CALIFORNIA PATH PROGRAM
INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Methods for Identifying High Collision Concentration Locations for Potential Safety Improvements

Judy Geyer, Elena Lankina, Ching-Yao Chan, David Ragland, Trinh Pham, Ashkan Sharafsaleh

California PATH Research Report
UCB-ITS-PRR-2008-35

This work was performed as part of the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation, and the United States Department of Transportation, Federal Highway Administration.

Final Report for Task Order 5215

December 2008

ISSN 1055-1425

CALIFORNIA PARTNERS FOR ADVANCED TRANSIT AND HIGHWAYS

Final Report for Task Order 5215

Methods for Identifying High Collision Concentration Locations For Potential Safety Improvements

Judy Geyer, Elena Lankina, Ching-Yao Chan,
David Ragland, Trinh Pham, Ashkan Sharafsaleh

Traffic Safety Center and California PATH
University of California, Berkeley

November 5, 2008

Abstract

The California Department of Transportation (Caltrans) uses Table C and related documents to identify and to investigate locations within the state highway system where a relatively large number of collisions occur. In earlier years, a task force evaluated the process of generating and using these reports and found that there was much room for improvements. A list of recommendations was made. The efforts undertaken within this project is part of the effort to make the process of safety investigations and improvements more efficient and productive. This report summarizes the work carried out in the first phase of Task Order 5215 and it provides guidelines for the second phase of the project.

Keywords

High Collision Concentration Location, Table C, Highway Safety, Safety Improvement

Executive Summary

This report is the final report for a Caltrans-PATH Project, Task Order 5215, which was continued with a second phase as Task Order 6215. The objectives of this project, "Methods for Identifying High Collision Concentration Locations (HCCL) for Potential Safety Improvements," are to investigate and improve upon the approaches used in searching for highway segments, ramps, and intersections that possess a high concentration of collisions. Through a reliable and accurate screening of these locations, safety investigation and safety improvements can be carried out effectively.

This report contains tow sub-reports: the first one focused on the literature report and the second one on methodologies of HCCL screening and identification.

The former is the review of HCCL identification methods adopted and suggested by other states or federal agencies across the country. We have reviewed research literature addressing this subject and have interviewed other states' transportation departments to learn about various approaches to identifying HCCL. The review includes information about SafetyAnalyst (http://www.safetyanalyst.org/), which is a framework and associated software funded by the FHWA and currently being developed as the most modern and automated method of identifying HCCL with promising safety improvement potential.

The second half of the report contain two major areas: (1) the studies of statistical methods and issues that are potentially significant in HCCL screening methods, based on literature review and ongoing research, and (2) the review of highway, collision, and traffic database for their use in statistical analysis. The latter involves the use of Traffic Accident Surveillance and Analysis System (TASAS) and the Performance Measurement System (PeMS).

This report summarizes the work carried out in the first phase of Task Order 5215 and it provides guidelines for the second phase of the project.

Table of Contents

1. Background

There are approximately 190,000 reported collisions on California State Routes annually. One of the Department's goals is to reduce the number and severity of these collisions. In achieving this goal, every quarter the Department publishes a list, so called "Table C", of high concentration collision locations (HCCL). There are 170 traffic safety investigators in Caltrans who review about 10,000 locations annually. Roughly 700 improvements are initiated annually as a result of the HCCL program (1). Traffic investigators also receive an annual "Wet Table C" that identifies high wet pavement collision concentration locations (1).

Recently Caltrans conducted a review of the HCCL investigation process, making the following short-term and long-term recommendations:

Short-term Table C recommendations:

1. Identify and Eliminate Repeat Locations
Repeat locations are defined as 100% the same postmile limits as any "required" location identified during the previous 3 quarters. Repeat locations will be screened out and will not be included in the list sent to the districts for investigations.

2. Identify and Eliminate Overlap Locations
Overlap locations are defined as an overlapping segment of 51% to 99.99% with any "required" location identified during the previous 3 quarters. Overlap locations will be screened out and not sent to the districts.

3. Combine Adjacent Highway Locations
These locations are defined as highway segments that are adjacent to one another. The adjacent locations will be combined in the report to the districts and will be done in a single investigation. Combined locations will not exceed 1 mile in length.

4. Send out only "Required" Locations
Only those locations marked with a "Req" will be sent to the districts.

5. Update Intersection Traffic Volume
Update intersection traffic volume.

Long-term Table C Recommendations:

1. Modify the selection criteria – Minimum number of collisions and statistical significance threshold could be evaluated.

2. Weigh the severity of collisions: fatal, injury, property damage only – Should there be a prioritization for investigations by placing a weighted factor on collisions by severity?

3. Analyze the segment by collision or revise length – Should the selection of location be made on the location of collisions and/or collision rate and not constrained by the segment length of 0.2 mile?

From this review and in light of the long-term recommendations, Caltrans initiated Task Order 5215 with the California Partners for Advanced Transit and Highways (PATH) and the University of California, Berkeley Traffic Safety Center (TSC). PATH and TSC proposed to evaluate the methodologies used for the identification of high-concentration collision locations.

The first task within Task Order 5215 is the review of HCCL identification methods adopted and suggested by other states or federal agencies across the country. We have reviewed research literature addressing this subject and have interviewed other states' transportation departments to learn about various approaches to identifying HCCL. The review includes information about SafetyAnalyst (http://www.safetyanalyst.org/), which is a framework and associated software funded by the FHWA and currently being developed as the most modern and automated method of identifying HCCL with promising safety improvement potential. The findings of the review are summarized in this report.

2. Critical Issues in Methods for HCCL identification

Even though the basic concept of identifying for HCCL is seemingly straightforward, the process of properly defining HCCL and the execution of identifying HCCL are challenging and complex. Stemming from the review of the Table C Task Force Report, the literature review, and communications with several other states' departments of transportation, several critical issues characterize the HCCL methods:

Frequency versus Rate
Some approaches select locations that are characterized by the highest frequency of collisions in a given roadway in a given time period; other approaches select locations characterized by the highest number of collisions per vehicle mile in a given roadway in a given time period.

Quality Control
There are three primary approaches to determining high collision locations. One is simply rank different locations based on their number or rate of collisions. The second consists of comparing the actual number or rate of collisions with an expected number based on the entire set of locations or a probabilistic model. The third approach, the so-called "Empirical Bayes" approach, compares the actual number or rate of collisions with a "true" number or rate based on a weighted sum of the expected and actual number. The latter two are in the category of "quality control" approaches.

Weighting by Severity
Weighting by collision severity has the advantage of addressing the most serious collisions, but, since severe collisions are substantially rarer than less severe collisions, estimates of high collision locations are less stable.

Segment Type
Intersections, ramps, and highways exhibit very different collision patterns, as do subcategories within these types. The classification of these roadway types, followed by the separate HCCL analysis for each group, may affect how HCCL are selected.

Segment Length
The size of segment analyzed will affect the specificity and the stability of the HCCL identification process.

Analysis Period
The length of collision history being analyzed affects the specificity and stability of the HCCL identification process.

Classification
One of the main challenges of HCCL is the ability to identify locations that have promising potential for engineering improvement. The classification of HCCL into collision-specific groups may assist the engineer in identifying roadway improvements (for example, an HCCL with an unusually high rate of DUI, rear-end, side-swipe, or roll-over).

Traffic Volume Adjustment
Whether studying frequency or rate, most models require data on the vehicle volume at each potential HCCL. The accuracy of the vehicle volume data, and the assumptions of how vehicle volume affects collision rate, greatly affect the HCCL identification process.

Consideration of Non-Highway Factors
Most of the methods, including the SafetyAnalyst method, focus on highway factors. However, non-highway factors that are clustered geographically may produce collision clusters on certain roadways or roadway segments. We will explore this dimension to see whether considering non-highway factors may help address high collision locations.

As can be seen above, there are a number of issues that should be carefully evaluated to have a robust and thorough method to generate well-grounded results in HCCL search. Our objective in the current literature stage and the follow-up task of evaluating HCCL methodologies is to work with the highway safety improvement team at Caltrans to identify the most critical parameters and variables to be incorporated into potential revisions of Table C Methodology.

3. Literature Review of Critical Issues

3.1 Frequency Versus Rate

Caltrans Current Table C: One of the main issues in identifying high collision areas is the choice to use collision frequency or collision rates. Currently the California Table C

methodology uses frequency. The first criterion for screening is that a particular segment must have at least four or more collisions. The second criterion is that the number of collisions in the segment must be statistically significantly in either the 3, 6, or 12-month period. The significance test is based on the comparison with the *expected* number of collisions. In turn, the expected number of collisions is based on the *collision rate* in the type of segment (i.e., rate group) and the *annual daily traffic volume (ADT)* estimated for the particular segment involved. Therefore, although a rate is involved in the calculation of the expected number, selection is based on a comparison of the actual number with the expected number.

Other Approaches: Other states use a variety of measures to defined high collision locations. In South Dakota and Nevada, engineers plot collisions on maps, visually identify HCCLs with a high number of collisions, and conduct collision analyses on these locations to find more information for possible remedies. The states of Iowa and Kentucky use the crash frequency/density method (2,3). Iowa uses a combination of crash frequency and crash rate (4). Kansas selects sites, intersections and roadway segments, based on collision rates. Idaho ranks sites based on both collision frequency and collision rate. The state of Washington has a minimum frequency criterion but also selects sites based on rate. Methods using frequency might just use "raw" frequency, but some methods model crash frequency as a function of volume (and other variables) (5-9).

Discussion: Approaches based on frequencies and rates have different advantages and disadvantages. The crash frequency method is simple to understand, and does not require additional information beyond number and location of crashes. It has the advantage of identifying sites which have a higher proportion of the overall crashes, and, if countermeasure cost and efficiency is relatively independent of traffic volume, then choosing sites based on frequency will be highly effective.

Of course, a disadvantage of frequency based methods is that traffic volume, and therefore, rate of crashes per vehicle, is not accounted for. Methods based on rates are able to identify site where the risk per vehicle is greatest, even if these locations don't have the highest absolute number of collisions. Since the risk to individual road users is higher, it may not be equitable or ethical to ignore high rate but low frequency locations in favor of high frequency but low rate locations.

The crash frequency and the crash rate method each have strengths and weaknesses. An ideal solution might involve using both methods; as a primary method choose sites that are based on frequency—this would enhance the efficiency of the process by allowing a focus on sites with a higher percentage of collisions; as a secondary method, choose sites with very high rates—this would address those locations where risk for individual road users is very high (12).

3.2 Quality Control

Caltrans Current Table C: "Quality Control" refers here to any procedure that uses a test of statistical significance to identify "unsafe" locations, as opposed to choosing locations that simply have a high number or rate of collision. The current Caltrans method of identifying high

collision concentration locations is to test whether the observed number of collisions significantly exceeds the expected number of collisions at each location. The expected number of collisions is calculated using the observed average collision rate for similar segment areas (i.e., those in the same "rate group") and the VMT (Vehicle Miles Traveled) estimated for that location. The calculation of the expected number of collisions from these two values is based on as assumption about the relationship between accident rates and traffic volume. Several decades ago Caltrans engineers computed the mathematical relationship between vehicle volume and crash rate for each type of segment. In some cases the rate increases with volume, in some the rate decreases with volume, and for others there is no relationship. The strong points of the current Caltrans approach are that it includes a statistical test for significance and incorporates vehicle volume and collision data. The presumed volume-rate relationships, having been derived several decades ago, are a potential source of error. A clear task is to update the volume-rate relationships using current TASAS data.

Other Approaches: There are three broad levels, or approaches, to "quality control." The first level does not use statistical tests to verify that high collision locations are due to change. Several states use an approach that falls in this category. These states include Georgia, Idaho, Nebraska, Kansas, Nevada, and others.

The second level involves a test of whether the actual number of collisions is significantly higher than the expected number of collisions. There are a number of ways to do this. In California, the expected number of collisions is based on an average rate of collisions in the entire set of locations in the same category of locations (i.e., in the same rate group). Another way is to calculate an expected number of collisions for a particular type of location based on a statistical model of the available collision data and roadway factors. This calculation is derived empirically and depends on the roadway types, the number of years of data used, the length of the segment being analyzed, the vehicle volume data for that segment, underlying assumptions about the relationship between vehicle volume and the collision rate, and assumptions about the probability distribution of collisions (usually either Poisson or Negative Binomial). Several of these issues in calculating the expected numbers of collisions are considered in this report (see other issue headings). The statistical test comparing the observed and expected number (or rate) of collisions often looks something like the following (when based on the assumption that collisions are Poisson random variables) (2):

$$Fc \geq Fa + k(Fa/M)^{1/2} + 0.5(1/M)$$

Fc = critical crash frequency or rate
Fa = average crash frequency or rate for that segment group
k = 3.090 for 99.9% conf, 2.576 for 99.5% conf, etc.
M = millions of vehicle miles for sections/spots

Thus, any crash site with greater than Fc collisions (or, Fc multiplied by segment length if using rate) would be identified as a high collision density area.

