

UC Berkeley

Working Papers

Title

Identifying Freeway Traffic Instability from PeMS Data Using Rolling-Window Speed and Occupancy Variability

Permalink

<https://escholarship.org/uc/item/2wn1n8b8>

Authors

Pan, Dangran

Xue, Zhiyuan

Shahriar, Muhammad Shajid

Publication Date

2026-10-06

Data Availability

The data associated with this publication are available at: <https://pems.dot.ca.gov>

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Identifying Freeway Traffic Instability from PeMS Data Using Rolling-Window Speed and Occupancy Variability

Dangran Pan
Zhiyuan Xue
Muhammad Shajid Shahriar

University of California, Berkeley

Corresponding author:
Muhammad Shajid Shahriar
msshahriar@berkeley.edu

October 2026

ABSTRACT

This study examines short-term freeway traffic instability using station-level 5-minute data from the California Performance Measurement System for Caltrans District 4. The dataset covers April 7, 2026. After restricting the sample to retained mainline observations, rolling 15-minute measures of mean speed, speed variability, occupancy variability and net speed change were constructed for 652,080 station-time observations. An instability-congestion proxy state was defined as the joint occurrence of speed variability at or above its 75th percentile and mean speed at or below its 25th percentile. An exploratory logistic classification model was then used to describe how the four traffic measures separate observations assigned to this proxy state. Occupancy variability and positive net speed change were associated with higher classification odds. However, mean speed and speed variability partly define the outcome and their coefficients are therefore definitional rather than independent evidence of safety effects. In addition, the model does not account for dependence among repeated observations from the same detector station. The results demonstrate a transparent approach for screening unstable congestion from high-resolution detector data but do not establish crash probability or causal safety effects. Validation against observed crashes, longer observation periods and models accounting for station-level and temporal dependence are required before operational deployment.

Keywords: Freeway Traffic Instability; PeMS; Speed Variability; Occupancy Variability; Traffic Congestion; Freeway Operations; Rolling-Window Analysis; Logistic Regression; Caltrans District 4; San Francisco Bay Area.

1. INTRODUCTION

Freeway traffic is not a passive stream of vehicles but a dynamical system whose state changes continuously across space and time. The theoretical foundation for understanding these dynamics was established by Lighthill and Whitham (1955) and Richards (1956) through the kinematic wave formulation, commonly known as the LWR model. This framework demonstrates that disturbances in traffic flow and density propagate as waves and when these wave characteristics intersect, shock waves emerge, producing abrupt transitions in speed and density that are experienced as sudden congestion (Lighthill & Whitham, 1955; Richards, 1956). Congestion onset is therefore not an isolated local event but a structured transition that typically originates at bottlenecks and propagates upstream. Traffic instability should accordingly be understood as an inherent feature of freeway dynamics rather than an exceptional condition.

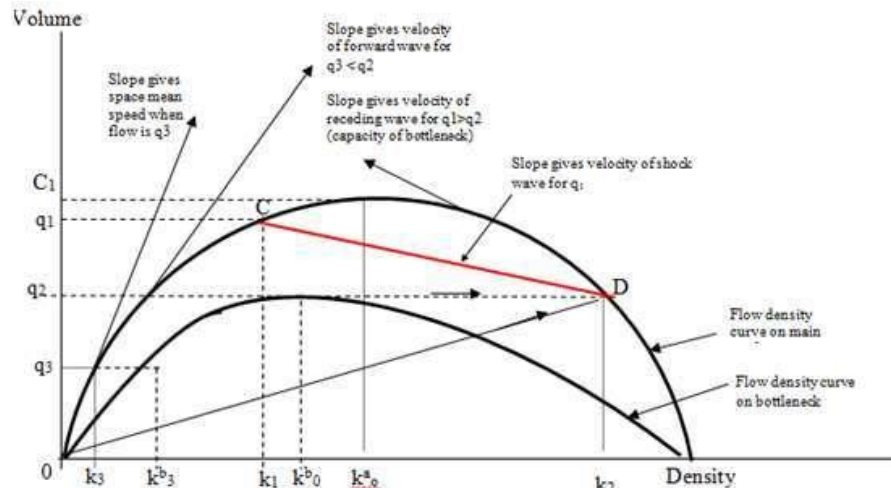


Figure 01: Kinematic and shock-wave measurements in relation to the flow-density curve.

Source: adapted from Saxena and Jain (2017).

Empirical studies reinforce this theoretical interpretation. Kerner and Rehborn (1996) demonstrated that congested traffic can exist in multiple distinct states, while Treiber, Hennecke and Helbing (2000) showed that similar freeway conditions can produce different congestion regimes depending on the relationship between demand and capacity. At bottlenecks, this instability frequently manifests as queue formation accompanied by a capacity drop, where discharge flow falls below pre-breakdown levels once congestion begins (Cassidy & Bertini, 1999). Chung, Rudjanakanoknad and Cassidy (2007) further observed that traffic density increases prior to this capacity drop, indicating that breakdown is preceded by measurable changes rather than occurring abruptly. These findings suggest that detector data may contain observable early signals of instability, such as rising occupancy and declining speed.

Once congestion develops, it rarely remains uniform. Instead, it often exhibits oscillatory patterns characterized by repeated backward-propagating cycles of acceleration and deceleration. Mauch and Cassidy (2002) documented such oscillations on California freeways, while Schönhof and Helbing (2007) identified similar unstable patterns in extended motorway observations. In multi-lane freeway environments, Ahn and Cassidy (2007) and Zheng et al. (2011) showed that lane-changing maneuvers near ramp interchanges can trigger deceleration waves that amplify as they propagate upstream. Together, these findings indicate that fluctuations in speed and occupancy are not random noise but meaningful signatures of underlying traffic instability.

However, measuring instability from loop detector data requires careful interpretation. The Federal Highway Administration notes that detectors directly measure occupancy rather than density and speed estimates derived from single-loop detectors become less reliable during traffic state transitions (Coifman, 2000; Lu et al., 2012). These limitations are particularly important because transition periods are precisely when instability is most relevant. In California, the PeMS system provides a critical data infrastructure for

studying such processes through high-resolution detector observations collected across a large freeway network (Chen et al., 2001). Despite challenges such as missing data, sensor failure and uncertainty in derived variables (Bickel et al., 2007), prior research indicates that these datasets remain suitable for identifying meaningful instability patterns when processed appropriately.

The significance of traffic instability extends beyond operational performance, as unstable traffic conditions have been consistently linked to elevated crash risk. Lee, Hellinga and Saccomanno (2003) introduced the concept of crash precursors using loop detector data and showed that rapid changes in traffic conditions are associated with increased crash likelihood. Abdel-Aty et al. (2004) found that speed variance, occupancy and traffic volume in the minutes preceding a crash are significant predictors of crash occurrence, while Abdel-Aty and Pande (2005) demonstrated that speed differentials across adjacent lanes strongly influence crash propensity. Hourdos et al. (2006) showed that such conditions can be detected in real time using upstream and downstream detector data. Most directly, Zheng, Ahn and Monsere (2010) demonstrated that crash-prone periods are significantly more likely than control periods to exhibit oscillatory traffic patterns. Collectively, this body of literature establishes that instability is not only observable but also critically relevant to freeway safety.

