Relationships among Urban Freeway Accidents, Traffic Flow, Weather and Lighting Conditions

UCI-ITS-TS-WP-02-3

Thomas F. Golob ¹
Will Recker ²

¹ Institute of Transportation Studies
University of California, Irvine, tgo1ob@uci.edu

² Department of Civil and Environmental Engineering and Institute of Transportation Studies
University of California, Irvine, wwecker@uci.edu

June 2002

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

Forthcoming, ASCE Journal of Transportation Engineering
Relationships Among Urban Freeway Accidents, Traffic Flow, Weather, and Lighting Conditions

Thomas F. Golob
Institute of Transportation Studies
University of California, Irvine
tgolob@uci.edu

Wilfred W. Recker
Department of Civil and Environmental Engineering and
Institute of Transportation Studies
University of California, Irvine
wwrecker@uci.edu

Abstract

Linear and nonlinear multivariate statistical analyses are applied to determine how the types of accidents that occur on heavily used freeways in Southern California are related both to the flow of traffic and to weather and ambient lighting conditions. Traffic flow is measured in terms of time series of 30-second observations from inductive loop detectors in the vicinity of the accident prior to the time of its occurrence. Results indicate that the type of collision is strongly related to median traffic speed and to temporal variations in speed in the left and interior lanes. Hit-object collisions and collisions involving multiple vehicles that are associated with lane-change maneuvers are more likely to occur on wet roads, while rear-end collisions are more likely to occur on dry roads during daylight. Controlling for weather and lighting conditions, there is evidence that accident severity is influenced more by volume than by speed.

June 6, 2002

To appear in Journal of Transportation Engineering (ASCE)
Introduction

Our objective is to quantify relationships between the type of traffic accidents (crashes) that occur on urban freeways and the configuration of the traffic flow, while taking into account weather and lighting conditions. We have data on approximately 1200 crashes that occurred on six freeway routes in Southern California during in 1998. These crashes are characterized by: (a) the type and location of the primary collision, (b) the movement of the involved vehicles prior to collision, the number of vehicles involved, and (c) the accident severity (in terms of injury versus property damage only). Traffic flow is measured in terms of time series of 30-second observations from inductive loop detectors in the vicinity of the accident prior to the time of its occurrence.


Previous studies typically used such aggregate traffic flow data as daily or hourly traffic counts and volume to capacity measures. Types of collisions are generally not distinguished, except in terms of severity. Specification of a functional relationship between accident probabilities and the ambient traffic flow at the time of the accidents, as measured by commonly available traffic monitoring devices, has remained elusive. Mensah and Hauer (1998) cite two key problems of averaging associated with using aggregated data — argument averaging and function averaging. Argument averaging relates to the use of average traffic flow data, rather than data measuring traffic conditions at the time of the accident. The second problem, function averaging, is caused by using the same functional relationship for all types of collisions under all conditions (e.g., day or night, dry or wet weather). By using traffic flow data prevailing just prior to the time of each accident and by including the conditions of the accident in the analysis, we are able to avoid these two problems.

Our analysis method involves several steps. First, to reduce collinearity in the traffic data, principal components analysis (PCA) is performed to identify relatively independent measurements of flow conditions. Nonlinear (nonparametric) canonical correlation analysis (NLCCA) is then conducted with three sets variables. The first set is comprised of a seven-category segmentation variable defining lighting and weather conditions; the second set is made up of accident characteristics (collision type, location and severity; and the third set made up of the traffic flow variables identified using PCA. NLCCA is a form of canonical correlation analysis in which categorical variables are
optimally scaled as an integral component in finding linear combinations of variables with the highest correlations between them. These analyses show clear patterns relating accident characteristics and prevailing flow conditions.

Data Description

Fusion of Accident and Traffic Flow Data

The accident data were obtained from the Traffic Accident Surveillance and Analysis System (TASAS) maintained by the California Department of Transportation (Caltrans, 1993). The database contains those collisions that occur on the California State Highway System for which there are police reports. Most of the collisions included in the TASAS database were investigated in the field, but some were reported after the fact, usually for insurance reasons. The database does not cover collisions for which there are no police reports. Since the focus is on collisions that involved vehicles traveling on the main lanes of urban freeways, we were concerned only with what are defined as "highway" collisions in the TASAS database. For calendar year 1998, 9341 such collisions are recorded in the database for six major freeway routes in Orange County, California: Interstate Route 5, State Route 22, State Route 55, State Route 57, State Route 91, and Interstate Route 405.

