear Referencing), and the other is to use similarity-based matching (see the later section) for local streets (Local Street Matching). Two different types of matching lead to the different scales in which an accident can be positioned (Figure 2). Linear Referencing identifies the exact point of location, while Local Street Matching identifies the most probable road segments.

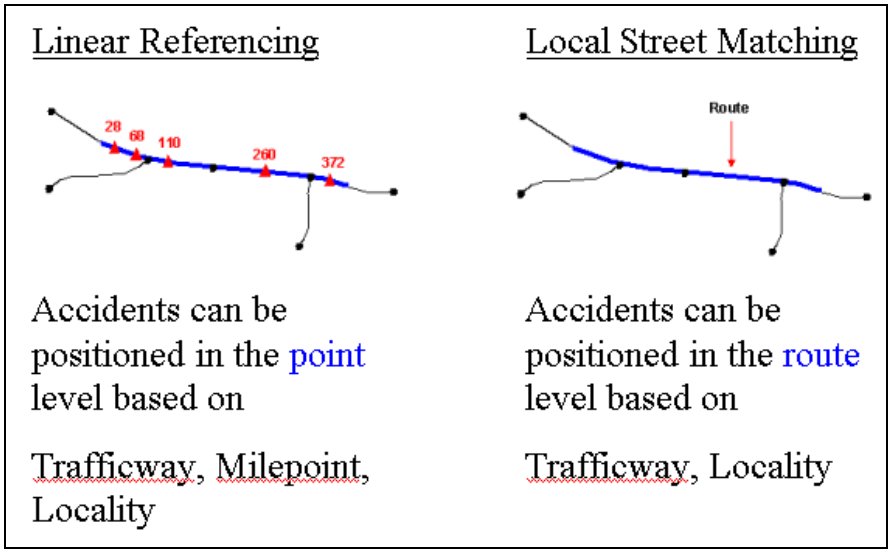


Figure 2. Two Different Types of Geo-referencing

These two different matching procedures are conducted in a serial manner (Figure 3) such that cases that cannot be handled in Linear Referencing can be handled in Local Street Matching. It has been noted that data quality is a critical

[4] Linear referencing locates events along a linear feature with only one parameter (usually known as *measure*) instead of two (such as *latitude/longitude* or *x/y* in Cartesian space). A certain point (ie. accident location) within a linear feature (ie. highway) can be referenced and created dynamically by indicating *measure* (ie. milepoint, milepost) measured from the reference point (ie. intersection between administrative boundaries and highway).

element that affects the reliability of geo-referencing results. Consequently, we adopt a scoring scheme to measure the uncertainty involved in geo-referencing, so that we can either accept or reject the results based on their scores.

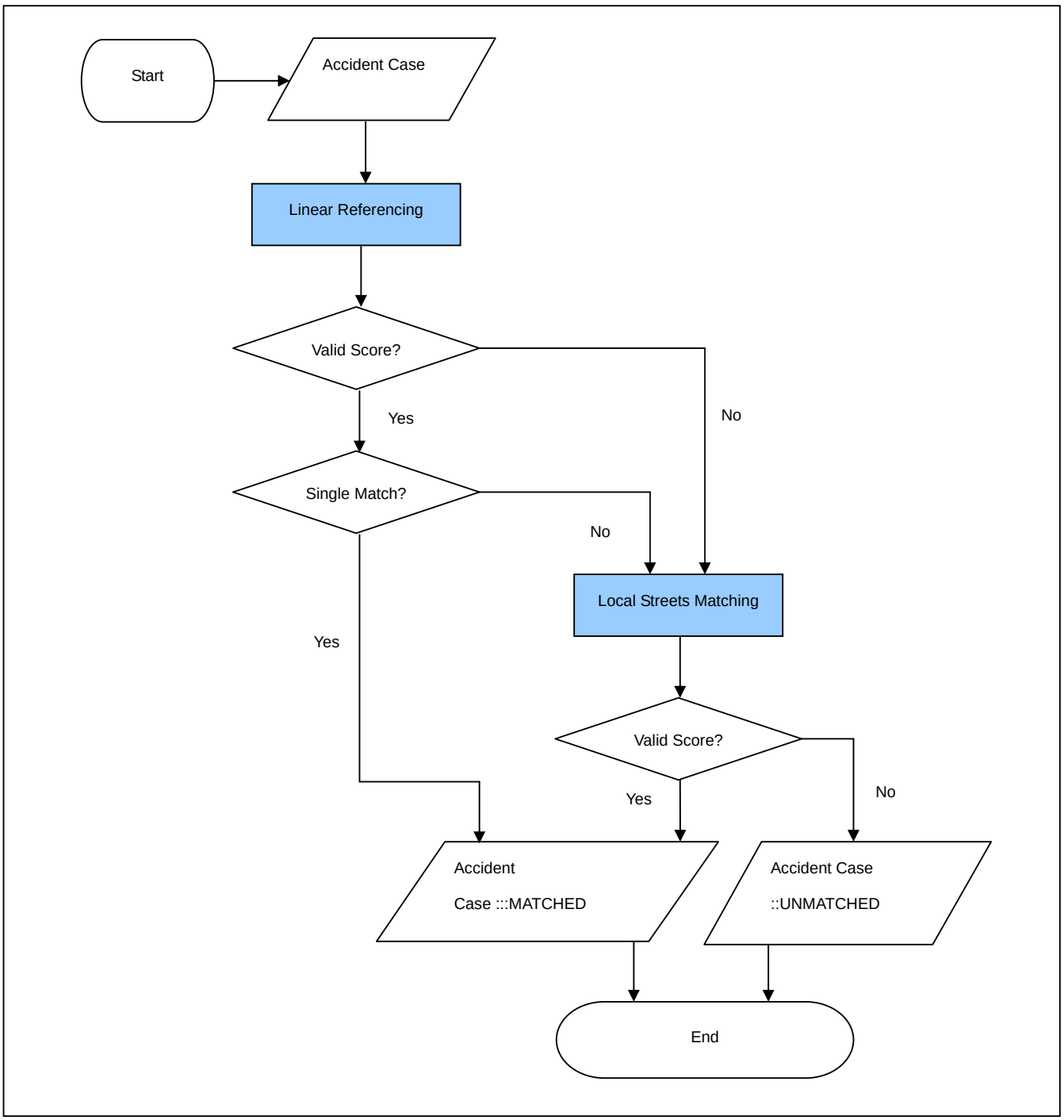


Figure 3. Flow Chart of the Geo-referencing Algorithm

2. What to Match?

Table 3 specifies the reference layers and their attributes against which FARS accident records are matched. A different set of reference data have been chosen depending on the matching types, either Linear Referencing or Local Streets Matching. FARS accident records that are linearly referenceable, that is, contain the complete value of both highway name (TWAY_ID) and measure (MILEPT), are matched against NHPN. PLACE_PT is, if necessary, considered for verifying and pruning multiple highway routes with the same name. Next, FARS records that are screened out from Linear Referencing modules are matched against TIGER/LINE roads. One of PLACE_PL and PLACE_PT is, at all times, used for verifying and pruning road segments. Finally, Scoring criteria (Table3) summarize how scores are computed when matching source and reference data. Generally speaking, scores measure the reliability of the matching, but an individual matching is dependent on features to be matched as shown by different criteria. The complete description of different scoring criteria is deferred to the later section.

