Transferability and robustness of real-time freeway crash risk assessment

Cameron Shew [a], Anurag Pande [a,*], Cornelius Nuworsoo [b]

A B S T R A C T

Introduction: This study examines the data from single loop detectors on northbound (NB) US-101 in San Jose, California to estimate real-time crash risk assessment models. *Method:* The classification tree and neural network based crash risk assessment models developed with data from NB US-101 are applied to data from the same freeway, as well as to the data from nearby segments of the SB US-101, NB I-880, and SB I-880 corridors. The performance of crash risk assessment models on these nearby segments is the focus of this research. *Results:* The model applications show that it is in fact possible to use the same model for multiple freeways, as the underlying relationships between traffic data and crash risk remain similar. *Impact on Industry:* The framework provided here may be helpful to authorities for freeway segments with newly installed traffic surveillance apparatuses, since the real-time crash risk assessment models from nearby freeways with existing infrastructure would be able to provide a reasonable estimate of crash risk. The robustness of the model output is also assessed by location, time of day, and day of week. The analysis shows that on some locations the models may require further learning due to higher than expected false positive (e.g., the I-680/I-280 interchange on US-101 NB) or false negative rates. The approach for post-processing the results from the model provides ideas to refine the model prior to or during the implementation.

1. Introduction

Much progress has been made in recent years in shifting from reactive (incident detection) to proactive (real-time crash risk assessment) traffic strategies, as traffic safety on freeways continues to be a growing concern. Reliable models that can take in real-time loop detector information and discern normal flow conditions from crash-prone conditions are keys to implementing crash preventative measures. This area of research has gained increased attention with increased capabilities to collect, archive, and analyze these data throughout the world. The examples of this work from around the world include: Canada (Lee, Hellinga, & Saccomanno, 2003), Europe (Christoforou, Cohen, & Karlaftis, 2011), Japan (Hossain & Muromachi, 2010), and several U.S. states including Florida (Pande & Abdel-Aty, 2006a, 2006b), Southern California (Oh, Oh, Ritchie, & Chang, 2001), and Colorado (Yu, Abdel-Aty, & Ahmed, 2013).

These studies, with ITS-related archived data from different sources and advanced modeling techniques, have demonstrated the feasibility of this proactive approach. A number of these studies have set up the problem in the form of binary classification (Pande & Abdel-Aty, 2006a, 2006b; Oh et al., 2001; Yu et al., 2013) and have developed sophisticated classifiers to separate crash prone conditions on uninterrupted flow facilities from 'normal' traffic

conditions. A classification approach is applied in the present study as well, with classification trees providing a preliminary assessment. The classification tree models identified the most significant traffic variables for separating the conditions before historical crashes from traffic conditions drawn from random time/location on NB US-101 freeway in the San Jose (California) area. These significant traffic variables are then subject to Levenberg-Marquardt (LM) training algorithm to estimate MLP neural network based classifiers. The focus of this work is to examine the performance of these models in a variety of settings including at other freeways segments (SB US-101 and NB and SB I-880) in the area and at different times of day/days of week. This is a follow-up of one of the author's most recent work in this area, where models developed from I-4 data in the central Florida region were assessed with data from I-95 corridor (Pande, Das, Abdel-Aty, & Hassan, 2011).

1.1. Background

Past studies have already demonstrated that statistical links can be established between real-time traffic flow variables (such as average speed, volume, occupancy, and their respective standard deviations) and crash likelihood (e.g., Golob & Recker, 2004; Xu, Liu, Wang, & Yu, 2011). However, most of these previous studies have focused on a single highway corridor. Some of the studies that did use data from multiple corridors combined the data from all freeways without transferability assessment (e.g., Abdel-Aty, Pande, Das, & Knibbe,

2008; Golob & Recker, 2004). In a recent study that did assess the transferability of the crash risk assessment models, the results did not appear to be promising, in terms of transferability for the model estimated from I-4 data onto the I-95 corridor in the central Florida region (Pande et al., 2011).

In this regard, the authors view transferability as one critical issue that still needs further addressing. This study advances the current body of knowledge by exploring whether traffic and crash patterns are similar enough in close geographic proximity to apply the estimated classification models from one roadway segment onto another. Since gathering data from different sources and combining them is a significant effort, it would be worthwhile to know whether models developed from one freeway can be applied to the data from other freeways. While it may be unreasonable for models developed with data from a dense urban freeway environment to perform well on a rural freeway corridor, no studies have yet tested models from the same geographical area to other freeways in close proximity. This study makes an effort in that direction. The effort would be helpful, for example, since the overall objective of the work recently done in Japan was to layout the infrastructure for traffic surveillance data collection (Hossain & Muromachi, 2010). Additionally, classification errors by the models (both in the form of false positives and false negatives) are analyzed for any pattern, with regards to the day of the week, time of the day, and physical location on the route.

This paper is organized as follows: In the next section, data collection and preparation, including the study area, are described. The following section details the steps of the data mining-based classification process. The section after that briefly discusses the model estimation process, along with the performance of the model on the validation dataset from the same freeway segment (NB US-101) as the training dataset. Note that since the focus of this study is on transferability and robustness, the model estimation and validation process (essentially testing the models with data from the same freeway corridor) is not very detailed. The section after that demonstrates the classification performance of the model on nearby freeway sections. After a discussion on transferability, the model outputs are examined in different settings. This examination provides insights into settings where the neural network-based model may require further learning.

2. Data collection and preparation

2.1. Study area

For this study, crash data from a 16-month period (January 2010 through April 2011) were collected on four freeway corridors: US-101 NB and SB and I-880 NB and SB in the San Jose area. These corridors in the city of San Jose run through dense urban development and are among the busiest in the South Bay Area. The mile point boundaries and number of crashes for these corridors during the aforementioned period are provided in Table 1.

