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

Thomas F. Golob[1] and Wilfred W. Recker[2]

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-s 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.

DOI: 10.1061/(ASCE)0733-947X(2003)129:4(342)

CE Database subject headings: Traffic safety; Traffic flow; Nonlinear analysis; Urban areas; Weather; Light.

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 1,200 crashes that occurred on six freeway routes in Southern California during 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.

There is strong empirical evidence of functional relationships between accident rates and traffic flow, conditional upon roadway characteristics (e.g. Gwynn 1967; Cedar and Livneh 1982; Frantzeskakis and Lordanis 1987; Sullivan and Hsu 1988; Hall and Pendleton 1989; Garber and Gadiraju 1990; Sullivan 1990; Stokes and Mutabazi 1996; Aljanahi et al. 1999; Sandhu and Al-Kazily 1996; 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 (Vickery 1969; Newberry 1988; Jones-Lee 1990; Vitaliano and Held 1991; Jansson 1994; O'Reilly et al. 1994; Johansson 1996; Shefer and Rietveld 1997; Dickerson, Peirson, and Vickerman 2000). In situations of congested flow, most studies have shown that both accident risk and the cost per accident (however defined) are nonlinearly related to traffic speed and density.

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.

[1]Researcher, Institute of Transportation Studies, Univ. of California, Irvine, CA 92687-3600. E-mail: tgolob@uci.edu

[2]Professor, Dept. of Civil and Environmental Engineering and Director, Institute of Transportation Studies, Univ. of California, Irvine, CA 92687-3600. E-mail: wwrecker@uci.edu

Note. Discussion open until December 1, 2003. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on November 27, 2001; approved on July 17, 2002. This paper is part of the ***Journal of Transportation Engineering***, Vol. 129, No. 4, July 1, 2003. ©ASCE, ISSN 0733-947X/2003/4-342–353/$18.00.

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, 9,341 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-s observations from inductance loop detectors buried at intervals along the freeways. These detectors provide information on two variables for each 30-s 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 min of 30-s 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 12 5-min intervals that comprise an hour. Because of this obvious reporting bias, reported accident times are treated as likely being rounded to the nearest 5-min interval. Since it is important in this study that the traffic data represent preaccident conditions (rather than conditions arising from the accident itself), the period of observations used in the analysis is cut off 2.5 min 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-accident traffic conditions are measured by up to 55 sequential 30-s loop-detector observations, beginning 30 min 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: (1) 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; (2) an interior lane, being lane two on three- and four-lane freeway sections and lane three on five- and six-lane sections; and (3) 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-min 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-min 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-s loop-detector observations were identified as being potentially invalid. In situations where one 30-s 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 min 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 270 m and the median distance is 190 m. Fully 78% of the 1,191 collisions were located within 400 m (0.25 miles) of the detector station, 95% were located within 800 m, and 99% within 1,200 m.

Accident Characteristics

Available information regarding the characteristics of each collision included: (1) the number of parties (usually vehicles) involved; (2) movements of each vehicle prior to collision; (3) the location of the collision involving each party; (4) the object(s) struck by each vehicle; and (5) 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.

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

Variable and category	Percent of sample
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.

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 m (18 ft), a volume/occupancy (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 12 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.

- 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 0.366 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

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	1,035	156	1,191

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

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 volume/occupancy (V/O) left lane	0.896					
	Median volume/occupancy (V/O) interior lane	**0.907**					
	Median volume/occupancy (V/O) right lane	0.909					
Block 2	Variation in volume/occupancy left lane				0.836		
	Variation in volume/occupancy interior lane				**0.875**		
	Variation in volume/occupancy 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**

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 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.

