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

A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways

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

California PATH Working Paper
UCB-ITS-PWP-2003-12

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.

Report for Task Order 4117

August 2003

ISSN 1055-1417

A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways

Thomas F. Golob

Institute of Transportation Studies
University of California, Irvine

+1.949.824.6287 voice
+1.949.824.8385 fax
tgolob@uci.edu

and

Wilfred W. Recker

Department of Civil and Environmental Engineering *and*
Institute of Transportation Studies
University of California, Irvine

+1.949.824.5642 voice
+1.949.824.8385 fax
wwrecker@uci.edu

August 19, 2003

To appear in
Transportation Research Part A: Policy and Practice

A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways

by

Thomas F. Golob

and

Wilfred W. Recker

Abstract

A method is developed to determine how crash characteristics are related to traffic flow conditions at the time of occurrence. Crashes are described in terms of the type and location of the collision, the number of vehicles involved, movements of these vehicles prior to collision, and severity. Traffic flow is characterized by central tendencies and variations of traffic flow and flow/occupancy for three different lanes at the time and place of the crash. The method involves nonlinear canonical correlation applied together with cluster analyses to identify traffic flow regimes with distinctly different crash taxonomies. A case study using data for more than 1,000 crashes in Southern California identified twenty-one traffic flow regimes for three different ambient conditions: dry roads during daylight (eight regimes), dry roads at night (six regimes), and wet conditions (seven regimes). Each of these regimes has a unique profile in terms of the type of crashes that are most likely to occur, and a matching of traffic flow parameters and crash characteristics reveals ways in which congestion affects highway safety.

1 Background

Understanding the benefits of improved traffic flow (reduced congestion) is critical to the assessment of investments in infrastructure or traffic management and control. Improved flow should lead to reductions in travel time, vehicle emissions, fuel usage, psychological stress on drivers, and improved safety. However, the manner in which safety is improved by smoothing traffic flow is not well understood. The documented research is aimed at shedding light on the complex relationships between traffic flow and traffic accidents (crashes).

The immediate objective of this research is to determine how crashes are related to traffic flow conditions at the time of their occurrence. Crashes here are depicted in terms of the type of the collision (e.g., rear end, sideswipe, hit object), collision location (designated lane or off-road location), number of involved vehicles, movements of these vehicles prior to collision, and severity. Traffic flow is characterized by parameters describing temporal distributions of variables available from single inductive loop detectors, e.g., central tendencies and variations of traffic flow and density for different freeway lanes. The ultimate goal is a safety performance measurement tool that can be used to measure the effects of changes in traffic flow patterns on traffic safety. Such a tool could be used to forecast future conditions or to evaluate the effectiveness of advanced transportation management projects.

Benefit/cost comparisons have long been a standard in assessing the effectiveness of investment of limited resources, and have served as an essential element in determining the most effective allocation of such resources. Developing these comparisons has presented a very perplexing problem in evaluating projects, primarily because hard numbers for benefit/cost ratios associated with traffic management operations cannot be obtained practically. For example, the costs of such management strategies as ramp metering or freeway service patrols (FSP) are easily determined. However, a true measurement of the benefits of these strategies can be determined only by shutting down all of the ramp metering or curtailing FSP for a period of time and measuring any adverse consequences. This direct approach, of course, is not feasible due to liability reasons. This measurement problem is heightened dramatically when issues of safety are involved, yet one of the most compelling arguments for implementation of Intelligent Transportation System (ITS) and Advanced Transportation Management System (ATMS) elements is their presumed enhancement of traffic safety.

Assessment of benefits of ITS and ATMS improvements largely translates into a problem of quantifying the benefits of improved traffic flow. Improved flow ostensibly leads to reductions in travel time, vehicle emissions, fuel usage, psychological stress on drivers, and improved safety. However, the manner in which safety is improved by smoothing traffic flow is not well understood at this time. Due to observed nonlinear relationships between and traffic flow, speed, and density, it is unclear whether projects aimed at mitigating congestion will have a positive or negative effect on safety (Garber and Subramanyan, 2001). This is especially true because safety can be measured in a variety of ways, depending on the choice of accident statistic (e.g., number of injuries,

injury crashes, total crashes, or cost of crashes) and the choice of exposure (e.g., travel distance or time)(Chang, 1982). The present research is aimed at shedding light on the complex relationships between traffic flow parameters and traffic crash parameters.

This research uses a disaggregate approach, in which the units of analysis are the crashes themselves, rather than aggregations of crashes over time and space. Disaggregate analyses represent a new form of traffic safety research, made possible by the need for data in support of ITS developments and the advent of Transportation Management Centers (TMCs), which have led to the availability of archived data on traffic flow from such sensor devices as inductive loop detectors. The three sets of variables in a disaggregate approach are: (a) the characteristics of the crash, (b) the characteristics of the traffic flow the time of the crash, measured at a location as close to the site of the crash as possible, and (3) environmental conditions, such as highway geometry, roadway and weather conditions, and visibility. Aggregate studies have been useful in identifying relationships between crash rates and traffic flow parameters such as mean flow, mean density, mean speed, and speed variance measured in various ways. However, as pointed out by Davis (2002), aggregate studies can be susceptible to the problem of ecological fallacy, in which an observed statistical relationship between aggregated variables is falsely attributed to the units over which were aggregated (Robinson, 1950). In principle, disaggregate studies avoid this problem, but there are several difficulties encountered in matching crashes with data on traffic flow at the time and location of the crash, as discussed in Section 3.

Disaggregate traffic safety studies have been reported by Lee and his colleagues (Lee, Saccomanno and Hellinga, 2002; and Lee, Hellinga and Saccomanno, 2003) and by Oh, *et al.* (Oh, Oh, and Chang, 2001; Oh, Oh, Ritchie and Chang, 2001). These studies have demonstrated how freeway traffic flow conditions prevailing at times and places where crashes occur differ statistically from "normal" conditions. The common goal of these studies was to develop a real-time crash prediction model. Based on analyses of archived data from traffic monitoring devices, combined with historical crash records, several traffic flow measures were identified as precursors of crashes: standard deviations of speed (Oh, et al., 2001), coefficients of variation of speeds compared across lanes, and traffic density (Lee, at al., 2002, 2003). The research documented here and in Golob, Recker and Alvarez (2002) (2003) takes a different approach, but it is also aimed at developing a real-time safety performance monitoring tool. The present approach is to determine how any traffic flow condition on an urban freeway can be classified into mutually exclusive clusters (called Regimes) that differ in terms as much as possible in terms of likelihood of crash by type of crash. The methodology underlying the proposed safety performance monitoring tool is the subject of this paper.

When operational, a real-time crash prediction model will complement existing tools that measure roadway productivity based throughput, average travel time, average speed or total delay (e.g., Chen, *et al.*, 2001; Choe, Skabardonis, Varaiya, 2002; and Varaiya, 2001). Inputs to both the safety and productivity performance tools would be streams of volume and occupancy data from ubiquitous single inductive loop detectors.

2 Methodology

We employ a series of multivariate statistical methods to find patterns in the relationship between crash and traffic flow characteristics. Two of the methods used are linear and well known: (a) principal components analysis, the most common form of factor analysis, and (b) cluster analysis. Principal components analysis (PCA) is used to interpret the correlation structure among traffic flow variables in terms of a smaller number of independent linear combinations, called factors. It is used to reduce the dimensionality of the data by accounting for redundancy among sets of highly correlated traffic flow variables. PCA was developed in the 1930s (Hotelling, 1933) and is covered in all textbooks on multivariate statistical methods.

Cluster analysis is a similarly widely used method of grouping observations based on similar data structure. Here we use cluster analysis to find homogenous groups of traffic flow conditions, which we call "regimes." There are many versions of cluster analysis, as described in textbooks such as Hartigan (1975) and Everitt (2001). We employ the non-hierarchical clustering algorithm known as k-means clustering (MacQueen, 1967). This algorithm starts with a fixed number (k) of random clusters, and then moves objects between those clusters with the goal of minimizing variability within the clusters and maximizing variability between the clusters. This is analogous to analysis of variance in reverse. Forcing passes and random restarts are used to ensure that determination of a global. One problem is to determine k, the best number of "natural" clusters. We use a unique method for finding the optimal number of clusters by comparing how well each clustering solution for traffic flow regimes explains crash typology. To measure the strengths of the relationships between different clustering solutions and crash characteristics, we employ a third type of multivariate analysis: nonlinear (nonparametric) canonical correlation analysis (NLCCA).

NLCCA needs some explanation because it is not commonly used in transportation research. Conventional linear canonical correlation analysis (CCA) can be viewed as an expansion of regression analysis to more than one dependent variable; there are two sets of variables, and the objective is to find a linear combination of the variables in each set so that the correlation between the linear combinations is as high as possible. The linear combinations are defined by optimal variable weights. Depending on the number of variables in each set and their scale types, further linear combinations (canonical variates, similar to principal components in factor analysis) can be found that have maximum correlations subject to the conditions that all canonical variates are mutually orthogonal or independent. CCA can also be generalized to more than two sets of variables, and with a single set of variables, CCA is essentially equivalent to principal components analysis.

NLCCA is designed for problems with variable sets that contain categorical or ordinal (nonlinear, or nonparametric) variables. The linear combinations can be defined only when there is a metric to quantify the categories of each nonlinear variable. NLCCA simultaneously determines both (1) optimal re-scaling of the categories of all categorical

and ordinal variables and (2) component loadings (variable weights), such that the linear combination of the weighted re-scaled variables in one set has the maximum possible correlation with the linear combination of weighted re-scaled variables in the second set. In our applications, one set of variables is comprised of a single categorical variable defining traffic flow regimes, while the other set of variables is always composed of categorical variables describing crash characteristics.

The NLCCA method we use is based on the alternating least squares (ALS) algorithm, which is described in detail in De Leeuw (1985), Gifi (1990), Michailidis and de Leeuw (1998), Van der Burg (1988), van Buren and Heiser (1989) and van der Boon, 1996). In ALS both the variable weights and optimal category scores are determined by minimizing a meet-loss function derived from lattice theory. ALS includes category quantifications for each variable as well as each variable's component loading, as parameters in the objective function. The meet-loss objective function is minimized by means of an algorithm that iterates between adjusting the category scores of the ordinal and nominal variables and adjusting the variable weights, subject to appropriate constraints. In many ways, the ALS algorithm is similar to the power method in singular value decomposition (Gifi, 1990; Israëls, 1987), which underlies most linear multivariate methods, such as principal components analysis and discriminant analysis. NLCCA output includes several overall measures of goodness-of-fit, component loadings, and optimal category scores. The selection of the number of canonical variates is based upon comparing the decay in the goodness of fit associated with each additional dimension, similar to the selection of the number of factors in factor analysis.