Matching Type	Source Data	Reference Data			Scoring Criteria
	FARS Field	Geographic Layer			
		Name	Feature Type	Fields	
Linear Referencing	TWAY_ID	NHPN	Route	SIGN1, SIGN2, SIGN3, LNAME	equivalence
	TWAY_ID	NHPN	Route	INVROUTE	equivalence
	MILEPT	NHPN	Route	BEGMPT, ENDMPT	logical consistency
	CITY	PLACE_PT	Point	CITY_GSA	proximity
	COUNTY	COUNTY	Polygon	COUNTY_FIPS	inclusion
Local Streets Matching	TWAY_ID	TIGER	Line	FEDIRP, FENAME, FETYPE, FEDIRS	equivalence, similarity
		NHPN	Route	LNAME	
	CITY	PLACE_PL	Polygon	CITY_GSA	fuzzyProximity
		PLACE_PT	Point	CITY_GSA	fuzzyProximity
	COUNTY	COUNTY	Polygon	COUNTY_FIPS	inclusion

Table 3. Layers and Attributes Used in FARS Data Matching.

A set of matching attributes is generated as shown in Figure 4 based on a pair of attributes in Table 3. Figure 4 can be read using the following notation: X.Y where

X is defined as table or layer and Y is defined as the field contained in X. For example, FARS.TWAY_ID = NHPN.SIGN1 means that field TWAY_ID in table FARS is matched with field SIGN1 in layer NHPN. Here the notation "≅" is roughly defined as equivalence, similarity, consistency, and fuzzy proximity depending on the case in hand, as suggested by the diversity of scoring criteria listed in Table 3.

■ ***Linear Referencing***

[(FARS.TWAY_ID ≅ NHPN.SIGN1) OR (FARS.TWAY_ID ≅ NHPN.SIGN2) OR (FARS.TWAY_ID ≅ NHPN.SIGN3) OR (FARS.TWAY_ID ≅ NHPN.LNAME)]

AND (NHPN.BEGMPT ≤ FARS.MILEPT ≤ NHPN.ENDMPT)

AND (FARS.CITY ≅ PLACE_PT.CITY_GSA)

AND (FARS.COUNTY = COUNTY.COUNTY_FIPS)

■ ***Local Streets Matching***

[FARS.TWAY_ID ≅ TIGER.(FEDIRP+FENAME+FETYPE+FEDIRS)] OR (FARS.TWAY_ID ≅ NHPN.LNAME)

AND (FARS.CITY ≅ PLACE_PL.CITY_GSA) OR (FARS.CITY ≅ PLACE.PT.CITY_GSA)

AND (FARS.COUNTY = COUNTY.COUNTY_FIPS)

Figure 4. Matching Attributes Used to Geo-reference FARS accidents

3. How to Match?

Here we clarify on two major modules, Linear Referencing and Local Streets Matching as shown in Figure 3.

A. Linear Referencing Modules

Linear Referencing modules consist of the following three steps:

■ *Step (i): Trafficway Identifier Match*

Select NHPN routes where any value of signing-related fields are matched[5] to TWAY_ID and passed through COUNTY.

(i)-1. If match (either multiple or single) found, go to step (ii).

(i)-2. Otherwise, select NHPN routes where the value of LNAME is matched to TWAY_ID and passed through COUNTY.

(i)-2-1. If match (either multiple or single) found, go to step (iii).

[5] It is not an equality match, but rather matching rules are required to account for incompatibility between two matching sets. For example, US-10 in FARS is equivalent to U10 in NHPN according to naming conventions.

(i)-2-2. Otherwise, go to Local Streets Matching modules.

■ *Step (ii): Milepoint Match*

Check if the value of MILEPT is consistent in reference to the measures[6] range of the selected NHPN route. (i.e., startMeasure <= MILEPT/10 <= endMeasure)

(ii)-1. If MILEPT turns out to be consistent, and

(ii)-1-1. If single route found, the selection is reserved for linear referencing.

(ii)-1-2. If multiple routes found, go to step (iii).

(ii)-2. Otherwise, go to Local Streets Matching modules.

■ *Step (iii): Locality Match*

Prune multiple routes based on their proximity[7] to locality of the accident record in hand.

(iii)-1. If pruned, the selection is reserved for linear referencing.

(iii)-2. Otherwise, go to Local Streets Matching modules.

