This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

Impact of Traffic Oscillations on Freeway Crash Occurrences

By Zuduo Zheng[a], Soyoung Ahn[a,*], Christopher M. Monsere[b]

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 paper 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 for varying control-case ratios. 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 likelihood of a (rear-end) crash 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.

Keywords: Traffic oscillations, Stop-and-go driving, Crash, Matched case-control design, Conditional logistic regression

[*] Corresponding author. Tel.: +1 480 965 1052; Fax: +1 480 965 0557; E-mail address: sue.ahn@asu.edu

[a] Department of Civil, Environmental and Sustainable Engineering, Arizona State University, Tempe, AZ 85281, USA

[b] Department of Civil and Environmental Engineering, Portland State University, P.O. Box 751, Portland, OR 97207-0751

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

1. 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 traffic condition 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 the presence of congestion. The exposures of interest are the average traffic state (e.g., average speed) and magnitude of oscillations (e.g., speed variations). 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 manuscript is organized as follows. The following section discusses previous research related to the present study. Section 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 Sections 4 and 5, respectively. Section 6 evaluates and interprets the models, while Section 7 offers concluding remarks and suggestions for future research.

2. Background

Three types of relevant literature are reviewed. We first discuss previous studies that examined the impacts of speed variance and/or congestion on traffic safety. The review revealed that contradictory results have been reported primarily due to limitations in data resolution and analysis methodology. 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 speed. However, his findings received little attention until Lave (1985) claimed in a controversial paper that once one controls for speed variations, average speed has little impact on highway safety. 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). Moreover, 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, UK, 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, UK. 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.

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

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 the speeds of selected drivers at a particular location and then analyzing their crash histories. This approach has been criticized for selection bias in the data collection as well as privacy issues (Kloeden et al. 1997).

Finally, some traffic safety studies borrow the case-control design pioneered in the field of epidemiology (Abdel-Aty et al. 2005; Kloeden et al. 1997). 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, crash) 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.

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 at the crash scene or in similar situations (but not involved in a crash) are taken as controls. The first notable case-control study in the traffic literature is Kloeden et al. (1997), which examined the relationship between speed in uninterrupted traffic flow and the risk of a casualty crash. More recently, Davis et al. (2006) used a case-control design to investigate the effect of speed in run-off-road crashes.

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 5-minute aggregated data from inductive loop detectors to develop a logistic model for multi-vehicle crashes. However, they did not distinguish un-congested and congested traffic regime and 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. Study site and data processing

The study site was a 12-mile (19-km) section (mileposts 296.26 to 307.9) in the northbound direction on I-5 in Portland, OR (see Figure 1 for a schematic). This location was selected among other freeway corridors in the region based on several criteria: extent of congestion and oscillations, spacing of loop detector stations, traffic data quality, crash data availability, and crash sample size. 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 to 6 p.m. on weekdays. Spacings between loop detector stations range 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 Portland Oregon Regional Transportation Archive Listing (PORTAL) (2009).