Table 4. Interpretation of Principal Components Results and Variable Selection

Factor	Interpretation	Represented by
1	Central tendency of speed	Median volume/occupancy interior lane
2	Central tendency of volume	Mean volume left lane
3	Variation in volume—left and interior lanes	Variation in volume left lane
4	Variation in speed—left and interior lanes	Variation in volume/occupancy interior lane
5	Variation in speed—right lane	Variation in volume/occupancy right lane
6	Variation in volume—right lane	Variation in volume right lane

The correlations among these six variables are relatively small, allowing a more clear understanding of their separate contributions in later analyses.

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).

Table 5. Variables in Nonlinear Canonical Correlation Analysis

Set	Variable	Scale type	Categories
1	Segmentation by lighting and weather	Nominal	7
2	Collision type	Nominal	6
	Collision location	Nominal	5
	Severity of the Collision	Nominal	2
3	Median volume/occupancy interior lane	Ordinal	10
	Variation in volume/occupancy interior lane	Ordinal	10
	Variation in volume/occupancy right lane	Ordinal	10
	Mean volume left lane	Ordinal	10
	Variation in volume left lane	Ordinal	10
	Variation in volume right lane	Ordinal	10

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 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 rescaling of the nominal and ordinal variables and (2) variable weights (coefficients), such that the linear combinations of the weighted rescaled 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 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. 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 6. Proportions of Variance Accounted for by Canonical Variates

	Dimension	
Set	1	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

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.

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.

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

Table 7. Weights of Variables Comprising Canonical Variates

		Dimension		
Set	Variable	1	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 volume/occupancy interior lane	−0.397	0.257	0.224
	Variation in volume/occupancy interior lane	0.074	−0.418	0.180
	Variation in volume/occupancy right lane	−0.009	0.151	0.023
	Mean volume left lane	0.593	0.011	0.351
	Variation in volume left lane	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.

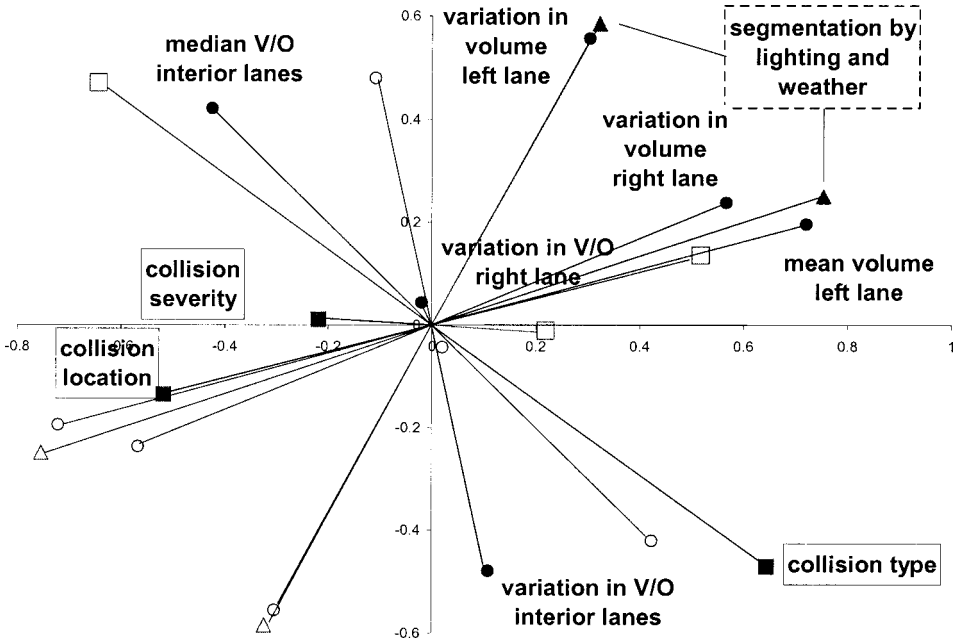

Fig. 1. Component loadings plot (triangle markers designate coordinate locations for variables in first set, squares for second set, and circles for third set)

loadings in PCA. The loadings for all variables are plotted in Fig. 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° 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 (Ter Braak 1990).

