OKLAHOMA TRANSPORTATION CENTER

ECONOMIC ENHANCEMENT THROUGH INFRASTRUCTURE STEWARDSHIP

FATALITY ANALYSIS REPORTING SYSTEM AND ROADWAY INVENTORY CORRELATION

RONALD D. BARNES, PH.D.
JOSEPH P. HAVLICEK, PH.D.

OTCREOS11.1-06-F

Oklahoma Transportation Center
2601 Liberty Parkway, Suite 110
Midwest City, Oklahoma 73110

Phone: 405.732.6580
Fax: 405.732.6586
www.oktc.org

TECHNICAL REPORT DOCUMENTATION PAGE

1. REPORT NO. OTCREOS11.1-06-F	2. GOVERNMENT ACCESSION NO.	3. RECIPIENTS CATALOG NO.
4. TITLE AND SUBTITLE Fatality Analysis Reporting System and Roadway Inventory Correlation		5. REPORT DATE January 6, 2013 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Ronald D. Barnes and Joseph P. Havlicek		8. PERFORMING ORGANIZATION REPORT
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Electrical and Computer Engineering The University of Oklahoma Norman, OK 73019		10. WORK UNIT NO. 11. CONTRACT OR GRANT NO. DTRT06-G-0016
12. SPONSORING AGENCY NAME AND ADDRESS Oklahoma Transportation Center (Fiscal) 201 ATRC, Stillwater, OK 74078 (Technical) 2601 Liberty Parkway, Suite 110 Midwest City, OK 73110		13. TYPE OF REPORT AND PERIOD COVERED Final October 1, 2011 – December 31, 2012 14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES University Transportation Center		
16. ABSTRACT In this project, we developed an integrated database to provide new analysis capabilities for discovering correlations between roadway characteristics and the occurrence of fatality collisions. Specifically, the aim of this data analysis project was to support the determination of relationships between Oklahoma roadway inventory data and NHTSA Fatality Analysis Reporting System (FARS) data. An approach had to be developed to match location data from the two different coordinate mapping systems used in these databases. The understanding of relationships between roadway characteristics and fatal crashes will support highway planning decisions and improve utilization of available resources targeting roadway safety. Initial blackspot and systematic analyses were performed as well as analyses utilizing additional data sources.		
17. KEY WORDS Traffic Records, Roadway Characteristics, Safety	18. DISTRIBUTION STATEMENT No restrictions. This publication is available at www.oktc.org and from NTIS.	
19. SECURITY CLASSIF. (OF THIS REPORT) Unclassified	20. SECURITY CLASSIF. (OF THIS PAGE) Unclassified	21. NO. OF PAGES 49 + covers 22. PRICE

SI (METRIC) CONVERSION FACTORS

Approximate Conversions to SI Units

Symbol	When you know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	25.40	millimeters	mm
ft	feet	0.3048	meters	m
yd	yards	0.9144	meters	m
mi	miles	1.609	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm
ft^2	square feet	0.0929	square meters	m^2
yd^2	square yards	0.8361	square meters	m^2
ac	acres	0.4047	hectares	ha
mi^2	square miles	2.590	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.0283	cubic meters	m^3
yd^3	cubic yards	0.7645	cubic meters	m^3
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.4536	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams	Mg
		TEMPERATURE (exact)		
°F	degrees Fahrenheit	(°F-32)/1.8	degrees Celsius	°C
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.448	Newtons	N
lbf/in^2	poundforce per square inch	6.895	kilopascals	kPa

Approximate Conversions from SI Units

Symbol	When you know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.0394	inches	in
m	meters	3.281	feet	ft
m	meters	1.094	yards	yd
km	kilometers	0.6214	miles	mi
		AREA		
mm^2	square millimeters	0.00155	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.196	square yards	yd^2
ha	hectares	2.471	acres	ac
km^2	square kilometers	0.3861	square miles	mi^2
		VOLUME		
mL	milliliters	0.0338	fluid ounces	fl oz
L	liters	0.2642	gallons	gal
m^3	cubic meters	35.315	cubic feet	ft^3
m^3	cubic meters	1.308	cubic yards	yd^3
		MASS		
g	grams	0.0353	ounces	oz
kg	kilograms	2.205	pounds	lb
Mg	megagrams	1.1023	short tons (2000 lb)	T
		TEMPERATURE (exact)		
°C	degrees Celsius	9/5+32	degrees Fahrenheit	°F
		FORCE and PRESSURE or STRESS		
N	Newtons	0.2248	poundforce	lbf
kPa	kilopascals	0.1450	poundforce per square inch	lbf/in^2

Acknowledgements

The authors express their appreciation to the Oklahoma Transportation Center (OkTC) for their support of this project. We would also like to thank the Oklahoma Highway Safety Office (OHSO) for providing development funding for the Statewide Analysis for Engineering & Technology (SAFE-T) system. We also thank Professor Jon Comer, Department of Geography, Oklahoma State University and his students for their collaboration on this work.

FATALITY ANALYSIS REPORTING SYSTEM AND ROADWAY INVENTORY CORRELATION

Final Report

9/9/2013

Ronald D. Barnes, Ph.D.
Associate Professor

Joseph P. Havlicek, Ph.D.
Professor

School of Electrical and Computer Engineering
University of Oklahoma
110 West Boyd, Room 150
Norman Oklahoma

Sponsoring agency:
Oklahoma Transportation Center
2601 Liberty Parkway, Suite 110
Midwest City, Oklahoma 73110

Table of Contents

Figures

Tables

Executive Summary

A new capability was developed to discover correlations between roadway characteristics and the occurrence of fatality collisions. Specifically, the aim of this data analysis project was to support the determination of relationships between Oklahoma roadway inventory data and national Fatality Analysis Reporting System (FARS) data. This information is desired by traffic engineers in the state of Oklahoma and will support highway planning decisions that will result in lives saved through improved utilization of available resources targeting roadway safety.

The Oklahoma roadway inventory database contains roadway characteristics for all highways in Oklahoma and is maintained by the Planning and Research Division in the Oklahoma Department of Transportation. The FARS database is a national record of details concerning all fatality collisions in the country and is maintained by the National Highway Traffic Safety Administration (NHTSA). Performing data analysis against these two different datasets enables correlative studies that are not previously possible. However this analysis requires merging these databases based on common geographical information. Complicating this data merger, each of these databases uses a different geographical location system, resulting in a current inability to directly analyze the relationships between the datasets.

We overcame this obstacle by first translating both databases into a common format and then developing algorithms to translate location information between the two geographical location systems. For the first time, stakeholders can now effectively fuse the two databases seamlessly and automatically, thereby providing them capability to carry out exhaustive data analysis and mining to reveal and discover important correlations between the actual physical highway characteristics and fatality collisions. Demonstrating this capability, two analysis approaches were implemented, Blackspot analysis and systemic analysis. Preliminary results of these analyses show the strength of this newly developed correlation capability.

We anticipate that capability will be used by transportation planners, engineers and managers. By identifying roadway characteristics that have historically exhibited strong correlation with fatality crashes, it will become possible prevent future fatalities through wide scale deployment of preemptive improvements. These improvements will ultimately save lives by effectively targeting available resources for the highest safety impact.

Introduction

As a part of their efforts to improve the safety of Oklahoma roadways, Oklahoma traffic engineers desire to know whether there is any correlation between roadway characteristics and the occurrence of fatality collisions. Although it has yet to be shown in scientific analysis of Oklahoma roadways, it makes sense to conjecture that roadway characteristics such as the degree of curve, average daily traffic (ADT), or the number of lanes may have an impact on where and how often fatal collisions occur. If such a correlation exists, there is a need to understand exactly how roadway characteristics influence the occurrence of fatality collisions. Specifically, information regarding which roadway characteristics are highly correlated with the occurrence of fatality collisions would be extremely useful to highway planners as they allocate funds for highway improvement.

Described in this report is a data analysis project which had the aim of determining relationships between Oklahoma roadway data and national Fatality Analysis Reporting System (FARS) data. Specifically, the purpose of the project was to discover any correlations between roadway characteristics and the occurrence of fatality collisions. This information can be used by traffic engineers in the state of Oklahoma and will support highway planning decisions that will result in lives saved throughout the state.

Prior to this project, the data needed to determine this information existed within three separate databases. One is the Fatality Analysis Reporting System (FARS) database, a national database maintained by the National Highway Traffic Safety Administration (NHTSA). Another is the Oklahoma roadway inventory database, which is maintained by the Planning and Research division of the Oklahoma Department of Transportation. The last is the Oklahoma roadway condition database, which is also maintained by ODOT Planning and Research.

The FARS database is a national record of details concerning all fatality collisions in the country. The Oklahoma roadway inventory database contains dense records of roadway characteristics for all highways in Oklahoma. The Oklahoma roadway condition database consists of a very dense set of roadway condition measurements over multiple years.

Although the data needed to find the desired correlation information was available, there was no way to query or analyze the three databases simultaneously. Apart from being maintained by different organizations, the databases use different geographical location systems, contain different fields and coding rules, and use different database formats. As such, there was no direct method by which correlation data between roadway characteristics and the occurrence of fatality collisions could be discovered. Prior to performing data analysis, these different datasets would first need to be merged based on common geographical information.

A significant portion of the work done in this project was related to overcoming this challenge. All databases were first translated into a common format so that they could be used simultaneously. Algorithms were then developed that were capable of resolving the different geographical location systems so that location information used in one database could be translated into location information that is used by the others. In addition to supporting our own analysis objectives, these algorithms will for the first time enable stakeholders to effectively and powerfully fuse the two databases seamlessly and automatically, thereby providing the capability to carry out exhaustive data analysis and mining to reveal and discover important correlations between the actual physical highway characteristics and fatality collisions. Once this process was complete, we were able to develop queries and analytical techniques to discover correlations between the datasets.

A two pronged approach to analysis was taken in this project: a) Blackspot analysis and b) systemic analysis. Examples of the preliminary results of this analysis are provided in the last section on this report.

The technical and analytical work proposed above was carried out by a qualified team at the Intelligent Transportations Laboratory (ITS) at the University of Oklahoma. Our laboratory has considerable experience and expertise in the database manipulation, algorithm development, and data analysis techniques needed to solve the problem described above. The lab has worked closely with the Oklahoma Department of Transportation for several years, and has implemented a large number of projects for the state. These projects include SAFE-T, a statewide collision analysis and reporting tool used by traffic engineers across the state. As a result of our past projects, our laboratory personnel are very familiar with the geographical location system that is used in the Oklahoma roadway inventory database. These factors contribute to a skill set in place at the ITS lab that greatly contributed towards the successful completion of this project.

As an important component of this project, we iterated with ODOT traffic engineers and their designated municipal stakeholders to ensure that the compiled statistics and reports generated by this project meet local, regional, and state needs and achieve the widest possible impact in terms of enhancing public safety by revealing previously unrecognized trends and correlations in existing fatality related traffic records data.

We anticipate that this report will be used by transportation planners and engineers as well as municipal and regional managers to identify roadway sections exhibiting high-risk characteristics directly in terms of fatality collisions. By identifying roadway characteristics that have historically exhibited strong correlation with fatality crashes, it will become possible to prevent future fatalities through wide scale deployment of preemptive improvements, including low-cost safety improvements such as shoulder rumble strips, cable barriers, and improved roadway delineation [1]. These improvements will ultimately save lives by effectively targeting available resources for the highest safety impact.

Background

Oklahoma Roadway Inventory Database

The Oklahoma roadway inventory database provides detailed roadway characteristics for all Oklahoma highways. The database is maintained by the Research and Planning division within ODOT. Roadway characteristics within the database are only recorded for non-turnpike highways, but as Oklahoma is an expansive state, this database still covers a significant number of highway miles. The roadway characteristics in the database are organized by geographical location on the roadways using Oklahoma's control section/milepoint location system.

The control section/milepoint location system is used in many different agencies within Oklahoma to reference highways and highway collisions. All highways in the state are segmented into distinct control sections. These control sections typically terminate at county lines and highway-highway intersections. This results in control sections that cover open segments of highway which are bounded by county lines and intersections. As such, control sections are uniquely identifiable stretches of roadway that do not suffer from the ambiguity of the traditional highway naming system. Figure 1 illustrates the structure of the control sections for Garfield County in Oklahoma. In this figure, the control sections are indicated by colored arrows, and each control section starts and stops at either the county line or a highway-highway intersection. Each highway in Figure 1 can be seen to belong to one and only one control section.

Control sections in Oklahoma are uniquely identified by their county number and control section number. Furthermore, all locations on control sections are identifiable by the milepoint of the location on the control section. The milepoint values for a control section start at 0.00 at one end of the control section and end at the other end of the control section with a value equal to the length of the control section. Thus locations for all significant roadway characteristics of Oklahoma roadways can be uniquely described by recording the county number, control section number, and milepoint of each characteristic.

