



## Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: A combined approach

Athanasios Theofilatos, George Yannis, Constantinos Antoniou, Antonis Chaziris & Dimitris Sermis

**To cite this article:** Athanasios Theofilatos, George Yannis, Constantinos Antoniou, Antonis Chaziris & Dimitris Sermis (2018) Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: A combined approach, *Journal of Transportation Safety & Security*, 10:5, 471-490, DOI: [10.1080/19439962.2017.1301611](https://doi.org/10.1080/19439962.2017.1301611)

**To link to this article:** <https://doi.org/10.1080/19439962.2017.1301611>



Published online: 24 Apr 2017.



Submit your article to this journal [↗](#)



Article views: 359



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 7 View citing articles [↗](#)

# Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: A combined approach

Athanasios Theofilatos<sup>a</sup>, George Yannis<sup>a</sup>, Constantinos Antoniou <sup>b</sup>,  
Antonis Chaziris<sup>c</sup>, and Dimitris Sermpis<sup>d</sup>

<sup>a</sup>Department of Transportation Planning and Engineering, School of Civil Engineering, National Technical University of Athens, Zografou Campus, Zografou-Athens, Greece; <sup>b</sup>Laboratory of Transportation Engineering, Department of Surveying Engineering, Technical University of Munich, Munich, Germany; <sup>c</sup>Traffic Management Center, Athens, Greece; <sup>d</sup>Attikes Diadromes S.A., Athens, Greece

## ABSTRACT



Predicting road accident probability by exploiting high-resolution traffic data has been a continuously researched topic in the last years. However, there is no specific focus on powered-two-wheelers. Furthermore, urban arterials have not received adequate attention so far because the majority of relevant studies considers freeways. This study aims to contribute to the current knowledge by utilizing support vector machine (SVM) models for predicting powered-two-wheeler (PTW) accident risk and PTW accident type propensity on urban arterials. The proposed methodology is applied on original and transformed time series of real-time traffic data collected from urban arterials in Athens, Greece, for 2006 to 2011. Findings suggest that PTW accident risk and PTW accident type propensity can be adequately defined by the prevailing traffic conditions. When predicting PTW accident risk, the original traffic time series performed better than the transformed time series. On the other hand, when PTW accident type is investigated, neither of the two approaches clearly outperformed the other, but the transformed time series perform slightly better. The results of the study indicate that the combination of SVM models and time-series data can be used for road safety purposes especially by utilizing real-time traffic data.

## KEYWORDS

accidents; powered-two-wheelers; real-time data; support vector machines; time series

## 1. Introduction and background

Road safety is a major concern for societies, as accidents impose serious problems to societies in terms of human costs, economic costs, property damage costs, and medical costs. Annually, there are 1.25 million fatalities, though one half of

**CONTACT** Athanasios Theofilatos  [atheofil@central.ntua.gr](mailto:atheofil@central.ntua.gr)  Research Associate, National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Heron Polytechniou str., GR-15773 Athens, Greece.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/utss](http://www.tandfonline.com/utss).

© 2018 Taylor & Francis Group, LLC and The University of Tennessee

fatalities on the world's roads are "vulnerable road users": pedestrians, cyclists, and motorcyclists (World Health Organization, 2015). Furthermore, vulnerable road users' safety (riders, pedestrians) is going to face challenges in the coming years (Tiwari, 2015a, 2015b).

A significant increase in motorcycling activities is observed in many countries worldwide in the last years. Over the last 20 years, the number of mopeds and motorcycles together referred to as powered-two-wheelers (PTWs) in Europe has almost doubled (Yannis, Vlahogianni, Golias, & Saleh, 2010). This shift in mode choice is likely to be attributed to economic, mobility, and flexibility benefits offered by PTWs. It is also notable that motorcyclists usually drive faster than car drivers (Jevtić, Vujanić, Lipovac, Jovanović, & Pešić, 2015).

Furthermore, PTW fatalities accounted for 18% of the total number of road accident fatalities in 2013 in the European Union-23 (EU-23) countries (European Road Safety Observatory [ERSO], 2015). The majority of moped fatalities occurred in urban areas whereas the majority of motorcycle fatalities occurred in rural areas (ERSO, 2015). Per vehicle mile travelled, motorcycle riders have a 34-fold higher risk of death in an accident than the other motor vehicles users (Lin & Kraus, 2009).

Huge efforts have been made by researchers to explain PTW accident risk, and numerous PTW accident-related factors have been identified in international literature so far. For example, the road environment such as road type, road geometry, and roadside installations have been found to have an influence on PTW accident occurrence (Harnen, Wong, Radin Umar, & Wan Hashim, 2003; Wanvik, 2009). Haque, Chin, and Huang (2009) found that several geometrical and environmental factors were linked with non-at-fault crashes of motorcyclists. Schneider, Savolainen, Van Boxel, and Beverley (2012) stated that younger motorcyclists, riders under the influence of alcohol (DUI), riders without insurance or not wearing helmet are more likely to be at fault in a crash. Human errors play of course a very important role as well (Penumaka, Savino, Baldanzini & Pierini, 2014). For a more complete list of PTW relevant risk factors, the reader is encouraged to refer to Vlahogianni, Yannis, and Golias (2012b) and Theofilatos and Yannis (2015).

The impact of traffic characteristics on PTW safety has not been investigated in a large extent. Various studies have addressed the effect of traffic on vehicle accidents, but the literature regarding PTW accidents is limited (Abdul Manan & Várhelyi, 2012; Sharma, Landge, & Deshpande, 2013). The investigation of PTW safety in relation to traffic characteristics on a real-time basis is considered highly important due to the vulnerability of PTWs, but also due to the conflicting interactions with other vehicles on the road, which complicates the understanding PTW drivers' behavior (Barmounakis, Vlahogianni, & Golias, 2016).

Recently, there is a trend in predicting and explaining road accident occurrence with real-time traffic data (Abdel-Aty & Pande, 2005; Abdel-Aty, Pande, Lee, Gayah, & Dos Santos, 2007; Ahmed & Abdel-Aty, 2012; Imprialou, Orfanou, Vlahogianni, & Karlaftis, 2014; Vlahogianni, Karlaftis, & Orfanou, 2012a; Vlahogianni,

Karlaftis, Golias and Halkias, 2010; Yu & Abdel-Aty, 2013a). However, to the best of our knowledge, there are few relevant studies dedicated to PTWs (Theofilatos & Yannis, 2017).

Accident type (referred also as “collision type” or “crash type”) is identified as another important parameter with a significant role in road safety, as underlined by a number of studies (Kim, Washington, & Oh, 2006; Pande & Abdel-Aty, 2006). However, various types of collisions are generally not distinguished in most studies, possibly because of the difficulties in collecting the necessary data (Christoforou, Cohen, & Karlaftis, 2011). Christoforou et al. (2011) also stressed that in most of the studies that considered different accident types, a distinction between single- and multivehicle accidents is made. The majority of these studies utilize aggregated traffic data (Ceder & Livneh, 1982; Zhou & Sisiopiku, 1997). However, some studies exploit real-time data (Abdel-Aty & Pande, 2005; Lee, Hellinga, & Saccomanno, 2003; Pande & Abdel-Aty, 2006). For example, Zhiqing, Du, Guo, and Liu (2007) analyzed the influence of the dynamic changing of weather, road, and transportation on freeway operating safety. Abdel-Aty and Pande (2005) exploited real-time traffic from the I-4 corridor in Orlando, Florida, and attempted to investigate the factors determining the accident type. The authors found that variation in speed 10 to 15 min prior to an accident is among the most significant factors. Golob, Recker, and Pavlis (2008) stated that congestion had a considerable influence on vehicle involvement. For example, the authors mention that congestion in the left and interior lanes distinguishes single- from multivehicle accidents.