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

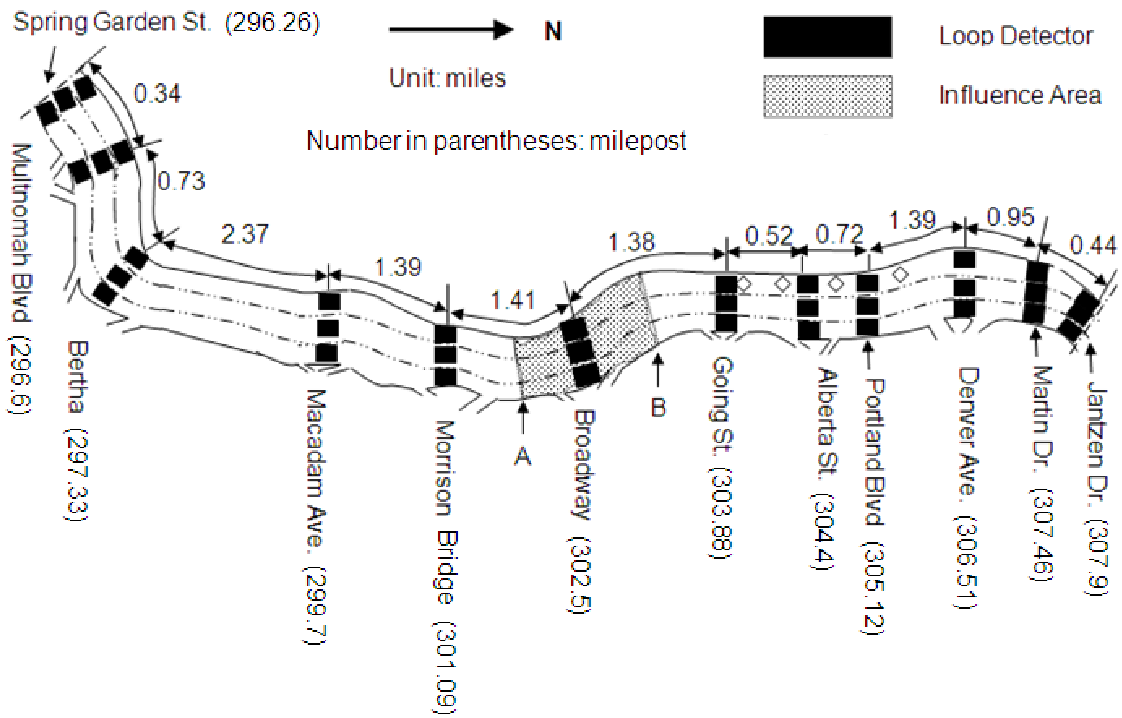


Figure 1 Schematic of the study site, northbound I-5, Portland, OR

Loop detectors (black squares in Figure 1) measure traffic conditions at 12 locations on the selected freeway segment. Notably, we define “influence areas” that are bounded by the midpoints between neighboring stations. The shaded area in Figure 1 is an example of an influence area. *A* corresponds to the midpoint between Broadway St. and Morrison Bridge, and *B* corresponds to the midpoint between Broadway St. and Going St. Table 1 provides more detailed information on the influence areas of the study segment.

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 later in this section.

Table 1 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	Martin Dr.	(306.99 - 307.68)
12	307.9	Jantzen Dr.	(307.68 - 307.9)

The speed contour plot in Figure 2 illustrates the average traffic conditions of the study section during weekdays in January 2006. A distinct congestion pattern, which is marked by the dark time-space regions, occurs during p.m. peak hours (defined as 4:00 – 6:00 p.m.). Congestion starts around 3:00 p.m. and ends around 6:00 p.m. The bottleneck for this congestion is located downstream of milepost 307, and the resulting queue propagates to about milepost 296. Moreover, congestion is present recurrently on this segment such that more than 85 percent of the p.m. peak hours on weekdays exhibit congestion on this freeway stretch.

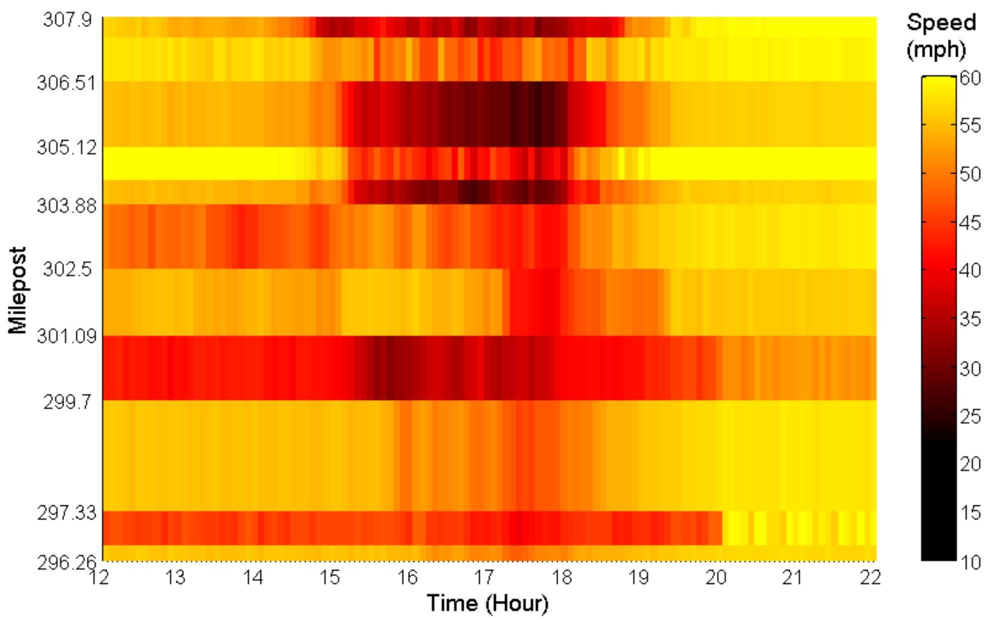


Figure 2 Speed contour for the study segment: Weekdays in January 2006; Time and space (in terms of mileposts) are shown on the x- and y-axes, respectively, and the color scale represents the estimated time-mean speeds over the month according to the legend in the figure.

