

A Simulation Study of Predicting Conflict-prone Traffic Conditions in Real-time

Christos Katrakazas

PhD Student

School of Civil and Building Engineering

Loughborough University

Loughborough LE11 3TU

United Kingdom

Tel: +44 (0)7789656641

Professor Mohammed Quddus*

School of Civil and Building Engineering

Loughborough University

Loughborough LE11 3TU

United Kingdom

Tel: +44 (0) 1509228545

Email: m.a.quddus@lboro.ac.uk

Professor Wen-Hua Chen

School of Aeronautical and Automotive Engineering

Loughborough University

Loughborough LE11 3TU

Tel: +44 (0)1509 227 230

Email: w.chen@lboro.ac.uk

*Corresponding author

Word Count: 6,979 words + 1 Table*250+ 1* Figure=7,479

Paper submitted for presentation at the 96th Annual Meeting of Transportation Research Board (TRB),
Washington D.C., USA

ABSTRACT

Current approaches to estimate the probability of a traffic collision occurring in real-time primarily depend on comparing the traffic conditions just prior to collisions with the traffic conditions during normal operations. Most studies acquire pre-collision traffic conditions by matching the collision time in the national crash database with the time in the aggregated traffic database. Since the reported collision time sometimes differs from the actual time, the matching method may result in traffic conditions not representative of pre-collision traffic dynamics. This may subsequently lead to an incorrect calibration of the model used to predict the probability of a collision. In this study, this is overcome through the use of highly disaggregated vehicle-based traffic data (i.e. vehicle trajectories) from a traffic micro-simulation (i.e. VISSIM) and the corresponding traffic conflicts (i.e. dangerous concurrences between vehicles) data generated by the Surrogate Safety Assessment Model (SSAM). In particular, the idea is to use traffic conflicts as surrogate measures of traffic safety, and data on traffic collisions are therefore not needed. Two classifiers are then employed to examine the proposed idea: (i) Support Vector Machines (SVMs) – a sophisticated classifier and (ii) k -Nearest Neighbors (k NN) – a relatively simple classifier. Substantial efforts are devoted to making the traffic simulation as representative to real-world as possible by employing data from a motorway section in England. Four temporally aggregated traffic datasets (i.e. 30-second, 1-minute, 3-minute and 5-minute) are examined. The main results demonstrate the viability of using traffic micro-simulation along with the SSAM for real-time conflicts prediction and the superiority of 3-minute temporal aggregation in the classification results. Attention should be however given to the calibration and validation of the simulation software so as to acquire more realistic traffic data resulting in more effective conflicts prediction.

Keywords: *Traffic safety, Traffic conflicts, Traffic micro-simulation, Support Vector Machines (SVMs), k -Nearest Neighbours (k -NN).*

1 INTRODUCTION

2 Over the past decades, the estimation of *unsafe* traffic conditions in real-time has been studied
3 by many researchers working in the area of Intelligent Transport Systems (ITS). This is because
4 predicting hazardous traffic conditions is an integral part of proactive highway safety management
5 that has the potential to reduce road traffic fatalities and injuries. In particular, predicting where and
6 when a traffic collision is likely to occur in real-time and preventing the collision by adjusting the
7 traffic dynamics through a range of traffic management interventions (e.g. variable message signs) are
8 beneficial to highway safety. Previous research on this topic has established an underpinning theory
9 suggesting that there exists a relationship between specific traffic dynamics (e.g. interactions between
10 speed, flow, congestion) and spatio-temporal collision risk (1). Based on this principle, a dominating
11 approach for detecting *unsafe* traffic conditions is the comparison of traffic situations just prior to
12 traffic collision occurrences on a segment with the traffic conditions at normal situations on the same
13 segment. Specifically, in the current era where various advanced driver assistance systems (2) and
14 autonomous vehicles (3) are massively developed, it becomes essential to effectively identify these
15 traffic fluctuations in real-time and enhance collision-free decision making of such technologies.

16 Perhaps the most important factor in the development of real-time collision-prone traffic
17 conditions models is the temporal aggregation of traffic data and the selection of important variables
18 which would lead to the correct distinction between collision-prone and normal traffic conditions.
19 Temporal aggregation of traffic data is available at pre-defined time intervals (e.g. 30-second or 1-
20 minute, 5-minute and 15-minute). Highly disaggregated traffic data (e.g. 30-second or 1-minute of
21 temporal aggregation) may not be suitable for implementing a timely intervention by the relevant
22 authorities to intervene and prevent both the collision and the collision-related congestion. In the
23 majority of recent studies (4–6), traffic conditions at 5-10 minutes before the collision have been
24 found to be the most suitable time period to identify such events in a timely manner and initiate an
25 intervention by the responsible traffic agencies. Highly disaggregated traffic data may not be available
26 in many countries. Furthermore, even if highly disaggregated traffic data are available, an error exists
27 between the reported collision time and the actual time of a collision. This is because the reported
28 time and location largely depend on the subjective volition of the police officers attending the site of
29 the collision (7). As a result, inaccurately reported collision time leads to misrepresentative pre-
30 collision traffic dynamics resulting in an inaccurate calibration of the collision prediction models.

31 Traffic micro-simulation can be utilised as a powerful tool to overcome the inherent
32 difficulties with the recorded collision time and temporal data aggregation issues. Recent research on
33 traffic micro-simulation and road safety (e.g. (8, 9)) showed that it is now possible to estimate
34 surrogate measures of safety performance based on dangerous vehicle interactions. If these risky
35 vehicle interactions are filtered with established risk indicating thresholds, they are termed as “traffic
36 conflicts”. According to the definition by Amundsen and Hyden (10) traffic conflicts occur when two
37 or more road vehicles are in such a collision course that a high probability of a collision exists if their
38 motion remains uninterrupted.