The third level or approach is often called the Empirical Bayes (EB) method. First the EB method computes the expected number of collisions based on a Safety Performance Function (SPF) specific to a certain roadway type. Instead of simply ranking the observed rate or number

(the first approach) or comparing the observed with an expected rate or number (the second approach), the EB method combines the observed and expected to produce a "true" number or rate. The underlying logic is that the expected number or rate contains information based on the entire set of segments and the observed contains additional information about a particular site that is not contained in the expected number. Therefore, combining these should produce a better estimate. The general formula for calculating the "true risk," or weighted expectation, is:

> Weighted Expectation = Weight x (Expected Safety for that segment based on SPF) + (1 – Weight) x (Actual number of collisions for that segment).
>
> where 0 < Weight < 1

The weight, a value between zero and one, depends on the relative variability of the expected and actual number of collisions. After this weighted expectation is calculated, it is compared to the actual number of collisions. If the actual number of collisions significantly exceeds this number, the site is identified as an HCCL. Thus, in the EB approach the observed number is compared to the weighted expectation whereas in the previous approach the observed is compared to the expected derived only from all similar locations combined or from a statistical model of all locations.

Discussion: In the next phase of the project, the research team will analyze how Caltrans' current method for calculating the expected number of collisions can be improved and whether the Empirical Bayes approach would result in a significantly better approach.

3.3 Weighting by Severity

Caltrans Current Table C: Caltrans currently does not weight collisions based on severity in determining high collision concentration locations.

Other Approaches: Most approaches to injury severity weighting use variations on the "equivalent property-damage-only" (EPDO) method. In this method, weights of fatal and injury crashes are compared to the weight of a PDO collisions. For example, the state of Iowa currently weights PDO collisions by 1, injury collisions by 5, and fatal collisions by 8 (2). Researchers at the University of Limburgh suggest weights 1, 3, and 5, respectively.

Another approach to using weights is to use numbers that reflect the actual cost of each collision. For example, Washington State assigns a weight of $1,100,000 for each fatal collision, $70,000 for each evident injury collision, $35,000 for each possible injury collision, and $6,500 for each property damage only collision. With these weights, the Washington DOT essentially analyzed "collision-dollars" per mile instead of collisions per mile. This method of course is formally equivalent to the EPDO since the essential feature is not the absolute amounts, but the ratio. Whether EPDO or cost approach is used, the ratio of the weights is usually based on average total costs of property, injury, and fatality collisions. Using these weights, a severity index is developed for each highway segment using the following formula:

$$SI = [WfF + WmM + WcC + P]/T$$

where ***SI*** is severity index, ***Wx*** are weights for fatal, major, and complaint of pain collisions, ***P*** is PDO collisions, ***T*** is total crashes at site (2).

Highway segments can be ranked by severity index, or the severity index of other criteria such as the crash frequency, crash rate, or can be integrated as part of the quality control methods discussed previously (2).

There are other methods to incorporate injury severity. For example, Kentucky ranks highway segments by a number of different dimensions; one ranking is of the total number of collisions, another is a rank by percentage of the total collisions in each segment that were injury and fatality collisions (3). Then, each segment's rank in each dimension is summed, and an overall ranking of highway segments is obtained (3). Another approach is to focus on fatal and injury collisions only; Colorado currently creates two versions of their "Table C": one for all collisions, and another for fatal and injury collisions.

Discussion: It is quite common to weight by injury severity, although the current Table C methodology does not.

The primary reason to weight collisions is to account for the increased burden or cost of specific types of collisions. For example, putting a larger weight on fatalities will mean that locations with fatal collisions are more likely to be identified as high risk locations.

However, there are several issues. One issue arises if severe collisions are more heavily weighted. Severe collisions tend to be rarer, and therefore the stability of estimates will be reduced, i.e., some locations might be identified based on one or two fatalities that arose "by chance" at those particular locations, and not because of something inherent in the locations. Clearly, in weighting by severity there is a trade-off with statistical stability.

Another issue is how to determine the weighting. Weighting by severity of injury is the approach used most often. However, other factors might be used in weighting, such as the cost of congestion or delay, which may be high even in PDO collisions.

A third issue is the relevance of the weighting by severity to highway factors. Severity does not always result from, nor is it sensitive to, highway factors, since severity depends on many other factors such as such as vehicle speed, vehicle type, seat-belt use, and other non-highway characteristics (2).

In the evaluation of the impact of different weighting of fatalities on the Iowa DOT safety improvement candidate location methodology, four different scenarios were considered:

- First fatality assigned value loss of a major injury (per location)
- An only fatality assigned value loss of a major injury (per location)
- All fatalities assigned value loss of a major injury (per location)
- Count only the first fatality per accident as a fatality, treat others as major injuries (per collision)

The results from their study clearly showed that the HCCL ranking process was impacted by the different scenarios. Two main recommendations were offered to minimize the bias towards fatalities in the final ranking. The first recommendation is to adopt a policy that minimizes the impact of a single fatality by adjusting the value loss contribution of that fatality. The adjustment is made to capture the random nature of crashes. For example, it is unlikely that the occurrence of only one fatality at a location in a five-year period indicates geometric or operation deficiencies at that location. The second recommendation is to treat the severity of multiple fatalities differently than single fatality crashes. A location that experiences several fatalities at once may be viewed differently than a location that regularly experiences fatalities. For instance, five fatalities at a location during a winter storm indicate a different problem than a location that experiences one fatality a year for five years (4).

The research team will consider analyzing previous Table C evaluations to examine if locations with fatal collisions were more, or less, easily remedied than non-fatal HCCLs. Specifically, the team will examine various methods of determining appropriate weights, the difference weighting can make as compared to not weighting, and how weighting impacts the stability, or robustness, of the HCCL identification process.

3.4 Segment Type

Caltrans Current Table C: Segment types are defined by a two-level set of categories. At the first level, segments are differentiated by (i) highway type, (ii) intersections, and (iii) ramps. Within each of these categories, sub-categories, or "rate groups" are defined. For highway type, rate groups are defined by traits such as the number of lanes, the speed limit, the ADT, the mix of vehicles on the road, and other factors. A particular route or highway is divided into smaller segments where the "border" between each segment represents a discrete location where any aspect of the road type changes. Caltrans has pre-defined 67 different types of highway segments based on the classifications described in the table below. Also, segments are defined by geographic location based on county lines; no segment contains roadway in more than one county.

Highway Type	1. Conventional 2 lanes or less 2. Conventional 3 lanes 3. Expressway 3 lanes or less 4. Undivided 4 lanes 5. Undivided 5-6 lanes 6. Divided 4 lanes 7. Divided 5 lanes or more 8. Divided expressway 4 lanes or more 9. Freeway 4 lanes or less 10. Freeway 5-6 lanes 11. Freeway 7 lanes of less 12. Freeway 7-8 lanes 13. Freeway 9-10 lanes 14. Freeway 11 lanes ore more

Terrain or ADT	1. [unspecified] 2. Flat 2. Rolling/mountainous 3. ADT less than or equal to 15,000 4. ADT greater than 15,000
Design Speed	1. [unspecified] 2. Less than or equal to 55 mph 3. Less than 45 mph 4. Less than or equal to 65 mph 5. Greater than 65 mph 6. Greater than 55mph 7. Greater than or equal to 45 mph
Area	1. Rural 2. Suburban 3. Urban

The second category of locations is intersections. There are 30 classifications for intersections that vary based on control type (no control, stop and yield signs, flashers, signals), intersection design (four-legged, multi-legged, offset, tee, others), and area (rural, suburban, urban).

Intersection Type	1. Four-legged, multi-legged, and offset 2. Tee, Y WYE, Other
Control Type	1. No control 2. Stop & yield signs (except 4-way) 3. 4 way stop 4. Signals 5. 4 way flashers
Area	1. Rural 2. Suburban 3. Urban

The third category of locations is ramps. There are 80 classifications for ramp type based on ramp design (frontage road, collector road, diamond, slip, loop, buttonhook, etc.), on- versus off-ramps, and area (rural, suburban, urban).

Ramp Type	1. Frontage Road 2. Collector Road 3. Direct, Semi-Dir Conn (LT TRN) 4. Direct, Semi-Dir Conn(RT TRN) 5. Diamond 6. Slip 7. Loop with left turn 8. Loop without left turn 9. Buttonhook

	10. Scissors 11. Split 12. Two-way ramp 13. Rest Area, Vista Pt, Truck Scale 14. Other
Ramp Areas	1-4
On/Off	1. On 2. Off
Area	1. Rural 2. Urban

Other Approaches: Most states analyze highways and intersections separately, however in the literature, no list of categorical segment variables was found. SafetyAnalyst considers different combinations of variables such as terrain, speed limit, and others. In the literature, no exhaustive list of categorical intersection or ramp variables was found.

Discussion: In the Department's review of the current HCCL methodology, 54% of investigators surveyed felt that the classification criteria of the rate-groups should be evaluated. The research team plans to further consult with traffic safety experts such as Ezra Hauer (consultant for this project) and Jake Kononov (Colorado DOT) to learn how the most modern methods categorize roadway types and what consequences result from these categories.

3.5 Segment Length

Caltrans Current Table C. Currently the Caltrans Table C methodology uses fixed locations to define what "rate group", or characteristics, a roadway belongs to. Within that area, Caltrans examines 0.2-mile segments. The process begins at the start of a highway segment and the first 0.2 mile segment is analyzed. If the segment is not found to be significantly unsafe, the algorithm dismisses the first 10% of that segment, and moves the 0.2 segment by 0.02 mile. For example, if "0" is the mile marker of the beginning of the segment, and the 0-0.2 mile segment is not significant, the algorithm continues to the 0.02-0.22 mile segment. If the segment is found to have a statistically significant high number of collisions, it is added to the output table and the algorithm moves ahead to the end of that mini-segment and begins considering the next 0.2 segment. This is generally called the "moving window" approach.

Other Approaches: Many states, such as Kansas, use pre-determined segment borders, or fixed locations. Using this method, segments are of variable length and might be defined by jurisdiction boundaries, major intersections, or other locations. These "fixed points" are permanent and a HCCL, or corridor, test is conducted for the entire segment at once, instead of breaking each segment into smaller categories.

There are several other approaches to highway segment length choice. In Iowa, highway segments have historically been generated using a link-node system for crash location.

> The link-node system involved the placement of nodes at locations including intersections, grade separations, bridges, ramp termini, severe curvature, and railroad crossings. These locations all have a unique identifier for its geographic location. Each crash at these locations is referenced to this unique location, or reference node. Crashes between these locations are referenced to both the nearest node (the reference node) and the node at the other end of the roadway link (the direction node), with a distance from the reference node specified as well. The total number of crashes that occur at each reference node and reference node/direction node pair can then be easily tabulated. However, only a list for reference node crashes is generated. However, the link-node system has been abolished [2002] and a switch to a coordinate-based system is in effect. Adjusting the Iowa SICL [HCCL] method to reflect this is one of the challenges for the Office of Traffic and Safety. (2)

Utah uses a similar fixed-point system for mile-long segments but also has the ability to use the "sliding-window" approach when necessary. The state of Washington uses a 0.1 mile segment, or less, in the case of small distances between highway segment types ("rate groups"). The state of New York uses 0.3-mile (0.5 km) segment.

Discussion: In the Department's review of the current Table C methodology, 68% of traffic engineers strongly endorsed a request to analyze segments currently ignored by the Table C methodology because they are less than 0.2 miles in length and lie between a positively identified HCCL in one rate group, and the beginning of a new highway segment classified as a different rate group. Also, 77% agreed that they often encounter pairs of HCCLs requiring separate investigation where the pair consists of two adjacent roadway segments.

There are two issues regarding segmentation length. The first, if segments lengths are fixed, is the segment length itself. A short segment length is appropriate if risk is localized. A number of studies suggest that risk conditions can vary rapidly over a fairly short highway length (5). A longer-length will generate a more stable estimate, and may be appropriate when highway conditions are fairly constant over an extended distance. Therefore, with a fixed-length system, the segment length is a judgment balancing specificity and stability.

The second, stemming from above considerations, is whether segments lengths should in fact be variable. A variable length approach would adjust to conditions where risk changes rapidly and adopt different lengths where risk is relatively constant for an extended distance. This would reflect the most realistic real-world condition, i.e., the fact that various roadway conditions that impact risk extend for variable distances.

A related issue is whether different types of collisions may take place on different types of roadway segments. For example, the frequency of run-off-the-road crashes over a long stretch of curvy mountain roads may be more significant over a longer segment, while the frequency of rear-end collisions may be higher in segments near intersections or specific spots on a corridor with lower speed limits under stop-and-go traffic conditions. We discuss this issue in a section below.

Using selective data set from the TASAS database, we will conduct a comparative analysis (i) to explore different segment lengths to determine how HCCL choices are impacted and (ii) explore variable length segments. One approach might be to model the state highway system to produce a continuous risk function, and then to determine how that continuous distribution might be segmented.

3.6 Analysis Period

Caltrans Current Table C: A roadway segment is selected if it meets the HCCL criteria (at least four collisions, number of collisions exceeds expected number of collisions based on 99.5% certainty) in any of the previous three, six, or twelve month periods. These selection criteria are defined as the crash frequency method and the crash rate method in the literature (1).

Other Approaches: In Kansas, where an automated state-wide analysis does not take place, different areas use different analysis periods. In more densely traveled areas, the two-year collision history is usually analyzed; in rural areas, a five-year history is typically analyzed; in some cases, a three-year history is used. Idaho calculates HCCL based on the most recent 3-year history; New York, a 2-year history. Washington uses a two-year period for high accident "locations" (segments less than 1 mile in length) and a 5-year period to identify high accident "corridors" (highway segments greater than one mile in length).

