Driver Speed vs. Crash Risk:
Pilot Study Summary

March
1998

Texas Transportation Institute

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle		5. Report Date
Driver Speed vs. Crash Risk: Pilot Study Summary		March 24, 1998
		6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Jennifer Harper Ogle, Olga Pendleton, Lindsay Griffin, Paul Carlson	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Texas Transportation Institute CE/TTI Bldg., Texas A&M University College Station, TX 77843-3135	
	11. Contract or Grant No.
	Project No. DTNH22-97-C-05141

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Office of Research and Traffic Records National Highway Traffic Safety Administration 400 7th Street SW Washington, DC 20590	Final Report September 1997-March 1999
	14. Sponsoring Agency Code

15. Supplementary Notes

Contracting Officers Technical Representative: Paul Tremont, Ph.D., NHTSA

16. Abstract

This report summarizes the development and pilot test of an automated data collection system that utilizes Global Positioning System (GPS) technology to continuously measure individual driver speed. With little previous research having been carried out on the measurement of normal individual driver behavior, the propensity to speed, and the relationship of these two factors on crash risk, it is hard to determine whether or not people who might be categorized as "speeders" are actually at greater risk of crash involvement. The National Highway Traffic Safety Administration sponsored this pilot study in an effort to advance the understanding of the relationship between the speeds at which motorists drive and crash probability. The resulting approach offers a unique microscopic view of individual driving behavior, and a source of accurate pre-crash travel speed information. Ultimately, this approach could be used on a larger sample to determine the crash risk relationship for various sub-populations and help to identify potential countermeasures for these groups.

For this pilot study, 50 vehicles were equipped with advanced vehicle tracking devices which facilitated the continuous monitoring and logging of speed and location data for a period of four months. While this sample of participants was not large enough to draw statistically significant conclusions about the relationship between speed and crash risk, it did provide for a proof-of-concept of the technology and a potential approach for quantifying the relationship in a full-scale study. This report will provide findings on: 1)proof-of-concept of technology for this application; 2) liability issues involved in obtaining vehicle tracking information; 3) effectiveness of various data acquisition rates; 4) elements of the relationship between driver behavior and the roadway environment; 5) issues associated with tracking crashes; and 6) recommendations for a full-scale study.

17. Key Words	18. Distribution Statement
Speed, crash risk, instrumented vehicles, global positioning system (GPS), pre-crash speed, safety, crash probability	No restrictions. This document is available to the public through NTIS: National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	28	

Form DOT F 1700.7 (8-72)   Reproduction of completed page authorized

LIST OF FIGURES

LIST OF TABLES

DRIVER SPEED VS. CRASH RISK: PILOT STUDY SUMMARY

This report summarizes the development and pilot test of an automated data collection system that utilizes Global Positioning System (GPS) technology to continuously measure individual driver speed. With little previous research having been carried out on the measurement of normal individual driver behavior, the propensity to speed, and the relationship of these two factors on crash risk, it is hard to determine whether or not people who might be categorized as "speeders" are actually at greater risk of crash involvement. The National Highway Traffic Safety Administration sponsored this pilot study in an effort to advance the understanding of the relationship between the speeds at which motorists drive and crash probability. The resulting approach offers a unique microscopic view of individual driving behavior and a source of accurate pre-crash travel speed information. Ultimately, this approach could be used on a larger sample to determine the crash risk relationship for various sub-populations and help to identify potential countermeasures for these groups.

PROBLEM STATEMENT

In September of 1997, NHTSA funded a pilot study to revisit the relationship between speeding and/or speed differential and crash risk. The focus of the pilot study was to examine the feasibility of obtaining speed profiles of individual drivers based on their overall driving pattern and then to associate this driving pattern (profile) with crash probability. The solicitation required a dual approach: 1) instrument a group of vehicles to continuously monitor and log driver speed and location; and 2) for crash-involved vehicles, compare measured pre-crash speed with actual speed distributions collected at the crash site using a case-control approach. At the end of the study, speed profiles would be developed for crash vehicles as well as non-crash vehicles based on continual tracking of the relationship between the individual's actual speed and the posted speed limit over the study period. The profiles should provide insight regarding driver speed behavior and help to answer some of the following questions:

- Are drivers who routinely speed more likely to be involved in crashes than drivers who observe speed limits?
- Do drivers who speed subject themselves to higher risks of crashing under certain circumstances only?
- What is the nature of the change in the relationship between crash risk and speed as other suspected key variables change (e.g., driver characteristics, roadway type, weather, traffic density, average traffic speed)?
- Under what conditions does exceeding the speed limit lead to crashes?

Answers to these and other questions are needed to come to a full understanding of the relationship between speeding and crash risk under varying conditions. This knowledge would allow for the development of varying countermeasure programs that could be targeted at controlling speeds in those situations when the risks of crashing associated with these circumstances are greatest. The information can also be used by state and local agencies to help

in setting speed limits that reflect the actual levels of risk associated with these circumstances and in setting penalties appropriate to the actual risks incurred by speeders.

STUDY DESIGN

The study design was limited by the short time-frame and limited budget associated with the pilot study. For these reasons, there were inherent design constraints that limit the usefulness of the final data set. The first design constraint was that of the selection and size of the sample population. Fifty test participants were obtained through a centralized volunteer (no monetary incentives) program at USAA Insurance Company in San Antonio, Texas. This was done to reduce the logistical considerations, as well as expedite the installation of equipment and participant training under the short time-frame. The process of limiting the participants to USAA volunteers created an initial self-selection bias to the statistical conclusions of the pilot study. Despite these constraints, the pilot study did provide for a proof-of-concept of technology and a potential approach for quantifying the relationship in a full-scale setting.

The second major design constraint was the lack of observed accident data from which to develop relationships between accident risk and driving behavior. Theoretical expectations, based on past accident rates in San Antonio, predicted that 1.4 crashes would be observed within the four-month field test. The researchers were aware of the possibility that no accidents may occur during the field test. This possibility also extends to any full-scale study although much less likely due to the proposed sample size of 2,000 to 4,000 participants.

Participants

The criteria used in selecting study participants were as follows:
- Both male and female drivers, age 18 and older;
- Possession of a current, valid drivers license with no restrictions;
- Possession of a personal vehicle--driven approximately 90 percent or more of the time by the subject;
- Residence within the boundary of U.S. Loop 1604 in San Antonio, Texas;
- Residence in Texas for more than three years (preferred, but not required); and
- Ability to read and write (to answer written surveys).

Participant Survey

A pre-study survey was administered to all participants to obtain vehicle and driver characteristics. These surveys were used as supplemental information which was included in the analysis. The surveys gathered general information regarding the age, sex, race, marital status, and education level of the participants. With the realization that no crashes might be observed during the field test, the participants were also asked a series of questions regarding behavioral issues such as seatbelt use, accident experience, propensity to speed, as well as alcohol use. The data were used in lieu of observed accident data to establish risk-taking behavior patterns among the participants.

Use of Human Subjects Review

The continuous collection of speed and location data produces a detailed account of an individual's driving patterns, behaviors, and activities. Recording such information leads to concerns of potential liability on the part of the university, the researchers, and the participants. To protect all involved parties, the Texas A&M Internal Review Board (IRB) requested that the researchers seek consideration for a Certificate of Confidentiality from the National Institute of Health (NIH).

A certificate was awarded that protects the researchers from being summoned to court, and also protects against the summons of data regarding any participant that is maintained in the research files. Unfortunately, the protection of data maintained in the vehicle recording device and data being transferred between the vehicle and the research center is not clearly defined. Such unknowns were noted in the consent form that participants were required to sign.