During the congested periods, traffic oscillations were examined visually using time-series speed plots to confirm their presence and amplitudes. For instance, Figure 3 shows speed over time (4:30 – 5:30 p.m.) at Jantzen Dr. 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 – 5:00 p.m.) than others (e.g., 5:20 – 5:30 p.m.).

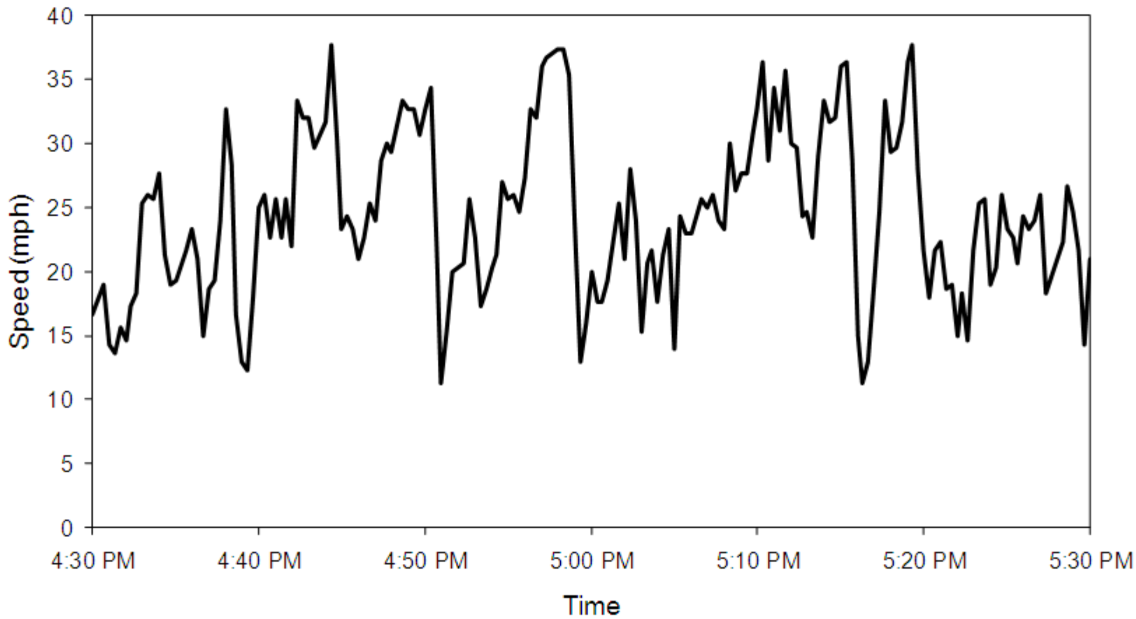


Figure 3 Speed time series plot at milepost 307.9 on May 14, 2007

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. Figure 4 presents monthly percentages of valid traffic data during p.m. peak hours in 2004-2007 for each detector station. The figure indicates that the percentage of valid data is reasonably good (mostly larger than 90 percent) for the selected study segment. Note that days with invalid traffic data prior to crashes were excluded in the further analysis.