Discussion: The primary issue in determination of the analysis period is having a length of sufficient duration to generate a stable estimate, and yet short enough to spot trends happening over a shorter period of time. One issue in the selection of the analysis period is the significance and stability of data gathered in the accident database. The total number of accidents may reflect location-specific factors as well as a random component caused by a variety of other factors. If the sampling period is short the random component will be dominant and mask the underlying true risk. If the sampling period is longer, the random component will even out, allowing the true risk component to emerge. Another issue is that a shorter sampling period will lead to a larger number of "zeroes" in the data sets, which has complications on the selection of statistical models for the representation of accident data. (12)

The ideal sampling period may depend on the specific areas being studied, and particularly on variations in roadways and traffic patterns. In areas where the traffic level is steady and the roadway geometry has not shifted meaningfully, the longer analysis period will yield more reliable data. However, in high-growth or recently renovated areas, the monitoring of data over shorter time spans will be more likely to reflect current conditions.

Using a selective data set from the TASAS database, we will conduct a comparative analysis illustrating the impact of different analysis periods.

3.7 Classification

Caltrans Current Table C: Table C does not contain detailed information of each HCCL. Thus, investigators face a difficult task in researching the collisions at a particular location and attempting to identify an appropriate remedy.

Other Approaches: There are three types of approaches to providing more information in HCCL reports. The first type is a collision type-specific approach where the Department would decide to create a "Table C" for specific kinds of collisions. This approach is already used for "wet" highway collisions in order to generate a Wet Table C. The goal of this approach is to help engineers identify where drainage or slippery pavements might be the cause of an unusually high number of collisions. If desired, similar tables could be created such as a Roll-Over Table C, Broadside Table C, Rear-End Table C, a DUI Table C, etc.

The second type is a report with enhanced analysis of identified sites. Currently, Caltrans uses this approach because each of the HCCLs is investigated separately. Further automation could be added to the Caltrans method so that summaries or possible remedies could be identified in the Table C report. Such a method could simply summarize the primary collision factors (PCFs) at each location. Or, the method could be more complex and test if the proportion of PCFs was unusual. The main challenge to studying specific collision types is the additional complexity required in computing *base*, or *expected*, number of collisions of the specific type for each highway segment. The Empirical Bayes approach provides a technique that can handle this kind of analysis accurately.

Finally, the third approach is to add another component to the HCCL selection. Instead of identifying HCCL and then suggesting possible remedies, this third approach identifies locations as HCCL *only* if they meet the criteria of an HCCL *and* the location exhibits an unusual collision pattern. For example, if on average 20% of intersections collisions in California were rear-end collisions, but a particular intersection saw 60% rear-end collisions, that intersection would be selected as a "promising" HCCL because remedies are more obvious: separate turning lanes, etc. The state of Colorado currently uses this approach, and automates the process by adding pattern-recognition software to their HCCL identification algorithm. (13-14)

Discussion: In the Department's review of the current HCCL methodology, 66% of the respondents felt that the current methodology does not adequately identify locations that need improvement. The development of a pre-location selection scheme would be a high priority if the Department wishes the flexibility to examine a particular collision type (for example, DUI, speeding, or other) across the state. This capability could be an integral tool in collaborating with the CHP. However, pre-location selection is not an efficient means of identifying and treating all HCCL. Automated post-location analysis would be a very helpful tool if investigators continued to request a list of all of the HCCL. If the Department's goal is to decrease the number of HCCL listed in Table C, the third approach might be preferred. Like the

second approach, it provides the investigator with more information, but unlike the second approach, it only identifies a location as an HCCL if the collision pattern is unusual.

3.8 Traffic Volume Adjustment

Caltrans Current Table C: The current HCCL method relies on ADT estimates for each location. This information is updated by the Department's Traffic and Vehicle Data Systems Unit. The current HCCL methodology selects a location if the number of observed/known collisions significantly exceeds the number of expected collisions for a particular rate group and vehicle volume. Thus, this methodology depends not only on updated volume estimates but also on the assumptions on the relationship between volume and collision rates.

Discussion: In the Department's review of the current HCCL methodology, 53% of the respondents strongly agreed that HCCLs frequently result in "no action" (remedy) because the majority of collisions were related to peak-hour congestion. Although not applicable system-wide, PATH and TSC propose to investigate how the actual measured traffic data from PeMS (Performance Measurement System) can be utilized to facilitate analysis of traffic incidents that are tied to peak hours. The regions with reliable PeMS information can provide frequently updated traffic data with a higher fidelity. Another suggestion was also made to use traffic data from the Caltrans Weigh–in-Motion systems, which has a high degree of precision as well as vehicle classification information. Even though these data sets may not be available system-wide for the whole state, it will enable the exploration of accident distribution on some major freeways. If collision frequencies can be assessed for specific congestion times, the methodology might no longer identify areas that result in "no action" due to variable traffic volumes that are not accounted for the current methodology.

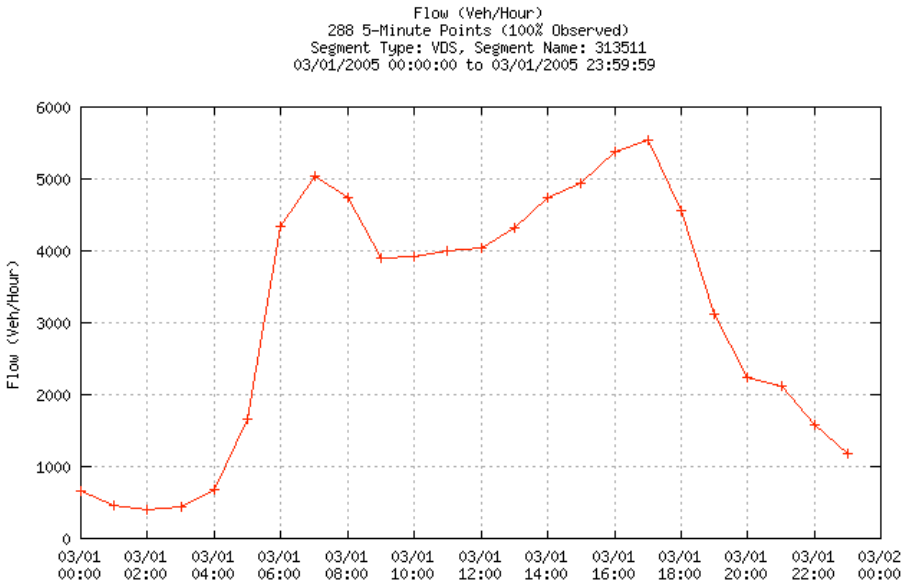

Figure 1. 24-Hour Traffic Flow on One Sample Data Station on I-80

The two graphs obtained from PeMS illustrate the degree to which traffic volumes can fluctuate greatly within a day or a week. The data was obtained on-line from PeMS by selecting a vehicle data station on eastbound Highway-80 near the city of Roseville. Figure 1 shows the fluctuation of traffic volume (vehicle counts) over a 24-hour span in a day. The peaks are clearly identifiable during the morning and afternoon rush hours. Figure 2 is an illustration of the daily traffic volume fluctuation during a one-month period. The peaks on this chart occurred repeatedly on Fridays, when the traffic traveling in the Lake Tahoe and Reno direction was considerably higher than the other days. The initial steps to take for the analysis of commuting related incidents will be to examine the number of incidents during selective hours of the day or selective days in a week. The total numbers of accidents or the distributions of accident types in the selective windows versus the overall distribution will provide the basis for evaluating the contribution of traffic volume and congestion related factors on the occurrence of incidents.

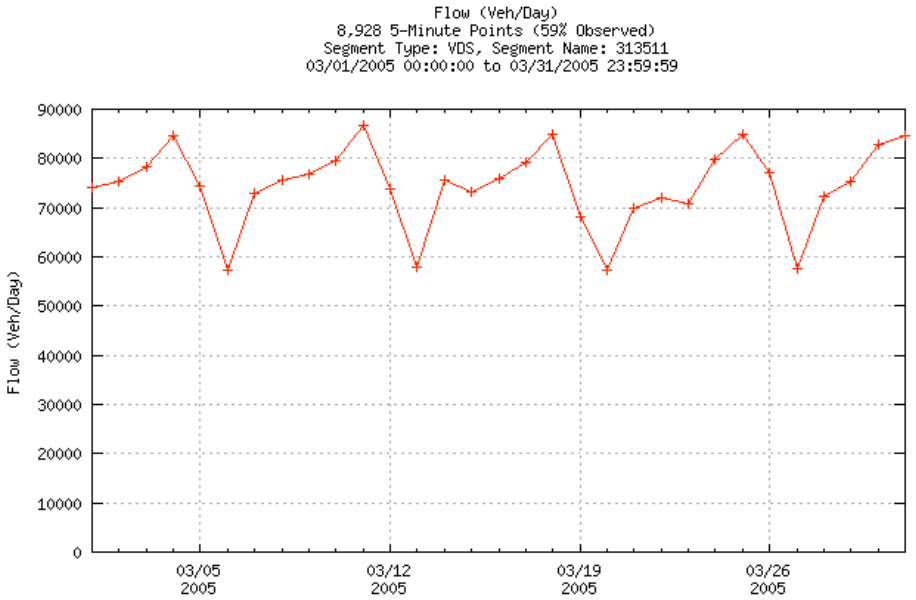

Figure 2. 7-Day Traffic Flow on One Sample Data Station on I-80

3.9 Non-Highway Factors

Caltrans Current Table C: The current HCCL method does not involve the screening of other factors to search for HCCL.

Discussion: A variety of factors can contribute to the occurrence of highway collisions, including at least the following major categories:

(1) Driver
(2) Vehicle
(3) Environment
(4) Traffic condition
(5) Demographics

(6) Alcohol outlets
(7) Events that cause unusual traffic patterns

For example, driver fatigue and inattention can result in a collision. Inclement weather may lead to poor visibility and slippery road surface may lead to loss of vehicle control. Sometimes, there are interactions among multiple factors. For instance, highway congestion and long delay may make drivers impatient, thus resulting in aggressive actions that directly or indirectly lead to a collision. Due to the potential complexity of accident causation, it is important to note that a search of HCCL without consideration of non-highway factors may lead to misrepresentation of HCCL that are really roadway related.

To overcome this potential deficiency, we suggest that the primary factor and contributing factors recorded in collision database should be used as a diagnostic tool to either pre-select before or to down-select after the statistical significance tests to reveal the locations that are more relevant and promising.

4. Findings based on Literature Review

This section summarizes the discussions in the previous section and provides an outline of the recommendations on the major issues for HCCL and follow-up actions in the next phase of the project.

Frequency versus Rate
The crash frequency and the crash rate method each have strengths and weaknesses. An ideal solution might involve using both methods. As a primary method, sites are chosen based on frequency — this would enhance the efficiency of the process by allowing a focus on sites with a higher percentage of collisions. As a secondary method, sites with very high rates are chosen — this would address those locations where risk for individual road users is very high (12).

Quality Control
In the next phase of the project, the research team will analyze how Caltrans' current method for calculating the expected number of collisions can be improved and whether the Empirical Bayes approach would result in a significantly better approach.

Weighting by Severity
The research team will consider analyzing previous Table C evaluations to examine if locations with fatal collisions were more, or less, easily remedied than non-fatal HCCLs. Specifically, the team will examine various methods of determining appropriate weights, the difference weighting can make as compared to not weighting, and how weighting impacts the stability, or robustness, of the HCCL identification process.

Segment Type
The research team plans to further consult with traffic safety experts such as Ezra Hauer (consultant for this project) and Jake Kononov (Colorado DOT) to learn how the most

modern methods categorize roadway types and what consequences result from these categories.

Segment Length
Using selective data set from the TASAS database, we will conduct a comparative analysis (i) to explore different segment lengths to determine how HCCL choices are impacted and (ii) explore variable length segments. One approach might be to model the state highway system to produce a continuous risk function, and then to determine how that continuous distribution might be segmented.

Analysis Period
In the upcoming tasks, we suggest that we use a selective data set from the TASAS database and conduct a comparative analysis with several analysis periods. Preferably, the selected data will contain roadways of different categories or from different regions. The variability of the resulting HCCL, compared to previously identified HCCL, will provide us with information to explore the effects of analysis period.

Classification
If the Department's goal is to decrease the number of HCCL listed in Table C, the third approach (discussed in subsection 3.7 above) might be preferred. It only identifies a location as an HCCL if the collision pattern is unusual.

Traffic Volume Adjustment
The initial steps to take for the analysis of commuting related incidents will be to examine the number of incidents during selective hours of the day or selective days in a week. The total numbers of accidents or the distributions of accident types in the selective windows versus the overall distribution will enable us to evaluate the contribution of traffic volume and congestion related factors on the occurrence of incidents.

Consideration of Non-Highway Factors
To overcome this potential deficiency, we suggest that the primary factor and contributing factors recorded in collision database should be used as a diagnostic tool to either pre-select before or to down-select after the statistical significance tests to reveal the locations that are more relevant and promising.