Equipment

As noted earlier, there were two distinct data collection activities: 1) continuous measurement of driver speed and location; and 2) measurement of speed behavior of the population at observed crash sites. These data collection activities required two different and independent equipment sets and field test methods – one for vehicle data collection, and the other for crash site data collection.

Vehicle Data Collection Equipment

Advanced technologies were used to obtain continuous speed and location data as well as accurate pre-crash speeds for the 50 test vehicles. Since the 1960s and 1970s, there have been significant technological advances in vehicle tracking equipment. The vehicle tracking system used in the pilot study was primarily based upon GPS technology. GPS is a satellite-based navigation system operated and maintained 24 hours a day by the U.S. Department of Defense.

The tracking technology product used in the pilot study, named "OnGuard LS(D), " is designed as an asset location and tracking system. The device may be added to vehicles, trucks, tractors, trailers, and earth moving equipment. The regular function on the equipment is to determine the location of an asset anywhere in the country that is covered by cellular service.

The system consists of an internal three-watt cellular phone, a 12-channel Rockwell Jupiter GPS receiver, a modem, a differential corrections signal receiver, two microprocessors, and a collection of input/output ports and miscellaneous electronic circuitry. The system used in the pilot study also offers an extended memory option. This feature added data storage capability for the logging of positions, speed, and other vehicle parameters.

The unit was originally designed to be housed in a single enclosure. Due to accuracy requirements, however, a differential corrections receiver was added through a resident serial port on the central processing unit. The differential corrections receiver upgrades the accuracy of the GPS receiver from 100-meter spherical error probability to 2-to 5-meter spherical error probability (this met the map-matching needs of the study without requiring extensive

algorithms). The size of the unit was kept as small as possible with a minimal amount of external wiring to facilitate installation. External hookups were limited to power, GPS and cellular antennas, an ignition sensor, and a backup battery. The equipment was controlled through the ignition sensor and required no interaction from the participants.

Crash Site Data Collection Equipment

The type of equipment used to measure the speed behavior of the population at observed crash sites was dependent on the type of facility (freeway or non-freeway), and the availability of Advanced Traffic Management System (ATMS) technology deployment. Three types of equipment were used in the pilot study: 1) ATMS technology (consisting of inductance loop detectors), 2) laser speed detection guns, and 3) piezo sensors. The first equipment option, detection technologies deployed within the San Antonio TransGuide ATMS coverage area, were used to obtain speed data for crash sites located within the freeway limits shown in Figure 1. Laser speed detection guns were used for crashes occurring on freeway locations with no ATMS sensor deployment. The final type of crash site speed detection device, the piezo-electric sensor, was used to collect crash site speed data on all non-freeway type facilities.
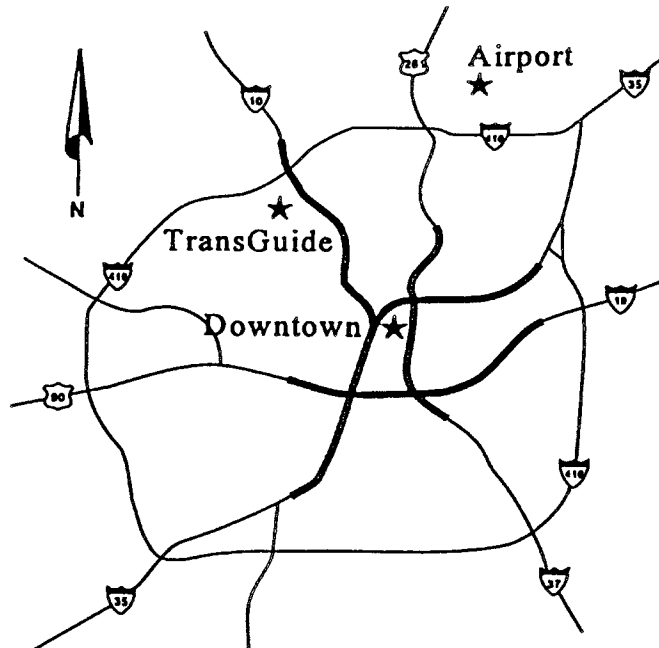
Figure 1. San Antonio TransGuide ATMS coverage area.

ATMS Technology

For freeway sites, the San Antonio TransGuide Advanced Transportation Management System (ATMS) was utilized. The TransGuide ATMS monitors 42 kilometers (26 miles) of freeway. With traffic data stations located approximately every 0.8 kilometer (one-half mile), this coverage establishes nearly 120 permanent data collection sites where travel speed, volume,

6

and inductive loop occupancy data are collected every 20 seconds. The traffic data stations consist of inductance loop detectors buried in the road surface.

Laser Speed Detection Gun

For freeway sites outside of the ATMS coverage area, a portable laser gun technology was used to collect speed data. The laser gun is manually controlled, similar to those used by police in regular speed enforcement. This type of device provided high accuracy under short-term use without the need and expense of shutting down freeway operations to install a more permanent type of speed measurement device. The device can be used from freeway shoulders or overhead bridges, and provides a safe and inexpensive detection source. Unlike the ATMS detection technologies that continuously log data, this device is employed in a case-control approach where data are manually collected during the day and time of the crash during the weeks following the crash.

Piezo Sensors

To collect the vehicle speed, headway, and classification data for non-freeway sites, class II piezo-electric sensors were used in conjunction with traffic counters/classifiers (TCC). The piezo sensors are also utilized in a case-control methodology following the crash. These sensors are semi-permanent and can be installed and left in the field for several weeks. This allows for the development of a more comprehensive data source than the laser speed gun technology, but they too are limited to post-crash implementation.

Data Collection

Vehicle Data Collection

The vehicle data collection field test was conducted for four months between April 1, 1998 and July 31, 1998. The subcontractor, ATX Technologies, Inc., managed the day-to-day operations of the field test. They were responsible for downloading the data from the vehicles, changing the sampling frequencies, and attending to equipment maintenance when necessary.

Data Downloading Operation

Data were downloaded from the vehicle through a cellular phone/modem connection in each participant's vehicle. A call is placed from the control center to the vehicle cell phone during the time period when the unit is "awake." Each unit turns on for two minutes every 20 minutes to ensure system operability. This time period is different for each vehicle, and only the control center operators will know when to place the call. Once a connection is made, the operator issues a keyboard command that commences automatic download. After download, the registers are reset, and all but the last trip is deleted from memory. Saving the last trip helps the researchers to ensure data continuity. If the download is interrupted, the operator will call back,

and the download will continue where it left off.

All 50 units were downloaded each week on Sunday evening. Download times varied from approximately 10 minutes to more than one hour, depending on logging frequency, amount of driving done by the participant, and data transfer rate. The download time was, by far, the most complicated of the unknown variables in estimating the cost of the project. The complication was primarily due to the unknown travel behavior of the participants.