39 Using traffic conflicts can, therefore, address the issues related to traffic collisions as
40 discussed above. Furthermore, studying conflicts can enhance the understanding of how road users
41 fail to drive safely and cause collisions (8). Approaches that use traffic conflicts are also criticised in
42 the literature because the correlation between traffic conflicts and traffic collisions on a segment may
43 be low (9). It is however admitted that the mechanism that triggers collisions and conflicts is
44 analogous (8, 9).

45 Additionally, because of the technological advances in the area of automated driving, the
46 concept of real-time collision prediction should not necessarily relate to a timely intervention from
47 traffic authorities but rather should concentrate on improving the speed of the prediction as well as
48 their implementation at the vehicle-level. Therefore, the exploitation of highly disaggregated traffic
49 data should be taken into account. In that direction machine learning and data mining approaches can
50 prove advantageous over traditional (e.g. logistic regression (11)) or sophisticated techniques (e.g.
51 Neural Networks(12)) for real-time collision prediction. Since collisions and conflicts are more rare
52 events than normal traffic conditions, attention should also be given to the handling of imbalanced
53 data (13) (e.g. data used for the classification where one class has significantly more instances than
54 the other).

The combination of traffic micro-simulation and machine learning classifiers to detect conflict-prone traffic conditions on motorways from highly disaggregated data form the motivation for the current paper. This study explores the application of the commonly employed Support Vector Machines (SVM) and k -Nearest Neighbours (k NN), which is a simple but an effective non-parametric classifier, for the classification of simulated traffic data with regards to traffic conflicts. The traffic data used in this study come from the PTV VISSIM micro-simulation software (14) and consist of speed, flow and acceleration data aggregated at different temporal units (e.g. 30-second, 1-minute, 3-minute and 5-minute time intervals) to compare the effectiveness of the temporal aggregation on the classification results. The conflict data are acquired through the Surrogate Safety Assessment Model (SSAM) (15), a software which uses the trajectories of the vehicles from the traffic micro-simulation and outputs traffic conflicts. A matched-case control data structure is used, in which traffic conditions before each conflict are matched with normal traffic conditions coming from three other simulation runs. The number of additional runs was chosen in order to cope with the imbalance between conflict and safe conditions which can prove essential for classification purposes (13).

The rest of the paper is organised as follows: firstly, the existing literature and its main findings are synthesised. An analytic description of the k NN and SVM classification algorithms is described next. This is followed by a presentation of the data used in the analysis along with the pre-processing methodology and the results of the classification algorithms. Finally, the last section summarises the main conclusions of the study and offers some recommendations for future research.

LITERATURE REVIEW

The purpose of this review was to synthesize existing studies on traffic conflicts-based safety assessment by comparing and contrasting their findings and identify whether there is any important or interesting knowledge gap. Focus was also given on the methods employed in real-time collision prediction algorithms so as to select the appropriate methods for predicting conflict-prone traffic conditions.

Safety assessment using simulated conflicts

The use of traffic conflicts in road safety assessment using traffic micro-simulation has gained popularity within the ITS research community over the recent years. For example, Minderhoud and Bovy (16) suggested that traffic micro-simulation can overcome the need to collect collision data and provide alternatives to the safety evaluation of ITS technologies.

In a traffic micro-simulation tool, simulating traffic collisions may not be possible because all micro-simulation software is programmed according to the car-following models such that vehicles cannot collide. However, Huguenin et al. (17) indicated the fact that vehicles can come very close to each other and the information on vehicles' exact positions, speeds, headings and accelerations can provide a relevant safety index for vehicle interactions.

In order to identify metrics which can help in identifying conflicts in a traffic micro-simulation environment, Gettman et al. (18) suggested to use a range of variables as a traffic conflict indicator. These include Time-to-Collision (TTC), Post-Encroachment-Time (PET), the maximum speed of the vehicles, the deceleration rate and the speed differential between the vehicles. Their work led to a profusion of studies investigating the safety of vehicles in traffic micro-simulation models. Likewise, El-Basyouny and Sayed (8) indicated that traffic conflicts can also be employed as an applicable predictor for traffic collisions instead of typical measures such as exposure because they are based on vehicle interactions. This is further justified by Archer (19) who stated that the traffic conflict technique based on the results from micro-simulation could have a practical impact and provide an insight on the identification of safety problems in real-world traffic environments. A detailed overview of approaches concerning safety-related traffic simulation is given by Young et al. (20). In their review, it was revealed that there exists a correlation between the number of simulated conflicts with the number of expected real-world collisions. On the other hand, it is observed that this correlation is used for before-and-after analyses of intersection or motorway sections in order to identify which interventions can improve safety at these specific sites.

For instance, a recent study from Shahdah et al. (9) used VISSIM with SSAM to develop a statistical relationship between conflicts and collisions for signalised intersections. Traffic conflicts

were estimated by using two thresholds for TTC (i.e. 1.5 and 0.5 seconds). Their results concluded that conflict-based surrogate safety measures can be used to identify the number of collisions at a particular site as well as for a before-and-after safety evaluation of crash modification factors. Essa and Sayed (21) however emphasised that the link between conflicts and collisions depends heavily on the calibration of the simulation model. In the same principle, Fan et al. (22), who investigated the safety of motorway merging areas, suggested that SSAM should be used with caution because of the purely stochastic nature of real-world collisions.

Machines Learning Classifiers

In order to reliably identify conflict-prone traffic conditions, potential classifiers need to be fast, accurate and suitable for real-time applications. Classifiers used to detect real-time traffic collisions can also be applied for this purpose. Most popular approaches in real-time collision detection include logistic regression (e.g. 4, 30) and Neural Networks (e.g. 31, 32). However, logistic regression models rely on distribution assumptions for the collision frequency and the corresponding collision precursors and Neural Networks heavily incorporate the black-box effect and present over-fitting problems (27). Bayesian Networks (4) and Genetic Programming (28) approaches to real-time collision prediction deal with the aforementioned problems but face difficulties with regards to their transferability and practical implementation.