Note that the crash counts listed in Table 1 are total crash counts and that not all of these crashes could be used for the analysis, since mechanical failure occasionally rendered the corresponding loop detectors unable to record any data for some crash observations. The segments are shown in a map in Fig. 1.

2.2. Data preparation

Both traffic and crash data were downloaded from Caltrans' PeMS (Performance Measurement System) clearinghouse (Varaiya, 2007). The downloaded traffic data included the following variables for each vehicle detection station (VDS): time and date, milepost, average speed, volume, and lane-occupancy information measured every 30 seconds at the corresponding VDS. It is worth mentioning that among these variables, volume and lane-occupancy are measured variables. The 30-second average speed is calculated using these two measurements along with a predetermined effective vehicle length. This is in contrast with dual loops which can measure speeds directly. In this study we have chosen to use only the data reported by the loop detectors independently and have not use speed information.

Since the modeling approach adopted here was binary classification, the same traffic data were also collected for randomly-generated non-crash cases. The traffic data corresponding to the 'non-crash' cases would be representative of the 'normal' conditions on the freeways, as opposed to the traffic data corresponding to the historical crash cases, representing crash prone conditions. Each crash and non-crash event was merged with corresponding traffic data from six VDS locations, the three nearest in both the upstream and downstream directions. The authors note that the Caltrans loops are typically spaced between 0.5 and 0.8 miles apart in this study area.

One of the previous studies by one of the authors (Pande & Abdel-Aty, 2008) noted that there is significant noise in the raw 30-second loop detector data and therefore they are not suited for modeling purposes. Hence, for each of the six VDS locations, individual variables were averaged across all lanes, and aggregated into five minute intervals. These intervals are: 0-5 minutes after the crash (time slice 0), 0-5 minutes before the crash (time slice 1), 5-10 minutes before the crash (time slice 2), 10-15 minutes before the crash (time slice 3), and 15-20 minutes before the crash (time slice 4). For each of the time slices, standard deviations of the variables were also calculated, since past studies (e.g., (Lee et al., 2003) note that variation in traffic parameters is critically associated with the freeway crash potential. Based on the results documented in Pande, Nuworsoo, and Shew (2012), it was found that time slice 2 is the most appropriate period for real-time crash risk estimation. Therefore, the analysis presented here focuses on models developed with variables aggregated for the interval 5-10 minutes before the crash.

The aim of this research is to develop models with the ability to separate conditions prone to crashes on these San Jose area freeways from 'normal' freeway traffic. The normal conditions were represented by the random sample of non-crash loop detector data. To generate a random non-crash sample, the total study period was divided into one minute periods from which a random sample of times could be selected as the time of the non-crash 'event.' Similarly, milepost location for non-crash cases could also be drawn from any milepost from the beginning to the end of the corresponding corridor. From all possible combinations of date-time and mileposts, a sample of non-crash cases were derived. To ensure that the neural network based classifier adequately learns 'normal' conditions, there were 10 non-crash 'events' for every crash event used in the analysis.

The nearest three VDS locations in both upstream and downstream directions of the event location milepost were also identified for all of these non-crash events. The time horizon (5-10 minutes

Table 1
Freeway sections details and crash counts.

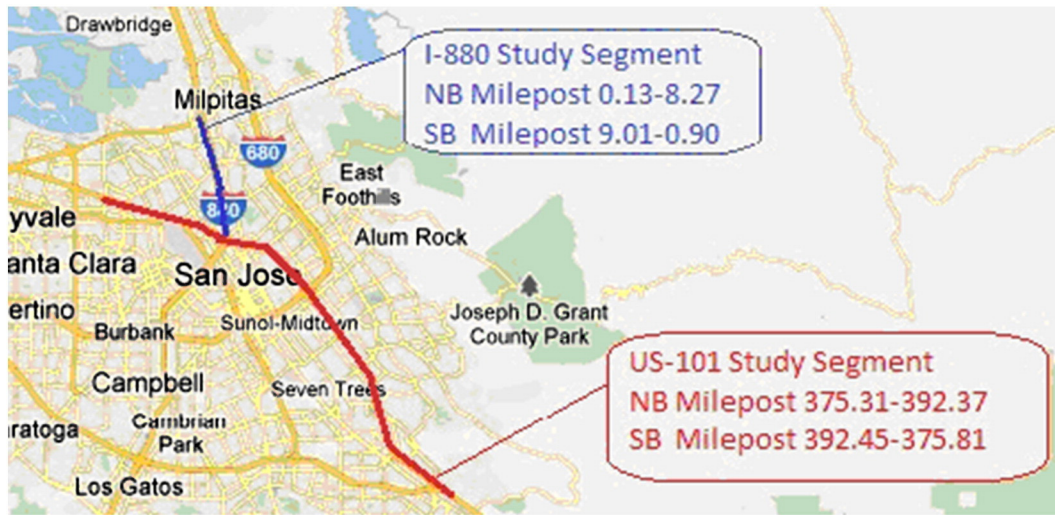| Freeway	Starting Milepost	Ending Milepost	Study Segment Length	Crash Counts
US-101 NB	375.31	392.37	17.1 mi.	2176
US-101 SB	392.45	375.81	16.6 mi.	1903
I-880 NB	0.13	8.27	8.1 mi.	937
I-880 SB	9.01	0.9	8.1 mi.	1128

Fig. 1. Map of the San Jose freeways with highlighted study segments (Source: Google Maps™).

before the time of event) was also the same as the crash events and was used to calculate the averages and standard deviations of all traffic variables. To relate these data with crash and non-crash cases, a data mining approach was adopted.

