OTREC

FINAL REPORT

OREGON TRANSPORTATION RESEARCH AND EDUCATION CONSORTIUM

Empirical Observation of the Impact of Traffic Oscillations on Freeway Safety

OTREC-RR-10-13
September 2010

A National University Transportation Center sponsored by the U.S. Department of Transportation's Research and Innovative Technology Administration

EMPIRICAL OBSERVATION OF THE IMPACT OF TRAFFIC OSCILLATIONS ON FREEWAY SAFETY

Final Report

OTREC-RR-10-13

by
Christopher M. Monsere (Principal Investigator)
Civil and Environmental Engineering
Portland State University

Soyoung Ahn, Ph.D. (Co-Principal Investigator)
Zuduo Zheng
Civil, Environmental and Sustainable Engineering
Arizona State University

for

Oregon Transportation Research
and Education Consortium (OTREC)

P.O. Box 751
Portland, OR 97207

OTREC
OREGON TRANSPORTATION RESEARCH AND EDUCATION CONSORTIUM

September 2010

Technical Report Documentation Page

1. Report No. OTREC-RR-10-13	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Empirical Observation of the Impact of Traffic Oscillations on Freeway Safety		5. Report Date September 2010
		6. Performing Organization Code
7. Author(s) Christopher M. Monsere; Soyoung Ahn; Zuduo Zheng		8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Civil and Environmental Engineering, Portland State University Civil, Environmental and Sustainable Engineering, Arizona State University		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 08-108
12. Sponsoring Agency Name and Address Oregon Transportation Research and Education Consortium (OTREC) P.O. Box 751 Portland, Oregon 97207		13. Type of Report and Period Covered Final Report
		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Traffic oscillations are typical features of congested traffic flow that are characterized by recurring decelerations followed by accelerations (stop-and-go driving). The negative environmental impacts of these oscillations are widely accepted, but their impact on traffic safety has been debated. This report describes the impact of freeway traffic oscillations on traffic safety. This study employs a matched case-control design using high resolution traffic and crash data from a freeway segment. Traffic conditions prior to each crash were taken as cases, while traffic conditions during the same periods on days without crashes were taken as controls. These were also matched by presence of congestion, geometry and weather. A total of 82 cases and about 80,000 candidate controls were extracted from more than three years of data from 2004 to 2007. Conditional logistic regression models were developed based on the case-control samples. To verify consistency in the results, 20 different sets of controls were randomly extracted from the candidate pool. The results reveal that the standard deviation of speed (thus, oscillations) is a significant variable, with an average odds ratio of about 1.08. This implies that the odds of a (rear-end) crash occurring increases by about 8 percent with an additional unit increase in the standard deviation of speed. The average traffic states prior to crashes were less significant than the speed variations in congestion.

17. Key Words Traffic oscillations, Stop-and-go driving, Crash, Matched case-control design, Conditional logistic regression		18. Distribution Statement No restrictions. Copies available from OTREC: www.otrec.us	
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 44	22. Price

ACKNOWLEDGEMENTS

This project was funded by the Oregon Transportation Research and Education Consortium (OTREC) and supported by matching funds from Arizona State University and Portland State University. The authors would like to thank Dr. Kristin Tufte, Sandeep Puppala and Chengyu Dai for their help in retrieving data from Portland Oregon Regional Transportation Archive Listing (PORTAL). The National Science Foundation supported the development of PORTAL. Finally, we acknowledge the anonymous reviewers of this report and manuscripts derived from this work for their helpful and instructive comments.

DISCLAIMER

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Traffic oscillations are typical features of congested traffic flow that are characterized by recurring decelerations followed by accelerations (stop-and-go driving). The negative environmental impacts of these oscillations are widely accepted, but their impact on traffic safety has been debated. The primary objective of this study is to examine the impact of freeway traffic oscillations on traffic safety.

The safety implications of oscillations have been studied largely by evaluating the significance of speed deviations on the likelihood of crash occurrence, though most existing studies did not make an explicit distinction between congested and uncongested traffic. Existing studies report contradictory findings and debates continue today, as existing studies exhibit shortcomings in data resolution and methodology. Moreover, only a few studies have investigated the characteristics of crashes in traffic congestion, and they used low-resolution traffic data (e.g., hourly or daily flow), which do not capture oscillatory driving conditions.

This study employed a matched case-control design using high-resolution traffic and crash data from a freeway segment. The case-control design is prevalent and well established in epidemiology due to its simplicity, cost-effectiveness and theoretical soundness.

A 12-mile stretch of Interstate 5 in Portland, OR, was selected as our study site since this segment exhibits fairly extensive recurrent congestion and oscillatory flows during morning and evening peak hours. Traffic conditions, including oscillations, were measured using 20-second aggregated data from inductive loop detectors. Crash data for the study corridor are available from the statewide crash database, and crash occurrence times were obtained or estimated from the incident database and time-series traffic data.

Traffic conditions prior to each crash were taken as cases, while traffic conditions during the same periods on days without crashes were taken as controls. These were also matched by the presence of congestion, geometry and weather. In particular, the averages and the standard deviations of count, occupancy and speed during the 10 minutes prior to crash occurrences were adopted as potential measures of average traffic states and traffic oscillations, respectively. A total of 82 cases and about 80,000 candidate controls were extracted from more than three years of data from 2004 to 2007.

Conditional logistic regression models were developed based on the case-control samples with two exposure variables, average traffic state and amplitude of oscillations. To verify consistency in the results, we evaluated our model by re-sampling controls for each case, repeating the model development process (described in Chapter 5) to find the best models based on the newly drawn samples, and conducting a sensitivity analysis with respect to the case-to-control ratio.

Two major findings from our data analysis are: (1) Oscillations have a significant impact on crash occurrence: an additional unit increase in the standard deviation of speed increases the odds of (rear-end) crashes by about 8 percent; (2) The average traffic states in congestion are less

significant than deviations in speed. Nevertheless, their odds ratios are qualitatively consistent and suggest that the likelihood of crash occurrence increases as congestion becomes more severe.

Our findings are notable given that the impact of speed variation on crash occurrence has been debated for decades. This study addresses the shortcomings of existing studies by using high-resolution traffic and crash data and adopting the case-control design proven to be effective in the field of epidemiology. Given that oscillations are becoming more common in everyday traffic, the findings from this study may help prioritize countermeasures, such as ramp metering or adaptive speed control, to improve traffic safety and estimate their expected benefits.

1.0 INTRODUCTION

Traffic oscillations (also known as stop-and-go driving) on freeways, which arise in congested traffic, are characterized by recurring patterns of decelerations followed by accelerations. These oscillations are known to increase fuel consumption, engine emissions, and vehicle wear and tear (*Bilbao-Ubillos 2008; Greenwood and Bennett 1995*). They also decrease driving comfort, as drivers are forced to repeatedly adjust their acceleration rates. However, the externalities associated with traffic oscillations are yet to be assessed systematically, including the impacts on traffic safety.

The safety implications of oscillations have been studied largely by evaluating the significance of speed deviations on the likelihood of crash occurrence, though most existing studies did not make an explicit distinction between congested and uncongested traffic. Some early studies (*Lave 1985; Solomon 1964*) report that larger speed deviations increase the probability of certain types of crashes (e.g., rear-end crashes). Many researchers have challenged this finding; Davis (*2002*) provides contradictory empirical evidence. Debates continue today, as existing studies exhibit shortcomings in data resolution and methodology. Moreover, only a few studies (*Noland and Quddus 2005; Shefer 1997; Wang et al. 2009*) have investigated the characteristics of crashes in traffic congestion, and they used low-resolution traffic data (e.g., hourly or daily flow), which do not capture oscillatory driving conditions.

The present study seeks to understand the impact of traffic oscillations on the likelihood of freeway crash occurrences by analyzing event-based crash data and high-resolution (20 seconds) traffic data on a freeway segment. Oscillations are measured as variations in congested flow, speed and occupancy (a dimensionless measure of density) over a certain period. We adopted a matched case-control design in which a case corresponds to the average traffic state (e.g., average speed) and magnitude of oscillations prior to a crash. The matched control corresponds to the conditions in the same location on a day when a crash did not occur after controlling for geometry, weather and traffic states. Conditional logistic regression models were developed based on samples of cases and matched controls. The modeling results show not only that speed variations have a significant impact on (rear-end) crash occurrences but that these variations are more significant than average traffic states.

This report is organized as follows: The following chapter discusses previous research related to the present study, and Chapter 3 describes in detail the features of the study site and our efforts to process the crash and traffic data. The design of the case-control study and the subsequent modeling efforts are explained in Chapters 4 and 5, respectively. Chapter 6 evaluates and interprets the models, while Chapter 7 offers concluding remarks and suggestions for future research.

2.0 BACKGROUND

Three types of relevant literature are reviewed. We first discuss the previous studies that examined the impacts of speed variance and/or congestion on traffic safety. The review of these studies revealed that contradictory results have been reported primarily due to limitations in data resolutions and analysis methodologies. In light of this finding, various methods to relate crash characteristics to traffic features are reviewed. Finally, we provide a brief background of the case-control design and a rationale for selecting this method for our study.

2.1 IMPACT OF SPEED VARIANCE AND CONGESTION ON TRAFFIC SAFETY

Solomon (*1964*) appears to be the first study to report that speed variations have a larger impact on crash occurrences than average speeds. In his study, information from the accident records of nearly 10,000 drivers was extracted from the selected rural highways in 11 states during the three to four years prior to June 30, 1958. Their pre-crash speeds were obtained from the accident reports which were usually collected from drivers, police or witnesses, with about 20 percent missing speeds. For comparisons, speeds were measured for 290,000 crash-free drivers from the selected rural highways in these states. The study sites were divided into 35 sections according to characteristics of main rural highways in the United States. One representative location was selected from each section, and crash-free speeds were collected at this location. A sample of the drivers were also stopped and interviewed for more information.

In Solomon's model the dependent variable was the accident involvement rate, which was defined as involvements per hundred million vehicle-miles. He found that 1) the relationship between crash risk and speed has a U-shaped curve; 2) vehicles with speeds of around 65 mph (105 km/h) have the lowest crash risk; and 3) the crash risk increases with increases in speed variations. This trend is consistent for both daytime and nighttime.

Solomon's findings received little attention until Lave (*1985*) claimed in his controversial paper that once one controls for speed variations, average speed has little impact on highway safety. He used the state average data from 1981 and 1982 based on highway statistics from the U.S. Department of Transportation. The analysis was carried out for six different highway types (Rural Interstate, Rural Arterial, Rural Collector, Urban Freeway, Urban Interstate and Urban Arterial) using the difference between average speeds and 85^{th} percentile speeds to measure speed variations. In total, 12 linear models were obtained (six for each year). The model results show that average speed is consistently insignificant and that the fatal rate would increase with the increase of speed variations. Therefore, the author claimed that speed laws should force drivers to coordinate their speeds instead of just limiting their speeds.

Several studies (*Davis 2002; Fowles and Loeb 1989; Lave 1989; Levy and Asch 1989; Synder 1989*) followed, debating Lave's claim. Levy and Asch (*1989*) found that the mean and variance

of speeds were actually correlated, thus disputing Lave's finding. Synder (*1989*) distinguished between fast and slow vehicles, examining the significance of the speed variance in each group and comparing the average speeds between the two groups. He found that average traffic speed is an important determinant of highway fatalities and that speed variance is important for fast vehicles only.

More recently, Malyshkina and Mannering (*2008*) conducted a before-and-after study to assess the severity of accidents in relation to the increased speed limits (65 to 70 mph (105 to 113 km/h)) in 2005 on some rural interstates in Indiana. They found that the increased speed limits did not have a statistically significant effect on the severity of accidents.

The aforementioned studies used data that were highly aggregated temporally and spatially (e.g., annual statewide data) and may be subject to an ecological fallacy, as noted by Davis (*2002*). By focusing on the foundation of statistical inference and using a relationship between vehicle speed and pedestrian accident risk on residential streets as an example, Davis demonstrated how contradictory results can be obtained by using data generated at different levels of detail. A simulation model was used to generate hypothetical data of crash risk of heedless pedestrians on residential streets from three levels of aggregation: the individual vehicle/pedestrian encounter, the population of encounters at a given site, and the population of sites. Based on the simulated data, he found that at the individual encounter level and at a given site level, the collision probability increases with the increase of mean speeds, which is consistent with the observations. However, at the highest aggregated level (the population of sites), the mean speed essentially appears to have no significant impact on the collision risk. In this way, he demonstrated the potential danger introduced by highly aggregated data used in traffic safety studies.

Moreover, in the aforementioned studies, no distinction was made between congested and uncongested traffic, even though traffic properties in the two regimes are not the same and may have different impacts on traffic safety.