It can be concluded that PTW safety has not received significant attention when real-time traffic data are used. Moreover, there are a few more gaps of knowledge in terms of data. First of all, the vast majority of relevant international literature focused on real-time traffic data from freeways and not from urban roads. Secondly, time series of real-time traffic data were not extensively exploited when investigating road accidents.

In terms of methodology, traffic time series could be exploited by means of predictive models. Support vector machines (SVM) specifically is a relatively new statistical method, which can handle multicollinearity, and is successfully used and evaluated for predictive purposes in road safety (Li, Lord, Zhang, & Xie, 2008; Yu & Abdel-Aty, 2013b). Moreover, Li et al. (2008) suggest the assessment of SVM models' performance when only traffic flow is considered. For these reasons, the SVM were selected for this study. It is noted that other classifiers could also be further evaluated, such as random forests and artificial neural networks (ANNs).

In that context, the research presented in this article aims to add to the current knowledge by investigating the possibility of utilizing SVM models for predicting PTW accident risk (PTW involvement in an accident or not) and PTW accident type propensity on urban arterials by exploiting real-time traffic data. Each case

constitutes a typical classification problem yet addressed using an advanced techniques.

## 2. Method

To predict PTW risk and PTW accident type propensity, a combined methodological approach was followed. More specifically, a time-series classification was performed by applying SVMs, which is a powerful machine learning technique. Firstly, the SVMs were applied by utilizing the original time-series data and secondly by utilizing the discrete wavelet transform (DWT) transformed data. Then, the findings of the study are compared and conclusions are drawn.

### 2.1. Time-series approach

#### 2.1.1. Original time series

Time-series classification is used when it is desired to build a classification model based on labelled time series and is then aimed to use the model to predict the label of unlabeled time series. Classification of unlabeled time series to existing classes is a further traditional data mining task. By *labelled time series*, it means that a training data set with correctly classified observations is used, and then the developed models are used to predict the labels of a test data set (Kleist, 2015).

It is possible to extract new features from time series to potentially improve the performance of classification models. There are various such techniques for feature extraction such as the singular value decomposition (SVD), discrete Fourier transform (DFT), DWT, piecewise aggregate approximation (PAA), perpetually important points (PIP), piecewise linear representation and symbolic representation (Zhao, 2013).

In this approach, the original time-series data are used, namely, data that have been sampled at equispaced points in time, without applying any techniques for feature extraction.

#### 2.1.2. Wavelet transformed time series

The wavelet transform provides a multiresolution representation using wavelets. In this article, a DWT is used to extract features from time series and then build a classification model (Burrus, Gopinath, & Guo, 1998). The time series and its transform can be considered to be two representations of the same mathematical entity.

The very name *wavelet* originates from the requirement that they should integrate to zero, “waving” above and below the  $x$  axis (Vidakovic & Mueller, 1994). A DWT is any wavelet transform for which the wavelets are discretely sampled and is an orthonormal transform. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures frequency and location information (location in time). DWT has been previously implemented in

the traffic flow analysis and forecasting (Jiang & Adeli, 2004; Vlahogianni, Geroliminis, & Skabardonis, 2008).

A very good description of DWT is demonstrated in McLeod, Yu, and Mahdi (2012). A time series of dyadic length is considered  $z_t, t = 1, \dots, n$ , where,  $n = 2^J$ . The DWT decomposes the time series into  $J$  wavelet coefficients vectors,  $W_{i,j} = 0, \dots, J-1$  each of length  $n_j = 2^{J-j}$ ,  $j = 1, \dots, J$  plus a scaling coefficient  $V_J$ . Each wavelet coefficient is constructed as a difference of two weighted averages each of length  $\lambda^j = 2^{j-1}$ . Similarly to DFT, the DWT provides an orthonormal decomposition,  $W = WZ$ , where,  $W' = (W'_1, \dots, W'_{J-1}, V'_J)$ ,  $Z = (z_1, \dots, z_n)'$ .

There are two functions that play a primary role in wavelet analysis: the scaling function (father wavelet) and the wavelet (mother wavelet). The simplest wavelet analysis is based on Haar scaling function, (Haar wavelet transform) (Struzik & Siebes, 1999). The Haar wavelet is a sequence of rescaled “square-shaped” functions that together form a wavelet family or basis. The Haar sequence is recognized as the first known wavelet basis and was proposed in 1909 by Alfréd Haar (1910). The Haar scaling function  $\varphi(x)$  is defined as:

$$\varphi(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The Haar wavelet's mother function is then defined as  $\psi(x) = \varphi(2x) - \varphi(2x-1)$

$$\psi(x) = \begin{cases} 1, & 0 \leq x < 1/2 \\ -1, & 1/2 \leq x < 1 \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Graphical representations of a Haar wavelet transform, which is the simplest DWT, can be found in Vidakovic and Mueller (1994) and Zhao (2013).

## 2.2. Support vector machines

Traditional statistical modelling has been widely used for transportation data analysis. However, such approach contains some limitations, for example, modelling assumptions that may not always be true. Nonparametric and artificial intelligent methods could then be applied to overcome such limitations.

SVMs constitute a relatively new modelling technique, which is useful for classification problems (Keckman, 2005). In transportation science, the studies having used SVMs are relatively rare (Li et al., 2008; Li, Liu, Wang, & Xu, 2012), especially in real-time crash risk evaluation (Yu & Abdel-Aty, 2013b, 2014).

SVMs have originated from statistical learning theory (Vapnik, 1998) and have been developed by Cortes and Vapnik (1995) mainly for binary classification. Basically, when building a SVM model, the aim is the optimal separating hyperplane

between two classes by maximizing the margin between the classes' closest points (Meyer, 2001). Therefore, different classes are separated by the hyperplane:

$$\langle w, \Phi(x) \rangle + b = 0 \quad (3)$$

which corresponds to the decision function

$$f(x) = \text{sign}(\langle \Phi(x_i), w \rangle + b) \quad (4)$$

The points lying on the boundaries are the support vectors, whereas the middle of the margin is the optimum separating hyperplane. From all the available kernel-based algorithms (kernels) (e.g., linear, polynomial, Gaussian radial-basis function and sigmoid), the Gaussian radial-basis function kernel (RBF) was considered in this article (Karatzoglou, Smola, Hornik, & Zelis, 2005), which is defined as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i, x_j\|^2), \gamma > 0 \quad (5)$$

where,  $\gamma$  is the kernel parameter.