3. Modeling methodology

Data mining is the analysis of "observational" databases to find potentially useful relationships (Hand, Mannila, & Smyth, 2001). Freeway traffic surveillance data, collected through loop detectors, is one such "observational" database maintained for various ITS (Intelligent Transportation Systems) applications, such as travel delay estimation and dissemination. In this research, a data mining process is used to relate the measures of traffic conditions (data from VDS) with historical crashes on freeways. Note that data mining based analysis is preferred here, since techniques from traditional statistics are more suitable for handling the data obtained through an experimental design, which is clearly not the case here. The data mining process applied for this study has two key components, namely a variable selection procedure (based on classification tree) and a neural network-based modeling procedure, with parameters identified through the preceding classification tree as inputs. These components of the data mining process are described in the ensuing sections.

3.1. Classification tree application for variable selection

The basic idea in classification tree model is to recursively split each (non-terminal) node such that the descendent nodes are 'purer' or further segmented with even larger (smaller) proportion of the majority (minority) class, than the parent node. In the context of binary classification the 'purest' node will have all observations belonging to only one of the two classes. Each rule assigns an observation to a group based on the value of an input. One rule is applied after another, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The final or terminal nodes are called leaves. For each leaf, a posterior probability of the observations belonging to each class is calculated based on the proportion of the two classes (crash and non-crash cases in the present study) in that leaf.

All possible splits for all variables are evaluated and ranked based on one of three criteria, namely, Chi-square test (used in this study), Entropy reduction, or Gini reduction, to choose amongst the available splits at every non-terminal node. According to Chi-square test criterion, the split resulting in the cross-frequency table with maximum $-log$ $(p$-vlaue$)$ (i.e., minimum p-value) is selected. Note that the selection of the split with the minimum p-value would ensure that Child nodes resulting from the selected split are more homogeneous in nature. The splitting process is continued until there is no (or less than a pre-specified minimum) reduction in impurity and/or the limit for minimum number of observation in a leaf is reached (SAS Institute, 2011).

Breiman (1984) devised a variable importance measure (VIM) for trees. This measure may be applied as a criterion to select a promising subset of variables for tools such as neural network. In this study, the VIM used has been scaled by maximum importance for the tree so that it lies between 0 and 1 with the VIM > 0.05, indicating a significant variable worthy of being included in neural network-based classification models. A detailed discussion of the application of VIM in this context may be found in Pande and Abdel-Aty (2006a).

3.2. MLP neural network architecture and training

A neural network is defined as a parallel-distributed processor made up of simple processing units having natural propensity for learning from an available training dataset and making generalizations for future datasets (Christodoulou & Georgiopoulos, 2000). Generalization refers to the ability of a "trained" network to provide satisfactory responses even for the inputs that it has not seen during the training process. Neural network models may usually be specified by three entities, namely the model of processing elements themselves, the model of interconnections and structures (i.e., network topology), and the learning rules. In this study, a multi-layer perceptron (*MLP*) network with feed-forward connections was used. The procedure adopted for training, crucial in the performance of a neural network, was the LM (Levenberg–Marquardt) algorithm. LM algorithm is an iterative optimization procedure to minimize the error of the neural network model. The theoretical details of the algorithm may be found in Wilamowski, Iplikci, Efe, et al. (2001). Details of the procedure for binary classification of traffic patterns in the present context may also be found in Pande and Abdel-Aty (2006a). According to Wilamowski et al. (2001), the LM algorithm becomes computationally impractical for large size neural networks with an increase in the number of independent variables. Therefore, the variable selection process through classification tree is quite useful for this application.

4. Model estimation

These modeling techniques were used to estimate 2-VDS, 4-VDS, and 6-VDS models with traffic information aggregated from 5-10 minutes before the crashes. 2-VDS model here refers to the

fact that one VDS station in each direction upstream and downstream of the crash (or non-crash) location contributed traffic information as input to the model. Similarly, 4-VDS and 6-VDS model indicates that two or three stations, respectively, in each direction upstream and downstream of the crash (or non-crash) location contributed traffic information to the model. While this paper documents the models for time slice 2 only, further details of models for other time slices and variables found significant by the classification tree may be found in Pande et al. (2012). Once the significant variables were identified, the variables were subjected to the LM algorithm-trained neural network model. It was found that 6-VDS model did not improve on the classification performance of the 4-VDS model. Combined with the fact that it required six VDS locations to report data simultaneously to produce the crash risk estimate, 6-VDS models were not considered practical for application. Hence, from here on, we will only discuss 2-VDS and 4-VDS models. The variables found significant by the classification tree for these sets of models may be found in Pande et al. (2012). The significant variables identified in this study were consistent with the findings from literature from past research.

The output for the LM trained neural network model was the posterior probability (between 0 and 1) of an observation being a crash, with a higher number indicating a greater risk of crash. It should also be noted the posterior probability is not the probability of crash occurrence at a given point in time/location but is a measure providing the relative likelihood of crash occurrence given the composition of the sample. The model's classification performance can be assessed by the proportion of crashes identified within certain percentiles of observations with highest posterior probability. A 30 percentile criterion was used in this study to assess classification performance. If one were to randomly classify observations (referred to as a random 'baseline' model) from a dataset, within any set of 30% observations one would expect to capture 30% of all crashes in the dataset. Any model can be assessed for its classification based on the difference between crashes it identifies within the top 30 percentile observations with highest posterior probability versus 30%. This comparison on the validation dataset is provided in the next section. Note that the validation dataset comprises of the NB US-101 data not used in the training of the neural network model or classification tree-based variable selection.