Crash frequency and severity are direct measures of road safety, but crash-based analysis is inherently reactive and may be constrained by sparse observations, reporting limitations and limited information about the mechanisms preceding a crash. Surrogate-safety analysis attempts to address these limitations by examining observable non-crash events or operating conditions that may be related to safety. However, a surrogate cannot be treated as an equivalent substitute for crashes merely because it appears hazardous. Its empirical relationship with observed crash outcomes must be established and validated (Tarko et al., 2009).

Existing freeway research shows that short-term speed and occupancy fluctuations can occur before crashes and that high-resolution detector data can characterize unstable operating conditions. Nevertheless, the present dataset does not include matched crash events. The unresolved question addressed here is therefore narrower: whether a transparent set of rolling-window variables can be used to identify and describe freeway observations exhibiting the joint conditions of low mean speed and high speed variability. This is an operational classification question rather than a crash-prediction or causal-safety question.

Accordingly, this study uses one day of station-level 5-minute PeMS data from Caltrans District 4 to construct rolling 15-minute measures of mean speed, speed variability, occupancy variability and net speed change. It then defines an instability-congestion proxy state using high speed variability and low mean speed and uses an exploratory logistic model to describe how the measured traffic variables separate observations assigned to that state. The study contributes a transparent demonstration of short-term traffic-state screening. It does not estimate observed crash probability, implement a hierarchical Bayesian model or validate the proxy against SWITRS crash records.

2. Methodology

2.1 Research Objective and Analytical Scope

The objective of this study is to construct and describe a short-term instability-congestion proxy state using station-level freeway detector data. The analytical logic is:

Observed PeMS measurements → rolling 15-minute traffic indicators → instability-congestion proxy classification

Traffic instability is represented through short-term speed variability, occupancy variability and net speed change, while mean speed represents the prevailing operating condition. An observation is classified into the proxy state when speed variability is at or above its sample 75th percentile and mean speed is at or below its sample 25th percentile.

This proxy is motivated by crash-precursor literature, but it is not an observed crash outcome and has not been validated as a crash surrogate in the present study. The analysis therefore does not estimate crash probability or test whether instability causes crashes. Instead, it provides an exploratory description of how the selected traffic indicators behave within a researcher-defined classification of unstable congestion.

The logistic model should also be interpreted with two important qualifications. First, speed variability and mean speed are used to define the dependent variable and are then included as model inputs. Their estimated coefficients therefore partly reproduce the classification rule and cannot be treated as independent empirical discoveries. Second, repeated observations from the same detector stations are not statistically independent, while the implemented model uses conventional model-based standard errors. Coefficients and nominal p values are consequently reported for descriptive purposes rather than confirmatory inference.

2.2 Data Structure and Sample Construction

The study uses station-level 5-minute data obtained from the California Department of Transportation Performance Measurement System for District 4, which covers the San Francisco Bay Area. The analyzed date is April 7, 2026. Available variables include total flow, average occupancy and average speed.

The raw dataset contains 1,120,032 station-time observations across several lane types. After restricting the sample to retained mainline observations and removing records with missing values in average speed, average occupancy or total flow, 656,640 observations remain. Rolling 15-minute variables require three retained observations within each station. The first two retained observations for each station therefore have incomplete rolling statistics and are omitted, resulting in 652,080 complete station-time observations.

The data have a time-series cross-sectional structure because each detector station is observed repeatedly throughout the day. The current analysis treats these observations as an exploratory classification sample. It does not adjust model uncertainty for station-level or temporal dependence.

2.3 Data Cleaning and Filtering

The raw PEMS data include multiple lane types, including mainline, freeway-to-freeway, on-ramp, off-ramp and high-occupancy vehicle lanes. Because the study focuses on freeway mainline traffic conditions, the data are filtered to mainline lanes only. Observations with missing values in average speed, average occupancy, or total flow are removed. Timestamps are converted into datetime format and the data are sorted by detector station and time. The implemented cleaning procedure removes missing measurements before calculating rolling statistics. It does not separately verify that every pair of retained observations is exactly five minutes apart. Consequently, a rolling window may span more than 15 clock minutes if an intervening record was removed. This issue is retained as a reproducibility limitation and should be addressed in a future implementation through explicit timestamp-continuity checks.

This procedure creates an ordered analytical dataset for rolling-window construction. However, it should not be interpreted as a complete detector-quality screening procedure because no additional filter for implausible or malfunction-related measurements was implemented.

2.4 Time Window Definition

Prior studies indicate that short time windows can capture operational changes associated with freeway instability and crash-prone conditions (Lee et al., 2002; Pande & Abdel-Aty, 2006). The present study does not observe crash times and therefore does not define a pre-crash interval. Instead, it uses a rolling window

based on the current retained 5-minute observation and the two preceding retained observations for the same detector station.

Under uninterrupted data, these three observations represent 15 minutes. Because the implemented code does not explicitly verify timestamp continuity after missing records are removed, the window should be described as a three-observation rolling window that ordinarily corresponds to 15 minutes.

2.5 Variable Construction

To characterize short-term traffic dynamics, rolling statistics are constructed from the current retained observation and the two preceding retained observations within each detector station.

2.5.1 Short-Term Mean Speed

$$\bar{v}_i = \frac{1}{3} \sum_{t=i-2}^i v_t \quad (1)$$

where

- $\bar{v}_i$ is the mean observed speed across the three-observation window for station-time observation i and represents the prevailing operating condition.
- v_t is the observed vehicle speed at time interval t .
- i is the current time index.
- $t = i - 2, i - 1$ and i denotes the three retained station observations used in the rolling calculation.

This variable captures the overall operating condition of traffic. Lower average speed may indicate congestion, shockwave formation, or traffic breakdown. Although average speed alone does not fully represent instability, it is an important baseline traffic condition and helps distinguish stable free-flow from disturbed traffic states.

2.5.2 Short-Term Speed Variability

$$\sigma_{v,i} = \sqrt{\frac{1}{3} \sum_{t=i-2}^i (v_t - \bar{v}_i)^2} \quad (2)$$

where $\sigma_{v,i}$ is the standard deviation of speed across the three selected intervals for observation i .

A large value of $\sigma_{v,i}$ indicates substantial variation in observed speeds within the rolling window. It may reflect acceleration-deceleration cycles or changing traffic states, although aggregated station-level data cannot identify individual vehicle maneuvers. In this study, speed variability is also used to define the dependent classification and its regression coefficient must therefore be interpreted as partly definitional.

2.5.2 Occupancy Variability

$$\sigma_{o,i} = \sqrt{\frac{1}{3} \sum_{t=i-2}^i (o_t - \bar{o}_i)^2} \quad (3)$$

where

- $\sigma_{o,i}$ is the variation in occupancy, representing fluctuations in traffic density.
- σ_i is the occupancy (percentage of time a detector is occupied by vehicles).

- σ_i is the mean occupancy over the 15-minute window.