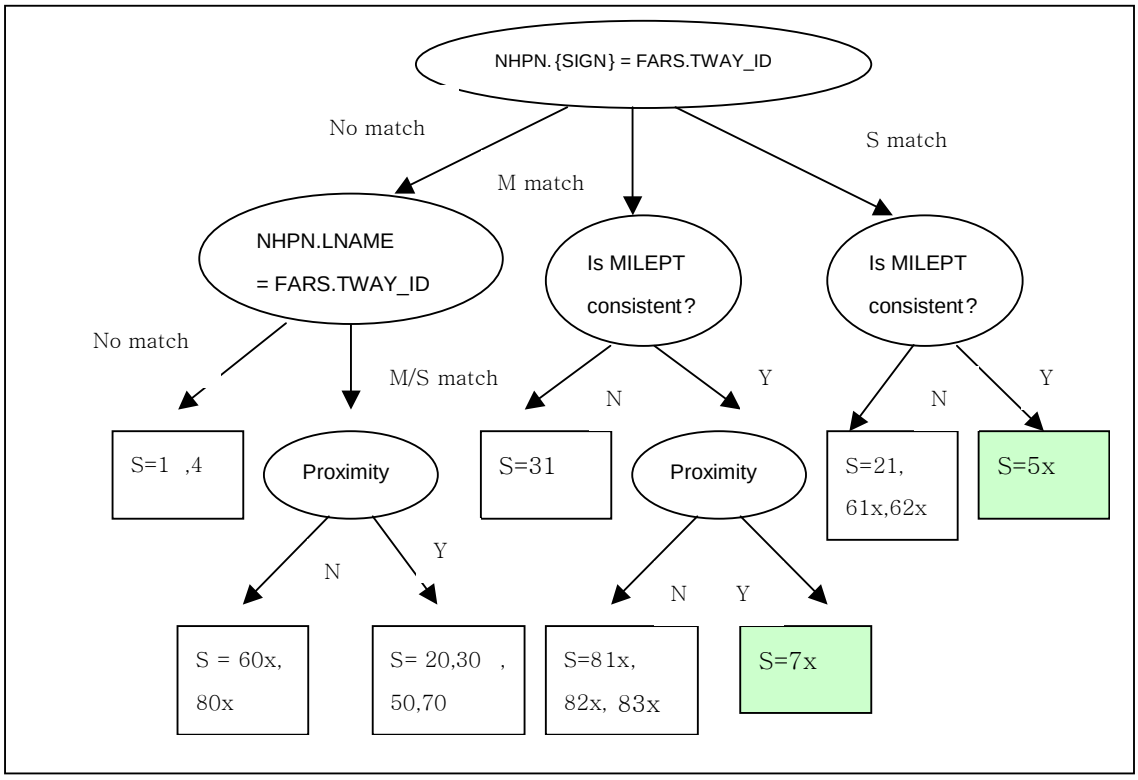


Figure 5. Decision Tree Used in the Linear Referencing Modules

[6] Each route is associated with a measurement system, a linear method consisting of a starting value and ending value along the route.

[7] Select nearest one among multiple routes

The outputs of these three steps are either (a) linearly referenced since they are believed to be eligible for linear referencing (marked as filled rectangle in Figure 5) or (b) sent to Local Streets Matching modules since they are not believed to be eligible for linear referencing, but rather suited for relaxed matching. These steps can be represented as a decision tree shown in Figure 5. Decision nodes represented as a circle correspond to the condition to test, while a decision leaf marked as a rectangle represents the results in the form of score. The scores are assigned in a way that they can track the process to check for logical consistency. Only cases whose scores start with 5 and 7, which are called 'valid score' here, are eligible for linear referencing. Others are sent to the Local Streets Matching module, which will be described in the next section.

Even though we screen out the data, we find that there are very few instances whose relation between FARS and NHPN is not one-to-one when linearly referenced. Such a problem, presumably, is attributable to the error in NHPN data. Only the FARS case with one-to-one relation to NHPN is accepted. Figure 6 illustrates such a case. Case number 361104(year '98) whose milepoint is 2.7 is linearly referenced to two locations, which is considered to be ambiguous.

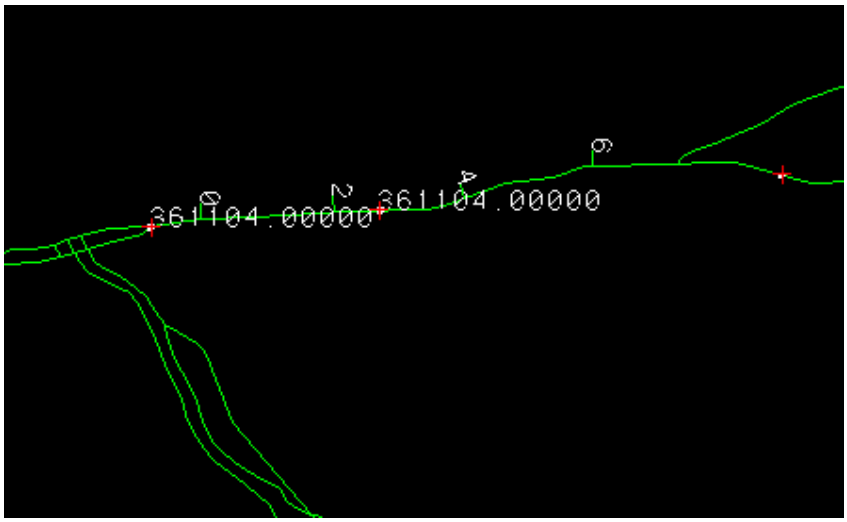


Figure 6. Example of Ambiguous Linear Referencing Result

B. Local Streets Matching Module

When it comes to Local Streets Matching, the task is to identify the road segments on which an accident is most likely to occur given *imperfect* information. For example, naming a road is highly variable (e.g. it could be misspelled, abbreviated, and could have an alternative name), and locality is not well defined

(e.g. it could refer to a city, town, village, place name, and could involve the perception gap between *fiat* boundary and mental representation). The more uncertainty is involved, the less likelihood that an exact match occurs. However, it is, at least, possible to increase a match rate by relaxing the matching criteria, from an equality matching (i.e. matching based on whether or not they are the same) to a *similarity* matching (i.e. matching based on how similar they are). A similarity measure is a very useful concept along this line. The similarity between source and target can be used to compute the quality of matching, and then we can determine whether or not to accept the matching based on that measure.

It is necessary to identify feature vectors in order to define the similarity measures. Figure 7 presents the feature vectors in a tree-like notation. Two important features are identified: One is a trafficway identifier (TWAY_ID) and the other is a locality (CITY, COUNTY). For simplicity, we use the term *trafficway* for the former and *locality* for the latter. The difference between solid and dashed line in Figure 7, is the relation between super- and sub-elements. For instance, *Trafficway* is either *Highway* OR *Local* whereas *Local* is the combination of Direction Prefix, Street Name, Street Type, AND Direction Suffix. Similarly, *Place_PL* consists of two kinds of spatial relations, either *In* or *Near*, where *In* indicates a crisp relation between road and city (i.e. a road is *in* the boundary of city) while *Near* indicates a fuzzy one (i.e. a road is *near* the boundary of city).

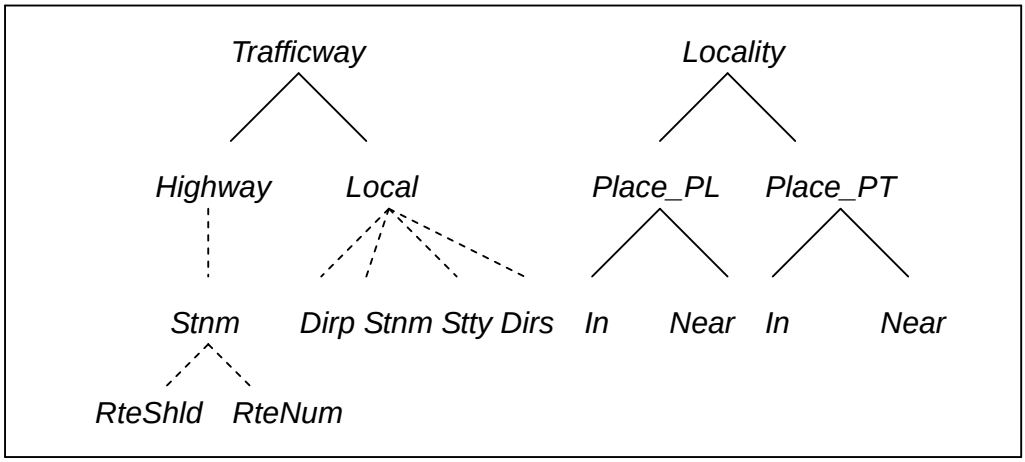


Figure 7. Feature Vectors of Similarity Measures Used in Local Streets Matching Modules