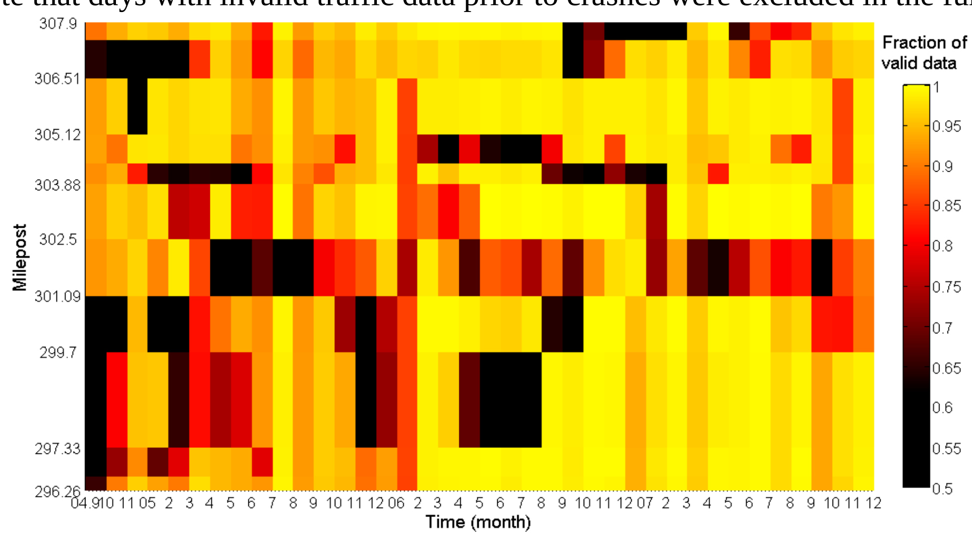
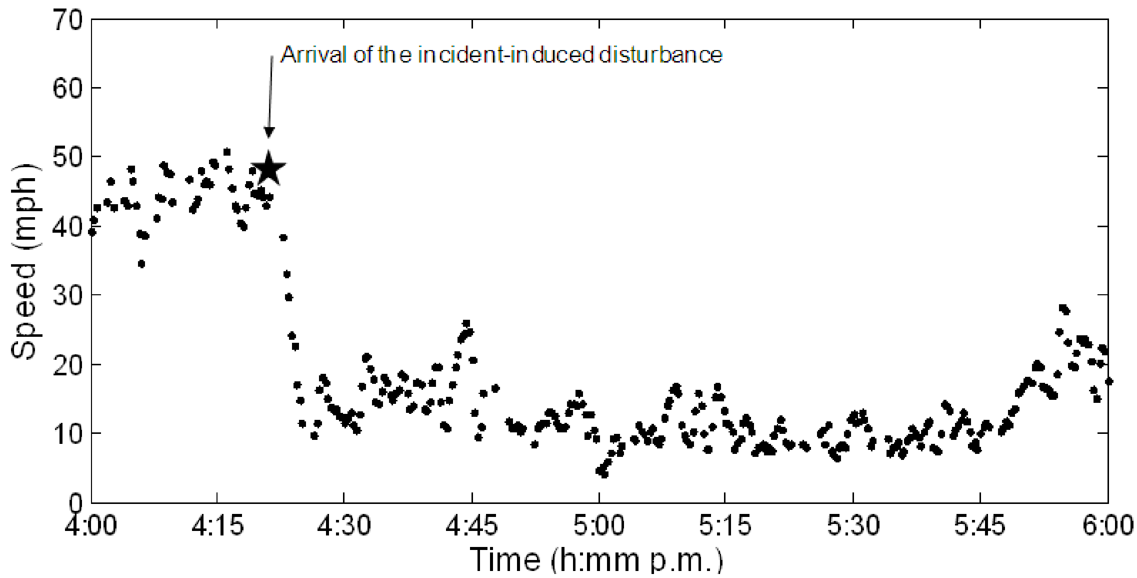


Figure 4 Data quality map of northbound I-5 during p.m. peak hours; the monthly percentages of valid data are shaded according the color scale shown in the legend: darker regions correspond to lower data quality; traffic data for 2004 were available only in September, October, and November.

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

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 Metropolitan area as logged by the Traffic Management and Operations Center (TMOC). 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 were 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 5 demonstrates this assumption using time-series speeds at the detector stations immediately upstream (Macadam Ave.) 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 (Figure 5(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 (Figure 5(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 Ave. (see Table 1), 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).