Occupancy is commonly used as a proxy for traffic density (May, 1990). Large occupancy fluctuations are often interpreted as evidence of unstable traffic transitions, such as queue growth, breakdown onset, or oscillatory congestion (Abdel-Aty & Pande, 2005; Pande & Abdel-Aty, 2006).

2.5.4 Short-Term Speed Change

$$\Delta v_i = v_i - v_{i-2} \quad (4)$$

where

- $v_{i,\text{start}}$ is the speed in the first retained observation of the rolling window,
- $v_{i,\text{end}}$ is the speed at the end of the selected pre-crash profile,
- Δv_i therefore measures the net speed change over the interval.

This signed variable measures net speed change across the rolling window. A negative value indicates a net speed reduction, while a positive value indicates net speed recovery or acceleration. Because it measures only the difference between the first and last observations, it does not capture every fluctuation occurring inside the window.

2.5.3 Binary High-Risk Traffic Indicator

The dependent variable is a binary instability-congestion proxy indicator. An observation is assigned a value of 1 when its rolling speed standard deviation is at or above the sample 75th percentile and its rolling mean speed is at or below the sample 25th percentile. All other complete observations are assigned a value of 0.

$$y_i = \begin{cases} 1, & \text{if } \sigma_{v,i} \geq Q_{0.75} \text{ and } \bar{v}_i \leq Q_{0.25} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where

- $y_i = 1$ denotes membership in the instability-congestion proxy state,
- $\sigma_{v,i}$ is rolling speed variability,
- $Q_{0.75}$ is the sample 75th percentile of rolling speed variability,
- $\bar{v}_i$ is rolling mean speed,
- $Q_{0.25}$ is the sample 25th percentile of rolling mean speed.

This definition is informed by literature linking oscillatory traffic and low-speed turbulent conditions with elevated crash propensity (Golob & Recker, 2003; Abdel-Aty & Pande, 2005; Zheng et al., 2010). However, the threshold combination is constructed by the authors and is not validated against observed crashes in this study. It should therefore be interpreted as an operational traffic-state classification rather than a measured high-risk or crash-prone state.

2.6 Exploratory Logistic Classification Model

Because the proxy indicator is binary, the following logistic model is estimated:

$$\text{logit}(p_i) = \beta_0 + \beta_1 \sigma_{v,i} + \beta_2 \sigma_{o,i} + \beta_3 \bar{v}_i + \beta_4 \Delta v_i$$

where:

- $p_i = P(y_i = 1 | X_i)$, the probability that observation i is assigned to the instability-congestion proxy state;
- β_0 is the intercept;
- $\sigma_{v,i}$ is rolling speed variability;
- $\sigma_{o,i}$ is rolling occupancy variability;
- $\bar{v}_i$ is rolling mean speed; and
- Δv_i is rolling net speed change.

The model is used as a descriptive classification diagnostic. Because $\sigma_{v,i}$ and $\bar{v}_i$ directly define y_i , their coefficient signs and statistical significance are mechanically influenced by the construction of the outcome. The coefficients for occupancy variability and net speed change are more informative about whether these additional variables separate the constructed states, although they still do not demonstrate crash effects.

The implemented model uses conventional model-based standard errors and does not adjust for repeated observations within detector stations. Reported p values and intervals are therefore nominal and should not be treated as fully calibrated inferential evidence.

2.7 Interpretation Strategy

The analysis addresses four descriptive questions:

1. How prevalent is the constructed instability-congestion proxy state in the analytical sample?
2. Do occupancy variability and net speed change provide additional separation between observations inside and outside the constructed state?
3. How strongly are rolling speed variability and occupancy variability correlated?
4. Which limitations prevent the resulting coefficients from being interpreted as estimates of observed crash risk?

Odds ratios are reported to describe classification differences. They are not interpreted as causal safety effects. Statistical significance is treated cautiously because the sample contains repeated station observations and two model inputs directly define the outcome.

3. RESULTS

3.1 Descriptive Statistics

Table 1 reports the summary statistics of the key raw traffic variables. These raw measures provide the empirical basis for the rolling-window indicators defined in Equations (1)–(4).

The descriptive statistics are important because they show whether the dataset contains enough variation to support a short-term instability analysis. If speed, flow and occupancy were nearly constant, the rolling-window indicators would have limited analytical value. In this dataset, however, all three variables show substantial variation across stations and time periods.

Table 1: Summary Statistics of Key Variables

Variable	Mean	Std. Dev.	Min	Median	Max
Total Flow	168.63	168.46	0.00	102.00	1042.00

Avg Occupancy	0.0480	0.0557	0.0000	0.0349	0.9989
Avg Speed	62.97	8.31	3.00	64.80	82.50

Several important features of the data can be observed from Table 1. First, the mean flow is 168.63 vehicles per 5-minute interval, while the standard deviation is 168.46. The fact that the dispersion of flow is nearly as large as its mean indicates substantial heterogeneity across time and stations. This suggests that the dataset includes both lightly loaded freeway conditions and highly congested periods, which is desirable for studying the transition from stable to unstable states.

Second, average occupancy has a mean of 0.0480 and a maximum of 0.9989. The extreme maximum may represent near-continuous detector actuation, an unusual operating condition or a detector-quality problem. Because the cleaning procedure removes missing values but does not apply an additional plausibility filter, the maximum should not be treated as verified evidence of traffic saturation. This observation reinforces the need for detector-quality checks in future analyses.

Third, the speed distribution also spans a broad range. The average speed is 62.97 mph and the median is 64.80 mph, which indicates that a large portion of the observations reflect normal freeway operating conditions. However, the minimum speed falls to 3 mph, indicating severe congestion or stop-and-go traffic. This contrast between high central tendency and extremely low minimum values shows that the dataset captures both free-flow and breakdown conditions. Such variation directly motivates the use of the short-term mean speed in Equation (1) and the short-term variability measures in Equations (2) and (3).

Overall, the variables display substantial cross-station and temporal variation, which permits construction of the rolling indicators. These descriptive statistics do not independently validate the instability-congestion proxy or its relationship with crash occurrence.

3.2 Exploratory Logistic Classification Results

Table 2 reports the exploratory logistic classification model. The dependent variable indicates whether a station-time observation meets the joint sample-based thresholds for high speed variability and low mean speed. The model therefore describes separation of a constructed traffic state rather than the probability of an observed crash.

Speed variability and mean speed directly define the dependent variable. Their estimates are included to document the internal classification structure and must not be interpreted as independent predictors. Occupancy variability and net speed change are not part of the threshold rule and provide the more informative descriptive comparisons.

Table 2: Exploratory Logistic Classification Results

Variable	Coefficient	Std. Error	Lower 95%	Upper 95%	Odds Ratio	Nominal p value
Intercept	3.2262	0.038	3.1517	3.3007	25.18	< 0.001
Speed Std (15 min)	1.5963	0.007	1.5826	1.6100	4.93	< 0.001
Occupancy Std (15 min)	6.2273	0.651	4.9513	7.5033	506.36	< 0.001
Mean Speed (15 min)	-0.1228	0.001	-0.1248	-0.1208	0.88	< 0.001
Speed Change (15 min)	0.0530	0.002	0.0491	0.0569	1.05	< 0.001

Note: The dependent variable is the constructed instability-congestion proxy indicator. Speed variability and mean speed directly define that indicator. Standard errors and p values are conventional model-based estimates and are not adjusted for repeated observations within detector stations. The p values are therefore nominal.