Data on traffic flow during the time period leading up to each accident was matched to the accident. These data come from an archived database of 30-second observations from inductance loop detectors buried at intervals along the freeways. These detectors provide information on two variables for each thirty-second interval: the number of vehicles that pass over the loop (count) and the proportion of time that the loop is covered by a vehicle (occupancy). Although these two variables can be used (under very restrictive assumptions of uniform speed and average vehicle length, and taking into account the physical installation of each loop) to infer estimates of space mean speeds at a point, we avoid making any such assumptions, and use only these direct measurements in our analyses. We assume only that the ratio of count to occupancy has a monotonic relationship to space mean speed. After testing different lengths of time for monitoring of traffic conditions, we determined that we needed approximately 30 minutes of 30-second observations at the loop detector station closest to the location of the accident to establish stable measures of traffic conditions prior to the accident.

The time of each accident is not known with precision. An inspection of the accident times, presumably obtained from eyewitness accounts documented in police reports, reveals that 85.6% of the 9,341 collisions have reported times in minutes that fall precisely on the twelve five-minute intervals that comprise an hour. Because of this obvious reporting bias, reported accident times are treated as likely being rounded to the nearest five-minute interval. Since it is important in this study that the traffic data represent pre-accident conditions (rather than conditions arising from the accident itself), the period of observations used in the analysis is cut off 2.5 minutes before the "nominal" accident time to help remove any "cause and effect" ambiguities associated
with the apparent round-off of reported times. Consequently, for each accident, pre-
事故交通条件 are measured by up to 55 sequential thirty-second loop-
detector observations, beginning 30 minutes before the nominal accident time.

At each mainline loop detector station, data typically are collected for each freeway
lane; the minimum number of lanes at any mainline freeway section in Orange County
in 1998 was three. To standardize traffic flow data for all collisions independent of the
number of freeway lanes involved, data were compiled for three lane designations: (a)
the left lane, always being the lane designated as the number one lane according to
standard nomenclature of numbering lanes in succession from the median to the right
shoulder; (b) an interior lane, being lane two on three- and four-lane freeway sections
and lane three on five- and six-lane sections; and (c) the right lane, always being the
highest numbered (right-most) lane.

Missing data proved a major problem in dealing with the loop detector data used in this
study. Complete data for all 55 intervals (a 27.5-minute period) was available for 24.5%
of the stations; another 11.4% of the stations had missing data for one or more of the 55
time slices. The remaining 34.1% of the loop detector stations reported no data at all
for the entire 27.5-minute period. Presumably these latter stations were inoperative at
that time, or there was some other problem in retrieving the data.

Filtering of observations by content was still necessary for the loop detector stations
with either full or partial data. We reviewed all data sequences based on time series
deviations, deviations across lanes, and logical rules derived from feasible volume and
occupancy relationships (i.e., from properties of plausible fundamental traffic flow
diagrams). Based on these tests, approximately 16% of the available 30-second loop-
detector observations were identified as being potentially invalid. In situations where
one 30-second observation was missing or out-of-bounds but where the data for the
adjacent time slices were valid, the data for the missing time slice were interpolated
from the adjacent observations.

Implementation of the filtering and interpolation operations resulted in a sample of 1,191
collisions with a full 27.5 minutes of ostensibly valid loop detector data for the
designated three lanes at the closest detector station. This represents 12.8% of the
9,341 highway collisions on the six major Orange County freeways that are recorded in
the TASAS database for 1998. For this final sample, the average distance from the
accident location to the closest detector station is 270 meters and the median distance
is 190 meters. Fully 78% of the 1,191 collisions were located within 400 meters (0.25
miles) of the detector station, 95% were located within 800 meters, and 99% within
1200 meters.

Accident Characteristics
Available information regarding the characteristics of each collision included: (a) the
number of parties (usually vehicles) involved, (b) movements of each vehicle prior to
collision, (c) the location of the collision involving each party, (d) the object(s) struck by
each vehicle, and (e) the severity, as represented by the numbers of injured and fatally
injured parties in each involved vehicle. No information was available to us concerning
drivers or vehicle makes and models. The characteristics used here are listed in Table
1. For each of these characteristics, contingency table chi-squared tests revealed that
there is no statistically significant difference (at the 95% confidence level) between the
subset of 1,191 accidents for which we have traffic flow data and the complementary
subset of 8,150 accidents on Orange County freeways for which we have no traffic data.

Table 1. Accident Characteristics Used in the Analyses (N = 1192).

<table>
<thead>
<tr>
<th>Variable and category</th>
<th>Percent of sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collision type</td>
<td></td>
</tr>
<tr>
<td>Single vehicle hit object or overturn</td>
<td>14.2</td>
</tr>
<tr>
<td>Multiple vehicle hit object or overturn</td>
<td>5.9</td>
</tr>
<tr>
<td>Two-vehicle weaving accident *</td>
<td>19.3</td>
</tr>
<tr>
<td>Three-or-more-vehicle weaving accident *</td>
<td>5.5</td>
</tr>
<tr>
<td>Two-vehicle straight-on rear end</td>
<td>33.8</td>
</tr>
<tr>
<td>Three-or-more-vehicle straight-on rear end</td>
<td>21.3</td>
</tr>
<tr>
<td>Collision Location</td>
<td></td>
</tr>
<tr>
<td>Off-road, driver's left</td>
<td>13.8</td>
</tr>
<tr>
<td>Left lane</td>
<td>25.8</td>
</tr>
<tr>
<td>Interior lane(s)</td>
<td>32.7</td>
</tr>
<tr>
<td>Right lane</td>
<td>19.3</td>
</tr>
<tr>
<td>Off road, driver’s right</td>
<td>8.3</td>
</tr>
<tr>
<td>Severity</td>
<td></td>
</tr>
<tr>
<td>Property damage only</td>
<td>71.9</td>
</tr>
<tr>
<td>Injury or fatality †</td>
<td>28.1</td>
</tr>
</tbody>
</table>