As a result, alternative methods should be sought to overcome the existing methodological drawbacks. According to Dreisettl and Ohno-Machado (29) SVMs are flexible and have less over-fitting problems while *kNN* provides a case-based explanation on classification results, address the black-box problem and are easily transferrable because they do not require prior knowledge of any datasets. The recent work on SVMs proves that they are an efficient classifier as well as a successful predictor when applied to traffic collisions prediction. Hence, it is a potential candidate for detecting conflict-prone conditions effectively. Moreover, the simplicity of *kNN* and its real-time applicability as suggested by studies on real-time traffic prediction provides an alternative algorithm that can be used for classifying traffic conditions.

SVM models have been applied to real-time collision and traffic flow predictions. For instance, Li et al. (30) compared the findings from the SVMs with the findings of the popular negative binomial models in predicting motorway collisions. Their results showed that SVM models have a better goodness-of-fit in comparison with negative binomial models. Their findings were in line with the study of Yu & Abdel-Aty (27) who compared the results from the SVM and Bayesian logistic regression models for evaluating real-time collision risk demonstrating the better goodness-of-fit of the SVM models. The prediction of side-swipe accidents using SVMs was evaluated in Qu et al. (31) by comparing SVMs with Multilayer Perceptron Neural Networks. Both techniques showed similar accuracy but SVMs led to better collision identification at higher false alarm rates. More recently, Dong et al. (32) demonstrated the capability of SVMs to assess spatial proximity effects for regional collision prediction.

On the other hand, *kNN* has recently been applied in the area of short-term traffic prediction due to the fact that it is one of the simplest data mining algorithms. Zhang et al. [34] made a first attempt to use *kNN* for traffic flow prediction using occupancy rate, vehicle speed and weather data. They compared the results from *kNN* with the results of backpropagation Neural Networks and showed that *kNN* classification was more accurate and transferable. Furthermore, Xiaoyu et al. (33) argued that although *kNN* has a relatively slow computing speed it is suitable for real-time applications. Lastly, in comparison with SVM, *kNN* have better transferability as suggested by Zhang et al (34).

Another issue that needs further investigation by the application of classification algorithms is the temporal aggregation of traffic data. Previous work on segment-based collision prediction indicated that 5-10-minute aggregated traffic data (e.g. (5)) offer an ideal balance between capturing the microscopic traffic fluctuations and enabling sufficient time to traffic authorities for introducing interventions. Such temporal aggregation may not be optimal for the case of (semi)autonomous vehicles which need a reliable prediction of unsafe traffic conditions as fast as possible. Therefore, different temporal aggregation intervals (i.e. 30-second, 1-minute, 3-minute and 5-minute) will be tested in order to identify the aggregation offering the best results in real-time conflict-prone traffic conditions estimation.

In summary, it can be concluded that data from a traffic micro-simulation tool (e.g. VISSIM) and relevant traffic conflicts from the SSAM have the potential to improve real-time highway safety assessment. Although vehicles in micro-simulation do not collide, they have abundant interactions with each other and their motions are realistic because of the built-in car-following models. Consequently, if proper attention to the correct calibration of the micro-simulation model is given, traffic conditions before a traffic conflict can be used as a surrogate measurement to identify traffic collisions. However, existing studies utilising VISSIM/SSAM concentrate on the investigation of the correlation between traffic collisions and traffic conflicts so as to evaluate the impact of interventions through the use of traffic conflicts. In this paper, simulated traffic conditions and the corresponding conflicts data are utilised to estimate conflict-prone traffic conditions in real-time by the use of machine learning classifiers (i.e. SVMs and k NN).

SUPPORT VECTOR MACHINES AND K-NEAREST NEIGHBOURS

The objective of this study is to identify conflict-prone traffic conditions from highly disaggregated data by using the SVM and k NN classifiers. It should be noted that this study is concerned with the binary classification problem of distinguishing between “safe” and “conflict-prone” traffic conditions, although in the real-world a road segment may have varying degrees of safety.

SVMs belong to the larger group of supervised learning algorithms and kernel methods. In supervised learning, there exists a set of example input vectors $\{x_n\}_{n=1}^N$ along with corresponding targets $\{t_n\}_{n=1}^N$, the latter of which corresponds to class labels. In this study, the two classes are defined as ‘*dangerous*’ when $t=1$ and ‘*safe*’ when $t=0$. The purpose of learning is to acquire a model of how the targets rely on the inputs and use this model to classify or predict accurately future and previously unseen values of x .

An SVM classifier is based on the following functional form:

$$y = f(x; w) = \sum_{i=1}^N w_i K(x, x_i) + w_0 = w^T \varphi(x) \quad (1)$$

In equation (1), $K(x, x_i)$ is a kernel function, which defines a basis function for each data point in the training dataset, w_i are the weights (or adjustable parameters) for each point, and w_0 is the constant parameter. The output of the function is a sum of M basis functions $\varphi(x) = [\varphi_1(x), \varphi_2(x), \dots, \varphi_M(x)]$ which is linearly weighted by the parameters w .

SVM, through its target function, tries to find a separating hyper-plane to minimize the error of misclassification while at the same time maximize the distance between the two classes (27). The produced model is sparse and relies only on the kernel functions associated with the training data points which lie either on the margin or on the wrong side. These data points are referred to as “Support Vectors” (SVs).

k NN is a non-parametric learning algorithm which is simple but effective in many cases (35). For a data record t to be classified its k nearest neighbours are retrieved and this forms a neighbourhood of t . During training, each t is assigned to a class if the majority of the k neighbours of t belong to this particular class. However, an appropriate value for k is needed to apply a k NN approach and the success of classification is very much dependent on this value (36).