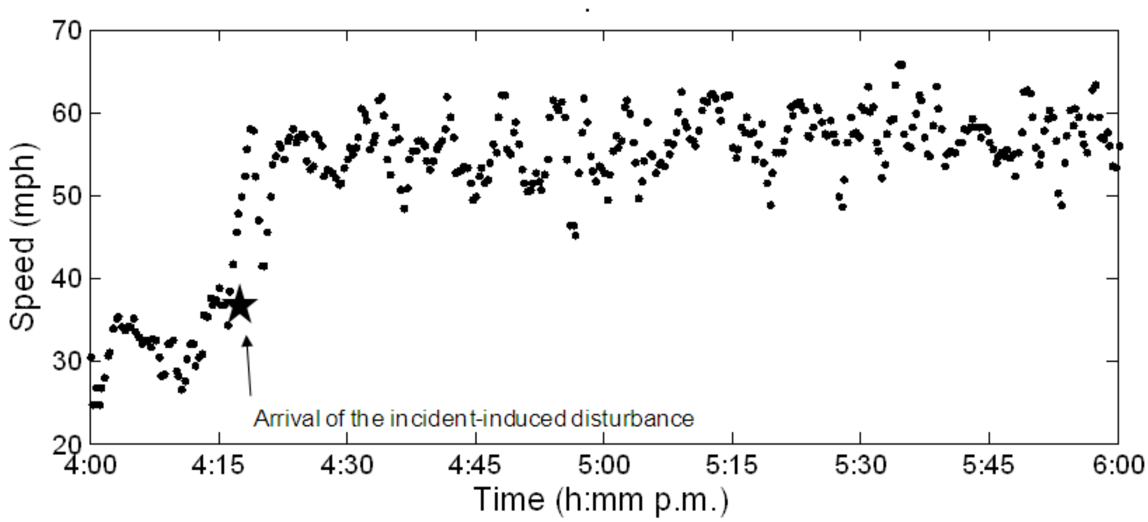


Figure 5 Illustration of crash occurrence time estimation

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.

The majority of the 82 crashes were 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 mis-coded. 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 these crashes can occur due to oscillations, all 82 crashes were included in our analysis[1].

Finally, the 82 crash events were linked to traffic data based on their mileposts and the corresponding influence areas. Of note, we used 1-minute moving averages of traffic measures (e.g., speed) to smooth out noise while preserving the underlying trend. Measurements from all lanes were also combined to incorporate influences from adjacent lanes and maximize the sample size in view of Abdel-Aty et al. (2005)[2]. Weather information (e.g., rainfall) was obtained from PORTAL (2009) and incorporated into the 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

---

[1] 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.

[2] A sample size can be maximized by using measurements taken in other lanes in case of a malfunctioning loop.

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 immediately upstream station 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 6 shows the basic statistics (e.g., average and standard deviations) of oscillations amplitude at each detector station for all crash-free days, as well as the spatial distribution of the identified crashes. The figure shows that the average amplitudes of oscillations are reasonably stable over space, such that they did not deviate more than 0.4 mph within each influence area on average. The largest difference of 0.8 mph is observed at the Multnomah station based on linearly interpolated amplitudes, though only 4 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.

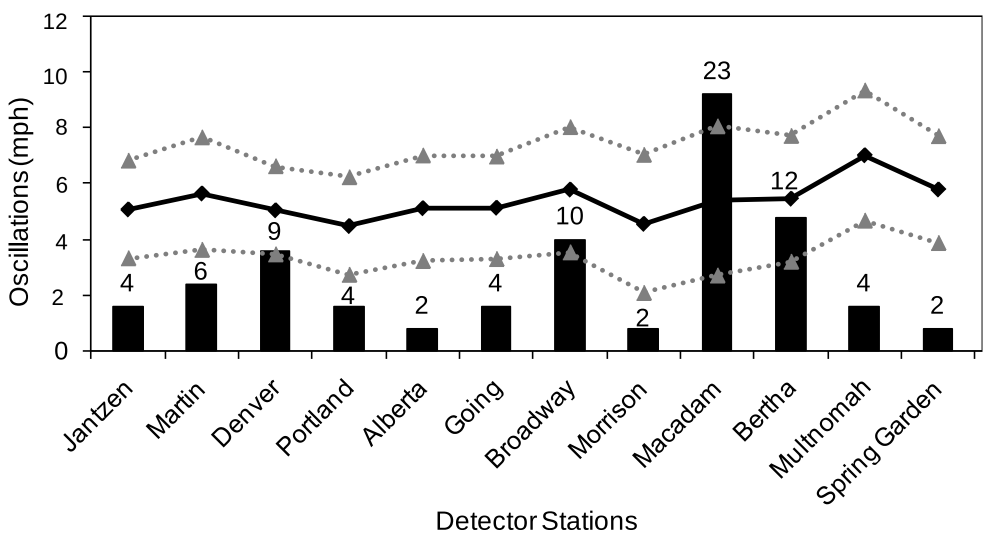


Figure 6 Spatial distributions of oscillations in a ten-minute period from 2004 to 2007 and number of crashes sampled for this study. The bars stand for number of crashes; the solid line is for average oscillations and the dotted lines are variations of oscillations over time bounded by one standard deviation.