Fewer studies have examined the effect of traffic congestion on crash occurrences. Shefer et al. (*1997*) suggest that crash frequency increases in congestion due to increased interactions among vehicles, while crash severity decreases because of the lower speeds of congested traffic. However, they used simulated data to test the performance of their model, which undermines its validity. Noland and Quddus (*2005*) report conflicting results based on the analysis of enumeration district data for London, which imply that congestion in urban areas is unlikely to reduce crash severity and frequency. Wang et al. (*2009*) confirm this finding using the congestion index (CI), which is defined as the ratio of difference in actual and free-flow travel times to the free-flow travel time, to measure the congestion level of a motorway near London. However, the congestion index was calculated from hourly data averaged over a year. Thus, the data resolutions of these studies were not adequate to analyze the effect of oscillatory traffic flow in congestion.

It is evident that existing studies report inconsistent findings on the effect of speed variance and congestion on crashes; this seems attributable to limited data resolution and analytical methods. Moreover, it appears that there has been limited effort, if any, to investigate the effect of traffic oscillations on freeway traffic safety.

In the following section, we briefly review methods to relate crash characteristics to traffic features in order to adopt a statistically sound methodology.

2.2 METHODS TO RELATE CRASH CHARACTERISTICS TO TRAFFIC FEATURES

For decades, researchers have used various methods to relate crash characteristics to traffic features. Many pioneering studies tried to establish a connection between crash and traffic conditions based on aggregated data (e.g., *Lave 1985; Solomon 1964*). In essence, they related aggregated indicators of traffic conditions, such as speed deviations (e.g., the annual 85th percentile speed minus the annual average speed), to crash frequency. As mentioned previously, using highly aggregated data, such as annually averaged data, can result in biased findings that are open to interpretation.

Another common method uses the speeds of individual vehicles (e.g., *Cirillo 1968; Research Triangle Institute 1970*) to compare a crash-involved vehicle to other vehicles. This method imposes a great challenge on data collection endeavors, as the trajectories of vehicles in crashes are necessary to measure pre-crash speeds (*Xin et al. 2008*).

Some researchers have attempted to use individual drivers' accident histories (*Fildes and Rumbold 1991; Tilden et al. 1936*) by sampling speeds of selected drivers at a particular location and then analyzing their crash histories. The speeds of selected vehicles were measured using unobtrusive devices (e.g., a radar gun), and information on drivers (e.g., demographics, trip, vehicle data, and viewpoint to speeding) was collected by stopping them downstream of the road. Accident history details were accessed through the government database by matching vehicles' registration numbers. Fildes and Rumbold (*1991*) found that drivers who were traveling above the mean speed of the traffic or the posted speed limit were often found to have been involved in crashes over the past five years. Contrary to Solomon's findings, no U-shaped relationship between crash involvement and speed deviations was found in their study, which might be partially due to the insufficient sample size in excessively fast and slow speeds, as they acknowledged.

Comparatively, this study approach is easy to implement in practice. However, this approach has been criticized by Kloeden et al. (*1997*) for privacy issues and selection bias in the data collection as elaborated below:

(1) This approach measures the pre-crash speed of a driver by sampling speeds independently at the crash location. The pre-crash speed measured in this way is highly likely to be biased due to different traffic conditions at different times;

(2) This approach inherently cannot cover fatal accidents in which drivers were killed since it is impossible to sample deceased drivers on the location of interest and then access their accident history.

Golob and Recker (*Golob and Recker 2004*) developed a novel approach to determine how crash characteristics are related to traffic flow conditions at the time of crash occurrence. They resorted to data mining techniques to study this issue from a pure data analysis perspective. To reduce the

dimensionality of the traffic data by accounting for correlations among traffic flow variables, the principal components analysis (PCA) was used in their study. Then a cluster analysis was deployed to find homogenous groups of traffic flow conditions, which were called "regimes." For the data from more than 1,000 crashes in Southern California, they identified 21 traffic flow regimes for three different ambient conditions: dry roads during daylight (eight regimes), dry roads at night (six regimes), and wet conditions (seven regimes). Then the nonlinear canonical correlation analysis (NLCCA) based on the alternating least squares (ALS) algorithm was employed to evaluate the link strength between crash characteristics and each traffic flow regime.

Their analysis has demonstrated that by controlling for environmental effects, the descriptive characteristics of crashes are distinguished by distinct traffic flow regimes. They have concluded that characteristics of freeway accidents are associated with the traffic flow conditions that prevail before an accident.

The analysis techniques employed in their study are somewhat unconventional to this domain of study. Rather than building upon the foundation of traffic engineering principles, they instead viewed the problem as essentially a data analysis problem, and relied on statistical techniques to help reveal the structure of the underlying phenomena. One major advantage of this methodology is that creating a panel data (non-crash data) can be avoided and it is comparatively easier to implement the method. However, as mentioned before, this method is purely derived from data analysis techniques rather than being based on the theoretical background of traffic flow. In addition, too many traffic regimes were identified and criteria to meaningfully distinguish them are not straightforward.

Finally, some traffic safety studies borrow the case-control design pioneered in the field of epidemiology (*Kloeden et al. 1997; Abdel-Aty et al. 2005*). This method is suitable for safety-related research, as a crash is a rare event. The next section discusses case-control design in greater detail.

2.3 REVIEW OF THE CASE-CONTROL DESIGN

The case-control design is an efficient method to study rare events that is particularly prevalent in epidemiology (*Manski 1995; Schlesselman and Stolley 1982*) due to its simplicity, cost-effectiveness, and theoretical soundness. The central idea of the case-control study is to compare two groups, one with the outcome of interest (such as disease, death and crashes) and one without it by incorporating potential explanatory factors (exposures) (*Cornfield et al. 1959; Cornfield 1951; Fisher 1958a; Fisher 1958b*). Cornfield (*1951*) proved theoretically the validity of the case-control method and demonstrated its effectiveness using epidemiological examples. The theoretical details of the case-control method are beyond the scope of this paper; interested readers should refer to Breslow and Day (*1980*) and Cornfield (*1951*) for more in-depth discussions.

Traffic safety studies using a case-control design are scattered in the traffic literature. These studies can be approximately put into two categories according to their data collection methods: (1) case-control designs using data for individual vehicles that are involved in crashes or present at accident scenes and (2) case-control designs using loop detector data to measure traffic

conditions. Of note, these two types of case-control studies require different interpretation to identify the results.

The most natural way to use a case-control design to study traffic safety is at the individual vehicle level. A vehicle involved in a crash is taken as a case, and other vehicles in the crash scene or in similar situations (but not involved in a crash) are taken as controls.

It should be pointed out that case-control studies were also adopted in Solomon (*1964*), Cirillo (*1968*) and Research Triangle Institute (*1970*). However, the selection of controls in these studies is rather problematic as noted previously. Following these studies, the first notable case-control study in the traffic literature is Kloeden et al. (*1997*). The primary aim of their study was to quantify the relationship between freely traveling speed and the risk of involvement in a casualty crash for sober drivers in 38 mph (60 km/h) speed limit zones in the metropolitan area of Adelaide, Australia. Other aims were to examine the effect of hypothetical speed reductions on crashes and to study the impact of driver blood alcohol concentration (BAC) on traveling speeds.

Kloeden et al. (*1997*) developed a case-control study by treating cars involved in casualty crashes as cases and cars not involved in crashes as controls by matching the location and time. In order to collect valid data on cases and controls, they carefully imposed the conditions in the data collection endeavor. Data were mainly collected from 9:30 a.m. to 4:30 p.m. Monday through Friday, as these times had the highest number of non-alcohol-related crashes in Adelaide. In order to isolate potential contributions to crash occurrence by confounding variables (e.g., driving under the influence of alcohol), drivers whose BACs were above zero were filtered out. For the same reason, cases related to illegal maneuvers prior to a crash also were not included in the analysis.

Based on detailed information obtained from on-site crash investigations, crashes selected as cases were reconstructed using the computer reconstruction programs (M-SMAC) to estimate their pre-crash speeds. Furthermore, an expert panel reviewed crash-related data which were collected or estimated; cases that did not pass this review procedure were excluded from the analysis.

A brief description on their data collection process follows. When the Road Accident Research Unit (RARU) received a notification of a crash either from the ambulance radio frequency or from a paging service provided by the South Australian Ambulance Service, RARU went to the crash scene and conducted a quick survey to decide whether the crash met the data collection criteria. If the crash indeed met the criteria, the crash investigation team documented the scene and collected information on the crash. Meanwhile, supplemental information on the crash (e.g., BAC) was extracted from the police accident report. Based on such information, the crash was reconstructed using M-SMAC to obtain the pre-crash speed. The validity of the reconstruction of the crash was evaluated by an expert panel.

For each selected crash (case), vehicles in similar conditions were randomly selected and their speeds were measured unobtrusively by using a laser speed meter. Then, the selected vehicles were stopped further down the roadway by a police officer, who conducted breathalyzer tests. In their study, four selected vehicles whose drivers were confirmed by the test to have a zero BAC formed the control group for that case.

By statistically comparing speeds of the cases and the controls, they found that when a driver travels above 38 mph (60 km/h) in a 38 mph (60 km/h) speed zone, the likelihood of being involved in a fatal crash doubles with each 3 mph (5 km/h) increase in traveling speed.

Overall, the study by Kloeden et al. (*Kloeden 1997*) was well done. The case-control study was carefully designed and an enormous data collection effort was executed. However, they acknowledged that it was difficult in the real world to arrange for police officers to do the breathalyzer test on selected drivers. This attempt was given up later, and thus the majority of controls were not subject to BAC verification. The authors argued that BAC had no meaningful effect on the results since only a small proportion (3%) of the controls tested had a positive BAC. However, their results were possibly biased. Moreover, this study only focused on uncongested traffic flow conditions and only used speed to explain occurrences of fatal crashes.

More recently, Davis et al. (*2006*) used a case-control design to investigate the effect of speed in run-off-road crashes since they believed that different crash types may be caused by fundamentally different processes. In their study, two datasets were used. One consisted of 14 run-off-road fatal crashes in Adelaide, Australia, (the same data as in Kloeden et al. (*1997*)) and the other consisted of 10 run-off-road fatal crashes in Minnesota. A Bayesian approach was adopted in their analysis to address the small sample-size issue and pre-crash speed estimation uncertainties in crash reconstructions.

Prior to modeling, they demonstrated the soundness of using logistic regression to model crash occurrences. Then they proposed probabilistic versions of the crash reconstruction procedure by Kloeden et al. (*1997*) to estimate pre-crash speeds.

In their case-control design, four controls were selected per case. They explicitly tested the U-shaped relationship between fatal run-off-road crash occurrences and speed variations proposed by Solomon (*1964*). They found that the likelihood of fatal run-off-road crashes increases with an increase in speed. However, the probability does not increase with speed decreases. Therefore, they claim that the U-shaped relationship hypothesized by Solomon (*1964*) does not exist for fatal run-off-road crashes in the two datasets.

Although employing a case-control design at an individual vehicle level is ideal, collecting data for individual vehicles is difficult in practice. For example, the pre-crash speed of a vehicle is usually estimated using crash reconstruction methods. However, these reconstruction methods are often quite complex and the estimated speed is subject to bias, as noted by Davis et al. (*2006*).

In an effort to remedy this shortcoming, Abdel-Aty et al. (*2005*) developed a case-control design using five-minute aggregated data from inductive loop detectors to develop a real-time crash prediction logistic model for multivehicle crashes on a 36-mile stretch on Interstate 4 in Central Florida. Their primary goal was to link crash occurrences with real-time traffic patterns observed through loop detector data, weather conditions, and geometric factors based on crash data and traffic data from 1999 to 2002.

For each crash as a case, five controls were selected by controlling location, time of day and day of week in their study. Traffic data for a half-hour prior to the estimated crash occurrence were

extracted for each crash from the loop detector station nearest to the crash location as well as the four nearest upstream stations and two nearest downstream stations. Similarly, traffic data for five crash-free controls during the same period from these stations were extracted. Of note, the detailed estimation technique for crash occurrence time was not provided in their final report despite their claim that "findings… are based on accurately estimated time of the crash thereby evading the 'cause and effect' fallacy."

Based on the preliminary analysis, they included five-minute aggregated loop data with only multivehicle crashes to develop a final model based on 1,528 strata (each stratum consists of one crash and five corresponding non-crash controls). Their analysis on the spatio-temporal pattern of the hazard ratio shows that the speed variation, the average occupancy, and the standard deviation of volume 5-10 minutes prior to the crash are significantly associated with the likelihood of crashes on the freeway studied.

However, they did not distinguish between the uncongested and congested traffic regime, and they did not focus on oscillations in congestion.

Of note, some studies have employed case-control designs to examine the safety impact of certain risk factors. Gross and Jovanis (*2007*) provide a review of case-control studies used in the highway safety literature. They also developed a matched case-control design to quantify the effect of road geometry such as lane and shoulder widths. However, traffic operational features (oscillations in particular) were not examined in their study.