DATA DESCRIPTION AND PROCESSING

This study aims to examine the effectiveness of SVM and k NN classifiers in identifying conflict-prone traffic conditions using data from a traffic micro-simulation (i.e. VISSIM) and the SSAM. As discussed in the literature review section, the fundamental issue relating to this approach is the building and calibrating of a traffic micro-simulation model using real-world traffic data. For this purpose, link-level disaggregated traffic data from loop detectors and GPS-based probe vehicles were obtained from the UK Highways England Journey Time Database (JTDB). Link-level data correspond to every day of the years 2012 and 2013 and include average travel speed, volume and average journey time at 15-minute intervals. It should be noted here that 15-minute traffic data correspond to link-based average speed, volume and journey time of all vehicles between two junctions.

A 4.52-km section of motorway M62 between junctions 25 and 26 in England was selected as the study area. In order to build a robust micro-simulation model, the JTDB data were split into four scenarios for each year:

- Morning peak hours (06:00 – 09:30) = 14 “15-minute” intervals
- Morning off-peak hours (09:30-13:00) = 14 “15-minute” intervals
- Afternoon off-peak hours (13:00-15:45) = 11 “15-minute” intervals
- Afternoon peak hours (15:45-19:15) = 14 “15-minute” intervals

For each of these scenarios the 15-minute traffic volumes and the cumulative speed distribution of the roadway segment were extracted and employed as input to VISSIM. Furthermore, the vehicle composition for 2012 and 2013 was also obtained from the UK Department of Transport (37) and was used to build a micro-simulation model. The road segment was manually coded in VISSIM using a background image from OpenStreetMap (38). It was decided to allocate data collection detectors every 300m in order to acquire detailed traffic data. The spacing of the detectors was inspired by previous studies on real-time collision prediction on motorways (e.g. (4, 27))

In order for the micro-simulation to be initiated, the car-following model needs to be defined in VISSIM. According to the software manual (14), the Wiedemann 99 model was selected because it applies to motorway scenarios. The Wiedemann model is characterised mainly by three parameters in VISSIM; the standstill distance, the headway time and the following variation (14). The standstill distance describes the average standstill distance between two vehicles. The headway time is the time gap (in seconds) which a driver wants to maintain at a certain speed. On the other hand, the following variation defines the desired safety distance a driver allows before moving closer to a car in front.

In order to validate the simulation results the travel time and the GEH-statistic (39), which correlates the observed traffic volumes with the simulated volumes, was used as shown below:

$$GEH = \sqrt{\frac{(V_{sim} - V_{obs})^2}{\frac{V_{sim} + V_{obs}}{2}}} \quad (2)$$

where V_{sim} is the simulated traffic volume and V_{obs} is the observed traffic volume.

After a number of trial simulations (~1000 for every scenario) the best GEH values came by using the following parameters for the Wiedemann 99 car following model:

- Standstill distance: 1.5 m
- Headway time: 0.9 sec
- Following variation: 4 m

For the simulation to efficiently resemble real-world traffic it is essential that the GEH statistic takes a value below five for 85% of the simulated intervals (39). In the simulations that were undertaken, the GEH values for the majority of the time intervals were found to be less than five. However, there were 3 out of 14 time intervals from the scenarios of the morning peak hours of the year 2012 where GEH values were found to be between 5 and 10. According to (40) these values indicate either a calibration problem or a data problem. Because of the large number of simulations undertaken it is assumed that the bad GEH values relate to the bad quality of the available data (i.e. 15-minutes aggregated road-level traffic data). Therefore it was decided to keep the simulation results for the corresponding morning peak time scenario of the year 2012.

After calibrating the simulations for every scenario by year, three additional simulations with different random seeds were run and validated with the GEH values resulting in a total of four different simulation results for each of the scenarios. The number of additional runs was chosen in order to cope with the imbalance between conflict and safe conditions which can prove essential for classification purposes (13). The four different simulations are used for the matched-case control structure, where the first simulation is used to acquire the traffic conflicts and the other three are used to resemble the normal traffic conditions.

For the extraction of traffic conflicts, the vehicle trajectory files exported from VISSIM were inserted to the SSAM. Conflicts were detected if the TTC value between two vehicles was below 1.5 seconds and the PET value was below 4 seconds which are the default values used in SSAM (15). In the last step of the data processing, a MATLAB (41) code was developed in order to match the conflicts (exported from the SSAM) with the traffic conditions (acquired from VISSIM). The estimated conflicts were filtered again in order to obtain conflicts with TTC below 1.3 seconds and

PET below 1 second in order to identify conflicts which are difficult to avoid. That is because TTC below 1.3 seconds is lower than the average human reaction time (42) and PET values close to zero show imminent collisions (15).

For every conflict, the nearest upstream detector on the road segment was identified by comparing the time of the conflict with the time the vehicles passed from every detector. This specific detector was marked as “conflict detector”. Traffic data were extracted for every conflict detector, the corresponding upstream and downstream detectors on the same lane and the detector in the adjacent lane for every time interval. The traffic measurements for these detectors were marked as “conflicts” because they represent the traffic conditions near the time when the conflict occurred. For each of the detectors and for every time interval the average number of vehicles, the average vehicle speed and the average vehicle acceleration were extracted. The traffic data exported from VISSIM were then aggregated in 30-second, 1-minute, 3-minute and 5-minute intervals prior to the conflict occurrence.

As mentioned before, three additional simulation runs were performed in order to acquire conflict-free traffic conditions for the detectors. Traffic conditions collected at the same conflict detectors in the three additional simulation runs were marked as “safe”, after checking that no conflicts happened on the same detectors during these additional runs. A total of 3,513 traffic conflicts and the corresponding conflict-prone traffic conditions were gathered for further analysis. Figure 1 illustrates the procedure which was followed to prepare the data for the classification.