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 Fig. 2. The pattern among these segments is clearly defined. The contrast between dry and wet weather conditions is consistently in the 120 versus 300° polar orientation (compass directions ESE versus WNW). The contrast between daylight and darkness is consistently in the 45° versus 225° rotation (NE versus SW). The (almost) parallel relationships evident in Fig. 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 Fig. 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 Fig. 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.

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 Fig. 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 superceded by the visual limitations imposed by darkness. Conversely, rear-end collisions are

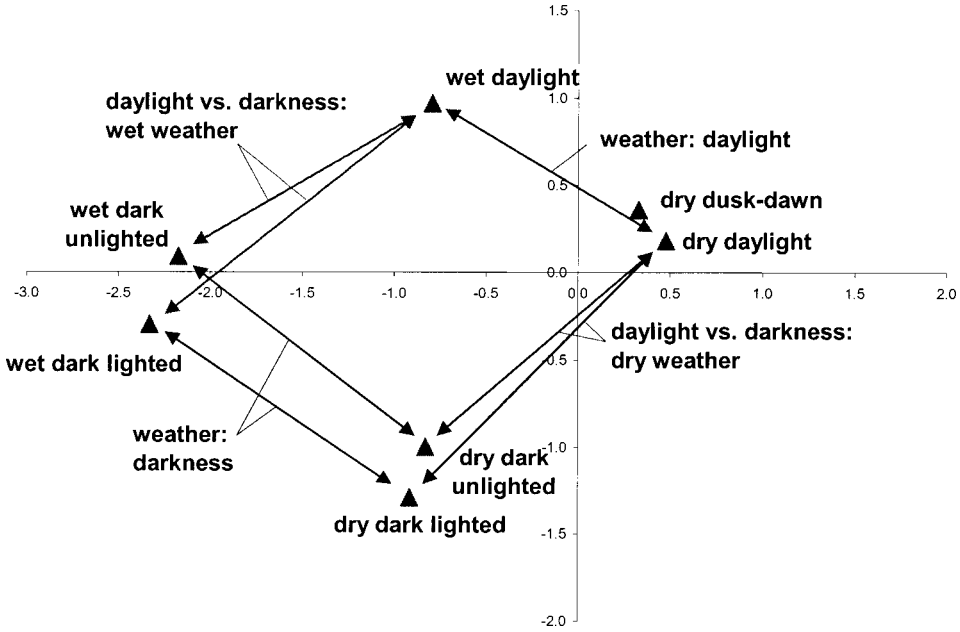

Fig. 2. Category centroids of segmentation variable

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.

The category centroids of the segmentation and collision type variables are plotted in Fig. 4. As shown in Fig. 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 Fig. 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 cat-

