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

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

Thomas F. Golob, Wilfred W. Recker
University of California, Irvine

California PATH Working Paper
UCB-ITS-PWP-2001-19

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

Final Report for MOU 3007

December 2001

ISSN 1055-1417

CALIFORNIA PARTNERS FOR ADVANCED TRANSIT AND HIGHWAYS

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

Thomas F. Golob

Institute of Transportation Studies
University of California
Irvine, CA 92687-3600 USA
tgolob@uci.edu

and

Wilfred W. Recker

Department of Civil and Environmental Engineering and
Institute of Transportation Studies
University of California
Irvine, CA 92687-3600 USA
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 to both the flow of traffic and 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.

Keywords: Traffic safety, Traffic flow, Loop detectors, Nonlinear multivariate analysis, Canonical correlation

1. Introduction

The objective of this research 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. Our data cover approximately 1200 crashes that occurred on six freeway routes in Southern California during calendar year 1998. These crashes are characterized by the type and location of the primary collision, the movement of the involved vehicles prior to collision, the number of vehicles involved, and 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.

There is strong empirical evidence of functional relationships between accident rates and traffic flow, conditional upon roadway characteristics (e.g., Aljanahi, *et al.*, 1999, Cedar and Livneh, 1982, Frantzeskakis and Iordanis, 1987, Garber and Gadiraju, 1990, Gwynn, 1967, Hall and Pendleton, 1989, Maher and Summersgill, 1996, Sandhu and Al-Kazily, 1996, Stokes and Mutabazi, 1996, Sullivan, 1990, Sullivan and Hsu, 1988, and Zhou and Sisiopiku, 1997). A series of studies have also dealt with quantification of the safety component of the marginal costs of roadway use, as a function of traffic speed, flow and density (Dickerson, Peirson and Vickerman, 2000, Jansson, 1994, Johansson, 1996, Jones-Lee, 1990, Newberry, 1988, O'Reilly, *et al.*, 1994, Shefer and Rietveld, 1997, Vickery, 1969, and Vitaliano and Held, 1991). In situations of congested flow, most studies have shown that both accident risk and the severity or cost per accident (however defined) are nonlinearly related to traffic speed and density.

These previous studies typically used aggregate traffic flow data, such 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. 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 the two problems of averaging cited by Mensah and Hauer (1998) – "argument" averaging and "function" averaging. The "argument" averaging problem is caused by using aggregate 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).

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 emerging from the relationships between accident characteristics and prevailing flow conditions.

2. Data Description

2.1. Fusion of Accident and Traffic Flow Data

The accident data were obtained from the TASAS (Traffic Accident Surveillance and Analysis System) 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 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. The analysis reported here uses data for 30 minutes prior to the reported time of each accident, at the loop detector station closest to location of 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 must be 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 round-off. Consequently, for each accident, pre-accident traffic conditions 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. In order 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.

The extent and quality of loop detector data is reliant upon the performance of the data retrieval and processing system. Missing data proved a major problem in dealing with the loop detector data used in this study. Complete data for all 55 time slices (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 64.1% 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 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 closet 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 0.17 miles and the median distance is 0.12 miles. Fully 78% of the 1,191 collisions were located within 0.25 miles of the detector station, 95% were located within 0.5 miles, and 99% within 0.75 miles.

2.2. Accident Characteristics

We have available the following information regarding the characteristics of each collision: the number of parties (usually vehicles) involved, movements of each vehicle prior to collision, the location of the collision involving each party, the object(s) struck by each vehicle, and 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 accident characteristics used here are listed in Table 1.

Table 1 Accident Characteristics Used in the Analyses.

	Percent of sample (N= 1192)
Collision type	
Single vehicle hit object or overturn	14.2
Multiple vehicle hit object or overturn	5.9
Two-vehicle weaving accident [a]	19.3
Three-or-more-vehicle weaving accident [a]	5.5
Two-vehicle straight-on rear end	33.8
Three-or-more-vehicle straight-on rear end	21.3
Collision Location	
Off-road, driver's left	13.8
Left lane	25.8
Interior lane(s)	32.7
Right lane	19.3
Off road, driver's right	8.3
Severity	
Property damage only	71.9
Injury or fatality [b]	28.1

[a] Sideswipe or rear end accident involving lane change or other turning maneuver

[b] There were only five fatal accidents

2.3. Weather and Lighting Conditions

Included in the documentation of each collision is information on lighting, weather, and pavement conditions. Southern California weather is typically dry for most of the year, and ice and snow is an extremely rare event at freeway elevations. 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.

There are at least 30 accidents for each combination of weather and lighting, with the exception of (three) dusk-dawn accidents on wet roads., and thirty is 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.

Lighting	Weather [a]		Total by lighting
	Dry	Wet	
Daylight	789	101	890
Dusk or dawn	30	3 [b]	33
Dark – street lights	95	32	127
Dark – no street lights	121	20	141
Total by weather condition	1035	156	1191

[a] Based on condition of the roadway surface (wet or dry)
[b] Eliminated from further analyses

2.4. 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.[1] 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 the ratio of volume to occupancy is, 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.

1 For example, under assumptions of stationary flow, and an average vehicle length of 18 feet, a V/O ratio of 90 would translate to a space mean speed estimate of 30.6 mph.