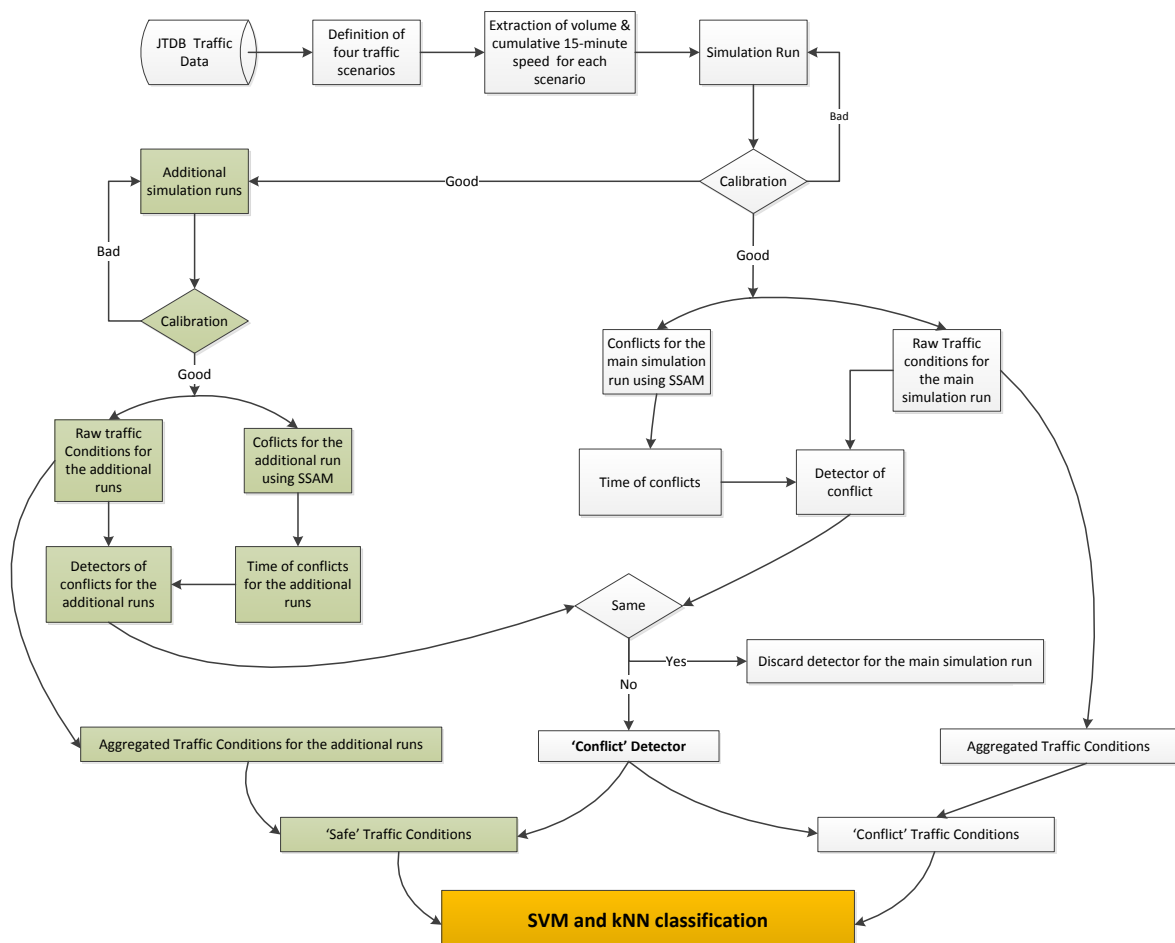


Figure 1: Flow chart of the procedure followed to perform the classification

In order to validate the classification results, a ten-fold cross-validation procedure was used (43). The original sample was randomly divided into ten equal sub-samples and of those ten sub-

samples, one sub-sample is chosen as the validation dataset for testing the classification accuracy and the rest of the sub-samples were used as the training data.

RESULTS AND DISCUSSION

Classification methods of SVM and k NN have been applied to a unified dataset containing all the cases (“*conflicts*” and “*safe*”) as discussed above.

SVMs depend on the kernel functions to perform the classification. The most popular kernels used in the SVM classification are the linear, polynomial and Gaussian or radial basis function (RBF). In this study, the Gaussian kernel has been used because existing research suggests that it provides more accurate results (27). The Gaussian kernel is calculated through the equation:

$$K(x_i, x_j) = \exp \left(-\gamma \|x_i - x_j\|^2 \right) \quad (2)$$

where γ determines the width of the basis function. The coefficient γ was set to 0.5 because the targets of the classification lie in the interval $\{0,1\}$.

The k NN classifier, on the other hand, requires tuning the important parameter - the number of nearest neighbours (k). A usual approach is to perform tests with different k values starting from 1 and ending at the square root of the number of observations (44). In this paper after trying different k values, the best results came from using $k=40$ and $k=119$ (which is the square root of the sample size). The results for those two k NN runs are therefore presented and compared with the SVM classification.

