Secondary Accident Data Fusion for Assessing Long-Term Performance of Transportation Systems

Carlos Sun
Department of Civil and Environmental Engineering
University of Missouri-Columbia
E2509 Lafferre Hall
Columbia, MO 65211
csun@missouri.edu

Venkat Chilukuri
University of Missouri-Columbia
Department of Civil and Environmental Engineering
E2509 Lafferre Hall
Columbia, MO 65211
vbc271@mizzou.edu

ABSTRACT

The research that is reported describes two critical steps in developing a methodology for extracting secondary accidents from police accident databases. The eventual goal of the research is to produce an incident progression model from which the number of secondary incidents could be determined. These steps include the following: (1) the development of models for incomplete incident data and (2) an analysis of site differences in maximum queue lengths and incident duration. The first step is important because it allows the use of intranet incident data that is complementary to police accident data. The second step is also important because it helps determine whether site-specific models are necessary instead of a single general model. The methodology involves the use of data fusion from accident data and intranet traffic reports. The project results in a near-term technology for analyzing the safety impacts of transportation assets.

Key words: accident data—data fusion—traffic safety
RESEARCH OBJECTIVES AND BACKGROUND

The main objective of this research is to develop a readily deployable methodology for extracting secondary accidents from an accident database. This methodology should be robust so that it can be used for analyzing the safety benefits of different transportation systems. In order to accomplish this main objective, several other objectives need to be achieved. First, a convention for formatting accident reports needs to be developed so that the appropriate information from the accident database is extracted and formatted for the derivation of secondary accidents. Second, a methodology for processing intranet traffic reports need to be developed. This methodology transforms the readily available traffic reports into useful information that tracks the progression of individual incidents. Third, incomplete incident reports need to be modeled using an appropriate methodology. Fourth, site-specific data is analyzed to determine whether different models are needed from different sites. Finally, the main objective of extracting secondary accidents is achieved through the use of a master incident progression curve.

This methodology has many potential applications. Long-term safety evaluation is critical in the decision making process for maintaining and operating equipment in incident management activities. For example, there have been legislators in Missouri who question the value of such programs and desire for such assets to be maintained by the private sector. The Motorist Assist (MA) program or Freeway Service Patrol, for example, consist of assets such as the maintenance garage, fuel station, field-deployable equipment such as variable message sign (VMS) trailers, MA and emergency management vehicles (specialized trucks), records management computing systems, communications infrastructure, and personnel. Since the inception of MA in Missouri in 1993, the feedback regarding MA has been mostly informal, in the appearance of public comment forms returned from motorists serviced by the program. There is a need to assess the value of this program in a systematic and logical fashion. It is often difficult to make statistically significant conclusions concerning the short-term using accident data because of small sample sizes. But accident data is useful for long-term asset management and planning because they capture long-term trends. The use of secondary accident data for evaluating the MA program provides concrete evidence. Since safety is arguably one of the most important transportation issues, the impact of the MA incident management approach on safety is a primary concern for all asset management systems.

Secondary accidents are accidents that result from an existing primary incident. Many times these accidents occur at the end of queues that had developed from the primary incident. Quickly opening the highway after an incident reduces the potential for secondary accidents. It is easy then to see the value of analyzing secondary accidents when considering traffic incident management strategies such as intelligent transportation systems and the MA program. On the other hand, the effects of such systems on primary accidents would be much smaller because many of these accidents are caused by driver error, such as fatigue, intoxication, or aggressive driving. Therefore, traditional analysis of primary accidents and accident rates will not reveal the full potential of such systems.