5. GOALS AND PERFORMANCE MEASURES OF HCCL IDENTIFICATION

The project of TO 5215 is initiated with specific near-term and longer-term objectives and goals in seeking improvements of HCCL identification methods. This section contains a brief discussion of the general goals, performance measures, and critical issues and variables related to HCCL.

5.1 Goals of HCCL Screening

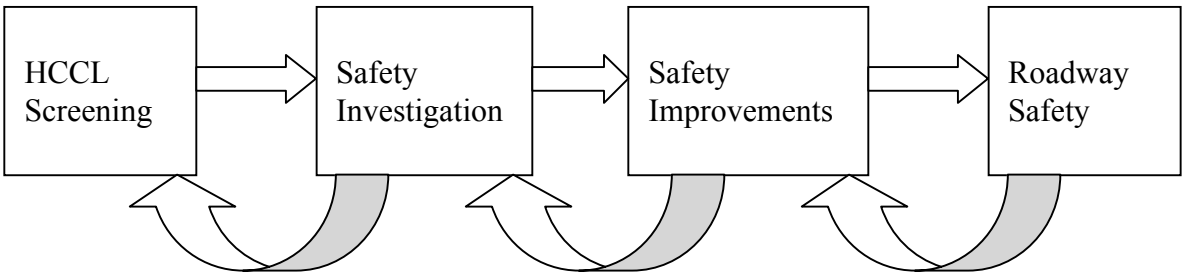

Figure 3. Safety Investigation and Improvements Process

The identification of HCCL, reported in Caltrans Table C quarterly after screening, provides a list of locations that deserve or require safety investigation. The accuracy and reliability of these lists have direct impact on the use of time and manpower that are dedicated to the safety investigation process. Furthermore, the results of the safety investigation then significantly dictate the allocation of resources for safety improvements and ultimately safety performance of roadways. A representation of this safety improvement process is depicted in Figure 3. As the starting point of the chained process, the screening of HCCL plays a critical role and has subsequent ripple effects on roadway safety.

Given the essential role of the HCCL screening module within the whole safety improvement process and the sequential effects caused by its outcome, it is evident that the overall system can be more robust and efficient if positive feedback is provided from the other functional modules or sub-systems in the process, which is illustrated by the curved arrows in Figure 3. The quality of the HCCL screening outcome can be enhanced if the performances of subsequent sub-systems are evaluated and their results are used for the revisions and improvements in the HCCL screening module.

5.2 Performance Measures of HCCL Screening

The functional performance of a HCCL identification system, with a given set of roadway and collision data, can be evaluated by a direct performance measure, the validity of these "hot spots" revealed by the safety investigation efforts that follow. In other words, the performance of a HCCL identification system is judged by the relevancy and consistency of location-specific causes that are found among the list of identified HCCL. If there is a lack of matching between the outcome of a HCCL screening module and the results of field investigation, then there is room for potential improvements in the HCCL screening methods.

In a hypothetical situation when all conditions remain the same, if two HCCL outcomes (two Table C generated by different screening methods, for instance) are available for comparison, one is considered to perform better if the subsequent investigation allows a better identification of location issues. This is particularly meaningful from the perspective of a highway operation agency, such as Caltrans, if the location-specific issues are roadway-related. Therefore, a

comparison of the percentage of meaningful locations revealed by the "two" Table C is a direct and immediate measure of performance. Furthermore, the measure of performance can be extended to include the implementation of safety improvements and to be evaluated for the cost-benefit returns of safety investment. Similarly, the accountability of the HCCL system will be ultimately examined by the system-wide roadway safety indicators, such as the reduction of collisions and fatalities.

In a realistic situation, the feasibility to repeat the safety investigation process for "two Table C" is limited on the available resources even if the roadway or traffic conditions have not changed at all. Thus, several alternative performance measures can be considered:
(1) Limit the full-scale investigation to a subset of locations from the outcome of the screening module.
(2) Survey field investigators, who are familiar with the roadways, with an alternative Table C and solicit their feedback on the validity of "uncommon locations," which are not present in the previous Table C.
(3) Compare the percentage of identified HCCL that lead to fruitful investigation results from the alternative Table C to see if there is a better matching of valid HCCL. A higher percentage is desirable.

A related issue in the evaluation of HCCL performance is the effectiveness and thoroughness of field investigation. There are established guidelines and proper training that are given to field investigators. Besides the emphasis of sufficient support being provided, an alternative perspective is to seek additional information that can be extracted from the roadway and collision database to enhance the effectiveness of field investigation. For example, if a particular type of collisions and a specific type of primary factors are prevalent in incidents at one location, the inclusion of such data can be additional information or tools that field investigators can use for pin-pointing potential safety improvements that can be most cost-effective.

5.3 Other Factors and Considerations for HCCL Screening

There are many other factors or variables that are significant in the identification of HCCL. The following items were identified to be of critical values. These should be taken into considerations in seeking the improvements of Table C in conjunction with the statistical methods and the associated technical approaches.

(1) Analysis Period
The current procedure in HCCL screening is to generate an updated Table C every quarter, or four times a year. The intention of this relatively frequent updating period is to capture emerging trends if existent. Other agencies generally adopt a much longer update interval, for example on a yearly basis. Dr. Hauer indicated that collision statistical trends tend to appear relatively slow and they may not be reflected in collision data quickly enough for the transition in trends to be detected.

(2) Weighting of Collision Severity

The weighting of collision severity provides a linkage to the direct and indirect costs associated with the consequences of collisions. This is a meaningful approach if the choices of weighting ratios are meaningful and equitable. A separation of property-damage-only (PDO) and injury and fatal accidents can by useful information, even if no specific weighting is assigned to each accident type. However, if the HCCL are to be ranked by their priority a proper weighting system will be desired.

(3) Segment Length
In order to understand the effects of segment length selection, selective case studies will be necessary to compare the outcome of Table C with different segmentation. Once the simulation screen models are functional with the TASAS database, optional choices of segment length will be experimented to compare the results of HCCL screening.

(4) Highway Factors
A variety of factors can contribute to the occurrence of highway collisions, including at least the following major categories:

- Roadway geometries and configurations
- Road attributes, such as curves, grades, medium types, shoulder widths, etc.
- Roadway modification history
- Safety improvement history

Since TASAS contain a set of highway information, it will be beneficial to explore any specific factor or correlations. This particular subject is not included in the current project, but should be investigated in future efforts.

(5) Non-Highway Factors
A variety of factors can contribute to the occurrence of highway collisions, including at least the following major categories:

- Driver
- Vehicle
- Environment
- Traffic condition
- Demographics

Since TASAS contain a rich set of highway and incident information, it is beneficial to search and identify potential patterns in the variables above for diagnostic purposes. This in turn will facilitate an effective follow-through of safety investigation and safety improvements. Depending on the fidelity and accuracy of data, a variety of environmental conditions and driver information can be distilled and built into a detailed screening of collision data. The traffic conditions can be at best captured by associating highway traffic volume that can be potentially extracted from other data bases, for example those highlighted in Section 4.

6. TECHNICAL APPROACHES AND ISSUES IN HCCL IDENTIFICATION

This section provides an overview of technical approaches and issues in the method of HCCL identification.

6.1 Existing Method for Generating Table C

The current method of Table C (HCCL screening reports) generation is based on a procedure that has been adopted and implemented by Clatrans in the last several decades. Currently, the actual software codes have been transitioned from a legacy system into a relational database. Part of the efforts in this project is to emulate at least a portion of the software functions based on statistical data analysis tools so that the outcome of Table C reports can be evaluated if certain parameters and variables are adjusted or altered.

The exact and actual procedure of generating Table C can be consulted with Caltrans, and is not given here. However, an outline of the process can be explained as follows for highway segments. The method for intersections and ramps are done separately but use the same concept.

(1) In the highway database, the categories of roadway segments are divided into a number of rate groups. Each group has its own characteristics of average accident rate as a function of roadway geometries and traffic volume.
(2) The average numbers of collisions are calculated for each segment type accordingly.
(3) A screening window is moved through each state highway system in searching for segments that indicate high concentration. The analysis is done for every 0.2 miles segments of highway at a time.
(4) To be flagged as a high-concentration site, the actual collision occurrence numbers and the average number are compared in a significance test. The significance test is to determine if the defined highway segments, ramps or intersections have an accident count that is significantly higher than *the number of accidents required for significance*.
(5) The numbers of accidents required for significance are derived from the Poisson's Distribution Curve with appropriate correction factors. 99.5% confidence level is used for Table C.
(6) For Table C, if the locations have *4 or more collisions* and *are significant* in either 3, 6, or 12 months period then the locations are labeled "REQ" in the output table.
(7) For wet Table C, if the locations have *3, 6, 9 or more collisions* and *are significant* in either 12, 24, or 36 months period respectively then the locations are labeled "REQ" in the output table.
(8) Accident investigators are required to investigate the "REQ" locations.

6.2 New Methods for Locating HCCL Using Crash Models

During recent years, many studies have been conducted to use crash models in order to help detecting sites with abnormally high frequency of accidents. One popular method is the following. It can be decomposed into two major steps. The first step is to estimate the average safety of intersections, highway segments or ramps of a specific type depending on their traffic volumes. This estimate is given by the so called Safety Performance Functions, determined by using generalized linear model fitting. The next step is estimating the expected accident frequency of a particular site using the empirical Bayes method. This method uses both observed accident count and Safety Performance Functions to come up with the expected accident count at the site of interest. The advantage is that the resulting variance of the weight average of both quantities is inferior to the one that would be obtained by using only one of these two.

6.2.1 Safety Performance Functions

Safety Performance Functions model crash frequencies as a function of traffic volumes and other variables thought to be significant. Many different functional forms for these SPF have been postulated and, although some are more popular than others, none have been universally acknowledge to be the correct one to use. The variables used can also be very different from a SPF to another. These variables can be related to design, such as horizontal and vertical alignment, traffic speed, weather or other factor.

The parameters of SPF are estimated using generalized linear model fitting, with a chosen error distribution (usually a negative binomial error distribution), on a particular set of data. The data used is checked, modified and corrected to improve its quality and thus to allow for more precise estimators. This data review is critical and must be carefully undertaken.

Safety Performance Functions provide valuable information about the expected number of accidents for a specific type of highway, intersection or ramp. They are usually provided with an overdispersion parameter that accounts for the precision of the estimator that can be obtained using these. As it is usually defined, the bigger the overdispersion parameter, the smaller the error.

6.2.2 The Empirical Bayes method

The empirical Bayes Method is a way to combine two different clues: the observed count of accidents at a specific site and the results given by SPF. Both clues are valuable as the first one gives the real number of accidents that occurred at the particular site studied and the second one gives the average number of accidents at similar sites. Using only one of these two clues may lead to results of lesser quality.

The major issue of using only the observed count is what is called the regression-to-the-mean bias. This comes from the insufficient number of observations available. Indeed, it is hard to obtain more than 10 years of crash data for a site. Taking into account only a few observations provides a bad precision and leaves plenty of room for randomness. The results provided might suggest investigating locations for improvements that had an unusual accident crash frequency only due to random phenomenon.

As maintained earlier, Safety Performance Functions provide only an estimate of the expected number of accidents for intersections of a particular kind. The categories are defined by the choice of variables included in the model. All the factors influencing crash frequencies cannot be considered and thus the model does not distinguish two intersections, most probably different in terms of safety, within the same category.

The estimator given by the Empirical Bayes Method is obtained by a weighted average of the two clues. The weight is calculated as a function of the overdispersion parameter in a way that minimizes the resulting variance. The Empirical Bayes Estimator is considered to perform better

than traditional estimators and can be used to improve the accuracy of identifying dangerous locations.

6.3 Quality Control

In the current Caltrans approach, the locations of 99.5% confidence level in a Poisson's distribution model are identified for required field investigation. One potential problem in the use of a fixed percentage is the limitation of the imposed boundary defined by the selected numbers. If the sampling size is huge, then the locations contained in Table C can be significant large even if only 0.5% is identified. One suggestion from Dr. Ezra Hauer, as in the practice of the state of Colorado, is the ranking of potential HCCL instead of a fixed percentage of statistically significant locations. With the alternative way of ranked locations, even though the number of locations can still be capped, the investigation efforts are scheduled and conducted according to their priority within the list.

The other main issue in quality control is the reliability of data. Here, an Empirical Bayes (EB) method is preferred because of its ability to handle two specific problems. [16] It increases the precision of estimates, and it corrects for the regression-to-mean bias. See Appendix A for a summary of mathematical formulation and descriptions. The use of EB methods corrects for the random fluctuations of data, which allows for a more reliable test of statistical significance. This is accomplished by a weighting allocation between the observed numbers of actual collisions and the expected numbers of collisions at a roadway location.