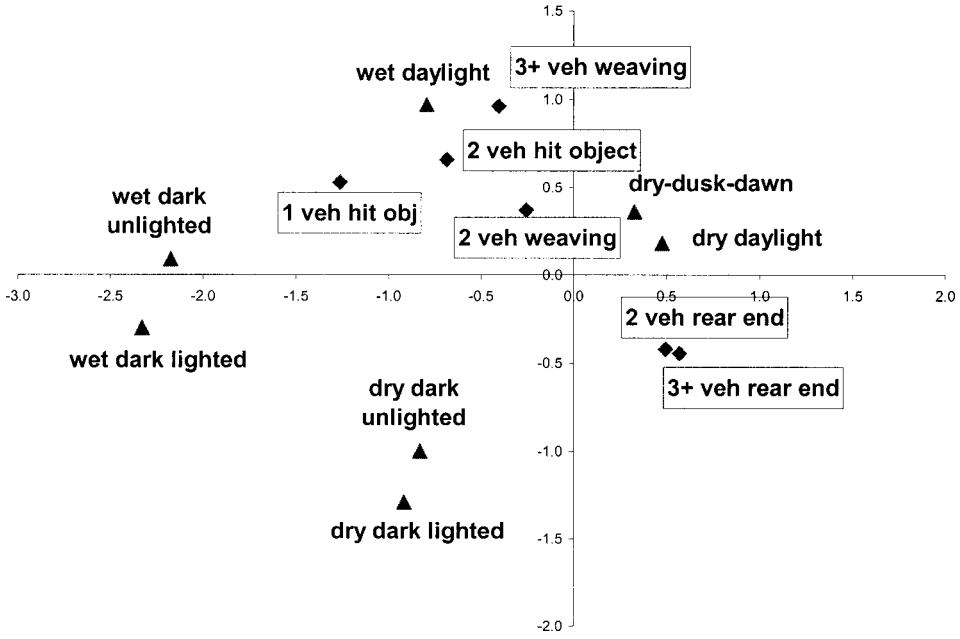

Fig. 3. Category centroids of collision type and segmentation variables

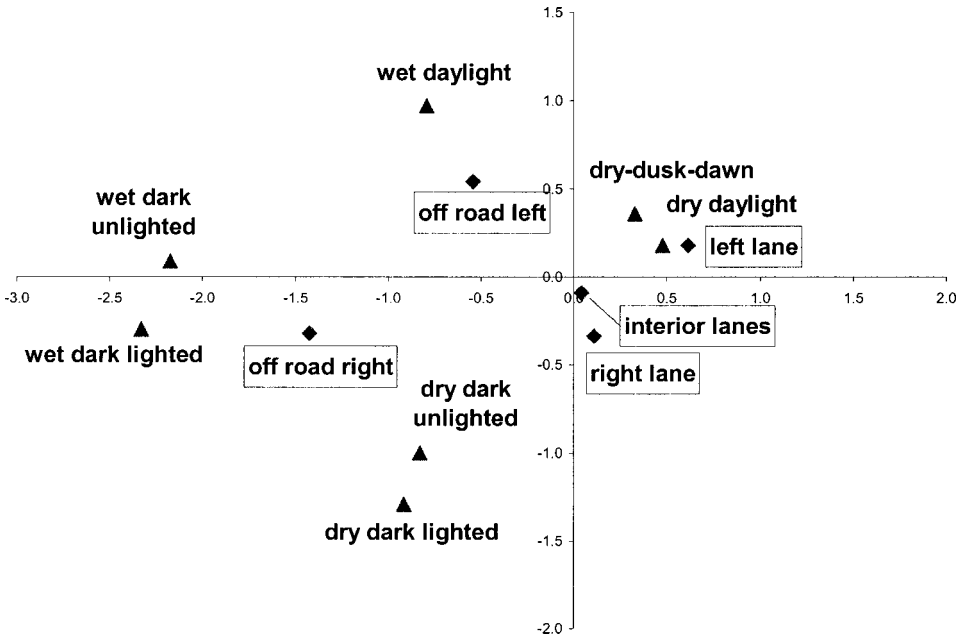

Fig. 4. Category centroids of collision location and segmentation variables

egory centroids of Fig. 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 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.

In terms of collision location, the results shown in Fig. 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 Fig. 4), while off-road left accidents more likely associated with light