Similarity measures of *trafficway* are computed based on the rule base that is selectively shown in Table 4. The rule base takes into account (a) the incompatibility between source and target strings (e.g. naming convention varies with data), (b)

typographic errors commonly found in source strings (i.e. misspelling, abbreviation, truncation), (c) commonly made mistake in identifying *trafficway* (e.g. not sure about route shield like county, state, US route in contrast to route number), and (d) the thesaurus of source strings (e.g. Most of major roads have alternative names – Niagara Falls Blvd is also known as 62). In Table 4, the fourth column, MATCHTYPE specifies the condition to match and SCORE is assigned accordingly. The more relaxed the condition to match is, the lower the score is. The two columns on the right show examples of FARS and TIGER that turn out to be matched according to the specified rules.

ReferTo	ROADCLASS	SUBSTRING	MATCHTYPE	SCORE	FARS.TWAY_ID	TIGER.StreetName
TIGER	Highway	Rtsld + Rtnum	ExactMatch	1.0	I-95	I-95
		Rtsld + Rtnum	Alias	1.0	SR-35	State Highway 35
		Rtsld + Rtnum	SR=US	0.8	SR-44	United States Highway 44
		Rtnum	RouteNumber	0.7	US-231	Route 231
	Local Street	StreetName	ExactMatch	1.0	WILLIAMS	Williams Dr
		StreetName	IgnoreSpace	0.9	FDR DR	F D R Dr
		StreetName	Alias	0.9	FOURTH ST	4th St
		StreetName	Soundex	0.8	MESSENGER RD	Messinger St
		StreetName	Abbreviation	0.7	MDLNECK RD	Middle Neck Rd
		StreetName	Similarity	string_similarity	MEADOWBKPK	Meadowbrook Pkwy
NHPN	Both?	StreetName	AlternateName	0.7	SR-347	Alexander Ave

Table 4. Selective Rules Used to Define Similarity Measures of Trafficway

Before we describe how to define similarity measures of *locality*, we need to define *fuzzy proximity,* which can be thought of as the relaxed version of crisp proximity in GIS. The rational is that locality is not necessarily perceived in a crisp manner because (a) the boundary is indeterminate in some cases, and (b) its mental representation (or perception) varies by an agent even if the boundary is known to be determinate. As shown in Figure 7, there are two datasets against which source datasets are matched. One is polygon boundary data (Place_PL), and the other is presumably a centroid (otherwise, the location of annotation in source map) with indeterminate boundary (Place_PT). They are different not only in their feature types, but also their contents. Place_PL contains more commonly used place names while Place_PT covers less frequently used names (e.g. Buffalo vs. Snyder).

As a part of defining *fuzzy proximity, vague regions* such as locality are

broken down into two parts. The determinate part of vague regions, called *kernel*, is delineated by the boundary for Place_PL, and Thiessen polygon of centroid for Place_PT. The vague part of vague regions is approximated up to the second lag boundaries surrounded by the determinate part. This division can be thought of an egg-yolk representation. The fuzzy proximity membership, *f*(x,y) is defined as follows:

$$f(x,y) = \begin{cases} 1 & \text{if } x \text{ is IN } y \\ [0.5, 1] & \text{elseif } x \text{ is IN lag1}(y) \\ [0, 0.5] & \text{elseif } x \text{ is IN lag2}(y) \end{cases}$$

where x= {x∈X|X is a set of road segments}, y= {y∈Y|Y is the determinate parts of vague regions}

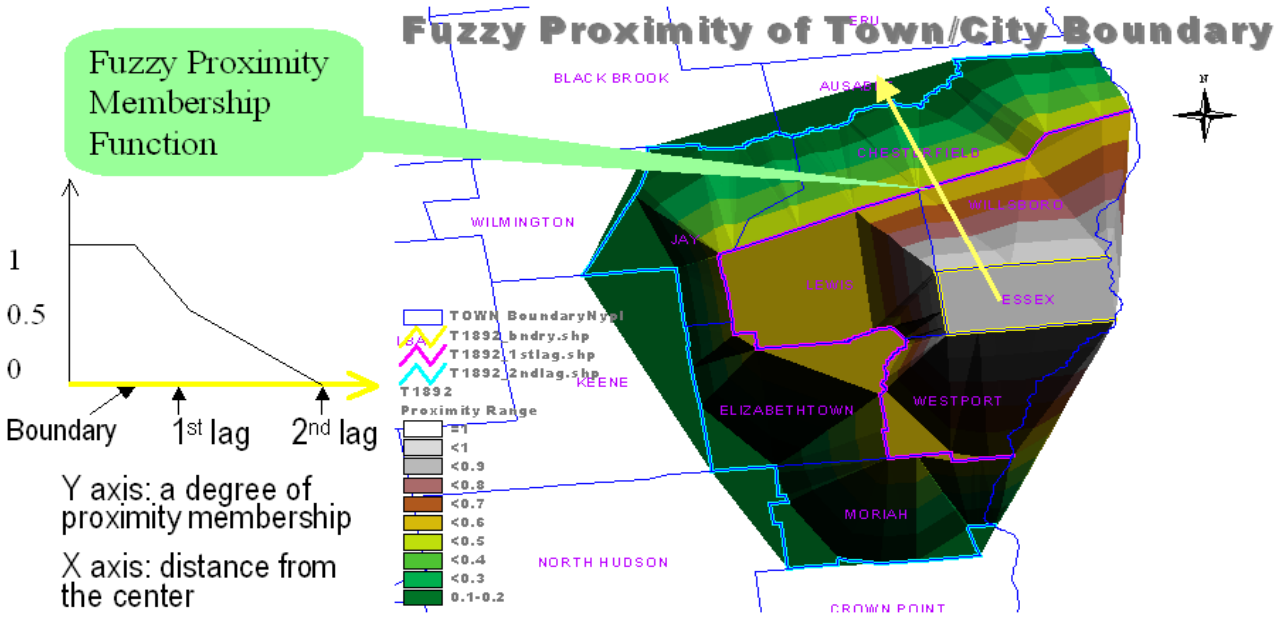
[a,b] is defined as a real number that is linearly interpolated between a and b,