One additional issue to consider in terms of quality control in HCCL screening is the estimation of expected numbers of collisions for various roadway segments, ramps and intersections. For example, Caltrans currently categorize various roadways into a number of rate groups, which are associated with certain forms of equations with specific parameters to estimate the expected number of collisions. This becomes the baseline of significant test against the actual number of collision occurrence. A further investigation into the use of "expected numbers" will be useful at several levels:

(1) The use of rate groups and their associated equations are based on historical patterns of safety characteristics of roadway segments or locations. It will be beneficial if the calculation formulae are validated with current traffic attributes.
(2) There are additional roadway attributes that may not be fully captured by the rate group classification. For example, grade and alignment are not part of existing highway database even though they have strong effects on the accident rates.
(3) Certain traffic attributes may not be existent or available. For example, cross street traffic volume are generally not available even though they are closely related to safety performance of intersections.
(4) In recent years, there have been significant developments in modeling roadway safety by the use of safety performance functions (SPF). See Appendix C and D for further explanations of SPF. In order to systematically generate reasonable expected numbers for significance tests in HCCL screening, the modeling of SPF and/or joint validation of rate groups will be strongly desirable.

7. HIGHWAY, COLLISION AND TRAFFIC DATA REVIEW

In order to accomplish a realistic evaluation of HCCL, we have acquired highway and accident database from Caltrans and subsequently revised and implemented a format compatible with SAS – a statistical tool. This is a significant endeavor, as the work involved in exploring, understanding, merging, and utilizing the data has turned out to be much more than expected. Nevertheless, we consider this a critical step in laying the foundation for further tasks.

The criticality of data is indisputable. The reliability of HCCL screening, Table C, can only be confidently trusted if the data integrity and accuracy is ensured. Furthermore, many embedded parameters and functions, such as expected number of collisions on a certain type of roadways, can only be reasonably estimated if the highway information is correct and the traffic flow data corresponds well to the real-world conditions. Therefore, in the course of this project, we are very interested in understanding and investigating various aspects of the data, even though our focus of the project remains the evaluation of HCCL methodologies.

7.1 TASAS Data

Traffic Accident Surveillance and Analysis System (TASAS) is a computerized database that contains the historical records of accidents on roadways under the state highway system. Along with collision data, a highway database includes a set of geometric and location data for highways, ramps and intersections.

7.1.1 Data Description

Traffic Accident Surveillance and Analysis System (TASAS) data was received in four separate excel tables for accident, highway, intersection, and ramp files. The total data set contains ten years of historical data. For statistical analysis of the data, SAS®, a comprehensive statistics package for analysis and data manipulation, will be used. For the utilization of data by SAS, the raw data was imported into SAS using internal SAS command interface. There were some special characters in the highway file that was deleted before conversion from Excel formats into SAS.

4.1.2 Issues Identified in Data Review

Some deficiency in data was found in all four files. Some noticeable problems include, but are not limited to:

- In the accident file, some accidents are identified as “ramp” incidents, but their post miles fields are marked at locations before the post mile in the ramp file starts.
- In the accident file, there are ramp accidents that do not match any post mile in the ramp file.
- In the highway file, there are a large number of overlapping segments.
- The highway accidents at some post miles fall in two segments of the highway data due to overlapping highway segments.
- The post mile values of some intersection do not match any segment in the highway data.
- There are intersection accidents that do not match with any location in the intersection data.
- Some sites do not have any rate groups

For further testing of HCCL screening, it will be necessary to clean up the data and resolve aforementioned issues in order to obtain reliable results. However, the efforts required to thoroughly correct the complete data set is beyond the scope of this project. Therefore, selective modifications may be performed if it is necessary to do so for the related tasks. For example, if a particular highway in a certain county or district will be evaluated for specific studies, it will be prudent to ensure the data integrity before reliable outcome can be achieved.

7.1.3 Data Analysis

To analyze data and simulate table C, accident file was merged with highway and intersection files. In merging the data, one important step was to define an appropriate segment from highway file to which a corresponding post mile marker from the accident file belongs.
The rate groups were assigned to each type of roadways according to their definition from Table C and Wet Table C Overview. The corresponding rates associated with each roadway type were then used to calculate the average and expected number of collisions based on annual average daily traffic volume (AADT).

An SAS program for analyzing highway segments was developed. One problem was discovered during highway analysis. This problem occurs when a segment in a Highway Rate Group that is less than 0.2 miles is currently ignored or not documented in the Table C and Wet Table C Overview. For example, if a Highway Rate Group is 0.5 miles long. If the first and second 0.2 miles segments are significant, then the last segment in the analysis for this Highway Rate Group will include 0.1 mile of the next Highway Rate Group. In this case, the analysis will stop and restart at the beginning segment of the next Highway Rate Group, and the last 0.1 mile of the previous Highway Rate Group will be ignored.

Another problem during Highway analysis appears when moving window is reaching the "N" area of an intersection—250 feet beyond the intersection. The analysis process will stop and restart beyond the "N" area, since accidents at intersections have already been analyzed in Intersection Analysis and will not be analyzed in the Highway Analysis. The collisions coded outside the intersection but within the 'N Area' (usually 250 feet) will have a File Type = 'H' however they are also included with the Intersection analysis. It means that some collisions are included twice as in highway file as in intersection file.

7.1.4 Next Steps in Data Analysis

The research team has initiated an effort in investigating the issues, some of detail, and fidelity of data, in the process of emulating the method of generating Table C. We will continue these efforts in constructing a simulation model as well as communicating to Caltrans in further understanding the use of data

7.2 PeMS (Performance Measurement System) Data

One question in the evaluation of collisions on roadways is whether there is a high-concentration phenomenon during the rush hours due to heavy commuting traffic volume and/or congestion. The answers to such question will provide additional clues for conducting investigation of specific locations and strategizing safety improvements if such hypotheses on congestion or commute-related phenomena are valid. In order to understand the importance of roadway-specific factors among all incidents, it will require an evaluation of traffic data in conjunction with collision database.

Typically detailed and higher-fidelity traffic data are not systematically stored or available for analysis. However, Caltrans over the years have implemented a Performance Measurement System (PeMS) for statewide highway systems at locations where vehicle detection or traffic measurement is available. Information about PeMS is available at the following link: http://pems.eecs.berkeley.edu/Public/. For the task of investigating the correlation between traffic flow fluctuation and collision occurrence, a set of traffic data was downloaded from the PeMS system and subsequently reviewed in preparation for further analysis.

7.2.1 Exemplar Traffic Data

The consistency and availability of data vary at different districts and highways across California. When actual measurement data is not available, PeMS utilized existing data to extrapolate and estimate the expected values. For the initial evaluation, a stretch of Interstate Highway I-5 in District 11 is selected due to its data availability. Vehicle detection stations are existent from California mileposts 17.02 to 48.14 in the northbound direction, and 16.12 to 53.29 in the southbound direction. A total of 365 days of data in Year 2003 was downloaded. The data was then segmented into one-hour intervals for each day with the overall average calculated for the whole year. In the following sections, the following plots were shown:

(1) Flow (vehicle count)
(2) Speed (mph)
(3) Vehicle-miles-traveled
(4) Vehicle-hours-traveled

7.2.1.1 I-5 Northbound

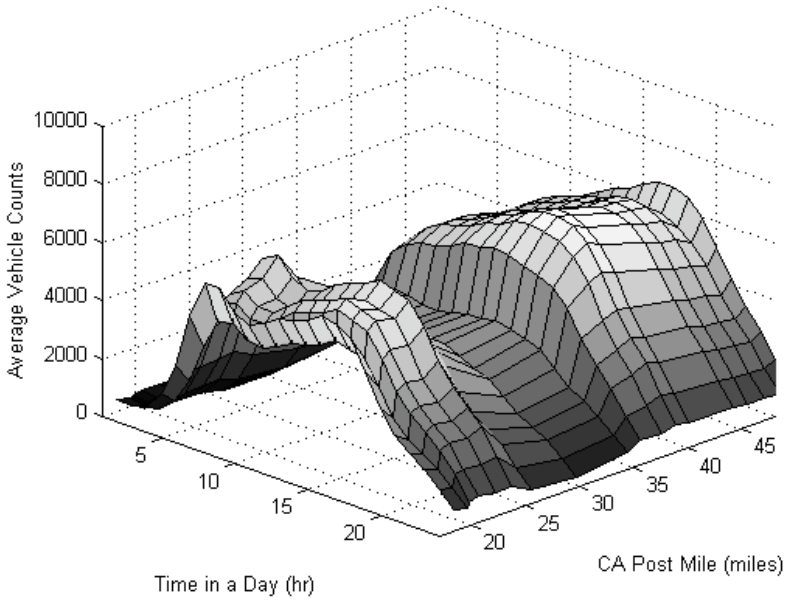

Figure 4a. I-5 Northbound Flow (Vehicle Count)

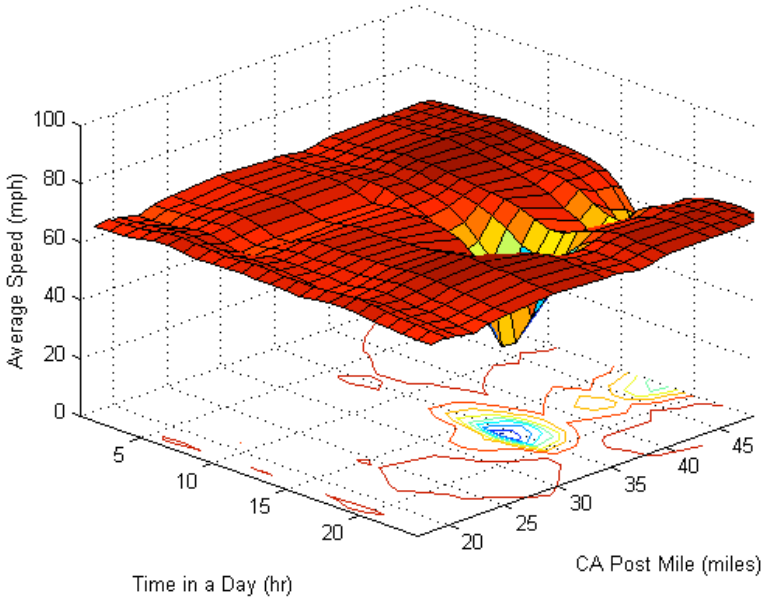

Figure 4b. I-5 Northbound Speed (mph)

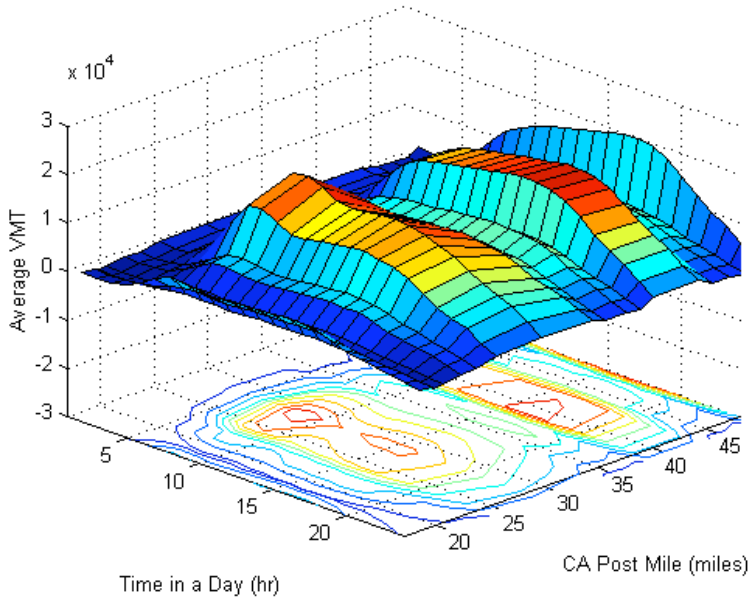

Figure 4c. I-5 Northbound Average Vehicle-Mile-Traveled

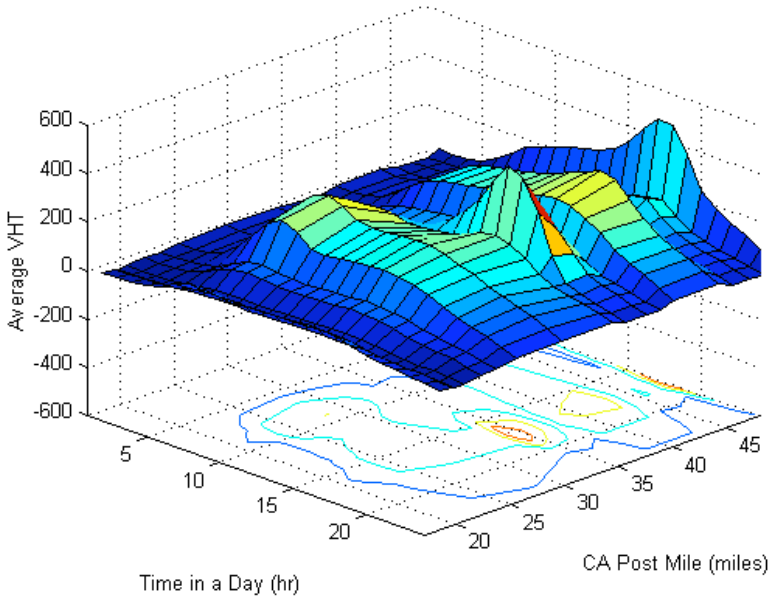

Figure 4d. I-5 Northbound Average Vehicle-Hour-Traveled

7.2.1.2 I-5 Southbound

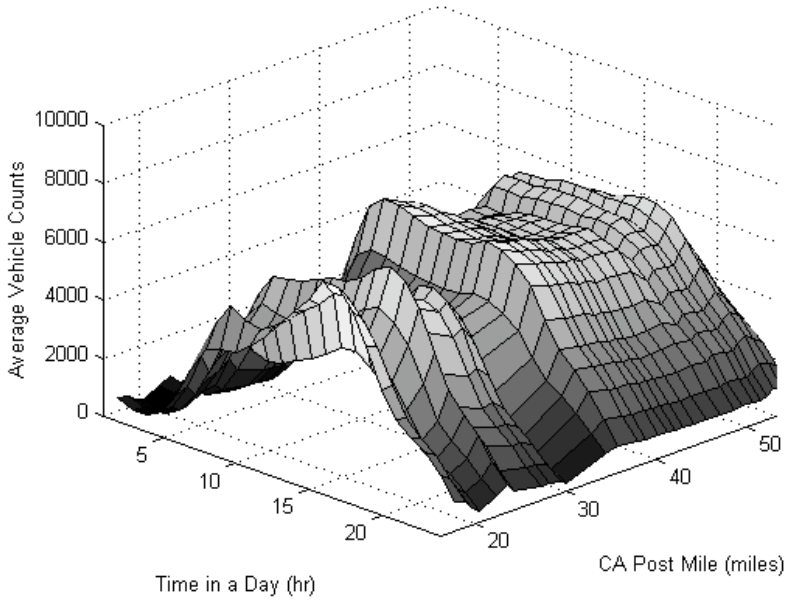

Figure 4e. I-5 Southbound Flow (Vehicle Count)