In the present study, a case-control design is implemented in combination with event-based crash data and high-resolution (20-second) traffic data to examine the impact of traffic oscillations in congestion on the likelihood of crash occurrence.

3.0 STUDY SITE AND DATA PROCESSING

This section describes the selection of a study site and processing of crash and traffic data. The traffic data used in this research were retrieved from the Portland Oregon Regional Transportation Archive Listing (PORTAL) (*2009*). PORTAL (http://portal.its.pdx.edu) is the official Intelligent Transportation Systems (ITS) data archive for the Portland metropolitan region. PORTAL has been archiving 20-second speed, count, and occupancy data from inductive loop detectors on Portland-area freeways since July 2004. Crash-related data are obtained from two databases maintained by the Oregon Department of Transportation (ODOT). The statewide Crash Data System (CDS) contains all reported crashes on public roads in Oregon, and the incident database records all freeway incidents in the Portland metro area as logged by the Traffic Management and Operations Center (TMOC).

3.1 STUDY SITE SELECTION

A study site was selected among several freeway corridors in the Portland metro area (see Figure 3.1), which consists of several interstates, U.S. highways and state routes. Several of them experience recurrent congestion during a.m. and/or p.m. peak hours. I-5 North was selected for this study based on a number of criteria: extent of congestion and oscillations, spacing of loop detector stations, traffic data quality, crash data availability, and crash sample size.

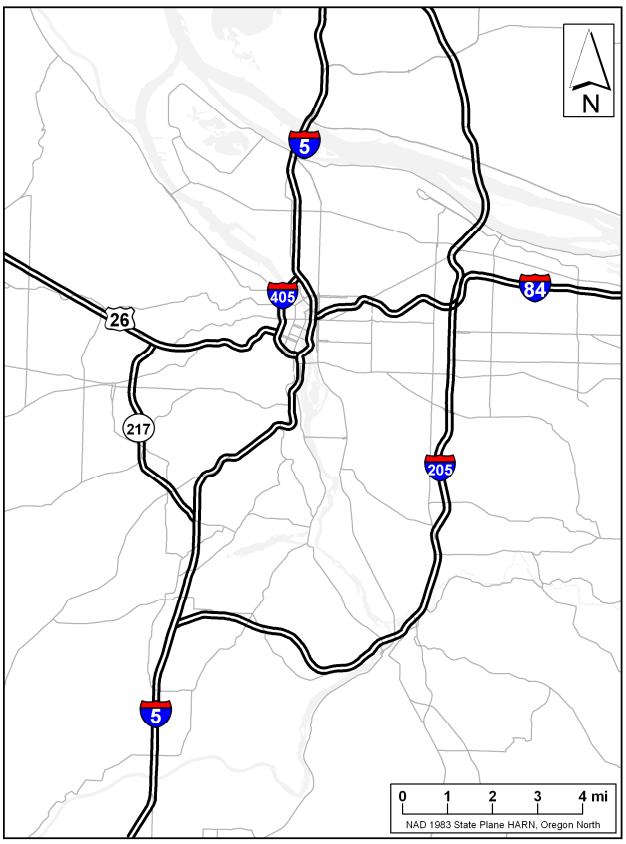


Figure 3.1: Freeway network in the Portland metro area

The speed contour plot in Figure 3.2 illustrates the average traffic conditions of the study section during weekdays in January 2006. Time and space (in terms of mileposts) are shown on the x- and y-axes, respectively, and the color scale represents the estimated time-mean speeds over the month according to the legend in the figure. From the figure, two congestion patterns are identified. One started around 7 a.m. and ended around 9 a.m. This congestion was caused by a bottleneck located at about milepost 300 and propagated backward to as far as milepost 286. The second congestion period started around 4 p.m. and ended around 6 p.m. The bottleneck responsible for this congestion was located somewhere beyond milepost 307.9. The resulting queue propagated upstream to about milepost 296. These patterns on I-5 North were typical for most weekdays between 2004 and 2007.

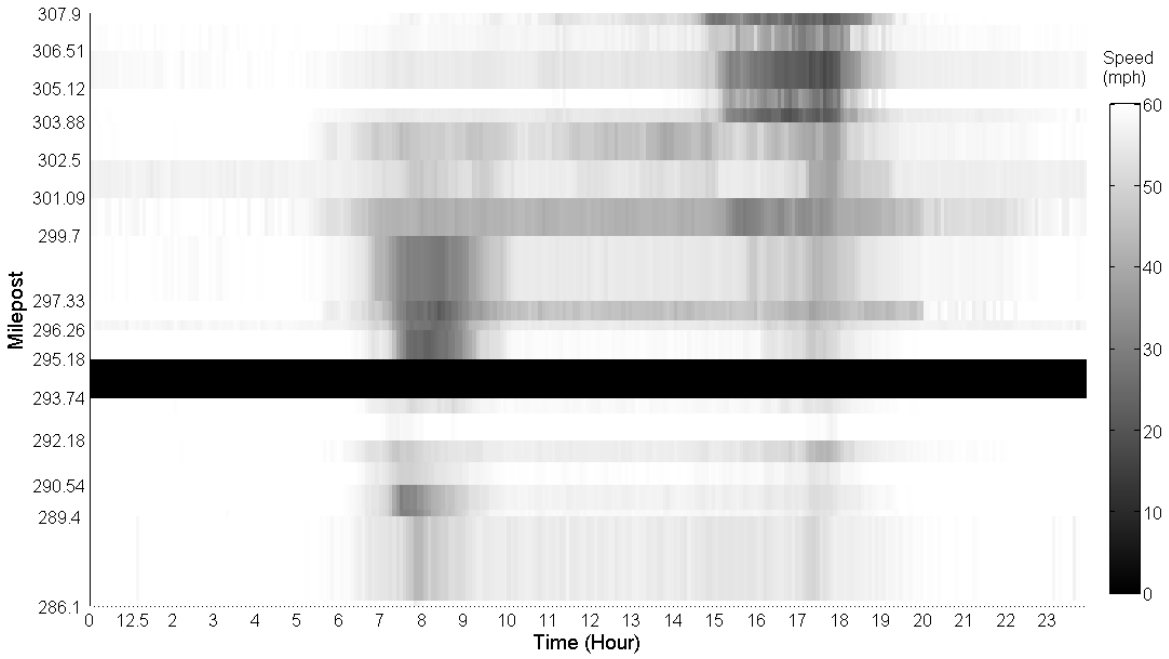


Figure 3.2: Speed contour for I-5 North on weekdays in January 2006

Traffic data from loop detectors often contain invalid data (e.g., missing value, negative value, and non-zero speed with zero count) due to malfunctioning detectors, communication failures, and other reasons (e.g., the loop detectors at milepost 293.74 in Figure 3.2 were malfunctioning for the entire month of January 2006.). Thus, it is necessary to evaluate the quality of traffic data collected from inductive loop detectors by the following steps:

Step 1: Download data for I-5 North from PORTAL;

Step 2: Extract data for a.m. (7-9 a.m.) and p.m. (4-6 p.m.) peak hours;

Step 3: For each day, aggregate data from the detectors at the same stations based on mileposts;

Step 4: For a.m. and p.m. peak hours for each day and each station, calculate the following indices:

(1) Average speed, *Avgspd*, during each peak:

$$Avgspd = \frac{\sum_t spd(t) * vol(t)}{\sum_t vol(t)}$$

Where, spd(t): time-mean-speed between time (t-15) minutes and t;
vol(t): volume between (t-15) minutes and t.

Any records with no readings were disregarded in the computation. If data were missing throughout the entire peak period, Avgspd is set to N/A.

(2) percentage of no-readings, *P_no* :

$$P\_no = \frac{\sum_{t} Count\_no\_data(t)}{sum(Count)}$$

Where,Count_no_data(t): number of no readings between (t-15) minutes and t;
sum(Count): total number of readings during the peak period.

(3) percentage of valid data, *P_ok*:

$$P\_ok = \frac{\sum_{t} Count\_ok(t)}{sum(Count)}$$

Where, Count_ok(t): number of readings with a good quality between (t-15) minutes and t;
sum(count): total number of readings during the peak period.

(4) percentage of data with uncertain quality, *P_suspect*:

$$P\_suspect = \frac{\sum_{t} Count\_sus(t)}{sum(Count)}$$

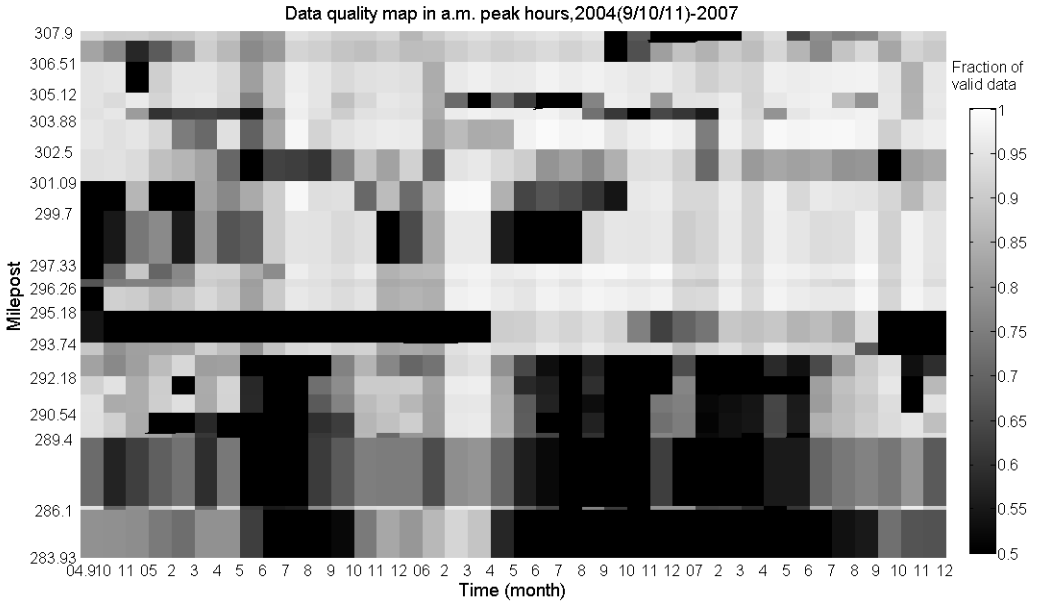
Where, Count_sus(t): number of suspect readings between (t-15) minutes and t;
sum(count): total number of readings during the peak period.

Step 5: **Output results in the format as shown in**
Table 3.1(Avgspd as an example):

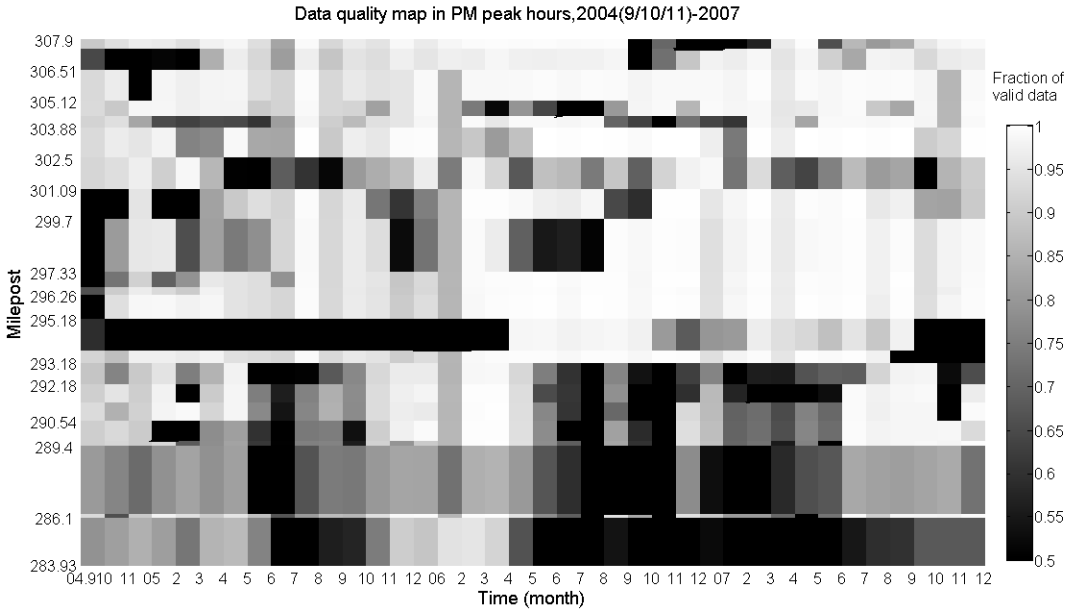
Table 3.1: A typical output structure for data quality matrix

	Milepost 1	Milepost 2	…	Milepost n
Day 1	Avgspd	Avgspd	…	Avgspd
Day 2	Avgspd	Avgspd	…	Avgspd
…	…	…	…	…
Day 31	Avgspd	Avgspd	…	Avgspd

To visually check monthly data quality at each station of I-5 North from 2004 to 2007, data quality maps were generated using the *P_ok* and *P_suspect* data for the a.m. and p.m. peak hours (see Figure 3.3 (a) and (b), respectively). The monthly percentages of valid data (defined as sum of *P_ok* and *P_suspect*) are shaded according the color scale shown in the legend; darker regions correspond to lower data quality. The figure reveals that the percentages of valid traffic data are below 85 percent for the segment between mileposts 295.18 and 283.93 in the a.m. and p.m. peak hours. For the other segment of I-5 North (mileposts 296.26 - 307.9), the percentages of valid traffic data are generally larger than 85 percent. Furthermore, the traffic data have better quality in the p.m. peak hours than in the a.m. peak hours. Note that days with invalid traffic data prior to crashes were excluded in the further analysis.