lag1(y) is a set of polygons surrounded by y in the first lag,

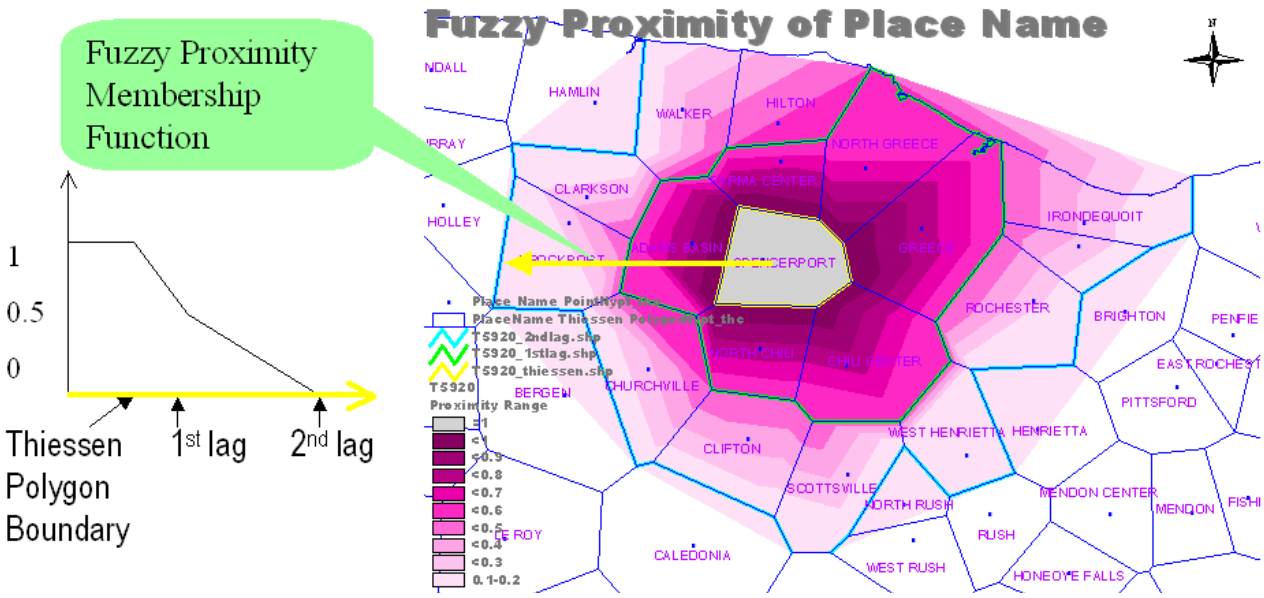
lag2(y) is a set of polygons surrounded by y in the second lag.

Figure 8. Definition: Fuzzy Proximity Membership

Figure 9 demonstrates how similarity measures of locality are computed based on fuzzy proximity defined above. The difference between Place_PL and Place_PT is in the boundary.



(a) Locality Similarity Measures of Place_PL



(b) Locality Similarity Measures of Place_PT

Figure 9. Similarity Measures of Locality based on Fuzzy Proximity

The similarity measure between FARS and TIGER is the average of two similarity measures of *trafficway* and *locality* (Figure 10). They are equally weighted such that both features can be compensated with the same weight. Road segments with the maximum score are selected as best candidates.

$$sim = \text{Similarity(FARS,TIGER)} = 1/2*(1*t + 1*l)$$

where t, l are the similarity measure of *trafficway*, and *locality* respectively

$t = \{\in \mathscr{R} | 0 \leq t \leq 1\}$, $l = \{\in \mathscr{R} | 0 \leq l \leq 1\}$ ($\mathscr{R}$ indicates a real number)

Figure 10. Similarity Measures in Local Streets Matching Modules

The best candidates are dissolved into the *route(s)* based on their attributes. The results can be either a single route or multiple routes. It is obvious that a single route can claim more certainty than multiple routes. Suppose matching a FARS accident record whose TWAY_ID is “bayridge” turns out three routes, “Bay Ridge Ave”, “Bay Ridge Pky”, and “Bay Ridge Pl”. In that case, it cannot be simply decided which one is better. The incomplete input (i.e. street type is missing) leads to an ambiguous output. Given that, we redefine the similarity measure in a way that is

weighted by a certain measure of ambiguity. A new similarity measure is computed as follows:

$$vsim = sim*\boldsymbol{w} = sim*[1/n*\sum_{i=1}(|p_i\text{-}1|/q_i)]$$

where i is the index of elements, $\{\in \text{TWAY\_ID} | \ i = dirp, stnm, stty, dirs\}$,
n is the number of elements with incomplete values where $n \neq 0$
p_i is 1 if i is complete, and 0 otherwise,
q_i is the number of routes that yield when dissolved by i

Figure 11. Verified Similarity Measures in Local Streets Matching Modules

To illustrate how the new similarity measure is computed, suppose similarity measure denoted by *sim* in Figure 10 is 0.8. In the same example above, *n* is 3 because dirp, stty and dirs have incomplete values. *p* can be denoted as 0100 in the order of four elements, that is, dirp, stnm, stty, and dirs since only stnm has a complete value. Similarly, *q* can be denoted 1131 because there are three routes resulted from dissolving by stty while other elements result in a single route. The new similarity measure is 0.8*[1/3*(1/1+0/1+1/3+1/1)] = 0.8*0.78 = 0.624. The multiple routes associated with incomplete input are penalized. Finally, we check if best candidates with the new similarity measure are above the preset threshold to qualify for being accepted.

4. Score Computation

Scores are used to determine whether or not a FARS record is suitable for linear referencing in the case of Linear Referencing. For Local Streets Matching, scores are used to rank candidate matches so that we can select the best one, provided that it is above a preset threshold.

There are some differences to be noted between the scores for Linear Referencing and Local Streets Matching. The score for Linear Referencing is binary, which means that it determines whether or not to accept the record for linear referencing. It indicates the type of treatment of each FARS record. On the other hand, the score for Local Streets Matching is set on an ordinal scale, so that the higher the assigned score is, the more the match is likely to be acceptable. It measures the degree of reliability of each record when it comes to matching against reference data. It is expected that the former is stricter than the later. The former invokes a binary decision – only an unambiguous case is chosen. It makes sense

because the Linear Referencing modules screen out the data with invalid score, and these data are handled in Local Streets Matching modules where other matching approaches are tested.