The development of this methodology is based on previous work in the area of accident analysis and safety. An important paper on the analysis of secondary accidents is Raub (1997), in which the author presents a methodology for the temporal and spatial analysis of incidents on urban freeways in order to identify the secondary crashes. He found that more than 15% of the crashes reported by police may be secondary in nature. He also found that such crashes result from external distractions instead of internal distractions or driver perception error. For the analysis, Raub assumed an accident effect duration of 15 minutes plus the clearance time. He also assumed a distance of effect of less than 1,600 meters (1 mile). In other words, if an accident occurred within these temporal and spatial boundaries, the accident is considered to be secondary.
More recently, Moore et al. (2004) examined secondary accident rates on Los Angeles freeways using accident records from the California Highway Patrol’s First Incident Response Service Tracking system, as well as data from loop detectors on Los Angeles freeways. They defined secondary accidents as accidents occurring upstream of the initial incident in either direction within or at the boundary of the queue formed by the initial incident. A fixed threshold of 3,218 m (2 miles) and 2 hours was used for forming this boundary. Several levels of filters served to eliminate erroneous data.

These two studies exemplify the use of fixed thresholds for extracting secondary accidents. Figure 1 graphs the progression of an actual incident, including fixed queue length and time thresholds superimposed on this progression. If an accident falls within the influence of the primary accident, i.e., the accident happened within the queue of the primary accident, then the accident is considered to be secondary. Progression refers to the growth and decline of the queue length as the incident progresses through the various stages. In general, the various stages of an incident that influence the traffic queues include the onset, the arrival of response teams, the clearance to the shoulder, the completion of clearance, and the normalization of traffic. The progression is a function of both the demand (traffic) and the supply (road capacity). With the demand changing constantly, it is clear that the assumption of fixed thresholds would not capture field conditions properly. Some would argue that, on average, the total number of secondary accidents can still be estimated accurately with fixed thresholds if the area of the fixed threshold rectangle is the same as the area under the progression curve. This argument also requires the assumption that accidents are independent from the location and time of the primary accident. For example, Figure 1 shows that the same number of accidents (three) is classified as secondary using a fixed threshold or the actual incident progression. However, by definition, secondary accidents differ in cause from primary accidents. Therefore, even if the average number of accidents is captured accurately with fixed thresholds, the accidents themselves are still misclassified. Referring back to the example and looking at the fixed thresholds, the number of secondary accidents is estimated correctly, even though accident B is a false positive and accident E is a false negative. The inadequacy in the use of fixed thresholds is one primary motivation for this research.

![Figure 1. Fixed thresholds versus actual incident progression](image-url)
Other articles relate to secondary accidents but do not address the extraction process directly. Karlaftis et al. (1999) examined the primary crash characteristics that influence the likelihood of secondary crash occurrence. They suggested that clearance time, season, type of vehicle involved, and lateral location of the primary crash are the most significant factors. The economic benefit of secondary crash reduction for the Hoosier Helper freeway service patrol program was also presented. There are several articles that address the magnitude and impact of incident delays. These include Garib et al. (1997), Giuliano (1989), Skabardonis et al. (1996), Morales (1987), Sullivan (1997), Smith et al. (2003), Lindley (1987), and Lee et al. (2003). Many of these articles try to estimate the impact of accidents. The present research, on the other hand, proposes the direct derivation of the impacts (queue lengths over time) from traffic reports.

RESEARCH METHODOLOGY

The primary source of data used in the methodology is the accident database obtained from the highway patrol in Missouri. The accident database for freeways is compiled by police agencies and consolidated by the state highway patrol. Table 1 shows an accident record snippet from I-70 in Missouri in 2002. The electronic accident records contain much more information than what is shown in Table 1. In addition, images of the four-page accident reports are also kept in the accident database. The primary fields that will be used in this analysis include name, direction, continuous log, date, severity, time, image number, and traffic condition. This is the consistent format used for accident analysis.