All four coefficients have nominal p values below 0.001. Given the very large sample, the repeated station observations and the direct inclusion of two outcome-defining variables, these p values should not be interpreted as confirmatory evidence. The substantive interpretation instead focuses on coefficient direction, measurement scale and whether a variable is part of the classification rule.

Occupancy variability has the largest coefficient in the model, estimated at 6.2273. This variable is derived from Equation (3), which measures short-term instability in occupancy. The very large positive coefficient indicates that density-related fluctuations are especially powerful predictors of risk. In operational terms, this result suggests that traffic states characterized by rapidly changing vehicle concentration are much more likely to become unstable. Because occupancy is closely related to congestion formation and breakdown, this finding reinforces the view that density oscillations are a key mechanism behind hazardous traffic conditions.

The coefficient for rolling speed variability is 1.5963. Its positive sign is expected because observations must have high speed variability to be assigned to the proxy state. The associated odds ratio of 4.93 therefore documents the classification rule rather than providing independent evidence that speed variability increases crash risk.

The coefficient for occupancy variability is 6.2273. Occupancy variability is not used directly to define the dependent variable, so its positive coefficient indicates that observations assigned to the proxy state also tend to show greater variation in occupancy. Interpretation must account for its measurement scale. The reported odds ratio of 506.36 corresponds to a full 1.00-unit increase in occupancy standard deviation, which is far larger than an ordinary change in the data. For a 0.01-unit increase, the corresponding odds ratio is $\exp(6.2273 \times 0.01) = 1.064$, or approximately 6.4% higher classification odds. This scaled interpretation is more meaningful than the full-unit odds ratio.

The coefficient for rolling mean speed is -0.1228 . Its negative sign is also expected because observations must have low mean speed to be assigned to the proxy state. The coefficient should therefore be understood as a definitional feature of the classification. It does not independently demonstrate that a 1-mph speed reduction increases crash risk.

The coefficient for net speed change is 0.0530. Because the variable is calculated as final-window speed minus initial-window speed, its positive sign indicates that larger positive net speed changes are associated with higher odds of proxy-state membership. A 5-mph increase corresponds to an odds ratio of $\exp(0.0530 \times 5) = 1.30$. This result should not be described as evidence that speed drops increase risk. A positive coefficient is more consistent with speed recovery or acceleration occurring within some observations that remain classified as low-speed and highly variable.

Taken together, the model reproduces the threshold-based roles of speed variability and mean speed and shows that occupancy variability and positive net speed change provide additional descriptive separation. It does not estimate crash probability, establish causal effects or validate the proxy as a safety outcome.

3.2.1 Correlation Diagnostic Between Speed Variability and Occupancy Variability

Because both speed variability and occupancy variability represent short term traffic instability, an additional correlation diagnostic was conducted to examine whether the two indicators capture overlapping information. The Pearson correlation between 15-minute speed standard deviation and 15-minute occupancy standard deviation is 0.6269, with $p < 0.001$. The Spearman rank correlation is 0.5772, also with

$p < 0.001$. These results indicate a moderate to strong positive relationship between the two instability indicators.

This relationship is expected from traffic flow theory because unstable congestion often involves both speed oscillation and density fluctuation. However, the two variables should not be treated as identical. Speed variability captures repeated acceleration and deceleration, while occupancy variability captures changes in vehicle concentration. A variance inflation factor check also shows that speed variability has a VIF of 1.65 and occupancy variability has a VIF of 1.99, both well below conventional concern thresholds. Therefore, the model does not show evidence of severe multicollinearity between these two predictors.

The diagnostic indicates that speed variability and occupancy variability contain related but nonidentical information. The VIF values do not indicate severe linear multicollinearity under conventional thresholds. However, this diagnostic does not correct the outcome-construction issue or the dependence among repeated station observations.

3.3 Discussion

The results describe a sample-defined state characterized by the joint occurrence of low mean speed and high speed variability. They should not be read as a model of observed crashes. Because mean speed and speed variability directly define the outcome, the corresponding coefficient signs largely confirm that the logistic specification reproduces the classification rule.

Occupancy variability provides a more informative additional result. Its positive coefficient indicates that observations meeting the speed-based instability and congestion thresholds also tend to exhibit greater short-term fluctuation in detector occupancy. This is consistent with traffic-flow theory, under which unstable congestion may involve simultaneous changes in movement speed and vehicle concentration. However, the model does not establish that occupancy variation causes crashes or that the estimated relationship would remain unchanged after matching the detector data to observed crash events.

The positive coefficient for net speed change also requires careful interpretation. The variable is signed, with positive values representing speed recovery or acceleration. The estimate therefore does not support a claim that larger speed drops increase classification odds. A more complete future analysis should separate acceleration and deceleration, consider the largest within-window speed decrease and examine oscillation measures that retain the sequence of all observations rather than only the difference between the first and last values.

Operationally, rolling detector measures may help agencies screen locations and times experiencing unstable congestion. Such screening could inform further investigation or contribute to decision rules for variable speed management, ramp metering and targeted monitoring. Variable speed limits have been shown to moderate flow oscillations under congested conditions, while ramp metering can regulate demand at bottlenecks (Papageorgiou et al., 2008; Papageorgiou & Kotsialos, 2002). The present model does not evaluate these interventions, so its outputs should be treated as candidate operational signals rather than deployment-ready safety triggers.

Operational screening can influence safety practice only when its outputs are communicated in forms that decision makers can interpret and act upon. Shahriar et al. (2025) transformed multimodal road-user evidence into a digital narrative and evaluated changes in transportation planners' priorities before and after exposure to that narrative. Although the present study does not implement a communication intervention, a future pilot could translate detector-identified instability episodes into comparable visual narratives for agency review and structured stakeholder feedback.

Detector measurements can identify unstable operating states but cannot directly explain the behavioral or institutional mechanisms producing them. Shahriar (2026) developed a compositional workflow that moves from citizen-sourced evidence curation and concern prioritization to strategic diagnosis and feasibility screening of potential interventions. A future extension of the present study could adapt that sequence by using recurrent detector-instability locations as initial screening targets, incorporating citizen reports to identify locally perceived behaviors or enforcement problems and then evaluating feasible operational responses. Such evidence would complement rather than replace validation against observed crashes.

Operational effectiveness should also be distinguished from distributional fairness. In a Bay Area analysis of fuel-tax and VMT-fee alternatives, Shahriar et al. (2026) showed that aggregate policy effects can conceal different burdens across income groups and urban-rural contexts, particularly when behavioral responses are considered. The relevance to the present study is not the pricing mechanism itself but the broader evaluation principle: future variable-speed, ramp-control or enforcement interventions should be assessed for how they distribute delay, compliance costs and safety benefits across corridors and communities.

The analysis remains exploratory because it covers one day, uses a researcher-defined proxy, includes two outcome-defining variables in the classifier and applies conventional logistic standard errors to repeated station observations. The results nevertheless clarify what can and cannot be learned from the implemented workflow and identify the methodological steps required for a validated freeway-safety application.