Our objective is to relate these traffic flow variables to accident characteristics. As a first step in accomplishing this, we wish to remove unnecessary redundancy from this set of twelve variables, so that we can interpret results accurately. The three variables in each of the four blocks might be highly correlated if the flow characteristic being measured is consistent across all three freeway lanes. However, 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, 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).

		Principal component					
	Traffic flow variable	1	2	3	4	5	6
	Percentage of original variance accounted for	21.6%	19.8%	14.5%	14.2%	8.7%	8.7%
Block 1	Median vol./occupancy (V/O) left lane	0.896					
	Median vol./occupancy (V/O) interior lane	**0.907**					
	Median vol./occupancy (V/O) right lane	0.909					
Block 2	Variation in V/O left lane				0.836		
	Variation in V/O interior lane				**0.875**		
	Variation in V/O right lane				0.308	**0.929**	
Block 3	Mean volume left lane		**0.928**				
	Mean volume interior lane		0.941				
	Mean volume right lane		0.742			-0.315	0.394
Block 4	Variation in volume left lane			**0.924**			
	Variation in volume interior lane			0.839			0.312
	Variation in volume right lane			0.366			**0.883**

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 also 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). "Mean volume in the left lane" is chosen to represent this factor in all further analyses. Although the factor loading for mean volume in the interior lane is greater, the left lane is chosen to represent both factors two and three based on consistently strong loadings on both factors.

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 0.366 with factor three, is captured by a separate, sixth factor. 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: The PCA results show that temporal variations in speeds 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 Four.

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 fifth factor 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 last factor is comprised mostly of "Variation in volume in the right lane." The distinction between the fourth and sixth factors 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 commuting traffic, on the other hand, will be characterized by vehicles traversing greater distances on the freeways, thus scoring low on this Factor.

A summary of these PCA results is given in Table 4. These results show that both (1) the central tendencies of the traffic volumes and speeds, and (2) their temporal variances, play separate roles in the traffic flow conditions present during collisions.

There are separate effects in the temporal variances that distinguish between the right lane effects and those of other lanes (left and interior). The six variables chosen to represent the factors are used in subsequent nonlinear statistical models. The correlations among these variables are small, allowing a more clear understanding of their separate contributions in the analysis that follows.

Table 4 Interpretation of Principal Components Results and Variable Selection.

Factor	Interpretation	Represented by
1	Central tendency of Speed	Median V/O interior lane
2	Central tendency of Volume	Mean volume left lane
3	Temporal Variation in Volume – Left & Interior Lanes	Variation in volume left lane
4	Temporal Variation in Speed – Left & Interior Lanes	Variation in V/O interior lane
5	Temporal Variation in Speed – Right Lane	Variation in V/O right lane
6	Temporal Variation in Volume – Right Lane	Variation in volume right lane

3. Nonlinear Canonical Correlation Analysis with Three Variable Sets

3.1. 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 an 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.[2] Also, 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 of Leiden University in the 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.

3.2. Model Specification

A NLCCA was 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. Finally 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 they were each 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)

[2] A nominal variable with only two categories is a special case of a numerical variable, a dummy variable, since only one interval is involved.

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

Set	Variable	Scale type	Number of categories
1	Segmentation by lighting and weather	Nominal	7[a]
2	Collision type	Nominal	6
	Collision location	Nominal	5
	Severity of the Collision	Nominal	2
3	Median V/O interior lane (“median speed”)[b]	Ordinal	10
	Variation in V/O interior lane (“variation in speed left and interior lanes”)	Ordinal	10
	Variation in V/O right lane (“variation in speed right lane”)	Ordinal	10
	Mean volume left lane (“mean volume”)	Ordinal	10
	Variation in volume left lane (“variation in volume left and interior lanes”)	Ordinal	10
	Variation in volume right lane (“variation in volume right lane”)	Ordinal	10

[a] 4 lighting conditions x 2 surface conditions, minus the “wet road - dusk or dawn” category
[b] Labels in parentheses are the factor interpretations represented by the variable

3.3. 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 variables on both dimensions. The first dimension is generally more effective than the second in explaining each of the segmentations.

Table 6 Proportions of Variance Accounted for by the Canonical Variates

Set	Dimension 1	Dimension 2
1. Segmentation by lighting and weather	0.57	0.34
2. Accident characteristics	0.50	0.39
3. Traffic flow characteristics	0.77	0.60

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

Set	Variable	Dimension 1	Dimension 2	R^2
1	Segmentation by lighting and weather	- [a]	- [a]	
2	Collision type	0.513	-0.694	0.746
	Collision location	-0.257	-0.471	0.288
	Severity of the Collision	-0.183	0.020	0.034
3	median speed	-0.397	0.257	0.224
	variation is speed left and interior lanes	0.074	-0.418	0.180
	variation is speed right lane	-0.009	0.151	0.023
	mean volume	0.593	0.011	0.351
	variation in volume left and interior lanes	0.082	0.482	0.239
	variation in volume right lane	0.256	0.041	0.067
	Canonical Correlation	0.424	0.165	

[a] 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.