Moreover, with the Gaussian RBF function, the SVM model has two parameters ( $C, \gamma$ ) which need to be determined. The cost parameter  $C$  controls the penalty for misclassifying a training point and consequently the complexity of the prediction function (Karatzoglou, Meyer, & Hornik, 2006). A high cost value  $C$  will result in a complex prediction function to misclassify as few training cases as possible. On the other hand, a low cost parameter  $C$ , results in simpler prediction functions. Thus, this type of SVM model is called "C-SVM" (Karatzoglou et al., 2006).

Karatzoglou et al. (2006) provide the primal form of the bound constraint C-SVM formulation:

$$\begin{aligned} \text{minimize } t(w, \zeta) &= \left(\frac{1}{2}\right) \|w\|^2 + \left(\frac{1}{2}\right) \beta^2 + \left(\frac{C}{m}\right) \sum_{i=1}^m \zeta_i \\ \text{subject to } y_i(\langle \Phi(x_i), w \rangle + b) &\geq 1 - \zeta_i \end{aligned} \quad (6)$$

where  $i = 1, \dots, m$ , and  $\zeta_i \geq 0$ , where,  $i = 1, \dots, m$ .

The dual form of the bound constraint C-SVM formulation (Karatzoglou et al., 2006) is:

$$\begin{aligned} \text{maximize } W(\alpha) &= \sum_{i=1}^m a_i - \frac{1}{2} \sum_{i,j=1}^m a_i a_j (y_i y_j + k(x_i, x_j)) \\ \text{subject to } 0 &\leq a_i \leq \frac{C}{m}, \text{ where, } i = 1, \dots, m \end{aligned} \quad (7)$$

and

$$\sum_{i=1}^m a_i y_i = 0.$$

Finally, SVMs can be enhanced to tackle nonlinear classification problems, regression and outlier detection, like other traditional machine learning models (Karlaftis & Vlahogianni, 2011). The major limitation of SVMs is that the models cannot be used directly to identify the relationships between the dependent and the independent variables. Therefore, SVMs can be considered as a “black box” technique. Jiang, Zou, Zhang, Tang, and Wang (2016) state that though machine learning models use a “black box” approach that lacks a good interpretation of the model, they are more flexible with no or little prior assumptions for input variables. Efforts to improve model prediction and interpretability have been discussed also in Zhang et al. (2016). For example, a well-applied solution is to carry out sensitivity analysis (Chen, Zhang, Qian, Tarefder, & Tian, 2016; Li et al., 2012; Yu & Abdel-Aty, 2013b). A recent study by Chen et al. (2016) provides a very good description of a sensitivity analysis and crash injury severity of rollover crashes. More specifically, the authors suggest that each explanatory variable has to be changed by a user-defined amount (while other variables remain unchanged) and afterwards the probabilities of each injury severity level before and after this perturbation were simulated in the SVM model. Other techniques to handle this issue exist and are also described in Orfanou, Vlahogianni, and Karlaftis (2012).

### 2.3. General framework

The general methodological framework is outlined in this section. More specifically, for each classification problem:

- Step 1. Choose the original or apply a DWT on the time series.
- Step 2. Split the data set into training set (80%) and test set (20%).
- Step 3. Tune the SVM model to find the best parameters ( $C$ ,  $\gamma$ ) on the basis of the highest 10-fold crossvalidation mean accuracy performed on the training set.
- Step 4. Apply the best SVM model on the test set.

Firstly, our approach suggests to choose if the original or the transformed traffic time series are applied (Step 1). As in every prediction attempt, when building SVM models a training and a testing set have to be defined. The 80% of the data set is used as training set and the 20% as a test set. This applies for each classification problem.

The models are calibrated on the training set (Step 3) by using a 10-fold crossvalidation technique was applied on each data set, to have a measure of the overall classification performance of the SVM models. Generally, in  $k$ -fold crossvalidation, the original sample is randomly divided into  $k$  equal-sized subsamples. Of the  $k$  subsamples, a single subsample is used as the validation data set for testing the prediction performance of the model while the remaining  $k-1$  subsamples are used as training data to calibrate the model. The crossvalidation process is then repeated  $k$  times (the folds), with each of the  $k$  subsamples used exactly once as the validation data (Kohavi, 1995). Consequently, 10 subsamples were created. The next step (Step 4) consists of the application of the SVM model on the test set to predict the



dependent variable, on the basis of the best parameters ( $C$ ,  $\gamma$ ) identified in Step 3. The need for an investigation of SVMs performance by applying different values of  $C$  and  $\gamma$  was emphasized by recent literature in the field (Yu & Abdel-Aty, 2013b), and thus it is another step made by this article. The performance of SVM models on the test set is evaluated on the basis of various metrics such as total accuracy, sensitivity, specificity, and so on.

### 3. Data preparation

The urban roads that were chosen were the Kifisias and Mesogeion avenues in Athens, Greece, mainly due to the fact that they had very similar characteristics, because they are signalized arterial corridors with high volumes and fluctuating traffic conditions during a typical day.

Kifisias Avenue has a total length is about 20 km, beginning 4 km northeast of downtown Athens and ending by the municipal boundary of Nea Erythraia north of Kifisia. The total amount of lanes is three, up to Kifisia, then two through Kifisia, before it turns to a one-lane (per direction) road for the rest of its length. The avenue begins at the intersection of Alexandras and Mesogeion Avenues. The avenue has a bus lane for a significant section of its length, close to its start.

Mesogeion Avenue is also a main road in Athens and its eastern suburbs. The total length is approximately 8 km. Mesogeion Avenue also intersects with Michalakopoulou Street, Katechaki Avenue, and Perikleous Avenue.

The required accident data were collected from the Greek accident database SANTRA, which is provided by the Department of Transportation Planning and Engineering of the National Technical University of Athens. A 6-year period was considered for the analyses of the present thesis, namely, 2006 to 2011.

Traffic data were extracted from the Traffic Management Centre (TMC) of Athens, which operates on a daily basis from July 2004 covering various major arterials in the city of Athens. To apply the SVM models, a number of different data sets had to be prepared. Having known the time and location for each accident, the 3-h time series of traffic flow, occupancy, and speed (in 5-min intervals ending at the time of the accident) from the closest upstream as well as the closest downstream loop detector were utilized. For example, if an accident occurred in Kifisias Avenue on Wednesday 12 August 2009 at 13:00, then traffic data from Wednesday 12 August 2009 10:00 to 13:00 are extracted from the closest upstream and downstream loop detector measured in 5-min intervals. There were rare cases when loop detectors suffered from problems that might have resulted in unreasonable values for speed, volume, and occupancy. Such unrealistic values (e.g., occupancy > 100%, speed > 200 km/h or speed > 0 along with flow = 0) were discarded from the database. Accidents with traffic data unavailability were also discarded.

Thus, each data set contains one of the following time series: traffic flow upstream, traffic flow downstream, speed upstream, speed downstream, occupancy

upstream, occupancy downstream. Then, by following the DWT procedure described earlier all data sets are then transformed. Consequently, for each data set containing the original time-series data a secondary data set is created with the transformed time-series data.

Then for each dataset, the SVM models were applied to predict:

- a) PTW accident risk
- b) PTW accident type propensity.