Table 1. Sample accident record of I-70 in Missouri

<table>
<thead>
<tr>
<th>NAME</th>
<th>DIR</th>
<th>CONT LOG</th>
<th>DATE</th>
<th>ACCIDENT TYPE</th>
<th>ACCIDENT CLASS</th>
<th>SEVERITY</th>
<th>TIME</th>
<th>VEHICLE NO</th>
<th>VEHICLE TYPE</th>
<th>CIRCUMSTANCE</th>
<th>TRAFFIC COND</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>E</td>
<td>223.409</td>
<td>1/1/02</td>
<td>MOTOR VEHICLE IN TRAFFIC</td>
<td>REAR END</td>
<td>PROPERTY DAMAGE ONLY</td>
<td>350</td>
<td>1</td>
<td>PASSENGER CAR</td>
<td>DRINKING</td>
<td>NORMAL</td>
</tr>
<tr>
<td>70</td>
<td>E</td>
<td>223.409</td>
<td>1/1/02</td>
<td>MOTOR VEHICLE IN TRAFFIC</td>
<td>REAR END</td>
<td>PROPERTY DAMAGE ONLY</td>
<td>350</td>
<td>1</td>
<td>PASSENGER CAR</td>
<td>INATTENTION</td>
<td>NORMAL</td>
</tr>
<tr>
<td>70</td>
<td>E</td>
<td>223.409</td>
<td>1/1/02</td>
<td>MOTOR VEHICLE IN TRAFFIC</td>
<td>REAR END</td>
<td>PROPERTY DAMAGE ONLY</td>
<td>350</td>
<td>2</td>
<td>VAN</td>
<td>NOT STATED OR UNKNOWN</td>
<td>NORMAL</td>
</tr>
<tr>
<td>70</td>
<td>E</td>
<td>250.312</td>
<td>1/1/02</td>
<td>MOTOR VEHICLE IN TRAFFIC</td>
<td>REAR END</td>
<td>PROPERTY DAMAGE ONLY</td>
<td>1110</td>
<td>1</td>
<td>PASSENGER CAR</td>
<td>IMPROPER PASSING</td>
<td>NORMAL</td>
</tr>
<tr>
<td>70</td>
<td>E</td>
<td>250.312</td>
<td>1/1/02</td>
<td>MOTOR VEHICLE IN TRAFFIC</td>
<td>REAR END</td>
<td>PROPERTY DAMAGE ONLY</td>
<td>1110</td>
<td>2</td>
<td>PASSENGER CAR</td>
<td>NOT STATED OR UNKNOWN</td>
<td>NORMAL</td>
</tr>
<tr>
<td>70</td>
<td>W</td>
<td>15.57</td>
<td>1/2/02</td>
<td>MOTOR VEHICLE IN TRAFFIC</td>
<td>REAR END</td>
<td>INJURY</td>
<td>815</td>
<td>1</td>
<td>PICKUP</td>
<td>FOLLOWING TOO CLOSELY</td>
<td>NORMAL</td>
</tr>
<tr>
<td>70</td>
<td>W</td>
<td>15.57</td>
<td>1/2/02</td>
<td>MOTOR VEHICLE IN TRAFFIC</td>
<td>REAR END</td>
<td>INJURY</td>
<td>815</td>
<td>2</td>
<td>PASSENGER CAR</td>
<td>NOT STATED OR UNKNOWN</td>
<td>CONGESTION AHEAD</td>
</tr>
<tr>
<td>70</td>
<td>E</td>
<td>234.229</td>
<td>1/2/02</td>
<td>OUT OF CONTROL</td>
<td>PROPERTY DAMAGE ONLY</td>
<td>847</td>
<td>1</td>
<td>PASSENGER CAR</td>
<td>NOT STATED OR UNKNOWN</td>
<td>NORMAL</td>
<td></td>
</tr>
</tbody>
</table>
In order to use secondary accidents as a performance measure, it is necessary to separate such accidents from the rest of the accidents. The police accident report contains a field that describes downstream conditions as “accident ahead” or “congestion ahead.” The difficulty with the police determining whether the accident is secondary is that they are limited spatially (at one location) and temporally (responding to the current accident). Since the effect of primary accidents can persist long after they have been cleared, it is difficult to determine at the scene of an accident whether it is due to recurrent or non-recurrent congestion. The use of the category “accident ahead” for finding secondary accidents would undercount the number of accidents while adding the category “congestion ahead” would severely overcount the number of accidents. Because of such problems, it is necessary to produce a methodology that can reliably distinguish the secondary accidents.