Using the Oklahoma control section/milepoint location system, the roadway inventory database catalogs in high resolution the roadway characteristics for all Oklahoma highways. These roadway characteristics cover a large number of roadway features, including Average Daily Traffic (ADT), number of lanes, access control, surface width and type, inner and outer shoulder width and type, median width and type, whether the roadway is divided or undivided, degree of roadway curve, and roadway grade.

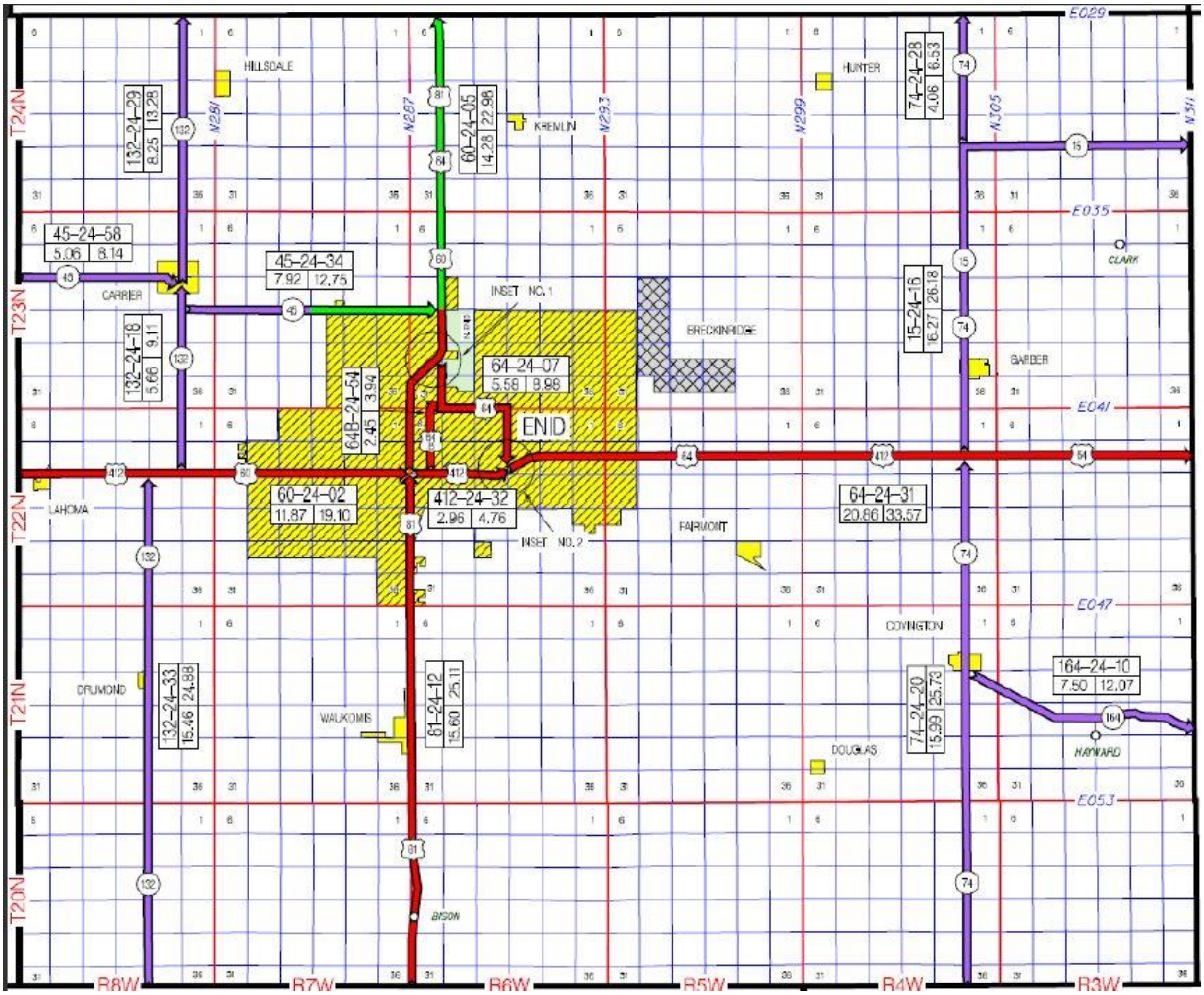

Figure 1. Control Section map for Garfield County, Oklahoma. Each control section is represented as a single colored arrow on the map.

Oklahoma Roadway Condition Database

The Oklahoma roadway condition database contains information about Oklahoma highway conditions and characteristics, such as rut depth and radius of curvature. The roadway condition database is also maintained by the ODOT Planning and Research division, and uses the control section/milepoint location system, but is separate from the roadway inventory database. The roadway condition data is collected by the Pavement Management Branch within ODOT, who contracts for a vendor to collect pavement data across Oklahoma's system of highways. The data is collected using semi-automated collection methods on a 2-year cycle, where about half of the mileage is collected in the first year, and the other half of the mileage is collected in the second year. The vendor delivers a raw condition database that is formatted with one record for every 0.01 mile of pavement. Each record of roadway condition data includes:

- Location Information: includes county and control section numbers, chainage (milepoint), latitude and longitude coordinates
- Pavement Type: a code which describes whether the pavement is asphalt, composite, jointed concrete pavement, continuously reinforced concrete pavement, or brick
- Sensor Data: includes roughness, rutting (for asphalt), and faulting (for jointed concrete)
- Distress Data: includes measuring or counting of cracks, potholes, punch-outs, etc.

Since data is collected on a 2-year cycle, collected data from pairs of years are grouped together to form a combined database that covers all of our highway mileage. The available roadway condition collections maintained at the time of this project were:

- 2001/2002: collected by Roadware
- 2004/2005: collected by Roadware
- 2006/2007: collected by Roadware
- 2008/2009: collected by Pathway Services
- 2009/2010: collected by Pathway Services

Similar roadway inventory databases have been utilized in other states by correlating their contents to other data sources. For example, in Wisconsin, a relatively general framework was developed for matching such data to image-based databases [2]. In this work, it was shown that road features could be located with a high degree of accuracy on Wisconsin Department of Transportation GIS base maps. However, their framework was focused on the problem of spatial mismatch with image data rather than the problem of correlating different datasets with different mapping coordinate systems addressed in this project.

FARS and SAFE-T Databases

The Fatality Analysis Reporting System (FARS) database is a national database of fatal traffic collisions on all public roadways in all 50 states, the District of Columbia, and Puerto Rico [3] [4]. This database is maintained by the National Highway Traffic Safety Administration (NHTSA). Within this database, data from the years 1975 to 2009 is publicly available. Before FARS was introduced, traffic engineers and other traffic safety agents often lacked adequate data for data-driven planning and evaluation [5]. FARS contains data on traffic accidents involving fatalities, personal injuries or property damage. It is generally accepted as the most complete source of data on fatal traffic accidents and supports a variety of analyses. Often, FARS data is used for statewide comparisons and problem identification while data sources containing data about individual crashes is used for identifying problems in specific areas [6]. SAFE-T, described later in this section is an example of a data source of the latter type. Figure 2 shows the introductory summary page for the online FARS Encyclopedia, which is a public website that may be used to obtain and query data in the FARS database.

File Versions	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994
Motor Vehicle Traffic Crashes																
Fatal Crashes	30,797	34,172	37,435	38,648	39,252	38,444	38,477	38,491	37,862	37,526	37,140	37,107	37,324	37,494	37,241	36,254
Traffic Crash Fatalities																
Vehicle Occupants																
Drivers	17,640	19,279	21,717	22,831	23,237	23,158	23,352	23,625	22,914	22,914	22,971	22,654	22,730	22,572	22,370	21,596
Passengers	6,770	7,441	8,716	9,187	9,750	10,042	10,171	10,370	10,227	10,451	10,325	10,327	10,765	10,860	10,576	10,294
Unknown	64	71	94	101	83	76	104	110	102	86	96	107	114	102	118	108
Sub Total1	24,474	26,791	30,527	32,119	33,070	33,276	33,627	34,105	33,243	33,451	33,392	33,088	33,609	33,534	33,064	31,998
Motorcyclists	4,462	5,312	5,174	4,837	4,576	4,028	3,714	3,270	3,197	2,897	2,483	2,294	2,116	2,161	2,227	2,320
Nonmotorist																
Pedestrians	4,092	4,414	4,699	4,795	4,892	4,675	4,774	4,851	4,901	4,763	4,939	5,228	5,321	5,449	5,584	5,489
Pedalcyclists	630	718	701	772	786	727	629	665	732	693	754	760	814	765	833	802
Other/ Unknown	150	188	158	185	186	130	140	114	123	141	149	131	153	154	109	107
Sub Total2	4,872	5,320	5,558	5,752	5,864	5,532	5,543	5,630	5,756	5,597	5,842	6,119	6,288	6,368	6,526	6,398
Total**	33,808	37,423	41,259	42,708	43,510	42,836	42,884	43,005	42,196	41,945	41,717	41,501	42,013	42,065	41,817	40,716

Figure 2. Summary page of Online FARS Encyclopedia [7].

Within the FARS database are over 100 data fields that describe characteristics of each fatal collision in the database. These fields cover a large amount of information that can be of interest in the analysis of fatality collision data, including the time and location of the crash, the statuses of the drivers, the types of vehicles involved, and the likely causes of the collision. Figure 3 shows many of the FARS database fields as they are listed in the online FARS Encyclopedia. Beginning in 2006, the fields in FARS have begun a process of standardizing [8] with the National Automotive Sampling System General Estimates System (NASS GES) [9], a broader source of crash data based on a nation-wide sampling of police traffic reports. Because the GES includes data from crashes of all types including non-fatal, it can be used to estimate how many crashes of different kinds take place. While the GES was not exploited during this project, the standardized data fields with FARS would enable the work done here to be directly leverage if correlation between GES and the Oklahoma Roadway Condition Database was desired.

Figure 3. Fields available within FARS data [7].

The location information available in the FARS database resides within the latitude, longitude, milepoint, and trafficway identifier fields. The latitude and longitude fields contain the latitude and longitude coordinates of the collision, while the trafficway identifier field contains the name of the roadway upon which the collision occurred. The milepoint field contains the mile point on the roadway at which the collision occurred. This mile point refers to the mile point along the roadway over the entire state, with the mile point values typically starting at 0.00 at either a state boundary or at one end of the highway.

The Statewide Analysis for Engineering & Technology (SAFE-T) database is a complete collection of all Oklahoma collision data from the year 1998 to present. This database is used within the SAFE-T online crash reporting system [10], and was readily available for usage in this project. The fields available in the database are a direct copy of the primary collision database maintained by ODOT, and as such all data within the SAFE-T database has passed through ODOT quality control. The ODOT collision database is used as the basis for submissions to the federal FARS database, and contains all of the information found in the FARS database, making the FARS database a subset of the SAFE-T collision database. Furthermore, while the location system within the FARS database is based off of latitude/longitude coordinates or statewide milepoints, the SAFE-T database makes use of the same control section/milepoint location system that is used by the Oklahoma roadway inventory and roadway condition databases. Because of the availability, completeness, and compatibility of the SAFE-T database, it was used for all integration and analysis efforts in this project.

Procedure

The first step in achieving the goals of this project was to combine the SAFE-T and Oklahoma roadway databases. This required that all database be obtained in their entirety (all fields for all Oklahoma fatality collisions in the case of the SAFE-T database) and then translated into a common database format. The latest copies of the Oklahoma roadway inventory database and the Oklahoma roadway condition database were obtained from the Planning and Research division within ODOT, while the SAFE-T database was available via the SAFE-T online system. The SAFE-T database was already stored using the open-source, powerful, and highly reliable MySQL database format [11], and the other two databases were translated into the MySQL database format.

Once the databases shared a common database format, they were joined based on common geographical data. While the location information in the FARS database consists of latitude/longitude coordinates, the SAFE-T database uses Oklahoma's control section/milepoint system, which is also used by each of the roadway databases. This allowed the usage of a single geographic location system for the three databases, however in order for the databases to be joined based on common geographical data, the different locations of the data points within each dataset needed to be resolved against each other so that related information between the datasets could be identified. To this end, we developed matching algorithms between the different data point locations that are capable of locating the closest, most relevant, and most reliable data points from a target database to a data point from a source database. While reduced storage might have resulted from dynamically querying the intersection of stored geographic shapes [12] or from storing only piecewise segments, the point-to-point matching algorithm used in this work was conducive to the formation of a fused database that could be both analyzed and shared with other researchers as later described.

With their completion, the matching algorithms have allowed us to determine what roadway characteristics are associated with each collision in the FARS database. As a result, fatality collision data and associated roadway characteristic data have successfully been stored in a new, combined database from which analysis operations could be carried out.