To enhance the classification performance of SVMs, PTW accident type was classified as a binary outcome, namely, single- and multivehicle accidents, transforming the classification problem to a two-category classification. This approach was followed because literature indicates that typical multiclassification problems have been very commonly observed when methods such as SVMs, ANNs, or classification trees are applied (Delen, Sharda, & Bessonov, 2006; Li et al., 2012). For example, Li et al. (2012) developed SVMs to model injury severity, and the SVM model ignored severity categories with small proportions (i.e., fatal and incapacitating injuries) to improve the overall classification accuracy.

The final accident data set consists of 527 accidents. Figures 1 and 2 present the respective percentage of PTWs as well as the accident type of PTWs.

It can be seen that the PTWs were involved in 326 of them (61.9%). Regarding PTW accident type, PTWs were involved in 107 single-vehicle accidents (32.8%) and 219 multivehicle accidents (67.1%).

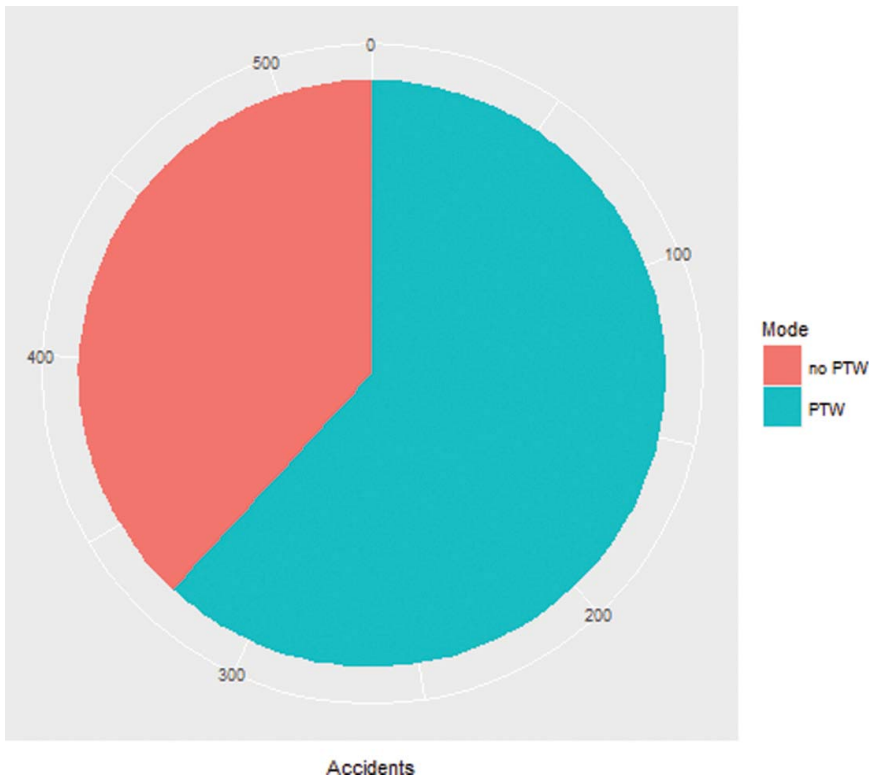
## 4. Results and discussion

### 4.1. Prediction of PTW accident risk

In general, the developed SVM models showed relatively good classification accuracy compared to other similar studies (Li et al., 2012). However, it is noted that this is the first attempt to incorporate real-time traffic time series in SVMs. The development of the models showed that by modifying the two parameters,  $C$  and  $\gamma$  accordingly, the classification accuracy can be substantially influenced. Figure 3 demonstrates a contour plot of the error resulting from the search for the best parameters.

Tables 1 and 2 illustrate the total classification performance of SVM models when predicting PTW accident risk by utilizing the original and the transformed time series, respectively.

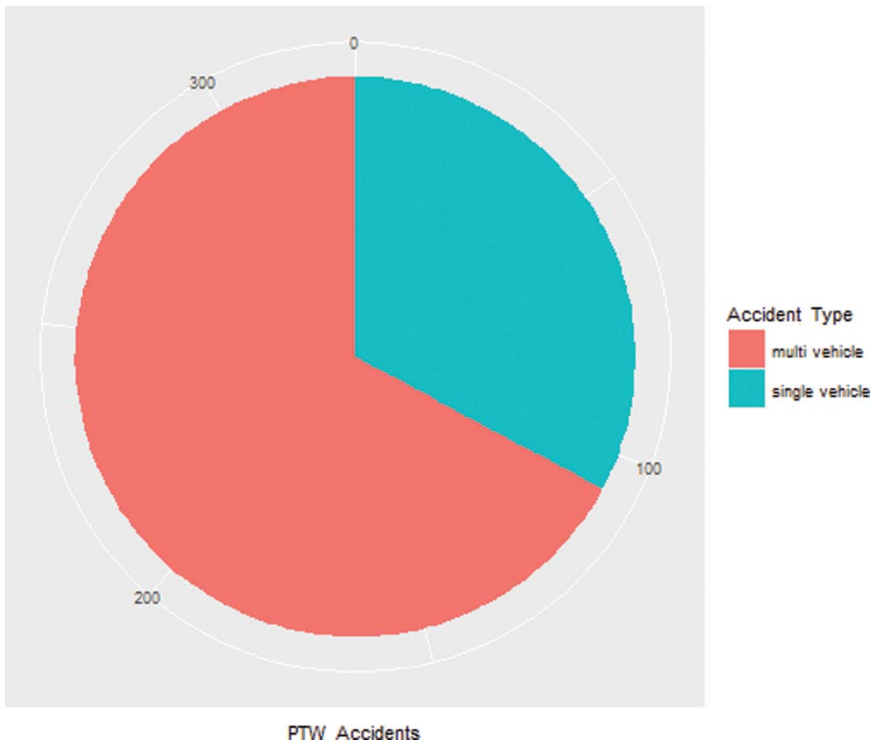
When the original time series are considered, the prediction accuracy on the test set is consistently higher than 60%. The best accuracy on the test set is achieved by the flow downstream (63.55%) and occupancy upstream (64.22%). On the other hand, the time series of occupancy downstream of the accident location was the worse predictor of PTW accident risk (60.00%). One core characteristic of the final original time-series models is that that they are very capable of correctly identifying PTW accidents, as the true positive rates



**Figure 1.** Share of powered-two-wheeler (PTW) involvement in accidents.

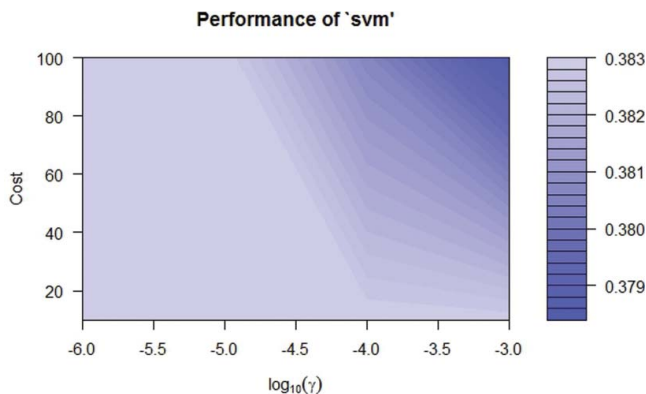
is very high almost in all cases. Indeed, false negative rates are very low as they usually range from 9.52% (speed upstream) to 16.90% (flow downstream). Only speed downstream has a relatively high false negative rate (32.47%). However, the original time series are not equally strong when it comes to predict accidents without a PTW, as it can be observed by the increased false positive rates that reach up to 82.22% (flow upstream). On the other hand, speed downstream time series show the lowest false positive rate (57.14%) and therefore have an adequate specificity value (42.86%). Therefore, it is suggested that the speed downstream is the best predictor for PTW accidents when original time series are considered.