The following formula was developed to estimate the amount of time needed to transfer the data:

$$D = [(m)(s)(0.1777)]/h \tag{1}$$

$$M = [(h)(5760)(D)]/s \tag{2}$$

$$t = M/r \tag{3}$$

where, D = number of days between downloads;
        m = amount of memory in tracking unit (512K);
        s = number of seconds between samples;
        h = ignition hours per day;
        M = total memory used over 'D' number of days;
        t = download time (secs); and
        r = current transfer rate (approx. 100 bps - actual number may vary)

For example, using Equation (1), a feasible number of days of ignition time can be determined based on the capacity of the tracking unit memory. The number of days of ignition time refers to the amount of time that the vehicle can be in operation (logging data) before the logging device runs out of memory. This time limit is based on the amount of memory in the device, sampling rate, and the amount of time per day that the vehicle is in operation. During the initial term of the field test vehicle speed and location were recorded every five seconds ($s = 5$ sec). The value of ignition time, h, was estimated and set at 3.5 hours/day/driver based on observations from similar driving studies undertaken by the subcontractor.

$$D = [(512 \ kb)(5 \ sec)(0.1777)]/3.5 \ hrs = 12.997 \ days$$

Completing equation 1 with the given logging rate and ignition time, we find that the unit will need to be downloaded within 13 days to avoid losing data due to a full memory. Once the limit is determined, D can be set to any number smaller than that calculated and entered into Equation (2) to estimate the total memory used over that time period (based on average daily ignition time). To be conservative and allow for uniform reporting periods, the researchers set the number of days between downloads at seven days. Knowing memory usage, we will then be able to estimate the budget for cellular charges necessary for downloading the data.

$$M = [(3.5 \ hrs)(5760)(7 \ days)]/5 \ sec = 28,224 \ bytes$$

8

Equation 2 shows that approximately 28,224 bytes of memory will be used in the seven day logging period. Finally, a total download time can be calculated using Equation (3).

$$t = 28,224 \; bytes/100 \; bps = 282 \; sec$$

The time needed to download this data, based on an estimated transfer rate of 100 bps yields a cellular call 282 seconds (4.7 min) in duration. Additional cellular time would also be needed for connection with the modem and to clear and reset the memory.

Sampling Rates

In order to determine the optimal logging frequency for future applications, sampling frequencies were varied over the course of the field test. The first 10 weeks of data were collected at 5-second intervals. The next 3-week time period was logged at 1-second intervals, followed by 3-second, and back to 1-second intervals to finish out the four-month period.

Crash Site Data Collection Field Test

To compare operating speed data of individual drivers with that of the driving population, it was necessary to establish data collection sites on roadways within the study area. The original idea was to collect data from sites where participants were involved in crashes. The likelihood that a crash would occur during the short four-month pilot study time frame was minimal. However, one minor crash was observed. Other "fictional" crash sites had to be chosen by the researchers to allow for sufficient data for analysis. These sites were chosen based on known locations of participant vehicular activity so that the data would have a meaningful application. The roadway data collection sites were divided into three categories: freeway sites, non-freeway sites, and one crash site. Data were collected at these sites simultaneous to the vehicle speed data collection activities.

Freeway Sites

Three of the study sites incorporated information from the TranGuide ATMS data archive system. The TransGuide system writes the volume and speed data to a file accessible to the research agencies. Researchers at the Texas Transportation Institute access this volume and speed data via the Internet. An example of the data output collected at these freeway sites is shown in Table 1. The loop data are aggregated on a 5-minute interval that results in a weighted average speed.

A database has been established at TTI that aggregates the volume and speed data by hour of the day. Since the data are collected continuously and archived, data can be retrieved from the archives to show relative speeds at the exact date and time of a crash occurrence. This application of the data may be more or less accurate based on the location of the crash in relation to the location of the sensor stations.

Table 1. Sample of archive data from TransGuide ATMS loop detectors.

Date	Time	Location	Speed[a]	Volume	Occupancy[b]
7/15/97	7:00:03	L1-0L10E-568.241	75	9	7
7/15/97	7:00:03	L1-0U10E-568.248	64	7	5
7/15/97	7:00:03	L2-0L10E-568.241	63	6	6
7/15/97	7:00:03	L2-0U10E-568.248	72	6	4
7/15/97	7:00:03	L3-0U10E-568.248	57	6	6
7/15/97	7:00:04	EN1-0U10E-568.845	-1	6	18
7/15/97	7:00:04	EX1-0U10E-568.764	-1	2	3
7/15/97	7:00:04	L1-0L10E-568.802	67	5	4
7/15/97	7:00:04	L1-0U10E-568.807	62	6	5
7/15/97	7:00:04	L2-0L10E-568.802	67	1	1
7/15/97	7:00:04	L2-0U10E-568.807	60	8	7
7/15/97	7:00:04	L3-0U10E-568.807	46	8	8
7/15/97	7:00:04	EN1-0010E-569.671	-1	3	6

[a] Speed = -1 means that no speed has been measured (single loop detector).

[b] Occupancy is the percentage of time the loop detector is occupied.

Non-freeway Sites

During the four-month test period, speed distributions, headway, and roadway geometrics were collected at four non-freeway sites: one major arterial, one minor arterial, one collector street, and one frontage road. The speed distributions for non-freeway sites were collected using the piezo sensors. The sensors were each six feet in length and were placed in the roadway to cover half of one lane. Each study site was 200 meters (660 feet) in length. Sensors were installed at the beginning, middle, and end of each study site. This procedure produced 100-meter (330 foot) intervals between the speed measurement devices. Therefore, if the average prevailing speed at the study site was 72 km/h (45 mph), then the time space intervals between speed measurements was approximately 5 seconds.

The class II piezo sensors that were used are of a temporary type. They are secured to the pavement with glue and tape. The use of temporary sensors allowed the research team to use the same set of data collection equipment at different sites, thereby reducing the amount of equipment to be purchased. Enough equipment was purchased so that data could be collected at two sites simultaneously.

Crash Site

One minor crash was recorded during the period of one-second sampling frequency. Crash specific data and speed data were collected in the roadway and the specific lane of the crash-involved vehicle . The lane chosen was based on the lane the vehicle was traveling in

10

immediately before lateral lane repositioning occurred. The speed data collection was initiated within one week of the crash. The researchers collected data for the same day of the week, time, and location as the crash using a case-control approach. The crash occurred on a section of freeway outside of the ATMS coverage area -- this was the only instance of speed data collection using the laser speed detection gun. The data collection continued for three weeks to ensure minimization of the variability associated with comparisons of data from one time period to another.

Data Analysis

In the pilot study, several types of analyses were carried out. For this summary report, we will only look at three of the more definitive analyses. Each analysis will be explained and the experimental and statistical assumptions underlying those analyses documented. The first analysis describes (in general) the performance of the vehicle speed data collection equipment, while the other two analyses are directed at two of the four research questions listed in the problem statement.

Analysis 1: Speed Data Collection Equipment Performance, Storage Requirements, and Sources of Error

During the course of the 4-month field test, voluminous amounts of speed data were collected from 50 participants. The data were collected at different frequencies to include 1-second, 3-second, and 5-second sampling rates. General statistics will be tabulated to show various aspects of the data size and storage requirements for each frequency. A review of the overall equipment performance will indicate the potential sources of error as well as data formatting problems. Additional information will be provided in the use and effect of filters applied to the data sets.

Analysis 2: Speeders vs. Non-Speeders and Relationship to Crash Risk

Research Question 1: How does the risk of crashing for drivers who exceed the speed limit differ from drivers who observe speed limits?

Suppose that through the use of in-vehicle recording devices we are able to define two groups of drivers: one group ("speeders") typically exceeds the posted speed limit; the other group ("non-speeders")[1] typically stays within the posted limits. The drivers referred to as speeders, and the drivers referred to as non-speeders, are defined by some decision rule (e.g., speeders drive in excess of the posted limits 30+ percent of the time; non-speeders exceed the

---

[1]Due to the properties of the constrained speed definitions, records from freeway sections will be filtered if the associated speeds are less than 46 mph. This precludes the study of drivers whose speeds vary on the lower side of the average driver. This lower variance has been cited as a factor to incidents in previous research.

posted limit less than 30 percent of the time). Note that drivers are not randomly assigned to their respective groups (as in a true experimental study), rather these drivers are simply separated into two mutually exclusive groups based upon some decision rule or operational definition.