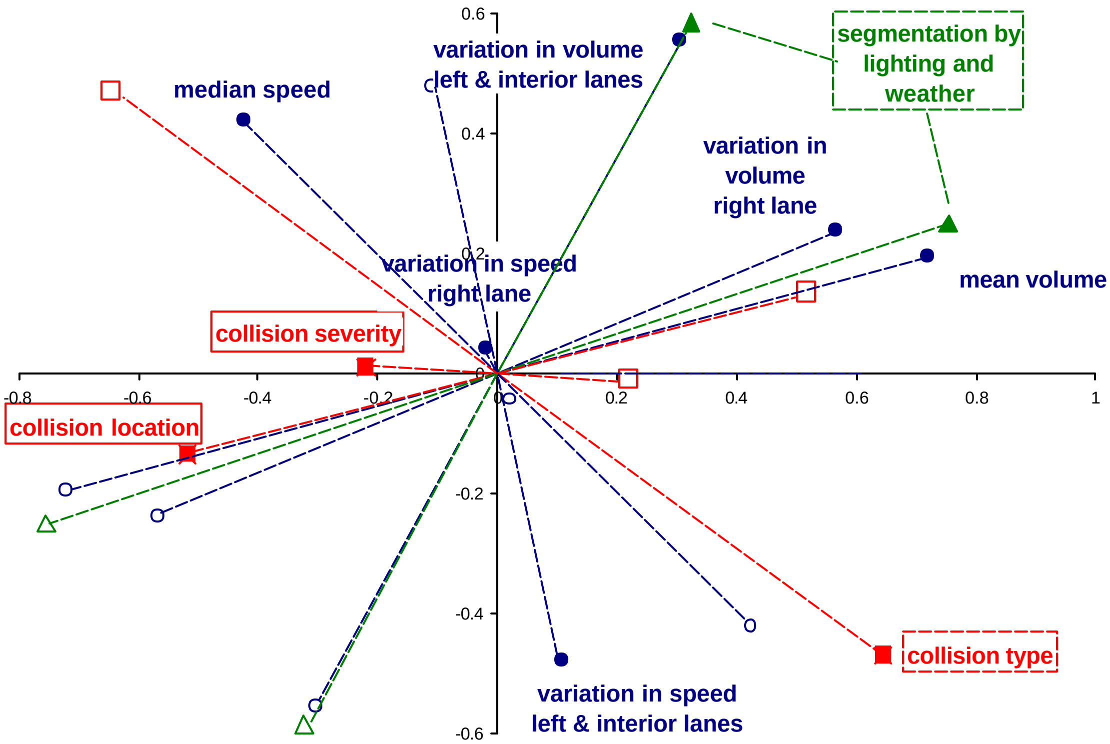

Figure 1 Component Loadings Plot (Triangle markers designate coordinate locations for variables in the first set, squares for the second set, and circles for the third set)

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 most well-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. An explanation of the variation in right-lane speed is not captured by the model.

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 are captured mostly by the second canonical variate.

3.4. 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 road 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 superceded 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.