4.1. All crash model comparison: US 101 validation dataset

The performance of best 2-VDS and 4-VDS model are shown along with a random baseline model in Fig. 2.
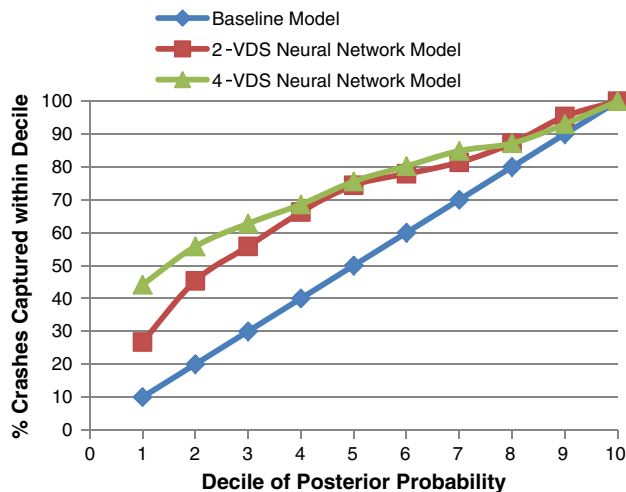

Fig. 2. Model performance comparison on validation dataset from US 101 NB section.

The curves in Fig. 2 show the percentage of the crashes in the validation dataset captured within various deciles (10 percentiles) of posterior probability by 2-VDS and 4-VDS model on the y-axis. On the x-axis the percentiles are shown at equal intervals of 10. Fig. 3 also demonstrates 'performance' of the random baseline model that represents the expected percentage of crashes identified in the validation sample if one randomly assigns validation dataset observations as crash and non-crash. A model can be assessed by examining the separation of its corresponding lift curve from the random baseline curve. The crash identification performance of the models (up to 50 percentiles) corresponding to these curves is provided in Table 2. The threshold may be selected at the application stage, based on the desired number of warnings.

The margin in the parentheses shows the differential between crashes identified by the corresponding model and the Baseline model. It may be seen from Table 2 that the 4-VDS model identifies 62.79, 68.60, and 75.58% crashes in the validation dataset, respectively, at 30, 40, and 50 percentile thresholds. For the 2-VDs model, the percentages are 55.81, 66.28, and 74.42% respectively. As we increase the percentage of observations declared as crash (i.e., higher percentile threshold), the crash identification will obviously improve, but the percentage of non-crash cases correctly identified would decrease. Hence, there is a trade-off involved, since as one assigns more patterns as crashes, the 'false alarms' increase. These percentages, when compared to the baseline model, clearly show that the both 2-VDS and 4-VDS neural network models are capable of identifying relative crash risk based on the posterior probability output measure. While the 4-VDS model does appear to be more capable with higher proportion of crashes being identified at various thresholds, the performance needs to be seen in the context of the higher data requirement. Note that the output from this model requires data to be available from 4 VDS locations simultaneously. Therefore, it was decided that the robustness and transferability analysis will be carried out for both models.

5. Are these models robust and transferable?

5.1. Transferability evaluation

Transferability evaluation, that is the potential to apply the predictive model developed on one freeway segment to other similar facilities in the region, is the focus of this research. The two models, 2-VDS and 4-VDS, were applied to the datasets from the US-101 SB, I-880 NB, and I-880 SB freeway sections described earlier in the paper. As was discussed in the literature review, previous studies have either NOT addressed the issue (which is critical to real-time application in a network) or tried to apply the model on dissimilar facilities (such as in a different study area) and subsequently reported lack of transferability (Pande et al., 2011).
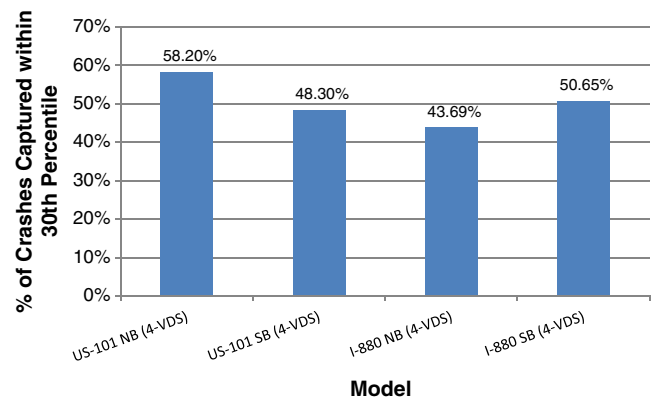

Fig. 3. Transferability Analysis for the Models on other freeways.

Table 2
Performance of the classification models over the validation dataset.

Percentiles of Posterior Probability	Percentage of crashes identified in the validation dataset		
	Baseline model	4-VDS	2-VDS
10	10	44.19 (+34.19)	26.74 (+16.74)
20	20	55.81 (+35.81)	45.35 (+25.35)
30	30	62.79 (+32.79)	55.81 (+25.81)
40	40	68.60 (+28.60)	66.28 (+26.28)
50	50	75.58 (+25.58)	74.42 (+24.42)

The posterior probability of the observation being a crash was estimated for each observation in the datasets. We then examined the proportion of crashes in the dataset captured within the 30% observations having the highest posterior probability. Note that the model that identifies higher proportion of crashes within 30th percentile is considered a better model. The cumulative percentages of identified crashes for the best model (either 2-VDS or 4-VDS) on each of the three corridors are depicted in Fig. 3.

In addition to these three nearby freeway corridors, the model was also applied on the complete set of US-101 NB data. It was done since the results shown in Table 1 and Fig. 2 are based on applying the tree model on the validation dataset (which is only 30% of observations from US-101 NB). Applying the model on the US-101 NB dataset allows us to compare all four models based on a consistent measure. It is worth mentioning that these percentages of captured crashes for transferability analysis are obtained from different models (4-VDS or 2-VDS), identifying higher percentages of crashes depending on the model. The figure also provides the model type with the percentage of crashes identified.