4. LIMITATIONS AND FUTURE WORK

Several limitations determine how the findings should be interpreted.

First, the dependent variable is a researcher-defined instability-congestion proxy rather than observed crash occurrence. Surrogate-safety literature emphasizes that a proposed surrogate must be empirically connected to and validated against the meaningful safety outcome it is intended to represent (Tarko et al., 2009). The present analysis has not completed that validation. Future work should match PeMS observations with geocoded and time-matched SWITRS crash records and assess whether the proxy discriminates crash and noncrash periods.

Second, rolling speed variability and rolling mean speed are used both to define the dependent variable and as explanatory variables. Their coefficient signs, odds ratios and nominal p values are therefore partly determined by construction. Future models should either exclude the defining variables from the explanatory set, use an independently observed outcome or estimate performance on a held-out sample using a prospectively specified classification rule.

Third, the model treats 652,080 station-time observations as if they were independent even though observations are repeatedly collected from the same detector stations and adjacent time intervals overlap. Conventional standard errors may consequently be too small. A revised inferential model should use station-clustered standard errors, generalized estimating equations, a station random-intercept model or a hierarchical spatiotemporal framework. Overlapping rolling windows should also be explicitly addressed.

Fourth, the data cover only April 7, 2026. A large number of observations from one day does not substitute for temporal replication. The findings may depend on day-of-week conditions, incidents, weather, seasonal demand or detector availability. Future work should include multiple weekdays and weekends across seasons and should document the number and spatial distribution of included stations.

Fifth, missing observations are removed before the rolling variables are calculated, but the current implementation does not verify that successive retained timestamps are exactly five minutes apart. Some nominal 15-minute windows may therefore span longer periods. Future code should reindex each station to a complete 5-minute sequence and calculate a rolling value only when all three expected timestamps are present.

Sixth, the cleaning procedure does not conduct a full detector-quality assessment. The observed maximum occupancy of 0.9989 may represent an extreme operating state, prolonged detector actuation or sensor malfunction. Future work should apply PeMS detector-health information, plausibility thresholds and sensitivity analyses with and without extreme observations.

Seventh, net speed change uses only the first and last observations in the window. It cannot distinguish a steady increase from an oscillatory sequence that ends at the same speed. Future models should examine maximum deceleration, acceleration and deceleration separately, coefficient of variation, adjacent-interval changes and spatial differences between upstream and downstream detectors.

Eighth, detector data cannot directly capture user perceptions, enforcement failures or the behavioral mechanisms behind unstable operations. Future validation could combine detector and crash records with geolocated complaints, structured road-user reports or agency observations. These complementary sources may help determine whether recurrent instability reflects operational conditions alone or also involves behavioral and institutional factors.

Finally, future intervention studies should evaluate distributional consequences by testing whether alerts, speed controls, ramp-management strategies or enforcement changes distribute safety benefits, delay and compliance burdens differently across communities.

These extensions would move the analysis from exploratory traffic-state classification toward a validated, interpretable and operationally useful freeway-safety framework.

5. CONCLUSION

This study demonstrates an exploratory method for classifying short-term unstable congestion using one day of station-level 5-minute PeMS data from Caltrans District 4. Rolling measures of mean speed, speed variability, occupancy variability and net speed change were constructed for 652,080 complete station-time observations. A binary proxy state was defined by the joint occurrence of high speed variability and low mean speed.

The logistic results should be interpreted as a description of this constructed classification. The positive speed-variability coefficient and negative mean-speed coefficient are partly definitional because those variables directly determine proxy-state membership. Occupancy variability provides additional separation between the classified states. Positive net speed change is also associated with higher classification odds, indicating that the result reflects speed recovery or acceleration rather than evidence that speed drops increase risk.

The framework does not estimate observed crash probability, demonstrate causal safety effects or implement Bayesian inference. Its principal contribution is methodological transparency: it shows how a simple detector-based classification can be constructed while making explicit the assumptions that limit its safety interpretation.

Before operational use, the proxy must be validated against time- and location-matched crash records. Future analysis should incorporate multiple days, timestamp-continuity checks, detector-quality screening,

station-level dependence and spatial interactions among adjacent detectors. Behavioral, participatory and distributional evidence may then be integrated to support interventions that are not only operationally effective but also understandable, context-sensitive and equitable.

6. ACKNOWLEDGMENTS

The authors thank Dr. Ipsita Banerjee and Dr. Jean Doig Godier for providing resources that supported data collection. The authors also acknowledge the California Department of Transportation (Caltrans) and the Caltrans Performance Measurement System (PeMS) for maintaining and providing public access to the freeway detector data used in this study, including observations from Caltrans District 4. Appreciation is extended to the University of California, Berkeley for institutional support. ChatGPT 5.6 was used to assist with grammar editing and language clarity. The authors reviewed and verified all AI-assisted revisions and take full responsibility for the manuscript's content, analysis and conclusions.

7. DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this work.

8. FUNDING

The authors received no financial support for the research, authorship and/or publication of this work.

9. REFERENCES

- Abdel-Aty, M., & Pande, A. (2005). Identifying crash propensity using specific traffic speed conditions. *Journal of Safety Research*, 36(1), 97–108.
- Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record*, 1897(1), 88–95.
- Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention*, 38(3), 618–625.
- Ahn, S., & Cassidy, M. J. (2007). Freeway traffic oscillations and vehicle lane-change maneuvers. In R. E. Allsop, M. G. H. Bell, & B. G. Heydecker (Eds.), *Transportation and Traffic Theory 2007* (pp. 691–710). Elsevier.
- Bickel, P. J., Chen, C., Kwon, J., Rice, J., van Zwet, E., & Varaiya, P. (2007). *Measuring traffic*. *Statistical Science*, 22(4), 581–597.
- California Department of Transportation. (2023). Performance Measurement System (PeMS). Retrieved from <https://pems.dot.ca.gov>
- California Highway Patrol. (2022). Statewide Integrated Traffic Records System (SWITRS). Sacramento, CA.
- Cassidy, M. J., & Bertini, R. L. (1999). Some traffic features at freeway bottlenecks. *Transportation Research Part B: Methodological*, 33(1), 25–42.
- Chen, C., Petty, K., Skabardonis, A., Varaiya, P., & Jia, Z. (2001). Freeway performance measurement system: Mining loop detector data. *Transportation Research Record*, 1748(1), 96–102.
- Chung, K., Rudjanakanoknad, J., & Cassidy, M. J. (2007). Relation between traffic density and capacity drop at three freeway bottlenecks. *Transportation Research Part B: Methodological*, 41(1), 82–95.
- Coifman, B. (2000). Using dual loop speed traps to identify detector malfunctions. *Transportation Research Record*, 1683(1), 47–58.
- Golob, T. F., & Recker, W. W. (2003). Relationships among urban freeway accidents, traffic flow, weather and lighting conditions. *Journal of Transportation Engineering*, 129(4), 342–353.
- Hauer, E. (1997). *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon Press/Elsevier Science.
- Hosmer, D. W., Jr., & Lemeshow, S. (2000). *Applied Logistic Regression (2nd ed.)*. John Wiley & Sons.
- Hourdos, J., Garg, V., Michalopoulos, P., & Davis, G. A. (2006). Real-time detection of crash-prone conditions at freeway high crash locations. *Transportation Research Record*, 1968(1), 83–91.