(a) The a.m. peak hours



(b) The p.m. peak hours

Figure 3.3: Data quality map for I-5 North Portland, OR

Figure 3.4 shows the crash counts for the a.m. and p.m. peak hours from 2004 to 2007. More crashes evidently occurred in the p.m. peak hours. Therefore, a larger crash sample size might be obtained from the p.m. peak hours.

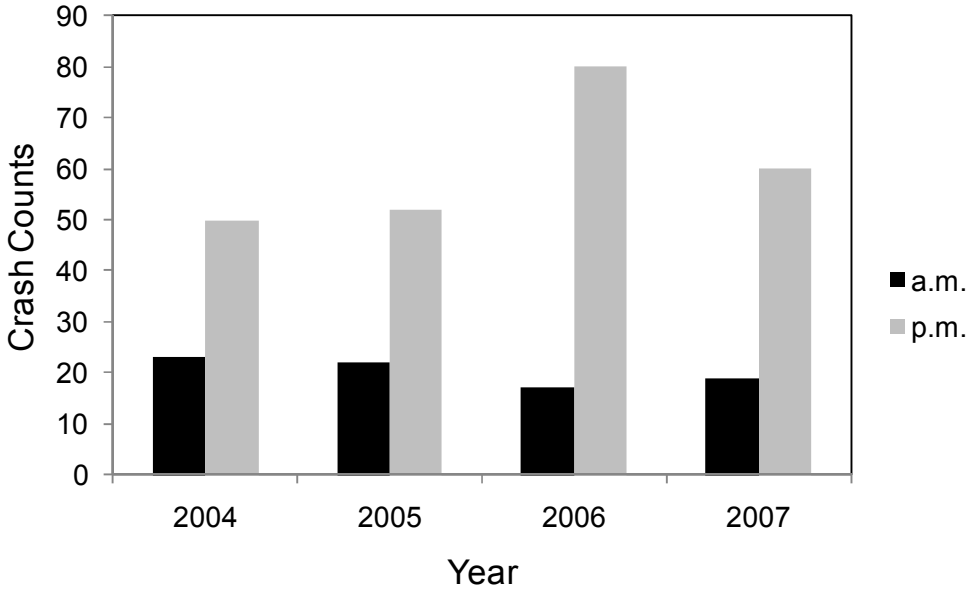


Figure 3.4: Crash number in p.m. peak hours vs. in a.m. peak hours in 2004-2007

During the congested periods, traffic oscillations were examined visually using time-series speed plots to confirm their presence and amplitudes. For instance, Figure 3.5 shows speed over time (4:30-5:30 p.m.) at Jantzen Drive on May 14, 2007. The plot exhibits a recurring pattern of acceleration followed by deceleration, with the speed ranging from around 12 to 37 mph (19 to 60 km/h). The figure also shows varied oscillation amplitudes, such that some periods exhibit larger speed variations (e.g., 4:50-5 p.m.) than others (e.g., 5:20-5:30 p.m.).

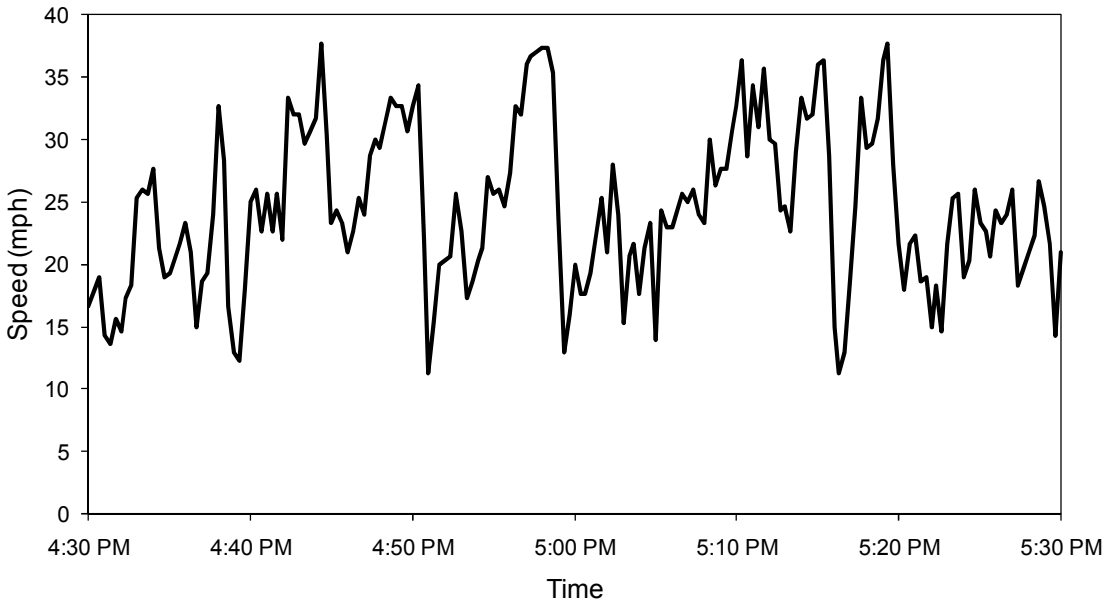


Figure 3.5: Speed-time series plot at Milepost 307.9 on May 14, 2007

Based on the analysis above, a 12-mile (19-km) section (mileposts 296.26 to 307.9) of I-5 North (see Figure 3.6 for a schematic) has been selected. The most suitable time period for our study is the p.m. peak hours (from 4-6 p.m.) from 2004 to 2007.

Note that this segment contains three lanes, and from mileposts 303 to 306 the left-most lane is dedicated to High Occupancy Vehicles (HOV). The HOV lane is limited to vehicles with at least two passengers from 3-6 p.m. on weekdays. Spacing between loop detector stations ranges from 0.34 to 2.37 miles (0.6 to 3.9 km) with an average of about 1.06 miles (1.7 km). Data from the loop detectors (vehicle count, occupancy and time-mean speed during each 20-second interval) are available from PORTAL (*2009*).

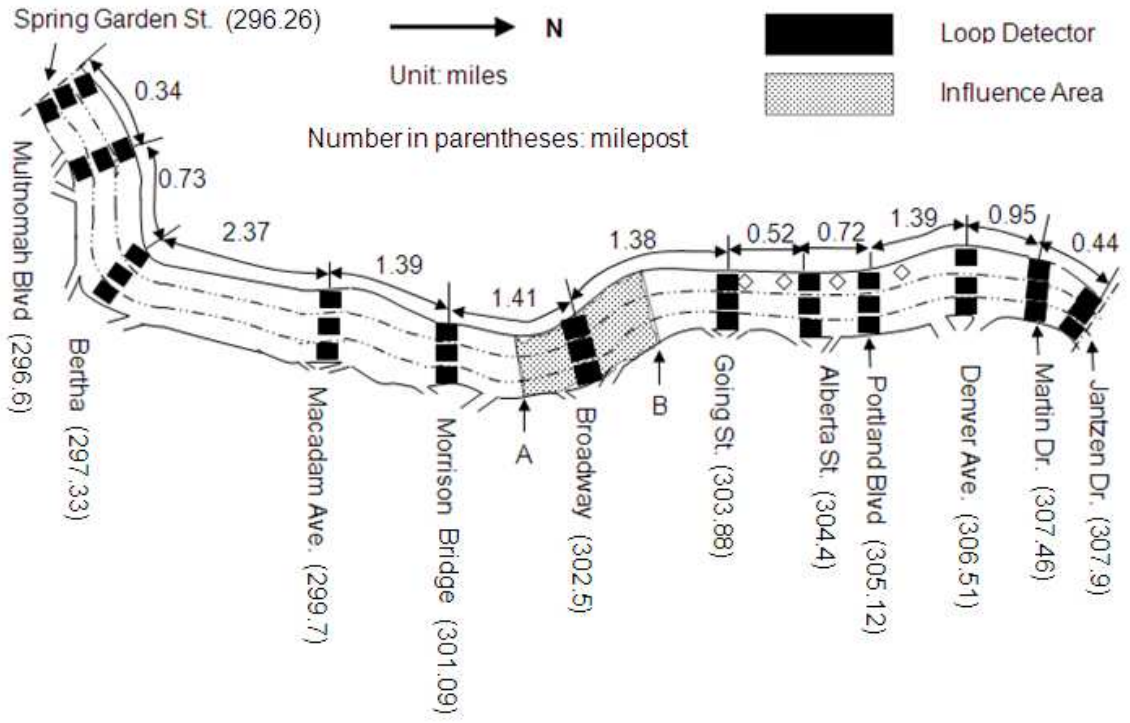


Figure 3.6: Schematic of the study site, northbound I-5, Portland, OR

Loop detectors (black squares in Figure 3.6) measure traffic conditions at 12 locations on the selected freeway segment; these conditions are extrapolated to represent the conditions between measurement locations. Notably, we define "influence areas" that are bounded by the midpoints between neighboring stations. The shaded area in Figure 3.7 is an example of an influence area. *A* corresponds to the midpoint between Broadway Street and the Morrison Bridge, and *B* corresponds to the midpoint between Broadway and Going Street. Table 3.2 provides more detailed information on the influence areas of the study segment.

Table 3.2: Mileposts of loop detector stations and influence areas, I-5 Northbound, Portland, OR

ID	Milepost	Location Description	Influence Area
1	296.26	Spring Garden St.	(295.72 -296.43)
2	296.6	Multnomah Blvd.	(296.43 - 296.97)
3	297.33	Bertha Blvd	(296.97 - 298.52)
4	299.7	Macadam Ave.	(298.52 - 300.40)
5	301.09	Morrison Bridge	(300.40 - 301.80)
6	302.5	Broadway St.	(301.80 - 303.19)
7	303.88	Going St.	(303.19 - 304.14)
8	304.4	Alberta St.	(304.14 - 304.76)
9	305.12	Portland Blvd.	(304.76 - 305.82)
10	306.51	Denver Ave.	(305.82 - 306.99)
11	307.46	Marine Dr.	(306.99 - 307.68)
12	307.9	Jantzen Dr.	(307.68 - 307.9)

We assume that traffic conditions (including oscillations) are the same within each influence area and are represented by the measurements taken at the corresponding detector station. This assumption was made due to the inherent limitation of loop detectors, which measure traffic conditions at specific points rather than over space. Moreover, the exact locations of crashes are unknown, and thus, it is not straightforward to estimate oscillations at the exact crash location. Nevertheless, we found that the assumption of uniform oscillations within an influence area is quite reasonable as demonstrated in the next section.

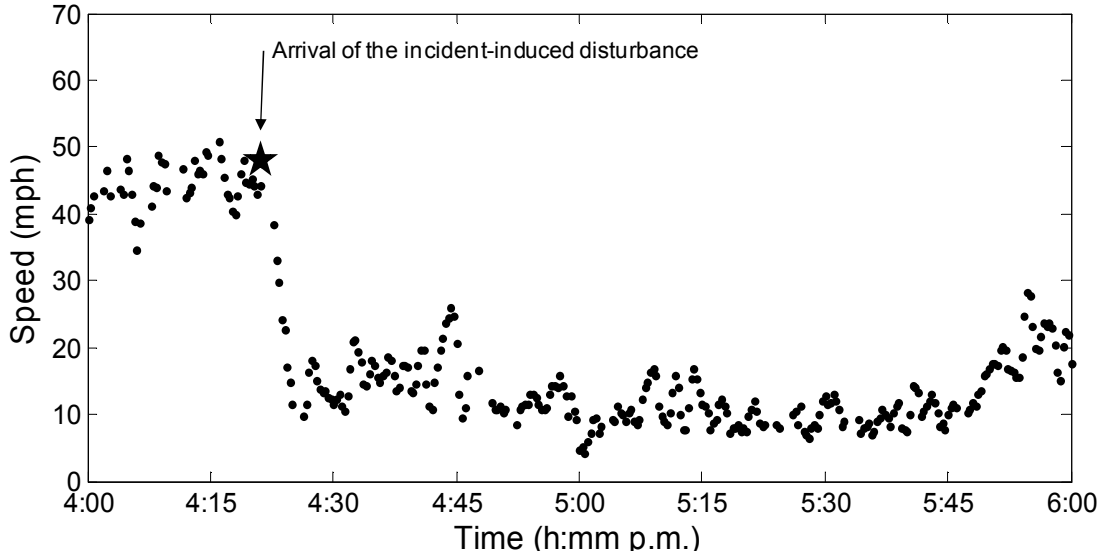
3.2 CRASH DATA PROCESSING

Crashes in the CDS are recorded to the nearest hour while the incident data is at much higher resolution (minutes). CDS was used to identify crashes, and the incident data were used to confirm the crashes and retrieve more detailed information about the crashes.