I-880 SB is the corridor where the model estimated from US-101 NB data seems to be the most readily transferable, based on the classification performance. Overall, all three models are better than the random baseline performance, but are not as good as the US-101 NB data. Hence, the conclusion from this analysis is that while the models do transfer, the performance may be improved by adding data from the same freeways at the model calibration (training and validation) stage.

5.2. Analysis of robustness: classification errors

Of course, no model will be a perfect classifier of crashes and non-crash cases. However, it is important to identify situations in which a model is more 'perfect' than other situations. Hence, these models' outputs for US-101 NB, as well as to data from the three other freeway segments, were then assessed for their classification performance in a variety of situations. This analysis of robustness has not been carried out in the similar studies and may help in identifying location and times of day/days of week for which additional training of the neural network may be warranted. To study the robustness of the models, for each model (2-VDS and 4-VDS models discussed above), all cases (crash and non-crash) were sorted in descending posterior probability output so the ones most likely to be a crash were at the top and the least likely ones at the bottom. All non-crashes in the top 10% observations (most likely to be crashes according to the model) were labeled as "false positives" and all crashes in the bottom 10% of observations (least likely to be crashes according to the model) were labeled as "false negatives." This process was repeated for cases in all four segments of the freeway.

To examine the robustness of the model, we examined patterns in these "false positives" and "false negatives:" day of week/time of day (off-peak, morning peak, or afternoon peak), and location of the crash/non-crash case. While potentially significant, incident duration could not be analyzed in this framework since the California Highway Patrol database from PeMS was missing this information for most

of the cases. The findings for the false positives and false negatives for each model were compared to the model performance on all crash and non-crash cases. It should be noted that the false negative (crash cases deemed safe by the model) is less conclusive due to the smaller sample size, although there are clearly observable trends. The trends shown below are from 4-VDS model for US-101 NB.

5.2.1. Time of day and day of week

Fig. 4 shows that while more than 80% of overall data were from the off-peak locations, among the "false positives" and "false negatives" off-peak periods represented a smaller proportion. The morning peak is overrepresented in "false positives." It indicates that while the model deems the morning peak conditions to be crash prone, there are fewer crashes. It may be caused by the fact that the drivers are more attentive during morning peak periods and are able to successively navigate through crash prone conditions.

While the trends is not as pronounced in the afternoon, it appears that there are more false negatives indicating that in the afternoon drivers end up in crashes even when the model is not detecting these conditions. While drivers' fatigue may play a role here, it could also be caused by the fact that congestion in the afternoon can back up much faster and those conditions are not captured by the model, since it uses data from up to 10 minutes before the crash.

5.2.2. Location

We next evaluated whether there are any locations that were over-represented in the misclassifications. The first upstream VDS location for all "false positives" and "false negatives" was determined as a subset of the original spatial distribution of all incidents. While most locations had the false positives and false negatives consistent with their proportion in the overall data, there were three locations that were noteworthy on US-101 NB:

VDS 401890: Higher percentage of "false positives:" Fig. 5 shows that this VDS is located at the US-101/I-280/I-680 interchange, where a large amount of weaving, merging activity may lead to higher speed variations. Higher level of turbulence prevailing in this location means that the drivers need to carefully navigate through this section, since the model deems this location to be crash prone more often than others.
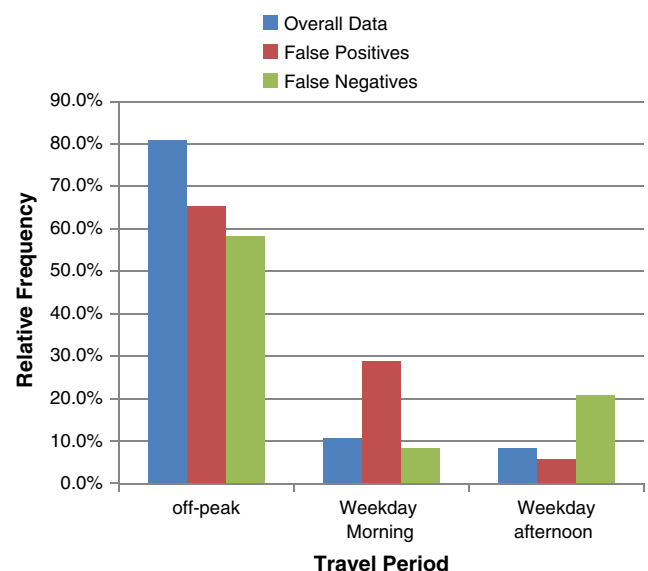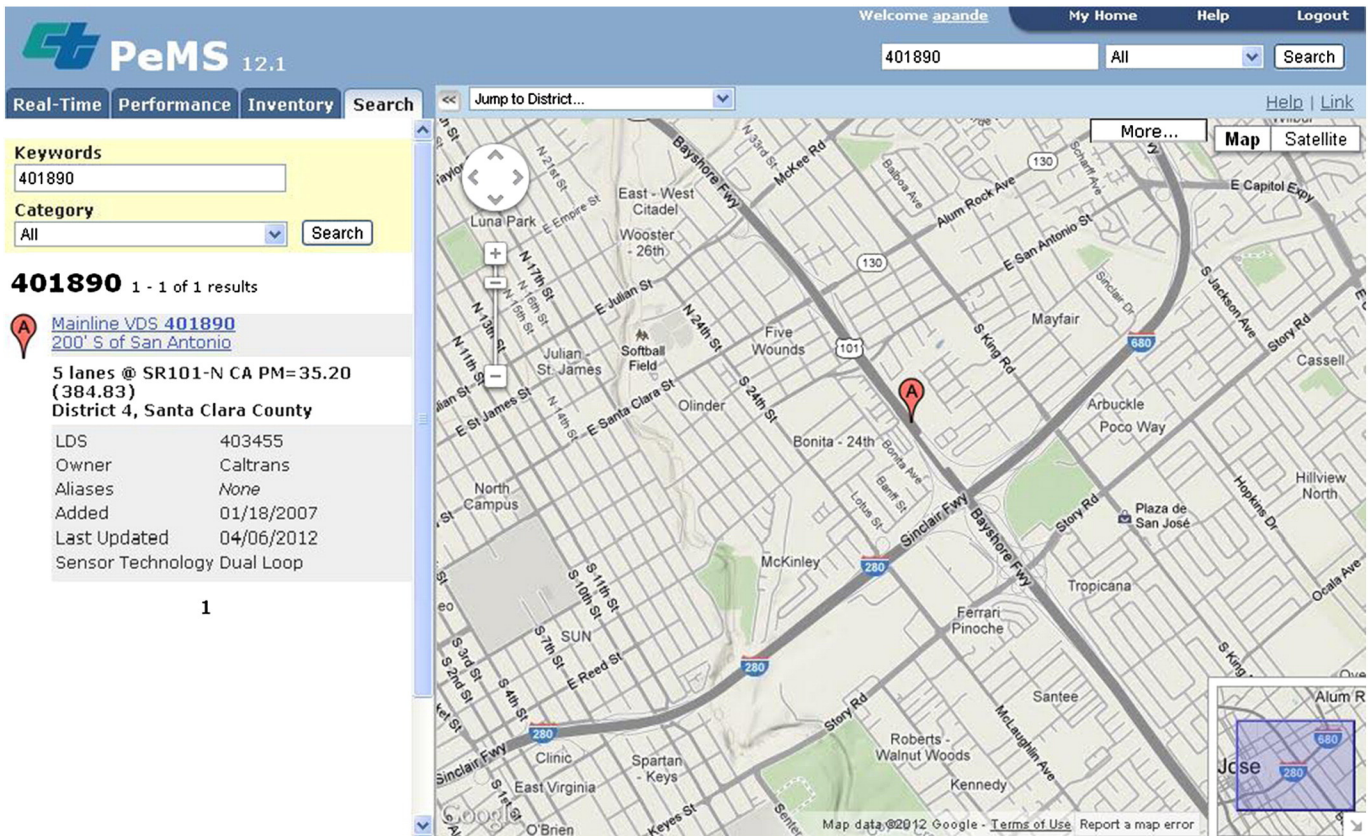