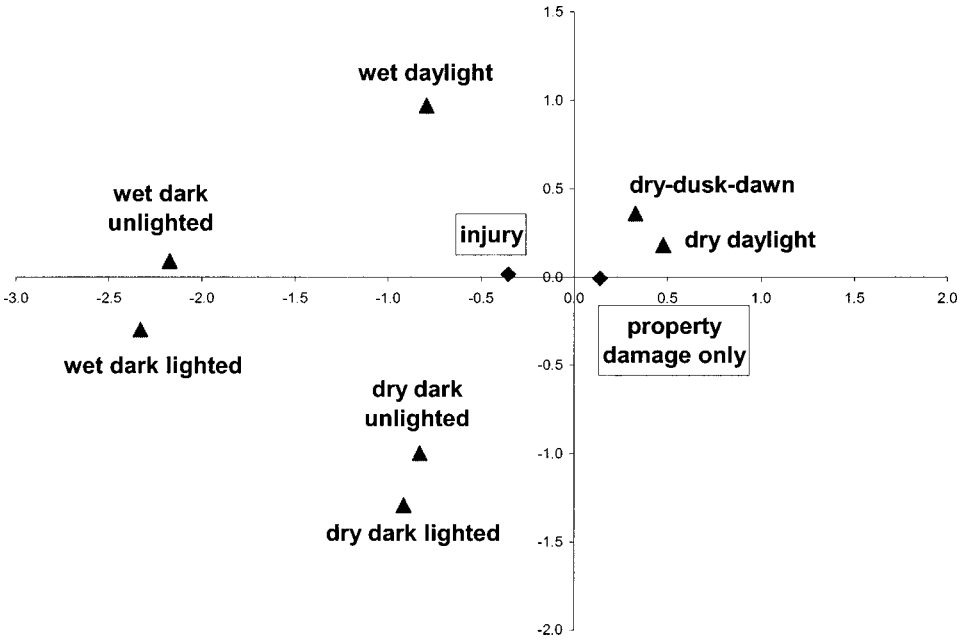

Fig. 5. Category centroids of severity and segmentation variables

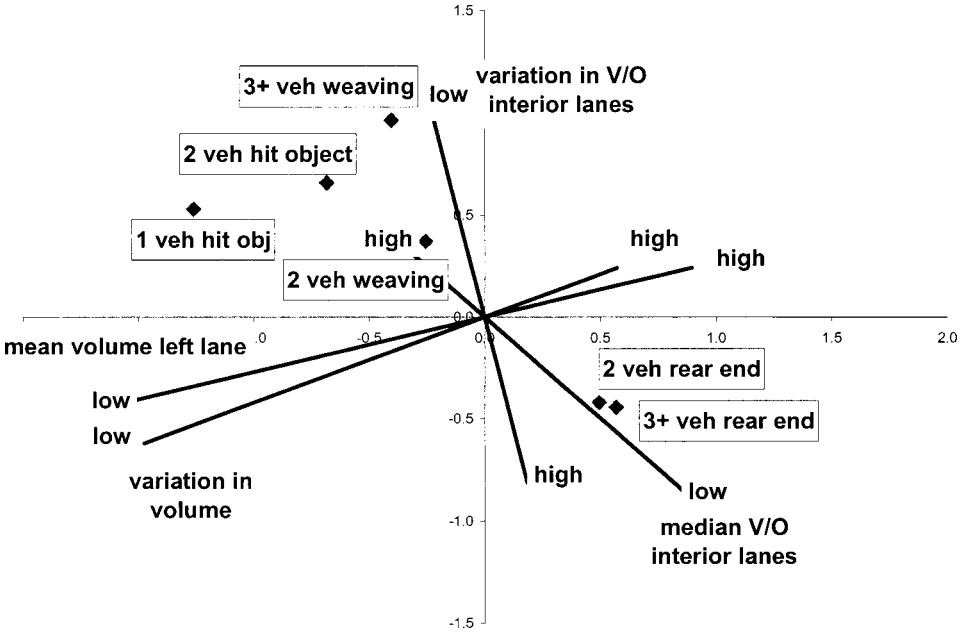

Fig. 6. Category centroids of collision type variable and projections of category centroids of four most effective traffic flow variables

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.

As expected, Fig. 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.

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 (Fig. 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.