Hensher and Golob (1999) use a geometric perspective to describe NLCCA as applied in transportation research. Component loadings, in the absence of missing data, are equivalent to product-moment correlations between the optimally scaled variables and the canonical variates (similar to factor loadings in principal components analysis). Geometrically, the sum of squared loadings (the length of the vector from the origin to the component loadings of a given variable in the orthogonal space of the canonical variates) indicates how much of the variable was explained by the canonical variates in total, and the square of the projections onto an axis reveals how much of the explanation was due to that canonical variate. For any two variables, the scalar (dot) product of the two vectors is an approximation of the correlation between the two optimally scaled variables (Ter Braak, 1990; van de Geer, 1986). These NLCCA results are useful in interpreting the relationships between traffic flow characteristics and crash typology.

3 The Data

3.1 Fusion of Crash and Traffic Flow Data

Accident data were drawn from the TASAS database (Caltrans, 1993), covering police-reported crashes that occurred on mainline sections of the six major freeways of Orange County California for the year 1998. Orange County is an urban area of about

three million population located between Los Angeles and San Diego. There were 9,341 crashes on these six freeway routes in 1998. From these, a sample of 1,192 crashes, 12.8% of the 9,341 highway crashes, was selected based on having sufficient corresponding valid loop detector data to perform the analysis – the criterion being having ostensibly valid loop detector data for a full 30 minutes preceding the accident for three designated lanes at the nearest detector station. Loop data for the 2.5-minute period immediately preceding the time of the accident were discarded, because accident times are typically rounded off to the nearest five minutes.

The loop detector data come from an archived database of 30-second observations from single inductance loop detectors maintained throughout the State Highway System. Each observation provides count and occupancy time for a 30-second time slice. At each mainline loop detector station, data are available for each lane, and there were at least three lanes in each direction on our freeways. In order to standardize traffic flow data for all crashes 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 being the number one lane according to standard nomenclature; (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 corresponding total number of loop detector observations sought for the analysis reported here is given by the product of 9,341 crashes, 55 time slices, and 3 lanes per location, or 1,541,265 distinct 30-second counts and occupancies. For this sample, the average distance from the crash location to the closest detector station is 0.17 miles and the median distance is 0.12 miles; 78% of these crashes were located within 0.25 miles of the detector station.

The following major crash characteristics available in the TASAS dataset were used in this analysis: (1) the type of collision (rear-end, sideswipe, broadside, head-on, overturn), (2) the location of the collision involving each vehicle (e.g., left lane, interior lanes, right lane, right shoulder area, off-road beyond right shoulder area), and (3) injuries and fatalities per vehicle. Other factors included in the data base were either treated as separate dimensions of the problem, e.g., such environmental conditions as lighting, weather, and pavement conditions, or as embellishments to a major crash characteristic, e.g., movement prior to collision and number of vehicles involved. For example, a new six-category coding of collision type was constructed that incorporates the most important information from three TASAS variables: collision type, movement prior to collision, and number of vehicles. The new coding avoids problems of structural relationships if the three TASAS variables were used separately (e.g., rear-end and sideswipe crashes by definition involve more than one vehicle, and rear-end crashes almost always involve a vehicle slowing or stopped.

3.2 Segmentation Based on Weather and Lighting Conditions

The data can be used to distinguish six sets of environmental conditions, defined by the combination of weather and ambient lighting. In Golob and Recker (2003), we report on an application of NLCCA to determine how crash typology is related to weather and

ambient lighting conditions. The exogenous variables represented weather and lighting conditions with five categories defined by all combinations of two types of road conditions (dry and wet), combined with three lighting conditions (daylight, darkness, and dusk-dawn), with the exception of wet dusk-dawn crashes, for which there were too few crashes to analyze. The other side of the problem was composed of the three crash characteristics listed in Table 1 (Collision Type, Collision Location, and Severity). The objective was to determine similarities among the five segments of weather and ambient lighting conditions in terms of their explanation of the crash typology. We found that it is best to combine the Wet-Night and Wet-Day segments into a single Wet segment, and to combine the relatively sparse Dry-Dusk-Dawn segment can be combined with the adjacent Dry-Day segment (Golob and Recker, 2003). The resulting segmentation is: (1) Dry-Day (including Dusk-Dawn): 819 crashes, (2) Dry-Night, 217 crashes, and (3) Wet (any lighting condition): 156 crashes. The breakdowns of the three crash characteristics for each weather and lighting segment are listed in Table 1.

Table 1 Characteristics of the Crashes that Occurred under Three Conditions (percentage breakdowns)

Crash Characteristic	Daylight-Dry N = 819	Dark-Dry N = 217	Wet road N = 156
Collision type			
Single vehicle hit object or overturn	10.5%	18.9%	26.9%
Multiple vehicle hit object or overturn	5.6%	6.5%	6.4%
Two-vehicle weaving crash [a]	17.8%	20.3%	25.6%
Three-or-more-vehicle weaving crash [a]	5.1%	4.1%	9.6%
Two-vehicle straight-on rear end	38.2%	30.0%	16.0%
Three-or-more-vehicle straight-on rear end	22.7%	20.3%	15.4%
Total	100.0%	100.0%	100.0%
Collision Location			
Off-road, driver's left	12.3%	11.5%	25.0%
Left lane	30.4%	15.7%	16.0%
Interior lane(s)	32.5%	32.3%	34.6%
Right lane	18.7%	26.3%	12.8%
Off road, driver's right	6.1%	14.3%	11.5%
Total	100.0%	100.0%	100.0%
Severity			
Property damage only	75.0%	70.5%	57.7
Injury or fatality	25.0%	29.5%	42.3%
Total	100.0%	100.0%	100.0%

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

3.3 Traffic Flow Variables

We use raw detector data to provide information on two variables: count and occupancy for each thirty-second interval. 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 point speeds, we avoid making any such assumptions, and use only these direct measurements in the analysis that follows. After testing different lengths of time for monitoring of traffic conditions, we determined that we needed approximately 30 minutes of 30-second observations at the loop detector station closest to the location of the accident to establish stable measures of traffic conditions prior to the accident.

Based on preliminary analyses, four blocks of three variables (one measure for each of the three lane type designations, left, interior, and right) were identified as being potentially related to taxonomy of crash. The first of these blocks is an indicator of prevailing traffic speed, the second the temporal variation of the prevailing speed, the third the traffic flow, and the fourth the temporal variation in the traffic flow. The four blocks of three variables are listed in Table 2.

Table 2 Traffic Flow Variables

Block 1 Central tendency of flow/occupancy	Median flow/occupancy - left lane
	Median flow/occupancy - interior lane
	Median flow/occupancy - right lane
Block 2 Variation in flow/occupancy	Difference between 90^{th} and 50^{th} percentiles of flow/occupancy – left lane
	Difference between 90^{th} and 50^{th} percentiles of flow/occupancy – interior lane
	Difference between 90^{th} and 50^{th} percentiles of flow/occupancy – right lane
Block 3 Central tendency of flow	Mean flow left lane
	Mean flow interior lane
	Mean flow right lane
Block 4 Variation in flow	Standard deviation of flow left lane
	Standard deviation of flow left lane
	Standard deviation of flow left lane

For the first block, median, rather than mean, is used to measure the central tendency of this proportional indicator of space mean speed in order to avoid the influence of outlying observations. Similarly, we use the difference of the 90^{th} and 50^{th} percentiles to measure variation in the proportional indicator of speed, in order to minimize the influence of potentially invalid outlying observations. Because flow is not as sensitive to

outliers (ranging from zero through twenty-five vehicles per 30-second interval), we use mean and standard deviation, respectively, to measure the central tendency and variation of traffic flow.

4 Covariance Structure of the Traffic Flow Variables

To avoid multicollinearity problems further along in the analyses, principal components analysis (PCA) was performed on the twelve traffic flow variables for each of the three segments. Our objective was to extract a sufficient number of factors to identify independent traffic flow variables while simultaneously discarding as little of the information in the original variables as possible. For the segment of crashes that occurred during daylight or dusk-dawn on dry roads, we found that six factors accounted for 86.8% of the variance in the original twelve variables. The six-factor PCA solution is invariant under orthogonal rotations; varimax rotation, a standard technique in factor analyses, was performed to aid in interpreting these factors. The rotation results in a redistribution of the explanatory power of each factor while preserving the cumulative variance explained by all retained factors. The factor loadings, which are the correlations between the original variables and the rotated factors, are listed in Table 3. Also listed in Table 3 are the variances accounted for by each factor. For ease of interpretation, one variable was then selected to represent each factor in the subsequent stages of the analysis. The minimum correlation between a factor and its representative variable is 0.85 (in the case of the third factor), the maximum is 0.920, and the mean is 0.900. These correlations indicate that the representative variables are relatively independent and are good substitutes for the factors.

The factor loadings show that the central tendency of flow/occupancy (Variable Block 1) is consistent across all three lanes. Based on consistent PCA results for the other two lighting and weather segments (reported in Sections 4.2 and 4.3), the variable chosen to represent this central tendency of flow/occupancy factor is median flow/occupancy in the interior lane. The correlation between this variable and its factor is shown underlined in bold in Table 3.

A single factor also encompasses the central tendency of flow (Variable Block 3) in all three lanes, but the factor is more representative of flows in the left and interior lanes than in the right lane, as witnessed by the lower correlation between this factor and right lane mean flow (0.635). Mean flow in the left lane was chosen to represent this factor in all further analyses. (Although the factor loading for mean flow in the interior lane is greater, its higher correlations with other factors, not shown, resulted in the choice of flow in the left lane to represent this factor.)

Factor three represents the temporal variation in flow/occupancy on the left and interior lanes only. Variation in flow/occupancy in the right lane is captured by a separate, sixth, factor. Variation in flow/occupancy is represented by variation in the flow to occupancy ratio on the left and interior lanes in further analyses, and variation in flow to occupancy

ratio for the right lane represents the sixth factor. We interpret this to mean that the variation in flow/occupancy in the rightmost lane, which may be influenced significantly by merging behavior in the vicinity of freeway on- and off-ramps, relates to crash characteristics in a fundamentally different way than does the variation in flow/occupancy that is attributable primarily to mainline freeway flow.

Table 3 Rotated PCA Loadings for Traffic Flow Variables for the Daylight, Dry Roads (showing only loadings with absolute values > 0.3)

		Principal component					
	Traffic flow variable	1	2	3	4	5	6
	Percentage of original variance accounted for	22.5%	17.8%	14.6%	13.6%	9.4%	9.0%
Block 1	Median flow/occupancy left lane	0.904					
	Median flow/occupancy interior lane	**0.892**					
	Median flow/occupancy right lane	0.921					
Block 2	Variation in flow/occupancy left lane	-.308		0.832			
	Variation in flow/occupancy interior lane			**0.853**			
	Variation in flow/occupancy right lane						**0.911**
Block 3	Mean flow left lane		**0.920**				
	Mean flow interior lane		0.929				
	Mean flow right lane		0.635			0.392	-.418
Block 4	Variation in flow left lane				**0.902**		
	Variation in flow interior lane				0.821	0.323	
	Variation in flow right lane					**0.914**	

Finally, the PCA results also show that temporal variations in flows on the three lanes are partitioned into two factors: variations in flow on the left and interior lanes (factor 4), and variation in flow on the right lane (factor 5). The left lane is again chosen to represent the former factor. Here again, the implication is that the flow in the rightmost lane, which has a direct influence on the level of service in the vicinity of freeway on- and off-ramps, relates to crash characteristics in a fundamentally different way than does the mainline freeway flow.