A. Linear Referencing Scores

In this section, we present how scores are assigned to accident records in the Linear Referencing modules by converting the decision tree in Figure 5 into decision rules. Figure 12 enumerates conditions for each score in the order of valid (accepted for linear referencing) and invalid one (rejected). Each valid score indicates the condition satisfied to be accepted. The conditions with invalid scores do not meet the requirements of accuracy measures such as completeness, logical consistency, equivalence, and geographic proximity.

■ *Valid Scores*

If TWAY_ID has a single match to NHPN.{SIGN}, and MILEPT is consistent, score = 51, 52,53

If TWAY_ID has multiple matches to NHPN.{SIGN}, and MILEPT is consistent, and CITY can find nearest one among multiple candidates, score = 71,72,73

■ *Invalid Scores*

If TWAY_ID has a single match to NHPN.{SIGN}, and MILEPT is not consistent, score = 21, 611, 612, 613, 621, 622, 623

If TWAY_ID has multiple matches to NHPN.{SIGN}, and MILEPT is not consistent, score = 31

If TWAY_ID has multiple matches to NHPN.(SIGN), and MILEPT is consistent, and CITY cannot find nearest one among multiple candidates, score = 811, 812, 813, 821, 822, 823, 831, 832, 833

If TWAY_ID has no match to NHPN.{SIGN}, TWAY_ID has some matches to NHPN.{LNAME}, and CITY can find nearest one among single(multiple) candidate(s), score = 20, 30 , 50, 70

If TWAY_ID has no match to NHPN.{SIGN}, TWAY_ID has some matches to NHPN.{LNAME}, and CITY cannot find nearest one among single(multiple) candidate(s), score = 601, 602, 603, 801, 802, 803

If TWAY_ID has no match to NHPN.{SIGN}, TWAY_ID has no match to NHPN.{LNAME}, score = 1, 4

Figure 12. Linear Referencing Scores

B. Local Streets Matching Scores

As described earlier, the score in Local Streets Matching modules is computed as the weighted average of similarity measures of *trafficway* and *locality*. *Trafficway* similarity measures are computed by string matching rules. *Locality* similarity measures are the output of fuzzy proximity function. Figure 13 presents how scores are assigned to accident records.

■ *Valid Scores*

vsim = *w***sim* > 0.7

■ *Invalid Scores*

vsim = *w***sim* ≤ 0.7

where the definition of *sim* and *vsim* is given in Figure 10 and Figure 11 respectively

set *t* = Trafficway_SimMeas; set *l* = Locality_SimMeas

■ *Trafficway_SimMeas*

```
If TWAY_ID is Highway, go to Highway()
ElseIf TWAY_ID is Local, go to Local()
Else, go to NotSure()
Highway() {
        If TWAY_ID = TIGER.FENAME, Trafficway_SimMeas = 1
        ElseIf Alias(TWAY_ID) = TIGER.FENAME, Trafficway_SimMeas = 1
        ElseIf RteNum(TWAY_ID) = RteNum(TIGER.FENAME) {
            If RteShld(TWAY_ID) = 'SR' and Alias(RteShld(TIGER.FENAME)) = 'US',
                Trafficway_SimMeas = 0.8
            Else, Trafficway_SimMeas = 0.7}
        ElseIf RteNumWithoutSuffix(TWAY_ID) = RteNumWithoutSuffix(TIGER.FENAME),
                Trafficway_SimMeas = 0.5
        ElseIf TWAY_ID = AlternateName(TIGER.FENAME), Trafficway_SimMeas = 0.7
}
Local() {
        Parse TWAY_ID into DirPrefix, StreetName, StreetType, and DirSuffix
        Set a = 0; b = 0; c = 0; d = 0
        Set dirp_score = 0; stnm_score = 0; stty_score = 0; dirs_score = 0
        If DirPrefix is NotNull, set a = 1 and go to GetDirPrefixScore()
        If StreetName is NotNull, set b = 1 and go to GetStNameScore()
        If StreetType is NotNull, set c = 1 and go to GetStTypeScore()
        If DirSuffix is NotNull, set d = 1 and go to GetDirSuffixScore()
        GetDirPrefixScore() { If DirPrefix(TWAY_ID) = TIGER.FEDIRP, dirp_score = 1}
        GetDirSuffixScore() { If DirSuffix(TWAY_ID) = TIGER.FEDIRS, dirs_score = 1 }
        GetStTypeScore() {If StType(TWAY_ID) = TIGER.FETYPE, stty_score = 1 }
```

GetStNameScore() {

If StName(TWAY_ID) = TIGER.FENAME, stnm_score = 1

ElseIf IgnoreSpace(TWAY_ID) = IgnoreSpace(TIGER.FENAME), stnm_score = 0.9

ElseIf Soundex(TWAY_ID) = Alias(TIGER.FENAME), stnm_score = 0.8

ElseIf Soundex(TWAY_ID) = Soundex(TIGER.FENAME), stnm_score = 0.8

ElseIf Abbreviation(TWAY_ID) = Abbreviation(TIGER.FENAME), stnm_score = 0.7

Else, stnm_score = StringSimilarity(FARS.TWAY_ID, TIGER.FENAME) }

Trafficway_SimMeas = 1/10*(1*a*dirp_score + 6*b*stnm_score + 2*c*stty_score + 1*d*dirs_score)

}

NotSure() {

If RteNum(TWAY_ID) = RteNum(TIGER.FENAME), Trafficway_SimMeas = 0.7

ElseIf TWAY_ID = AlternateName(TIGER.FENAME), Trafficway_SimMeas = 0.7

Else, go to Local()

}

■ *Locality_SimMeas*

Locality_SimMeas = *f*(x,y) where *f*(x,y) is fuzzy proximity membership defined in Figure 8

Figure 13. Local Streets Matching Scores

V. Geo-referencing Algorithms

1. Pseudo Code of Linear Referencing Modules

See Appendix A.1.