* Sideswipe or rear end accident involving lane change or other turning manoeuvre
† There were only five fatal accidents

Weather and Lighting Conditions

Included in the documentation of each collision is information on lighting, weather, and
pavement conditions. Only 13% of all freeway accidents in Orange County in 1998
occurred during conditions of wet roads. A breakdown of the accidents by these
environmental conditions is displayed in Table 2. With the exception of (three) dusk-
dawn accidents on wet roads, there are at least 30 accidents for each combination of
weather and lighting, a number judged to be a sufficient cell size for analyses. The
three wet dusk-dawn accidents were dropped from the analyses, leaving seven segments defined by the cross-tabulation of Table 2.

Table 2. Breakdown of Sample by Weather and Ambient Lighting Conditions.

<table>
<thead>
<tr>
<th>Lighting</th>
<th>Weather*</th>
<th>Total by lighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dry</td>
<td>Wet</td>
</tr>
<tr>
<td>Daylight</td>
<td>789</td>
<td>101</td>
</tr>
<tr>
<td>Dusk or dawn</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Dark – street lights</td>
<td>95</td>
<td>32</td>
</tr>
<tr>
<td>Dark – no street lights</td>
<td>121</td>
<td>20</td>
</tr>
<tr>
<td>Total by weather condition</td>
<td>1035</td>
<td>156</td>
</tr>
</tbody>
</table>

*a Based on condition of the roadway surface (wet or dry)

Eliminated from further analyses

Traffic Flow Characteristics

Twelve variables were computed from the loop detector data. These were organized into four blocks of three variables each (one variable for each of the three lane type designations: left, interior, and right). The four blocks are as follows:

- The first of these blocks is an indicator of prevailing traffic speed. These three variables measure the central tendency of the ratio of volume to occupancy. This ratio is typically assumed to be proportional to the space mean speed. For example, under assumptions of stationary flow, and an average vehicle length of 5.49 meters (18 feet), a V/O ratio of 90 would translate to a space mean speed estimate of 49.2 km/h (30.6 mph). Median, rather than mean, is used in order to avoid the influence of outlying observations that can be due to failure of the loop detectors.

- The second block represents the temporal variation of the prevailing speed. Because we wish to minimize the influence of potentially invalid observations and the effects of outliers, we use the difference of the 90th percentile and 50th percentile of the distribution of volume over occupancy to capture variation.

- The third block measures the central tendency of traffic volume over the period. Volume alone is not as sensitive to outliers as is the ratio of volume to occupancy, so mean is used rather than median. Mean and median values are quite similar for these data, so either can be used without affecting results.

- The fourth and final block measures variation in volume over the period. Here we use standard deviation, but the difference between the 90th percentile and 50th percentiles can be used without affecting the results.
Our objective is to relate these traffic flow variables to accident characteristics. However, we recognize that the three variables in each of the four blocks might be highly correlated if the flow characteristic being measured is consistent across the three freeway lanes; yet, it is not known how well speed and volume variances in different lanes are linked. To better understand the correlation structure of these twelve variables, and to remove unnecessary redundancy from this set of twelve variables so that we can interpret results accurately, principal components analysis (PCA) was performed. The objective was to extract a relatively large number of factors in order to identify independent traffic flow variables while simultaneously discarding as little of the information in the original variables as possible. Six factors were found to account for 87.5% of the variance in the original twelve variables, and Varimax rotation was performed to aid in interpreting the factors. The factor loadings, which are the correlations between the original variables and the rotated factors, are listed in Table 3, together with the variances accounted for by each rotated factor. One variable was then selected to represent each factor in the subsequent stages of the analysis.

Table 3. Factor Loadings and Explained Variances for Six Principal Components of the Twelve Traffic Flow Variables (showing only loadings with absolute value greater than 3.0; factor loadings for variables selected to represent each factor are shown in bold and underlined).