When the DWT time series are considered, the total classification accuracy was generally lower than the original time series. Flow upstream and occupancy downstream had the highest accuracies (61.11% and 58.09%, respectively) and showed also a good balance between sensitivity (true positive rate) and specificity (true negative) values. Thus they are considered as the best predictors in this case. In general, the false negative rates of the DWT time series are considered generally adequate as they range from 18.03% (speed downstream) to 39.71% (speed upstream), however false positive rates have generally increased values. Consequently, the DWT approach seems to have the



**Figure 2.** Single and multivehicle accidents with powered-two-wheelers (PTWs).

same strengths and limitations as the original time-series approach. In addition, the performance is not as good as the original time series. The lower prediction performance of the transformed time series in predicting PTW accident risk may imply that it be not necessary to extract features from time series but use the original time series instead. Alternatively, a different transformation could be applied.



**Figure 3.** Graphical example of support vector machine (SVM) parameters tuning-flow upstream for powered-two-wheeler (PTW) accident involvement.

**Table 1.** Total classification performance of support vector machine (SVM) models to predict powered-two-wheeler (PTW) accident risk (original time series).

Original Time Series (PTW involvement in an accident)						
Total SVM Performance	Speed Downstream	Speed Upstream	Flow Downstream	Flow Upstream	Occupancy Downstream	Occupancy Upstream
10-fold cross validation mean accuracy % on the training set	61.52%	65.06%	63.71%	64.68%	63.74%	64.99%
Best parameters (C, $\gamma$ )	(100, 0.01)	(100, 0.001)	(100, 0.01)	(100, 0.01)	(100, 0.001)	(100, 0.001)
Accuracy % on the test set	60.90%	62.96%	63.55%	61.11%	60.00%	64.22%
Sensitivity (True positive rate)	67.53%	90.48%	83.10%	92.06%	87.10%	87.14%
Specificity (True negative weight)	42.86%	24.44%	25.00%	17.78%	20.93%	21.43%
False positive rate	57.14%	75.56%	75.00%	82.22%	79.07%	78.57%
False negative rate	32.47%	9.52%	16.90%	7.94%	12.90%	12.86%

#### 4.2. Prediction of PTW accident type

Tables 3 and 4 illustrate the total classification performance of SVM models regarding PTW accident type by utilizing original and transformed time series, respectively.

As a first remark, neither of the two approaches (original or DWT time series) clearly outperformed the other. However, the DWT transformation of the time series provides a slight better performance.

When the original time series are considered, the best performance on the test set varies from 53.38% (flow downstream) to 66.66% (speed downstream). Overall, the false negative rates are generally low. According to the other performance evaluation measures as well, the best performance is generally achieved when the speed downstream is considered. However, all models suffer from low specificity, meaning that there is a low rate of true negative accuracy (limited strength of identifying single vehicle PTW accidents), whereas

**Table 2.** Total classification performance of support vector machine (SVM) models to predict powered-two-wheeler (PTW) accident risk (transformed time series).

Discrete Wavelet Transform Time Series (PTW involvement in an accident)						
Total SVM Performance	Speed Downstream	Speed Upstream	Flow Downstream	Flow Upstream	Occupancy Downstream	Occupancy Upstream
10-fold cross validation mean accuracy % on the training set	61.28%	61.60%	60.37%	62.39%	61.90%	61.36%
Best parameters (C, $\gamma$ )	(100, 0.001)	(100, 0.02)	(100, 0.0001)	(100, 0.01)	(100, 0.01)	(100, 0.08)
Accuracy % on the test set	57.14%	50.00%	56.10%	61.11%	58.09%	57.79%
Sensitivity (True positive rate)	81.97%	60.29%	68.42%	65.43%	62.30%	65.22%
Specificity (True negative weight)	22.73%	23.50%	25.81%	48.15%	52.27%	45.00%
False positive rate	77.27%	67.50%	74.19%	51.85%	47.73%	55.00%
False negative rate	18.03%	39.71%	31.58%	34.57%	37.70%	34.78%

**Table 3.** Total classification performance of support vector machine (SVM) models using to predict powered-two-wheeler (PTW) accident type (original time series).

Original Time Series (Accident type)						
Total SVM Performance	Speed Downstream	Speed Upstream	Flow Downstream	Flow Upstream	Occupancy Downstream	Occupancy Upstream
10-fold cross validation mean accuracy % on the training set	69.23%	63.81%	63.60%	58.36%	64.03%	54.27%
Best parameters (C, $\gamma$ )	(100, 0.3)	(100, 0.1)	(100, 0.5)	(100, 0.3)	(100, 0.2)	(100, 0.1)
Accuracy % on the test set	66.66%	60.60%	55.38%	55.22%	57.81%	53.73%
Sensitivity (True positive rate)	76.16%	79.49%	68.18%	75.61%	73.17%	67.44%
Specificity (True negative weight)	28.57%	33.33%	28.57%	23.08%	30.43%	29.17%
False positive rate	71.43%	66.67%	71.43%	76.92%	69.57%	70.83%
False negative rate	23.81%	20.51%	31.82%	24.39%	26.83%	32.56%

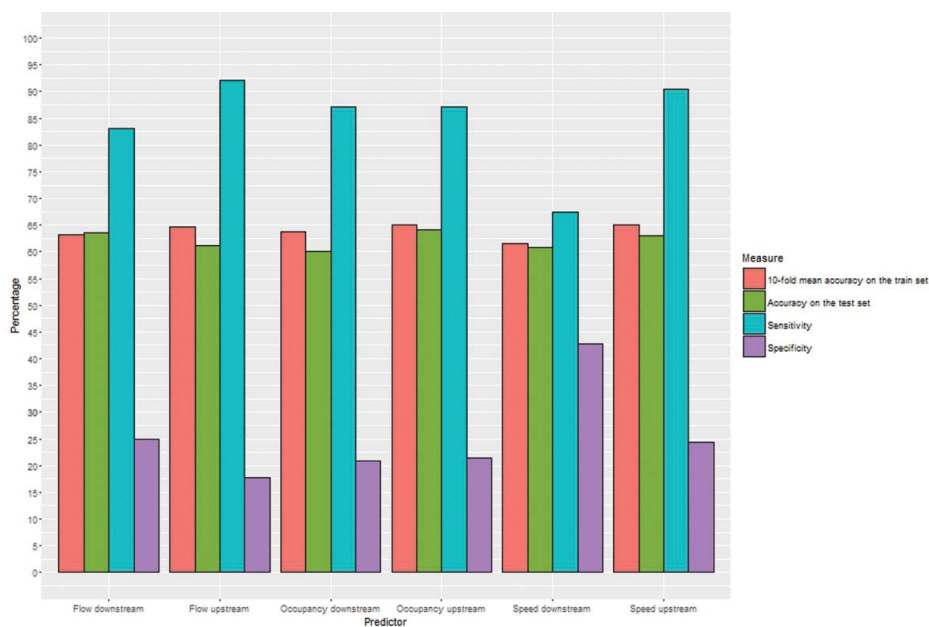
the false positive rates are high, varying from 66.67% to 76.92%.