Traffic management centers and traffic news agencies can provide wide spatial coverage of incidents as well as track the incidents over time. They can use information from aircrafts, elevated traffic cameras, MA, emergency management (fire, ambulance, HAZMAT), and motorist calls. They can also monitor and update this information throughout the course of an incident. Such intranet traffic information can be independent from police information; therefore, such information can complement the accident database from the police. Data fusion helps incorporate all available information sources, including intranet traffic reports and the accident database. By analyzing individual traffic reports in detail, the reporting times of the incident and the dynamic locations of the back of the queue can be found. The difference between the initial and final times estimates the total incident duration, and the distance from the location of the incident to the back of the queue estimates the length of the roadway affected by the incident.

However, intranet reports need to be processed significantly to be in a usable format. The methodology for processing such reports is as follows. Pages of traffic reports are downloaded daily at regular, e.g., three-minute, intervals. A Unix script has been written to perform this automatically. These reports are then consolidated and parsed to extract pieces of information into specific fields, such as incident reporting time, incident type, and incident description. A major task is to extract the traffic information for a particular highway along a particular direction in the sequence they are reported in the files on a particular day. A computer program saves the information pertaining to a single incident through multiple reports in a single day. Since there is no unique identifier associated with the information pertaining to a particular incident, the lines containing information related to a particular incident need to be extracted through the use of keywords present in those lines and absent in other lines. There can be difficulties in this process, since traffic reports are human-generated and can include syntax variability as well as errors. As an example, consider the primary route eastbound interstate 70. In the reports, eastbound can also be expressed as “EB,” “E/B,” or “east,” and interstate 70 as “70,” “I70,” or “I-70.” There can also be descriptions of the route expressed in phrases such as “eastbound lanes of 70” or “east and westbound lanes of 70.” Figure 2 shows the result of this processing. The information pertaining to a single incident is tracked throughout the incident, giving the queue length as the incident progresses. Even though the processing of intranet traffic reports is laborious, a valuable incident dataset is produced.
Modeling Incomplete Incident Data

Many of the analyzed traffic reports are incomplete because the time of normalization is missing even though the clearance time is reported. For this reason, it is necessary to derive a model to estimate the incomplete incident progression data. Since detailed incident data are rare and valuable, it would be desirable to continue using the incomplete traffic reports. One way of accomplishing this is to extrapolate or model the incomplete incident data. The incomplete data usually contain a significant number of samples after the clearance of the incident, so the modeling of the incomplete portion of the data should not produce significant errors.

Intranet traffic reports from I-70 and I-270 in St. Louis, Missouri are used in this investigation. A total of 49 accidents have complete information. In other words, the accident is tracked from the beginning until traffic is normalized. Ninety-four accidents have some traffic information after the accident clearance, but not all of the information needed to verify traffic normalization. The 49 complete accidents are then used as a ground truth for testing the modeling process of incomplete incident data.

There are three assumptions that need to be made when using regression to estimate models of the incomplete traffic reports. The three assumptions are that (1) the model error has a mean of zero, (2) the error has a constant variance over all observations, and (3) the error corresponding to different points in time are not correlated (Ostrom 1978). If autocorrelation is present, the variances would be underestimated, which would result in overconfidence in the model.

A test dataset was constructed by taking the complete accidents and then artificially eliminating data samples after the clearance time. This test set tried to replicate the incomplete data and provided the ground truth. The performance of the models in estimating the missing portion of the progression curve was evaluated using this test dataset. Since the backup queue may not decrease linearly, second-, third-, and fourth-order polynomials were used for modeling the incomplete accidents. By using the test dataset mentioned, it was found that a third-order polynomial provided the best fit when compared to the second- and fourth-order polynomials. The third-order polynomial was able to reproduce the total delay estimates (or areas under the queue length/time curves) to within ±10% with an average difference of 1.4 % from the true value. While the average difference between second-order and real data was 5.3%, the difference between the fourth-degree model and real data was 6.5%.