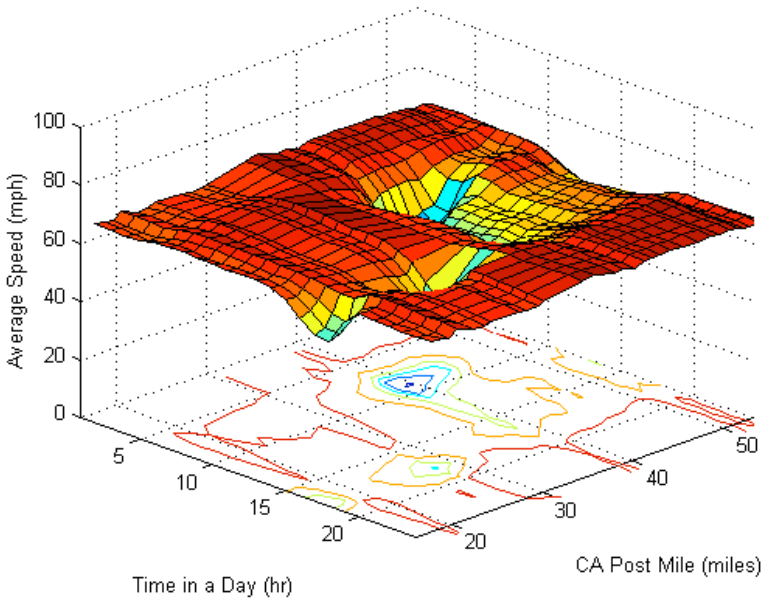

Figure 4f. I-5 Southbound Speed (mph)

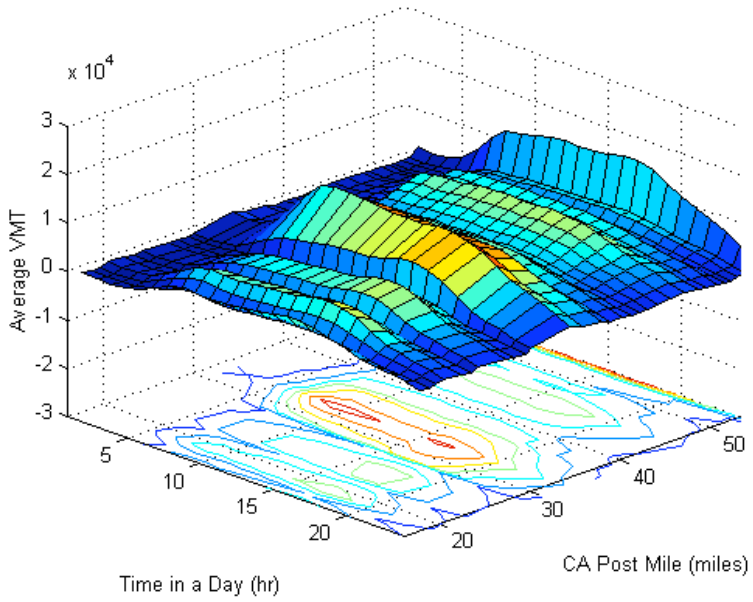

Figure 4g. I-5 Southbound Average Vehicle-Mile-Traveled

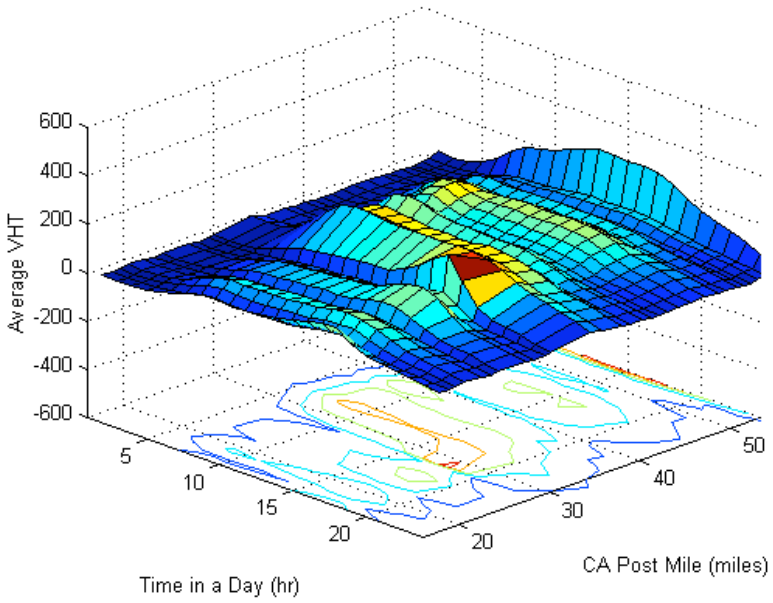

Figure 4h. I-5 Southbound Average Vehicle-Hour-Traveled

7.2.2 Discussions

As can be seen from the plots of exemplar data shown above, the variations of traffic volume are quite significant over different hours of a day and over different segments of the highways. In typical evaluation of collision occurrence, the accident rate is commonly expressed as a function

of the traffic volume, such as described in Safety Performance Functions (SPF). With the availability of detailed traffic data, the distribution of collisions in different time periods can be calculated. Furthermore, the relationship of traffic flows and the accident rates can be validated by analyzing the distribution of accident occurrence at different segments and in different hours. Moreover, if collision types can be identified from the collision database, it can also be determined whether a particular type of collisions manifests under heavy commute traffic at specific locations. These issues will be evaluated in the upcoming months in the project.

8. Summary

HCCL screening and identification is a critical step in the process of improving and ensuring roadway safety. [17-19] In this report, certain critical issues are highlighted for the proper selection of HCCL screening methods. Work on the evaluation of highway and collision data is also described.

The goal of this project is to enhance the current method implemented for the generation of Table C. This report represents the second deliverable of the project. At the end of the project, a set of recommendations will be made to indicate the most effective ways of seeking improvements in the performance of Table C.

References

1. Caltrans Table C Task Force Summary Report of Task Force's Findings and Recommendations. September, 2002.
2. Pawlovich, M.D. Safety Improvement Candidate Location (SICL) Methods. Iowa Department of Transportation, Highway Division, Engineering Bureau, Office of Traffic Safety. 2002.
3. Evaluation of High Traffic Crash Corridors. Kentucky Transportation Center, College of Engineering. 2002
4. Evaluation of the Iowa DOT's Safety Improvement Candidate List Process. Center for Transportation Research and Education (CTRE), Iowa State University. June 2002.
5. Markos Papageorgiou, *Fellow, IEEE*, and Apostolos Kotsialos. Freeway Ramp Metering: An Overview. IEEE Transactions on Intelligent Transportation Systems, Vol. 3, NO. 4, December 2002.
6. Geurts K., Wets G., Brijs T., Vanhoff K. Clustering and Profiling Traffic Roads by Means of Accident Data. Limburg University. 2003
7. Ivan J.N., Wang C., Bernardo N.R. Explaining Two-Lane Highway Crash Rates Using Land Use and Hourly Exposure. University of Connecticut, Connecticut Transportation Institute, Civil and Environmental Engineering. 1999.
8. Geurts K., Wets G., Brijs T., Vanhoof K. Identification and Ranking of Black Spots: Sensitivity Analysis. Limburg University 2003
9. Geurts K., Wets G., Brijs T., Vanhoof K. Profiling High Frequency Accident Locations Using Association Rules. 2003.

10. Sayet T., Navin F., Abdelwahab W., A Countermeasure-based Approach for Identifying and Treating Accident Prone Locations. 2003.
11. Hauer, E., Hardwood, DW, Council, FM., Griffith, MS. Estimating safety by the empirical Bayes method: a tutorial. http://ca.geocities.com/hauer@rogers.com/Pubs/TRBpaper.pdf Accessed 3/2005
12. Hauer, E. SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites: Task K: White Paper for Module 1-Network Screening. Federal Highway Administration Task No. DTFH61-01-F-00096. December, 2002.
13. Lord D., Washington S.P., Ivan J.N., Statistical Challenges with Modeling Motor Vehicle Crashes: Understanding the Implications of Alternative Approaches. Transportation Research Board, 2004.
14. Kononov, J. (2002) Use of Direct Diagnostics and Pattern Recognition Methodologies in Identifying Locations with Potential for Accident Reduction. Transportation Research Record. 2002.
15. Kononv,J. and Janson,B. (2002) Diagnostic Methodology for Detection of Safety Problems. Transportation Research Record. 2003.
16. Hauer, E., Hardwood, DW, Council, FM., Griffith, MS. Estimating safety by the empirical Bayes method: A tutorial. *Transportation Research Record 1784*, pp. 126-131. National Academies Press, Washington, D.C.. 2002. http://ca.geocities.com/hauer@rogers.com/Pubs/TRBpaper.pdf
17. Ezra Hauer, Bryan K. Allery, Jake Kononov and Michael S. Griffith. Cost Effective Evaluation of Screening Criteria. Transportation research Board. 2004. http://ca.geocities.com/hauer@rogers.com/Pubs/TRB2004CostEffectEvaluationofScreeningCriteria.pdf,
18. Hauer, E., J. Kononov, B.K. Allery and M.S. Griffith, Screening the road network for sites with promise. *Transportation Research Record 1784*, pp 27-31 National Academies Press, Washington, D.C.., 2002 http://ca.geocities.com/hauer@rogers.com/Pubs/ScreeningforSWIPs.pdf
19. SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites: Task K: White Paper for Module 1-Network Screening. Federal Highway Administration Task No. DTFH61-01-F-00096. December, 2002.

Appendix A: Table Comparing States' HCCL Methodologies

Tabulated Comparison and Highlights of Approaches Taken by Various States

	CA	SA*	FL	GA	IO	IL	KS	WA	CO	NB	NY	ND	OH	OR	PA	SC	SD	
Freq	√	√		√	√	√		√	√			√	√	√	√	√		**12**
Rate	√	√	√	√	√	√	√		√	√	√	√	√	√		√	√	**15**
Quality Control	√	√	√					√	√	√	√					√		**5**
Weight		√		√	√	√	√	√				√	√	√	√	√	√	**12**
Separate Ranking	√	√	√	√		√	√	√	√									**8**
Analysis Period	3,6,12 mon		1,2,3,5 yr	3 yr	3 yr	3 yr	2,3 or 5 yr	2 yr		2 yr	2 yr	1, 3 yr	3 yr					

SA* describes the SafetyAnalyst System

Appendix B: Sources of Information

Interviews:

Kansas DOT: David Church, Chief of Traffic Engineering at KSDOT, Church@ksdot.org
Georgia DOT: Jack Carver, Accident Analyst, Jack.Carver@dot.state.ga.us
Idaho DOT: Steve Rich, Principal Behavioral Data Analysis and Dissemination, Steve.Rich@itd.idaho.gov
Minnesota DOT: Loren Hill, Loren.Hill@dot.state.mn.us
Missouri DOT: Ron Beck, Director, Missouri Statistical Analysis Center. Ron.Beck@mshp.dps.mo.gov
Nebraska DOT: Randy Peters, Traffic Engineer/Division Manager, rpeters@dor.state.ne.us
Pennsylvania DOT: Bill Crawford, Highway Safety Engineer, wcrawfo@dot.state.pa.us
South Dakota DOT: Cliff Reuer, Traffic & Safety Engineer, Cliff.Reuer@state.sd.us
Oregon DOT: Tim Burks, Highway Safety Coordinator, Timothy.W.Burks@odot.state.or.us
New York DOT: Robert Limoges, rlimoges@dot.state.ny.us
Wisconsin DOT: Richard Lange, Richard.Lange@dot.state.wi.us
Florida DOT: Patrick Brady, Patrick.Brady@dot.state.fl.us
Washington DOT: Brian Limotti, LimottiB@wsdot.wa.gov
Utah DOT: Robert Hull, Director of Traffic and Safety, rhull@utah.gov
Colorado DOT: Jake Kanonov
Nevada DOT: http://www.geoplace.com/uploads/FeatureArticle/0505ta.asp

Appendix C: Notes on Review of SafetyAnalyst

This section provides an overview of the network screening module proposed for the FHWA SafetyAnalyst Program. The review is based primarily on the white paper for Module 1 of SafetyAnalyst – Network Screening, http://www.safetyanalyst.org/docs.htm. It is prepared by Midwest Research Institute, iTRANS Consulting, Inc., Human Factors North, Inc., Ryerson Polytechnic University, and Dr. Ezra Hauer, and it is submitted to FHWA on December 2002

The purpose of the network screening module is to use available data to review the entire roadway network under the jurisdiction of a particular highway agency and identify and prioritize those sites that have promise as sites for potential safety improvements, and therefore merit further investigation. The basic function of the network screening module will be to rank sites by one or more selected measures or indices based on a consideration of each site's accident history, traffic volume, and roadway characteristics. The module will also have other complementary capabilities.