A similar PCA was performed for all crashes that occurred during darkness on dry roads. Here six factors account for 87.7% of the total variance in the twelve original variables, a slightly more effective solution than for daylight crashes. (For purposes of brevity, we decline to show the factor loadings and breakdown of the explained variance, which can be found in Golob, Recker and Alvarez, 2002) The factor structure

is similar to that found for daylight conditions on dry roads, and the variables chosen to represent the factors are identical.

A third and final PCA was performed for all crashes that occurred on wet roads. Results show that the correlation structure among the twelve traffic flow variables is nearly identical for the three weather and lighting conditions. Here six factors account for 87.1% of the total variance in the twelve original variables, versus 86.8% and 87.7% for dry-daylight and dry-nighttime, respectively (Golob, Recker and Alvarez, 2002). The factor structure is essentially the same as found previously, so the same six variables were chosen to represent the factors.

5 Traffic Flow Regimes

Cluster analyses were performed in the space of the six traffic flow variables representative of the principal components in order to establish relatively homogenous traffic flow regimes. A k-means clustering algorithm was used. The objective was to determine the best grouping of observations into a specified number of clusters, such that the pooled within groups variance is as small as possible compared to the between group variance given by the distances between the cluster centers. We repeated runs of the clustering algorithm with different initial cluster centers to avoid local optima.

The optimal number of clusters is usually determined by inspecting various clustering criteria, most of which are developed from eigenvalues of the characteristic equation involving the ratio of the pooled within-groups and between-groups variance matrices. Two of the commonly used criteria are: (1) Wilk's Lambda, given by the ratio of the determinants of the within-groups and total variance matrices (equivalent to the product of the eigenvalues of the characteristic equation), and (2) Hotelling's Trace, given by the sum of these eigenvalues (Everitt, 2001). Selection of the optimal number of groups using such criteria is relatively arbitrary; as in many applications with well-distributed continuous data on many variables, there is no natural number of clusters based on clustering criteria alone. However, in the present application, we can apply an external criterion to the clustering problem to identify the optimal number of clusters. We conducted nonlinear canonical correlation analysis (NLCCA) for each clustering solution (from 4 to 18 clusters). The NLCCA problem was configured with the multiple-nominal cluster variable on one side and the three single nominal crash variables described in Table 1 on the other side. The criteria that describe how well each of the cluster variables explained the crash characteristics are the canonical correlations between the two sets of variables, one for each of the variates of the two-dimensional solution. The results of these analyses for each of the three environmental conditions are described below.

5.1 Traffic Flow Regimes for Crashes During Dry-Day Conditions

The results of the two-dimensional NLCCA solution for the dry-day segment are displayed in Figure 1. Canonical correlation for the first dimension reaches a maximum at eight clusters. The fit for the second dimension has a local maximum at eight clusters and then does not improve until the 13-cluster level is reached. Based on these results, and corroborative evidence from Wilk's Lambda and Hotelling's Trace (not shown), eight clusters (representing eight distinct traffic flow regimes, hereafter simply referred to as "Regimes") were selected.

The distribution of the 819 dry-day crashes over the eight regimes is as shown in Table 4, together with a brief qualitative description of each regime based on the deviation of the regime center (in terms of standard deviations) from each variable's grand mean for the entire sample of dry daylight crashes (N = 819).

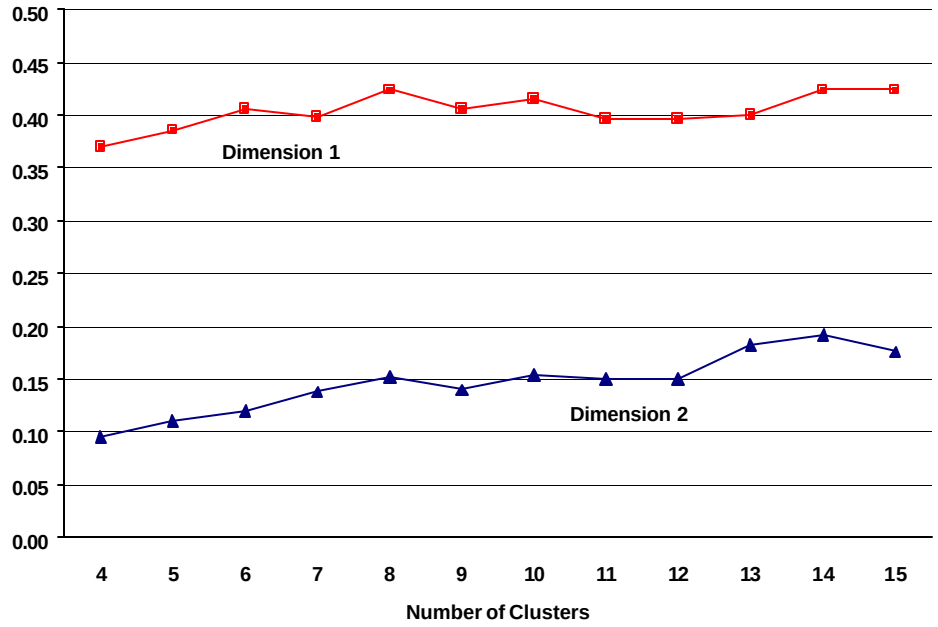

Figure 1 Canonical Correlations for Two-dimensional Nonlinear Discriminant Analysis Solutions for Different Number of Clusters – Daylight, Dry Roads

Table 4 Summary of the Eight Traffic Flow Regimes, in Order of Mean Flows (from lowest to highest) – Daylight, Dry Roads

Regime	Traffic flow conditions	Dry-Day Crashes	
		N	%
D1	**Light free-flow:** Very low flow, high mean flow/occupancy, low variance of flow/occupancy in the right lane and about average variances of flow/occupancy in the other lanes.	71	8.7
D2	**Heavily congested flow:** Low flow and very low flow/occupancy. Low variances of flow in all lanes. Low variance of flow/occupancy, particularly in right lane.	68	8.3
D3	**Congested flow:** Moderately low mean flow and low mean flow/occupancy. High variances in flows and high variance in flow/occupancy except for the right lane.	99	12.1
D4	**Light, right-variable flow:** High mean flow/occupancy and moderately low mean flows. Left and interior lanes free-flowing, but right lane flow/occupancy variance high and flow variance low.	85	10.4
D5	**Flow at capacity:** Very high variances in flow/occupancy, average flows and variances in flow, and moderately low mean flow/occupancy.	159	19.4
D6	**Heavy, variable flow:** Very high flow variances, particularly in the right-lane, and moderately high flows. High mean flow/occupancy and relatively low flow/occupancy variances.	148	18.1
D7	**Heavy, steady flow:** High flow and high mean flow/occupancy, with low temporal variances of flow/occupancy on all lanes and near-average flow variances.	81	9.9
D8	**Flow near capacity:** High flow, and low flow variances. Mean flow/occupancy and flow/occupancy variations about average to moderately below average.	108	13.2

5.2 Traffic Flow Regimes for Crashes During Dry, Dark Conditions

Once again we clustered the crashes in the space of the six variables. The optimal number of clusters is six clusters, based on the internal clustering criteria (Wilk's Lamba and Hotelling's Trace) and the explanation of crash characteristics provided by the NLCCA. Detailed results can be found in Golob, Recker and Alvarez (2002). The distribution of the 217 dry-darkness crashes over the six regimes is as shown in Table 5, together with a brief qualitative description of each regime based on the deviation of

the regime means (in terms of standard deviations) from each variable's grand mean for the entire sample of dry daylight crashes (N = 217).

Table 5 Summary of the Six Traffic Flow Regimes, in Order of Mean Flow (from lowest to highest) – Nighttime, Dry Roads

Regime	Traffic flow conditions	Dry-Dark Crashes	
		N	%
N1	**Very light free-flow:** Very low mean flow and low variances in flow. High mean flow/occupancy and high variances in flow/occupancy on all lanes.	49	22.6
N2	**Light free-flow:** High mean flow/occupancy and moderately low flow/occupancy variances. Moderately low flow and low variance of flow in right lane.	47	21.7
N3	**Conservative nighttime driving:** Low mean flow/occupancy. Low variances of flow/occupancy. Average mean flow (for periods of darkness) and average variances of flow.	23	10.6
N4	**Sporadically congested flow:** Low mean flow/occupancy. High variances of flow/occupancy in interior lanes. Moderately high flow (for periods of darkness) and high variances of flow in all lanes.	30	13.8
N5	**Heavy, variable flow:** High flow and very high variances of flow in all lanes. Moderately high mean flow/occupancy and low variance of flow/occupancy.	32	14.7
N6	**Flow near capacity:** Very high flow. Slightly below average mean flow/occupancy and flow/occupancy variations. Also slightly below average variations in flows.	36	16.6

5.3 Traffic Flow Regimes for Crashes During Wet Conditions

For crashes on wet roads, the optimal number of clusters in the space of the six variables is seven, based on the internal clustering criteria and the explanation of crash characteristics. The explanation of crash characteristics peaks at seven clusters, and the seven-level cluster is also consistent with a break in Hotelling's Trace criteria, as described in Golob, Recker and Alvarez (2002).

The distribution of the 154 wet road collisions over the seven regimes is shown in Table 6; brief summaries of the flow characteristics of these regimes are also provided. Once again, the regimes are labeled in order of increasing mean flow.

Table 6 Summary of the Seven Traffic Flow Regimes in Order of Mean Flow (from lowest to highest) – Wet Road Crashes

Regime	Traffic flow conditions	Wet Crashes	
		N	%
W1	**Very light flow, variable flow/occupancy:** Very low flow and very low variations in flow. Mean flow/occupancy slightly below average for wet roads. Variations in flow/occupancy high, especially for right lane.	26	16.9
W2	**Light free-flow:** Low mean flow and moderately high flow/occupancy. Low variances in flow and flow/occupancy in right lane.	22	14.3
W3	**Moderate free-flow:** Moderately high flow/occupancy and near average flow, flow variances, and flow/occupancy variances for wet roads.	27	17.5
W4	**Moderate flow with right-lane concentration:** Moderately high flow/occupancy and near average flow, but high variance of flow and low variance of flow/occupancy in right lane.	26	16.9
W5	**Heavy, variable flow:** Moderately high flow/occupancy and flow. Very high variance of flow in left lane and high variance of flow in right lane.	13	8.4
W6	**Very heavy flow:** High flow and mean flow/occupancy with low variances in flow/occupancy. High flow variances, especially in left lane.	15	9.7
W7	**Flow approaching capacity:** Low flow/occupancy and high flow. Average to slightly below average variances of both flow/occupancy and flows.	25	16.2

6 Crash Taxonomy Explained by Traffic Flow Regime

Three separate NLCCA were then performed to determine how each nominal traffic regime variable (for dry-day, dry-night, and wet-road conditions) accounted for patterns in the three crash characteristics. Another way to view the problem is to ask how the crash characteristics distinguish among traffic flow regimes. NLCCA with a single categorical (segmentation) variable in one set is equivalent to nonlinear (nonparametric) discriminant analysis.