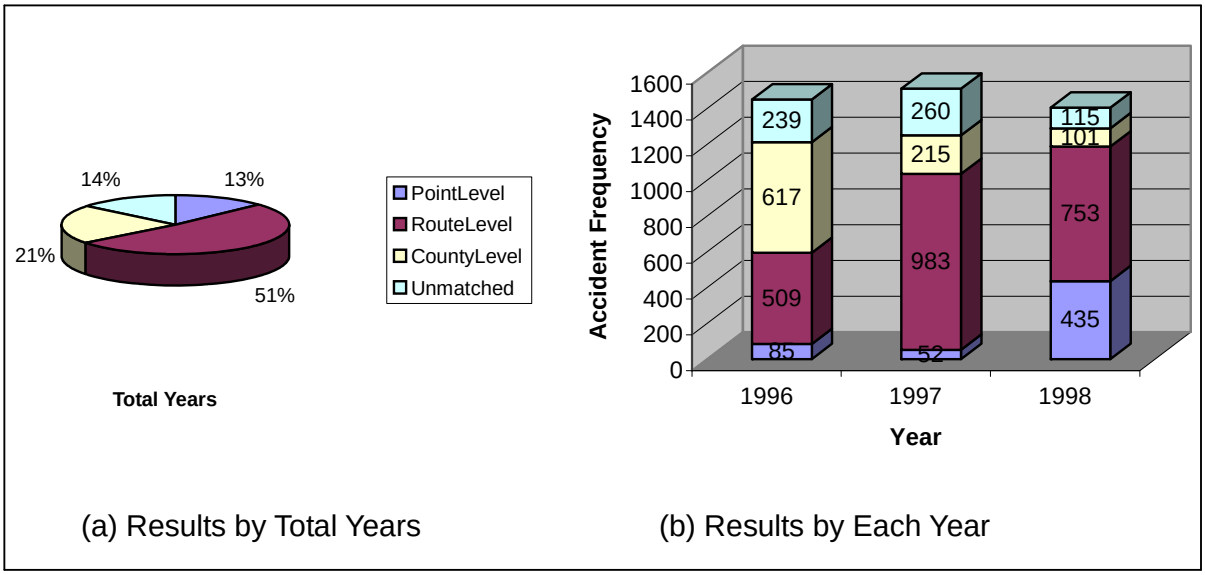
2. Pseudo Code of Local Streets Matching Modules

See Appendix A.2.

VI. Results

We tested our program on the FARS database covering the period 1996 to 1998 within New York State. We picked these years because of heterogeneous data quality among these years. The data quality of FARS since the year 1998 has significantly improved relative to the previous years. The results are presented in

different levels of positional accuracy due to the different matching criteria. (a) Point level results from unambiguous linear referencing; (b) Route level results from the local streets matching with a score greater than 0.7; (c) County level results from the local street matching with a score less or equal to 0.7.



(a) Results by Total Years (b) Results by Each Year

Figure 14. Results of Geo-referencing FARS '96 – '98 Presented By Different Levels of Positional Accuracy

Figure 14 (a) shows the results aggregated by three years. 13% of records are linearly referenced, that is positioned in the exact point of location. The rest are handled in Local Streets Modules. 51% was positioned in the route level with a verified accuracy. Figure 14 (b) shows how the data quality of FARS databases, specifically location-related attributes has changed over time. Significantly more accidents in the year 1998 are reported using linear referencing with complete values compared to the previous years. Higher percentage of county level in the year 1996 is due to the missing value in locality-related attributes.

The results of Linear Referencing Modules can be roughly broken down into three categories according to scores assigned. First, the exact point of location is identified in the case where scores are 51, 52, 71, 72, and 73. Second, the route is identified, but exact point along the route is not identifiable due to the inconsistent value of milepoint in the case where scores start with 2, 3, 6, and 8. Third, nothing can be identified in the case where scores are 1 or 4. These three positional categories are graphically presented in Figure 15. The graph suggests that data quality has improved over test periods as shown from the increasing portion of higher level of positional categories.

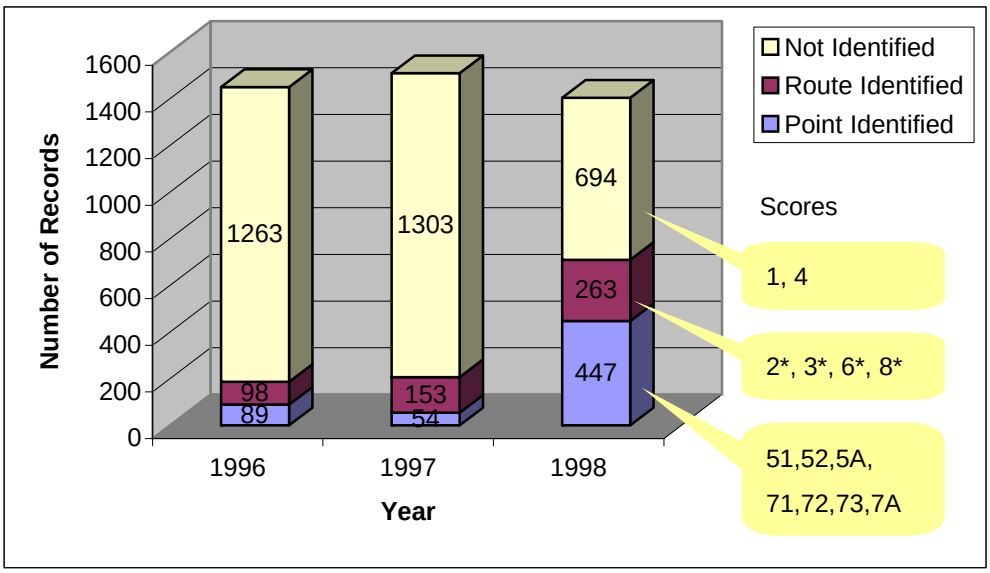


Figure 15. The Results of Linear Referencing Modules by Three Positional Categories Given by Scores Assigned

Local Streets Matching modules were capable of assigning scores to 1126, 1198, and 854 cases out of 1450, 1510, and 1404 respectively in the year 1996, 1997, and 1998. The matching results can be classified into three categories such as poor, right, and exact matching according to scores assigned as shown in Figure 16. The score implies the verified similarity measures (Figure 11). Right matching accounts for 26% on average, thus allowing for the increase in a match rate.

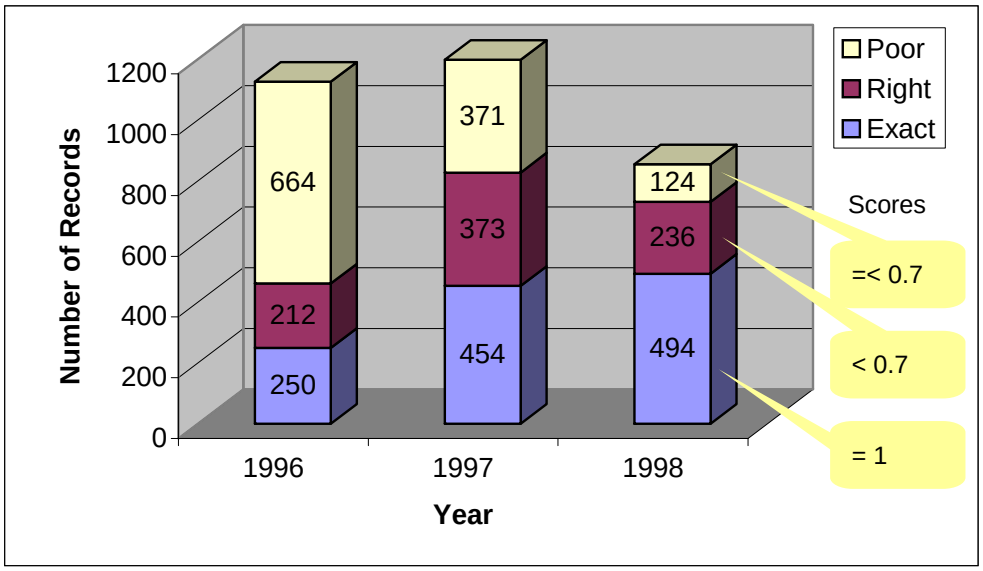


Figure 16. The Results of Local Streets Matching Modules by Matching Quality Classification Given by Verified Similarity Measures.