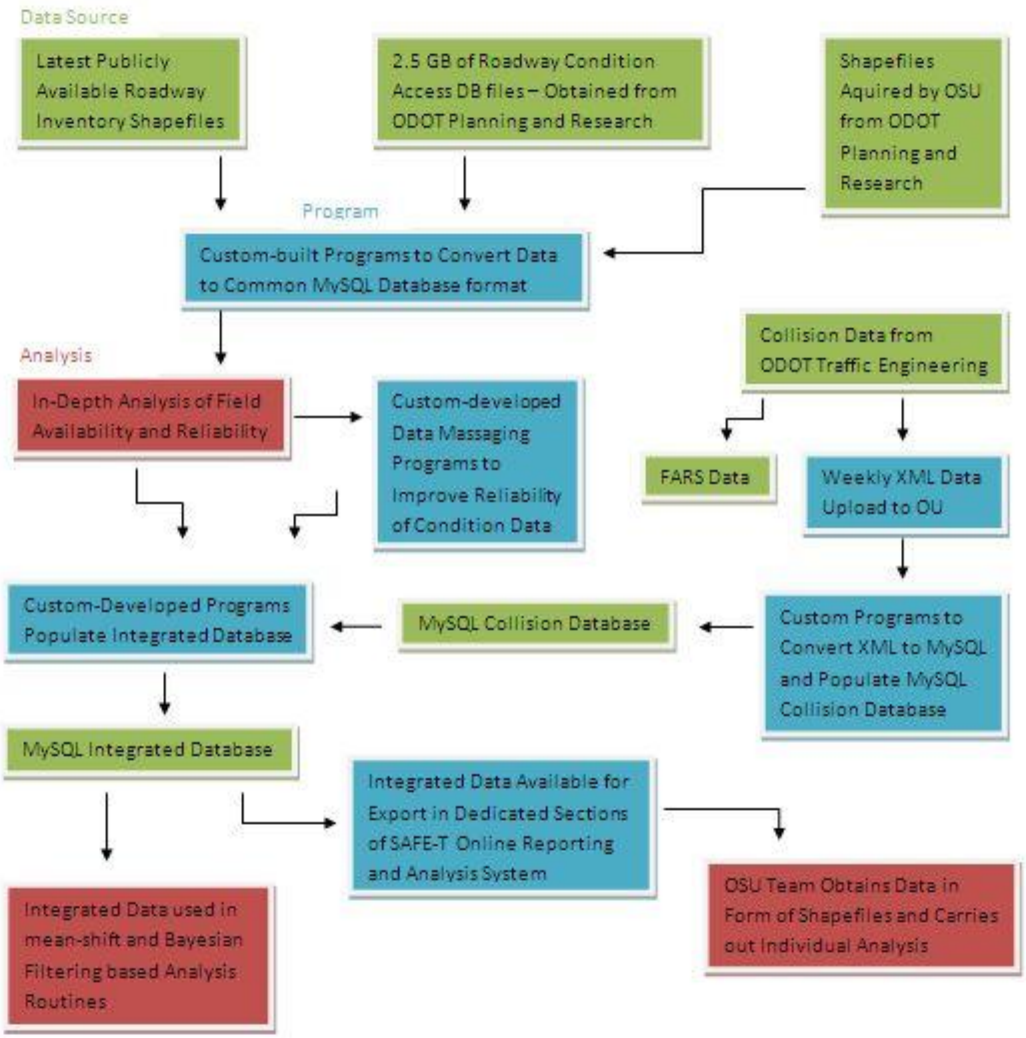

Figure 4. Flow Diagram of Acquired and Created Datasets

Definition of Project Milestones

The first event to occur related to this project was a meeting with a team from the Geography department at OSU, which had also been awarded this contract, on the 11th of August, 2011. Dr. Jon

Comer, of the OSU Geography Department, and his students Leo Bombom and Nick Rose, travelled to the ITS Lab for discussions. The proposal from the OSU team for this project focuses primarily on the analysis techniques to be applied to the data, and they were mostly unaware of the amount of work that would be required to first obtain and integrate the data.

Table 1. Project Milestones as distributed among OU and OSU teams.

Milestone	OSU	OU
1. Acquire roadway characteristics/inventory data	-	Lead
2. Translate FARS and roadway characteristics/inventory data to a common format	Support	Lead
3. Develop engine to query joint data	-	Lead
4. Reverse geocode FARS data to control section/ mile marker coordinate system.	Support	Lead
5. Create a library of GIS files encoding the joint data	Lead	-
6. Perform statistical analysis to identify and assess correlations between fatality crashes and roadway characteristics	Lead	Support
6A. Exploratory Data Analysis - Spatial - Temporal	Lead	Support
6B. Low complexity parametric methods	Lead	Support
6C. High complexity spatial correlation techniques	Lead	-
6D. Bayesian inference	-	Lead
6E. Assess and interpret quantitative results	Support	Support
7. Iterate with ODOT designated stakeholders to refine analysis and maximize impact	Lead	Support
8. Deliver quarterly and final reports	Support	Support

Out of that meeting, a project milestone list, shown in Table 1, was created with roles for each team. The OU team was to be responsible for obtaining and integrating the data, which would then be provided to the OSU team, at which point both teams will proceed with independent analysis.

In the months after the meeting, a few initial tasks were carried out by the OU team in preparation for later work. Firstly, all immediately available data and documentation were given to the OSU team, including a login to the SAFE-T online which could be used to export Excel data and shapefiles and a data dictionary for the collision data. Some work was done to the shapefiles exported by the SAFE-T online system to ensure that they were compatible with the latest versions of ArcGIS.

It had been decided that the best way to deliver the integrated datasets to the OSU team would be in the form of shapefiles exported from the SAFE-T system, however the shapefiles at that time only exported latitude and longitude points without associated fields. This meant that an initial major objective was to modify the exported shapefiles in SAFE-T to include all available fields from the integrated dataset. That work was completed in early October, and from that point the OSU team has been using shapefiles exported from SAFE-T.

Acquisition of Datasets

The other initial undertaking was the acquisition of the needed roadway characteristic data from ODOT. Ginger McGovern, who had been the initial point of contact within ODOT Planning and Research and who had been involved in the proposal for this project, had retired, and we were referred to John Bowman, her successor. We contacted Mr. Bowman with an explanation of the project and a request for data. After some time and multiple contact attempts, Mr. Bowman referred us to Mark Brown within Planning and Research, and Mr. Brown provided a link to ODOT's publicly available highway shapefile as well as a link to an internal system they use for the data and some documentation.

The Road Inventory Manual from that system provides a good description of the fields available within the public shapefile. The document is large, but descriptions of the relevant fields can be found on pages 10-23.

We were operating on the assumption that the public shapefile represented all available roadway data, however on Nov. 14th we were forwarded a document from Ginger McGovern regarding the project. This document indicated that there existed relevant roadway condition data in addition to the roadway inventory data available in the public shapefile. We contacted Mark Brown directly requesting the additional data and were referred to Justin Calvarese within Planning and Research.

Mr. Calvarese, after checking the project details, promptly provided us with all available roadway condition data. Since the SAFE-T database contains collision data from 1998 to present, we requested all available years of roadway condition data (from 2001 to 2010), and Mr. Calvarese uploaded 2.5 GB of Microsoft Access files to a location for us to download, sent us a data dictionary for the data titled Pavement Management Condition Database Fields, and provided some information about the data. He recommended against using the collected Radius data from 2008 and 2009, and suggested that 2007 or 2011 data should be substituted for those years. He also informed us that some of the fields that appear in more recent years weren't part of older collections.

After downloading the Access files, the OU team was in possession of roadway condition data for the years 2001-2002, 2004-2005, 2006-2007, 2008-2009, and 2009-2010. An initial check of the radius data from 2008 and later did indicate that it was unreliable, with widely varying numbers including large spikes at single points within areas with normal readings of 0. Meanwhile data from 2007 and before appeared to be quite a bit more reliable.

Translation of Datasets Into Common Database Format

Once the roadway inventory and roadway condition databases had been acquired, they were translated into the MySQL database format to facilitate analysis of and operations on the datasets.

For the shapefile used in the roadway inventory dataset, a conversion program called shp2mysql was used to automatically create and populate MySQL tables from the fields and data defined in the shapefile. The converted data was tested against the original data to verify that the translations were performed accurately.

For the Microsoft Access files used in the roadway condition dataset, Access itself was used to export directly to MySQL. This involved installing MySQL, MySQL ODBC drivers, and Access on a single system. Then the Access 'Export to ODBC' options were used to create relevant tables directly within MySQL. This was done each set of the roadway condition data, with a MySQL table created for each. These new tables were also tested against the original data to verify the translations were performed accurately. The mysqldump command was then used to isolate and copy the newly created tables to the server housing the SAFE-T collision database.

Analysis of Data Availability

After all three databases had been placed on a single machine and translated into a single format, an analysis was performed to determine which fields described in the data dictionaries for the databases were actually available within the datasets. This proved challenging as the field names in the datasets were often different from those in the data dictionaries, but with some effort each field from the roadway inventory shapefile was matched with a field description in the Road Inventory Manual, and each column from each roadway condition Access file was matched with a field from the Pavement Management Condition Database Fields document.

The results of this analysis were that all listed fields within pages 10 to 23 of the Road Inventory Manual were available within the roadway inventory shapefile except for the fields in Table 2.

Table 2. Roadway Inventory Fields defined in Data Dictionary but not available in Roadway Inventory Dataset.

Alternate Name Route
Ramp or Frontage Road Code
Speed Limit
County FIPS Number
Urban FIPS Number
Percent Passing
Truck Route
STRAHNET
Maintenance Crew Type
Is Structure
Facility Type
Route Signing
Route Qualifier

Additionally, many fields from the Road Inventory Manual and roadway inventory shapefile were already in the collision database. These fields had already been matched to collisions by ODOT and were provided with all collision data updates. A list of these pre-existing roadway inventory fields is shown in Table 3.

Table 3. Pre-existing Roadway Inventory Fields within SAFE-T Database.

Access Control
Average Daily Traffic
County Number
Control Section Number
Inner Shoulder Type
Inner Shoulder Width
Maintenance Division
Median Type
Median Width
Number of Lanes
Outer Shoulder Type
Outer Shoulder Width
Route Number
Rural/Municipal
Subsection Start
Subsection End
Surface Width
Surface Type
FA Hwy Route #
NHS
UAT
National Functional Classification

Upon analyzing the data available in the roadway inventory shapefile data and the pre-existing data in the SAFE-T database, it was found that all data was the same except for a few minor differences:

- Entries for the 'Route' field were described in the data dictionary and seen in the roadway inventory shapefile as having single letter prefixes, such as 'U060'. In the pre-existing data, however, the same field would appear as 'US060'.
- The 'Surface_Type_CD', 'Surf_Primary', 'Surf_Original', 'Base_Type', 'Surf_Thickness' were individual fields of one character in the roadway inventory shapefile, but are a 5 character string (the 'Surface_type' field) in the pre-existing data. The data represented by the fields is the same.
- Values from the 'ADT' field in the roadway inventory shapefile values often differed from the pre-existing data. It was determined that the ADT values differed depending on the collision year, and that values from the shapefile matched pre-existing ADT values for roughly the years 2005 to 2009, effectively dating the shapefile. As such, the pre-existing ADT values can be considered to be more reliable.
- The pre-existing roadway inventory data contained one extra field, 'FA Hwy Route #', that was not in the roadway inventory shapefile or the data dictionary.

After removing the fields not found and the pre-existing roadway inventory fields, the remaining fields from the data dictionary could be isolated as newly available roadway inventory fields. These new fields are listed in Table 4.

Table 4. Available Roadway Inventory Fields from Roadway Inventory Shapefile.

Suffix
Subsection Type
NLF ID
Break Reason
Fabric
Base Thickness
Base Width
Right-of-Way Width
City FIPS Number
City Number
Jurisdiction
NHS Route Number
Status Control
Terrain Area Type
Population Group
State Functional Classification
Type Parking
Reservation
Connecting Link
Primary Route
Follow Route #1

Follow Route #2
Follow Route #3
Follow Route #4
Defense Route Number
Defense Route Sequence
Defense Route Direction
Date of Construction
Last Maintenance Date
Beginning Node Number
Ending Node Number
Subsection Ending Description
Project Ties
Maintenance Responsibility
Maintenance Division
HPMS Sample Number

From the analysis of the Pavement Management Condition Database Fields document and the roadway condition databases, all fields listed in the data dictionary appeared to be available within the datasets to varying degrees. In other words, the roadway condition dataset had been collected multiple times over several years, and not all fields were available for all years. The differences that were found between the different years of the roadway condition dataset were that:

- The field 'MacroTexture' in the data dictionary was called 'Texture' in all datasets except 2009-2010
- The fields 'Radius' and 'Grade' were unreliable for the datasets from 2008-2009 and 2009-2010
- The field 'Sensors' in the data dictionary was not in the dataset from 2009-2010
- The field 'Coll_Yr' in the data dictionary was called 'ImageID' in the datasets from years 2006-2007, 2004-2005, and 2001-2002
- The fields 'Altitude' and 'Heading' in the data dictionary were only available in the dataset from the years 2009-2010. The fields were also shown in the data set from the years 2008-2009 but were blank for all 2008 data.

The available roadway condition fields are listed in Table 5.

Table 5. Available Roadway Condition Fields.