The two groups of drivers are followed for a period of four months and involvement in traffic crashes is recorded for both groups. The ratio of the estimated probabilities of crashes for speeders (p1) and non-speeders (p2) is a measure of their relative risks (RR) of crash involvement. For 20 speeders who are involved in 3 crashes in four months and 80 non-speeders who are involved in 4 crashes in the same time frame, the relative risk of crash involvement is 3 times as great for speeders as non-speeders.

$$RR = \left(\frac{p_1}{p_2}\right) = \left(\frac{\dfrac{3 \text{ crashes}}{20 \text{ speeders}}}{\dfrac{4 \text{ crashes}}{80 \text{ non-speeders}}}\right) = 3.0 \quad (4)$$

This definition of relative risk assumes, of course, that speeders and non-speeders are comparable on all factors that might contribute to the likelihood crash involvement (e.g., age, sex, health, alcohol consumption, etc.). It also assumes that the conditions under which speeders and non-speeders drive are comparable (i.e., the same kinds and types of roads, the same weather conditions, the same kinds and types of traffic, etc.).

Granting the assumptions of the previous paragraph, it should also be pointed out that this RR of 3 is not significantly different from 1.0 (z = 1.52).Obviously, the numbers of drivers at risk (100) and the numbers of crashes (7) in this example are relatively few—too few, in fact, to achieve statistical significance even though speeders appear to be three times more likely than non-speeders to be involved in crashes.

The relative risk analysis discussed in question 1 is, for example, a statistical estimate of probabilities and is considered a categorical method. The two variables of interest are crash involvement (two categories, yes and no) and driver speed (two categories -- speeders and non-speeders). This measure of relative risk is highly sensitive to the definition of the categories. In the case of driver speed, we have taken a continuous variable, speed, and categorized it according to some decision rule which defines "speeders" and "non-speeders". Due to the sensitivity of the relative risk measure, analyses will be performed using different decision rules in the definition of "speeders."

Analysis 3: Driver Speed vs. Prevailing Speed Measurements

Research Question 2: What is the nature of the association between travel speed at the time of the crash and prevailing travel speed at the crash site?

This discussion will be prefaced by defining case-control studies. The *case-control* study, also commonly called a *retrospective study*, follows a paradigm that proceeds from effect

12

to cause. In a case-control study, individuals with a particular condition(the *cases*) are selected for comparison with a series of individuals in whom the condition is absent (the *controls*). Cases and controls are compared with respect to existing or past attributes or exposures thought to be relevant to the development of the condition under study.

In response to the posed question, we start with an outcome (a crash) and work backwards. One of the 50 vehicles involved in the study is in a crash. The crash occurred at 7:00 p.m. on a Thursday. For this particular crash-involved vehicle, recorded travel speeds during the time leading up to the crash are available.

Returning to the crash site at 7:00 p.m. on the Thursday following the crash, speeds are recorded upstream from the crash site for N vehicles. These recorded speeds constitute the matched control against which the travel speed for the crash-involved vehicle (case) is assessed. If the traveling speed of the crash-involved vehicle was not of consequence in producing the crash, then we would expect that the speed for the crash-involved vehicle to equal the mean speed for the N matched control vehicles. To the extent that the observed speed of the crash-involved vehicle exceeds the mean speed of the control vehicles, speed may be of consequence in the production of crashes.

The logic of this methodology is presented graphically in Figure 2. Three crashes (data points) are shown for purposes of illustration. The horizontal axis is the mean speed for the matched controls. The vertical axis depicts the speeds of the crash-involved vehicles (cases). To the extent that the data points lie systematically above the diagonal, speed is seen to be a contributing factor to the recorded crashes. To the extent that the data points scatter around (and on both sides of the diagonal), speed is not a contributing factor to the production of these crashes involving the case vehicles.
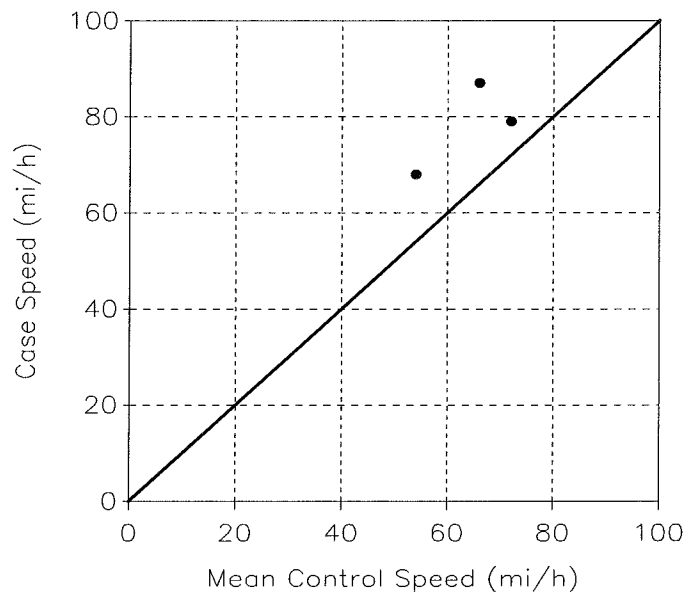
Figure 2. Pre-crash speed vs. mean control speed

FINDINGS

During the course of the study, voluminous amounts of speed data were collected along with a single crash observation data point. These data were reduced and are being used to answer a variety of questions, including two of the four suggested in the Problem Statement. Reference is made in the data analysis to three speed measurement data sets. The differences between the data sets is the frequency at which the data were collected. These are typically referred to in the increasing order of 1-second, 3-second, and 5-second speed data collection.

Analysis 1: Speed Data Collection Equipment Performance, Storage Requirements, and Sources of Error

The vehicle data collection hardware performed quite well, with few problems. Only one equipment failure was experienced, requiring the replacement of a cellular transmitter in one of the units. There was also one instance of tampering, which resulted in the disconnection of an antenna by one of the participant's children. There was no damage to the unit in this instance. More important to note was the loss of participant vehicles due to vehicle trade-ins. In this era of leased vehicles, car trading is a major issue with vehicle instrumentation studies. Three participants were ultimately lost because they traded in their vehicles and did not wish to have equipment installed in their new car for the short time left in the study. Allowing a change in vehicles may affect driving characteristics and driver consistency. These are all issues that need to be considered in a full-scale study.

Another issue regarding vehicle speed and data collection, via GPS, is the validity of the GPS data. Several types of potential sources of error should be noted:

- The first is the lack of the GPS receiver to achieve a positional fix during some short trips. This condition would return no valid GPS records, thus no data could be matched to the map and finally stored in the database.
- Secondly, some "good" GPS data points may not be processed with differential corrections. Due to the time constraints of the pilot study, the data recording equipment could not be reconfigured to log a parameter check for differential corrections reception. If the differential corrections sub-carrier signal could not be received due to bridge blockage, signal strength or urban canyons, GPS data was still recorded, although the quality could be greatly reduced.
- The third and final noted error in the GPS data was minimal, but worth mentioning. A small amount of records were logged that did not follow the logging format, but contained strings of unrecognizable characters (most likely attributable to a cough in data transfer). This, however, caused errors in the raw data conversion programs. All raw data had to be pre-filtered for unrecognized variables before being converted.

Combining all of the data collected over the three sampling frequencies, the participants logged over 300 megabytes of raw data. A quick calculation using an approximate value of 25 records per kilobyte yields an estimate of approximately 8 million raw data records that were collected in the 17-week data collection period by the 50 participants. Table 2 shows some basic statistics related to the raw data file size.