4. Methodology: case control design

The matched case-control method was selected to study the relationship between traffic oscillations and crash occurrences in congestion. Many previous studies (mentioned in the Background section) found that crash occurrences are significantly associated with exogenous factors such as traffic conditions (e.g. speed), roadway characteristics (e.g. pavement conditions, speed limits, geometric features), and environmental conditions (e.g., weather). Thus, we have incorporated these factors in our modeling. Since we are interested in the impact of traffic operational features, the factors related to roadway and environmental conditions are controlled in our matched case-control design.

Two types of exposure variables are considered for traffic conditions, one for average traffic states and one for oscillations during a period of time (10 minutes, as explained below) prior to the crashes. Potential measures of oscillations include the 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.[3] 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 7 presents an oblique curve of cumulative speed, $V(x, t) - V_0 * (t-t_0)$, versus time, t, at Jantzen Dr. 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.

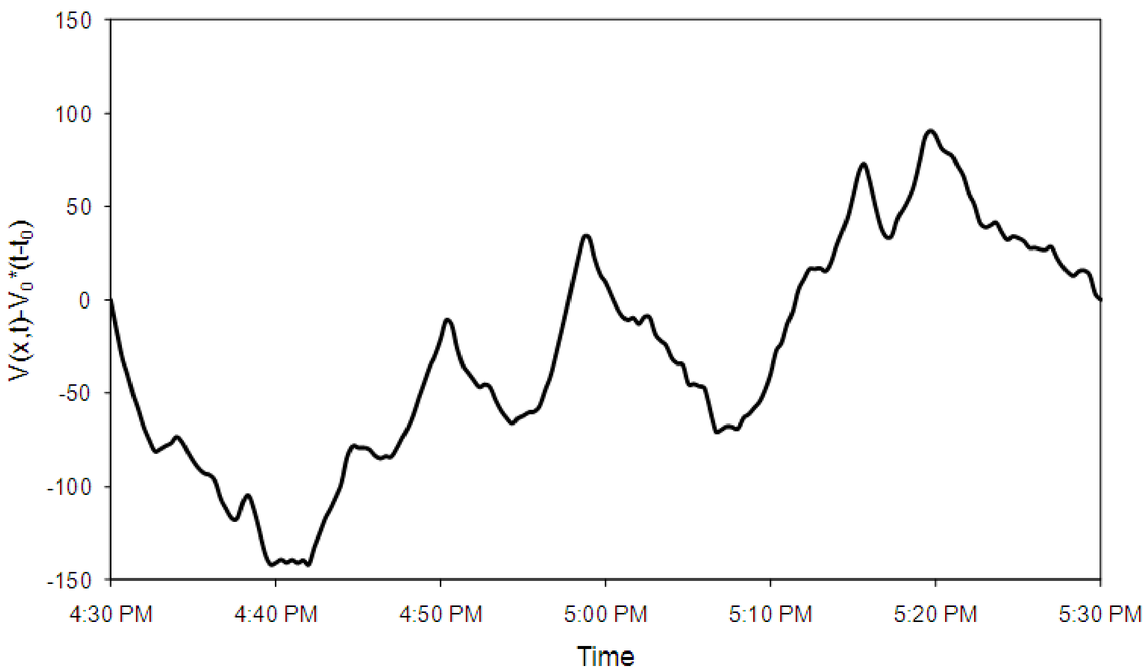


Figure 7 Oblique curve of cumulative speed at Jantzen Dr. between 4:30 and 5:30 p.m. on May 14, 2007

[3]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. To smooth the noise in the data and better reveal the underlying trends, cumulative curves of traffic observations are constructed on an oblique time axis. This technique is described in detail in Munoz and Daganzo (2002).

In the literature, it has been reported that crash occurrence times are often rounded to the nearest 5 minutes (Golob and Recker 2003; Kockelman and Ma 2007). In such cases, traffic data within the 5 minutes before each crash should be excluded from the analysis to avoid "cause and effect" ambiguity. We inspected our dataset to assess if the crash occurrence times were rounded to the nearest 5 minutes in Oregon. Our analysis shows that the majority (68 out of 82) of the crash occurrence times are not in multiples of 5 minutes. Therefore, we believe that the crash occurrence times used in this research were not rounded to the nearest 5 minutes and thus, did not further exclude any data.

The dependent variable is binary (crash-prone or not-crash-prone traffic conditions). Of note, the cases here are not the crashes themselves, but the traffic conditions immediately before the crash occurrences, while the 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 St. 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 Section 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 Section 5 and Section 6.