<table>
<thead>
<tr>
<th>Traffic flow variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of original variance accounted for</td>
<td>21.6%</td>
<td>19.8%</td>
<td>14.5%</td>
<td>14.2%</td>
<td>8.7%</td>
<td>8.7%</td>
</tr>
<tr>
<td>Median vol./occupancy (V/O) left lane</td>
<td>0.896</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median vol./occupancy (V/O) interior lane</td>
<td></td>
<td>0.867</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median vol./occupancy (V/O) right lane</td>
<td></td>
<td>0.909</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variation in V/O left lane</td>
<td></td>
<td></td>
<td>0.836</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variation in V/O interior lane</td>
<td></td>
<td></td>
<td></td>
<td>0.875</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variation in V/O right lane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.929</td>
<td></td>
</tr>
<tr>
<td>Mean volume left lane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.928</td>
</tr>
<tr>
<td>Mean volume interior lane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>Mean volume right lane</td>
<td></td>
<td></td>
<td></td>
<td>0.742</td>
<td></td>
<td>-0.315</td>
</tr>
<tr>
<td>Variation in volume left lane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.924</td>
</tr>
<tr>
<td>Variation in volume interior lane</td>
<td></td>
<td></td>
<td></td>
<td>0.839</td>
<td></td>
<td>0.312</td>
</tr>
<tr>
<td>Variation in volume right lane</td>
<td></td>
<td></td>
<td></td>
<td>0.366</td>
<td></td>
<td>0.883</td>
</tr>
</tbody>
</table>
Factor 1: The factor loadings show that the central tendency of speed (Variable Block 1) is highly correlated across all three lanes. The variable chosen to represent this central tendency of speed factor is “median volume/occupancy in the interior lane”.

Factor 2: A single factor encompasses the central tendency of volume (Variable Block 3) in all three lanes, but the factor is more representative of volumes in the left and interior lanes than in the right lane, as witnessed by the lower correlation between this factor and right lane mean volume (0.742). Although the factor loading for “mean volume in the interior lane” is greater, “mean volume in the left lane” is chosen to represent this factor in all further analyses based on its consistently strong loadings on both this factor and Factor 3 (see below).

Factor 3: The third factor represents the temporal variation in volume in the left and interior lanes. Variation in volume in the right lane, which has a relatively low correlation of 0.368 with this factor, is captured by a separate factor (see Factor 6, below). Our interpretation is that the rightmost lane volume is influenced significantly by freeway on- and off-ramps, while traffic in the left and interior lanes is principally comprised of vehicles that are less impacted by weaving traffic in the vicinity of the ramps. “Variation in volume in the left lane” is chosen to represent temporal variations in volumes on the left and interior lanes.

Factor 4: As in the case of the temporal variation in volumes, the PCA results show that temporal variation in speed in the three lanes also is partitioned into two factors. Here again, the implication is that speed in the rightmost lane, which has a direct influence on the level of service in the vicinity of freeway on- and off-ramps, varies over relatively short periods of time in a different way than does mainline freeway speeds. “Variation in volume in the interior lane” is chosen to represent Factor 4.

Factor 5: “Variations in volume to occupancy ratio in the right lane” is relatively uncorrelated with any other factor, and by deduction relatively uncorrelated with any of the variables chosen to represent the other factors. There is a minor negative correlation between the Factor 5 and mean volume in the right lane, indicating that a high variation in speed in the right lane is associated with a lower traffic volume in that lane.

Factor 6: The final factor is comprised mostly of “variation in volume in the right lane.” The distinction between Factors 4 and 6 shows that flow on a section of freeway encompassing a series of ramp junctions may score high on Factor 6 during a weekend period during which there is substantial short-distance, discretionary travel that makes intensive use of freeway exits and entrances. Weekday peak-period traffic, on the other hand, will be characterized by longer trip lengths, thus scoring low on this factor.

The PCA results, summarized in Table 4, show that both the central tendencies of the traffic volumes and speeds, and their temporal variances, play separate roles in the traffic flow conditions present during collisions. For variances, we need to distinguish
between right lane effects and effects of the other lanes. Thus, six variables (two central tendency and four variances) can represent these factors in subsequent nonlinear statistical models. The correlations among these six variables are relatively small, allowing a more clear understanding of their separate contributions in later analyses.

Table 4. Interpretation of Principal Components Results and Variable Selection.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Interpretation</th>
<th>Represented by</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Central tendency of Speed</td>
<td>Median V/O interior lane</td>
</tr>
<tr>
<td>2</td>
<td>Central tendency of Volume</td>
<td>Mean volume left lane</td>
</tr>
<tr>
<td>3</td>
<td>Variation in Volume – Left &amp; Interior Lanes</td>
<td>Variation in volume left lane</td>
</tr>
<tr>
<td>4</td>
<td>Variation in Speed – Left &amp; Interior Lanes</td>
<td>Variation in V/O interior lane</td>
</tr>
<tr>
<td>5</td>
<td>Variation in Speed – Right Lane</td>
<td>Variation in V/O right lane</td>
</tr>
<tr>
<td>6</td>
<td>Variation in Volume – Right Lane</td>
<td>Variation in volume right lane</td>
</tr>
</tbody>
</table>