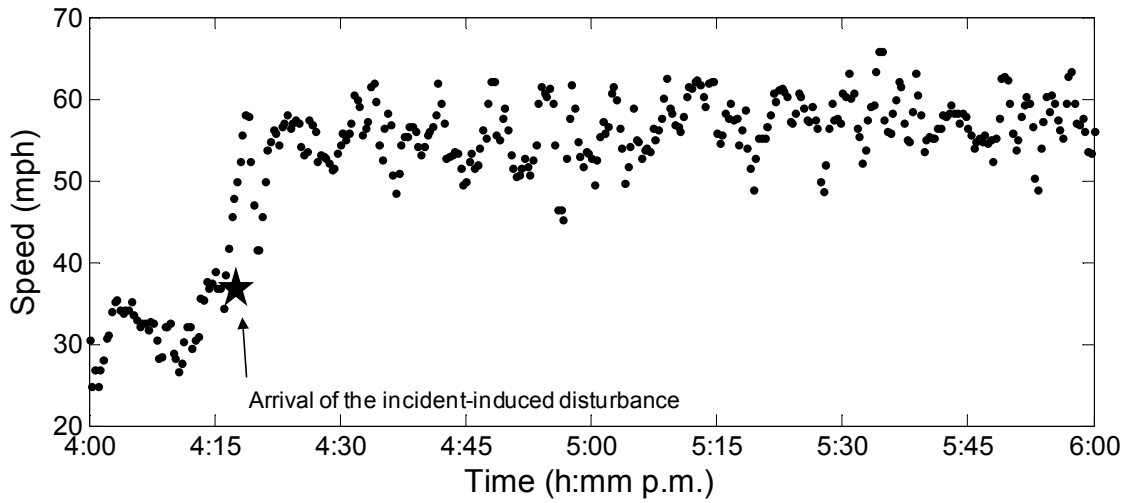
A total of 242 crashes were reported in the study segment during the evening peak periods in 2004 to 2007. Several steps were taken to obtain or estimate the occurrence times of these crashes. The occurrence times were extracted from the incident database since the incident database provides more accurate incident occurrence times than the crash database. For crashes missing from the incident database, 20-second traffic data are used to estimate the occurrence times based on the assumption that a crash in queued traffic will be reflected in the traffic characteristics (e.g., a sharp speed drop (increase) at an upstream (downstream) detector station).

Figure 3.7 demonstrates this assumption using time-series speeds at the detector stations immediately upstream (Macadam Avenue) and downstream (Morrison Bridge) of a crash that occurred at milepost 300.31. According to the crash database, this crash occurred sometime between 4 and 5 p.m. on April 26, 2006. The time-series speed at the upstream detector station (3.7(a) initially displays relatively moderate congestion with a speed of 45 mph (73 km/h)). However, around 4:20 p.m. (marked by a star in the figure), the speed decreases markedly to around 10-15 mph (16-25 km/h), marking the arrival of the disturbance due to the crash downstream. The opposite trend is observed at the downstream location (b), which is initially characterized by a congested speed of about 30 mph (49 km/h). However, around 4:18 p.m., the speed increases to about 55-60 mph (89-97 km/h), marking the arrival of the crash-induced disturbance. Since this crash occurred in the influence area of the loop station at Macadam Avenue (see Table 3.2), data from this upstream station taken prior to 4:20 p.m. are used to represent pre-crash traffic conditions. Crashes whose occurrence times could not be determined from the incident database or traffic data were excluded from our analysis. (Traffic data from the days of these crashes were also excluded.) A total of 194 crashes remained (out of 242).

The remaining 194 crash events were further scrutinized and removed if they met any of the following three criteria: (1) the crash occurred in an uncongested condition; (2) traffic data before the crash are invalid; or (3) an involved driver was impaired by drugs/alcohol. After these filtering processes, a total of 82 crashes remained; these were used in the analysis. Table 3.3 summarizes the filtered crashes from each year.



a) Speed-time series at Macadam Ave.



b) Speed time-series at Morrison Bridge

Figure 3.7: Illustration of crash occurrence time estimation

Table 3.3: Yearly sample size of crashes

Year	Crashes in p.m. peak hours	Crashes with occurrence time matched in the incident database	Crashes with occurrence time estimated from the traffic data	Crashes with no confirmed occurrence time	Crashes filtered out due to other factors	Filtered crashes
2004[a]	50	34	0	16	31	3
2005	52	37	9	6	28	18
2006	80	49	16	15	25	40
2007	60	26	23	11	28	21
Total	242	146	48	48	112	82

[a] Traffic data in 2004 were available only in September, October and December.

The majority of the 82 crashes are rear-end crashes (71 out of 82), which seems reasonable. Nine sideswipe-overtaking crashes (vehicles traveling in the same direction on parallel paths collide) and one fixed or other object crash (a vehicle strikes a fixed or other object on the roadway or off roadway) were observed. One crash was coded as a turning movement crash, but inspection of the entire crash record indicates that this was probably miscoded. The crash occurred on the mainline in the vicinity of an exit ramp. Both vehicles were traveling in the same direction and likely collided when a vehicle made a maneuver towards the exit ramp (the reason for the turn code). As all of these crashes can occur due to oscillations, all 82 crashes were included in our analysis. Of further note, the descriptions of the turning-movement and fixed-or-other-object crashes were not clear enough to determine if the crashes were potentially induced by oscillations. To be on the conservative side, these crashes were included in the analysis. Nevertheless, these crashes are unlikely to affect our results significantly since there were only two crashes of these crash types.

Finally, the 82 crash events were linked to traffic data based on their mileposts and the corresponding influence areas. Of note, we used one-minute moving averages of traffic measures (speed) in order to smooth out noises while preserving the underlying trend. Measurements were also combined for all lanes to incorporate influences from adjacent lanes and maximize the sample size (sample size can be maximized by using measurements taken in other lanes in case of a malfunctioning loop) in view of Abdel-Aty et al. (*2005*). Weather information (e.g., rainfall) was obtained from PORTAL (*2009*) and incorporated into crash and traffic data.

We now turn our attention to verifying the assumption of uniform oscillations within each influence area. We sampled oscillations during a 10-minute period (around 5 p.m.) each day at the most downstream location and measured the amplitude of oscillations by taking the standard deviation of speeds. Note that the duration of 10 minutes was used since oscillations typically exhibited a comparative period. This is further elaborated in the following section. The amplitudes of the sampled oscillations were then traced at all upstream locations by accounting for the oscillation propagation speed. As noted in Mauch and Cassidy (*2002*) and Ahn and Cassidy (*2007*), oscillations in congestion propagate upstream against traffic flow at nearly constant speeds independent of traffic states. We estimated the propagation speed from one detector station to the station immediately upstream using the cross-correlation technique. The basic idea is to search for a time lag which maximizes the correlation coefficient between the

time series from two detector stations. The optimal time lag corresponds to the travel time of oscillations between the two locations. The propagation speed corresponding to the optimal time lag was found to be around 12 mph.

Figure 3.8 shows the basic statistics (e.g., average and standard deviations) of oscillation amplitude at each detector station for all crash-free days, as well as the spatial distribution of the identified crashes. The bars stand for number of crashes, the solid line is for average oscillations, and the dotted lines are variations of oscillations over time bounded by one standard deviation. The figure shows that the average amplitudes of oscillations are reasonably stable over space, such that, on average, they did not deviate more than 0.4 mph within each influence area. The largest difference of 0.8 mph is observed at the Multnomah station based on linearly interpolated amplitudes, though only four crashes occurred within this influence area. Thus, it seems reasonable to assume that oscillations within an influence area can be reasonably approximated by the oscillations measured at the corresponding loop detector station.

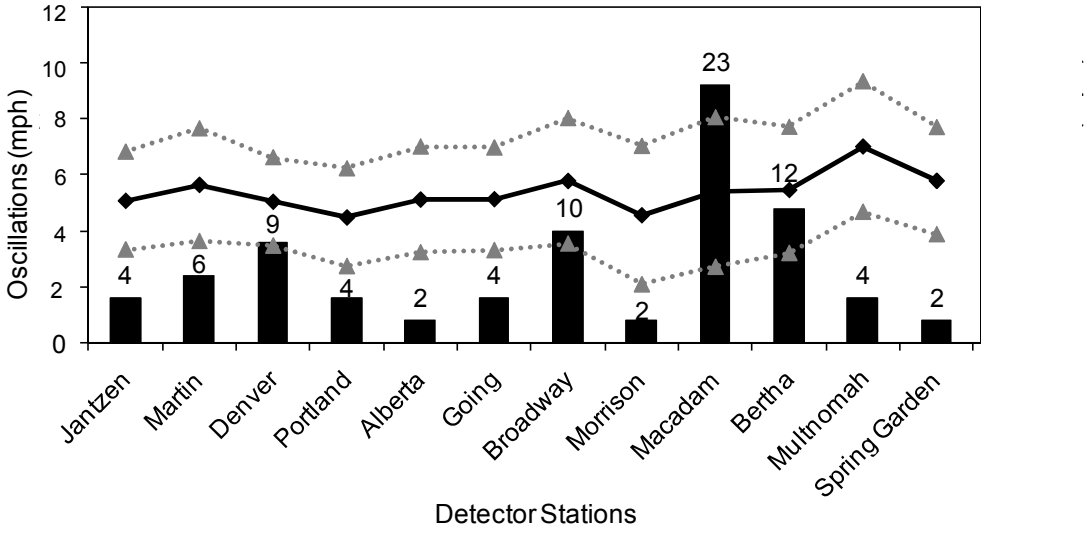


Figure 3.8 Spatial distributions of oscillations in a 10-minute period from 2004-07 and number of crashes sampled for this study.

4.0 METHODOLOGY: THE CASE-CONTROL DESIGN

The matched case-control method was selected to study the relationship between traffic oscillations and crash occurrences in congestion. The explanatory variables are oscillations and average traffic states during a period of time (10 minutes, as explained below) prior to crashes. Potential measures of oscillations include standard deviations of speed, count, and occupancy. The mean speed, count and occupancy are used as measures of average traffic states.

The measurement duration of the explanatory variables was determined based on their temporal characteristics: an ideal duration would be long enough to capture at least one oscillation cycle but short enough to preclude changes in longer-term average traffic states. We determined the ideal sampling duration using oblique curves of cumulative time-mean speed since time-series plots of count, occupancy and speed typically display a large amount of noise, making it difficult to identify different traffic states and the times of state changes. However, cumulative curves of traffic observations constructed on an oblique time axis can smooth the noise in the data and better reveal the underlying trends. This technique is described in detail in Munoz and Daganzo (*2002*).

On an orthogonal axis, an oblique curve is obtained by taking the difference between the cumulative speed at time t, $V(t)$, and its background reduction, $V_0 (t-t_0)$, where V_0 is a scaling factor and t_0 is the starting time of the curve. Figure 4.1 presents an oblique curve of cumulative speed, $V(x, t) – V_0 (t-t_0)$, versus time, t, at Jantzen Drive between 4:30 and 5:30 p.m. on May 14, 2007. The figure shows a series of oscillations with a cyclic pattern of an increase in slope (i.e., speed) followed by a decrease. Judging from the figure, oscillations typically have a period of about 10 minutes; we use this as the sampling duration for our study.