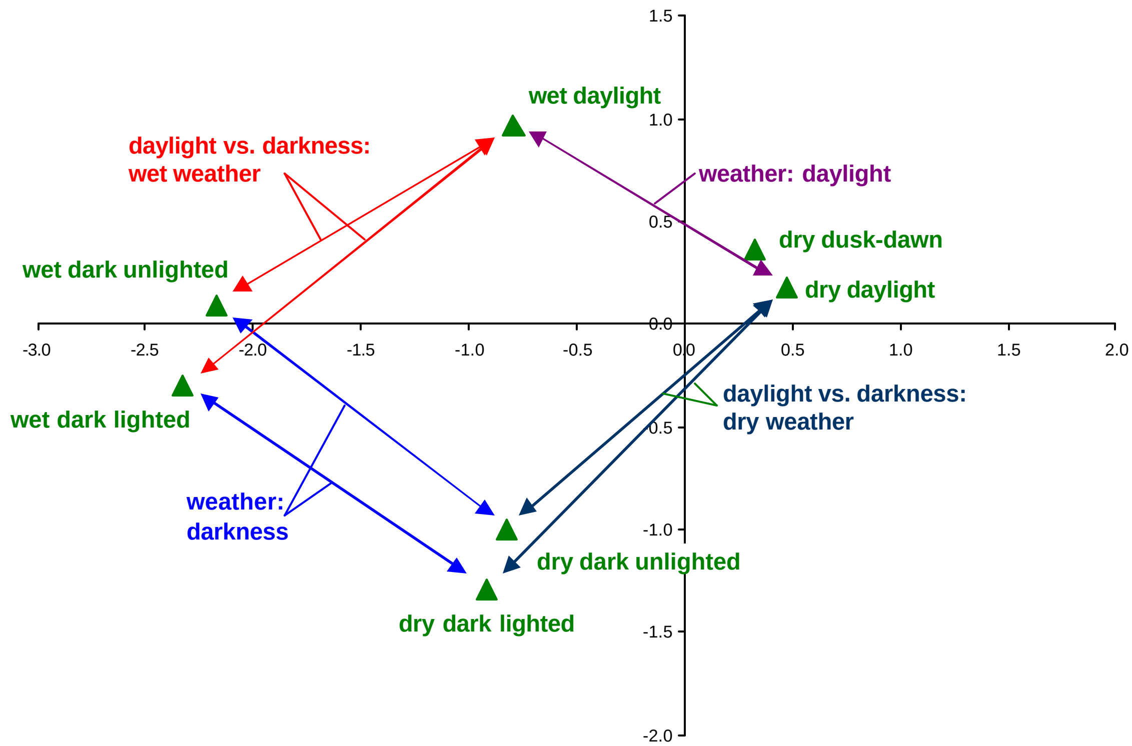

Figure 2 Plot of Category Centroids of the Segmentation Variable

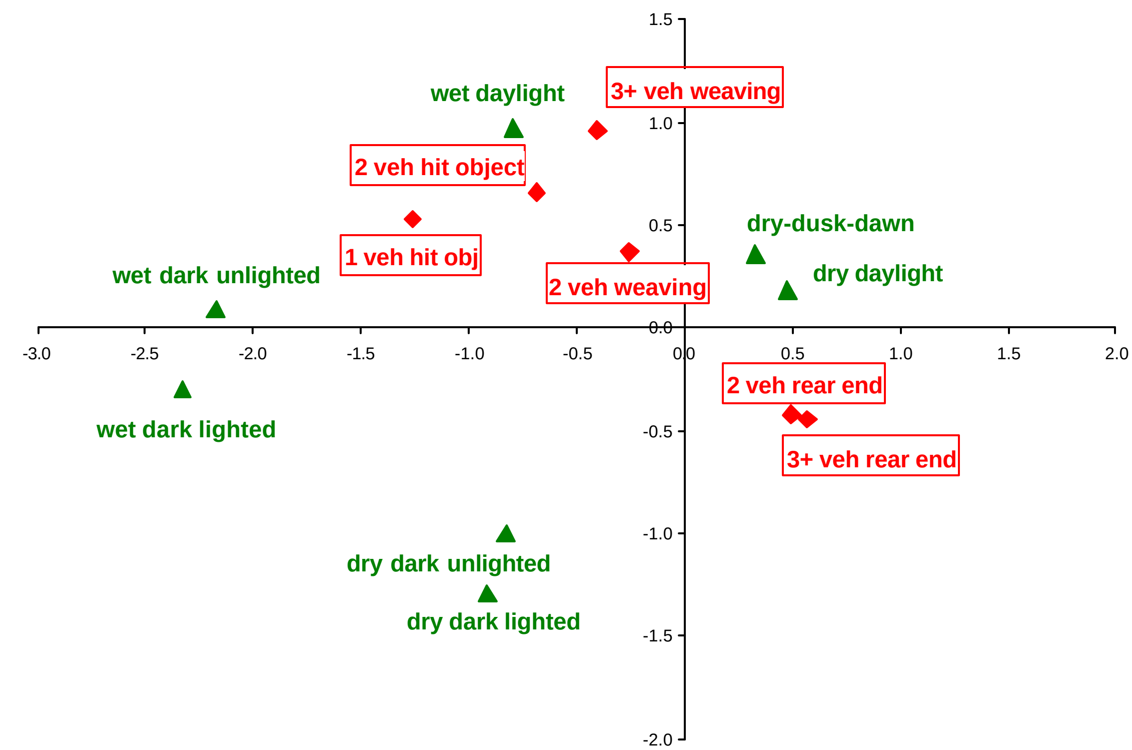

Figure 3 Plot of Category Centroids of the Collision Type and Segmentation Variables

The category centroids of the segmentation and collision type variables are plotted in Figure 4. As shown in Figure 4, the off-road –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.