5. Model development

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 using average and the standard deviations of flow, density or speed as potential explanatory variables of average traffic states and oscillations. 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 2 lists all the potential explanatory variables and their basic statistics. The average speed is about 25 mph (40 km/h), indicating heavy congestion. The mean amplitude of traffic oscillations in terms of the 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), but ranges 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 2 Basic statistics of the potential explanatory variables (4:1 control-to-case ratio)

Potential Variable	Mean	Std. Dev.	Min	25% Percentile	75% Percentile	Max
Average speed[a]	24.743	11.237	2.303	16.542	30.448	49.938
Average count[b]	5.294	1.656	0.703	4.072	6.444	9.559
Average occupancy[c]	20.955	14.913	0.486	4.968	33.401	64.294
Std. dev. of speed[d]	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

[a] The average speed across all travel lanes at all detector stations during the 10-min periods right before the crash occurrences.
[b] The average number of vehicles per lane per 20 s at all detector stations during the 10-min periods right before the crash occurrences.
[c] The average percentage of time that the detectors are occupied by vehicles across all travel lanes during the 10-min periods right before the crash occurrences.
[d] The standard deviation of speed during the 10-min 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. However, since flow, density and speed are highly correlated in congested traffic (e.g., transitioning into a more congested state is typically accompanied by decreases in flow and speed and an increase in occupancy), only one of them is used to represent average traffic state or oscillations in the modeling process. Therefore, we include two exposure variables (one for oscillations and one for average traffic state). Furthermore, as supported by previous studies, an interaction term is not considered for the lack of a strong theoretical insight.

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.[4] Of note, 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.

Table 3 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 if the standard

[4] There are many criticisms (Good and Hardin 2006; Johnson 1999; Sterne and Smith 2001) of over-emphasizing 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.

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

deviation of speed increases by one additional unit (1 mph in this case), the odds ratio of a crash occurrence would increase by 9.6 percent.

Table 3 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[a]	1.096	.0424	2.37	0.02	1.016	1.183

[a] 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. Our analysis indicates that keeping the continuous variables linear in the logistic model is appropriate for the range of data (e.g., speeds). The details of this analysis are provided in Appendix A.

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 potentially can 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. 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 (see footnote 4). Ideally, cases and controls should be free of any sources of bias, such as construction, lane closure, 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 models based on the newly drawn samples and conducting a sensitivity analysis with respect to the case-to-control ratio. For the former, twenty 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 Section 5. Table 4 presents the modeling results from the 20 different runs for the control-to-case ratio of 4:1.

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

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

Run	Variables[a]	Odds ratio	95% CI
1	std_spd	1.096	(1.016, 1.183)
2	std_spd	1.066	(0.988, 1.151)
3	avg_occ	1.030	(1.002, 1.058)
	std_spd	1.119	(1.029, 1.216)
4	std_spd	1.070	(0.988, 1.160)
5	std_spd	1.090	(1.010, 1.177)
6	std_spd	1.069	(0.994, 1.149)
7	std_spd	1.077	(1.001, 1.160)
8	std_spd	1.079	(1.006, 1.158)
9	avg_occ	1.020	(0.997, 1.043)
	std_spd	1.071	(0.991, 1.157)
10	avg_occ	1.025	(0.997, 1.054)
	std_spd	1.094	(1.010, 1.185)
11	avg_occ	1.028	(0.999, 1.058)
	std_spd	1.118	(1.029, 1.215)
12	avg_occ	1.019	(0.995, 1.042)
	std_spd	1.089	(1.008, 1.176)
13	avg_occ	1.020	(0.996, 1.044)
	std_spd	1.077	(0.995, 1.165)
14	std_spd	1.081	(1.002, 1.166)
15	N/A	N/A	N/A
16	avg_cnt	0.809	(0.646, 1.013)
	std_spd	1.089	(1.005, 1.179)
17	avg_occ	1.017	(0.994, 1.041)
	std_spd	1.069	(0.994, 1.150)
18	avg_occ	1.021	(0.994, 1.050)
	std_spd	1.081	(0.997, 1.172)
19	avg_occ	1.026	(0.999, 1.053)
20	std_spd	1.078	(0.998, 1.165)
Average	std_spd	1.084	(1.003, 1.171)

[a] avg_occ is average occupancy; std_spd is standard deviation of speed; avg_cnt is average count, and N/A means not available.