6.1 Crash Characteristics for Dry-Day Traffic Flow Regimes

The relationships between the traffic flow regimes and the categories of each of the three crash characteristics is captured graphically by a joint plot of the locations of the category centroids of each variable in the space of the canonical variates. Each canonical variate is a latent (unobserved) variable defined solely by the linear combinations of the optimally scaled variables, and the graphs of the category centroids in Figures 2 through 4 are used to interpret the variates (Ter Braak, 1990).

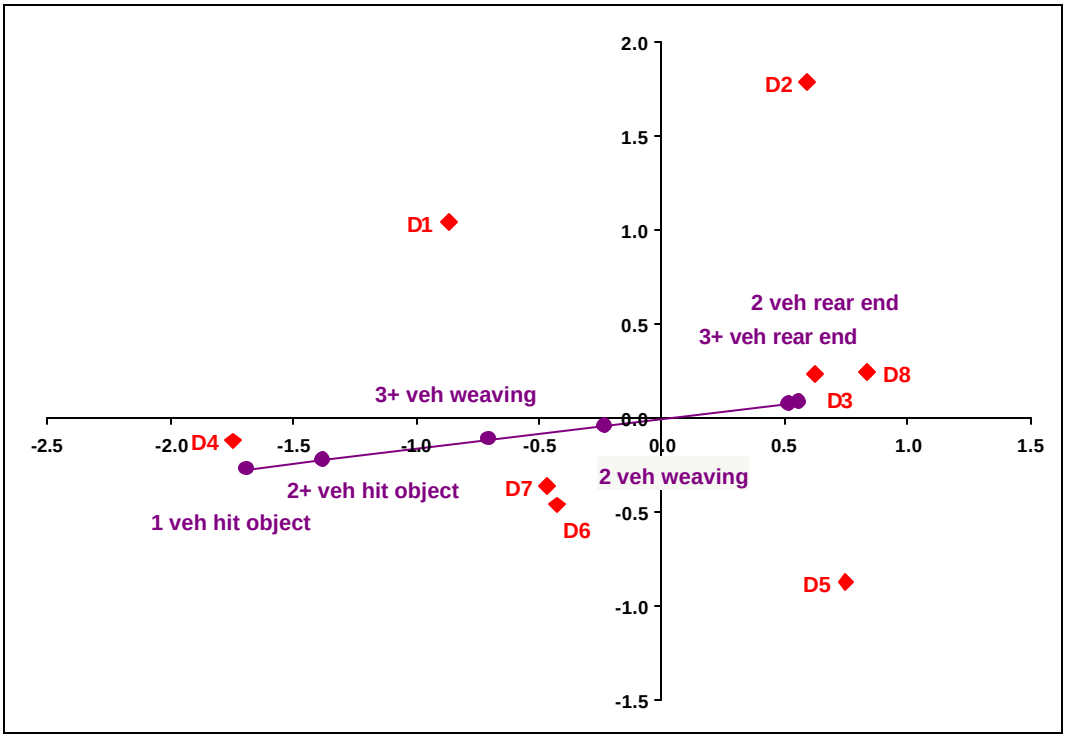

Figure 2 Category Centroids for the Traffic Flow Regime and Collision Type Variables – Daylight, Dry Roads

Focusing first on the locations of the traffic regimes in the two-dimensional space of the canonical variates, which is constant in Figures 2 through 4, we see that the first canonical variate, the x-dimension in these Figures, captures primarily (negative) mean flow/occupancy, and secondarily flow. In the negative domain of the first variate, the regimes are ordered from low to high in terms of decreasing mean flow/occupancy in the middle lane (D4, then D1, then D7 and D6). The four regimes that score in the positive domain of the first variate are more similar to one another; they all represent heavy traffic, and their ordering from low to high is according to mean flow, rather than mean flow/occupancy. The first dimension captures aspects of the density (concentration) dimension of the fundamental diagram of traffic flow versus traffic density (Prigogine and Herman, 1971).

The second canonical variate, which is independent of the first in terms of its functional relationships with the two sets of variables, primarily distinguishes high-flow Regimes D5, D6 and D7, from low-flow Regimes D2, and D1. This dimension captures, to a considerable degree, the flow dimension of the fundamental diagram.

The relationship between traffic flow regime and crash type is depicted in Figure 2. Collision type is almost entirely explained by the first canonical variate, which resembles the density dimension of the fundamental diagram. The optimal scaling of the crash type categories contrasts hit-object versus rear-end crashes, with weaving crashes in between. Thus, as expected, rear-ends are associated with high-density traffic; hit-object crashes are associated with low-density traffic. Weaving crashes (sideswipes and rear-ends caused by lane-change maneuvers) are associated with intermediate density traffic. High-density Regimes D8, D3 and D5 are most associated with rear-end crashes, while low-density Regimes D4 and D1 are associated with hit-object crashes. Intermediate-density Regimes D6 and D7 are most associated with crashes involving weaving maneuvers.

Both dimensions explain collision location (Figure 3), with the second canonical variate being stronger. We can interpret this to mean that collision location is primarily a flow phenomenon, and secondarily a density phenomenon. The optimal scaling of the categories of the location variable shows that left-lane crashes are associated with high density and high flow conditions, while other locations, especially interior lane crashes, are associated with low density and low flow conditions. Regime D5 is associated with left lane crashes, while Regimes D1 and D4 are associated with off-road crashes.

Both dimensions also explain crash severity (Figure 4), on an approximately equal basis. Thus, the difference between property-damage and injury crashes is a function both of flow and density. Injury crashes are more likely to occur in lower density conditions, and in higher flow conditions. Regimes D2 and D4 have the most extreme projections onto the vector defined by the category quantifications of the severity variable. Thus, the NLCCA model predicts that Regime D4 will have a higher proportion of injury crashes, and Regime D2 will have a higher proportion of property-damage-only crashes.

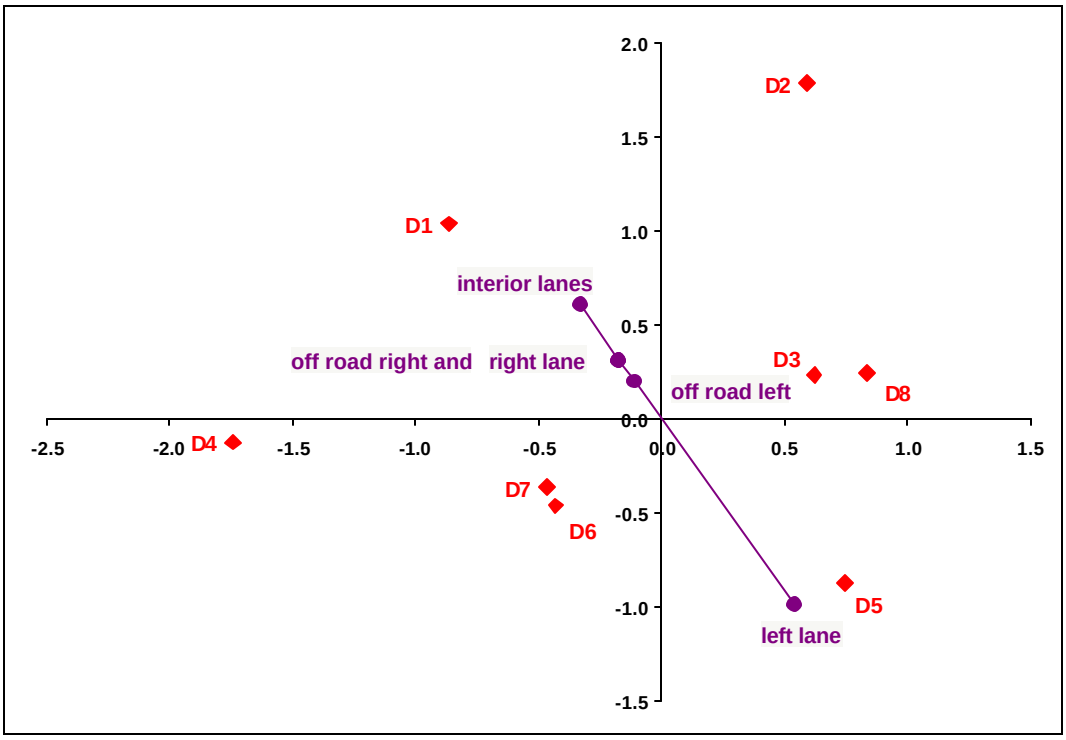

Figure 3 Category Centroids for the Traffic Flow Regime and Collision Location Variables – Daylight, Dry Roads

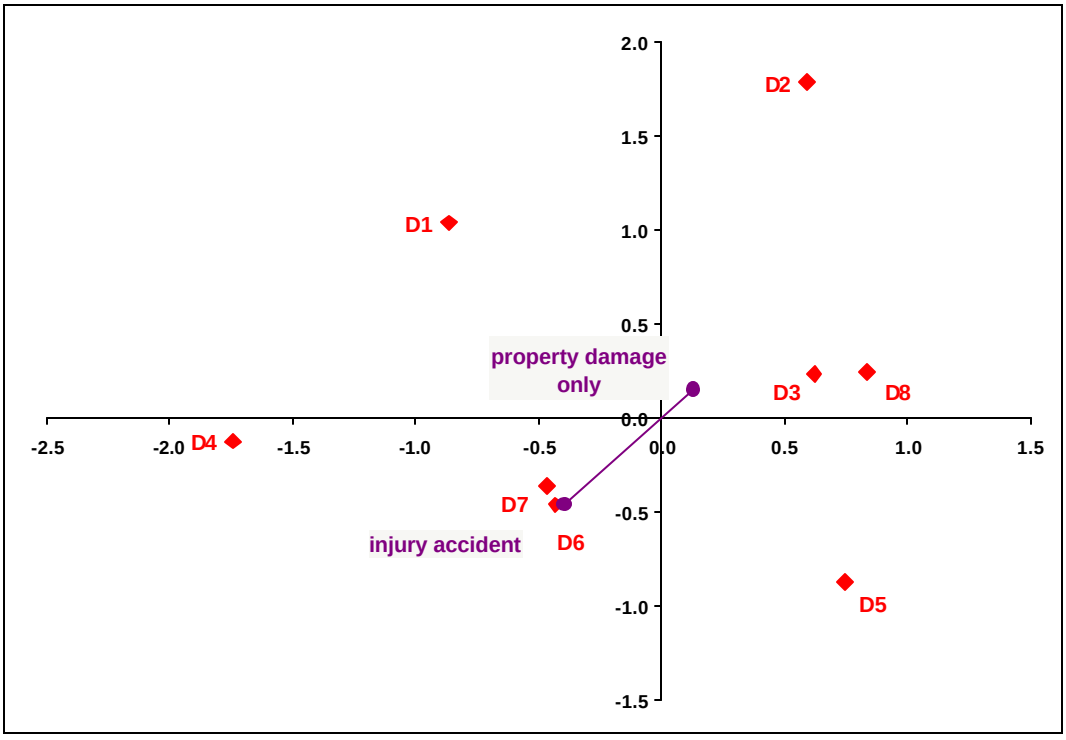

Figure 4 Category Centroids for the Traffic Flow Regime and Crash Severity Variables – Daylight, Dry Roads

The results of the NLCCA model were verified and refined by cross-tabulating each crash characteristic against the eight-category regime segmentation variable. The results were consistent. The traffic flow and crash conditions that define the eight traffic flow regimes for daylight, dry road conditions are summarized in Table 7.