Both SVM and k NN were developed using the Statistics and Machine Learning Toolbox of MATLAB. In order to test the performance of the three different algorithms (i.e. SVM, 40-NN and 119-NN) the classification accuracy was initially tested for each of the temporal aggregation intervals. To further investigate the performance of the classifiers for real-time conflict-prone traffic conditions identification as well as to cope with the imbalance of the dataset (because *conflict* to *safe* conditions ratio is 1:3) a number of metrics were employed to evaluate the performance of the classifiers. These metrics are sensitivity, specificity, precision, recall, G -means and F -measure and are defined in equations (3) - (8) according to (6):

$$\text{Sensitivity} = \frac{T_{\text{conflict}}}{T_{\text{conflict}} + F_{\text{safe}}} \quad (3)$$

$$\text{Specificity} = \frac{T_{\text{safe}}}{T_{\text{safe}} + F_{\text{conflict}}} \quad (4)$$

$$\text{Precision} = \frac{T_{\text{conflict}}}{T_{\text{conflict}} + F_{\text{conflict}}} \quad (5)$$

$$\text{Recall} = \frac{T_{\text{conflict}}}{T_{\text{conflict}} + F_{\text{safe}}} \quad (6)$$

$$G\text{-means} = \sqrt{\text{Sensitivity} * \text{Specificity}} \quad (7)$$

$$F\text{-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (8)$$

where T_{conflict} represents a correct detection of conflict-prone traffic conditions identified as *conflict-prone*, F_{conflict} represents an incorrect detection of conflict-prone traffic conditions identified as *safe*, T_{safe} is a safe traffic condition instance correctly identified as *safe* and F_{safe} is a safe traffic condition instance falsely identified as *conflict-prone*.

The sensitivity statistic shows the correct classification accuracy with respect to conflict-prone traffic conditions, while the specificity statistic shows the classification accuracy in terms of safe conditions. Precision and recall are used for identifying the classification accuracy among each class. G -means is used to check whether the use of an imbalance dataset (1:3; conflicts vs safe) has any negative impact on the balanced qualification accuracy. Lastly, the F -measure is a metric which resembles the conflict-prone classification ability of the classifier models.

Results for all the above mentioned performance metrics for the classifiers are summarised in Table 1.

**Table 1: Classification performance metrics for the temporal aggregation intervals
(Sample size 14,052 cases with 10-fold cross validation)**

Classifier	Accuracy			
	30-second	1-minute	3-minute	5-minute
SVM	76.4%	76.7%	78.1%	76.3%
40-NN	75.1%	75.4%	77.4%	75.9%
119-NN	75.0%	75.2%	77.0%	75.7%
	Sensitivity	Specificity	Precision	Recall
	30-second			
SVM	0.893	0.761	0.062	0.893
40-NN	0.542	0.752	0.015	0.542
119-NN	0.600	0.750	0.003	0.600
	1-minute			
SVM	0.875	0.765	0.080	0.875
40-NN	0.672	0.755	0.033	0.672
119-NN	0.692	0.752	0.015	0.692
	3-minute			
SVM	0.827	0.779	0.157	0.827
40-NN	0.758	0.775	0.140	0.758
119-NN	0.713	0.773	0.133	0.713
	5-minute			
SVM	0.835	0.762	0.067	0.835
40-NN	0.696	0.760	0.061	0.696
119-NN	0.807	0.756	0.037	0.807
	G-Means		F-Measure	
	30-secons			
SVM	0.825		0.116	
40-NN	0.638		0.029	
119-NN	0.671		0.005	
	1-minute			
SVM	0.818		0.146	
40-NN	0.713		0.063	
119-NN	0.722		0.030	
	3-minute			
SVM	0.803		0.264	
40-NN	0.766		0.236	
119-NN	0.742		0.224	
	5-minute			
SVM	0.797		0.125	
40-NN	0.727		0.113	
119-NN	0.782		0.071	

As can be seen from Table 1, SVM outperforms the three k NN classifiers with respect to the accuracy statistics regardless of the temporal aggregation of traffic data. This is an expected outcome because the SVM classification is a more sophisticated and powerful technique than the simplistic k NN. Surprisingly, the difference in the classification accuracy is relatively small (~1.0 - 1.5%) for the same temporal aggregation. Forty neighbours are the optimal number of neighbours which can be

used to classify an instance of traffic conditions as *conflict-prone* or *safe*. On the contrary, the 119-NN classifier falls short in classifying conflict-prone conditions at a level of 0.2% - 0.4% compared to the 40-NN classifier.

Unlike existing studies on estimating collision-prone traffic conditions where 5-10-minute aggregated traffic data provided the best performance (e.g. (6)), traffic data aggregated in 3-minute time interval have proved to be a better conflict precursor than any other temporal aggregation used in this study. This is probably related to the noise inherent to 30-second and 1-minute aggregated data and lack of 5-minute data to capture accurately the traffic dynamics leading to a conflict.

It can be observed that SVM demonstrates a higher sensitivity and specificity compared to k NN. This implies smaller Type I and Type II errors because both conflict-prone and safe conditions have a better chance of being correctly classified, especially when using 30-second traffic data. The performance of k NN classification regarding sensitivity and specificity improves with higher temporal aggregation reaching its best value when 5-minute traffic data are classified using the 119-NN classifier.

As far as precision and recall are concerned, it can be seen that the classifiers demonstrate high recall but low precision. Hence, the classifiers perform well in classifying traffic conditions, but most of the correct classifications correspond to safe traffic conditions (which form the majority of the sample) rather than conflict-prone conditions. The best precision (i.e. the best identification of conflict-prone conditions) was found for the case of 3-minute aggregated data.

According to the results presented in Table 1, the G -means metric shows that the balanced classification ability of the SVM classifier is higher in smaller temporal aggregation intervals but drops when 3-minute and 5-minute aggregation are utilised. On the other hand, the G -means metric improves with higher temporal aggregation in the case of 119-NN, while the 40-NN algorithm performs well when 3-minute temporal aggregation is used.

Finally, the F -measure results of the classifiers in Table 1 show that conflicts are difficult to be detected by all three algorithms. This is probably due to the class imbalance problem, as well as the noise included in lower temporal aggregation intervals. The 3-minute temporal aggregation interval again shows better results than the other two intervals.