Fig. 4. Robustness of the best model.

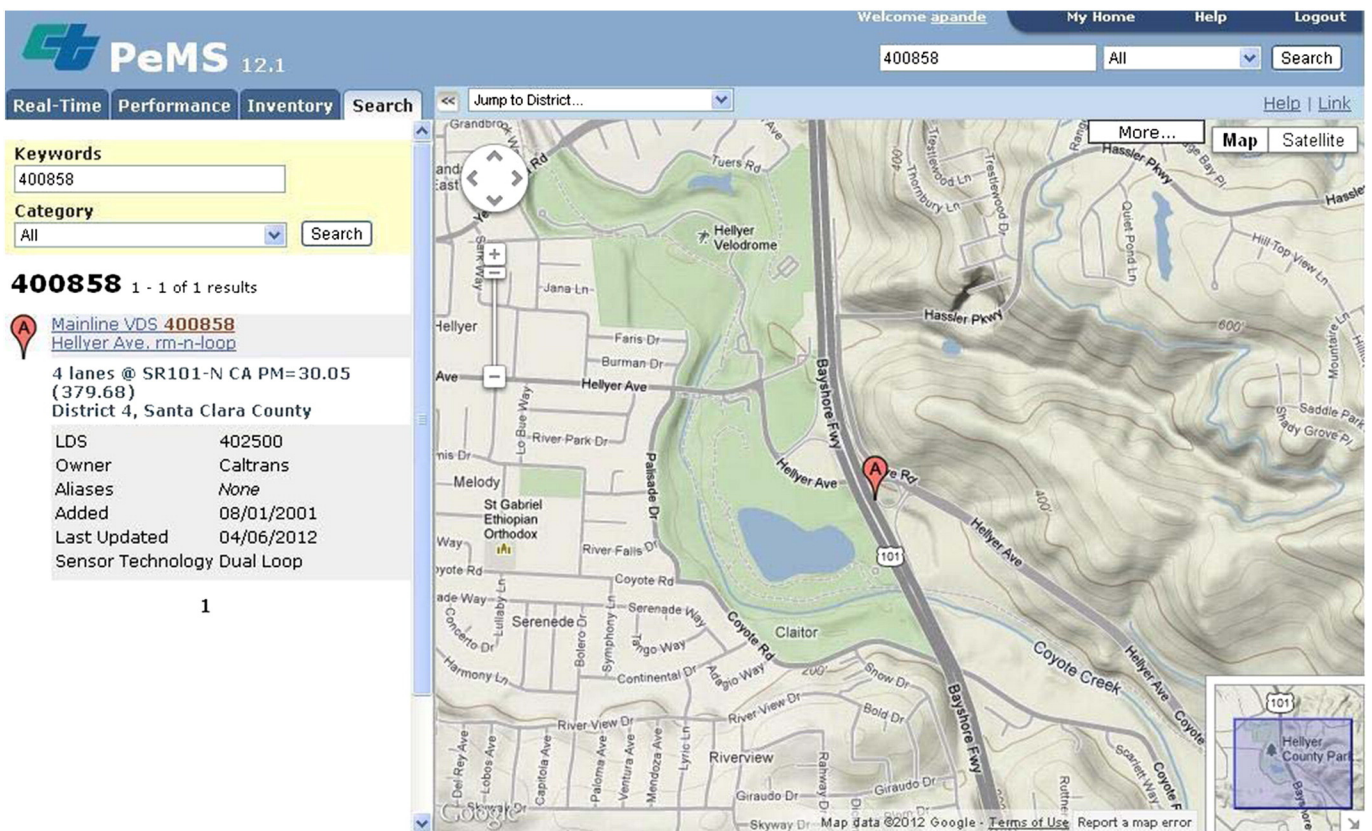**Fig. 5.** Location Map of VDS 401890 (High False Positives).