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 a minimum of 1.066 and maximum of 1.119, which is fairly a tight bound. Moreover, the average lower and upper limits of the 95 percent confidence intervals are 1.003 and 1.171, respectively. For the study segment, an additional unit increase in the standard deviation of speed increases the odds ratio of crash occurrence by an average of 8.4 percent. This indicates that an average day is about 1.49 times[5] 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 2). This magnitude of impact seems reasonable and quite significant

[5] This number comes from $(1+0.084)^{4.982} = 1.49$.

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 only 9 times in the 20 runs. Thus, it is reasonable to conclude that speed variations in queued traffic (i.e., oscillations) have a larger impact on crash occurrence than average traffic states. Nevertheless, the average odds ratio for occupancy (around 1.02) is qualitatively consistent: it implies that the likelihood of crash occurrence increases as congestion becomes more severe.

Now we turn our attention to the sensitivity analysis of the modeling results shown in Table 4 with increasing control-to-case ratios. If the model results are robust, we expect similar or more consistent results for larger control-to-case ratios since the statistical power is expected to increase. To verify this, 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 re-developed based on the criteria described in Section 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 5 for the control-to-case ratios of 4:1 to 7:1.

Table 5 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 the consistency improves with increasing ratios, as expected. 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 Table B1 in Appendix B.

Table 5 Summary of model evaluation results for different control-to-case ratios (4:1 to 7:1)

Control-to-case ratio	Standard deviation of speed			Number of significant runs out of 20[a]			
	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

[a] avg_occ is average occupancy; std_spd is standard deviation of speed; avg_cnt is average count, and N/A means not available.

7. 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 occurrence, 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 crash occurrences 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 to 2007.

Models were developed based on a conditional logistic regression with several exposure variables related to 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 further evaluated by randomly drawing 20 different sets of controls and also by varying the control-to-case ratios (4:1 to 7:1) to assess consistency. 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 likelihood 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. 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.

Acknowledgements

This research was funded by Oregon Transportation Research and Education Consortium, 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 PORTAL. The authors also thank two anonymous reviewers for their valuable comments.

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.

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

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.

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

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.

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

Appendix A: The fractional polynomial analysis

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, $FP_m\{std.dev.of\ speed; (p_1, \ldots, p_m)\}$, is defined as

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

Where,

$p_1 \leq p_2 \leq \cdots \leq p_m$ denote 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 $p_1, \ldots, p_m$ 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 A1 for the result) and a first-degree (see Table A2) fractional polynomials. Higher-degree fractional polynomials are not considered to keep our model parsimonious.

Table A1 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).

Table A1 Summary of the second-degree FP comparisons for std. dev. of speed

Degree of FP, m	Powers	Degree of freedom	Deviance	Dev. Dif.	p^*
0 (Not in model)	N/A	0	263.948	10.026	0.040
1	1	1	258.582	4.660	0.198
1	2	2	258.316	4.393	0.111
2	−0.5, 0	4	253.922	N/A	N/A

* p-value from deviance difference comparing reported model with $m = 2$ model.

Similarly, according to Table A2, 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). Thus, the analysis indicates that it is appropriate to keep the standard deviation of speed linear in the logistic regression model.

Table A2 Summary of the first-degree FP comparisons for std. dev. of speed

This is a draft version. Please cite this paper as: Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis and Prevention 42*(2), 626-636.

Degree of FP, m	Powers	Degree of freedom	Deviance	Dev. Dif.	p^*
0 (Not in model)	N/A	0	263.948	5.632	0.060
1	1	1	258.582	0.266	0.606
1	2	2	258.316	N/A	N/A

* *p*-value from deviance difference comparing reported model with $m = 1$ model.

This procedure has been applied to other variables, which resulted in the same conclusion.

Appendix B: Model evaluation results for other control-to-case ratios

Table B1 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[a]	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)	1.080	(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)
	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)
	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)
	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)
	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)
	avg_occ			1.017	(0.997, 1.038)	1.020	(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)
	avg_occ					1.020	(0.997, 1.044)
	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)
	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)
	avg_cnt	0.804	(0.645, 1.003)	0.822	(0.662, 1.019)		
	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)
	avg_spd	0.980	(0.958, 1.003)			0.981	(0.959, 1.004)
	avg_occ			1.019	(0.996, 1.043)		
18	std_spd	1.083	(1.001, 1.173)	1.074	(0.996, 1.16)	1.077	(0.997, 1.162)
	avg_occ	1.024	(0.997, 1.053)	1.021	(0.995, 1.05)		
	avg_spd					0.982	(0.960, 1.005)
19	std_spd					1.071	(0.998, 1.150)
	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)

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