- Kerner, B. S., & Rehborn, H. (1996). Experimental features and characteristics of traffic jams. *Physical Review E*, 53(2), R1297–R1300.
- Lee, C., Hellinga, B., & Saccomanno, F. (2003). Proactive freeway crash prevention using real-time traffic control. *Canadian Journal of Civil Engineering*, 30(6), 1034–1041.
- Lee, C., Saccomanno, F., & Hellinga, B. (2002). Analysis of crash precursors on instrumented freeways. *Transportation Research Record*, 1784(1), 1–8.
- Lighthill, M. J., & Whitham, G. B. (1955). On kinematic waves II: A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A*, 229(1178), 317–345.
- Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 291–305.
- Lu, S., Liu, X., & Dai, S. (2012). Applying k-nearest neighbor algorithm for missing traffic data imputation. In *Proceedings of the 2012 International Conference on Information Engineering and Management* (pp. 67–71).
- Mauch, M., & Cassidy, M. J. (2002). Freeway traffic oscillations: Observations and predictions. In M. A. P. Taylor (Ed.), *Transportation and Traffic Theory in the 21st Century* (pp. 653–674). Pergamon.
- May, A. D. (1990). *Fundamentals of Traffic Flow* (2nd ed.). Prentice-Hall.
- Miaou, S.-P., & Song, J. J. (2005). Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion and spatial dependence. *Accident Analysis & Prevention*, 37(4), 699–720.
- Pande, A., & Abdel-Aty, M. (2006). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, 38(5), 936–948.
- Papageorgiou, M., & Kotsialos, A. (2002). Freeway ramp metering: An overview. *IEEE Transactions on Intelligent Transportation Systems*, 3(4), 271–281.
- Papageorgiou, M., Kosmatopoulos, E., & Papamichail, I. (2008). Effects of variable speed limits on motorway traffic flow. *Transportation Research Record*, 2047(1), 37–48.
- Richards, P. I. (1956). Shock waves on the highway. *Operations Research*, 4(1), 42–51.
- Saxena, N., & Jain, S. S. (2017). *Traffic Engineering and Transport Planning* (8th ed.). Nem Chand & Brothers.
- Schönhof, M., & Helbing, D. (2007). Empirical features of congested traffic states and their implications for traffic modeling. *Transportation Science*, 41(2), 135–166.
- Shahriar, M. S. (2026). Compositional Co-Design of Traffic Law Reform: Integrating Citizen-Sourced Data with Strategic Game-Theoretic Analysis. *Transportation Research Record: Journal of the Transportation Research Board*. <https://doi.org/10.1177/03611981261473514>

Shahriar, M. S., P. Luengjitvatana, M. S. A. Baig, S. Albreiki, and A. Luther. 2026. "Distributional equity of replacing fuel taxes with vehicle miles traveled fees: Evidence from five San Francisco Bay Area cities." UC Berkeley: Institute of Transportation Studies at UC Berkeley.
<https://escholarship.org/uc/item/6j93d459>

Shahriar, M. S., Mehrabuzzaman, S. N. Sadeek, M. A. Raihan, M. Hossain, and M. A. Al Mamun. 2025. "Digital storytelling as a transportation planning tool." *Transportation Research Record: Journal of the Transportation Research Board* 2679 (11): 373–395. <https://doi.org/10.1177/03611981251347617>

Shi, Q., & Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, 58, 380–394.

Tarko, A., Davis, G., Saunier, N., Sayed, T., & Washington, S. (2009). Surrogate measures of safety. ANB20(3) Subcommittee on Surrogate Measures of Safety, Transportation Research Board.

Treiber, M., Hennecke, A., & Helbing, D. (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2), 1805–1824.

Zheng, Z., Ahn, S., Chen, D., & Laval, J. (2011). Freeway traffic oscillations: Microscopic analysis of formations and propagations using wavelet transform. *Transportation Research Part B: Methodological*, 45(9), 1378–1388.

Zheng, Z., Ahn, S., & Monsere, C. M. (2010). Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention*, 42(2), 626–636.

APPENDIX A: IMPLEMENTATION, OUTPUTS AND INTERPRETATION

This appendix documents the implemented workflow, including raw-data inspection, mainline filtering, rolling-variable construction, proxy-state definition, correlation diagnostics, exploratory logistic classification and odds-ratio calculation. The appendix supports computational transparency but does not by itself establish full reproducibility because the raw data file and executable code are not embedded in the document.

A.1 Data Loading and Initial Inspection

The first step loads the PeMS station-level dataset, assigns column names and provides initial inspection of the data structure, missing values, lane types and summary statistics.

```

1 import pandas as pd
2 import numpy as np
3
4 file_path = "data.txt"
5
6 df = pd.read_csv(file_path, header=None)
7
8 df.columns = [
9     "timestamp", "station", "district", "freeway", "direction",
10    "lane_type", "station_length", "samples", "pct_observed",
11    "total_flow", "avg_occupancy", "avg_speed"
12 ]
13
14 print(df.head())
15 print("shape:", df.shape)
16 print(df.isna().sum())
17 print(df["lane_type"].dropna().unique())
18 print(df[["total_flow", "avg_occupancy", "avg_speed"]].describe())

```

The code produced the following output:

```

1      timestamp station district freeway direction lane_type station_length samples
2 0 2026-04-07  400000      4    101         S      ML           NaN         50
3   ...          50.0      0.0112    63.7
4 1 2026-04-07  400001      4    101         N      ML           NaN         65
5   ...          65.0      0.0096    66.5
6 2 2026-04-07  400002      4    101         S      ML           NaN        236
7   ...        236.0      0.0390    68.4
8 3 2026-04-07  400006      4    880         S      ML           NaN         100
9   ...        100.0      0.0208    66.4
10 4 2026-04-07  400007      4    101         N      ML           NaN         132
11   ...        132.0      0.0254    68.7
12
13 shape: (1120032, 12)
14
15 timestamp          0
16 station            0
17 district           0
18 freeway            0
19 direction          0
20 lane_type          0
21 station_length    462240

```

```

17 samples          0
18 pct_observed      0
19 total_flow        200092
20 avg_occupancy      200092
21 avg_speed         462816
22 dtype: int64
23
24 ['ML', 'FF', 'FR', 'OR', 'HV']
25
26      total_flow  avg_occupancy  avg_speed
27 count  919940.000000  919940.000000  657216.000000
28 mean    168.626977      0.048044    62.974223
29 std     168.456961      0.055691     8.308572
30 min       0.000000      0.000000     3.000000
31 25%       27.000000      0.012000    61.600000
32 50%      102.000000      0.034900    64.800000
33 75%      282.000000      0.066900    67.600000
34 max     1042.000000      0.998900    82.500000

```

The initial inspection shows that the raw dataset contains 1,120,032 observations and multiple lane types. The raw data also contains missing values in station length, total flow, average occupancy and average speed. These results justify the subsequent cleaning and mainline filtering steps.