It is expected that the following capabilities will be of most interest to a majority of users. The module will be able to:

1. Rank sites by appropriate measures or indices related to:
   - Potential for safety improvement (PSI) based on expected accident frequency
   - PSI based on excess accident frequency (amount by which the expected accident frequency exceeds that expected at similar sites)
   - Prospective cost-effectiveness based on expected accident frequency, excess accident frequency, or both
   - Overrepresentation of specific accident types (e.g., a higher than expected proportion of rear-end accidents at signalized intersection may indicate the need to adjust the inter-green period, adjust the cycle length, or implement some other accident countermeasure)
2. Provide flexibility and guidance for the user to choose among available measures/methods for ranking sites
3. Provide flexibility for the user to apply default SPFs provided with the software or to apply user-supplied SPFs
4. Rank sites separately, or in combination, by:
   - Type of roadway elements (e.g., roadway segments, intersections, interchange ramps)
   - Area type (rural/urban)
   - Terrain type (level/rolling/mountainous)
   - Geographic areas (entire jurisdiction, or specific regions, counties, cities, etc.)
5. Permit ranking based either on the sum, or weighted sum, of property-damage-only (PDO), nonfatal injury (NFI), and fatal injury (FI) accidents
6. Provide an option for the user to choose whether or not to rank by accident costs and to accommodate either default or user-supplied values for accident costs
7. Provide a geographic distribution of accidents within a roadway segment by accident severity level and identify points of concentration of accidents
8. Screening sites for specific accident types/countermeasures (e.g., run-off-road accidents for shoulder rumble strip or left-turn collisions for turn-lane installation)
9. Screening for sites that show deterioration in safety over time

10. Identification of "corridors with promise" through review of the safety performance of extended roadway sections
11. Screening based on a sliding-window approach for roadway segments

SafetyAnalyst will be built on the concept of conducting screening based on *expected* accident frequencies. Expected accident frequencies can be estimated from safety performance functions (SPFs), which often take the form of negative binomial regression relationships to predict accident frequencies from traffic volumes and roadway characteristics. The Empirical Bayes (EB) method provides a means to combine SPFs predictions and observed accident frequencies into a single estimate of the expected accident frequency, so that the observed accident history of a site can be considered in the estimation process. The EB method used in *SafetyAnalyst* will be adapted from the approach currently being developed by the Colorado DOT. In addition, it is recommended that *SafetyAnalyst* include not only an EB approach to network screening based on the analysis of homogeneous roadway sections, as recently developed for the Colorado DOT, but also a traditional sliding-window approach to network screening for roadway sections that is updated to incorporate EB concepts.

Appendix D: Empirical Bayes Technique

D1. Prior and Posterior Probability

Bayesian statisticians claim that methods of Bayesian inference are a formalization of the scientific method involving collecting evidence which points towards or away from a given hypothesis. There can never be certainty, but as evidence accumulates, the degree of belief in a hypothesis changes; with enough evidence it will often become very high (almost 1) or very low (near 0).

Bayes theorem provides a method for adjusting degrees of belief in the light of new information.

Bayes' theorem is

$$P(H_0|E) = \frac{P(E|H_0)\ P(H_0)}{P(E)} \qquad (1)$$

For our purposes, H_0 can be taken to be a hypothesis which may have been developed *ab- initio* or induced from some preceding set of observations, but before the new observation or evidence E.

The term $P(H_0)$ is called the ***prior probability*** of H_0.

The term $P(E \mid H_0)$ is the *conditional probability* of seeing the observation E given that the hypothesis H_0 is true; as a function of H_0 given E, it is called the ***likelihood function***.

The term $P(E)$ is called the *marginal probability* of E; it is a *normalizing constant* and can be calculated as the sum of all mutually exclusive hypotheses $\sum P(E|H_i)P(H_i)$.

The term $P(H_0 \mid E)$ is called the ***posterior probability*** of H_0 given E.

The theorem may be paraphrased as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalizing constant}}$$

There is also a version of Bayes' theorem for continuous distributions. Bayes' theorem for probability densities is formally similar to the theorem for probabilities:

$$f(x|y) = \frac{f(y|x)\ f(x)}{f(y)}$$

and there is an analogous statement of the law of total probability:

$$f(x|y) = \frac{f(y|x)\, f(x)}{\int_{-\infty}^{\infty} f(y|x)\, f(x)\, dx}$$

D2. Examples

1. From which bowl is the cookie?

To illustrate, suppose there are two bowls full of cookies. Bowl #1 has 10 chocolate chip and 30 plain cookies, while bowl #2 has 20 of each. Our friend Fred picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a plain one. How probable is it that Fred picked it out of bowl #1?

Intuitively, it seems clear that the answer should be more than a half, since there are more plain cookies in bowl #1. The precise answer is given by Bayes' theorem. Let H_1 corresponds to bowl #1, and H_2 to bowl #2. It is given that the bowls are identical from Fred's point of view, thus $P(H_1) = P(H_2)$, and the two must add up to 1, so both are equal to 0.5. The datum D is the observation of a plain cookie. From the contents of the bowls, we know that $P(D \mid H_1) = 30/40 = 0.75$ and $P(D \mid H_2) = 20/40 = 0.5$. Bayes' formula then yields

$$\begin{aligned} P(H_1|D) &= \frac{P(H_1)\cdot P(D|H_1)}{P(H_1)\cdot P(D|H_1) + P(H_2)\cdot P(D|H_2)} \\ &= \frac{0.5\times 0.75}{0.5\times 0.75 + 0.5\times 0.5} \\ &= 0.6. \end{aligned}$$

2. Typical examples that use Bayes' theorem assume the philosophy underlying Bayesian probability that uncertainty and degrees of belief can be measured as probabilities. One such example follows.

We wish to know about the proportion r of voters in a large population who will vote "yes" in a referendum. Let n be the number of voters in a random sample (chosen with replacement, so that we have statistical independence) and let m be the number of voters in that random sample who will vote "yes". Suppose that $n = 10$ voters and only $m = 7$ voted yes, from Bayes' theorem we can calculate the probability distribution function r using

$$f(r|n=10, m=7) = \frac{f(m=7|r, n=10)\, f(r)}{\int_0^1 f(m=7|r, n=10)\, f(r)\, dr}.$$

From this we see that once we have in hand the prior probability density function $f(r)$ and the likelihood function $L(r) = P(m = 7|r, n = 10,)$, we can compute the posterior probability density function $f(r|n = 10, m = 7)$.

The prior summarizes what we know about the distribution of r in the absence of any observation. We will assume in this case that the prior distribution of r is uniform over the interval [0, 1]. That is, $f(r) = 1$. That assumption should be considered provisional -- if some additional background information is found, we should modify the prior accordingly.

Under the assumption of random sampling, choosing voters is just like choosing balls from an urn. The likelihood function for such a problem is just the probability of 7 successes in 10 trials for a binomial distribution.

$$P(m = 7|r, n = 10) = \binom{10}{7} r^7 (1 - r)^3.$$

As with the prior, the likelihood is open to revision -- more complex assumptions will yield more complex likelihood functions. Maintaining the current assumptions, we compute the normalizing factor,

$$\int_0^1 P(m = 7|r, n = 10)\, f(r)\, dr = \int_0^1 \binom{10}{7} r^7 (1 - r)^3\, 1\, dr = \binom{10}{7} \frac{1}{1320}$$

and the posterior distribution for r is then

$$f(r|n = 10, m = 7) = \frac{\binom{10}{7} r^7 (1 - r)^3\, 1}{\binom{10}{7} \frac{1}{1320}} = 1320\, r^7 (1 - r)^3$$

for r between 0 and 1, inclusive.

One may be interested in the probability that more than half the voters will vote "yes". The *prior probability* that more than half the voters will vote "yes" is 1/2, by the symmetry of the uniform distribution. In comparison, the posterior probability that more than half the voters will vote "yes", i.e., the conditional probability given the outcome of the opinion poll -- that seven of the 10 voters questioned will vote "yes" -- is

$$1320 \int_{1/2}^1 r^7 (1 - r)^3\, dr \approx 0.887$$

which is about an "89% chance".

D3. Application of Empirical Bayes Technique

Let's assume that
m is total number of accident locations
n_i is total number of accidents at location i, $i \leq m$
x_i is total number of the particular patterns under investigation, $i \leq m$

1. Establish x_i reference groups with a homogeneous in traffic conditions. The criteria to form the group can be: intersections with left-turn movements, un-signalized intersections, run-off-road, side-swipe, or other specific collision types.

2. For each location i calculate: $p._i = x_i/n_i$ of each accident pattern (x_i) at the location to the total number of accidents (x_i). If the value of $p._i$ is known $p._i = p$, then the probability of occurrence of certain events of x is given by the Binomial Distribution.

$$P\left(x_i = x \mid n_i, \tilde{p}_i = p\right) = \begin{bmatrix} n_i \\ x \end{bmatrix} p^x (1-p)^{n-x} \qquad (0 \leq x \leq n) \tag{6}$$

It is assumed that the prior distribution for *p* across the reference group is a Beta distribution (Maritz and Lwin 1989) given by

$$f_R(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad (0 < p < 1) \tag{7}$$

where α and β are parameters of the prior distribution, determined by fitting observations of all (*xi*, *ni*) pairs in the reference group to the Beta distribution. The mean and variance of the Beta distribution are

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mu(1-\mu)}{\alpha+\beta+1} \tag{8}$$

3. Method of moments is used to estimate the parameters α and β.

$$\bar{p} = \frac{1}{m}\left(\sum_{i=1}^{m}\frac{x_i}{n_i}\right)$$

$$s^2 = \frac{1}{m-1}\left[\sum_{i=1}^{m}\left(\frac{x_i^2 - x_i}{n_i^2 - n_i}\right) - \frac{1}{m}\left(\sum_{i=1}^{m}\frac{x_i}{n_i}\right)^2\right]$$

Then the values of parameters of the Beta *posterior* distribution α_i, and β_i for empirical Bayes method can be calculated:

$$\alpha_i = \frac{\frac{\bar{p}}{s^2} - (\bar{p}+1)^2}{(\bar{p}+1)^3} + x_i$$

$$\beta_i = \frac{\frac{\bar{p}^2}{s^2} - (\bar{p}+1)^2\bar{p}}{(\bar{p}+1)^3} + n_i - x_i$$

where $\bar{p}$ and s^2 are sample mean and sample variance that are derived using the method of moments. The considered location has overrepresentation of the left-turn pattern if the probability that left-turn ratio pi exceeds the reference group average p is significant:

$$P(p_i - \bar{p}) < 1 - 0.95$$

or when the following inequality is true:

$$\left[1 - \int_0^{\bar{p}} \frac{\Gamma(\alpha i + \beta i)}{\Gamma(\alpha i)\Gamma(\beta i)} p^{\alpha i - 1}(1-p)^{\beta i - 1}\,dp\right] > 0.95$$

where $\Gamma(\alpha_i)$ is the gamma function.

D4. Reference

T. Sayed, F. Navin, and W. Abdelwahab, "A countermeasure-based approach for identifying and treating accident prone locations," 1997 NRC Canada.
http://article.pubs.nrc-cnrc.gc.ca/ppv/RPViewDoc?_handler_=HandleInitialGet&journal=cjce&volume=24&calyLang=eng&articleFile=l97-015.pdf

Appendix E: Summary of Crash Prediction Models – Safety Performance Functions (SPF)

The purpose of developing crash prediction models is to enable us to provide a realistic estimate of expected accident frequency as a function of traffic volume and roadway geometries over a highway segment. Development of such estimates is a critical component in the consideration of safety in highway planning and design.

E1. Software

Development of the crash prediction models involved determining which explanatory variables should be used, whether and how variables should be grouped, and how variables should enter into the model, that is, the best model form. McGee et al from the National Cooperative Highway Research Program (NCHRP) Report 491 used generalized linear modeling (GLM) to estimate model coefficients using the software package GENSTAT and assuming a negative binomial error distribution, all consistent with the state of research in developing these models (*1*). In specifying a negative binomial error structure, a parameter, *K*, that relates the mean and variance of the regression estimate is iteratively estimated from the model and the data. The value of *K*, which is the inverse of the overdispersion parameter of the negative binomial distribution, is such that the larger the value of *K*, the smaller the variance of the model estimate and therefore the better the model.