Table 7 Distinguishing Crash Typology for Daylight, Dry Road Traffic Flow Regimes

Regime	Relatively more common types	Relatively less common types
1. Light free flow	27% run-offs (versus 16% overall) 11% off-road right (v. 6%)	11% 3+ vehicle rear-ends (v. 23%) 13% left-lane (v. 30%)
2. Mixed free flow	44% run-offs (v. 16%) 18% off-road right (v. 6%) 25% off-road left (v. 12%) 38% injury (v.25%)	28% rear-ends (v. 61%) 11% left lane (v. 30%)
3. Heavy, variable free flow	Near average distribution of collision types, locations and severity	
4. Flow approaching capacity	9% 3+ vehicle weaving (v. 5%) 19% 1 vehicle run-offs (v.11%)	
5. Heavy flow at moderate speed	52% 2 vehicle rear-ends (v. 38%)	4% run-offs (v. 16%) 1% off-road right (v. 6%)
6. Variable-speed congested flow	31% 3+ vehicle rear-ends (v. 23%) 45% left lane (v.30%)	4% 1 vehicle run-offs (v. 11%) 3% off-road right (v. 6%)
7. Variable-volume congested flow	50% 2 vehicle rear-ends (v. 38%)	5% run-offs (v. 16%)
8. Heavily congested flow	87% property damage only (v. 75%) 41% interior lane(s) (v. 32%)	3% run-offs (v. 16%)

6.2 Crash Characteristics for Dry-Nighttime Traffic Flow Regimes

NLCCA of the 6-category traffic regime variable versus the three crash characteristics again shows how the traffic flow regimes are related to patterns of crash characteristics. A two-dimensional NLCCA solution yielded canonical correlations of 0.526 for the first canonical variate and 0.278 for the second variate.

The relationship between traffic flow regime and crash type is depicted in Figure 5. We can interpret the two canonical variates (dimensions) based on the positions of the six traffic flow regimes in Figure 5. The most important canonical variate, the x-dimension, primarily contrasts Regimes N6 and N4 against Regime N1. It is consistent with the

flow dimension of the fundamental diagram. The y-dimension, which primarily distinguishes Regime N3 from all other regimes, is consistent with the density (concentration) dimension of the fundamental diagram. These two dimensions are similar to the canonical variates found for daylight, dry conditions, but they are reversed in terms of explanatory power. Density is more important than flow in explaining the effects of traffic on the types of crashes that occur during the day on dry roads, while flow is more important than density in explaining the effects of traffic on the types of crashes that occur at night on dry roads.

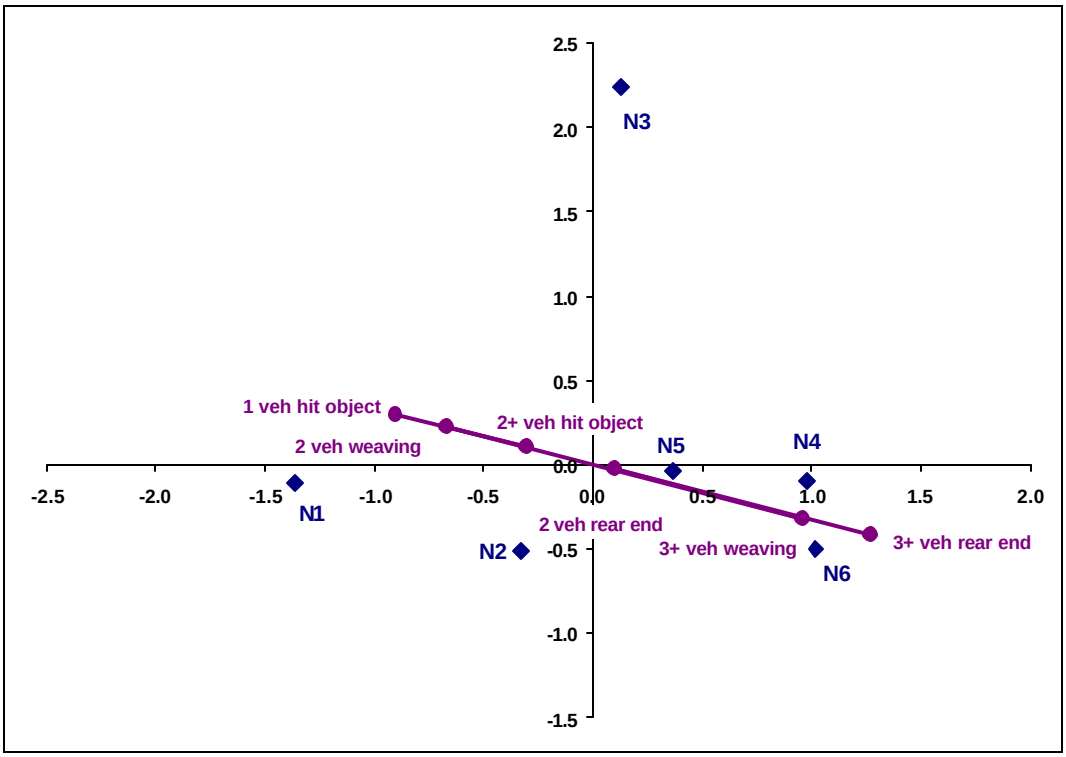

Figure 5 Category Centroids for the Traffic Flow Regime and Collision Type Variables – Nighttime, Dry Roads

As in the case of dry, daylight conditions (Figure 2), collision type is primarily explained by the first canonical variate, which in this case is consistent with the flow dimension of the fundamental diagram. We found previously that collision type for daylight conditions (Figure 2) was explained more by the canonical variate that was associated with traffic density. The optimal scaling of the collision type variable is also different for day and night. For nighttime conditions, the optimal scaling of the crash type categories contrasts single-vehicle hit-object (low flow) versus three-plus vehicle rear-end and weaving crashes (high flow), with two-vehicle crashes in between (Figure 5). The scaling for nighttime conditions is based more on the number of vehicles involved in the collision. For daylight conditions, the optimal scaling of type categories contrasts hit-

object (low density) versus rear-end (high density) crashes, with weaving crashes in-between (Figure 2). The scaling for daylight conditions is based more on kind of the collision, rather than the number of vehicles involved. Three-or-more-vehicle crashes are associated with Regimes N4 and N6, while single-vehicle crashes are associated with Regime N1.

Interpretations relative to crash location and severity were obtained using a similar analysis (Golob, Recker and Alvarez, 2002), but are omitted from detailed discussion. A summary of the traffic flow conditions and associated crash typology is presented in Table 8.

Table 8 Distinguishing Crash Typology for Nighttime, Dry Road Traffic Flow Regimes

Regime	Relatively more common types	Relatively less common types
1. Very light free flow	54% run-offs (versus 26% overall) 35% off-road right (v. 14%)	17% rear-ends (v. 50%) 2% left-lane (v. 16%)
2. Light free flow	38% injury (v.30%)	
3. Heavy, variable flow	9% 3+ vehicle weaving (v. 4%) 25% left lane (v. 16%)	
4. Flow approaching capacity	44% 3+ vehicle rear-ends (v. 20%) 42% right lane (v. 26%)	19% injury (v. 30%)
5. Sporadically congested flow	37% 3+ vehicle rear-ends (v. 20%)	3% run-offs (v. 26%) 3% off-road right (v. 14%)
6. Congested flow	39% 2 vehicle weaving (v. 20%) 48% 2 vehicle rear-ends (v. 30%) 52% interior lane(s) (v. 32%)	

6.3 Crash Characteristics for Wet-Road Traffic Flow Regimes

A two-dimensional NLCCA solution of the seven-category traffic regime variable for wet roads versus the three crash characteristics yielded canonical correlations between the two sets of variables of 0.532 (first canonical variate) and 0.298 (second canonical variate). The optimally scaled category centroids are plotted in Figure 6.

The first variate (the x-dimension) contrasts low flow regimes (W3, W1 and W2) against the high flow regime with low mean flow/occupancy, W7. Regimes W4, W5 and W6, which have moderate to heavy flows but average, or slightly above average flow/occupancy, score close to zero on this first variate. This variate may be capturing a

measure of exposure that is independent of traffic stream flow/occupancy effects. The second canonical variate contrasts Regime W6, then W2 and W5 (negative scores), against Regime W3, then W7 and W1 (positive). No interpretation for this variate is obvious; perhaps, this is an artifact of the relatively small sample size for this category of environmental conditions. Also, the nature of the environmental conditions is a potentially confounding effect for this particular segmentation. The category is defined by a simple binary variable (wet vs. dry): light rain conditions are indistinguishable from heavy downpours; it also covers the complete spectrum of lighting conditions. So, for example, although it is plausible to interpret that Regime W1 typifies light flow conditions under extreme weather (since the ratio of the "very low" flow to occupancy infers mean flow/occupancy slightly below average), only the traffic flow variables are directly measurable; any hypothesis regarding causal effects of weather on these variables requires additional environmental information.