Group
Division
Route Name
Direction
Chainage
Number of Lanes
NLF ID
Subsection ID

Pavement Type
Surface Type
Right IRI
Left IRI
Mean IRI
Mean Rut
Max Rut
Low Rutting (Percent)
High Rutting (Percent)
Mean Fault
Max Fault
Standard Dev of Faults
Faulted Joints (Number)
Joints (Number)
Low Sev. Cracks (Number)
Medium Sev. Cracks (Number)
High Sev. Cracks (Number)
Very High Sev. Cracks (Number)
Low Fatigue Cracking (Length)
Medium Fatigue Cracking (Length)
High Fatigue Cracking (Length)
Low Misc. Cracking (Length)
Medium Misc. Cracking (Length)
High Misc. Cracking (Length)
AC Patching (Area)
PC Patching (Area)
Raveling (Length)
Low Sev. Transverse Cracks (Number of Slabs with)
High Sev. Transverse Cracks (Number of Slabs with)
Low Sev. Longitudinal Cracks (Number of Slabs with)
High Sev. Longitudinal Cracks (Number of Slabs with)
Low Sev. Corner Breaks (Number of Slabs with)
High Sev. Corner Breaks (Number of Slabs with)
Low Sev. Spalling (Number of Joints with)
High Sev. Spalling (Number of Joints with)
Low Sev. D-Cracking (Number of Joints with)
High Sev. D-Cracking (Number of Joints with)
Low Sev. Multicracking (Number of Joints with)
High Sev. MultiCracking (Number of Joints with)
Low Sev. Longitudinal Cracks (Length)
High Sev. Longitudinal Cracks (Length)
Low Sev. Punchouts (Number)
Medium Sev. Punchouts (Number)
High Sev. Punchouts (Number)
MacroTexture
Radius of Curvature
Grade
Cross-Slope

Events
Over Under
Condition Latitude
Condition Longitude
Number of Sensors
Filename
Collection Year
Chainage Begin
Chainage End
FKey
Collection Date
Altitude
Heading

Design of Integrated Data Structures

With the analysis of data availability complete, a database structure was designed that could be used for the integration of the roadway inventory and the roadway condition data with the SAFE-T collision data. In order to allow researchers to better use the roadway data dictionaries, the names and order of the integrated fields were made to match those from the data dictionaries. Because the pre-existing roadway inventory fields present in the SAFE-T dataset were of higher quality than those in the roadway inventory shapefile, only the roadway inventory fields that were not already in the SAFE-T dataset (the fields listed in Table 4) needed to be integrated with the collision data.

Inventory Dataset Characteristics

The roadway inventory shapefile has fields and data for each subsection of Oklahoma highways. As such each collision matches a single subsection's data, and the shapefile contains just one set of fields per collision. Some subsections have multiple sets of data per subsection, with data for each side of the road, but after some analysis it was determined that the data was identical for both sides of the road, with the exception of ADT, which was 0 on the secondary side of the road. However, as the ADT field already existed within the SAFE-T database, this was a nonissue, and data from any single dataset for a subsection could be used matched to the collision data.

Condition Dataset Characteristics

Whereas the roadway inventory shapefile has just one set of fields per collision, analysis of the roadway condition tables showed that they have potentially two sets of fields per collision - one set each for the primary and secondary sides of the road, with differing data for each. The data dictionary defines the primary side as the side of the road that goes with the arrow on the control section maps (see Figure 1), while the secondary side is the side of the road that goes against the arrow. Further analysis of the data showed that a secondary side is often, but not always, recorded.

Integrated Dataset Characteristics

It was determined that the best way to store and integrate both the roadway inventory and roadway condition data was to create two additional tables similar to the existing roadway inventory table within the SAFE-T database, which has one entry per collision containing extra fields for the collision. The new secondary collision-roadway inventory table would contain all new roadway inventory shapefile fields that were not already in the SAFE-T database (the fields listed in Table 4). The new collision-roadway condition table would contain both the primary and secondary road information from the roadway condition database which matches the collision date.

Each of the new tables have the collision Document ID as the primary index of the table, followed by all relevant road characteristic fields. The field names in the new collision-roadway condition table are preceded by 'p_' for the primary side of the road, and by 's_' for the secondary side of the road.

Population of Integrated Data Structures

A program was designed and implemented to automatically populate the new secondary collision-roadway inventory and collision-roadway condition tables from the from the MySQL table containing the original sets of translated data. The program was set to be run for the entire collision database and for all new collisions as they are uploaded to the SAFE-T system. The program inserts rows into the new collision-roadway tables for any highway collisions for which rows do not already exist in the new tables.

Roadway Condition Dataset Cleaning

The population program was initially run over the entire available collision database, and while the program was able to match collisions to roadway inventory data rather quickly, matching the correct roadway condition data to the collision was more of a challenge, as each collection of roadway condition data has around 1.5 million entries. Because of this, and because some fields from the roadway condition tables have varying names or are unreliable for different years, a 'clean up' program was developed and run on the original translated roadway condition tables so that each collision could correspond to a single roadway condition table, rather than have the population program gather data from multiple condition tables for each collision. This program made the following modifications to the original translated roadway condition tables:

- 'Radius' and 'Grade' data were cleared from the 2008-2009 and 2009-2010 tables
- 'Radius' and 'Grade' data from 2006-2007 were copied to the 2008-2009 and 2009-2010 tables
- 'Altitude', 'Heading', and 'AvgSpeed' data were cleared from 2008-2009. Those fields were added to the 2001-2002, 2004-2005, and 2006-2007 tables
- The 'Altitude', 'Heading', and 'AvgSpeed' data from 2009-2010 was copied to all other tables
- The 'Sensors' field was added to 2009-2010 and data for the field was copied from 2008-2009
- The 'Texture' field was renamed to 'MacroTexture' as needed
- The 'ImageID' field was renamed to 'Coll_Yr' as needed

In all cases where data was 'copied' from other years, each entry from the target table was matched with the closest corresponding entry from the source table by milepoint. In cases where a close corresponding entry could not be found, the fields were populated with 0. This 'clean-up' program took around 36 hours to run, with the tables appropriately indexed.

Roadway Condition Matching

Once the condition data had been cleaned, the population program was run over the entire collision database. When matching roadway condition data to collisions, this program also matches each collision to the closest entry in the relevant condition table by milepoint. It in fact does so twice, once for each side of the road. After several hours, the population program succeeded in populating the new collision-roadway inventory and collision-roadway condition tables, effectively matching and merging the collision, roadway inventory, and roadway condition databases in a common format on a single server.

Within the roadway condition data, the secondary side of the roadway was not always collected by this program since it was not always available, so the corresponding fields are blank for many collisions, but include in the integrated dataset for the sake of completeness. In cases where both sets exist, a clear picture of the road conditions for the collision is obtainable by matching the reported vehicle directions with the heading field in the condition data. Alternatively, effective analysis can be performed by simply always considering only the primary condition data.

Release of Integrated Data

With the new collision-roadway tables populated, all newly available fields were added to the export functionality of the SAFE-T online crash reporting system. The new collision-roadway inventory fields were added to a new export section titled 'Additional Roadway Inventory', while the new collision-roadway condition fields were broken up into Primary and Secondary 'Roadway Condition' sections. The new sections on the SAFE-T export pages were very large, and so the user interface of the export pages was modified to allow the new sections to be collapsible.

The Lane ID field in the new collision-highway inventory section displays a random side of the road do to the design of the population program, and could be confusing since the field is related to the roadway inventory data but not to the collision. Since the program was designed with the assumption that the roadway inventory data was the same for both sides of the road, the LaneID field was removed from the export functions of the SAFE-T system.

With the newly merged collision-roadway characteristic data now available, the OSU team was notified of the work done thus far so that they might begin their analysis efforts.

Integration of Additional Datasets

After some time, additional questions were received from the OSU team regarding several fields that they could not seem to locate within the merged collision-roadway dataset.

It appeared as though the OSU team had seen data sources that the OU team had not, so a meeting was arranged to discuss the issue. The OU team travelled to Stillwater and held a very productive meeting with the OSU team, during which it was found that they did indeed have some additional data. This data was in the form of four shapefiles that had been given to them by Mark Brown from ODOT Planning and Research at some point, and these shapefiles included a few extra fields such as grade, curve, and passibility that were, somewhat sporadically, associated with control sections on Oklahoma highways. A copy of these shapefiles was given to the OU team, and it was agreed that the OU team would extract any possible new data from them, then matching that data to collisions as new fields.

During the meeting, there was also discussion about how a single roadway database might be created that contains all available roadway data, which could then be used for prediction of hotspots on highways. This is a problem that not been previously addressed, and is a challenging one since the various highway data sources all have different points of reference on the highways. Even the shapefiles provided to the OSU team by Mark Brown differ from the shapefile that is publicly available on ODOT's website in that the subsections appear to be defined very differently. Meanwhile, the condition databases represent an extremely dense dataset, while the shapefiles are much more sparse. It is a problem that could be addressed by selecting a dataset, then attempting to conform the other datasets to it. An alternative approach might be to simply attempt to do predictive analysis on each highway dataset individually and then attempt to merge the analysis results. Either approach, however, falls beyond the resources available to this project.

The new shapefiles were analyzed and processed in much the same way as the Oklahoma highway inventory shapefile had been analyzed and processed. The new shapefiles were first converted to the MySQL database format and were then analyzed to see what data could be extracted from them. The new fields available from the four shapefiles were determined after going through each and finding which fields had not already been present in the Oklahoma highway inventory shapefile. Those fields were then extracted and added to the secondary collision-roadway inventory table within the integrated database structure using the same techniques that had been used for the Oklahoma highway inventory shapefile. The new fields from the new shapefiles were then made available within the SAFE-T online system as additional fields in the 'Additional Highway Inventory Fields' section of the export pages. The new fields that were found to be available are listed in Table 6.

Table 6. Fields Available in Additional Highway Inventory Shapefiles.

Stop Sight Def
Stop Sight Pos/Neg Grade 2
Stop Sight Grade 1
Stop Sight Pos/Neg Grade 2
Stop Sight Grade 2
Stop Sight Design Spe.
Stop Sight Remarks
Passing Def
Grade Def
Grade Percent
Curve Def
Curve Degrees

Within each shapefile, there were also Direction and Deficiency fields that were not in the main highway inventory shapefile. The Deficiency field, however, only indicated which shapefile the record was in, and so was not included for extraction. The Direction field contained N,S,E,W characters for a few shapes, but for the vast majority of shapes the field was null, so it was also not included for extraction.

The new fields were extracted and matched to collisions with a high degree of accuracy, however the distribution of data within the fields was quite sparse. For each collision, data was gathered from the new shapefiles and matched to a collision if the shapefiles had shapes at the collision location. For most collisions, however, the new shapefiles did not have shapes, and the new fields are blank. Still, the new fields may provide some information or act as a useful check to the roadway condition data.

In addition, many locations in the new shapefiles had several shapes for the same stretch of road. The only apparent differences between the shapes were between definition fields and the date fields, while all other fields were identical. The values of the definition fields varied widely in such cases, and were thus either id numbers or unreliable. Unfortunately, without any data dictionaries to go with the new shapefiles, the exact meaning of such fields could only be inferred. As those were the only differences, collisions were associated with the values of the first matching shape at the collision location, if there were any.

Once these new shapefiles had been analyzed and the new fields from them made available, the OSU team was notified of the additional data for their analyses.

Support and Investigation of Data Characteristics

In response to further queries from the OSU team, additional investigation was carried out into a few issues. Resultantly, we found, and communicated to OSU, a few more characteristics of the data.

With the assistance of Justin Calvarese from ODOT Planning and Research, we determined that the 'Radius of Curvature' field could be described in degrees when reporting our results. Standard practice is to relate the radius of curvature to the degrees of change in the roadway over 100 feet. An approximation of this relation can be found in

$$D = 5729.57795 / R$$

where D is the degree of the curve in degrees, and R is the radius of the curve in feet. Once calculated, the value of D can be referred to as the ‘Degree of Curvature’.

Through our own investigation, we also verified that for the Radius field within the roadway condition data, positive values indicate the road is curving to the left, while negative values indicate the road is curving to the right. For the Grade field, positive values mean the road has an uphill slope, negative is downhill. For the Xfall field, positive means the left edge of the road is higher, negative means the right edge is higher.

The Curve Degrees field, which was extracted from the new shapefiles provided by the OSU team, appears to be the degree of curvature of the road, in arc degrees, without any indication of being to the left or right. The Curve Degrees field values for a collision are generally on the same order of magnitude as the Degree of Curvature values that are calculated from the Radius of Curvature field for the collision. The Curve Degrees field, however, is very sparsely populated, as the shapefile it was extracted from contained very few shapes.

Data Analysis

Data Analysis Overview

FARS data has been previously analyzed extensively using a variety of techniques. For example, a basic statistical analysis of fatal large truck crashes demonstrated that the majority of such crashes occurred with a passenger vehicle when the passenger vehicle was in the truck's lane [13]. Empirical Bayesian methods that attempt to estimate an underlying distribution based on observed information, have been used on FARS data to evaluate traffic improvements and rank locations in which improvements are most likely to be effective in reducing crashes. Examples of such studies evaluated the safety efficacy of red light cameras [14] and provided analysis for selecting the most effective sites for these camera installations [15].