Another possible software package to develop crash prediction models is R, which was used in developing the crash prediction model for the study of pedestrian safety in number in the city of Oakland (*2*). However, R cannot handle large data sets. Software package SAS is another powerful software that can handle large data sets. The question remains is do we want to include all observations in our model to obtain a crash prediction model, or do we want to come up with a different crash prediction model for each different type of road segment? SAS and GENSTAT can handle either direction we choose. R might be able to handle the latter.

(Note: in R, use the negative binomial model for predicting crash frequency on highway segments rather than using the quasipoisson model, since a quasi- mode does not have a likelihood and so does not have an Akaike Information Criterion (AIC), by definition. Having AIC statistics give us another criteria to choose the best model. Also, we can use the step-wise selection function only if we have AIC statistics. The reason for using a quasipoisson model to model collision frequency in the Oakland intersections study was because the distribution of the number of collisions followed the Poisson distribution fairly closely; the dispersion parameter was estimated to be 1.16, where 1 indicates that there is no over-dispersion. The dispersion parameters in the crash prediction models in the highway study might not, however, be close to 1.)

E2. Possible Models

Using existing databases of accidents, traffic volumes, and roadway geometries, examples of common functional forms for the crash prediction models or Safety Performance Functions (SPFs) are:

$$Accidents/year = (segment\_length)\alpha(x_1)\beta_1(x_2)\beta_2...(x_n)\beta_n \quad \textbf{(1)}$$

$$Accidents/year = (segment\_length)\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n) \quad \textbf{(2)}$$

where:

$x_1,\ldots,x_n$ = traffic and geometric variables such as average annual daily traffic (AADT) and lane width

α and $\beta_1,\ldots,\beta_n$ = coefficients estimated in the model calibration procedure

(Note: model (2) above was used in the Oakland intersections study.)

It is possible that in some cases the influence of a variable will be represented by a few regression parameters, not a function. For example, variations of the above equations may have several parameters α_{year}, one for every year used in the screening. Segment length also may have to be represented by a more complex function, and not just as a multiplier. Thus, the process of developing SPFs may be quite complex. Because of this complexity, verification is essential. One verification level is to show the function for each variable introduced as a graph, first before the parameters are entered in some generic form and then after the parameters are entered in their true form. Another verification level would be to make validation an integral part of the process of entering the SPF. The premise is that if the user has an SPF, it must be based on the user's data, the data that also serve for screening. Therefore, the user's SPF must fit the data and have a ratio of observed to predicted accidents of close to 1 (*3*).

In addition to the functional form and the coefficients of the regression equation, the user will also have to specify the value of *k,* an overdispersion parameter estimated during model calibration. The regression coefficients and the overdispersion parameter are essential in the network screening calculations.

E3. Example #1

An example in (*1*) of a model for all injury accidents at 4-legged stop controlled intersections is:

Accidents/year = α (Major Road AADT)^ b (Minor Road AADT)^c

where α, b, and c have calibrated values of 0.000426 (that is, e-7.76), 0.499 and 0.430 respectively. These estimates were obtained by using a generalized linear model (GLM) in the GENSTAT package, assuming a negative binomial error distribution.

The recalibration procedure for the model for each jurisdiction and for each year of the analysis period is in (Harwood). To apply this procedure requires yearly accident counts and AADTs for a sample of 4-legged stop controlled intersections in the jurisdiction that are typical of those that tend to be considered for signal installation. The default base model is first used to estimate

accidents each year for each intersection in the sample. For each year, the sum of the observed counts divided by the sum of the model estimates gives a calibration factor that is applied as a multiplier to the model to obtain a recalibrated value of α.

Table 8: Summary of Data and Example Calculations for ALL Injury Crashes

1) Year (y)	**1996**	**1997**	**1998**	**1999**	**Jan-Aug 2000**	**1999 (Signal)**
2) Crashes in year (X)	4	6	3	6	4	
	Sum = $\mathbf{X_b}$ = 23					
3) MAJAADT	41309	42169	43460	43891	44321	48441
4) MINAADT	3596	3671	3783	3821	3858	4295
5) Recalibrated $\alpha \times 10^{-4}$	4.26	4.40	4.01	4.20	4.36	4.30
6) Parameter K	2.30	2.30	2.30	2.30	2.30	3.1
7) Model Prediction $E\{\kappa_y\}$	2.897	3.049	2.858	3.021	2.110	3.337
8) $C_{i,y} = E\{\kappa_y\}/E\{\kappa_{99}\}$	0.959	1.009	0.946	1	0.698	1.105
9) Comp. Ratio for period	Sum = C_b = 4.613					C_a=1.105
10) Expected annual crashes without signalization (and variance) [based on last full year (1999)]	$\kappa(99) = C_a(k+X_b)/\{[K/E\{\kappa_{99}\}]+C_b\}$ = 1.105(2.30 + 23)/{(2.30/3.021) + 4.613) = 4.679 $Var\{\kappa(99)\} = C_a(K+X_b)/[\{K/E\{\kappa_{99}\}+C_b\}^2]$ = 0.865					
11) Expected annual crashes after signalization (and variance)[From model in Table 4 (based on 1999)]	$E\{\kappa_{99}\}_{signal} = \exp(-5.751)(48441)^{0.4911}(4295)^{0.1975}$ = 3.318 $Var\{\kappa_{99}\}_{signal} = E\{\kappa_{99}\}^2/K = 3.318^2/3.1$ = 3.551					

Step 1:
Assemble data and accident prediction models. The counts of all injury accidents in each year of the analysis period are shown in the second row of Table 8. Entering volumes for the major and minor roads are estimated for each year using suitable methods applied locally and are entered in the 3rd and 4th rows of Table 8.

Step 2:
a) Estimate the expected number of accidents each year using the recalibrated prediction model. For example, for 1996,

E{K1996}all = 0.000426(41302)^0.499(3596)^0.430 = 2.897

These estimates are shown in row 7 of Table 8. Note that for the last year, an estimate is also done for the anticipated volumes if the intersection were to be signalized (still using the stop controlled model).

b) Calculate the comparison ratio (*Ci,y*) of the model estimate for a given year divided by the model estimate for 1999. These ratios are shown in row 8 of Table 8 and summed in row 9.

c) Using the values in the previous rows and the formula shown in the Table 8 estimate the expected annual number of accidents without signalization (and its variance) for the last full year (1999). The values, shown in row 10 of Table 8, are

K(99)all = 4.679; Var{ K(99)all} = 0.865

E4. Example #2

A similar procedure for recalibrating default SPFs (provided by SafetyAnalyst) in (*3*) to a particular situation is as following. Consider an SPF for total accidents being recalibrated for urban 4-legged intersections. Suppose the default SPF for total accidents is:

Total accidents/year = 0.00005(major road AADT)^0.750(minor road AADT)^0.350

Assume the data to be screened consist of 200 urban 4-legged intersections with the following accident history (i.e., observed accident frequencies):

Year 1: 150 total accidents
Year 2: 130 total accidents
Year 3: 165 total accidents

Step 1:
Apply the SPF from the equation in this example to estimate the number of accidents, separately for years 1 to 3, at each of the 200 intersections. Use the AADTs for the respective year.

Step 2:
For each year, calculate a yearly calibration factor, Ci, by dividing the sum over all sites of the observed number of accidents in that year by the sum of the predicted number of accidents in that year:

$$C_i = \frac{\sum observed\_accidents_i}{\sum predicted\_accidents_i}$$

In this case, suppose the sums for all sites of the yearly predictions were:

Year 1: 134.50 total accidents
Year 2: 140.75 total accidents
Year 3: 150.55 total accidents

It follows that:

C1 = 150/134.50 = 1.12
C2 = 130/140.75 = 0.92
C3 = 165/150.55 = 1.10

Step 3:
Add the calibration factors to the SPF as a multiplier for each year. The recalibrated SPF is then:

Accidents/year = (Ci)(0.00005)(major road AADT)^0.750(minor road AADT)^0.350

that is, three SPFs in fact were developed in this particular case.

Step 4*:*
Using the recalibrated SPF for each year in step 3, estimate the predicted number of accidents, P, for each site and each year. Steps 1 through 4 are summarized in Table E-1.

Table E-1 Example Calculation of Yearly Calibraiton Factors and Final Predicted Accident Frequencies

Intersection No.	Observed accident frequency			Predicted accident frequency using default SPF (Eq. 3)			Predicted accident frequency using recalibrated SPF (Eq. 5)		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
1	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$				$P_{1,1}$	$P_{1,2}$	$P_{1,3}$
2	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$				$P_{2,1}$	$P_{2,2}$	$P_{2,3}$
3									
.									
.									
i									
.									
.									
200	$O_{200,1}$	$O_{200,2}$	$O_{200,3}$				$P_{200,1}$	$P_{200,2}$	$P_{200,3}$
Total	150	130	165	134.50	140.75	150.55			

Step 5*:*
Recalibrate the overdispersion parameter, *k. (It is possible that, as a result of the research, this step may not be necessary. It is documented here to provide a feel for what is involved should it be necessary.)*

1. For each site, calculate the total number of observed accidents, *0*, across all three years.

2. Similarly, for each site, calculate the total number of predicted accidents, *P*, across all three years. Also compute *P2* for each site.

3. For each site, determine the value of the squared residual (*SR*):

$$SR = (P - O)^2$$

4. Subtract the value of P from the squared residual (SR). This gives an estimate of *P^2/k*:

$$[\text{Estimate of } P^2/k\ ] = SR - P$$

5. Fit a straight line to the data with *P^2/k* as the dependent variable and *P^2* as the independent variable, forcing the line through the origin. Thus, in this example, a straight line, forced through the origin, will be fit to 200 pairs of [*P^2*, *(SR-P)*] data points. An ordinary least squared regression procedure such as that provided by Excel should suffice.

6. The inverse of the slope of the fitted regression line is an estimate of *k*. Table E-2 summarizes Step 5 calculations. The highlighted fourth and sixth columns show the data

used for estimating the slope of the straight line to obtain an estimate of the overdispersion parameter.

Table E-2 Example Calculation for Recalibrated Overdispersion Parameter

Intersection No.	Accident frequencies—total over 3 years			Squared residual $(P-O)^2$	Estimate of P^2/k
	Observed	Predicted	Predicted^2		
1	O_1	P_1	P_1^2	SR_1	$SR_1 - P_1$
2	O_2	P_2	P_2^2	SR_2	$SR_2 - P_2$
3					
.					
.					
i					
.					
.					
200	O_{200}	P_{200}	P_{200}^2	SR_{200}	$SR_{200} - P_{200}$

E5. Major Issues

Regression-to-the-Mean and the Problem with Using the Accident Count and Rate

Accident count method is the simplest of techniques, but it suffers from the regression-to-the-mean bias in which an unusually high count is likely to decrease even if no improvements were implemented. Therefore, a site with such counts may not be in need of improvement. On the other hand, a truly hazardous site may have a randomly low observed count and incorrectly escape detection as a result.

Accident rates are calculated as accident rate = accident frequency/AADT. If accident rates are based on the observed counts, then the regression-to-the-mean difficulty will still apply. The non-linearity relationship between accident frequency and AADT causes another problem in using accident rates. Often, when the slope of the accidents/AADT relationship is decreasing with increasing traffic volume levels, screening by accident rates will tend to identify low AADT sites for further investigation. The most valid basis of comparison using accident rates is for the cases when the traffic volumes are the same or when the relationship between the accidents and AADT is linear.

In the SafetyAnalyst report, it's recommended that we do not screen for sites based on observed accident frequencies and/or rates. As an alternative, it is recommended to conduct screening based on expected accident frequencies, which will be estimated as a weighted average of the observed accident frequency and the accident frequency predicted with an SPF.

In theory, this method helps minimize the regression-to-the-mean problem and gives a more realistic estimate of accident frequency. The difficulty is in determining the best crash prediction model or SPF. Another problem is to come up with an appropriate weight.

Adequacy of Data

It's not known how much data is enough. The more historical data we have, the better our analysis. The longest analysis period we have seen used 5 years of historical data. A preliminary list of desired data items is provided in Appendix B of the SafetyAnalyst report. These data items include, at a minimum, elements of traffic accident data, roadway segment, intersection, and interchange ramp data, cost data, and SPF data (*3*).

Development of Models

The main difficulties in developing SPFs are that we do not know which explanatory variables are relevant to be included in the model, and which functional form they have. Stepwise selection can solve the problem in choosing significant predictors. The best way to find out the relationship between predictors and the dependent variable is by exploring the data through graphs. Crash prediction Models might differ for each jurisdiction and data set, and no one model might serve all road type, ramp, or intersection. Therefore, the task of developing SPFs could be very time consuming and requires careful assessments.

E6. Reference

1. McGee, H., Taori, S. and Persaud, B. *Crash Experience Warrant for Traffic Signals.* NCHRP Report 491. Transportation research Board. 2003.

2. Geyer, J., Raford, N., Ragland, D. and Pham, T. *The Continuing Debate about Safety in Numbers—Data From Oakland, CA*. (Final paper in T drive.)

3. SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites. Task K. White Paper for Module 1—Network Screening. December 2002.

4. Harwood, D.W., Council, F.M., Hauer, E., et al. *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. FHWA-RD-99-207, U.S. Department of Transportation, December 2000.