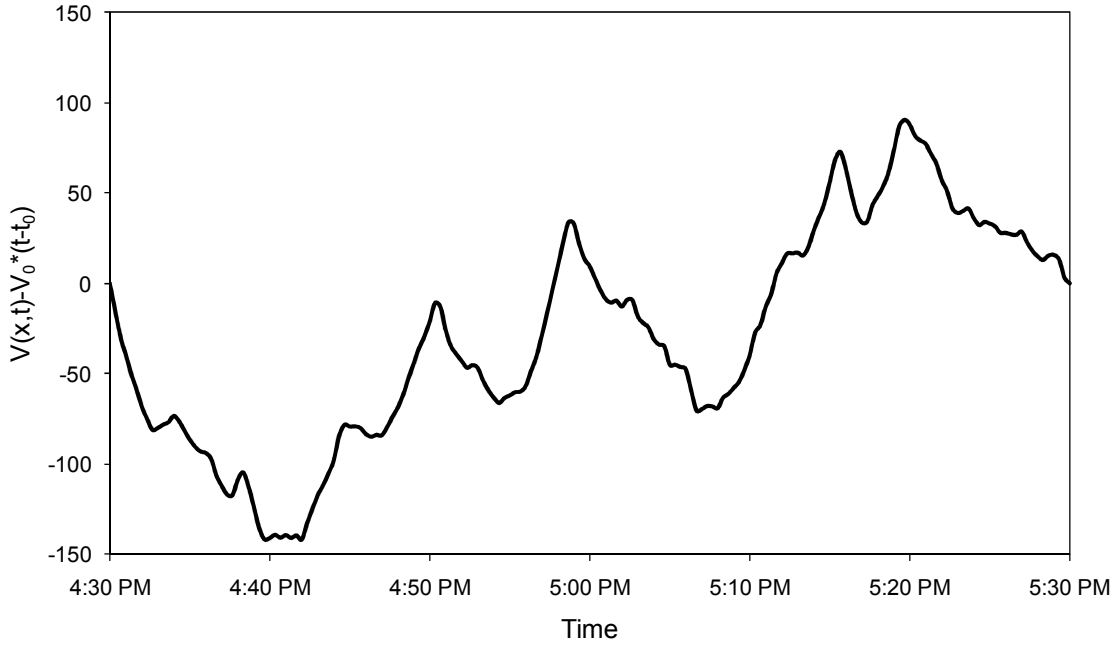


Figure 4.1: Oblique curve of cumulative speed at Jantzen Drive between 4:30 and 5:30 p.m. on May 14, 2007

In the literature, it has been reported that a crash occurrence time is often rounded to the nearest five minutes (*Golob and Recker 2003; Kockelman and Ma 2007*). In such cases, traffic data within five minutes before each crash should be excluded from analysis to avoid "cause and effect" ambiguity. We inspected our original dataset to assess if the crash occurrence times were rounded to the nearest five minutes in Oregon. Our analysis shows that the majority (68 out of 82) of the crash occurrence times are not in multiples of five minutes. Therefore, we believe that the crash occurrence times used in this research were not rounded to the nearest five minutes and thus did not further exclude any data.

The dependent variable is binary (crash-prone or not-crash-prone traffic conditions). Of note, cases here are not the crashes themselves, but the traffic conditions immediately before the crash occurrences while controls are congested traffic conditions under which no crash has occurred.

To control for the potential impacts of other factors on crash occurrences, several confounding variables (time of day, location, weather, and presence of congestion) are used to limit the selection of controls. Of note, each location is treated as a unique geometric factor to be conservative. For instance, a crash occurred in the influence area of the Broadway Street detector station at 4:11 p.m. on the rainy day of October 6, 2005. Measurements taken from 4:01 to 4:11 p.m. on this day are used as a case. Then, measurements are taken at the same location during the same period on the crash-free days in 2004 to 2007 that exhibited congestion (average speed less than 50 mph (81 km/h)) and similar weather conditions. These extracted measurements are used as potential candidates for controls. Among the candidates, n observations are randomly selected as the controls for that particular crash (case). In our study, about 1,000 candidate controls are available for each case. All legitimate controls were stored for the model evaluation, as discussed in Chapter 6.

The above procedure is applied to each crash, resulting in the 82 cases matched with $82 \times n$ controls. It is difficult to theoretically determine the optimal control-to-case ratio. As a rule of thumb, a control-to-case ratio around 4:1 is recommended since the statistical power generally does not increase significantly beyond a 4:1 ratio (*Ahrens and Pigeot 2005; Hennekens and Buring 1987; Rothman and Greenland 1998; Schlesselman and Stolley 1982*). Therefore, the control-to-case ratio of 4:1 was implemented in this study.

Our modeling efforts essentially consist of two main phases: 1) model development using a set of sampled controls and 2) model evaluation via repeated model developments using different samples of controls and sensitivity analysis with respect to the control-to-case ratio (4:1 to 7:1). The latter is designed to evaluate the consistency of modeling results with respect to multiple (unique) control samples and control-to-case ratios. These efforts are presented in detail in chapters 5 and 6.

5.0 MODELING EFFORT AND RESULT

Conditional logistic regression is employed to model the association between traffic oscillations and crash occurrences in the study segment. This is a standard technique for analyzing matched case-control data. Conditional inference can be adopted easily in the conventional (unconditional) logistic regression model if the constant for each stratum is excluded in the estimation by treating it as a "nuisance" parameter (*Hosmer and Lemeshow 2004*). Estimates from a conditional logistic regression have been proven to be consistent and asymptotically normally distributed. Cox and Hinkley (*1974*) provide the mathematical details of conditional likelihood analysis.

Conditional logistic regression analysis is implemented in Stata®/IC10 (*StataCorp 2009*) by taking average traffic states and traffic oscillations as the independent variables. It is well acknowledged in the literature that the control-to-case ratio should be around 4:1, as cited in the previous section, since 80 percent of the maximum efficiency can be obtained around that ratio. Table 5.1 lists all the potential explanatory variables and their basic statistics (the 4:1 control-to-case ratio is taken here as an example). The average speed is about 25 mph (40 km/h), indicating heavy congestion. The mean amplitude of traffic oscillations in terms of standard deviation of speed is 5 mph (8 km/h). The standard deviation of oscillations (also in terms of standard deviation of speed) is 3 mph (5 km/h), ranging from 0.3 mph (0.5 km/h) to nearly 20 mph (32 km/h), confirming that the samples (a total of 82 cases and 328 controls) contain variations in traffic oscillations.

Table 5.1: Basic statistics for the potential explanatory variables (4:1 control-to-case ratio)

Potential Variable	Mean	Std. Dev.	Min	25% Percentile	75% Percentile	Max
Average speed[b]	24.743	11.237	2.303	16.542	30.448	49.938
Average count[c]	5.294	1.656	0.703	4.072	6.444	9.559
Average occupancy[d]	20.955	14.913	0.486	4.968	33.401	64.294
Std. dev. of speed[e]	4.982	3.057	0.327	2.928	6.015	19.957
Std. dev. of count	0.894	0.402	0.0993	0.595	1.098	2.460
Std. dev. of occupancy	4.364	3.829	0.380	1.233	6.386	30.609

[b] the average speed across all travel lanes at all detector stations during the 10-minute periods right before the crash occurrences

[c] the average number of vehicles per lane per 20 seconds at all detector stations during the 10-minute periods right before the crash occurrences

[d] the average percentage of time that the detectors are occupied by vehicles across all travel lanes during the 10-minute periods right before the crash occurrences

[e] the standard deviation of speed during the 10-minute periods right before the crash occurrences; and std. dev. of count and std. dev. of occupancy are similarly defined.

We followed a sound model development practice. Namely, different combinations of potential explanatory factors were incorporated via conditional logistic regression, following a univariate analysis. Since a logistic regression generally requires a large sample size (*Hosmer and Lemeshow 2004; Good and Hardin 2006*) (and we have 82 cases), no more than two explanatory variables are considered simultaneously in the model and interactions between the variables are

also not considered. The performance of each model was evaluated based on the theoretical soundness (e.g., signs of estimated coefficients), the *p*-value for the overall model, R^2 and likelihood ratio tests, rather than the statistical significance of each estimated parameter. Our threshold *p*-value for the overall model performance is set to 0.1; models with overall *p*-values less than 0.1 were considered as candidates for the best model. Finally, a model with the best performance was selected. Of note, during our model development (in this section) and evaluation procedure (in Chapter 6), we emphasized replication/consistency rather than a statistical significance test. We did this because there are many criticisms (Good and Hardin 2006; Johnson 1999; Sterne and Smith 2001) of overemphasizing significance tests, such as the *p*-values of individual parameters and (Pseudo) R^2, to evaluate models, as these tests can be arbitrarily manipulated (by changing the sample size or adding unnecessary variables, for example) and are difficult to interpret correctly.

Table 5.2 shows the statistics of the best model for a set of sampled controls with 4:1 control-to-case ratio. The result shows that the odds ratio for the standard deviation of speed is about 1.096, with a 95 percent confidence interval of 1.016 to 1.183. This implies that an additional standard deviation unit of speed (1 mph in this case) increases the odds ratio of crash occurrence by 9.6 percent..

Table 5.2: Results from the conditional logistic regression analysis (4:1 control-to-case ratio)

Crash_prone	Odds Ratio	Std. Err.	Z	P>|z|	95% Confidence Interval (CI)	
Std. dev. of speed[f]	1.096	.0424	2.37	0.02	1.016	1.183

[f] Number of observations=410; LR chi2(1) = 5.37; Prob > chi2 =0.0205; Log likelihood = -129.29096.

Since the independent variables in our model are continuous, the linearity assumption in the logistic model has been tested using fractional polynomials (FP), as recommended by Hosmer and Lemeshow (*2004*). The fractional polynomials allow us to build several models with different power functions of a continuous explanatory variable and then to choose the most suitable one by conducting the partial likelihood ratio test.

For illustration purposes, we use the standard deviation of speed to explain the fractional polynomial analysis used in our research. For our model, a fractional polynomial of degree m for the standard deviation of speed, $FP_m\{std.dev.of\ speed; (p_1, \dots, p_m)\}$, is defined as

$$FP_m\{std.dev.of\ speed; (p_1, \dots, p_m)\} = \beta_0 + \beta_1(std.dev.of\ speed)^{p_1} + \cdots + \beta_m(std.dev.of\ speed)^{p_m}$$

Where $p_1 \le p_2 \le \cdots \le p_m$ denotes powers (integer or fractional); and

$$(std.sev.of\ speed)^{(p_j)} = \begin{cases} \ln(std.dev.of\ speed) & if\ p_j = 0 \\ (std.dev.of\ speed)^{p_j} & if\ p_j \neq 0 \end{cases}$$

For a given degree (e.g., m), the best fitting powers $p_1, \dots, p_m$ are obtained by choosing the model with the smallest deviance. Royston and Altman (*1994*) provide more technical details of the fractional polynomial analysis.

The simplest functional form of log-relative ratio in the standard deviation of speed is linear, which corresponds to an FP of first degree ($m = 1$) with a power of one ($p_1 = 1$). We have tested a second-degree (see Table 5.3 for the result) and a first-degree (see Table 5.4) fractional polynomials. Higher-degree fractional polynomials are not considered to keep our model parsimonious.

Table 5.3: Summary of the second-degree FP comparisons for standard deviation of speed

Degree of FP, m	Powers	Degree of Freedom	Deviance	Dev. Dif.	P[g]
0 (Not in model)	N/A	0	263.948	10.026	0.04
1	1	1	258.582	4.66	0.198
1	2	2	258.316	4.393	0.111
2	-0.5,0	4	253.922	N/A	N/A

g -P-value from deviance difference comparing reported model with m = 2 model.

Table 5.3 shows that $FP_2\{std.dev.of\ speed; (-0.5,0)\}$ model has the smallest deviance of 253.922 and is better than the model without any variables. However, no significant performance improvement is gained compared to $FP_1\{std.dev.of\ speed; (2)\}$ with deviance 258.316 at the 95 percent confidence level (p-value=0.111).

Similarly, according to Table 5.4, the performance of $FP_1\{std.dev.of\ speed; (2)\}$ model is not significantly better than the performance of the simple linear model at the 95 percent confidence level (p-value=0.606).

Table 5.4: Summary of the first-degree FP comparisons for standard deviation of speed

Degree of FP, m	Powers	Degree of Freedom	Deviance	Dev. Dif.	P[g]
0 (Not in model)	N/A	0	263.948	5.632	0.06
1	1	1	258.582	0.266	0.606
1	2	2	258.316	N/A	N/A

[g] P-value from deviance difference comparing reported model with m = 1 model.

The above analysis indicates that keeping the continuous variables linear in the logistic model is appropriate for the range of speeds we obtained. This procedure has been applied to other variables, which resulted in the same conclusion.

Several techniques (e.g., computing residual variation, leverage) can be used to detect potential outliers in a conditional logistic regression (*Hosmer and Lemeshow 2004*). However, removing outliers can potentially lead to over-fitting, which diminishes the effectiveness of the model (*Washington et al. 2003*). Therefore, we include all data points in our modeling efforts and evaluate the effectiveness of the model using different samples of controls, as described in the following section.