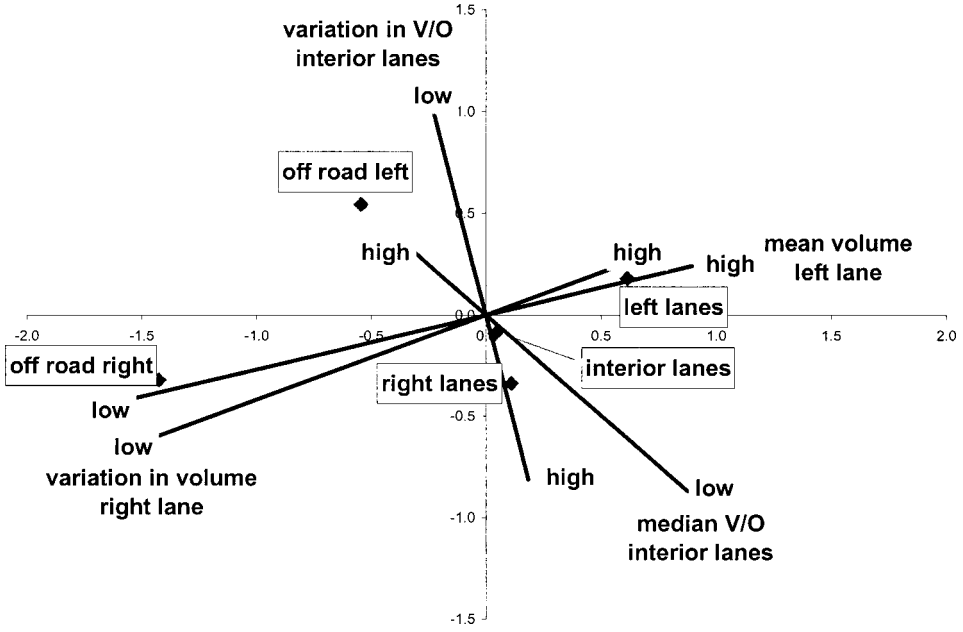

Fig. 7. Category centroids of collision location variable and projections of category centroids of four most effective traffic flow variables

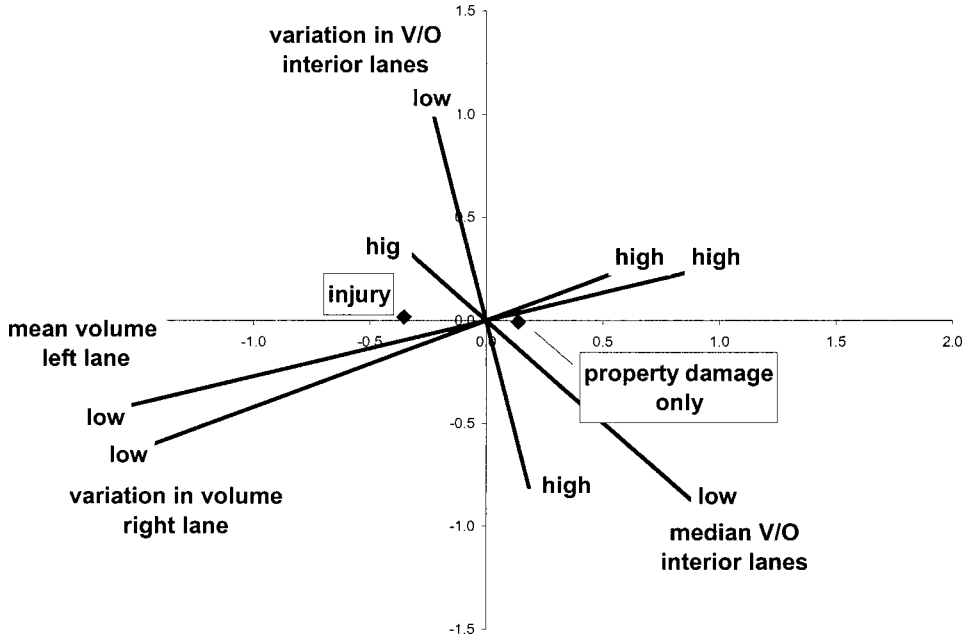

Fig. 8. Category centroids of severity variable and projections of category centroids of four most effective traffic flow variables

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 (noted earlier by Fridstrøm et al. 1995) 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