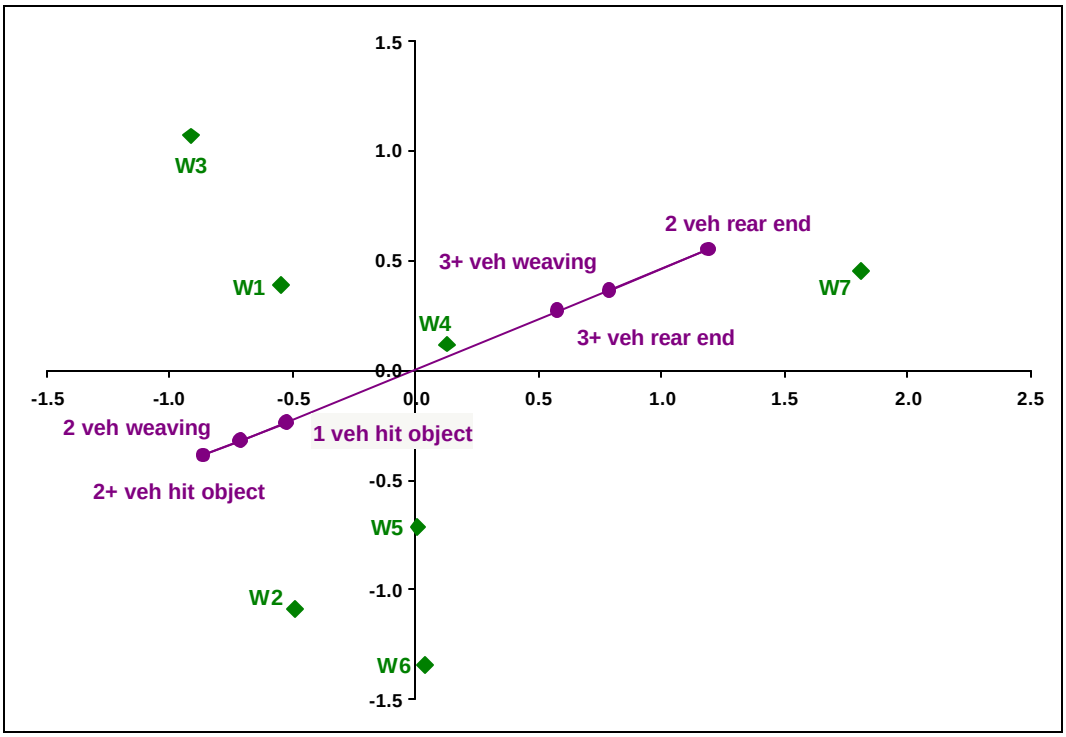

Figure 6 Category Centroids for the Traffic Flow Regime and Crash Type Variables – Wet Road Crashes

Crash type is explained by both canonical variates, but more so by the first. The optimal scaling of crash type contrasts hit-object crashes and two-vehicle weaving crashes against two-vehicle rear end crashes. Three-or-more-vehicle collisions are in-between but are more like two-vehicle rear end crashes. Hit-object crashes are more likely in low-flow Regimes W1, W2 and W3. Rear end crashes are more likely in flow conditions approaching capacity, Regime W7. Once again, Interpretations relative to

crash location and severity are omitted from detailed discussion. Table 9 presents a summary of the combined results

Table 9 Distinguishing Crash Typology for Wet Road Traffic Flow Regimes

Regime	Relatively more common types	Relatively less common types
1. Very light variable flow	46% run-offs (versus 32% overall) 35% off-road right (v. 14%)	0% left lane (v. 16%) 19% rear-ends (v. 33%)
2. Light free flow	46% run-offs (v. 32%)	14% rear-ends (v. 33%)
3. Moderate free flow	52% 1 vehicle run-offs (v. 27%)	19% rear-ends (v. 33%)
4. Moderate, right-concentrated flow	58% injury (v. 43%)	12% 1 vehicle run-offs (v. 27%) 19% off-road (v. 36%)
5. Heavy, variable flow	Near average distribution of collision types, locations and severity	
6. Flow approaching capacity	80% property damage only (v. 57%)	
7. Congested flow	76% rear-ends (v. 33%) 44% left lane (v. 16%)	8% run-offs (v. 32%)

7 Interpretation of Results

The eight dry-day Regime centroids can be plotted in the space of two of the six defining variables: mean speed and mean flow, using the ratio flow to occupancy as a surrogate for speed. The mean speed and flow dimensions are standardized (origin set at system mean, and scale in standard deviation units) for easy comparison among all dimensions and environmental conditions. By embedding four-dimensional plots at the location of each Regime in standardized speed-flow space, we expose the roles of the other four defining dimensions: the mean temporal variations in speed and flow, by through lanes versus right lane. The resulting compound plot is shown in Figure 7. In the embedded figures, the mean Regime variations in flow are plotted on the horizontal axes, with the mean variation in flow in the right lane plotted on the left axis, and variation in flow in the through lanes plotted on the right axis. The corresponding mean variations in speed in the right and through lanes plotted on the north and south axes, respectively. For reference, the sample means are plotted as the lightly hashed figure in each of the plots, while the solid figure represents each particular Regime's values. So, for example, a solid figure that is completely enclosed by the lightly hashed figure

would correspond to a regime for which all of the variations in speed and flow are less than the corresponding variations for the entire sample.

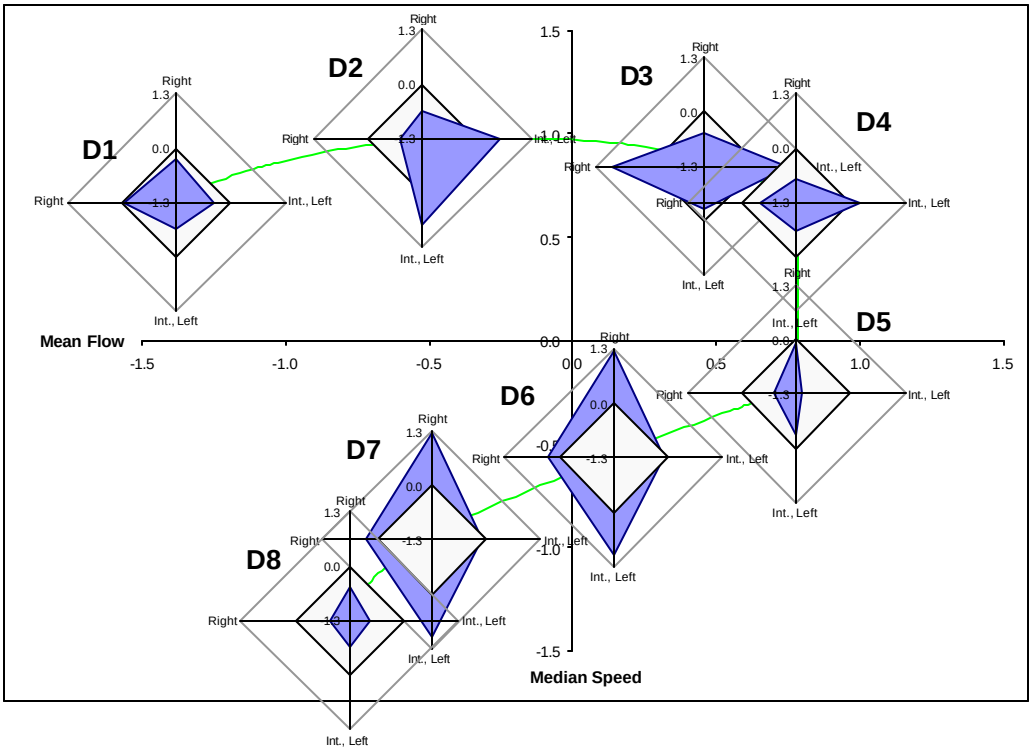

Figure 7 Variations in Flow and Speed for the Eight Dry-day Traffic Flow Regimes in Standardized Speed-Flow Space

It is tempting to speculate that the Regimes for dry-day conditions, and thereby the associated crash rates and characteristics, trace the entire range of a "standardized" speed-flow curve that is similar to that found in many empirical studies (Pushkar, Hall and Acha-Daza, 1994). The curve has three distinct branches (Figure 8): (1) a top nearly horizontal convex segment, generally known as "free flow," (2) a vertical segment near maximum observable flow, known as "queue discharge," and (3) a bottom segment known as "congested flow" or "within the queue" (Hall, Hurdle and Banks, 1992).

For the most part, Regimes in the upper "free flow" branch of the flow-speed curve, as defined by dry-day crashes, are characterized by less-than-average variations in speed relative to conditions present for all such crashes. An exception to this is Regime D2, which has a significantly higher variation in speed in the left lane; it is perhaps notable that this Regime also demonstrated the highest percentage of injury crashes, off-road left, and run-offs. The "queue discharge" portion of the curve generally exhibits less temporal variation in both speed and flow, while Regimes in the "congested flow" region

are marked by high relative variation in speed until the extreme point is reached in which presumably traffic is more or less at a standstill.

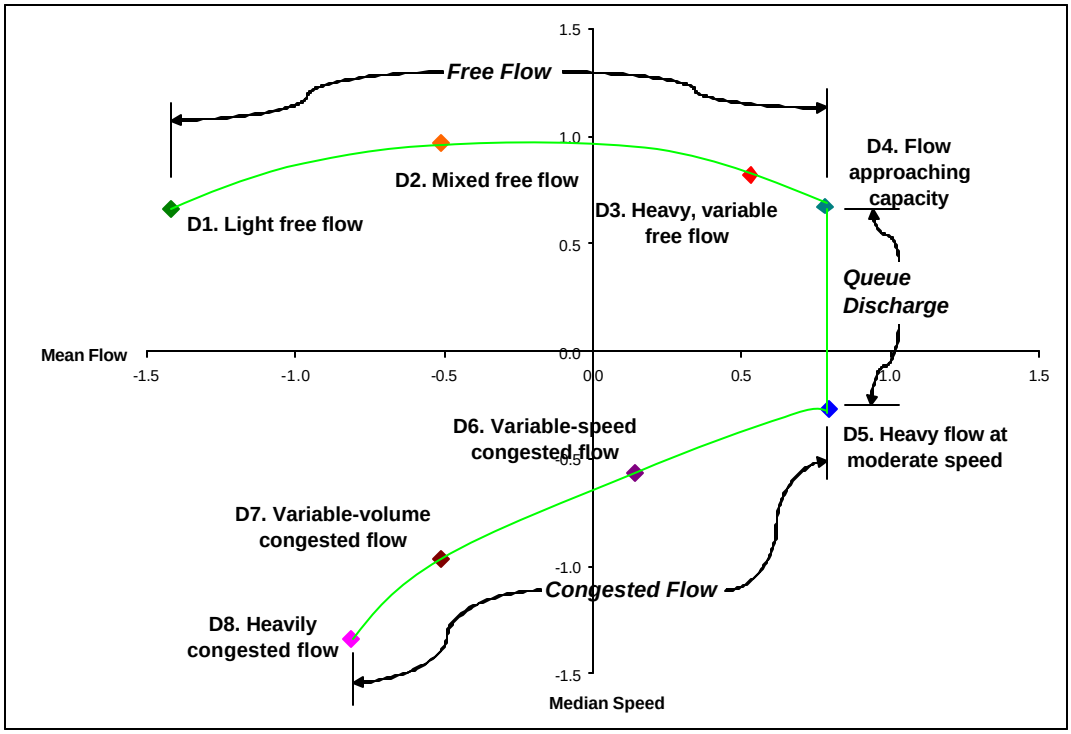

Figure 8 Centroids of the Eight Dry-day Traffic Flow Regimes on Implied Speed-Flow Curve in Standardized Speed-Flow Space

These plots of the Regime centroids in standardized speed-flow space are derived **only** from traffic conditions present at, or near, the time and location of crashes, rather than being a sampling of the population of all traffic conditions at a particular freeway location. In this sense, the plot in Figure 8 represents a sampling of flow – speed conditions conducive to freeway incidents. Distinct clusters of crashes relative to collision type, frequency, location and severity roughly arrange themselves along a standard traffic flow-density curve according to some measure analogous to level of service. This suggests that there may be a direct correspondence between level of service (a traffic performance measure) and crash typology (a traffic safety measure).