A.2 Data Cleaning and Mainline Filtering

The data are filtered to mainline freeway observations only. Observations with missing values in average speed, average occupancy, or total flow are removed. Timestamps are converted to datetime format and the data are sorted by station and time.

```

1 df = df[df["lane_type"] == "ML"].copy()
2 df = df.dropna(subset=["avg_speed", "avg_occupancy", "total_flow"])
3
4 df["timestamp"] = pd.to_datetime(df["timestamp"], errors="coerce")
5 df = df.sort_values(["station", "timestamp"]).reset_index(drop=True)
6
7 print(df.head())
8 print("shape after cleaning:", df.shape)

```

The code produced the following output:

```

1      timestamp  station  district  freeway  direction  lane_type  station_length  samples
2      pct_observed  total_flow  avg_occupancy  avg_speed
3 0 2026-04-07   400000         4      101         S      ML           NaN         50
4      ...         50.0         0.0112      63.7
5 1 2026-04-07   400000         4      101         S      ML           NaN         53
6      ...         53.0         0.0105      64.8
7 2 2026-04-07   400000         4      101         S      ML           NaN         51
8      ...         51.0         0.0094      65.2
9 3 2026-04-07   400000         4      101         S      ML           NaN         49
10      ...         49.0         0.0101      64.1
11 4 2026-04-07   400000         4      101         S      ML           NaN         47
12      ...         47.0         0.0122      63.1
13
14 shape after cleaning: (656640, 12)

```

After cleaning, the dataset contains 656,640 observations. This step removes incomplete or non-mainline records while preserving a large analytical sample for rolling-window construction.

A.3 Rolling-Window Variable Construction

Short-term traffic indicators are constructed from the current retained observation and the two preceding retained observations for each station. Under uninterrupted data, these observations represent a rolling 15-minute window. The variables include rolling mean speed, speed standard deviation, occupancy standard deviation and net speed change. Because missing measurements are removed before the rolling calculation, the current code does not verify that every successive timestamp is exactly five minutes apart.

```

1 grouped = df.groupby("station")
2
3 df["speed_mean_15"] = grouped["avg_speed"].rolling(3).mean().reset_index(level=0,
4   drop=True)
5 df["speed_std_15"] = grouped["avg_speed"].rolling(3).std().reset_index(level=0, drop=
6   True)
7 df["occ_std_15"] = grouped["avg_occupancy"].rolling(3).std().reset_index(level=0,
8   drop=True)
9
10 df["speed_lag2"] = grouped["avg_speed"].shift(2)
11 df["speed_drop_15"] = df["avg_speed"] - df["speed_lag2"]
12
13 print(df[[
14     "avg_speed",
15     "speed_mean_15",
16     "speed_std_15",
17     "occ_std_15",
18     "speed_drop_15"
19 ]].head(10))

```

The code produced the following output:

	avg_speed	speed_mean_15	speed_std_15	occ_std_15	speed_drop_15
1	0	63.7	NaN	NaN	NaN
2	1	64.8	NaN	NaN	NaN
3	2	65.2	64.566667	0.776745	0.001361
4	3	64.1	64.700000	0.556776	0.000917
5	4	63.1	64.133333	1.050397	0.002274
6	5	63.6	63.600000	0.500000	0.001626
7	6	63.3	63.333333	0.251661	0.001762
8	7	64.2	63.700000	0.458258	0.000945
9	8	63.5	63.666667	0.472582	0.000265
10	9	63.3	63.666667	0.472582	0.000321
11					

The first two retained observations for each station have missing rolling values because three observations are required. Subsequent values describe the retained three-observation sequence. They should not automatically be assumed to represent exactly 15 clock minutes when an intervening record has been removed.

A.4 Correlation Diagnostic Between Speed and Occupancy Variability

A correlation diagnostic was conducted to examine whether the 15-minute speed variability and 15-minute occupancy variability measures capture overlapping information. The analysis used the final complete model dataset after dropping observations with missing rolling window values.

```
1 grouped = df.groupby("station")
2
3 # Correlation diagnostic between speed and occupancy variability
4 corr_data = df[["speed_std_15", "occ_std_15"]].dropna()
5
6 pearson_r, pearson_p = pearsonr(
7     corr_data["speed_std_15"],
8     corr_data["occ_std_15"]
9 )
10
11 spearman_r, spearman_p = spearmanr(
12     corr_data["speed_std_15"],
13     corr_data["occ_std_15"]
14 )
15
16 X_vif = sm.add_constant(
17     df[["speed_std_15", "occ_std_15", "speed_mean_15", "speed_drop_15"]].dropna()
18 )
19 for i, col in enumerate(X_vif.columns):
20     print(col, variance_inflation_factor(X_vif.values, i))
21
22 print("Pearson correlation:", round(pearson_r, 4))
23 print("Pearson p value:", pearson_p)
24 print("Spearman correlation:", round(spearman_r, 4))
25 print("Spearman p value:", spearman_p)
```

The code produced the following output:

```
1 Pearson correlation: 0.6269
2 Pearson p value: 0.0
3 Spearman correlation: 0.5772
4 Spearman p value: 0.0
5 const 35.40
6 speed_std_15 1.65
7 occ_std_15 1.99
8 speed_mean_15 1.63
9 speed_drop_15 1.15
```

The Pearson correlation between speed variability and occupancy variability is 0.6269, while the Spearman rank correlation is 0.5772. Both correlations are statistically significant at $p < 0.001$. The VIF values are 1.65 for speed variability and 1.99 for occupancy variability, indicating that the two variables are related but do not create severe multicollinearity in the model.

A.5 Instability-Congestion Proxy Definition

The binary instability-congestion proxy indicator is constructed from high rolling speed variability and low rolling mean speed. An observation is assigned to the proxy state when speed standard deviation is at or above its sample 75th percentile and mean speed is at or below its sample 25th percentile. This is a researcher-defined operational classification and not an observed crash outcome.

```

1 speed_std_cut = df["speed_std_15"].quantile(0.75)
2 speed_low_cut = df["speed_mean_15"].quantile(0.25)
3
4 df["risk_flag"] = (
5     (df["speed_std_15"] >= speed_std_cut) &
6     (df["speed_mean_15"] <= speed_low_cut)
7 ).astype(int)
8
9 print(df[[
10     "avg_speed",
11     "speed_std_15",
12     "occ_std_15",
13     "speed_drop_15",
14     "risk_flag"
15 ]].head(10))
16
17 print(df["risk_flag"].value_counts())

```

The code produced the following output:

	avg_speed	speed_std_15	occ_std_15	speed_drop_15	risk_flag
0	63.7	NaN	NaN	NaN	0
1	64.8	NaN	NaN	NaN	0
2	65.2	0.776745	0.001361	1.5	0
3	64.1	0.556776	0.000917	-0.7	0
4	63.1	1.050397	0.002274	-2.1	0
5	63.6	0.500000	0.001626	-0.5	0
6	63.3	0.251661	0.001762	0.2	0
7	64.2	0.458258	0.000945	0.6	0
8	63.5	0.472582	0.000265	0.2	0
9	63.3	0.472582	0.000321	-0.9	0
risk_flag					
0	584924				
1	71716				
Name: count, dtype: int64					

After observations without complete rolling values are removed, the final model dataset contains 652,080 observations. The threshold rule assigns 71,716 observations to the instability-congestion proxy state and 580,364 observations outside that state. These counts describe the selected sample thresholds and should not be interpreted as counts of hazardous events or crashes.