When the DWT time series are considered, the total classification accuracy on the test set was generally around 60%. False negative rates are low, ranging from 24.32% (speed downstream) to 36.17% (flow upstream). On the other hand, false positive rates are lower than the respective values of the original time series but are still considered high as they are usually around 60%, whereas occupancy upstream has the lowest false positive rate (52%). Traffic flow upstream had the lowest accuracy on the test set (56.70%), whereas occupancy downstream showed the highest accuracy on the test set (62.50%). Therefore, occupancy downstream is considered as the most appropriate predictor because of the overall decent values of all performance measures.

Figures 4–7 graphically illustrate the performance of the SVM models when predicting PTW accident risk and PTW accident type, by utilizing original and DWT time series.

**Table 4.** Total classification performance of support vector machine (SVM) models using to predict powered-two-wheeler (PTW) accident type (transformed time series).

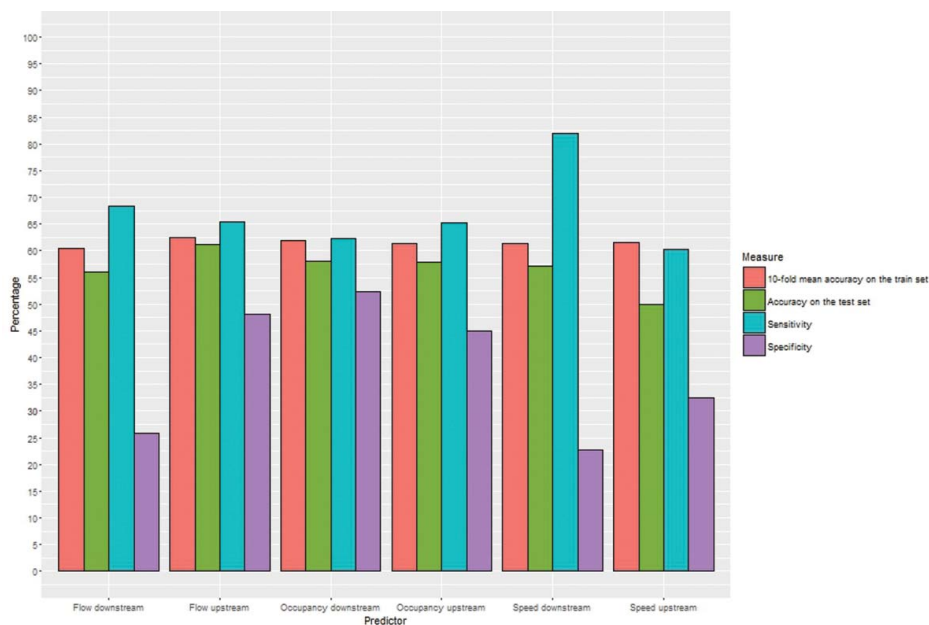
Discrete Wavelet Transform Time Series (Accident type)						
Total SVM Performance	Speed Downstream	Speed Upstream	Flow Downstream	Flow Upstream	Occupancy Downstream	Occupancy Upstream
10-fold cross validation mean accuracy % on the training set	61.33%	65.29%	64.36%	62.08%	66.41%	67.68%
Best parameters (C, $\gamma$ )	(100, 0.01)	(100, 0.01)	(100, 0.025)	(100, 0.01)	(100, 0.0001)	(100, 0.01)
Accuracy % on the test set	60.30%	59.09%	61.54%	56.70%	62.50%	59.09%
Sensitivity (True positive rate)	75.68%	68.18%	68.75%	63.83%	69.39%	65.85%
Specificity (True negative weight)	38.46%	40.91%	41.18%	40.00%	40.00%	48.00%
False positive rate	61.54%	59.09%	58.82%	60.00%	60.00%	52.00%
False negative rate	24.32%	31.82%	31.25%	36.17%	30.61%	34.15%



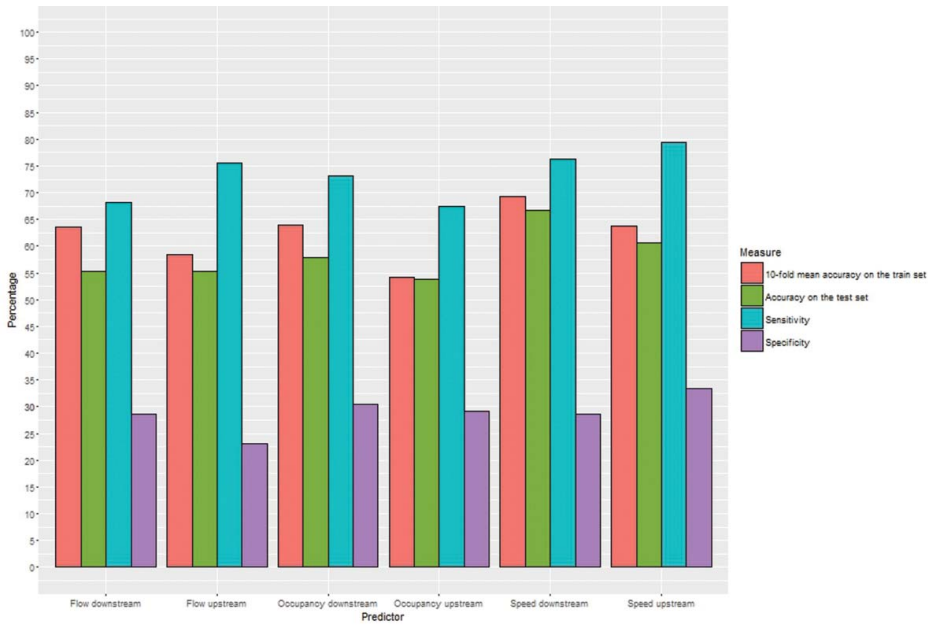
**Figure 4.** Graphical illustration of support vector machine (SVM) models performance when analyzing powered-two-wheeler (PTW) accident involvement with original time series.