14

Table 2. Basic statistics on raw data storage requirements.

	Sampling Frequency		
	1-second	3-second	5-second
Total Combined File Size (kb)	120,264	46,469	139,298
Collection Time Period (wks)	4	3	10
Total # of Drivers in Study	50 (1-wk) 49 (1-wk) 46 (2-wks)	46	50
Number of Driver-weeks[a] (wks)	191	138	500
Average File Size/driver/week (kb)	630	337	279
Adjusted Number of Driver-weeks[b] (wks)	150	101	442
Adjusted Average File Size/driver/week (kb)	802	460	315
Estimated data file size for full-scale study[c] (gb)	167	96	66

[a] Number of Driver-weeks = [number of drivers logging data (based on # of participants)]*[collection period (wks)]

[b] Adjusted Driver-weeks = [number of driver-weeks] - [number of driver-weeks with filesizes < 20 kb]

This corrects the file size for lack of logging due to vacations, business trips, and vehicle maintenance.

The majority of 1-second and 3-second data were logged in a non-school period, which could explain why there is more variability in the amount of data collected at these frequencies.

[c] Total amount of raw data expected during a full-scale study with 2000 drivers over a period of 24 months [(adjusted average file size/driver/week)(2000 drivers)(104 wks)]

The amount of data collected during the 1-second and 3-second logging periods was much smaller than the researchers had expected based on estimations using the volume of 5-second data. The 3-second data volume was lower than estimated by 13%, whereas the 1-second data was lower by 49%. During conversion of the data files, a couple of problems were noted in the format of the data taken from higher frequency logging rates. The researchers attribute part of the problem with the 1-second data to difficulties in data transmission. The voluminous amounts of data took much longer to transfer than did the data at lower frequencies. Instability in the cellular download connection with longer calls led to instances of premature terminations of downloads, as well as erroneous data entering the data stream. The pilot study was intended to be a test of the limits of cellular transmission -- these numbers clearly show that more frequent downloads with higher sampling rates would be beneficial to help reduce the errors in the data transfer process. Ultimately, 22 of the 693 total files were found to be unusable (approximately 3.2% of the total files).

Other differences noted between the sampling rates, were the temporal differences in number of drivers with minimal data logs (<20 kilobytes in size). During the 5-second data collection period (April 1, 1998 through June 10, 1998), the number of drivers with file sizes less

than 20 kilobytes ranged from 3 to 7. After June 10, 1998, however, the range increased to between 8 and 21 drivers with file sizes less than 20 kilobytes -- more than double that of the 5-second data collection period. This change correlates well with the release of school, and the family vacation period (mid June-July).

Finally, not all recorded data were used in the final analysis, even if they were valid data points. Upper and lower thresholds were set for the vehicle speed measurements. The lower speeds were censored for speeds that were considered to be constrained by other factors (i.e., traffic and signals). The lower speed filters were set at < 74 kph (46 mph) for urban interstates, < 26 kph (16 mph) for urban principal arterials, and < 21 kph (13 mph) for urban minor arterials. Data for collector and local streets were discarded because the researchers could not determine an appropriate way to estimate constrained speeds. Speeds on these types of facilities are a function of many different constraints such as school zones, speed humps, extreme roadway geometry, and an endless array of other constraints. Upper speeds were also filtered due to suspicion of data quality. Across the board, all speeds over 161 kph (100 mph) were censored. Occurrence of these high speeds were random, and were considered instances of GPS signal loss. Recording signal status would eliminate this unknown in a full scale study. The effects of these filters on the database size are shown in Table 3.

Table 3. Effects of filters on data base sizes.

	1-second	3-second	5-second
Total records in data set	1,493,543	636,721	2,255,245
Number of records censored for lower speeds	684,437	281,890	918,540
Number of records censored for upper speeds	50	7	1,027
Number of records left in evaluation data set	809,056	345,734	1,335,678
Number of records remaining after aggregating to one minute intervals	52,087	45,335	214,741

Analysis 2. Speeders vs. Non-Speeders and Relationship to Crash Risk

Using the 1-second filtered data set, two groups of drivers were defined. "Speeders" were defined as those drivers who drive greater than 5 mph over the speed limit 50 percent of the time or more. "Non-speeders" were defined as those exceeding the posted speed limit by 5 mph less than 50 percent of the time. Figure 3 shows the percent of speed observations greater than 5 mph over the speed limit for each participating driver. A bold horizontal line denotes the dividing line between speeders and non-speeders. Based on these decision rules, 20 out of 50 drivers (40%) are classified as speeders.

Having observed only one crash during the field test period, the proposed calculation of the relative risk of crash involvement using equation 4 was nulled. However, a similar calculation was made based on the 3-year accident experience reported in the participant surveys.

16
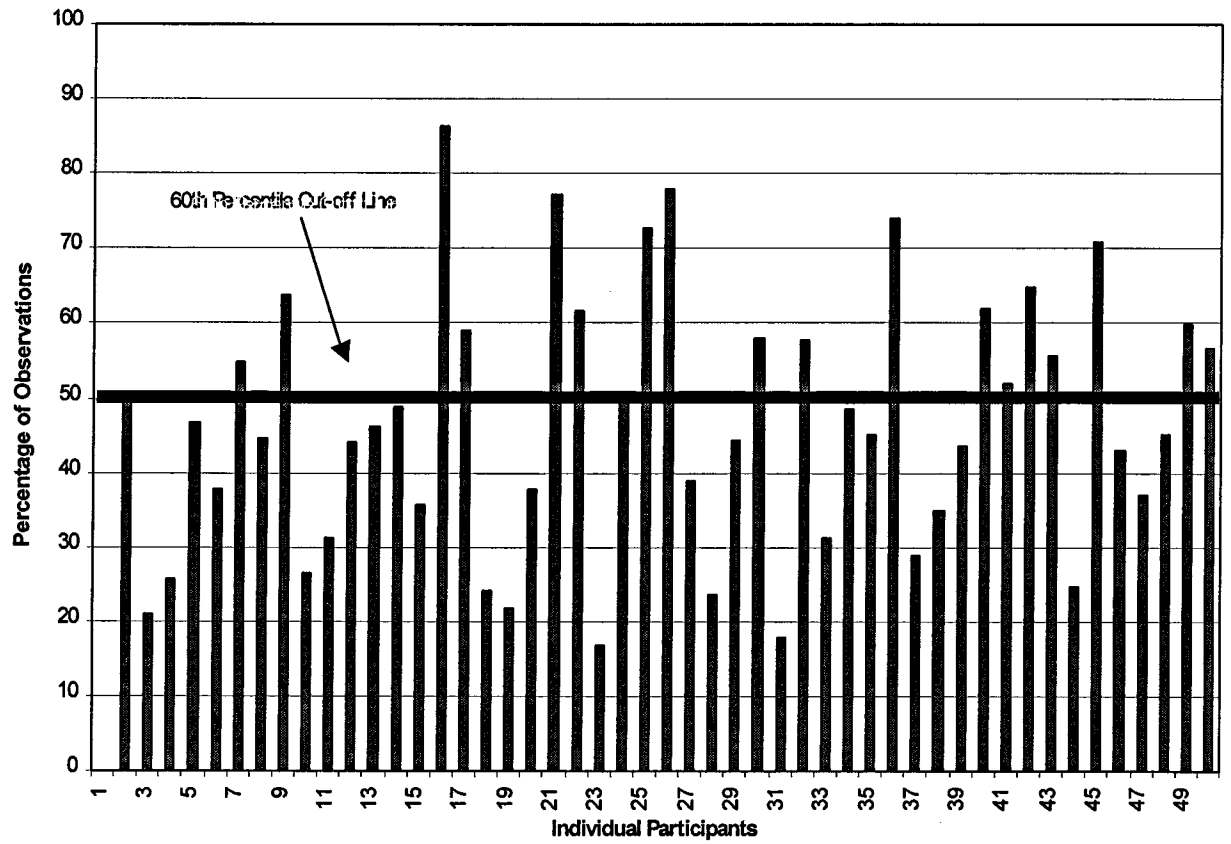
Figure 3. **Speed observations greater than 5 mph above speed limit (all roadway classes combined)**