Nonlinear Canonical Correlation Analysis with Three Variable Sets

Methodology

The objective of this step in the analysis is to find the best explanation of patterns in the three accident characteristics listed in Table 1 as a function of the six flow characteristics representing the factors listed in Table 4, controlling for the seven categories of lighting and weather conditions defined by the cross-tabulation shown in Table 2. If all of the variables were numerical (measured on a scale with equal intervals), and all functional forms expected to be linear, this could be accomplished using canonical correlation analysis (CCA). In CCA, which is an expansion of regression analysis to more than one dependent variable, the objective is to find a linear combination of the variables in each of two or more sets, so that the correlations among the linear combinations in each set are as high as possible. Depending on the number of sets and the number of variables in each set, multiple linear combinations (called canonical variates) can be found that have maximum correlations subject to the conditions that all canonical variates are mutually orthogonal (uncorrelated).

The present CCA problem involves nonparametric (nonlinear), rather than numerical variables. The variable defining the seven segments of weather and lighting conditions and the two accident characteristics with more than two categories are nominal (categorical) by definition. Because we expect to find nonlinear relationships involving
the traffic flow variables, they are also considered nonlinear (either nominal or ordinal) in order to determine the optimal functional forms. The nonparametric CCA problem is more complex than its linear counterpart, because the optimal linear combination of the variables is undefined until the categories of each accident characteristic are quantified and the most effective nonlinear transformations of the traffic flow variables are determined. The variable categories must be optimally quantified (scaled) while simultaneously solving the traditional linear CCA problem of finding variable weights (van de Geer, 1986, van Buren and Heiser, 1989, ver Boon, 1996).

An elegant solution to the nonparametric (nonlinear) CCA problem was first proposed by researchers at the Department of Data Theory, Leiden University, Netherlands. The Leiden team developed a suite of nonparametric methods for conducting canonical correlation analysis (CCA), principal components analysis, and homogeneity analysis with variables of mixed scale types: nominal, ordinal, and interval. Their nonlinear CCA (NLCCA) method was operationalized in a program called CANALS (Canonical Analysis by Alternating Least Squares), later extended to generalized canonical analysis with more than two sets of variables in a program called OVERALS. The Leiden method for nonlinear CCA is described in van der Burg and de Leeuw (1983), Israëls (1987), Michailidis and de Leeuw (1998), and (most extensively) in Gifi (1990). The method simultaneously determines both (1) optimal re-scaling of the nominal and ordinal variables and (2) variable weights (coefficients), such that the linear combinations of the weighted re-scaled variables in all sets are maximally correlated. The variable weights and optimal category scores are determined as an eigenvalue problem related to minimizing a loss function derived from the concept of "meet" in lattice theory.

Model Specification

A NLCCA is specified with three sets variables, as described in Table 5. The first set is comprised solely of the seven-category segmentation variable defining the environmental conditions. This variable was treated as being "multiple nominal" in NLCCA parlance. That is, it was allowed to have different optimal category quantifications for each dimension in the solution. The second set is made up of the three accident characteristics (collision type, location and severity), each treated as being nominally scaled with a single optimal quantification for all dimensions. The third set was made up of the six traffic flow variables that were selected to represent the respective factors identified in Table 3. These were all treated as being ordinal, in that each was constrained to have a single optimal scaling that was monotonically increasing or decreasing across ten deciles. Tests of the effects of releasing these constraints (single versus multiple quantification, in the case of the accident characteristics; and nominal versus ordinal, in the case of the traffic flow characteristics) revealed that the simplifications are justified in that no major improvement in model fit is obtainable by complicating the variable treatments. The model results are described in the remainder of the paper.
Table 5. Variables in the Nonlinear Canonical Correlation Analysis

<table>
<thead>
<tr>
<th>Set</th>
<th>Variable</th>
<th>Scale type</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Segmentation by lighting and weather</td>
<td>Nominal</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Collision type</td>
<td>Nominal</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Collision location</td>
<td>Nominal</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Severity of the Collision</td>
<td>Nominal</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Median V/I interior lane</td>
<td>Ordinal</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Variation in V/I interior lane</td>
<td>Ordinal</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Variation in V/I right lane</td>
<td>Ordinal</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Mean volume left lane</td>
<td>Ordinal</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Variation in volume left lane</td>
<td>Ordinal</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Variation in volume right lane</td>
<td>Ordinal</td>
<td>10</td>
</tr>
</tbody>
</table>

Model Fit

A two-dimensional NLCCA solution was chosen. Table 6 lists the fit of this two-dimensional solution in terms of the variance accounted for within each set of variables by each of the two dimensions (canonical variates). The fit is greatest for the traffic flow variates on both dimensions. The first dimension is generally more effective than the second in explaining each of the segmentsations.