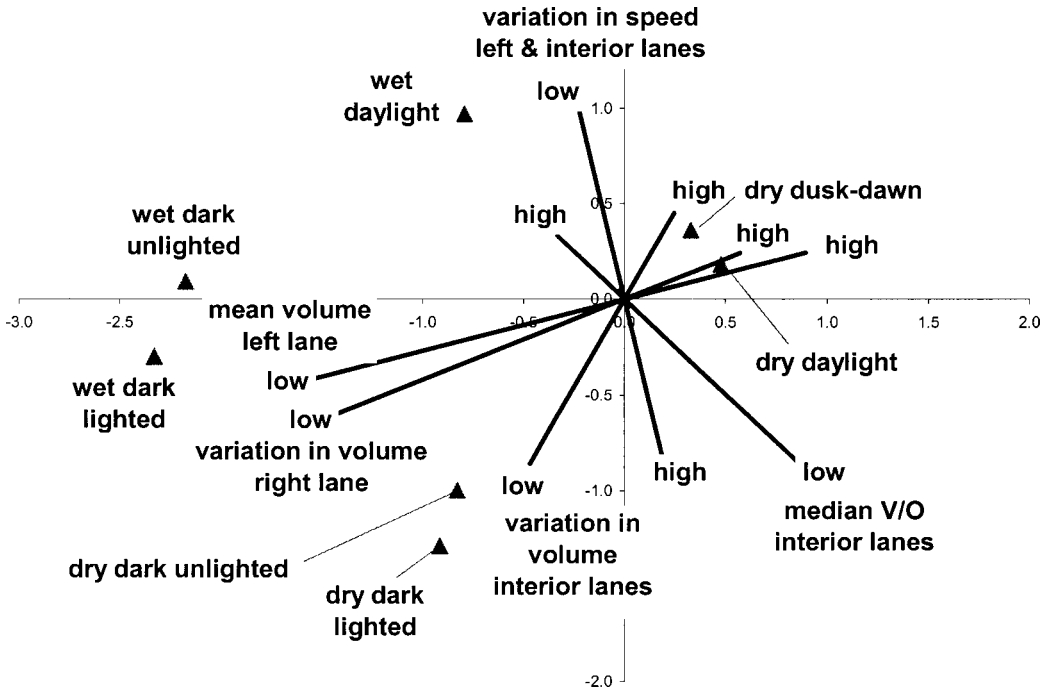

Fig. 9. Category centroids of segmentation variable and projections of category centroids of traffic flow variables

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-s 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

Aljanahi, A. A. M., Rhodes, A. H., and Metcalfe, A. V. (1999). ''Speed, speed limits and road traffic accidents under free flow conditions.'' *Accid. Anal Prev.*, 31(1–2), 161–168.

CALTRANS. (1993). *Manual of traffic accident surveillance and analysis system*, California Department of Transportation, Sacramento, Calif.

Cedar, A., and Livneh, L. (1982). ''Relationship between road accidents and hourly traffic flow.'' *Accid. Anal Prev.*, 14(1), 19–44.

Dickerson, A., Peirson, J., and Vickerman, R. (2000). ''Road accidents and traffic flows: An econometric investigation.'' *Economica*, 67(265), 101–121.

Frantzeskakis, J. M., and Iordanis, D. I. (1987). ''Volume-to-capacity ratio and traffic accidents on interurban four-lane highways in Greece.'' *Transp. Res. Rec.*, 1112, 29–38.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., and Thomsen, L. K. (1995). ''Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts.'' *Accid. Anal Prev.*, 27(1), 1–20.

Garber, N. J., and Gadiraju, R. (1990). ''Factors influencing speed variance and its influence on accidents.'' *Transp. Res. Rec.*, 1213, 64–71.

Gifi, A. (1990). *Nonlinear multivariate analysis*, Wiley, Chichester, England.

Gwynn, D. W. (1967). ''Relationship of accident rates and accident involvement with hourly volumes.'' *Traffic Q.*, 21(3), 407–418.