Another criterion for evaluating the performance of the polynomial models was the $R^2$ value, which measures the proportion of the data that can be explained by the model. Figure 3 shows the $R^2$ values of several accidents being modeled by a second-, third-, and fourth-order polynomial. Figure 3 shows that
the third-order polynomial provided the best fit for the accident data, since the third order results in the best $R^2$ over the entire test dataset.

![R-Square values of 2nd, 3rd and 4th order polynomials](image)

**Figure 3. Polynomial model fit**

**Analyzing Site Differences**

In analyzing accident data from different sites, it was helpful to determine whether there were any differences in the accident characteristics. Specifically, two techniques were used to examine accidents from I-70 as compared to I-270. These accidents were examined both spatially and temporally by looking at the maximum queue length and the duration of the accident. First, these accidents were plotted and examined visually to see if I-70 and I-270 accidents patterns were different. Even though no clean-cut boundary could divide accidents from these two freeways, Figure 4 shows that I-270 accidents tend to have longer queue lengths than I-70 accidents. Second, a statistical test was applied to confirm the visual observation.

A student-t test was applied to see if there were any differences in the means between the two freeway sites. Two variables were examined: maximum queue length and accident duration. The test showed that there was a statistically significant difference between the two freeway sites when the variable maximum queue length was considered ($p=0.041$). However, the test also showed that there was no statistically significant difference when the variable accident duration was considered ($p=0.95$). Therefore, the differences in the maximum queue length will need to be considered when the incident progression models are developed for I-70 and I-270.
Secondary Accident Extraction Process

The process to extract secondary accidents from the accident database is as follows. First, the accident database is sorted by route and by year (e.g., I-70, 2002). Each of the 28 fields that describe each accident record is parsed and stored. The time and date fields are translated for computation so that they can be added and subtracted. The entire accident file is then converted into a doubly-linked list so that the file can be collapsed into one record per accident. In other words, a file with one record per vehicle involved in an accident is consolidated into one record per accident. The secondary accidents are extracted by using the queue-time (incident progression) curves, as shown in Figure 5.

The total number of secondary accidents are tallied and recorded for each route and year. A routine can be written in the MATLAB programming language (a C-like programming language for engineering) for this purpose. Figure 5 gives an overview of the secondary accident extraction process.

Figure 4. Visual inspection of I-70 vs. I-270 accident data

Figure 5. Pseudo-flowchart of the process of extracting secondary accidents
CONCLUSIONS

This paper described the progress of research that would develop a methodology for determining the secondary accidents from a police accident database. The eventual goal is accomplished by data fusion of the police database with intranet traffic reports. An important data processing methodology and two key findings were described. First, the processing of intranet traffic reports was described. Since these reports were human-generated, they presented significant challenges to data processing. A valuable dataset of incident progression data was produced. Second, a third-order polynomial was found to be the best for modeling incident progression curves, as it resulted in the smallest average error and the best model fit. Third, site-specific differences were investigated and the maximum queue length was found to be statistically different between two freeway sites.

The research described seeks to improve asset management decision making by developing a methodology for extracting secondary accidents from police accident databases. The methodology involves the use of data fusion accident data and intranet traffic reports. There is great potential for the immediate technology transfer and the implementation of the results of this research in Missouri and the Midwest. This implementation is in the form of a standard method for extracting secondary accidents from the primary accident database maintained by the police. This standardization would guarantee that the evaluation of all asset management systems will consider secondary accident data in a consistent manner.
ACKNOWLEDGMENTS

The authors would like to recognize the support of the Midwest Transportation Consortium for this research project.

REFERENCES