Table 6. Proportions of Variance Accounted for by the Canonical Variates

<table>
<thead>
<tr>
<th>Set</th>
<th>Dimension 1</th>
<th>Dimension 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Segmentation by lighting and weather</td>
<td>0.57</td>
<td>0.34</td>
</tr>
<tr>
<td>2. Accident characteristics</td>
<td>0.50</td>
<td>0.39</td>
</tr>
<tr>
<td>3. Traffic flow characteristics</td>
<td>0.77</td>
<td>0.60</td>
</tr>
</tbody>
</table>

The weights defining the two dimensions in terms of the optimally scaled variables are listed in Table 7. These weights are unique only for the variables that are constrained to have unique category quantifications. The contribution of the segmentation variable (i.e., weather and lighting) to the canonical variates is allowed to be different for each variate, and the results are described in terms of the category scores on each dimension (discussed later in the Section). In terms of the variables of sets two and
three, the first canonical variate primarily relates collision type, and secondarily collision location, to mean volume and median speed, with some contribution of variance in right-lane volume. The second variate relates both collision type and location to variations in volume and speed in the left and interior lanes. Accident severity is poorly explained, and its explanation is solely in terms of the first dimension.

Table 7. Weights of the Variables Comprising the Canonical Variates

<table>
<thead>
<tr>
<th>Set</th>
<th>Variable</th>
<th>Dimension 1</th>
<th>Dimension 2</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Segmentation by lighting and weather</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Collision type</td>
<td>0.513</td>
<td>-0.694</td>
<td>0.746</td>
</tr>
<tr>
<td></td>
<td>Collision location</td>
<td>-0.257</td>
<td>-0.471</td>
<td>0.288</td>
</tr>
<tr>
<td></td>
<td>Severity of the Collision</td>
<td>-0.183</td>
<td>0.020</td>
<td>0.034</td>
</tr>
<tr>
<td>3</td>
<td>Median V/I O interior lane</td>
<td>-0.397</td>
<td>0.257</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>Variation in V/I O interior lane</td>
<td>0.074</td>
<td>-0.418</td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td>Variation in V/I O right lane</td>
<td>-0.009</td>
<td>0.151</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>Mean volume left lane</td>
<td>0.593</td>
<td>0.011</td>
<td>0.351</td>
</tr>
<tr>
<td></td>
<td>Variation in volume left lane</td>
<td>0.082</td>
<td>0.482</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>Variation in volume right lane</td>
<td>0.256</td>
<td>0.041</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>Canonical Correlation</td>
<td>0.424</td>
<td>0.165</td>
<td></td>
</tr>
</tbody>
</table>

*Weights are not unique for variables treated as multiple nominal

The canonical correlation for each of the two orthogonal dimensions is a measure of the correlations among the three sets of variables. The first dimension is approximately 2.5 times more effective than the second at capturing the relationships among the three sets. The component loadings of each variable are measures of the correlations between the optimally scaled variables and the two orthogonal canonical variates. These are similar to factor loadings in PCA. The loadings for all variables are plotted in Figure 1, in which the first dimension is measured along the abscissa, the second along the ordinate. The length of the vector from the origin to the coordinates of each variable (shown by the solid markers) indicates the extent to which the variable is explained by the two canonical variates (the square of the length being equal to the percent of variance explained by all the other variables). Each vector is also projected through the origin to a phantom coordinate (shown by the empty markers) of equal magnitude but rotated 180 degrees from the variable coordinates in order to visualize negative correlations. The lighting and weather segmentation variable has two locations in the canonical space because it is allowed to have a different quantification for each dimension. The scalar (dot) product between any two variable vectors is indicative of the correlation between the two optimally scaled variables.
The components loadings plot shows that mean volume and variation-of-volume in the right lane are highly related to differences among one of the lighting and weather segments (that most closely aligned with the first, and most powerful canonical dimension), while the variation-of-volume in the left and interior lanes is correlated with the other (less powerful) dimension. Collision location is also related to mean volume and right-lane variation in volume, as well as to the weather and lighting segmentation variable aligned with the first dimension. Collision severity is also aligned with these variables (and the first dimension), but is the least well-explained accident characteristic by the two canonical variates. Collision type, on the other hand, is the best-explained accident characteristic and is related to median speed, and to left- and interior-lane variations in speed; contributions to its explanation are derived almost equally from each of the two canonical deviates. The model does poorly at capturing variation in right-lane speed.

The centroids of the optimally scaled categories of the segmentation variable are located in the canonical space in Figure 2. The pattern among these segments is clearly defined. The contrast between dry and wet weather conditions is consistently in the 120- versus 300-degree polar orientation (compass directions ESE vs. WNW). The contrast between daylight and darkness is consistently in the 45 vs. 225-degree rotation (NE vs. SW). The almost parallel relationships evident in Figure 2 indicate that the
relative effects of lighting conditions (in terms of their explanations by the two canonical variates) are invariant with respect to road surface condition, as are the corresponding effects of road surface condition to lighting. The first canonical variate (abscissa in Figure 2) is aligned with the difference between accident and traffic conditions on dry freeways in daylight as opposed to conditions on wet freeways in darkness. The second canonical variate (ordinate in Figure 2) is aligned with the difference between accident and traffic conditions on wet freeways in daylight as opposed to conditions on dry freeways in darkness. Dry dusk-dawn conditions are most similar to dry daylight conditions (rather than dry dark conditions). Finally, minor differences between unlighted and lighted conditions are similar on both wet and dry roads, and are captured mostly by the second canonical variate.