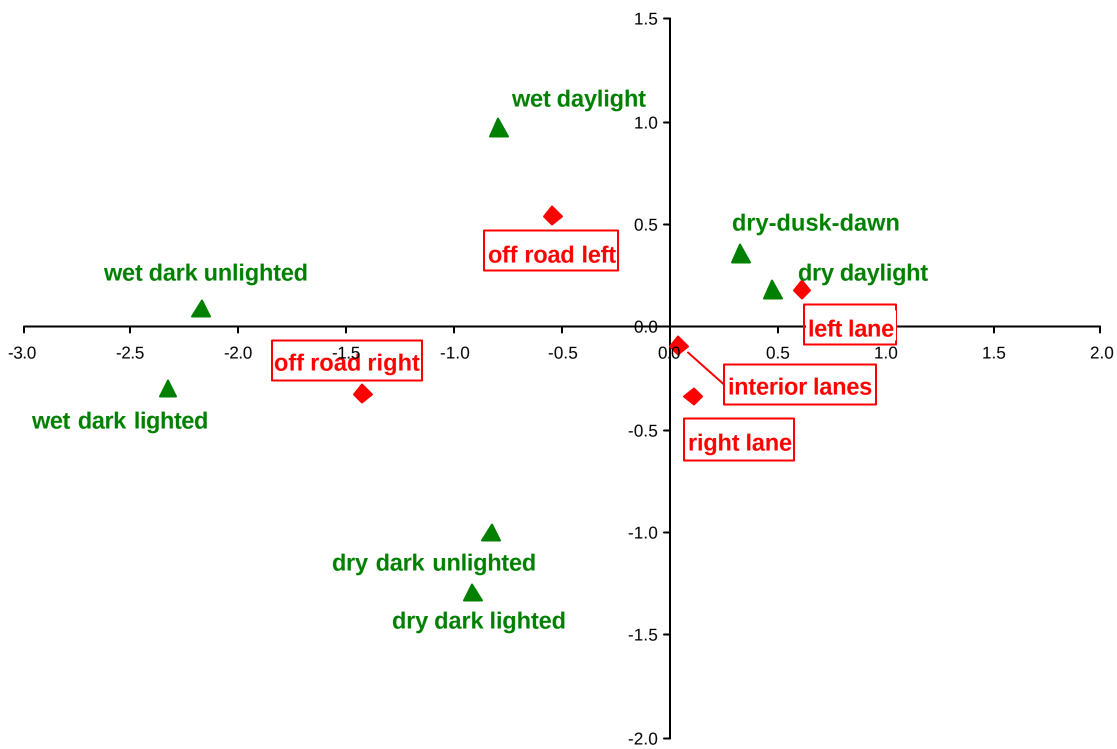

Figure 4 Plot of Category Centroids of the Collision Location and Segmentation Variables

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.

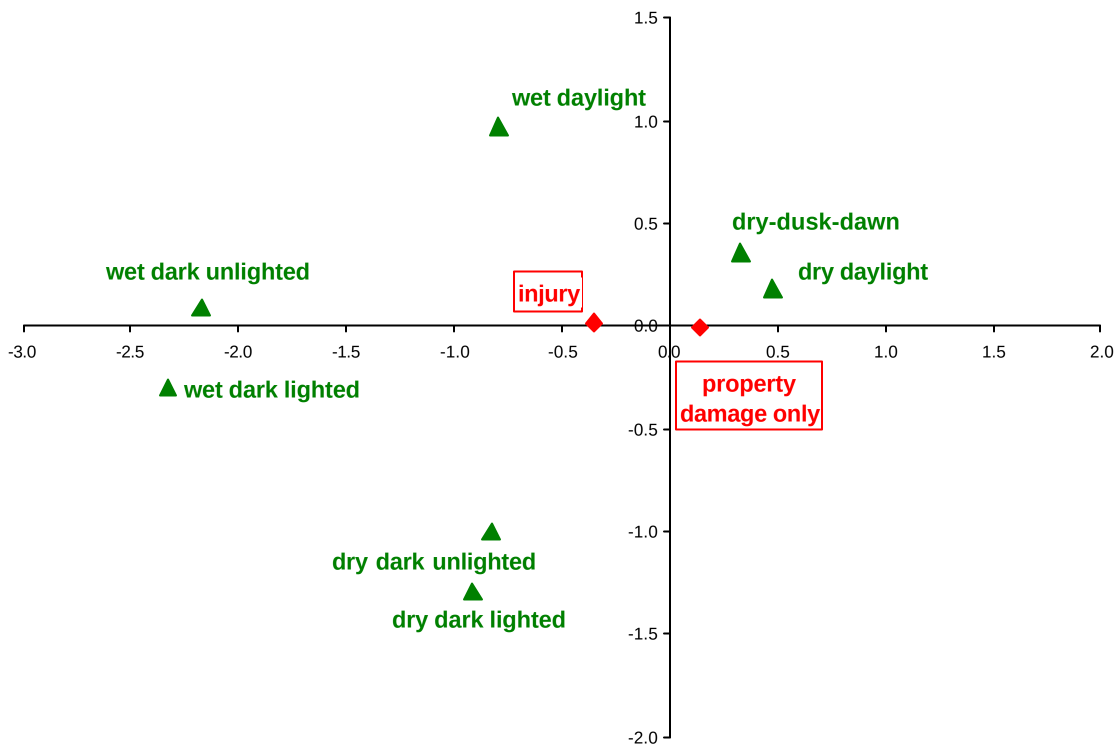

Figure 5 Plot of Category Centroids of the Severity and Segmentation Variables

3.5. 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 ten decile categories 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.

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 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.