The two primary types of analyses supported on the collision data set in this work are Black Spot Analysis (BSA) [16] [17] and Systemic Analysis. The analysis enabled by this project provides traffic engineers and city planners in the State of Oklahoma a tool to focus their attention on roadway locations that have an increased number of collisions or an unusually high number of collisions. Since traffic volumes and system use evolve over time as our state grows the needs of the traffic network change. Existence of higher collision incidences does not necessarily indicate either poor initial design or outdated configurations. A traffic engineer with appropriate simulation modeling software can make those analyses. The analysis provided here is an important tool to aid those processes.

Blackspot Analysis

Given that each data point in our data set is individually listed by latitude and longitude, some sense of locality has to be introduced to the data. This notion of locality is called data clustering and collisions that are deemed local to other collisions are considered to be elements of the same cluster.

Once the data is clustered the number of collisions at a given location can be used for further analyses. Each clustering algorithm has its benefits and detractors so no one algorithm is ideal for all applications. We chose a non-parametric algorithm using Gaussian kernel functions known as the Mean-Shift Algorithm (MSA) [18]. In the case where a Gaussian kernel is used the algorithm is

1. Given the covariance of an n-dimensional Gaussian Σ and the set X is the collection of collision locations then at each point $x \in X$ we evaluate $m(x) = \frac{\sum_{x_i \in X} x_i e^{-\frac{(x-x_i)^T \Sigma^{-1} (x-x_i)}{2}}}{\sum_{x_i \in X} e^{-\frac{(x-x_i)^T \Sigma^{-1} (x-x_i)}{2}}}$
2. If for some threshold ϵ, $1 - e^{-\frac{(x-x_i)^T \Sigma^{-1} (x-x_i)}{2}} < \epsilon$ we merge the two data points as they will travel the same trajectory and decimating the dataset in this manner easy the computational burden in subsequent iterations.

3. Then the energy of the motion of each data point is summed to get an overall "motion energy," $\mathrm{ME}= \ {}_{x_i \in X}(x_i\text{-}m(x))^2$ and we move each point towards its cluster mode $x \leftarrow m(x)$. If the motion energy is below some small threshold then stop, otherwise repeat the process.

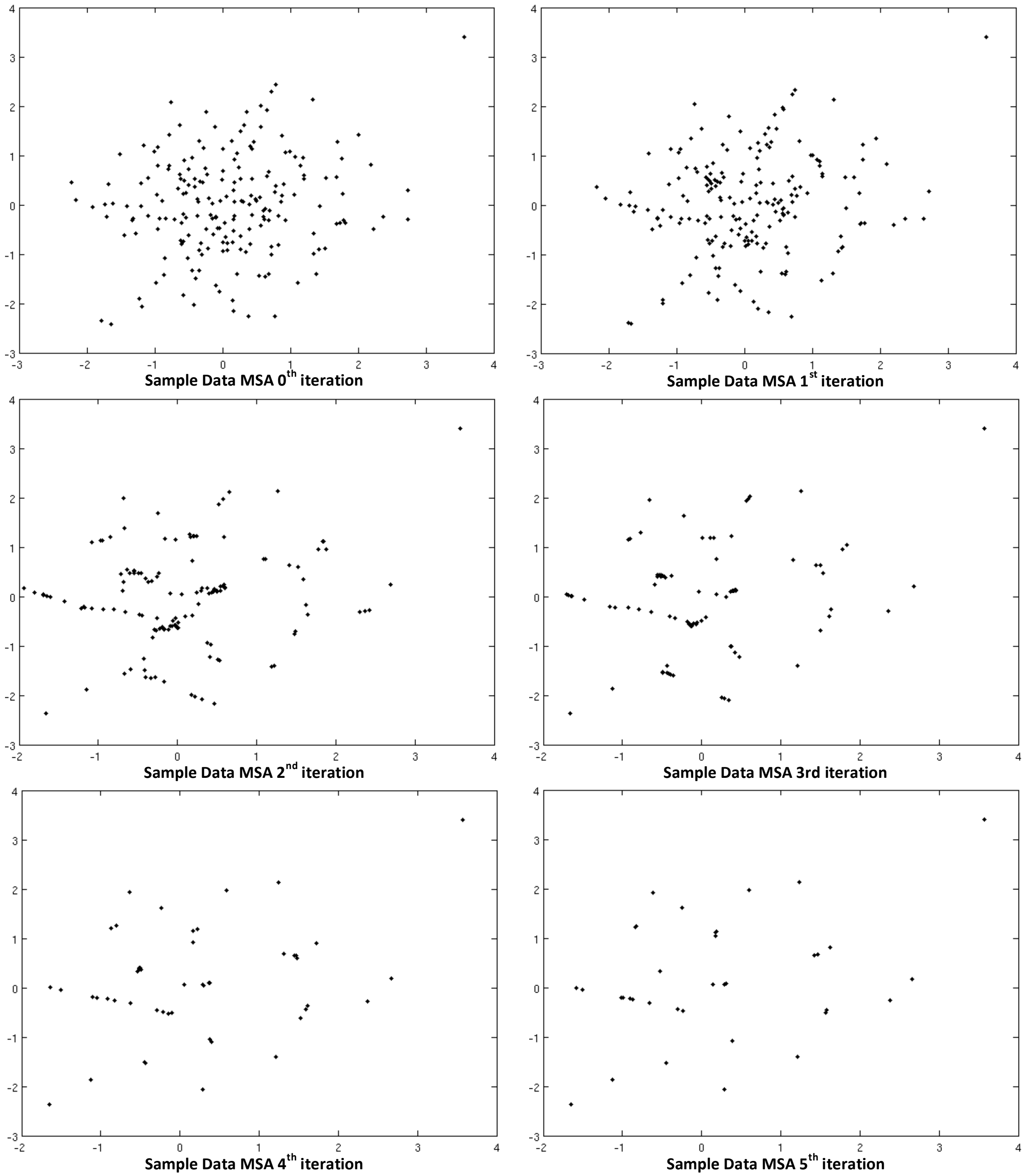

Figure 5: MSA Example using Random Sample Data and Bandwidth Kernel $10^{-1}/\ I_{2\times 2}$ for 0th through 5th iterations

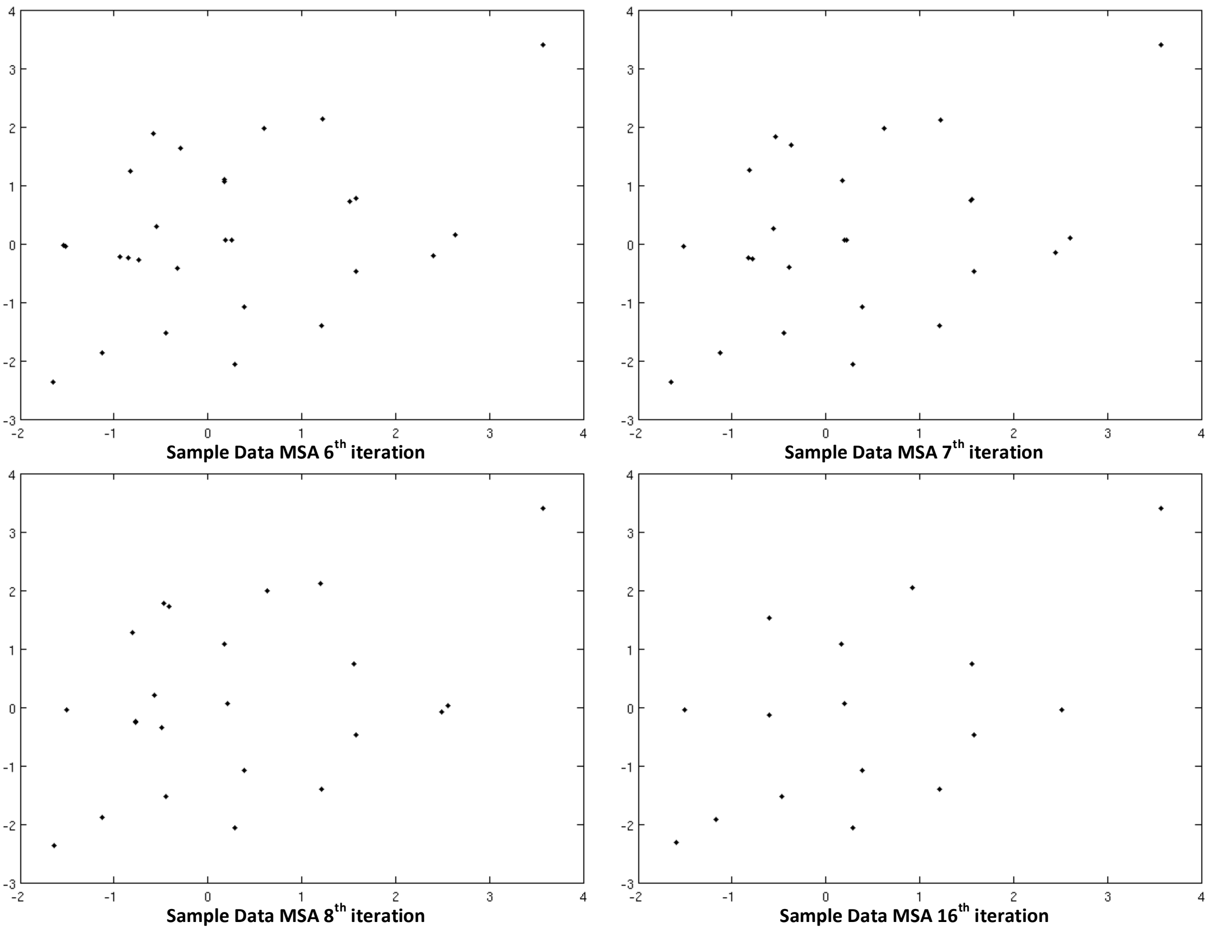

Figure 6: MSA Example using Random Sample Data and Bandwidth Kernel $10^{-1}/\ I_{2\times 2}$ for 6th, 7th, 8th and 16th iterations

This clustering algorithm was then applied to each of the 77 counties in the state of Oklahoma. The clustering algorithm could have been applied to the entire state, however the computational complexity would have been prohibitive and unnecessary since $e^{-\frac{(x-x_i)^T\Sigma^{-1}(x-x_i)}{2}}$ is effectively zero for an appropriate kernel size for this analysis.

The initial kernel selected was $2.6504\text{e}^{-05} \times \text{I}_{2\times 2}$ where $\text{I}_{2\times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. This worked for many of the rural counties, and was selected because 2.6504e^{-05} is approximately one third of a mile in orthodromic distance. In accident dense counties such as Oklahoma County or Tulsa County or in some urban areas 2.6504e^{-05} was adjusted to 2.6504e^{-07} due to over clustering. Over clustering is the effect of having too many underlying data-points densely located, as the algorithm associates more together and accretes too large an area to be considered local for our analysis purpose. A method was developed to overcome this limitation. Each cluster was evaluated for the maximal distance between the cluster centroid (geometric mean) and individual cluster elements. If this distance exceeds the desired upper limit for distance the clustering algorithm is reapplied decreasing the size of the kernel. In future development and research a more robust search algorithm will be developed to optimize the kernel bandwidth.

In this situation the Kernel bandwidth is adjusted down and the algorithm is rerun.

After the data is clustered the number of collisions associated with each cluster can then be analyzed. The raw counts may be plotted as in Figure 7 or as the ratio of the number of collisions and the Average Daily Traffic (ADT). Due to the mass of this data the researchers are looking at integrating this into already existing Intelligent Transportation Systems (ITS) data visualization systems.

One such asset is the Collision Explorer on the Statewide Analysis for Engineering & Technology (SAFE-T) which allows traffic management stake holders to visualize and browse such data. Techniques developed through this project can be applied to this system. Comma Separated Value (CSV) files for this data can also be obtained by contacting the ITS lab at the University of Oklahoma: http://its.ou.edu/.