## 5. Conclusions

This article presents the prediction results from the SVM models, which were applied on the original time series and on the DWT time series of flow, speed, and occupancy upstream and downstream of the accident location. This

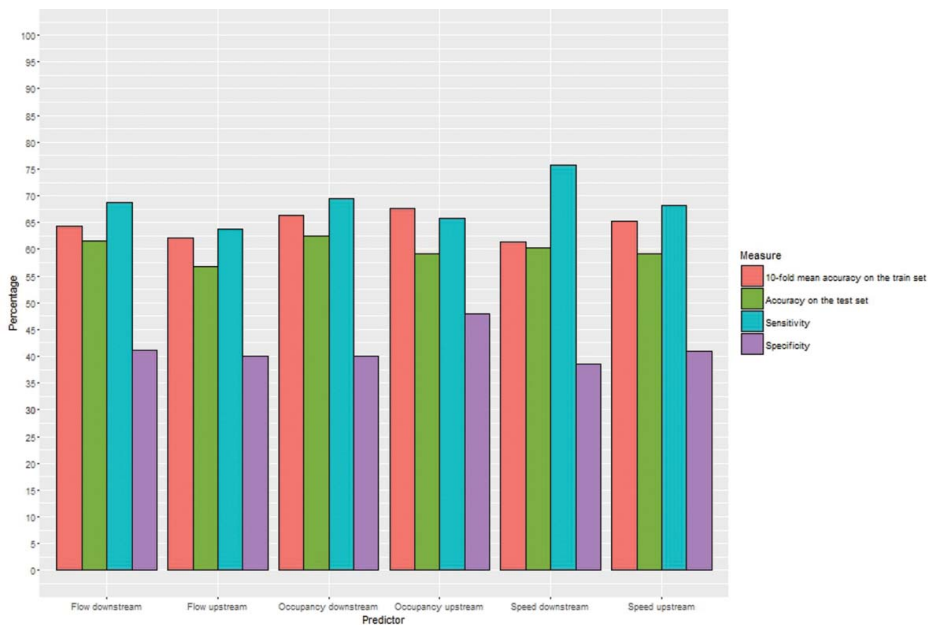


**Figure 5.** Graphical illustration of support vector machine (SVM) models performance when analyzing powered-two-wheeler (PTW) accident involvement with discrete wavelet transform time series.



**Figure 6.** Graphical illustration of support vector machine (SVM) models performance when analyzing powered-two-wheeler (PTW) accident type with original time series.

methodological approach is considered as a first attempt to incorporate time-series data when analyzing road safety with real-time traffic data. Moreover, the opportunity of applying of a relatively new and scientifically strong classification technique



**Figure 7.** Graphical illustration of support vector machine (SVM) models performance when analyzing powered-two-wheeler (PTW) accident type with discrete wavelet transform time series.



such as SVMs in road safety with real-time traffic data was only recently been explored.

The prediction of whether a PTW is involved in an accident given that the accident has occurred is a typical two-category classification problem. In some cases, the approach is very promising. In overall, the original time-series data performed better than the DWT time series. Consequently, original time series are preferred for predicting PTW accident risk.

When predicting PTW accident type, the dependent variable (single- and multivehicle accident) is considered as a two-category classification problem as well. In this case, original time series and DWT time series had similar performance, and neither of the two approaches clearly outperformed the other. However, the DWT transformation showed better overall performance measures. Consequently, the transformed time series should be utilized to predict PTW accident type.

It is interesting, despite the fact that Zhao (2013) suggests to extract features from time-series when performing time-series data mining, the performance of SVMs on the DWT data did not clearly outperformed original time series. This means that the transformation applied in the article is not always necessary. Alternatively, a different transformation could be applied, such as SVD, DFT, PAA, and PIP. Other classification methods that have been applied widely in literature (Elminity, Yan, Radwan, Russo, & Nashar, 2010; Harb, Yan, Radwan, & Su, 2009; Moon & Hummer, 2009; Yan, Richards, & Su, 2010; Yang, Lu, Gunarante, & Xiang, 2003) could also be potentially considered.

Summing up, this methodological approach showed promising results and produced a number of promising classification performances in some cases, despite the fact that there was a limited and inhomogeneous data set. The main conclusion of the present article is that the combination of SVM models and time-series data can be used for road safety purposes, especially by utilizing real-time traffic data. It is suggested that this direction has to be further explored for accident frequency and severity analyses. Moreover, further insight on the effect of traffic parameters could be gained. For instance, if the time series that predict accident risk are identified (e.g., traffic, speed or occupancy), then additional analyses could be carried out on that basis to gain more information about the temporal evolution of that traffic parameter. In other words, the temporal evolution of traffic time series could serve as a preliminary analysis and could be considered as a first step toward the identification of hazardous traffic conditions. For example, if a specific traffic parameter (e.g., flow) is capable of predicting accident risk, then more focus should be given on further exploration of this parameter by applying more traditional statistical modeling to extract direct relationships. Additionally, if the relevant road authority is aware of this information, then specific actions for better traffic monitoring in major urban arterials could be taken to reduce PTW accident risk in urban arterials. For instance, after traffic conditions that increase accident risk are identified

(e.g., hazardous traffic time series), advanced warning signs about PTWs or variable messages signs could be implemented.

Finally, it would be interesting to expand this research in other specific road segments, such as work zones and other critical elements of road network (Meng & Weng, 2011; Meng, Weng, & Qu, 2010). In that case, a possible future research direction is to focus on predicting rear-end collisions in work zones by utilizing real-time traffic data.

## ORCID

Constantinos Antoniou  <http://orcid.org/0000-0003-0203-9542>

## References

- Abdel-Aty, M., & Pande, A. (2005). Identifying crash propensity using specific speed conditions. *Journal of Safety Research*, 36, 97–108.
- Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., & Dos Santos, C. (2007). Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 11(3), 107–120.
- Abdul Manan, M. M., & Várhelyi, A. (2012). Motorcycle fatalities in Malaysia. *IATSS Research*, 36, 30–39.
- Ahmed, M., & Abdel-Aty, M. (2012). The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions of Intelligent Transportation Systems*, 13(2), 459–468.
- Barmounakis, E., Vlahogianni, E., & Golias, J. (2016). Intelligent transportation systems and powered two wheelers traffic. *IEEE Transactions on Intelligent Transportation Systems*, 17(4), 908–916.
- Burrus, C. S., Gopinath, R. A., & Guo, H. (1998). *Introduction to wavelets and wavelet transforms: A primer*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Ceder, A., & Livneh, M. (1982). Relationship between road accidents and hourly traffic flow—I. *Accident Analysis and Prevention*, 14, 19–34.
- Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis and Prevention*, 90, 128–139.
- Christoforou, Z., Cohen, S., & Karlaftis, M. (2011). Identifying crash type propensity using real-time traffic data on freeways. *Journal of Safety Research*, 42, 43–50.
- Cortes, C., & Vapnik, V. (1995). Support-vector network. *Machine Learning*, 20, 1–25.
- Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis and Prevention*, 38(3), 434–444.
- Elminity, N., Yan, X., Radwan, E., Russo, C., & Nashar, D. (2010). Classification analysis of driver's stop/go decision and red-light running violation. *Accident Analysis and Prevention*, 42, 101–111.
- European Road Safety Observatory (ERSO). (2015). *Traffic safety basic facts on motorcycles and mopeds*. Brussels: European Commission, Directorate General for Transport.
- Golob, T., Recker, W., & Pavlis, Y. (2008). Probabilistic models of freeway safety performance using traffic flow data as predictors. *Safety Science*, 46, 1306–1333.