18

$$RR = \cfrac{\left(\cfrac{6\text{ crashes}}{20\text{ speeders}}\right)}{\left(\cfrac{19\text{ crashes}}{30\text{ non}-\text{speeders}}\right)} = 0.5$$

For 20 speeders who have been involved in 6 crashes over a 3-year period and 30 non-speeders who were were involved in 19 crashes in the same time frame, the relative riskof crash involvement appears to be half as great for speeders as for non-speeders. As before, the numbers of drivers used in this example are too few to achieve statistical significance.

At this point, some observations on sample size, statistical power, and the definition of "speeders" seem in order. In 1994 some 922,896 vehicles were registered in Bexar County, Texas. During that same year (1994), 77,978 vehicles were involved in crashes in Bexar County. Or, in very gross terms, we might estimate that the probability that a given vehicle was involved in a crash in Bexar County in 1994 was 0.085. Applying that crash probability to the 50 vehicles participating in the proposed study (over a four-month period), we expected to see 1.4 crashes recorded for these vehicles during the course of the study.

$$1.4 = 50\left(\frac{4\text{ months}}{12\text{ months}}\right)0.085 \quad \textbf{(5)}$$

For the proposed follow-on study of 2,000 vehicles monitored over a two-year time period, expected crashes would equal 340.

$$340 = 2{,}000\left(\frac{24\text{ months}}{12\text{ months}}\right)0.085\,\textbf{(6)}$$

Now, these estimated crash frequencies (for the full study—2,000 vehicles involved in 340 crashes over a two year period) will be used to assess the significance of the relative risk of speeders (versus non-speeders) being involved in a crash under 25 different scenarios.

Scenario 1: Speeders are 1.5 times as likely as non-speeders to be involved in crashes and 50 percent of all drivers are speeders.			
Involved in a Crash	Speeders	Non-Speeders	Total
Yes	204	136	340
No	796	864	1,660
	1,000	1,000	2,000

The estimated relative risk (RR) of a crash for speeders (p_1) versus non-speeders (p_2) is:

$$RR = \left(\frac{p_1}{p_2}\right) = \left(\frac{\frac{204}{1,000}}{\frac{136}{1,000}}\right) = 1.50 \quad (7)$$

The natural logarithm (ln) of relative risk (RR) is asymptotically normal with a standard error that can be approximated as:

$$SE_{\ln(RR)} = \sqrt{\frac{1-p_1}{n_1(p_1)} + \frac{1-p_2}{n_2(p_2)}} = 0.1013 \quad (8)$$

Where p_1 and p_2 are as previously defined, and n_1 and n_2 represent the numbers of speeders and non-speeders at risk, respectively.

Transforming the calculated ln (RR) into a z score:

20

$$z = \frac{\ln(RR)}{SE_{\ln(RR)}} = \frac{\ln(1.50)}{0.1013} = 4.00 \quad (9)$$

Scenario 2: Speeders are 1.1 times as likely as non-speeders to be involved in crashes and 10 percent of all drivers are speeders.			
Involved in a Crash	Speeders	Non-Speeders	Total
Yes	37	303	340
No	163	1,497	1,660
	200	1,800	2,000

The relative risk (RR) of a crash for speeders vs non-speeders in this second scenario is 1.099 ($\approx$ 1.1). The z calculated from the data shown in this second scenario is 0.60. That is to say, under this scenario the risks of crash involvement for speeders and non-speeders are not statistically different (at traditional levels of α).

In similar fashion, standard normal (z) scores were calculated for all 25 cells (scenarios) in Table 4.

Table 4. Calculated Standard Normal (z) Scores for 25 Different Scenarios

	Percent of Drivers Classified as "Speeders"				
Over Involvement of "Speeders" in Crashes	10	20	30	40	50
1.1	0.60	0.75	0.91	0.97	0.95
1.2	1.21	1.50	1.69	1.82	1.78
1.3	1.82	2.26	2.61	2.67	2.60
1.4	2.44	3.02	3.40	3.40	3.30
1.5	3.07	3.78	4.06	4.12	4.00

If speeders are narrowly defined to include only a small (but presumably aggressive) percentage of drivers at risk (say 10 percent), we will sensitize our dependent variable by categorizing only the most egregious of drivers as speeders. But, by so doing, we reduce the statistical power of our analysis, as demonstrated by the relatively small z's in the first column shown above. If speeders are more broadly defined to include, say, 40 percent of the drivers at risk, the sensitivity of our dependent variable will be reduced, but statistical power will be enhanced, as demonstrated by the relatively larger z's in the fourth column. Clearly, within a certain range there is a "sensitivity-power" tradeoff in defining speeders.

Analysis 3. Driver Speed vs. Prevailing Speed Measurements

For this analysis, we will refer to data collected for the singular observed crash. The crash occurred at 5:24 p.m. on Monday, June 8, 1998. The facility was a three-lane eastbound section of IH 410 on the north side of San Antonio between West Avenue and Vance Jackson. The geometry of the roadway contained slight horizontal and vertical curvature in this area, but fully supported a posted speed limit of 97 kph (60 mph). The crash was of a rear-end type, and caused minor/moderate damage to the vehicles. The conditions at the time of the accident were congested with a stop-and-go flow. Figure 4 shows the progression of the participant into the accident area. The accident area is highlighted with a circle. A noticeable slow down is occurring in the participant's speed as they enter into the circled area. The slow-down is denoted by points that begin to get more closely spaced (i.e., with a constant logging interval, a vehicle will travel a shorter distance at a slower speed than it will at a higher speed).
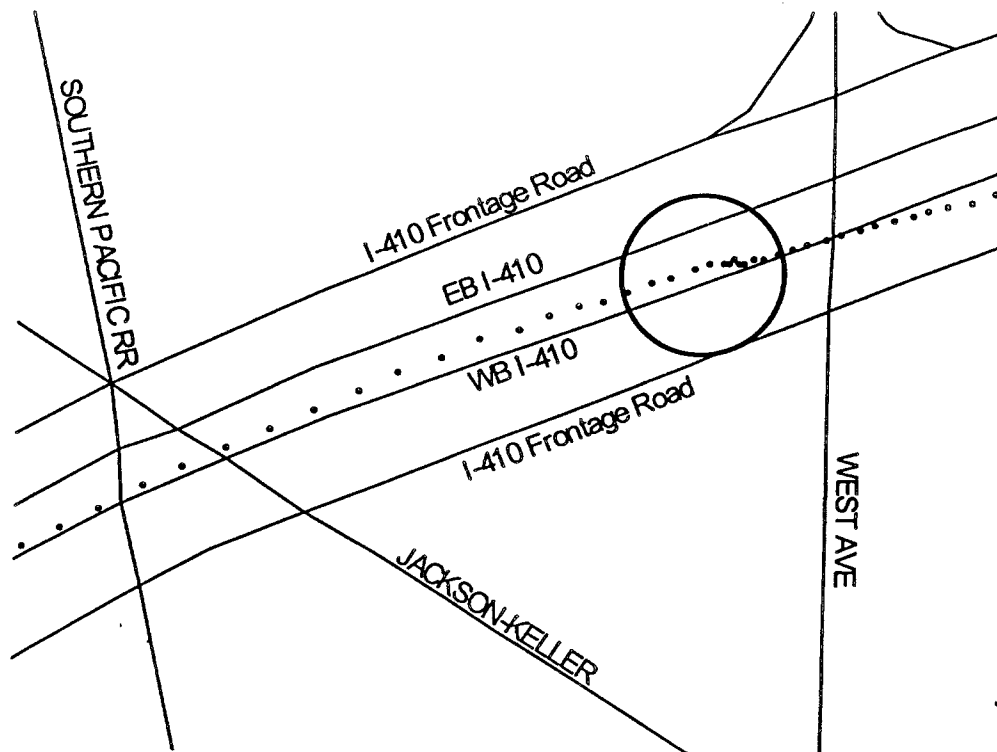
Figure 4. View of participant trip data involving a crash.