Figure 2. Category Centroids of the Segmentation Variable

Accident Typology and Lighting and Weather Conditions

Controlling for traffic flow differences (the third set of variables in the model), the relationships between weather and lighting conditions and collision type are revealed by the plot of category centroids of Figure 3. Hit object collisions and collisions involving multiple vehicles that are precipitated by weaving maneuvers are more likely on wet roads; this finding is consistent with the degradation of vehicle performance characteristics associated with wet road conditions (e.g., braking distance and skidding.
resistance). That all of these accident types, and particularly multiple vehicle collisions caused by weaving maneuvers, are more likely to occur on wet roads during daylight than on either dry or wet roads during darkness may be indicative of drivers' overconfidence in both their own and their vehicles' performance capabilities – a confidence that is superseded by the visual limitations imposed by darkness. Conversely, rear-end collisions are more likely to occur on dry roads during daylight, again perhaps reflecting the notion of a general driver overconfidence that succumbs to cautions dictated by adverse weather.

![Figure 3. Category Centroids of the Collision Type and Segmentation Variables](image)

The category centroids of the segmentation and collision type variables are plotted in Figure 4. As shown in Figure 4, off-road to drivers' right and left-lane collision locations are most associated with the first canonical variate (abscissa), which is also associated with the difference between dry freeways in daylight as opposed to wet freeways in darkness. Conversely, right-lane collisions are more closely aligned with the second variate (ordinate), separating wet daylight conditions from dry darkness conditions. Based on the optimal scaling of the categories of the collision location variable, this means that collisions off-road to drivers' right are associated with wet roads at night. Left-lane collisions are more associated with dry roads during daylight. There is also a moderate tendency for off-road-left collisions on wet roads during daylight.
Finally, category centroids of the segmentation and collision severity variables are plotted in Figure 5. Both of these category centroids fall directly on the axis defined by the first canonical variate, with the tendency toward increasing severity associated with wet road conditions under darkness.

**Accident Typology and Traffic Flow Conditions**

Controlling for lighting and weather conditions (the first set of variables in the model), the relationships between traffic flow characteristics and collision type are shown by the plot of category centroids of Figure 6. In case of the six traffic flow variables, for which the optimal scaling was restricted to be ordinal (monotonically increasing or decreasing), the centroids for the deciles are projected onto the coordinates of the variable. For clarity, the two traffic flow variables with the weakest relationships to the collision type variables (variation in speed right lane, and variation in volume left and interior lanes) are not included in the figures.
Figure 5. Category Centroids of the Severity and Segmentation Variables

Figure 6. Category Centroids of the Collision Type Variable and Projections of the Category Centroids of the Four Most Effective Traffic Flow Variables
The results indicate that differences in both the mean traffic volume and its variance are aligned with the first canonical deviate, while the second deviate is more closely associated with variance in speed effects. As expected, rear end collisions are generally associated with high variances in relatively low speeds—a condition commonly observed under heavily congested “stop-and-go” traffic. Conversely, hit-object and weaving collisions are predominately associated with relatively stable traffic characterized by low volumes and high steady speeds.

In terms of collision location, the results shown in Figure 7 identify off-road accidents with low volume conditions and relatively high speeds, with off-road right accidents more likely associated with the extremely light volumes of late night traffic (see Figure 4), while off-road left accidents more likely associated with light traffic coupled with high speed effects during daylight hours. Left lane collisions are more likely induced by volume effects, while right lane collisions are more closely tied to speed variances in adjacent lanes.

![Figure 7](image_url)

Figure 7. Category Centroids of the Collision Location Variable and Projections of the Category Centroids of the Four Most Effective Traffic Flow Variables

As expected, Figure 8 confirms that severity of accident generally tracks the inverse of the traffic volume. However, controlling for weather and lighting conditions, we find that
severity of accidents on urban freeways is influenced more by volume than by speed. One explanation for this is that, while relatively minor accidents are a direct byproduct of the low speed associated with congested traffic, it is the combination of moderate volumes with the relatively constant speeds associated with the high levels of service categories that produce conditions conducive to increased severity.