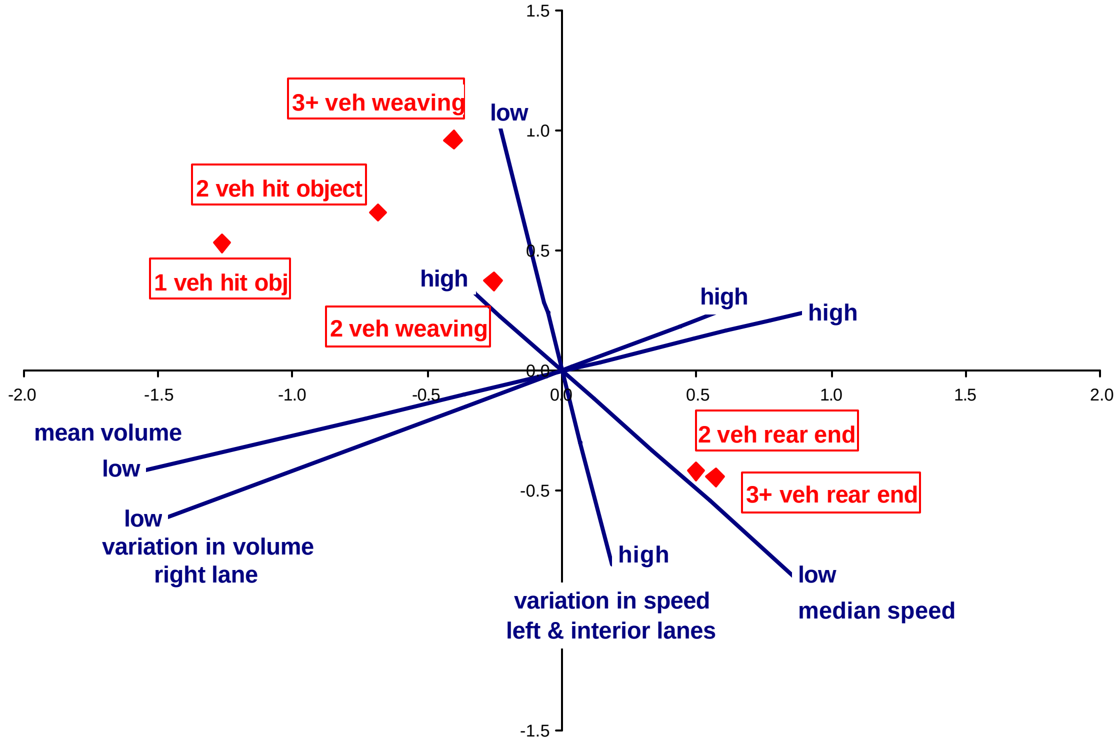

Figure 6 Plot of Category Centroids of the Collision Type Variable and Projections of the Category Centroids of the Four Most Effective Traffic Flow Variables

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.

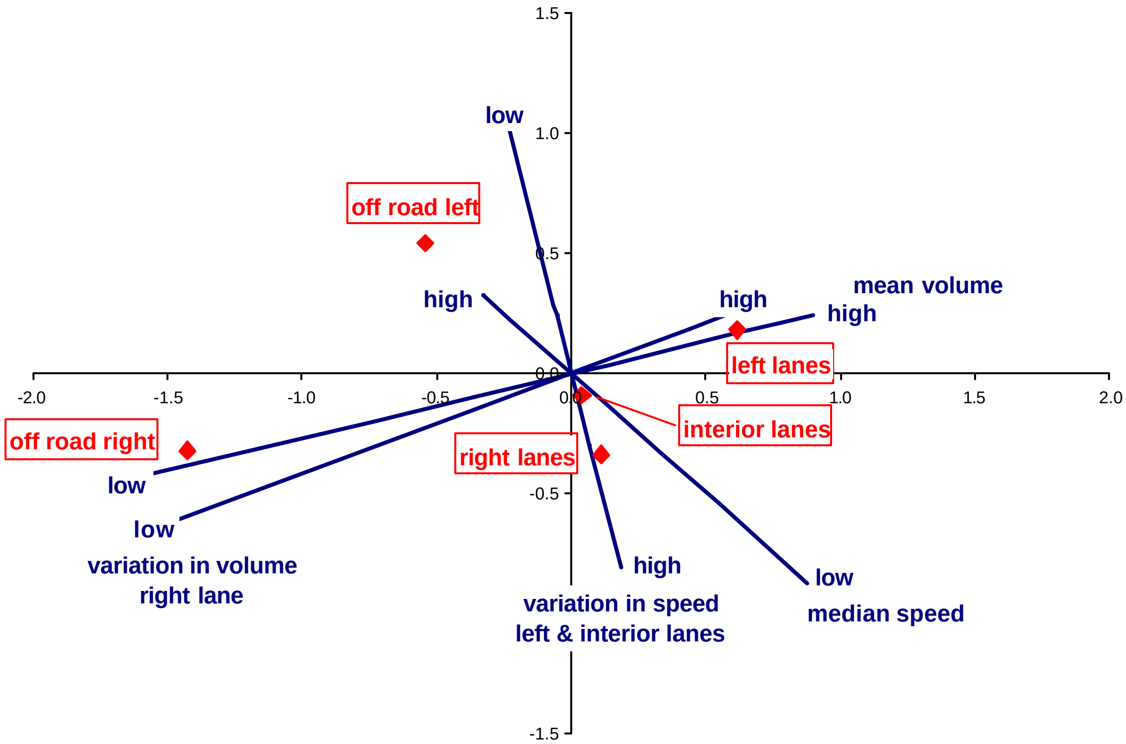

Figure 7 Plot of Category Centroids of the Collision Location Variable and Projections of the Category Centroids of the Four Most Effective Traffic Flow Variables

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.