Similarity with conventional traffic speed-flow curves is also apparent in the case of the dry-night Regimes (Figure 9), although there are not as many distinctly different accident types defined by the congested region of flow conditions. The three dry-night Regimes on the free-flow branch of the implied speed-flow curve demonstrate a transition from high speed variances in all lanes (Regime N1) to high flow variances in all lanes (Regime N3). The corresponding trend in crash types is from run-offs to

weaving crashes (Table 8). On the congested-flow branch of the curve, both speed and flow variances first increase, then decrease, with decreasing mean speed and flow. Rear-end crashes are the dominant type for Regimes N4 through N6, locations shift from the right lane to interior lanes, and injury crashes become less prevalent as mean speed decreases.

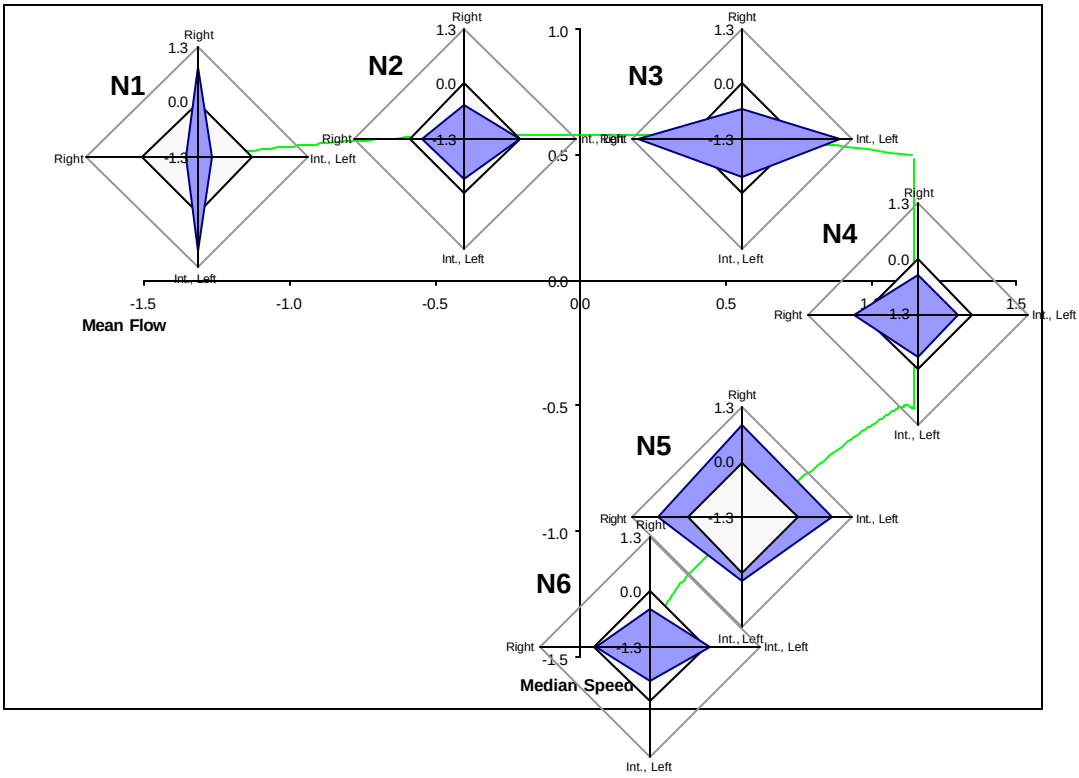

Figure 9 Variations in Flow and Speed for the Six Dry-Night Traffic Flow Regimes in Standardized Speed-Flow Space

In the case of Wet environmental conditions, distinctions in the flow-speed plane among crash characteristics are mainly confined to near free-flow conditions with traffic flow (rather than speed) being the principal discriminator. The lone congested Regime, which apparently is located in the "queue discharge" region of the flow-density curve, is dominated by rear-end collisions. There appear to be no stable patterns of crash typology within the "congested flow" segment of the flow-density curve, suggesting that under wet heavily congested conditions there is a paucity of the types of aggressive or inattentive behaviors that can lead to collisions.

The first three wet-road Regimes are dominated by vehicle run-off crashes (Table 9). This changes dramatically for Regime W4, which has about the same flow-speed conditions as Regime W3, but differs from Regime W3 in the "variations" dimension.

Presumably owing to the high temporal variation in flow in the right lane (Figure 15), perhaps indicative of merging traffic from/to on/off ramps, the character of crashes changes to a high proportion of injury accidents, a dramatic decrease in run-offs (compared to Regime W3). Conversely, Regime W6, which has the smallest temporal variation in speed (for all lanes), is marked by less severe collisions (i.e., property damage only vs. injury).

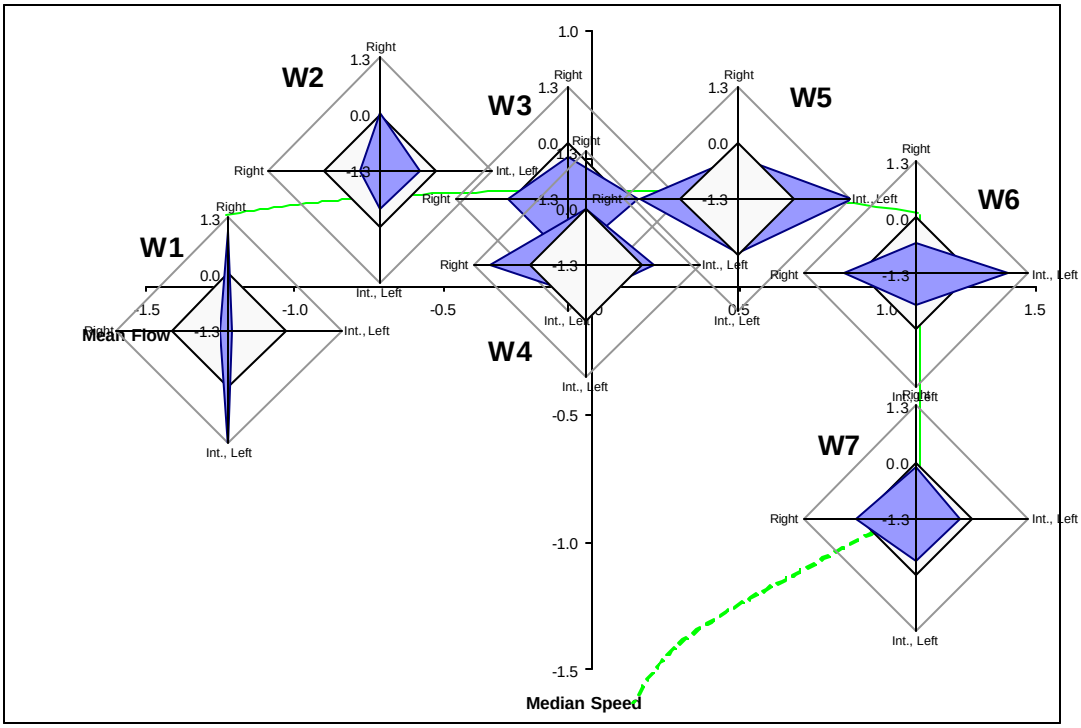

Figure 10 Variations in Flow and Speed for the Seven Wet Traffic Flow Regimes in Standardized Speed-Flow Space

8 Test Application

In an example application documented in Golob, Recker and Alvarez (2003), we estimated the distribution of 1998 AM peak period crashes across the eight dry-day Regimes. Here, we combine those estimates with total volumes associated with each observed Regime occurrence were calculated from total 30-second volumes across all freeway lanes. The results for the 1998 AM peak hour traffic reveal that the eight Regimes for daylight and dry road conditions were characterized by different patterns of crash types.

An estimate of the number of crashes involving lane-changing maneuvers per million exposed vehicles per Regime is plotted in standardized speed-flow space in Figure 11. Weaving crashes are most prevalent under congested flow conditions, and in conditions of light flow and high speeds. Conditions less conducive to weaving crashes are characterized by relatively high flow rates, but with relatively small temporal variation in speed (Regimes D3 and D4).

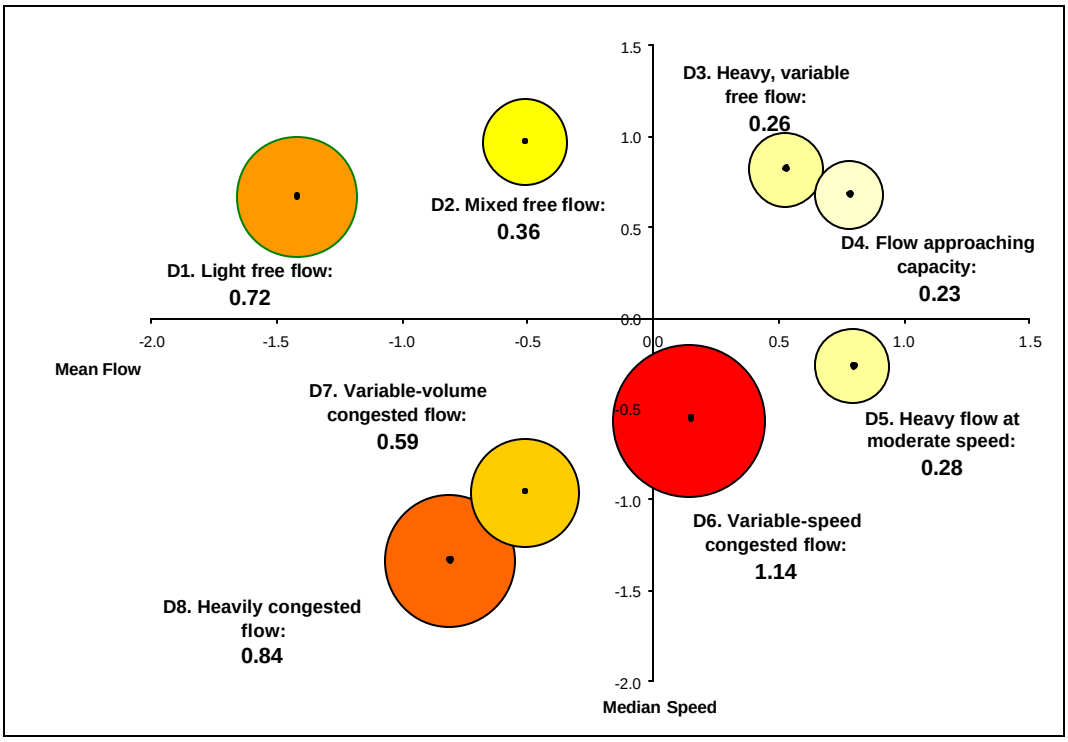

Figure 11 Estimated Lane-change Crashes per Million Vehicle Miles of Travel for the Eight Traffic Flow Regimes During AM Peak Hours, Plotted in Standardized Speed-Flow Space

The pattern for rear-end crashes (Figure 12) is generally characterized by a concentration of such crashes at high levels of congestion. Most notably (and as expected), the rates appear to accelerate dramatically under extreme "stop-and-go" conditions. Relatively lower rates of crashes for high density traffic flow conditions can indicate both the synchronized nature of these conditions and the lower likelihood of crashes resulting in police reports, due to severity of the crash and inability of the involved drivers to pull off the road.