A.6 Exploratory Logistic Classification

The implemented model estimates the probability that a station-time observation is assigned to the constructed instability-congestion proxy state. Inputs are rolling speed variability, rolling occupancy variability, rolling mean speed and rolling net speed change.

Speed variability and mean speed directly define the dependent variable. Their coefficients are therefore partly mechanical and are retained only to document the classification structure. The model also uses conventional standard errors without adjustment for repeated observations within detector stations. Coefficients, intervals and p values are consequently descriptive and nominal.

The code fits the logistic model and extracts coefficients, conventional standard errors, z statistics, nominal p values, confidence intervals and odds ratios.

```
1  import pandas as pd
2  import numpy as np
3  import statsmodels.api as sm
4
5  model_df = df[[
6      "risk_flag",
7      "speed_std_15",
8      "occ_std_15",
9      "speed_mean_15",
10     "speed_drop_15"
11  ]].dropna()
12
13  X = model_df[[
14      "speed_std_15",
15      "occ_std_15",
16      "speed_mean_15",
17      "speed_drop_15"
18  ]]
19  X = sm.add_constant(X)
20  y = model_df["risk_flag"]
21
22  model = sm.Logit(y, X).fit()
23
24  print(model.summary())
25
26  results_table = pd.DataFrame({
27      "variable": model.params.index,
28      "coef": model.params.values,
29      "std_error": model.bse.values,
30      "z_stat": model.params.values / model.bse.values,
31      "p_value": model.pvalues.values,
32      "lower_95": model.conf_int()[0].values,
33      "upper_95": model.conf_int()[1].values,
34      "odds_ratio": np.exp(model.params.values)
35  })
36
37  print(results_table)
```

The software output reports nominal p values below 0.001 for all four variables. Values displayed as 0.0 reflect numerical display precision rather than exact zero. Because the observations are clustered within detector stations and two variables directly define the outcome, these p values should not be interpreted as confirmatory statistical evidence.

The code produced the following output:

```

1 Optimization terminated successfully.
2 Current function value: 0.161604
3 Iterations 8
4
5 Logit Regression Results
6 =====
7 Dep. Variable: risk_flag No. Observations: 652080
8 Model: Logit Df Residuals: 652075
9 Method: MLE Df Model: 4
10 Date: Mon, 20 Apr 2026 Pseudo R-squ.: 0.5336
11 Time: 00:00:15 Log-Likelihood: -1.0538e+05
12 converged: True LL-Null: -2.2593e+05
13 Covariance Type: nonrobust LLR p-value: 0.000
14 =====
15 coef std err z P>|z| [0.025 0.975]
16 -----
17 const 3.2262 0.038 84.269 0.000 3.151 3.301
18 speed_std_15 1.5963 0.007 230.752 0.000 1.583 1.610
19 occ_std_15 6.2273 0.651 9.572 0.000 4.952 7.502
20 speed_mean_15 -0.1228 0.001 -196.985 0.000 -0.124 -0.122
21 speed_drop_15 0.0530 0.002 24.155 0.000 0.049 0.057
22 =====
23
24 Detailed coefficient table with p values:
25
26 variable coef std_error z_stat p_value lower_95 upper_95 odds_ratio
27 const 3.2262 0.0380 84.90 0.0 3.1517 3.3007 25.18
28 speed_std_15 1.5963 0.0070 228.04 0.0 1.5826 1.6100 4.93
29 occ_std_15 6.2273 0.6510 9.57 1.05e-21 4.9513 7.5033 506.36
30 speed_mean_15 -0.1228 0.0010 -122.80 0.0 -0.1248 -0.1208 0.88
31 speed_drop_15 0.0530 0.0020 26.50 6.55e-129 0.0491 0.0569 1.05

```

The positive coefficient for speed variability and negative coefficient for mean speed reproduce the threshold directions used to construct the dependent variable. They are not independent safety findings. Occupancy variability is positively associated with proxy-state membership. Net speed change is also positive, meaning that greater speed recovery or acceleration, rather than a larger speed drop, corresponds to higher classification odds under the implemented coding. None of these coefficients estimates observed crash probability.

A.7 Odds-Ratio Interpretation

The coefficients are exponentiated to obtain odds ratios. Odds ratios for continuous variables depend on their measurement units and should be interpreted using realistic increments.

The speed-variability odds ratio of 4.93 corresponds to a 1-mph increase but is partly definitional because speed variability determines proxy-state membership. The occupancy-variability odds ratio of 506.36 corresponds to a full 1.00-unit increase. A more meaningful 0.01-unit increase corresponds to an odds ratio of 1.064, or approximately 6.4% higher classification odds. The mean-speed odds ratio of 0.88 is also partly definitional. For net speed change, a 5-mph increase corresponds to an odds ratio of approximately 1.30.

```

1 odds = np.exp(model.params)
2 print(odds)

```

The code produced the following output:

```
1  const                25.184100
2  speed_std_15         4.934493
3  occ_std_15          506.364916
4  speed_mean_15        0.884410
5  speed_drop_15        1.054407
6  dtype: float64
```

The odds ratio for speed variability is approximately 4.93, indicating that higher short-term speed variability substantially increases the odds of a high-risk traffic state. The odds ratio for occupancy variability is approximately 506.36, showing that occupancy fluctuation is a very strong separator between normal and high-risk states in this model. The odds ratio for mean speed is 0.88, indicating that higher mean speed reduces the odds of the high-risk state.

APPENDIX B: DATA SOURCE DOCUMENTATION

B.1 Data Platform

The dataset was obtained from the California Department of Transportation Performance Measurement System (PeMS). PeMS provides high-resolution freeway detector data across California, including traffic flow, speed and occupancy at 5-minute temporal resolution.

B.2 Dataset Description

This study uses station-level 5-minute data from Caltrans District 4 for April 7, 2026. The raw dataset contains 1,120,032 observations and includes the following variables:

- timestamp;
- station identifier;
- freeway identifier and direction;
- lane type;
- total flow, measured as vehicles per interval;
- average occupancy; and
- average speed.

After filtering to retained mainline observations and removing records with missing flow, occupancy or speed values, 656,640 observations remain. After incomplete rolling observations are removed, the exploratory logistic sample contains 652,080 observations.

B.3 Access and Reproducibility

Data are available through the official PeMS clearinghouse at <https://pems.dot.ca.gov>. The dataset was downloaded from the Station 5-Minute Data module using the following selections:

Specifically, the dataset was downloaded from the “station 5-minute data” module under district 4. To reproduce the dataset, users can navigate to:

- **Data Type:** Station 5-Minute
- **District:** 4
- **Date:** April 7, 2026

Appendix A documents the implemented processing sequence. Exact reproduction additionally requires the downloaded raw file, the final executable code and the software/package versions used in the analysis.