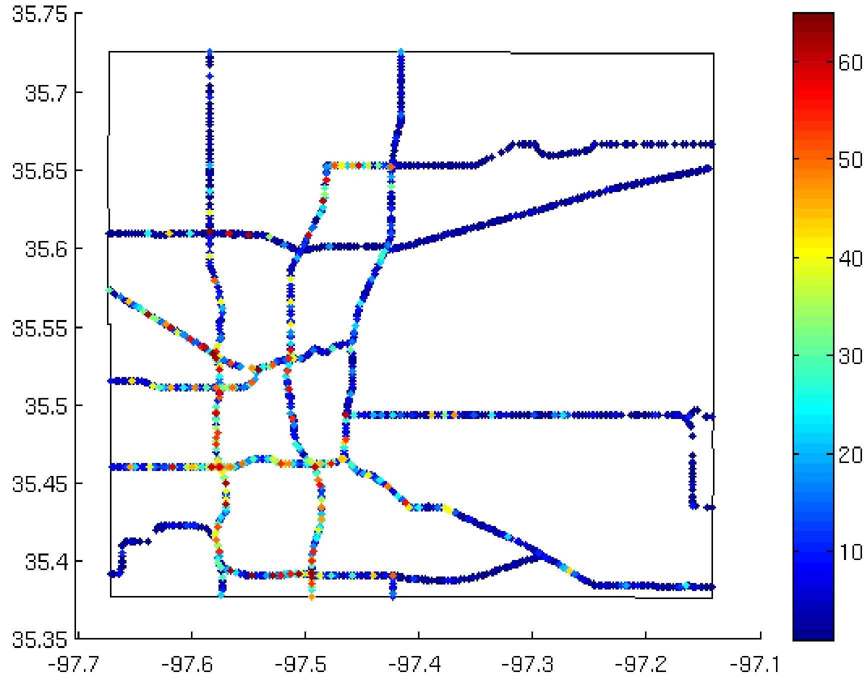

Figure 7: Oklahoma County Blackspot Locations and Intensity (in number of collisions) from 1998 to 2012 without ADT adjustment

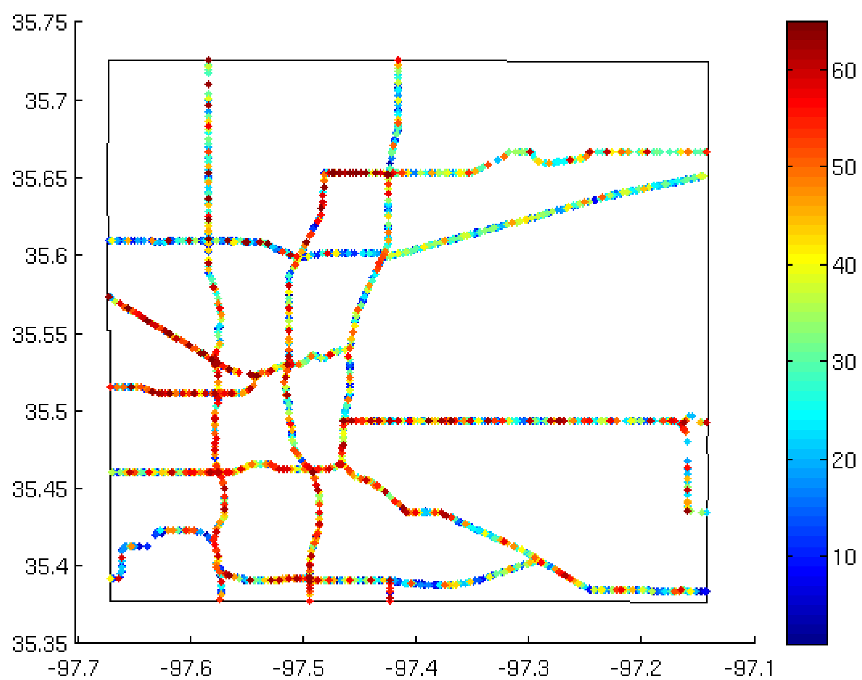

Figure 8: Oklahoma County Blackspot Locations and Intensity (in number of collisions) from 1998 to 2012 with ADT adjustment

This analysis has been complete for all counties and at the completion of this project a complete set of data as seen in Table 7 will be available for download in comma separated file format. This will allow the interested party to use this data with widely available tools like Google maps. For instance, a user of this data can copy and paste the latitude and longitude from a file and Google maps will show the location in question.

If the user has access to oksafe-t.org they can copy and paste the lat/long into the search bar of the collision explorer and analyze the nature of the local collisions around the clustering point. It should be noted that the point at which a cluster is located is not a precise location but indicates an area around which traffic dynamics should be investigated.

In Figure 9 we see a screenshot of the Collision Explorer which was developed for traffic stakeholders to explore collisions. The underlying mapping engine is Google Maps and each of the circles indicated in the mapping software are clusters of collisions and the color coding specifies which type of collision. For city streets the data is being clustered at the nearest intersections as reported by the municipal agent reporting the collision. For highways and major arterials a different clustering algorithm is applied. However, at the time of writing blackspot analyses have not been integrated into this tool. One possible model for this is the clustering algorithms discussed above. What can be done is that the clustering algorithm is run on the data set and the clustering information stored in a separate table. This data could then be explored. However, the details of the clustering and how they would be explored needs to be discussed with stakeholders to address the scale, kernel densities and usability of this information.

Figure 10 shows the collision explorer at the local scale. At this scale the object become rectangles with the number of collisions at each point being displayed within the color coding by type of collision. The user can click on an icon and the information on aggregate data will appear in a white comment box as can be seen for the intersection of Western & 96th. Also when an icon has been selected the panel on the right side of the screen lists pertinent collision information and a link which provides the complete collision information. This offers a promising new course of inquiry and development of assets for traffic management stakeholders on the Federal, state, regional and local levels for Oklahoma.

This tool as well as other mapping programs, such as Google Maps, offers the traffic manager to explore the black spots quickly via CSV files as the Lat/Long information can be taken from these files and pasted into the Enter Location bar of Collision Explorer, or the search bar of Google Maps.

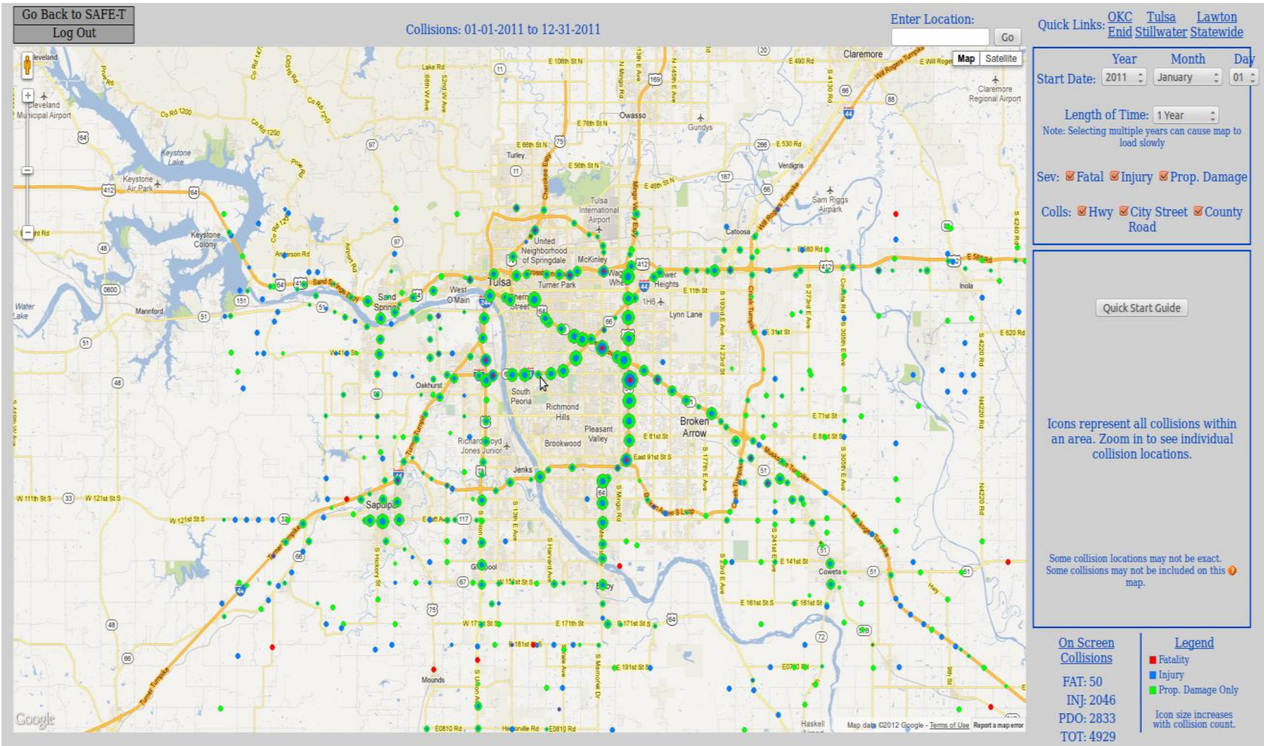

Figure 9: Screenshot of Tulsa County in Collision Explorer (http://www.oksafe-t.org/). Green indicates property damage only, blue indicates injuries and red indicates fatality collisions. Each radius indicates number of each type of collision.

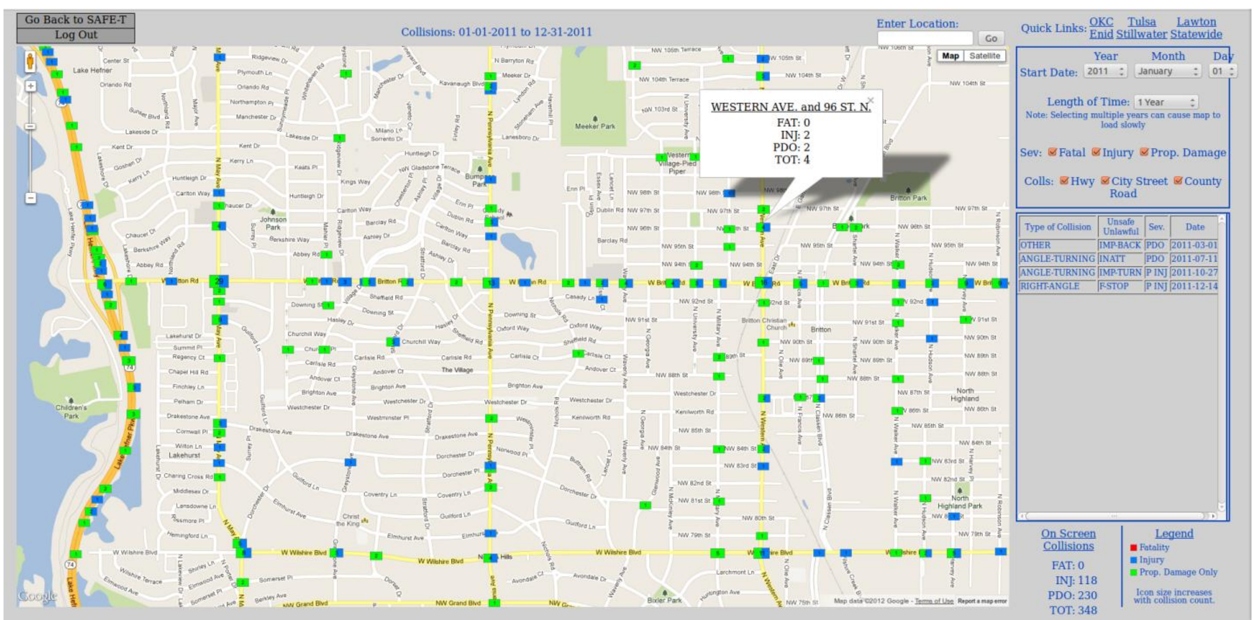

Figure 10: Screenshot of Oklahoma City Surface Streets in Collision Explorer (http://www.oksafe-t.org/). Green indicates property damage only, blue indicates injuries and red indicates fatality collisions. Size of indicates number of each type of collision.

Table 7: Sample Analysis Data

CC	CC/ADT	Lat.	Lon.	2007	2008	2009	2010	2011	2012	Trend
847	0.009049	35.522	-97.542	53	104	92	69	56	44	0.62214
782	0.007534	35.391	-97.548	42	62	63	61	40	37	0.62172
718	0.007067	35.393	-97.495	38	52	32	26	32	22	0.62129
685	0.012077	35.533	-97.582	38	52	41	57	49	41	0.62087
668	0.02672	35.611	-97.585	77	64	49	33	29	16	0.62044
649	0.006149	35.391	-97.53	41	57	49	52	34	32	0.62002
649	0.005337	35.392	-97.507	48	49	44	56	41	36	0.6196
637	0.007053	35.39	-97.495	56	62	122	136	66	2	0.61917
573	0.0076	35.394	-97.566	45	34	41	43	28	22	0.61875
565	0.01434	35.558	-97.639	38	36	38	44	44	59	0.61832
564	0.005698	35.53	-97.579	60	74	75	47	47	40	0.6179
549	0.010619	35.609	-97.55	62	62	56	23	29	10	0.61747
543	0.004973	35.45	-97.57	42	37	48	54	52	40	0.61705
538	0.011185	35.63	-97.484	46	31	45	29	28	21	0.61663
526	0.010988	35.549	-97.619	39	31	45	44	29	16	0.6162
524	0.005024	35.436	-97.57	34	37	55	41	31	34	0.61578
507	0.009054	35.537	-97.59	31	35	38	43	43	36	0.61535
499	0.004883	35.46	-97.583	26	31	32	22	22	21	0.61493
490	0.003182	35.495	-97.579	24	21	47	26	30	27	0.6145
490	0.00802	35.494	-97.465	30	39	34	33	26	16	0.61408
485	0.003617	35.39	-97.495	46	46	29	15	27	10	0.61366
464	0.009033	35.46	-97.492	19	35	32	43	56	38	0.61323
462	0.009027	35.609	-97.497	37	32	16	23	11	13	0.61281
454	0.005781	35.392	-97.495	15	21	25	17	45	94	0.61238
436	0.025497	35.653	-97.453	39	27	39	26	38	13	0.61196
420	0.004094	35.46	-97.617	42	34	31	31	28	22	0.61153
414	0.003753	35.522	-97.577	37	55	76	46	38	16	0.61111
414	0.010781	35.609	-97.568	35	35	37	21	10	5	0.61069
413	0.003719	35.46	-97.577	13	17	34	34	8	9	0.61026
408	0.002923	35.508	-97.576	34	19	32	36	45	27	0.60984

CC: Collision Count, CC/ADT: Collision Count Divided by ADT, Lat., Lon. Collision Counts by Year 2007-2012, Trend

Systemic Analysis

A second analytical technique is the systematic approach which is described in detail in [4] which discusses two off the shelf analytical software packages. The "Systemic" approach looks at the overall traffic network and helps determining areas where improvement will be beneficial.