6.0 MODEL EVALUATION AND INTERPRETATION

Evaluation of statistical models for consistency prior to interpretation is essential since they are often subject to measurement errors, selection bias, invalid design, and/or human errors. Ideally, cases and controls should be free of any sources of bias, such as construction, lane closures, and other incidents. This issue was not directly addressed in our analysis given the extensive data collection efforts involved. Instead, we evaluated our model by re-sampling controls for each case and repeating the model development process (described in Section 5) to find the best model based on the newly drawn sample and conducting a sensitivity analysis with respect to the case-to-control ratio. For the former, 20 different sets of controls were sampled from the pool of candidate controls for the control-to-case ratio 4:1. Note that sampling the same control more than once is unlikely since each case has about 1,000 candidate controls. The best model for each re-sampled dataset was obtained according to the criteria described in Chapter 5. Table 6.1 presents the modeling results from the 20 different runs for the control-to-case ratio of 4:1.

The result indicates that in 18 out of 20 runs (90 percent), the standard deviation of speed is significant and its odds ratio is fairly consistent throughout the runs. Notably, the average odds ratio is 1.084 with the minimum of 1.066 and the maximum of 1.119, which is a fairly tight bound. Moreover, the average lower and upper limits of the 95 percent confidence intervals are 1.003 and 1.171, respectively, indicating that the standard deviation of speed is a significant variable associated with the likelihood of a crash occurrence. For the study segment, an additional unit in the standard deviation of speed increased the odds ratio of crash occurrence by an average of 8.4 percent. This indicates that an average day is about 1.49 times more likely to have a crash than the day without any oscillations since the average standard deviation of speed on the study corridor is about 4.982 mph (see Table 5.1). This magnitude of impact seems reasonable since many other factors, such as human factors and vehicle conditions, also impact crash occurrences (*Hauer 1997*).

It is also notable that the average occupancy is significant: nine times in the 20 runs. These nine appearances have an average odds ratio of 1.023. However, the *p*-values are relatively large (>0.1) in most runs, and the average lower limit of the 95 percent confidence interval for its odds ratio is lower than one (0.997). Moreover, the average count appears significant in only one run, with an average odds ratio less than one (0.809). Thus, it is reasonable to conclude that speed variations in queued traffic (oscillations) have a larger impact on crash occurrence than average traffic states. Nevertheless, the average odd ratios for occupancy and count are qualitatively consistent: they both imply that the likelihood of crash occurrence increases as congestion becomes more severe.

Table 6.1: Model evaluation results (4:1 control-to-case ratio)

Run	Variables[h]	Odds Ratio	P-value	95% CI	
1	std_spd	1.066	0.1	0.988	1.151
2	avg_occ	1.03	0.04	1.002	1.058
	std_spd	1.119	<0.01	1.029	1.216
3	std_spd	1.07	0.1	0.988	1.16
4	std_spd	1.096	0.02	1.016	1.183
5	std_spd	1.09	0.03	1.01	1.177
6	std_spd	1.069	0.07	0.994	1.149
7	std_spd	1.077	0.05	1.001	1.16
8	std_spd	1.079	0.03	1.006	1.158
9	avg_occ	1.02	0.09	0.997	1.043
	std_spd	1.071	0.08	0.991	1.157
10	avg_occ	1.025	0.08	0.997	1.054
	std_spd	1.094	0.03	1.01	1.185
11	avg_occ	1.028	0.06	0.999	1.058
	std_spd	1.118	<0.01	1.029	1.215
12	avg_occ	1.019	0.12	0.995	1.042
	std_spd	1.089	0.03	1.008	1.176
13	avg_occ	1.02	0.1	0.996	1.044
	std_spd	1.077	0.07	0.995	1.165
14	std_spd	1.081	0.04	1.002	1.166
15	N/A	N/A	N/A	N/A	N/A
16	avg_cnt	0.809	0.06	0.646	1.013
	std_spd	1.089	0.04	1.005	1.179
17	avg_occ	1.017	0.15	0.994	1.041
	std_spd	1.069	0.07	0.994	1.15
18	avg_occ	1.021	0.13	0.994	1.05
	std_spd	1.081	0.06	0.997	1.172
19	avg_occ	1.026	0.06	0.999	1.053
20	std_spd	1.078	0.06	0.998	1.165
Average	std_spd	1.084	N/A	1.003	1.171

[h] avg_occ is for average occupancy; std_spd is for standard deviation of speed; avg_cnt is for average count, and NA means not available.

Now we turn our attention to the sensitive analysis of the modeling results shown in Table 6.1 with varying control-to-case ratios. If the model results are robust, we expect similar or more consistent results for larger or smaller control-to-case ratios since the statistical power is expected to increase. To verify this, two sampling strategies were adopted.

First, we randomly re-sampled totally different controls for each case in varying control-to-case ratios (3:1 to 5:1). The results for the control-to-case ratios of 3:1 and 5:1 are fairly consistent with the results for the 4:1 ratio (For a brief summary, see Table 6.2; for details, see Appendix A). The standard deviation of speed appears significant in more than 15 runs, and the average odds ratios are consistently around 1.08, which also supports the robustness of our findings. We think that the slight decrease in the number of significant runs (18 to 17) from ratios 4:1 to 5:1 is rather a result of uncertainty in the data than a decrease in statistical power. Therefore, to evaluate the model performance (consistency) with respect to statistical power, a different sampling strategy was developed, as described below.

We conducted an additional analysis by increasing control-to-case ratios from 4:1 to 7:1 with the set of controls used in 4:1 ratio as the base. In other words, additional controls were sampled and added to the controls for the 4:1 ratio. For example, for the 5:1 ratio, one additional control for each case was randomly sampled from the pool of candidates and added to the existing four controls. For each run and each increased control-to-case ratio, a best model was redeveloped based on the criteria described in Chapter 5 to examine the consistency of models in terms of significant explanatory variables, odds ratios, and confidence intervals. The results of this sensitivity analysis are summarized in Table 6.3 for the control-to-case ratios of 4:1 to 7:1. Table 6.3 reports the number of runs in which the standard deviation of speed (i.e., amplitude of oscillations) and average occupancy, speed, and count were significant over 20 different control samples for each control-to-case ratio. The table also reports the average odds ratios and average confidence intervals for the standard deviation of speed.

The model results for the standard deviation of speed are quite consistent for different control samples and, as expected, the consistency improves with increasing ratios. More specifically, the standard deviation of speed is significant in more than 18 runs for all ratios, and the number of significant runs increases from 18 (90 percent of the runs) to 20 (100 percent of the runs) as the ratio increases from 4:1 to 7:1. Moreover, the average odds ratios for the standard deviation of speed are consistently around 1.08 for different control-to-case ratios, and the average 95 percent confidence intervals tighten as the ratio increases. Thus, the results strongly support the existence of relation between oscillations (measured by the standard deviation of speed) and crash occurrences. To the contrary, other variables for average traffic states appeared to be significant in less than 10 runs, and the consistency does not improve with increasing control-to-case ratios. The results imply that traffic oscillations are more significantly associated with crash occurrences than average traffic states. The details of individual runs for each control-to-case ratio are presented in Appendix B.

Table 6.2: Model evaluation results for different control-to-case ratios (3:1 to 5:1)

	Standard Deviation of Speed			Number of Significant Runs out of 20[i]			
Control-to-Case Ratio	Average Odds Ratio	Average 95% CI		std_spd	avg_occ	avg_spd	avg_cnt
3:1	1.086	1.001	1.178	15	4	1	2
4:1	1.084	1.003	1.171	18	9	0	1
5:1	1.083	1.004	1.169	17	6	3	4

[i] avg_occ is average occupancy; std_spd is standard deviation of speed; avg_cnt is average count, and avg_spd is average speed.

Table 6.3 Summary of model evaluation results for different control-to-case ratios (4:1 to 7:1)

	Standard Deviation of Speed			Number of Significant Runs out of 20[j]			
Control-to-Case Ratio	Average Odds Ratio	Average 95% CI		std_spd	avg_occ	avg_spd	avg_cnt
4:1	1.084	1.003	1.171	18	9	0	1
5:1	1.080	1.002	1.165	18	4	1	1
6:1	1.080	1.003	1.163	19	8	1	1
7:1	1.081	1.005	1.162	20	5	3	0

[j] avg_occ is average occupancy; std_spd is standard deviation of speed; avg_cnt is average count, and avg_spd is average speed.

7.0 CONCLUSIONS AND DISCUSSION

The present study examines the impact on crash occurrences of freeway traffic oscillations that arise in queued traffic. This research employs a 12-mile stretch of freeway in Portland, OR, where fairly extensive recurrent congestion and oscillatory flow are observed. Traffic conditions, including oscillations, were measured using 20-second aggregated data from inductive loop detectors. Crash data for the study corridor are available from the statewide crash database, and crash occurrence times were obtained or estimated from the incident database and time-series traffic data.

A case-control study was designed using these data. A case was a traffic condition prior to a crash, and the controls were the conditions on days without any crashes matched for the confounding variables of time of day, presence of congestion, weather and geometry. The averages and standard deviations of count, occupancy and speed during the 10 minutes prior to a crash were adopted as potential measures of average traffic states and traffic oscillations, respectively. A total of 82 cases and around 80,000 candidate controls were extracted from data from 2004-07.

Models were developed based on a conditional logistic regression with two exposure variables, average traffic state and amplitude of oscillations. Of note, models were initially developed for the control-to-case ratio of 4:1. The model results were evaluated by randomly drawing 20 different sets of controls. To further assess the model's consistency for different control-to-case ratios, two sampling strategies were adopted (one sampling strategy for 3:1 to 5:1 control-to-case ratios and the other for 4:1 to 7:1 control-to-case ratios). The evaluation results show that the amplitude of oscillations was consistently significant for different control samples and control-to-case ratios. The evaluation results also displayed consistent average odds ratios (of about 1.08) and their confidence intervals. The consistency in the results demonstrates that oscillations have a significant impact on crash occurrence. An additional unit increase in the standard deviation of speed increases the odds of (rear-end) crashes by about 8 percent.

Of further note, in this study the average traffic states in congestion were less significant than deviations in speed. For example, in the control-to-case ratio of 4:1, the average occupancy appeared to be significant in less than half of the runs. Nevertheless, its odds ratio was qualitatively consistent and suggests that the likelihood of crash occurrence increases as congestion becomes more severe.

Our findings are notable given that the impact of speed variation on crash occurrence has been debated for decades. This study addresses the shortcomings of existing studies by using high-resolution traffic and crash data and adopting the case-control design proven to be effective in the field of epidemiology. Moreover, the present study elucidates the impact of oscillatory driving in queued traffic on safety, which is a rare contribution. Given that oscillations are becoming more common in everyday traffic, the findings from this study may help prioritize countermeasures, such as ramp metering or adaptive speed control, to improve traffic safety and estimate their expected benefits. Nevertheless,

future investigations are warranted to acquire a more complete understanding of the safety impact of oscillations, as elaborated below.

Although a sufficient number of samples (82 crashes) were obtained to study the impact of oscillations, the sample size in this study was substantially reduced (from 242 to 82) due to missing traffic data and the inability to accurately estimate crash occurrence times. As a result, our modeling efforts were limited to one or two exposure variables at a time, and interactions among the variables were not considered. This may be addressed by employing the Bayesian approach, and work in this regard is ongoing.

Furthermore, a single freeway corridor was selected in this study. Other freeway locations should be analyzed to confirm the current findings. It is likely that the impact of oscillations depends on various characteristics of roadways, such as traits of the driving population and the freeway geometry. Future investigations in this regard are necessary. Nevertheless, the present study provides a methodological framework for further investigations.

REFERENCES

Ahn, S. and Cassidy, M.J. (2007). Freeway Traffic Oscillations and Vehicle Lane-Change Maneuvers. The 17th International Symposium of Transportation and Traffic Theory, London. Elsevier.

Ahrens, W. and Pigeot, I. (2005). Handbook of Epidemiology. Springer.

Abdel-Aty, M., Pande, A., Uddin, N., Dilmore, J., Pemmanaboina, R. (2005). Relating Crash Occurrence to Freeway Loop Detectors Data, Weather Conditions and Geometric Factors. University of Central Florida.

Bilbao-Ubillos, J. (2008). The costs of urban congestion: Estimation of welfare losses arising from congestion on cross-town link roads. Transportation Research Part A 42, 1098-1108.

Breslow, N. E., and Day, N. E. (1980). Statistical Methods in Cancer Research. Volume I - the Analysis of Case-Control Studies. IARC Sci. Publ.,32 (32), 5-338.

Cirillo, J. A. (1968). Interstate System Accident Research: Study II, Interim Report II. 35(3), 71-75.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., Wynder, E. L. (1959). Smoking and Lung Cancer: Recent Evidence and a Discussion of some Questions. J. Natl. Cancer Inst., 22(1), 173-203.

Cornfield, J. (1951). A Method of Estimating Comparative Rates from Clinical Data; Applications to Cancer of the Lung, Breast, and Cervix. J. Natl. Cancer Inst., 11(6), 1269-1275.

Cox, D. R., and Hinkley, D. (1974). Theoretical Statistics, Chapman & Hall/CRC .