The driver of the striking vehicle was a male, 64 years of age, driving a 1988 Nissan pick-up truck. The study participant, ID D6040633, was driving a 1992 Mercedes with anti-lock braking system. No discernable skid marks could be found. Environmental conditions were daylight and clear with a dry black-top surface.

Figure 5 shows a more detailed account of the seconds just prior to the accident. The points are numbered in the order in which they occurred. Table 5 contains the data records for the numbered points. Looking across the columns, there is a speed column which contains spot speed information collected by the GPS receiver. Note that the participant's speed gets progressively slower from 34.6 kph (21.5 mph) down to 11.7 kph (7.3 mph) until the accident occurs, and the car remains in one location for 4 seconds (points 6-9). Point 6 denotes the final resting place of the impacted vehicle.
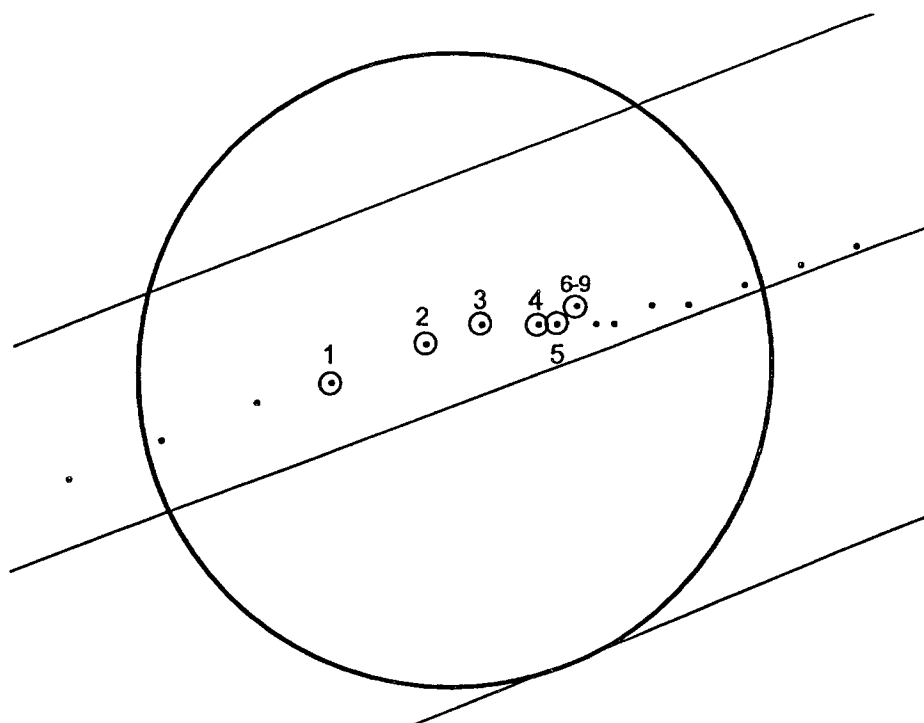
**Figure 5. Detail of actual crash data points.**

Table 5. Numerical detail of crash data points.

Point #	Unit ID	Trip #	Date	Time (GMT)	Latitude	Longitude	Speed (mph)	Heading

1	D6040633	3	8/6/98	22:24:46	29.516967	-98.526467	21.5	69.5
2	D6040633	3	8/6/98	22:24:47	29.517000	-98.526383	19.7	69.8
3	D6040633	3	8/6/98	22:24:48	29.517017	-98.526333	14.2	72.2
4	D6040633	3	8/6/98	22:24:49	29.517017	-98.526283	7.9	83.6
5	D6040633	3	8/6/98	22:24:50	29.517033	-98.526267	7.3	76.3
6	D6040633	3	8/6/98	22:24:51	29.517033	-98.526250	1.5	94.9
7	D6040633	3	8/6/98	22:24:52	29.517033	-98.526250	0	0.0
8	D6040633	3	8/6/98	22:24:53	29.517033	-98.526250	0	0.0
9	D6040633	3	8/6/98	22:24:54	29.517033	-98.526250	0	0.0

Further analysis was performed by looking at the participants speed versus speed measurements from the population at the crash site. Speed distributions were collected in a case-control methodology for three consecutive Mondays following the date of the accident. These data were collected using a laser gun to avoid the cost of shutting down a lane to install temporary piezo sensors. Hundreds of samples were taken between 5:00 P.M. and 6:00 P.M. on each of the weeks. The data were developed into cumulative frequency plots shown in Figure 6. The dates of June 15th and 22nd show very similar data trends with the majority of the data in the upper 50s to 60 mph, with few or no data falling below 45 mph. However, on June 29, there is a very different pattern with many samples below 45 mph. By only examining this one type of data, it is not obvious why the trend has shifted. It is not until the speed distribution data were merged with the weather data that the relationship in the shift of speeds due to rainy weather became obvious. Note that on these three particular weeks, there were no observations below 32 mph.
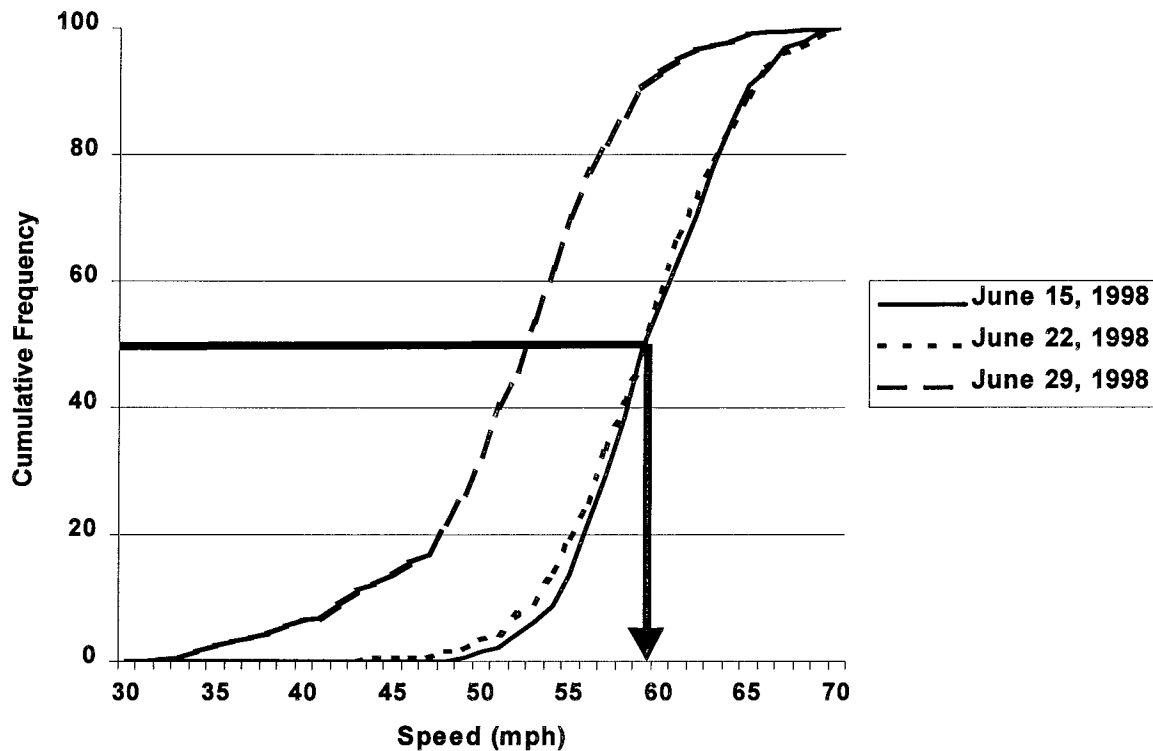
Figure 6. Case-control speed distributions gathered at the crash site