Finally the geocoded FARS accidents in New York State are presented as a map. In Figure 17, a random point of location is assigned along the identified route(s) within the positional tolerance in the case of RouteLevel and CountyLevel to visualize on the equal footing (point),

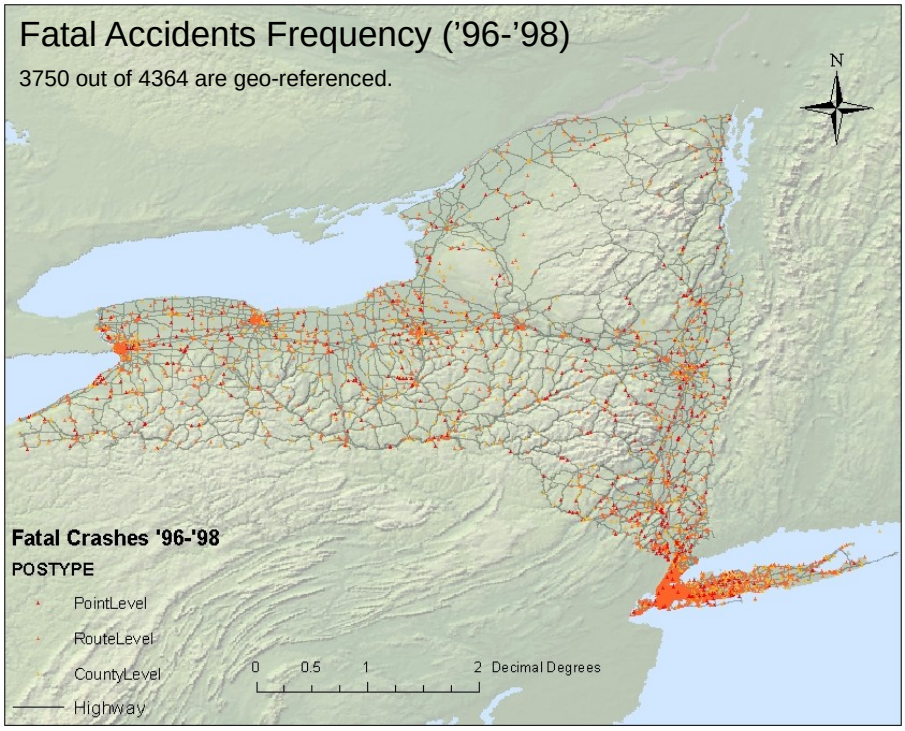


Figure 17. Map of Geocoded FARS Accidents Frequency in New York State

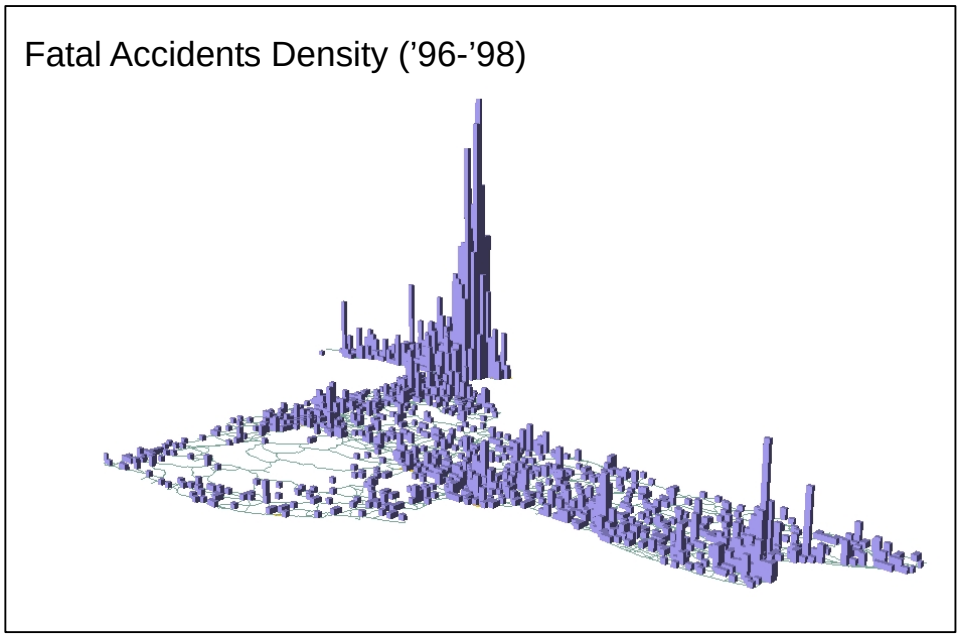


Figure 18. Map of Geocoded FARS Accidents Density in New York State

In Figure 18, the number of accidents that occurred within 0.05 by 0.04 (in decimal degree) grid is accumulated as the height over the test period.

VII. Conclusion

We match the FARS accidents against geographic data mostly open to general public such as NHPN, TIGER, boundary (county/city), and place. Our approach to geo-referencing the FARS accidents allows for uncertainty handling faced with significant amount of imperfect information.

The procedures start with the Linear Referencing modules, from which the data rejected due to the high degree of uncertainty are sent to Local Streets Matching modules, where rather approximate matches are performed. We use different matching criteria as well as different sets of reference data depending on available attributes in FARS accidents. In the case when an accident record contains highway name and milepoint, we geo-reference against NHPN using a linear referencing system. Accidents on local streets are matched against TIGER based on the similarity measures.

There are several tasks for the future extension of this system. First, we will proceed to evaluate our procedures with an independent source of higher accuracy. Obtaining such a dataset in recent years is under way. Second, the scope will be expanded to other states and periods. Third, clustering algorithms of geo-referenced accidents will be developed as a way of effectively presenting the spatial pattern of accidents.

References

Heather Kennedy (ed.), 2001, *ESRI Press Dictionary of GIS Terminology*, ESRI press

FARS 1996 Coding and Validation Manual, 1995, National Center for Statistics and Analysis, National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C.