Davis, G. A., Davuluri, S., Pei, J. (2006). Speed as a Risk Factor in Serious Run-Off-Road Crashes: Bayesian Case-Control Analysis with Case Speed Uncertainty. Journal of Transportation and Statistics, 9(1), 17.

Davis, G. A. (2002). Is the Claim that 'Variance Kills' an Ecological Fallacy? Accid. Anal. Prev., 34(3), 343-346.

Fildes, B., and Rumbold, G. (1991). Speed Behaviour and Drivers' Attitude to Speeding, Monash University.

Fisher, R. A. (1958a). Lung Cancer and Cigarettes. Nature, 182(4628), 108.

Fisher, R. A. (1958b). Cancer and Smoking. Nature, 182(4635), 596.

Fowles, R., and Loeb, P. D. (1989). Speeding, Coordination, and the 55-MPH Limit: Comment. Am. Econ. Rev., 79(4), 916-921.

Golob, T. F., and Recker, W. W. (2003). Relationships among Urban Freeway Accidents, Traffic Flow, Weather, and Lighting Conditions. J. Transp. Eng., 129, 342.

Golob, T. F., and Recker, W. W. (2004). A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways. Transportation Research Part A: Policy and Practice, 38(1), 53-80.

Good, P. I., and Hardin, J. W. (2006). Common Errors in Statistics (and how to Avoid them). Wiley-Interscience.

Greenwood, I.D., Bennett, C.R. (1996). The effects of traffic congestion on fuel consumption. Road and Transport Research, 5, 18-35, the Australian Road Research Board.

Gross, F., and Jovanis, P.P. (2007). Estimation of the safety effectiveness of lane and shoulder width: Case-control approach. Journal of Transportation Engineering, 133(6), 362-369.

Hauer, E. (1997). Observational before-After Studies in Road Safety. New York, NY: Elsevier.

Hennekens, C. H., and Buring, J. E. (1987). Epidemiology in Medicine, Lippincott Williams & Wilkins.

Hosmer, D. W., and Lemeshow, S. (2004). Applied Logistic Regression, Wiley-Interscience.

Johnson, D. H. (1999). The Insignificance of Statistical Significance Testing. The Journal of Wildlife Management, 763-772.

Kloeden, C., A. McLean, V. Moore, and G. Ponte (1997). Travelling Speed and the Risk of Crash Involvement, NHMRC Road Accident Research Unit, University of Adelaide,Adelaide, Australia.

Kockelman, K. M., and Ma, J. (2007). Freeway Speeds and Speed Variations Preceding Crashes, within and Across Lanes. Journal of the Transportation Research Forum, 46(1), 43-62.

Lave, C. (1989). Speeding, Coordination, and the 55-MPH Limit: Reply. Am. Econ. Rev., 79(4), 926-931.

Lave, C. A. (1985). Speeding, Coordination, and the 55 MPH Limit. Am. Econ. Rev., 75(5), 1159-1164.

Levy, D. T., and Asch, P. (1989). Speeding, Coordination, and the 55-MPH Limit: Comment. Am. Econ. Rev., 79(4), 913-915.

Malyshkina, N., and Mannering, F. L. (2008). Analysis of effect of speed-limit increases on accident causation and injury severity. Proc. of TRB 87th Annual Meeting Compendium of Papers DVD.

Manski, C. F. (1995). Identification Problems in the Social Sciences, Harvard University Press.

Mauch, M. and Cassidy, M.J. (2002). Freeway traffic oscillations: observations and predictions. The 15th Int. Symp. on Transportation and Traffic Theory, Pergamon-Elsevier, Oxford, UK.

Munoz, J.C., and Daganzo, C. F. (2002). Fingerprinting Traffic from Static Freeway Sensors. Cooperative Transportation Dynamics 1, 1.1–1.11.

Noland, R. B., and Quddus, M. A. (2005). Congestion and Safety: A Spatial Analysis of London. Transportation Research Part A: Policy and Practice, 39(7-9), 737-754.

PORTAL. (2009). Portland Oregon Regional Transportation Archive Listing.<http://portal.its.pdx.edu/> (07/01, 2008).

Research Triangle Institute. (1970). Speed and Accidents: Volume I, US Department of Transportation, Washington, DC.

Rothman, K.J. and Greenland, S. (1998). Modern Epidemiology. Lippincott- Raven, Philadelphia.

Royston, P., and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling (with discussion). Applied Statistics, 43, 429-467.

Schlesselman, J. J., and Stolley, P. D. (1982). Case-Control Studies: Design, Conduct, Analysis. Oxford University Press.

Shefer, D. (1997). Congestion and Safety on Highways: Towards an Analytical Model. Urban Stud., 34(4), 679-692.

Solomon D. (1964). Accidents on Main Rural Highways Related to Speed, Driver and Vehicle, US Department of Commerce & Bureau of Public Roads, Washington, DC.

StataCorp. (2009). Stata:Data analysis and statistical software.<http://www.stata.com/> (04/01, 2009).

Sterne, J. A. C., and Smith, G. D. (2001). Sifting the Evidence--what's Wrong with Significance Tests? Phys. Ther., 81(8), 1464.

Synder, D. (1989). Speeding, Coordination, and the 55-MPH Limit: Comment. Am. Econ. Rev., 79(4), 922-925.

Tilden, C. J., Morris, D. L., Martin, T. M. C., Russell, E. W.(1936). Motor Vehicle Speeds on Connecticut Highways, Committee on transportation, Yale University.

Wang, C., Quddus, M.A., Ison, S.G.(2009). Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. Accid. Anal. Prev.,41(4),798-808.

Washington, S., Karlaftis, M., Mannering, F. (2003). Statistical and Econometric Methods for Transportation Data Analysis, Chapman and Hall/CRC Press.

Xin, W., Hourdos, J., Michalopoulos, P. G. (2008). Vehicle trajectory collection and processing methodology and its implementation. Proc., TRB 87th Annual Meeting Compendium of Papers DVD.

APPENDIX A:

Model evaluation results for control-to-case ratios 3:1 and 5:1

Model evaluation results for control-to-case ratios 3:1 and 5:1 from the first sampling strategy (a totally different set of controls was randomly sampled) are summarized in the tables below.

Table A.1: Model evaluation results (3:1 and 5:1 control-to-case ratios)

Control-to-case ratio		3:1			5:1		
Run	Variables	Odds Ratio[k]	95% CI		Odds Ratio	95% CI	
1	std_spd	1.087	0.999	1.181	1.074	0.995	1.158
2	avg_spd				0.979	1.013	1.191
2	std_spd				1.098	0.997	1.043
3	std_spd	1.115	1.021	1.217	1.09	1.009	1.176
3	avg_occ				1.02	0.997	1.043
4	avg_spd	0.971	0.946	0.997	0.976	0.954	0.999
4	std_spd	1.111	1.011	1.22	1.071	0.993	1.155
5	avg_cnt	0.811	0.634	1.039			
5	std_spd	1.084	0.997	1.178	1.062	0.99	1.14
6	avg_occ	1.021	0.996	1.046			
6	std_spd	1.123	1.032	1.223	1.088	1.007	1.175
6	avg_cnt				0.825	0.664	1.024
7	avg_occ	1.025	0.994	1.057			
7	std_spd	1.075	0.998	1.157	1.088	1.009	1.173
7	avg_cnt				0.81	0.656	1.001
8	std_spd						
9	std_spd	1.071	0.993	1.155	1.064	0.986	1.147
9	avg_cnt				0.811	0.642	1.023
10	std_spd	1.085	1	1.176	1.109	1.026	1.198
10	avg_spd				0.977	0.955	1
11	std_spd	1.066	0.992	1.146	1.122	1.037	1.213
11	avg_occ				1.035	1.006	1.064
12	avg_occ	1.021	0.997	1.047	1.025	1	1.051
12	std_spd	1.067	0.988	1.153	1.078	0.997	1.165
13	std_spd				1.073	0.995	1.157
14	std_spd				1.064	0.987	1.146
14	avg_cnt				0.794	0.622	1.013
15	avg_occ	1.033	1.004	1.064	1.02	0.997	1.044
15	std_spd	1.096	1.005	1.194	1.106	1.026	1.192

[k] The blank cell means that the corresponding variable did not appear in the final model.

16	avg_cnt	0.819	0.648	1.035			
16	std_spd	1.071	0.986	1.164	1.097	1.016	1.186
16	avg_occ				1.023	0.996	1.05
17	std_spd	1.071	0.991	1.157			
18	std_spd				1.068	0.989	1.152
18	avg_occ				1.018	0.996	1.041
19	std_spd	1.099	1.012	1.194			
20	std_spd	1.069	0.992	1.153	1.066	0.99	1.148

APPENDIX B:

Model evaluation results for 5:1 to 7:1 control-to-case ratios

Model evaluation results for 5:1 to 7:1 control-to-case ratios from the second sampling strategy (additional controls were sampled and added to the controls for the 4:1 ratio to form a new set of controls) are summarized in the tables below.

Table B.1 Model evaluation results (5:1 to 7:1 control-to-case ratios)

control-to-case ratio		5:1		6:1		7:1	
Run	variables	odds ratio[1]	95% CI	odds ratio	95% CI	odds ratio	95% CI
1	std_spd	1.101	(1.020,1.190)	1.093	(1.018,1.172)	1.084	(1.008,1.165)
2	std_spd	1.062	(0.986,1.143)	1.062	(0.987,1.144)	1.08	(1.002,1.165)
3	std_spd	1.081	(1.003,1.165)	1.116	(1.033,1.206)	1.092	(1.016,1.174)
3	avg_occ			1.027	(1.000,1.054)	1.018	(0.997,1.040)
4	std_spd	1.079	(0.997,1.168)	1.075	(0.997,1.158)	1.093	(1.011,1.181)
5	std_spd	1.090	(1.013,1.174)	1.077	(1.002,1.157)	1.088	(1.012,1.170)
6	std_spd	1.069	(0.992,1.150)	1.080	(1.009,1.157)	1.084	(1.012,1.161)
7	std_spd	1.080	(1.005,1.161)	1.077	(1.002,1.157)	1.075	(1.002,1.153)
8	std_spd	1.076	(1.002,1.154)	1.085	(1.012,1.165)	1.066	(0.996,1.140)
9	std_spd	1.071	(0.994,1.154)	1.080	(1.001,1.165)	1.069	(0.994,1.150)
9	avg_occ	1.018	(0.996,1.041)	1.018	(0.995,1.042)		
10	std_spd	1.089	(1.009,1.176)	1.088	(1.006,1.177)	1.082	(1.005,1.164)
10	avg_occ			1.024	(0.999,1.051)		
11	std_spd	1.121	(1.034,1.215)	1.078	(1.002,1.16)	1.085	(1.011,1.164)
11	avg_occ	1.026	(0.999,1.054)				
12	std_spd	1.081	(1.005,1.161)	1.074	(0.999,1.155)	1.085	(1.009,1.167)
12	avg_occ			1.017	(0.997,1.038)	1.02	(0.998,1.043)
13	std_spd	1.066	(0.991,1.147)	1.072	(0.991,1.161)	1.068	(0.994,1.147)
13	avg_occ					1.020	(0.997,1.044)
13	avg_spd			0.981	(0.958,1.004)		
14	std_spd	1.066	(0.990,1.147)	1.074	(0.998,1.156)	1.062	(0.991,1.137)
15	std_spd			1.071	(0.994,1.154)	1.080	(1.001,1.164)
15	avg_occ			1.020	(0.995,1.044)	1.028	(1.000,1.056)
16	std_spd	1.096	(1.014,1.185)	1.078	(0.997,1.166)	1.107	(1.019,1.202)
16	avg_cnt	0.804	(0.645,1.003)	0.822	(0.662,1.019)		
16	avg_spd					0.978	(0.956,1.001)
17	std_spd	1.070	(0.994,1.151)	1.070	(0.996,1.149)	1.087	(1.010,1.170)
17	avg_spd	0.980	(0.958,1.003)			0.981	(0.959,1.004)
17	avg_occ			1.019	(0.996,1.043)		

[1] the blank cell means that the corresponding variable did not appear in the final model.

18	std_spd	1.083	(1.001,1.173)	1.074	(0.996,1.16)	1.077	(0.997,1.162)
18	avg_occ	1.024	(0.997,1.053)	1.021	(0.995,1.05)		
18	avg_spd					0.982	(0.960,1.005)
19	std_spd					1.071	(0.998,1.150)
19	avg_occ	1.022	(0.997,1.048)	1.020	(0.996,1.044)	1.023	(0.998,1.049)
20	std_spd	1.067	(0.989,1.150)	1.093	(1.013,1.178)	1.075	(1.002,1.153)

P.O. Box 751
Portland, OR 97207

OTREC is dedicated to stimulating and conducting collaborative multi-disciplinary research on multi-modal surface transportation issues, educating a diverse array of current practitioners and future leaders in the transportation field, and encouraging implementation of relevant research results.