An arrow has been drawn on Figure 6 to denote the control mean speed under dry weather conditions. If the relationship between pre-crash speed and mean control speed were plotted as in the design of Analysis 3, this data point would lie well below the diagonal. Thus, speed would be considered a contributing factor. All though the contribution of speed in this crash was due to slower speeds and not excessive speeds – a topic that we are not focusing on.

DISCUSSION AND RECOMMENDATIONS

The main objective of this pilot study was to determine the feasibility of conducting a full-scale study to show quantitative properties of the relationship between driving speeds and crash probabilities. Three focus areas were defined as: 1) proof-of-concept of the technology; 2) development of driver profiles; and 3) provision of recommendations for future research.

The pilot study has proven to be effective in all three focus areas. The vehicle tracking technology provided speed and location measurements with negligible equipment problems. Data obtained from the equipment were sufficient for a pilot study. However, minor modifications are recommended. Data relationships were defined and tested successfully,

although no practical statistically significant conclusions could be drawn from the small sample. Complications such as participants trading in vehicles and liability issues were experienced and provide great insight for the smooth operation of a full-scale study. The following sections provide a brief summary of major findings and recommendations for a full-scale study.

Vehicle Data Collection

The use of off-the-shelf GPS receivers to collect speed and location data is a viable option. The use of differential corrections in connection with the GPS receivers provides accurate speed and location data. The unit used in the pilot had an add-on differential corrections receiver which provided 2-to 5-meter accuracy. However, due to the short time-frame of the pilot study, the logging software was not modified to log parameters regarding the status of the differential corrections. This unnecessary unknown did affect the reliability of the data used in the pilot study, but would be an easy modification to make for a full-scale study. Another useful modification would be the addition of a separate speed detection device that is capable of logging speeds at smaller intervals. One-second data is sufficient for location data, but a more detailed portrayal of speeds prior to an accident could be beneficial – especially when observing accidents at high speeds. This type of device would be able to sense erratic acceleration/deceleration behavior, which in combination with speed data may be a better predictor of crash risk than speed alone.

There did not seem to be a large difference in the various sampling rates used (1-second, 3-second, or 5-second) to determine driver profiles. The 1-second sampling rate did, however, give a fairly well-defined portrayal of vehicle movement during the one observed crash. Software could be configured for different temporary and permanent logging rates to provide maximum data potential. The development of driver profiles would perhaps be easier if only a random sample of trips were recorded. This would not only provide for randomization of the data set, but also minimize the need for both massive storage and more robust statistical analysis tools.

Crash Site Data Collection

The San Antonio ATMS provided an excellent source of speed data for the analysis of driver speed versus average speeds. Accessibility to this type of high tech traffic data collection is limited in the national sense, but is becoming more common by the day. While a very good source of participant speed versus average driver speed at a given time, the probability that one of the participants would be involved in an accident in close proximity to one of the data collection sites is still minimal.

The use of piezoelectric sensors to collect speed distributions is very accurate. The sensors are semi-permanent, and can be laid out for several weeks without experiencing any problems. However, most data collection activity of this type is personnel and equipment

intensive. Tape and other supplies used to install equipment, as well as travel expenses, are not negligible. On heavily-traveled roadways and freeways, lane closures may also be necessary for installation of the equipment. Lane closures can range in price from a couple of hundred dollars to well over a thousand dollars for a one-lane closure. The equipment must also be maintained on a weekly basis by changing batteries and downloading data. Personnel requirements averaged 3.5 person-days per site. In full scale study with 200 crash sites, this expense may prove prohibitive to this type of collection.

Database Development

Groundwork for a couple of speed data relationships have been provided. Types and sources of error in the data have been outlined and will provide an insight to the development of more reliable data logging systems. Censoring and filtering mechanisms have also been explored and will provide a good starting point for the evaluation of data consistency in future studies.

Relationships between the data elements appear to exist, but to provide conclusive statistics on these relationships will require tightly defined research questions and targeted sampling plans. To choose a random sample based on the total accident rate may not be an efficient or definitive plan for studying accidents involving excessive speed. The overall accident rate of a location or nation includes all types of crashes. Possibly a more appropriate sampling strategy would be to consider what types of accidents are likely to be affected by excessive speed and pose specific questions within sub-populations based on demographics surrounding each crash type. These crash types may include rear-end, run-off-road, and roll-over crashes as well as others. Right-angle crashes coded in urban settings may be less likely to involve excessive speed as a causal factor. Using accident type as a pre-cursor filter along with driver accident histories could help to define smaller homogeneous samples, thus cutting down on the variability that was seen in the pilot study data.

Data Analysis

The analysis of the data is fairly straight forward, and detailed methods are included in the data analysis plan and the data analysis chapter. Obstacles associated with the data analysis include: time involved in sorting through the millions of records that this type of data collection produces; initial lag time involved with the amount of time needed for data conversion and development of full data sets; and, determination of the source of the anomalies that inherent in almost all data sets of this size as well as the development of filtering techniques to correct them.

While the volume of data is not a real storage issue, it does become an issue during the processing of the data. File type and format should be chosen in any full-scale study to minimize total size of the database. Stratified sampling plans would also allow the data to be analyzed in groups which would speed the analysis process.

The issue of statistical significance versus practical significance also warrants concern because with *so* many observations, significance is almost a certainty and decisions will have to be made as to what is *really* significant or just an artifact of the theory.

Issues Involving Human Subjects

The liability associated with collecting detailed speed and location data with instrumented vehicles will always be a major concern. All means of protecting the researchers, agencies, participants and data in connection with a full-scale study should be fully utilized. While consent forms are sometimes deterrents to participation, the forms must clearly portray the liabilities to the participant. Monetary or other incentives could be used to overcome this issue. Also, while Certificates of Confidentiality do not completely remove the risk of liability, they do provide the first level of security for the researchers as well as the participants. Techniques for scrambling and encrypting the data may also work to keep unauthorized users out of the data, and are recommended for use in the pending full-scale study.

Feasibility of Conducting a Full-Scale Study

A full-scale study is feasible from both a technical and monetary point of view. The probability of answering posed research questions will be only as good as the questions themselves and the associated sampling plans. Quantity does not necessarily constitute quality, and while the analysis shows most differences in the data to be significant, the issue of practical statistical significance must still be defined. The researchers feel that it would be a worthwhile endeavor to take this pilot study to the next level.