Hall, J. W., and Pendleton, O. J. (1989). ''Relationship between V/C ratios and accident rates.'' *Rep. FHWA-HPR-NM-88-02*, U.S. Dept. of Transportation, Washington, D.C.

Israëls, Z. (1987). *Eigenvalue techniques for qualitative DATA*, DSWO Press, Leiden, The Netherlands.

Jansson, J. O. (1994). ''Accident externality charges.'' *J. Transp. Econ. Policy*, 28(1), 31–43.

Johansson, P. (1996). ''Speed limitation and motorway casualties: A time series count data regression approach.'' *Accid. Anal Prev.*, 28(1), 73–87.

Jones-Lee, M. W. (1990). ''The value of transport safety.'' *Oxford Rev. Econ. Poli.*, 6(1), 39–60.

Mensah, A., and Hauer, E. (1998). ''Two problems of averaging arising from the estimation of the relationship between accidents and traffic flow.'' *Transp. Res. Rec.*, 1635, 37–43.

Michailidis, G., and de Leeuw, J. (1998). ''The GIFI system of descriptive multivariate analysis.'' *Stat. Sci.*, 13(4), 307–336.

Newberry, D. (1988). ''Road user charges in Britain.'' *Econom. J.*, 98, 161–176.

O'Reilly, D., Hopkin, J., Loomes, G., Jones-Lee, M., Philips, P., McMa-

hon, K., Ives, D., Sobey, B., Ball, D., and Kemp, R. (1994). ''The value of road safety: UK research on the valuation of preventing on-fatal injuries.'' *J. Transp. Econ. Policy,* 28(1), 45–59.
Sandhu, B., and Al-Kazily, J. (1996). ''Safety impacts of freeway traffic congestion.'' *Presented at Annual Meeting of Transportation Research Board*, January 7–11, Washington, D.C.
Shefer, D., and Rietveld, P. (1997). ''Congestion and safety on highways: Towards an analytical model.'' *Verfahrenstechnik,* 34(4), 679–692.
Stokes, R. W., and Mutabazi, M. I. (1996). ''Rate-quality control method of identifying hazardous road locations.'' *Transp. Res. Rec.,* 1542, 44–48.
Sullivan, E. C. (1990). ''Estimating accident benefits of reduced freeway congestion.'' *J. Transp. Eng.,* 116(2), 167–180.
Sullivan, E. C., and Hsu, C.-I. (1988). ''Accident rates along congested freeways.'' *Research Rep. UCB-ITS-RR-88-6,* Institute of Transportation Studies, Univ. of California, Berkeley, Calif.
Ter Braak, C. J. F. (1990). ''Interpreting canonical correlation analysis through biplots of structure correlations and weights.'' *Psychometrika,* 55(3), 519–531.
van Buren, S., and Heiser, W. J. (1989). ''Clustering N-objects into K-groups under optimal scaling of variables.'' *Psychometrika,* 54(4), 699–706.
van de Geer, J. P. (1986). ''Relationships among *k* sets of variables, with geometrical representation, and applications to categorical variables.'' *Multidimensional data analysis,* J. de Leeuw et al., eds. DSWO, Leiden, The Netherlands.
van der Burg, E., and de Leeuw, J. (1983). ''Non-linear canonical correlation.'' *Br. J. Math. Stat. Psychol.,* 36(1), 54–80.
ver Boon, P. (1996). *A robust approach to nonlinear multivariate analysis*, DSWO, Leiden, The Netherlands.
Vickrey, W. (1969). ''Congestion theory and transport investment.'' *Am. Econ. Rev.*, 59(2), 251–260.
Vitaliano, D. F., and Held, J. (1991). ''Road accident external effects: An empirical assessment.'' *Appl. Econom.*, 23(2), 373–378.
Zhou, M., and Sisiopiku, V. P. (1997). ''Relationship between volume-to-capacity ratios and accident rates.'' *Transp. Res. Rec.*, 1581, 47–52.