Figure 8. Category Centroids of the Severity Variable and Projections of the Category Centroids of the Four Most Effective Traffic Flow Variables

Traffic Flow Conditions and Lighting and Weather Conditions

The third set of relationships captured by the nonlinear canonical correlation model is between traffic flow and lighting and weather conditions (Figure 9). The more adverse conditions (in terms of visibility and road surface) are associated with the lowest volumes and variations in flow, while dry-daylight (or dusk-dawn) conditions are associated with high mean volumes and high variations in volumes. In terms of speed considerations, wet-daylight conditions are associated with low variations in speed on the left and interior lanes, while dry dark conditions are associated with high variations in speed.
Conclusions
The objective of this research is to find the best explanation of patterns in accident characteristics as a function of traffic flow characteristics, controlling for lighting and weather conditions. NLCCA results revealed that two independent dimensions (canonical variates), comprised of multiple linear combinations of the original accident, traffic flow and environmental conditions, effectively explained these relationships. The first canonical variate, which is approximately 2.5 times more effective than the second at capturing the relationships, primarily relates collision type, (and secondarily collision location) to mean volume and median speed. The second variate relates both collision type and location to variations in volume and speed in the left and interior lanes.

The results indicate that differences in certain aspects of lighting and weather (those aligned with the first canonical variate) are closely related to the mean volume and variation-of-volume in the right lane under accident conditions, which in turn influence the locations of the collisions. These conditions highlight the difference between accident and traffic conditions on dry freeways in daylight as opposed to conditions on wet freeways in darkness. Generally, off-road to drivers’ right and left-lane collision locations are most associated with such differences (i.e., between dry freeways in daylight as opposed to wet freeways in darkness); collisions off-road to drivers’ right are associated with wet roads at night, while left-lane collisions are more associated with
dry roads during daylight. There is also a moderate tendency for off-road to drivers' left
 collisions on wet roads during daylight. Off-road accidents generally are identified with
 low volume conditions and relatively high speeds, with off-road to drivers' right
 accidents more likely associated with the extremely light volumes of late night traffic,
 while off-road to drivers' left accidents more likely associated with light traffic coupled
 with high speed effects during daylight hours.

The second canonical variate is aligned with the difference between accident and traffic
 conditions on wet freeways in daylight as opposed to conditions on dry freeways in
darkness, and captures influences of the variation-of-volume in the left and interior
 lanes principally on right-lane collisions, separating wet daylight conditions from dry
 darkness conditions. Whereas left lane collisions are more likely induced by volume
 effects, right lane collisions are more closely tied to speed variances in adjacent lanes.

Collision type, the best-explained accident characteristic, is related to median speed,
 and to left-lane and interior-lane variations in speed. Hit object collisions and collisions
 involving multiple vehicles that are precipitated by weaving maneuvers are more likely
 on wet roads; rear-end collisions are more likely to occur on dry roads during daylight,
 and are generally associated with high variations in relatively low speeds – a condition
 commonly observed under heavily congested "stop-and-go" traffic. Conversely, hit-
 object and weaving collisions are predominately associated with relatively stable traffic
 characterized by low volumes and high steady speeds.

Finally, severity of accident generally tracks the inverse of the traffic volume. However,
 controlling for weather and lighting conditions, there is evidence that severity is
 influenced more by volume than by speed, an indication that the combination of
 moderate volumes with the relatively constant, speeds associated with the high levels of
 service categories, produce conditions conducive to increased severity.

The results of this investigation can begin to shed light on the complex relationships
 between traffic flow and traffic accidents (crashes). Although it is generally recognized
 that improved flow should lead to reductions in travel time, vehicle emissions, fuel
 usage, psychological stress on drivers, and improved safety, understanding the manner
 in which safety may be improved by smoothing traffic flow is not well understood. With
 such understanding, the potential safety benefits of improved traffic flow could be
 included, together with more traditional measures related to reduced congestion, in the
 assessment of investments in infrastructure or traffic management and control. The
 statistical procedures that have been developed can be used in conjunction with a data
 stream of 30-second observations from single inductive loop detectors to forecast the
 types of crashes that are most likely to occur for the flow conditions being monitored.
 Because the historical traffic flow data were not sufficiently representative of Orange
 County for an entire year (owing to systematic patterns in missing data as a function of
 freeway route, location along each route, day of week, and week of the year) we were
 unable to accurately calculate the rates, in terms of vehicle miles of travel, for crashes
 that happened to vehicles that were exposed to different traffic flow conditions.
Consequently, the current analysis provides information as to which types of crashes are more likely under different types of traffic flow, but does not forecast crash rates.

In spite of these limitations, we believe that we have demonstrated that procedures developed here can be used to gain insight into how changing traffic flow conditions affect traffic safety. To the extent that changed conditions are due to ATMS operations, or other projects that influence traffic operations, they can be used in evaluating the effectiveness of such projects. Or, as a forecasting tool combined with simulation studies of the likely future conditions, the relationships can be used to evaluate the safety conditions of alternative scenarios of operations with different ATMS or infrastructure treatments. The enhancement of these procedures in this direction, together with recalibration with more recent accident and traffic flow data, is necessary before any large-scale deployment of this tool, and is an important subject for future research.

References