The first mentioned in [4] is called SafetyAnalyst by AASHTOWare which prices at $15,000 *per annum* for a workstation license and $25,000 *per annum* for a site license. All of these software packages require that the data be centralized which during the project we have made significant progress in creating a system that does this with all available data. As per [19], the minimal data set is crash and traffic data with either roadway segment, intersection or ramp data. For crash data the relevant information is date, location, type, severity, relationship to junction and maneuvers by involved vehicles. All of this data is now easily accessible with the possible exception that relationship to junction and maneuvers by involved vehicles may require a translation between the current formats being used in the database to a format acceptable by the software.

Further this tool requires a numbering system for the segments, location, length, designation rural or urban, median type, access control (toll/freeway), average annual daily traffic, and one or two way traffic. All of this data are available to the end user. For surface streets of interest to regional and local planning agencies a similar data set is required for intersection information. The majority of this data exists, however Major-road traffic volume (AADT) and Minor-road traffic volume are not available for most municipalities. At the time of writing [19] lists 14 US states and one Canadian province as licensees.

The second program mentioned in [4] is a pilot project called United States Road Assessment Program which is a project of the AAA and a conglomeration of Federal, State and Local municipalities. USRAP provides mapping information to municipalities. They provide two types of information maps the first Star Rating which ranks roads based on engineering features and configurations. A second type of map is directed at the commuting public and is a risk assessment to the commuter given statistics about current problems. The AAA Foundation for Traffic Safety considers collaborations with public entities and can be contacted via usRAP@aaafoundation.org [20]. The current state of this project is the completion of a three phase pilot project and claims that the technologies are at a mature state. In the author's opinion contacting this information and exploring this as a potential low cost option is highly recommended. Phase III risk maps are currently available for IL, KT, NM, and UT [20].

Another possible third party solution is the use of a third party Dashboard management information system and integrating that into existing databases. These systems are designed for ease of use and integration into already existing data sets. They provide easy to access, generally web based visualizations such as tableau software [21]. They are useful and quick to set up but are not specialized for the types of data dealt with in Traffic Engineering. However, many can be adapted or modified to be effective on the data in question. However, they have a price range of around $2,000 and considering the customization costs may not be cost effective for the types of analyses required.

If a set of systematic analyses can be determined beforehand software can be developed to periodically perform the analysis and the results can be made available to stakeholders, for a relatively low cost though the first two options may highly cost competitive.

Other Data Sources and Analyses

Other data sources that are not often considered though may yield a lot of useful information is also available to the data analyst. One of the data sets archived at the Oklahoma University ITS lab is the output of the Remote Traffic Microwave Sensors (RTMS) located in the Oklahoma City and Tulsa area. RTMS measure traffic speed and volume data aggregated on a two minute basis. This aggregated data has been shown to be useful in predicting travel times [22]. One question that recently arose was whether it was feasible to find clearance time for accidents. One of the benefits of this knowledge is that the clearance time can be used to find out how many accidents are coincidental to other accidents. Another, analysis can be made as to how roadway incidents affect traffic given the event of a collision.

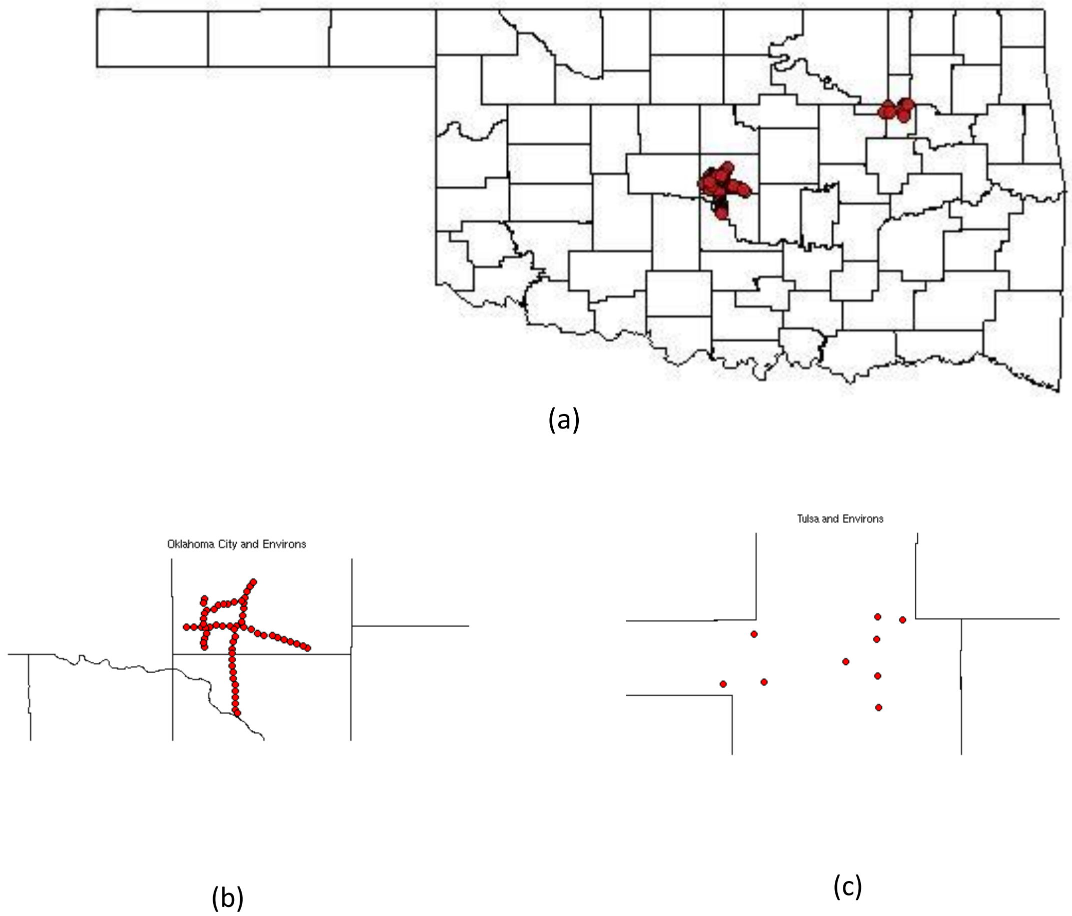

Figure 11: Locations of Remote Traffic Microwave Sensors (a) Statewide, (b) WIthin Oklahoma and Cleveland County, and (c) Within Tulsa County

The first problem that needs to be addressed is registering the collisions to the locations of data sensors. Since these sensors are collocated with the major arterials this data can be made attributable to collisions that can in some way be registered to a sensor. To begin researching integrating this data set collisions that where within one third of a mile of an RTMS sensor were selected for analysis. There is then still a challenge in determining which lane or direction the vehicles involved in the collision are traveling.

Then each of these collisions can be associated with the speed, volume and occupancy data that these sensors provide. For our purposes here we will discuss only collisions after 4-28-2011 when the data collection processes were restructured for ease of searching. These collisions are plotted along with their location with respect to the nearby RTMS sensor in Figure 12. However, this data is archived for several years prior to this date.

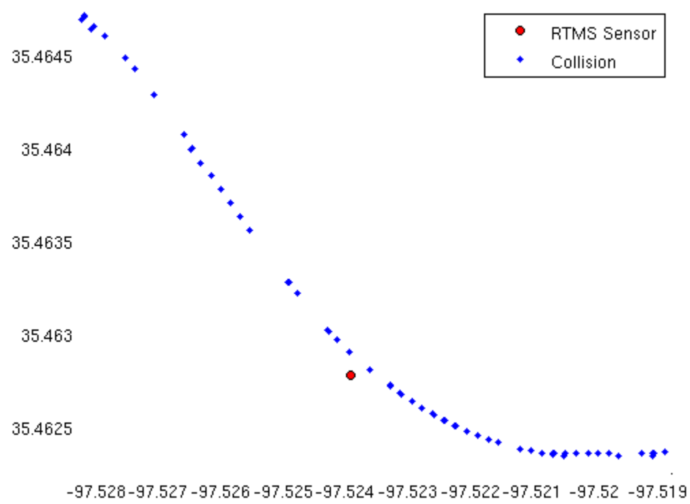

Figure 12: Plot of collisions near RTMS sensor

After the collisions are registered with the RTMS sensors the speed, volume and occupancy (SVO) data can be retrieved from the database. To explore the efficacy of this data association SVO data was collected 45 minutes prior to and 45 minutes after the reported time of the collision. All three of these data types can be visualized easily but we will restrict our focus on only speed data. Another useful piece of information is what the average data is for this location for the time of day. In this way, changes in traffic flow information can deseasonalized from naturally occurring daily traffic flow variation. An example of average daily traffic flow can be seen in Figure 13.

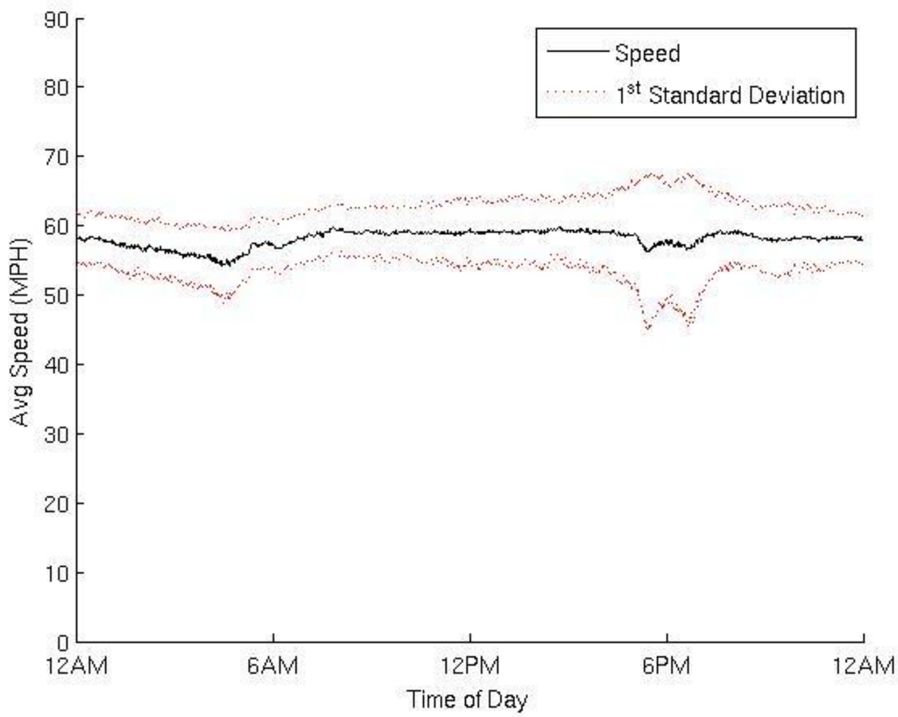

Figure 13: Daily average speed and standard deviation per time of day for one RTMS sensor and one lane of traffic

Then all six lanes of traffic can be visualized in the same plot, for ease of visualization one of the directions of traffic is negated so that the directions are visually separated. A study into the feasibility of using this data to determine accurately the most likely time of the collision as well as the lane and direction that most the collision most likely occurred in, follows.

At the time of writing there is not a ground truth data set to evaluate collision detection algorithms, however developing this data set and implementing this algorithm and making the results available to traffic managers and engineers would be a promising.

In many cases there is not a clear signal as to when the collision occurred as can be seen in Figure 14. While others present relatively clear signals as in Figure 15, in which it is not clear if the collision occurred while another collision was in the process of being cleared or that the report of the time of collision was in error. To estimate the time and lane of collision the cumulative change in speed per lane was calculated and zero crossings were used to estimate the time at which a collision occurred or traffic resumed free-flow after the collision. The area between the curve in Figure 16 and the x-axis was used to estimate the magnitude of the speed change for each event.