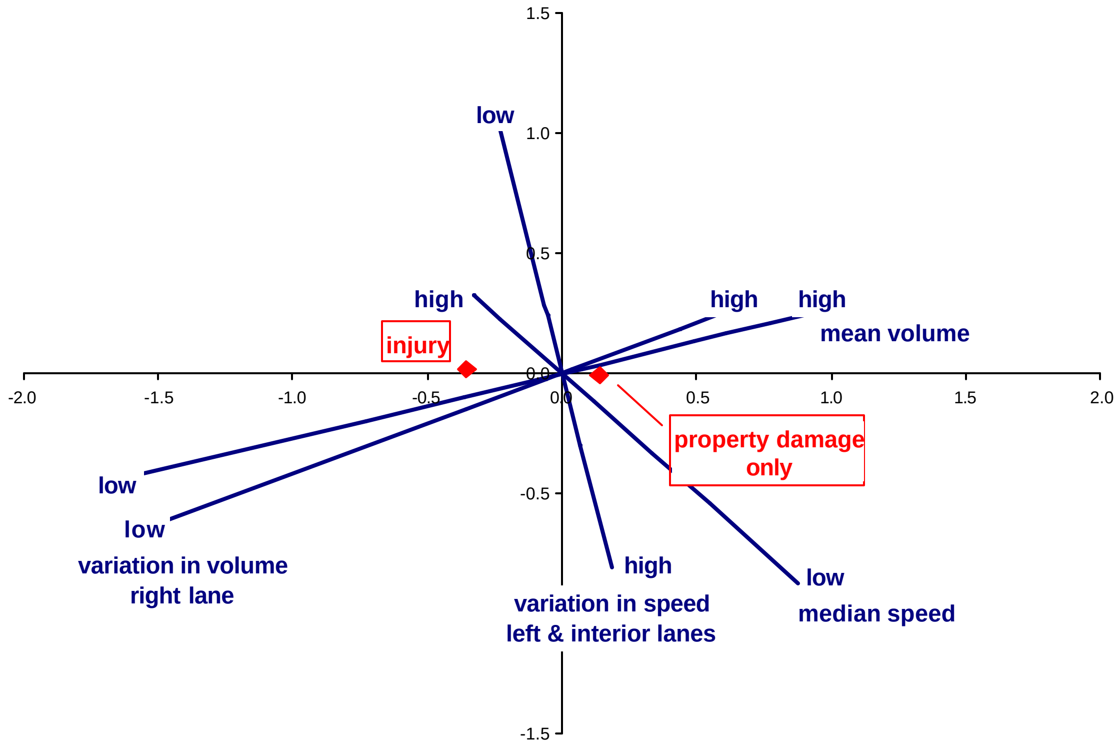

Figure 8 Plot of Category Centroids of the Severity Variable and Projections of the Category Centroids of the Four Most Effective Traffic Flow Variables

3.6. 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 most 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.

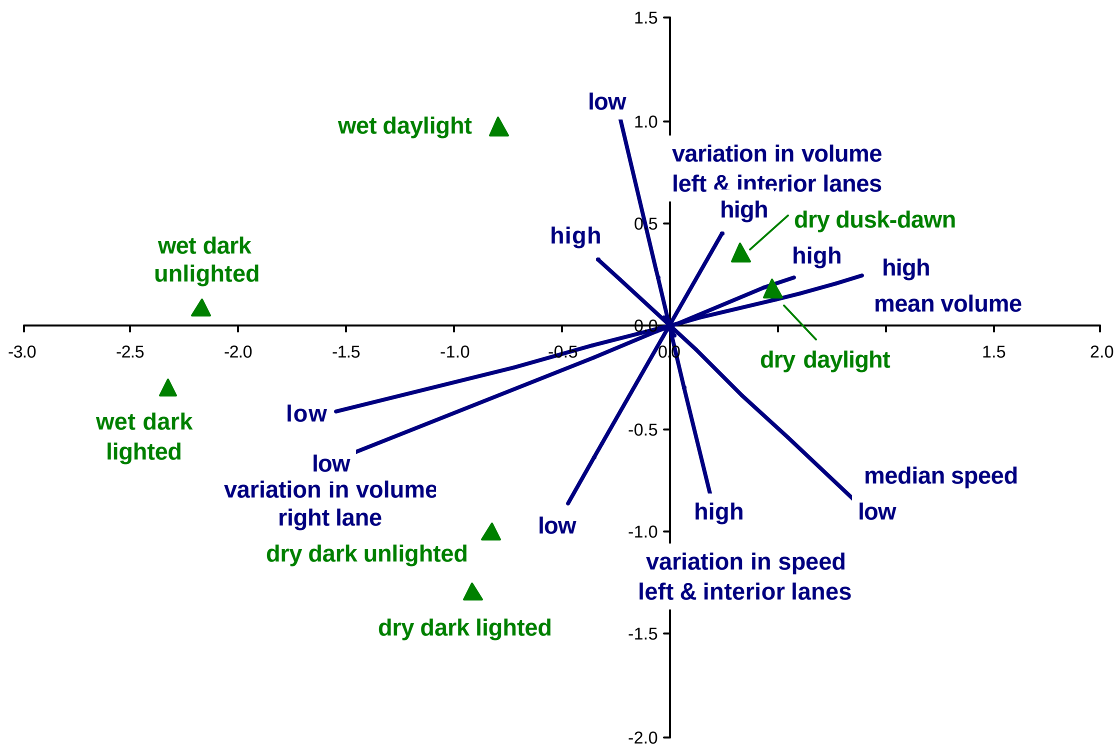

Figure 9 Plot of Category Centroids of the Segmentation Variable and Projections of the Category Centroids of the Traffic Flow Variables

4. 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 have general influence on 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 –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-left collisions on wet roads during daylight. Off-road accidents generally are identified 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, while off-road 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 most well-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.

Acknowledgments

This research was prepared in cooperation with the State of California, Business, Transportation and Housing Agency, Department of Transportation. The contents reflect the views of the authors, who are solely responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views of the State of California or the Federal Highway Administration. This report does not constitute a standard , specification, or regulation.