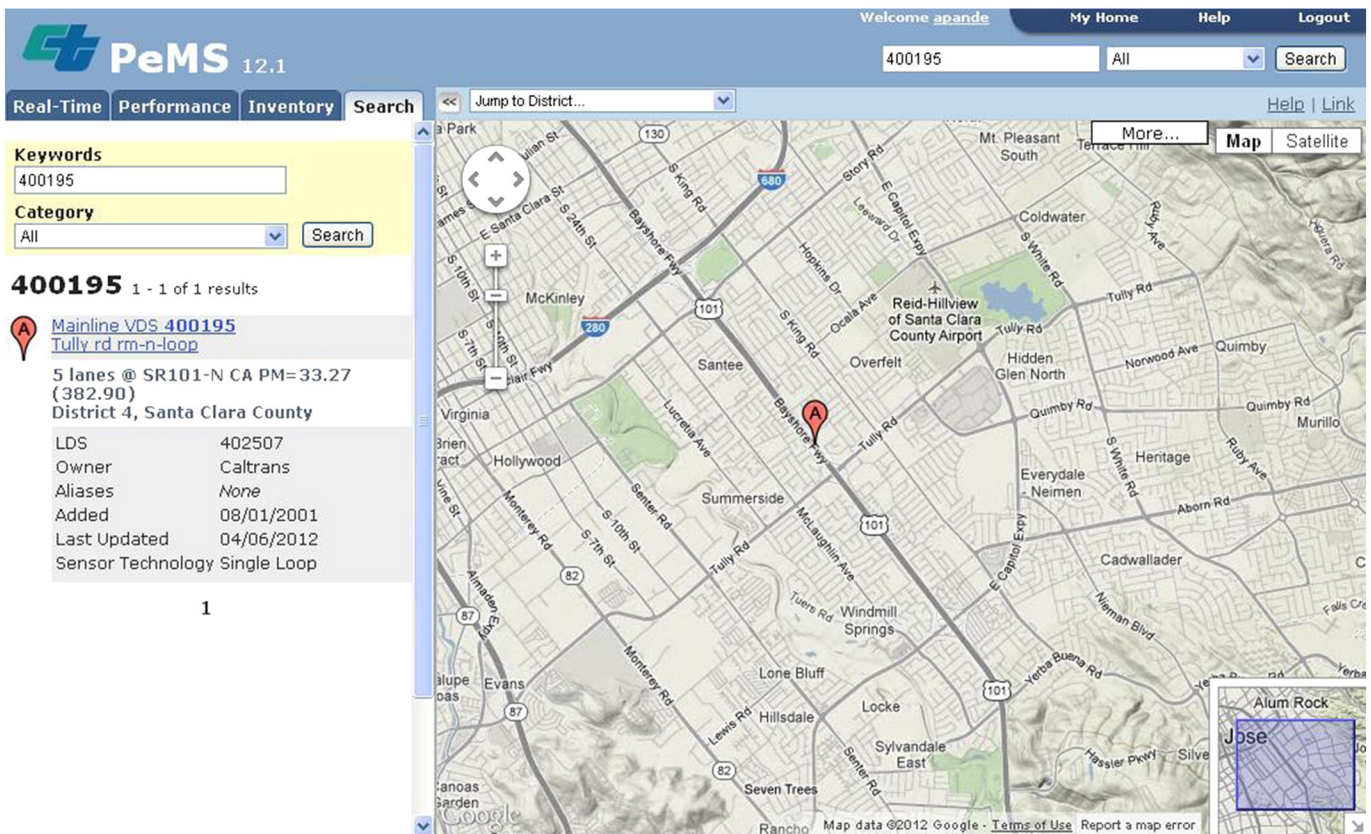**Fig. 6.** Location Map of VDS 400858 (High False Negatives).

Fig. 7. Location Map of VDS 400195 (High False Negatives).

VDS 400858 and 400195: Higher percentage of false negatives: Figs. 6 and 7 show that these locations are on long, straight US-101 NB segments, where other factors (driver errors at high speed) are likely to be responsible for more crashes.

It is worth mentioning that while results from all freeways demonstrated these trends; the trends from the other freeways mirror US-101 NB results to the degree of how well the original predictive US-101 NB model fit the other data. For example, I-880 NB was closest to the US-101 NB in terms of crash identification and hence the trends on I-880 matched most closely to the US-101 NB trends.

6. Conclusions

Research in the area of real-time crash risk estimation on uninterrupted flow facilities have been enhanced in the past 10 years with sophisticated classification algorithms applied to a variety of available traffic surveillance data. The authors identified that the literature had not yet conclusively answered the question of transferability of a crash risk estimation model from one freeway to another. The other question that remained unanswered was whether the misclassifications from crash risk estimation models are concentrated on certain situations of time of day/day of week or locations.