The classification accuracy for both classifiers agree with the results in the literature (e.g. (6, 27, 46)) which used actual collision data and more precise traffic data. The results of the classifiers regarding accuracy, G -Means and F -measure are comparable to the findings by Sun and Sun (6) who employed a Dynamic Bayesian Network (DBN) classifier. Their DBN classifier achieved an overall accuracy of 76.6% which is similar to the accuracy of most of the conflict-based classifiers in this paper. It should also be observed that using 3-minute traffic data with SVMs, the accuracy performance increases to 78.1%. Furthermore, it is shown that even though the two k NN classifiers are considered to be a simple algorithm their performance is similar when using 3-minute traffic data and is low only by approximately 1% for other temporally aggregated traffic data examined in this study. Regarding G -means the DBN classifier in (6) has a value of 0.76 which is similar to the k NN classifiers in the current study but lower when compared to SVM. This shows the supremacy of SVMs for balanced classifying for the case of temporally aggregated data. On the contrary, DBNs are better in detecting collision-prone conditions than SVMs and k NN with an F -Measure value of 0.512. This shows that further research is needed to overcome the data imbalance for better detection of conflict-safe traffic conditions.

In summary, it can be concluded that that traffic data aggregated in a 3-minute interval have proved to be the best temporal aggregation in classifying conflict-prone traffic conditions. However, improvements regarding the data imbalance problem need to be made to improve the F -measure metric for the classifiers.

CONCLUSIONS

This paper developed a simulation approach to detect traffic conflict-prone traffic conditions in real-time. This approach overcame two issues associated with the classification of collision-prone traffic conditions: (i) the temporal traffic data aggregation problem and (ii) the issues surrounding the incorrect reporting of collision time and the corresponding misrepresentative pre-collision traffic conditions. Since real-world data on traffic conflicts were not available, a simulation method

consisting of two widely used simulation tools - VISSIM for traffic information and SSAM for traffic conflicts - was adopted. VISSIM provided aggregated traffic data and information on individual vehicles' trajectories that were fed to SSAM which was capable of converting microscopic traffic information into meaningful safety-related information such as traffic conflicts. Significant efforts were devoted to calibrating the traffic simulation model in VISSIM. The performance of the algorithms was evaluated using their overall accuracy and the metrics of sensitivity, specificity, precision, recall, G-means and F-measure.

The classification results showed that traffic micro-simulation along with safety thresholds to detect conflicts from the SSAM model could be used in real-time safety assessment. The accuracy of both the SVM and *k*NN classifiers was found to be in-line with recent studies on real-time collision prediction which used actual collision data along with the corresponding traffic data. Thus, having overcome the misreported collision time simulation-based data can better represent traffic conditions before the occurrence of a dangerous vehicle encounter. Since the mechanism leading to a conflict and the mechanism leading to collision present similarities, the correct real-time identification of conflict-prone conditions would lead to safer real-time traffic because collisions are a fraction of the observed conflicts. Moreover, the superiority of 3-minute temporal aggregation in the classification results indicates that safety experts should utilise 3-minute aggregated data to understand the traffic fluctuations and the occurrence of traffic collisions. Researchers should be cautious if highly disaggregated traffic data (i.e. 30-second) are utilised in estimating real-time conflicts for the risk assessment of advanced driver-assistance systems (ADAS) and autonomous vehicles (AVs) which need to collect and process data as fast as possible from their on-board sensors.

It should, however, be noted that if simulated data are used, special attention shall be given in the validation using real-world data (e.g. video surveillance data or radar-based data). Further research shall be devoted to solving the issue with the data imbalance as identified in this study by the low F-measure metric.

ACKNOWLEDGEMENT

This research was funded by a grant from the UK Engineering and Physical Sciences Research Council (EPSRC) (Grant reference: EP/J011525/1). The authors take full responsibility for the content of the paper and any errors or omissions. Research data for this paper are available on request from Mohammed Quddus (m.a.quddus@lboro.ac.uk).

REFERENCES

1. Roshandel, S., Z. Zheng, and S. Washington. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis and Prevention*, Vol. 79, 2015, pp. 198–211.
2. Hummel, T., M. Kühn, J. Bende, and A. Lang. *Advanced Driver Assistance Systems: An investigation of their potential safety benefits based on an analysis of insurance claims in Germany- German Insurance Association, Research report FS 03*. 2011.
3. Thrun, S. Toward Robotic Cars. *Communications of the ACM*, Vol. 53, No. 4 (April), 2010, pp. 99–106.
4. Hossain, M., and Y. Muromachi. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident; analysis and prevention*, Vol. 45, Mar. 2012, pp. 373–81.
5. Abdel-aty, M., and A. Pande. The viability of real-time prediction and prevention of traffic accidents. 2005.
6. Sun, J., and J. Sun. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies*, Vol. 54, 2015, pp. 176–186.
7. Imprialou, M.-I. *Developing accident-speed relationships using a new modelling approach*. Developing Accident-Speed Relationships Using a New Modelling Approach by. 2015.
8. El-Basyouny, K., and T. Sayed. Safety performance functions using traffic conflicts. *Safety Science*, Vol. 51, No. 1, 2013, pp. 160–164.