References

Aljanahi, A.A.M., A.H. Rhodes and A.V. Metcalfe (1999). Speed, speed limits and road traffic accidents under free flow conditions. *Accident Analysis and Prevention*, 31: 161-168.

Caltrans (1993). Manual of Traffic Accident Surveillance and Analysis System. California Department of Transportation, Sacramento.

Cedar, A. and L. Livneh (1982). Relationship between road accidents and hourly traffic flow. *Accident Analysis and Prevention*, 14: 19-44.

De Leeuw, J. (1985). The Gifi system of nonlinear multivariate analysis. In E. Diday, et al., eds., *Data Analysis and Informatics, IV: Proceedings of the Fourth International Symposium.* North Holland, Amsterdam.

Dickerson, A., J. Peirson and R. Vickerman (2000). Road accidents and traffic flows: An econometric investigation. Economica, 67: 101-121.

Ekman, L. (1996). *On the Treatment of Flow in Traffic Safety Analysis*. Institutionen För Trafikteknik, Lund, Sweden.

Frantzeskakis, J.M. and D.I. Iordanis (1987). Volume-to-capacity ratio and traffic accidents on interurban four-lane highways in Greece. *Transportation Research Record,* No. 1112: 29-38.

Fridstrøm, L., J. Ifver, S. Ingebrigtsen, R. Kulmala and L.K. Thomsen (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis and Prevention*, 27: 1-20.

Garber, N.J. and R. Gadiraju (1990). Factors influencing speed variance and its influence on accidents. *Transportation Research Record* 1213: 64-71.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, Chichester.

Gwynn, D.W. (1967). Relationship of accident rates and accident involvement with hourly volumes. *Traffic Quarterly*, 21: 407-418.

Hall, J.W. and O.J. Pendleton (1989). Relationship between V/C ratios and Accident Rates. Report FHWA-HPR-NM-88-02, U.S. Department of Transportation, Washington, DC.

Israëls, .Z. (1987). *Eigenvalue Techniques for Qualitative DATA.* DSWO Press, Leiden.

Jansson, J.O. (1994). Accident externality charges. Journal of transport Economics and Policy, 28: 31-43.

Johansson, P. (1996). Speed limitation and motorway casualties: A time series count data regression approach. *Accident Analysis and Prevention*, 28: 73-87.

Jones-Lee, M.W. (1990). The value of transport safety. Oxford Review of Economic Policy, 6: 39-60.

Mensah, A. and E. Hauer (1998). Two problems of averaging arising from the estimation of the relationship between accidents and traffic flow. *Transportation Research Record* 1635: 37-43.

Michailidis, G. and J. de Leeuw (1998). The GIFI system of descriptive multivariate analysis. Statistical Science, 13: 307-336.

Newberry, D. (1988). Road user charges in Britain. *Economic Journal*, 98: 161-176.

Sandhu, B. and J. Al-Kazily (1996). Safety impacts of freeway traffic congestion. Presented at Annual Meeting of Transportation Research Board, January 7-11, Washington, DC.

O'Reilly, D., J. Hopkin, G. Loomes, M. Jones-Lee, P. Philips, K. McMahon, D. Ives, B. Sobey, D. Ball and R. Kemp (1994). The value of road safety: UK Research on the valuation of preventing on-fatal injuries. *Journal of Transport Economics and Policy,* 28: 45-59.

Sandhu, B. and J. Al-Kazily (1996). Safety impacts of freeway traffic congestion. Presented at Annual Meeting of Transportation Research Board, January 7-11, Washington, DC.

Shefer, D. and P. Rietveld (1997). Congestion and safety on highways: Towards an analytical model. Urban Studies, 34: 679-692.

Stokes, R.W. and M.I. Mutabazi (1996). Rate-quality control method of identifying hazardous road locations. *Transportation Research Record,* No. 1542: 44-48.

Sullivan, E.C. (1990). Estimating accident benefits of reduced freeway congestion. *Journal of Transportation Engineering*, 116: 167-180.

Sullivan, E.C. and C-I. Hsu (1988). *Accident rates along congested Freeways*. Research Report UCB-ITS-RR-88-6, Institute of Transportation Studies, University of California, Berkeley.

Ter Braak, C.J.F. (1990). Interpreting canonical correlation analysis through biplots of structure correlations and weights. *Psychometrika*, 55: 519-531.

van Buren, S. and W.J. Heiser (1989). Clustering N-objects into K-groups under optimal scaling of variables. *Psychometrika*, 54: 699-706.

van de Geer, J.P. (1986). Relationships among *k* sets of variables, with geometrical representation, and applications to categorical variables. In J. de Leeuw, *et al.*, (Eds.), *Multidimensional Data Analysis,* DSWO Press, Leiden.

van der Burg, E. and J. de Leeuw (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology,* 36: 54-80.

ver Boon, P. (1996). *A Robust Approach to Nonlinear Multivariate Analysis*. DSWO Press, Leiden.

Vickrey, W. (1969). Congestion theory and transport investment. *American Economic Review,* 59: 251-260.

Vitaliano, D.F. and J. Held (1991). Road accident external effects: An empirical assessment. *Applied Economics*, 23: 373-378.

Zhou, M. and V.P. Sisiopiku (1997). Relationship between volume-to-capacity ratios and accident rates. *Transportation Research Record,* No. 1581: 47-52.