Answering these questions is important since the uninterrupted flow facilities from the same region tend to have similar types of data collection infrastructure. Also, the objective of research conducted in Japan (Hossain & Muromachi, 2010), was to identify the setup of traffic surveillance infrastructure yet to be put in place. The conclusion from this study on the transferability of the same model can be beneficial to freeways where such infrastructure is either being currently or have been recently put in place. The crash risk on such sections can be estimated from a transferable model from the freeways nearby. It is also worth noting that while models for most

locations may be transferable from one freeway to the other, some locations on the same freeway may require additional training for crash risk estimation (e.g., the US-101 NB section near the I-680/I-280 interchange). This study provided a framework to flag these locations for additional model training, through analysis of "false positives" and "false negatives" by locations. On a system of freeways, these locations with higher "false positives" or "false negatives" may be combined together from different facilities by not restricting the freeway crash risk estimation model by the corridor.

Acknowledgments

The research was sponsored by Mineta Transportation Institute at San Jose State University and their support is thankfully acknowledged. The authors would also like to acknowledge the use of PeMS (Performance Measurement System) database in the conduct of this research. Dr. Alexander Skabardonis of the Institute of Transportation Studies (ITS) at the University of California, Berkley was very helpful in providing access to the data.

References

Abdel-Aty, M., Pande, A., Das, A., & Knibbe, W. J. (2008). Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems. *Transportation Research Record: Journal of the Transportation Research Board, 2083*, 153–161.
Breiman, L. (1984). *Classification and regression trees.* : Chapman & Hall/CRC.
Christodoulou, C., & Georgiopoulos, M. (2000). *Applications of neural networks in electromagnetics.* Artech House, Inc. (Retrieved from http://dl.acm.org/citation.cfm?id = 557934).
Christoforou, Z., Cohen, S., & Karlaftis, M. G. (2011). Identifying crash type propensity using real-time traffic data on freeways. *Journal of Safety Research, 42*(1), 43–50.
Golob, T. F., & Recker, W. W. (2004). A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A: Policy and Practice, 38*(1), 53–80.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: The MIT press.

Hossain, M., & Muromachi, Y. (2010). Evaluating Location of Placement and Spacing of Detectors for Real-Time Crash Prediction on Urban Expressways. *Presented at the 89th Annual Meeting of the Transportation Research Board, Washington D.C.*

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.

Oh, C., Oh, J. S., Ritchie, S. G., & Chang, M. (2001). Real-time estimation of freeway accident likelihood. *Presented at the 80th Annual Meeting of Transportation Research Board, Washington D.C.*

Pande, A., & Abdel-Aty, M. (2006a). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis and Prevention, 38*(5), 936–948. http://dx.doi.org/10.1016/j.aap.2006.03.004.

Pande, A., & Abdel-Aty, M. (2006b). Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board, 1953*, 31–40.

Pande, A., & Abdel-Aty, M. (2008). A computing approach using probabilistic neural networks for instantaneous appraisal of rear-end crash risk. *Computer-Aided Civil and Infrastructure Engineering, 23*(7), 549–559.

Pande, A., Das, A., Abdel-Aty, M., & Hassan, H. (2011). Estimation of Real-Time Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board, 2237*, 60–66.

Pande, A., Nuworsoo, C., & Shew, C. (2012). *Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways (No. 11-15).* San Jose, CA: Mineta Transportation Institute (Retrieved from http://transweb.sjsu.edu/project/1006.html).

SAS Institute (2011). *Getting started with Enterprise Miner 7.1.* Cary, NC: SAS Institute (Retrieved from http://books.google.com/books?hl = en&lr = &id = j_ERl5lKx-kC&oi = fnd&pg = PA1&dq = Getting + Started + with + Enterprise + Miner + Software,+Release + 4.1,+&ots = xHUGWft4WC&sig = j93Htmz7P-IIQ-jD1valVNJ63DQ).

Varaiya, P. P. (2007). *Freeway performance measurement system (PeMS), PeMS 7.0.* Berkeley, CA: California Center for Innovative Transportation, Institute of Transportation Studies, University of California, Berkeley (Retrieved from http://www.its.berkeley.edu/publications/UCB/2007/CWP/UCB-ITS-CWP-2007-11.pdf).

Wilamowski, B. M., Iplikci, S., Kaynak, O., & Efe, M. O. (2001). An algorithm for fast convergence in training neural networks. *In Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on, Vol. 3.* (pp. 1778–1782) (Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber = 938431).

Xu, C., Liu, P., Wang, W., & Yu, C. (2011). Exploration and Identification of Hazardous Traffic Flow States before Crash Occurrences on Freeways. *Transportation Research Board 90th Annual Meeting* (Retrieved from http://trid.trb.org/view.aspx?id = 1092773).

Yu, R., Abdel-Aty, M., & Ahmed, M. (2013). Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis & Prevention, 50*, 371–376.

Cameron Shew is a recent graduate of California Polytechnic State University, San Luis Obispo, having earned a Master of Science in Civil and Environmental Engineering. He is a certified engineer-in-training (EIT) and currently works for a transportation consulting firm in the Sacramento, CA area. His research and work experiences have covered entry-level transportation planning, geometric design, and traffic operations tasks, and left him with strong interests in transportation design, and traffic simulation and safety.

Anurag Pande, Ph.D. is Assistant Professor of Civil Engineering at California Polytechnic State University, San Luis Obispo. Dr. Pande has significant experience in collection and analysis of traffic data relevant to measurement of safety. His research interests also include driver behavior, traffic simulation, emergency evacuation, and scholarship of teaching. He received the "Young Researcher Award" by the Transportation Research Board's Committee on Safety Data, Analysis, and Evaluation in 2007. He has co-authored close to 30 manuscripts that have been either published or are forthcoming in peer reviewed journals such as AAP, TRR, and IEEE transactions on ITS. He is currently working on several sponsored projects including a study on driver behavior funded by the National Science Foundation.

Cornelius Nuworsoo, Ph.D. is Associate Professor of Transportation Planning at the California Polytechnic State University, San Luis Obispo. His areas of interest include transportation planning, bicycle and pedestrian related issues, and traffic safety.