9. Shahdah, U., F. Saccomanno, and B. Persaud. Application of traffic microsimulation for evaluating safety performance of urban signalized intersections. *Transportation Research Part C: Emerging Technologies*, Vol. 60, 2015, pp. 96–104.
10. Amundsen, F., and C. Hyden. Proceeding of First Workshop on Traffic Conflicts.
11. Hassan, H. M., and M. a Abdel-Aty. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. *Journal of safety research*, Vol. 45, Jun. 2013, pp. 29–36.
12. Abdel-Aty, M., and A. Pande. Identifying crash propensity using specific traffic speed conditions. *Journal of safety research*, Vol. 36, No. 1, Jan. 2005, pp. 97–108.
13. He, H., and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, 2009, pp. 1263–1284.
14. PTV Planug Trasport Verker AG. PTV VISSIM 6 User Manual. 2013.
15. Pu, L., and R. Joshi. Surrogate Safety Assessment Model (SSAM): Software User Manual. No. May, 2008, p. 96.
16. Minderhoud, M. M., and P. H. Bovy. Extended time-to-collision measures for road traffic safety assessment. *Accident; analysis and prevention*, Vol. 33, No. 1, Jan. 2001, pp. 89–97.
17. Huguenin, F., A. Torday, and A. Dumont. Evaluation of traffic safety using microsimulation. 2005.
18. Gettman, D., and L. Head. Surrogate Safety Measures from Traffic Simulation Models. *Transportation Research Record: Journal of the Transportation Research Board*, 2003, pp. 104–115.
19. Archer, J. *Indicators for traffic safety assessment and prediction and their application in micro-simulation modelling: A study of urban and suburban intersections*. 2005.
20. Young, W., A. Sobhani, M. G. Lenn??, and M. Sarvi. Simulation of safety: A review of the state of the art in road safety simulation modelling. *Accident Analysis and Prevention*, Vol. 66, 2014, pp. 89–103.
21. Essa, M., and T. Sayed. Simulated Traffic Conflicts Do They Accurately Represent Field-Measured Conflicts? *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2514, 2015.
22. Fan, R., W. Wang, P. Liu, and H. Yu. Using VISSIM simulation model and Surrogate Safety Assessment Model for estimating field measured traffic conflicts at freeway merge areas. *IET Intelligent Transport Systems*, Vol. 7, No. 1, 2013, pp. 68–77.
23. Oh, C., J.-S. Oh, S. G. Ritchie, and M. Chang. *Real-Time Estimation of Freeway Accident Likelihood, Report UCI-ITS-TS-WP-00-8, Department of Civil and Environmental Engineering and Institute of Transportation Studies, University of California, Irvine*. 2001.
24. Abdel-Aty, M., N. Uddin, A. Pande, F. Abdalla, and L. Hsia. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record, Journal of the Transportation Research Board*, 2004.
25. Pande, A., and M. Abdel-Aty. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident; analysis and prevention*, Vol. 38, No. 5, Sep. 2006, pp. 936–48.
26. Pande, A., A. Das, M. A. Abdel-aty, and H. Hassan. Real-Time Crash Risk Estimation: Are All Freeways Created Equal? 2011, p. 2011.
27. Yu, R., and M. Abdel-Aty. Utilizing support vector machine in real-time crash risk evaluation. *Accident; analysis and prevention*, Vol. 51, Mar. 2013, pp. 252–9.
28. Xu, C., W. Wang, and P. Liu. A Genetic Programming Model for Real-Time Crash Prediction on Freeways. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 2, Jun. 2013, pp. 574–586.
29. Dreiseitl, S., and L. Ohno-Machado. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, Vol. 35, No. 5–6, 2002, pp. 352–359.
30. Li, X., D. Lord, Y. Zhang, and Y. Xie. Predicting motor vehicle crashes using Support Vector Machine models. *Accident Analysis and Prevention*, Vol. 40, No. 4, 2008, pp. 1611–1618.
31. Wang, W., X. Qu, W. Wang, and P. Liu. Real-time freeway sideswipe crash prediction by support vector machine. *IET Intelligent Transport Systems*, Vol. 7, No. 4, 2013, pp. 445–453.
32. Dong, N., H. Huang, and L. Zheng. Support vector machine in crash prediction at the level of

- traf fi c analysis zones : Assessing the spatial proximity effects. Vol. 82, 2015, pp. 192–198.
33. Xiaoyu, H., W. Yisheng, and H. Siyu. Short-term Traffic Flow Forecasting based on Two-tier K-nearest Neighbor Algorithm. *Procedia - Social and Behavioral Sciences*, Vol. 96, No. Cictp, 2013, pp. 2529–2536.
34. Zhang, L., Q. Liu, W. Yang, N. Wei, and D. Dong. An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction. *Procedia - Social and Behavioral Sciences*, Vol. 96, No. Cictp, 2013, pp. 653–662.
35. Hand, D., D. Hand, H. Mannila, H. Mannila, P. Smyth, and P. Smyth. *Principles of data mining*. The MIT Press, 2001.
36. Guo, G., H. Wang, D. Bell, Y. Bi, and K. Greer. KNN Model-Based Approach in Classification. 2003, pp. 986–996.
37. Transport, U. D. for. GB Road Traffic Counts - Datasets - DGU. <http://data.gov.uk/dataset/gb-road-traffic-counts>.
38. OpenStreetMap®. OpenStreetMap. <https://www.openstreetmap.org>.
39. Transport For London. *Traffic Modelling Guidelines TfL Traffic Manager and*. 2010.
40. Transportation, W. D. of. Model Calibration - Traffic Analysis and Microsimulation. http://www.wisdot.info/microsimulation/index.php?title=Model_Calibration.
41. Mathworks, MATLAB.
42. Triggs, T. J., and W. G. Harris. Reaction Time of Drivers to Road Stimuli. *Medicinski Pregled*, Vol. 62, No. June 1982, 1982, pp. 114–9.
43. Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. 2012.
44. Hassanat, A. B., M. A. Abbadi, and A. A. Alhasanat. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 12, No. 8, 2014, pp. 33–39.
45. Yu, R., and M. Abdel-Aty. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. *Accident; analysis and prevention*, Vol. 62, Jan. 2013, pp. 161–7.
46. Abdel-Aty, M., A. Pande, A. Das, and W. J. Knibbe. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. *Transportation Research Record*, Vol. 2083, No. 2083, 2008, pp. 153–161.