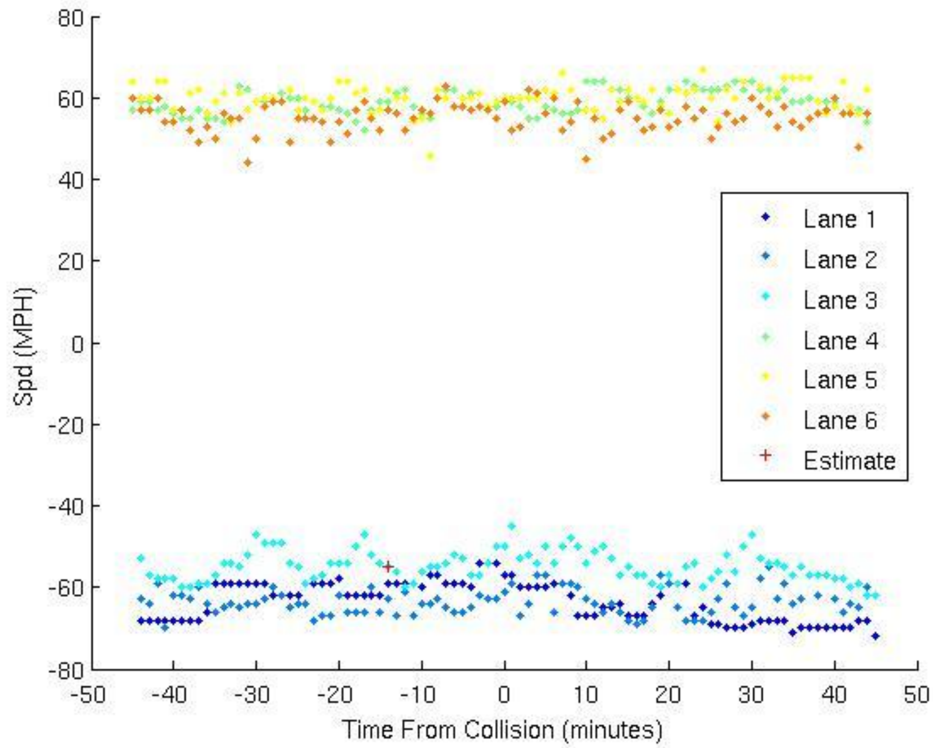

Figure 14: Speed plot of six lanes for sample collision with low apparent signal

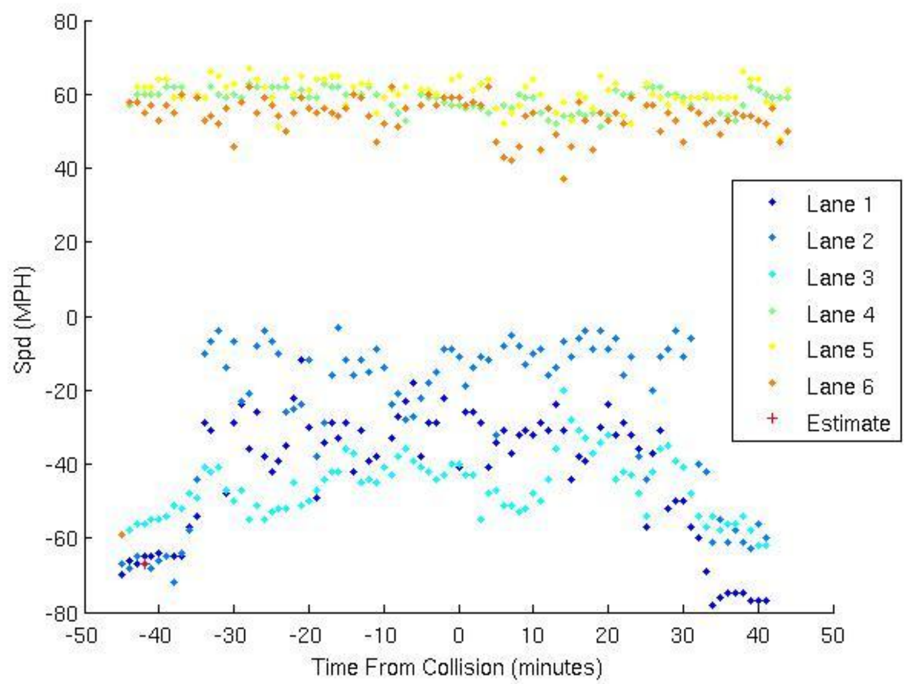

Figure 15: Speed plot of six lanes for sample collision with high apparent signal

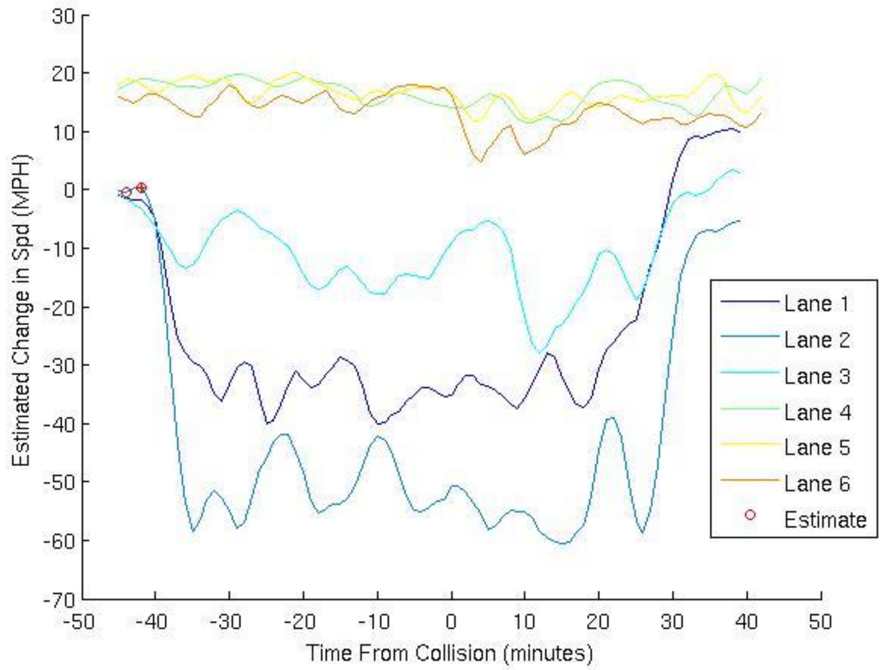

Figure 16: Cumulative speed change per lane

Building a ground truth data set would require investigating some subset of collisions and determining the conditions around the collision and looking at collisions possibly on the same roadway recent to the current collision in question. Much of this information can be extracted from crash diagrams but that would require a human to evaluate the data set for each point. This would be a fairly labor intensive undertaking.

Another commonly occurring event is that traffic comes to a complete halt, or a catastrophic failure of traffic flow. In this event the algorithm has to determine why the sensor is not reading any traffic flow. It is quite possible that on occasion due to the volume of traffic that the sensor is naturally not reading. It is also quite possible that this is coincidental with a traffic event. As a preliminary experiment any lane that read more than four consecutive no-reads was considered to have failed catastrophically, bicubic interpolation was used for no read intervals smaller less than four consecutive intervals. A better strategy for evaluating catastrophic failure is perhaps analyzing the statistical probability of the occurrence of non-reads at that time daily. Then the probability that a given event is a collision, i.e. catastrophic failure could be evaluated.

More advanced statistical and system identification techniques could be applied to this data set. For example a possible approach that was not explored fully was modeling the traffic flow at each minute as a three dimensional Gaussian distribution and thresholding the probability below some unlikely value, say two standard deviations. Fixing this threshold would require further investigation.

The University of Oklahoma ITS lab continues to develop methods to allow the maximal exploitation of data from out Intelligent Transportation Systems and to make this data available to traffic stakeholders.

Conclusions

In this project the Intelligent Transportation laboratory at the University of Oklahoma created a new resource enabling data analysis to determine correlations between roadway characteristics and fatality collisions in the state of Oklahoma. A new, single database based on corresponding geographical information was developed by translating roadway characteristics as recorded in the Oklahoma roadway inventory database and fatality collision data as recorded in the national FARS database into a common database format. These two data sets were then merged into a single database based on geographic location data. Each database uses different geographical location systems, so algorithms had to be developed to match location information between the two systems thus enabling the merger of the two databases.

The new, combined database was shared with researchers working on a similar project enabling them to extensively query and analyze the available data to discover relationships between roadway characteristics and the occurrence of fatality collisions without having to duplicate the significant we performed to bring the disparate data sets into a common format.

Example analyses including blackspot and systemic analysis as well as correlative studies with additional data sources were performed to demonstrate the utility of the newly developed capability. The results of this project will allow traffic engineers and highway planners to utilize this analysis capability to more selectively improve roadways based upon the likelihood of fatality collisions, thus using resources effectively to save lives throughout the state.

References

[1] C. W. Niessner, "Alternative Strategies for Safety Improvement Investments," *Transportation Research Board: Research Results Digest,* vol. 345, pp. 1-28, April 2010.

[2] G. Khan, K. R. Santiago-Chaparro, M. Chiturri and D. A. Noyce, "Development of Data Collection and Integration Framework for Road Inventory Data," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2160, pp. 29-39, 2010.

[3] NHTSA, "Fatality Analysis Reporting System (FARS)," 2012. [Online]. Available: http://www.nhtsa.gov/FARS. [Accessed October 2012].

[4] F. M. Council, D. W. Harwood, B. H. Delucia, R. C. Peck and T. R. Neuman, "Report 500: Guidance for Implementation of the AASHTO Strategic Highway Safety Plan Transportation Research," National Cooperative Highway Research Program, 2008.

[5] A. A. Sebastian and D. M. Flemons, "NHTSA'S FATAL ACCIDENT REPORTING SYSTEM," *Transportation Research News,* no. 97, pp. 5-9, 1981.

[6] B. Lovelace and J. Staha, "Traffic Records Analysis in Texas," *Transportation Research Circular,* no. 233, pp. 7-9, 1981.

[7] NHTSA, "FARS Encyclopedia," October 2012. [Online]. Available: http://www-fars.nhtsa.dot.gov/. [Accessed October 2012].

[8] National Highway Traffic Safety Administration (NHTSA), "FARS Analytic Reference Guide 1975 to 2009," US DOT HS 811 352, 2010.

[9] National Highway Traffic Safety Administration (NHTSA), "National Automotive Sampling System (NASS)," [Online]. Available: http://www.nhtsa.gov/NASS. [Accessed December 2012].

[10] OU ITS Laboratory, "SAFE-T: Statewide Analysis for Engineering & Technology," 2012. [Online]. Available: http://www.oksafe-t.org/. [Accessed October 2012].

[11] MySQL.com, "MySQL: The world's most popular open source database," 2012. [Online]. Available: http://www.mysql.com/. [Accessed October 2012].

[12] J. C. Sutton and M. M. Wyman, "Dynamic Location: An Iconic Model to Synchronize Temporal and Spatial Transportation Data," *Transportation Research Part C: Emerging Technologies,* vol. 8, no. 1, pp. 37-52, 2000.

[13] R. Moonesinghe, A. Longthorne, U. Shankar, S. S. R. Singh and J. Tessmar, "An Analysis of Fatal

Large Truck Crashes," US DOT HS 809 569, 2003.

[14] S. Hallmark, M. Orellana and E. J. M. T. a. M. D. Fitzsimmons, "Red Light Running in Iowa," *Journal of the Transportation Research Board,* vol. 2182, pp. 48-54, 2010.

[15] T. D. Walden, "Evaluation of Photographic Traffic Signal Enforcement Systems in Texas," Traffic Operations Division, Texas Department of Transportation, 2012.

[16] H. Preston, W. Stein, T. Maze and R. Souleyrette, "White Paper on Alternative Strategies for Safety Improvement Investments," Research results digest NCHRP Project 17-18(19), 2010.

[17] H. Chen, "The Use of Relationships Between Road Accidents and Traffic Flow in "Black-Spot" Identification," *Traffic Engineering & Control,* vol. 22, no. 8/9, 1981.

[18] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognitionc," *IEEE Transactions on Information Theory,* vol. 21, pp. 32-40, 1975.

[19] AASHTOWare, "Data, Hardware and Software Requirements," AASHTOWare, [Online]. Available: http://www.aashtoware.org/Safety/Pages/Hardware-Software-Requirements.aspx. [Accessed 29 December 2012].

[20] AAA Foundation For Traffic Safety, "usRAP (United States Road Assessment Program)," AAA, [Online]. Available: http://www.usrap.us. [Accessed 17 October 2012].

[21] "tableau software: Business Dashboards," [Online]. Available: http://www.tableausoftware.com/solutions/business-dashboards. [Accessed 29 December 2012].

[22] J. Junger, J. P. Havlicek, R. D. Barnes and M. P. Tull, "Prediction aggregation of remote traffic microwave sensors speed and volume data," in *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems*, 2009.

[23] AASHTOWare, "Data, Hardware and Software Requirements," AASHTOWare, [Online]. Available: http://www.aashtoware.org/Safety/Pages/Hardware-Software-Requirements.aspx. [Accessed 17 December 2012].