These preliminary results indicate that crash rates for rear end collisions for the same levels of flow are substantially higher in congested versus free flow conditions: 1.04 weaving crashes per million vehicles for Regime D2 "Mixed free flow" versus 2.42 for

Regime D7 "Variable volume congested flow;" and 0.21 for Regime D4 "Flow approaching capacity" versus 0.91 for Regime D5 "Heavy flow at moderate speeds." It is not possible to put confidence bounds on these estimates at this time, but the pattern is clear. Differences in the likelihood of rear end collisions appear to be the cause of the trend observed in a number of aggregate studies (notably those of Ceder, 1982; Sullivan, 1990; and Persaud and Dzbik, 1992) that crash rates for free congested conditions are higher than crash rates for free flow conditions. The conclusion from aggregate studies, that crash rates decline as flow approaches capacity (e.g., Sullivan, 1992; Garber and Subramanyan, 2001) appears to be caused by differences in the likelihood of lane-change crashes (Figure 11).

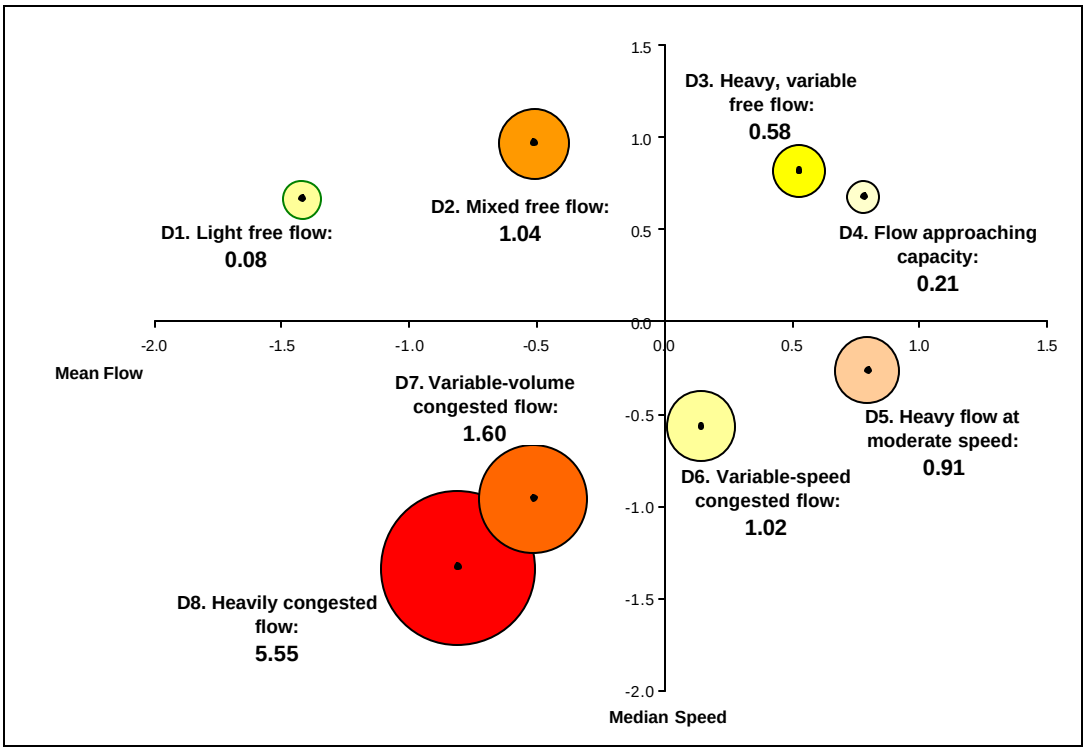

Figure 12 Estimated Rear-end Crashes per Million Vehicle Miles of Travel for the Eight Traffic Flow Regimes During AM Peak Hours, Plotted in Standardized Speed-Flow Space

It is emphasized that these estimates are for demonstration purposes only; additional research is needed before we can confidently assign safety levels to different traffic flow conditions. However, if these results hold up under a full-scale implementation, the approach could form a basis with which to directly quantify the safety benefits of improved traffic flow. By identifying the types of crashes that are most likely to occur under different traffic conditions, then identifying where and when on the freeway system these conditions occur, the most dangerous conditions could be highlighted for

mitigation and forecasts of crash reductions associated with alternative mitigation schemes evaluated.

9 Conclusions

The results in this paper evidence that there are well-defined associations between freeway accident characteristics and prevailing traffic flow conditions. Controlling for environmental effects, we demonstrate that the descriptive characteristics of crashes are distinguished by distinct traffic flow regimes, defined by specific combinations of central tendencies and temporal variations.

The analysis techniques employed here are somewhat unconventional to this domain of study. Rather than build upon a foundation of traffic engineering principles and constructs, we have instead viewed the problem as being one essentially of data analysis, and have relied on classical (and emerging) statistical techniques to help reveal the structure of the underlying phenomena. We believe that such an approach has the potential to cast the problem in terms of more appropriate explanatory variables that in turn can form the basis of richer engineering analysis.

These results offer encouragement that further investigations might uncover important, and well defined, linkages between traffic crash characteristics and accepted fundamental relationships in traffic engineering.

Such results can lay the groundwork for development of tools that can be used to assess the changes in traffic safety that result from changes in traffic flow. The only input that such tools require is a stream of 30-second observations from ubiquitous single inductive loop detectors. The tool can then be used as part of any evaluation that compares before and after traffic flow data, as measured by such detectors. Applications might involve assessing the benefits of ATMS operations or other projects that influence traffic operations.

This analysis applies only to urban freeways with at least three lanes in each direction, and the specific results apply to conditions during 1998 in Orange County, California. We presume that the relationships uncovered are indicative of all California urban freeways, particularly those in the San Francisco Bay, San Diego, and Sacramento Metropolitan Areas, but validation has not yet been conducted, so we cannot confirm the degree of spatial transferability.

Other limitations apply. First, due to the quality of the historical loop detector data that were used in calibrating the tool, we were unable to accurately estimate crash rates for different traffic flow Regimes. The historical traffic flow data were not sufficiently representative of Orange County for an entire year, because there were systematic patterns in missing data as a function of freeway route, location along each route, day of week, and week of the year. Thus, 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, we focused instead on which types of crashes are more likely under different types of traffic flow, while controlling for exposure by incorporating loop volume data to produce a rate in the form of crashes/vehicle. A more conventional and accurate calculation of crash rates as a function of vehicle miles traveled is an important subject for future research.

Acknowledgements

This research was funded by the California Partners for Advanced Transit and Highways (PATH) and the California Department of Transportation (Caltrans). The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the University of California, California PATH, or the California Department of Transportation.

References

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

Chang, M-S., 1982. Conceptual development of exposure measures for evaluating highway safety. *Transportation Research Record 847*: 37-42.

Chen, C; K. F. Petty, A. Skabardonis, P.P. Varaiya, and Jia, Z. (2001). Freeway performance measurement system: mining loop detector data. *Transportation Research Record 1748*: 96-102.

Choe, T., Skabardonis, A., and Varaiya, P.P., 2002. Freeway performance measurement system (PeMS): an operational analysis tool. Presented at Annual Meeting of Transportation Research Board, January 13-17, Washington, DC.

Davis, G.A., 2002. Is the claim that 'variance kills' an ecological fallacy? *Accident Analysis and Prevention,* 34: 343-346.

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.

Everitt, B. S., 2001. *Cluster Analysis.* Oxford University Press, New York.

Garber, N.J. and Subramanyan, S., 2001. Incorporating crash risk in selecting congestion-mitigation strategies. Transportation Research Record 1746: 1-5.

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

Golob, T.F. and Recker, W.W., 2003. Relationships among urban freeway accidents, traffic flow, weather and lighting conditions. *Journal of Transportation Engineering, ASCE,* 129: 342-353.

Golob, T.F., Recker, W.W. and Alvarez, V.M., 2002. *Freeway Safety as a Function of Traffic Flow: The FITS Tool for Evaluating ATMS Operations.* Final Report prepared for California Partners for Advanced transit and Highways (PATH). Institute of Transportation Studies, University of California, Irvine, CA.

Golob, T.F., Recker, W.W. and Alvarez, V.M., 2003. A Tool to Evaluate the Safety Effects of Changes in Freeway Traffic Flow. *Journal of Transportation Engineering – ACSE*, in press.

Hall, F.L., Hurdle, V.F. and Banks, J.H., 1992. Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transportation Research Record*, 1365: 12-18.

Hartigan, J.A., 1975. *Clustering Algorithms.* John Wiley, New York.

Hensher, D.A. and Golob, T.F., 1999. Searching for policy priorities in the formulation of a freight transport strategy: a canonical correlation analysis of freight industry attitudes towards policy initiatives. *Transportation Research Part E, Logistics and Transport Review*, 35: 241-267.

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

Hotelling, H., 1933. Analysis of a complex series of statistical variables into principal components. Journal of Educational Psychology, 24: 417-441, 498-520.

Lee, C., Saccomanno, F., and Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. *Transportation Research Record,* 1784: 1-8.

Lee, C., Hellinga, B. and Saccomanno, F., 2003. Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. Presented at the Annual Meeting of the Transportation Research Board, January 12-16, Washington, D.C.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and probability*, 281-297. University of California Press, Berkeley, CA.

Mensah, A. and Hauer, E., 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 de Leeuw, J., 1998. The GIFI system of descriptive multivariate analysis. *Statistical Science*, 13: 307-336.

Oh, C., Oh, J-S. and Chang, M., 2001. An advanced freeway warning information system based on accident likelihood. Presented at the 9th World Conference on Transport Research, July 22-27, 2001, Seoul, Korea.

Oh, C., Oh, J-S., Ritchie, S.G. and Chang, M., 2001. Real-time estimation of freeway accident likelihood. Presented at the Annual Meeting of the Transportation Research Board, January 8-12, Washington D.C.

Prigogine, I. And Herman, R., 1971. *Kinetic Theory of Vehicular Traffic*. American Elsevier, New York.

Persaud, B. and Dzbik, L., 1992. Accident prediction models for freeways. . *Transportation Research Record* 1401: 55-60.

Pushkar, A., F.L. Hall and J.A. Acha-Daza, 1994. estimation of speeds from single-loop freeway flow and occupancy data using cusp catastrophe theory model. *Transportation Research Record* 1457: 149-157.

Robinson, W., 1950. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15: 351-327.

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

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 Heiser, W.J., 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., 1988. *Nonlinear canonical Correlation and Some Related Techniques*. DSWO Press. Leiden.

van der Boon, P., 1996. *A Robust Approach to Nonlinear Multivariate Analysis*. DSWO Press, Leiden.

Varaiya, P. P., 2001. Freeway Performance Measurement System, PeMS V3, Phase 1: Final Report. Report UCB-ITS-PWP-2001-17, California PATH Program, Institute of Transportation Studies, University of California, Berkeley, CA.