- Haar, A. (1910). Zur zheorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69 (3), 331–371.
- Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis and Prevention*, 41, 98–107.
- Haque, M. M., Chin, H. C., & Huang, H. (2009). Modelling fault among motorcyclists involved in crashes. *Accident Analysis and Prevention*, 41, 327–335.
- Harnen, S., Wong, S. V., Radin Umar, R. S., & Wan Hashim, W. I. (2003). Motorcycle crash prediction model for non-signalized intersections. *IATSS Research*, 27(2), 58–65.
- Jevtić, V., Vujanić, M., Lipovac, K., Jovanović, D., & Pešić, D. (2015). The relationship between the travelling speed and motorcycle styles in urban settings: A case study in Belgrade. *Accident Analysis and Prevention*, 75, 77–85.
- Jiang, X., & Adeli, H. (2004). Wavelet packet-autocorrelation function method for traffic flow pattern analysis. *Computer-Aided Civil and Infrastructure Engineering*, 19, 324–337.
- Jiang, H., Zou, Y., Zhang, S., Tang, J., & Wang, Y. (2016). Short-term speed prediction using remote microwave sensor data: Machine learning versus statistical model. *Mathematical Problems in Engineering*, 2016, Article ID 9236156. doi:10.1155/2016/9236156
- Imprialou, M., Orfanou, F., Vlahogianni, E., & Karlaftis, M. (2014). Methods for defining spatio-temporal influence areas and secondary incident detection in freeways. *Journal of Transportation Engineering*, 140(1), 70–80.
- Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software*, 15(9), 1–28.
- Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2005). *Kernlab-kernel methods* (R package, Version 0.6-2). Retrieved from <http://CRAN.R-project.org/>
- Karlaftis, M., & Vlahogianni, E. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C*, 19(3), 387–399.
- Keckman, V. (2005). Support vector machines-an introduction. In L Wang, (ed.), *Support vector machines: Theory and applications* (pp. 1–48). Berlin, Heidelberg, Germany; New York, NY: Springer-Verlag.
- Kim, D., Washington, S., & Oh, J. (2006). Modeling crash types: New insights into the effects of covariates on crashes at rural intersections. *Journal of Transportation Engineering*, 132, 282–292.
- Kleist, C. (2015). Time series data mining methods: A review (Master Thesis), Humboldt-Universität zu Berlin, School of Business and Economics.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2, 1137–1143.
- Lee, C., Hellinga, B., & Saccomanno, F. (2003). Real-time crash prediction model for application to crash prevention in freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board*, 1840, 67–77.
- Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis and Prevention*, 40, 1611–1618.
- Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis and Prevention*, 45, 478–486.
- Lin, M. R., & Kraus, J. F. (2009). A review of factors and patterns of motorcycle injuries. *Accident Analysis and Prevention*, 41, 710–722.
- McLeod, A. I., Yu, H., & Mahdi, E. (2012). Time series analysis in R. In T. S. Rao, S. S. Rao, & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 30, pp. 661–712). Amsterdam, The Netherlands: Elsevier.

- Meng, Q., & Weng, J. (2011). Evaluation of rear-end crash risk at work zone using work zone traffic data. *Accident Analysis and Prevention*, 43, 1291–1300.
- Meng, Q., Weng, J., & Qu, X. (2010). A probabilistic quantitative risk assessment model for the long-term work zone crashes. *Accident Analysis and Prevention*, 42(6), 1866–1877.
- Meyer, D. (2001). Support vector machines. *R News*, 1(3), 23–26.
- Moon, J. P., & Hummer, J. E. (2009). Development of safety prediction models for influence areas of ramps in freeways. *Journal of Transportation Safety and Security*, 1(1), 1–17.
- Orfanou, F., Vlahogianni, E., & Karlaftis, M. (2012). Associating driving behavior with hysteretic phenomena of freeway traffic flow. *IFAC Proceedings Volumes*, 45(24), 209–214.
- Pande, A., & Abdel-Aty, M. (2006). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis and Prevention*, 38, 936–948.
- Penumaka, A. P., Savino, G., Baldanzini, N., & Pierini, M. (2014). In-depth investigations of PTW-car accidents caused by human errors. *Safety Science*, 68, 212–221.
- Schneider, W. H. IV, Savolainen, P., Van Boxel, D., & Beverley, R. (2012). Examination of factors determining fault in two-vehicle motorcycle crashes. *Accident Analysis and Prevention*, 45, 669–676.
- Sharma, A. K., Landge, V. S., & Deshpande, N. V. (2013). Modeling motorcycle accident on rural highway. *International Journal of Chemical, Environmental and Biological Sciences*, 1(2), 313–317.
- Struzik, R., & Siebes, A. (1999). The Haar wavelet transform in the time series similarity paradigm. In *Proceedings of the third European conference on principles and practice of knowledge discovery in databases* (pp. 12–22). Berlin Heidelberg: Springer.
- Theofilatos, A., & Yannis, G. (2015). A review of powered-two-wheeler behaviour and safety. *International Journal of Injury Control and Safety Promotion*, 22(4), 284–307.
- Theofilatos, A., & Yannis, G. (2017). Investigation of powered-two-wheeler accident involvement in urban arterials by considering real-time traffic and weather data. *Traffic Injury Prevention*, 18(3), 293–298.
- Tiwari, G. (2015a). Safety challenges of powered two-wheelers. *International Journal of Injury Control and Safety Promotion*, 22(4), 281–283.
- Tiwari, G. (2015b). The safety of vulnerable road users: The challenge for twenty-first century. *International Journal of Injury Control and Safety Promotion*, 22(2), 93–94.
- Vapnik, V. (1998). *Statistical learning theory*. New York, NY: Wiley.
- Vidaković, B., & Müller, P. (1994). *Wavelets for kids – A tutorial introduction*. Durham, NC: Institute of Statistics and Decision Science, Duke University. Retrieved from <http://www.isye.gatech.edu/»brani/wp/kidsA.pdf>
- Vlahogianni, E., Geroliminis, N., & Skabardonis, A. (2008). Empirical and analytical investigation of traffic flow regimes and transitions in signalized arterials. *Journal of Transportation Engineering*, 134(2), 512–522.
- Vlahogianni, E., Karlaftis, M., Golias, J., & Halkias, B. (2010). Freeway operations, spatiotemporal-incident characteristics, and secondary-crash occurrence. *Transportation Research Record*, 2178, 1–9.
- Vlahogianni, E., Karlaftis, M., & Orfanou, F. (2012a). Modeling the effects of weather and traffic on the risk of secondary incidents. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 16(3), 109–117.
- Vlahogianni, E., Yannis, G., & Golias, J. (2012b). Overview of critical risk factors in power-two-wheeler safety. *Accident Analysis and Prevention*, 49, 12–22.
- Wanvik, P. O. (2009). Effects of road lighting: An analysis based on Dutch accident statistics 1987–2006. *Accident Analysis and Prevention*, 41, 123–128.
- World Health Organization. (2015). *Road traffic injuries*. Retrieved from <http://www.who.int/mediacentre/factsheets/fs358/en/>

- Yan, X., Richards, S., & Su, X. (2010). Using hierarchical tree-based regression model to predict train-vehicle crashes at passive highway-rail grade crossings. *Accident Analysis and Prevention*, 42, 64–74.
- Yang, J., Lu, J. J., Gunarante, M., & Xiang, Q. (2003). Forecasting overall pavement condition with neural networks: Application on Florida highway network. *Transportation Research Record*, 1853(1), 3–12.
- Yannis, G., Vlahogianni, E., Golias, J., & Saleh, P. (2010). Road infrastructure and safety of powered-two wheelers. In *Proceedings of the 12th World Conference on Transport Research*, July 11–15 2010, Lisbon.
- Yu, R., & Abdel-Aty, M. (2013a). Investigating the different characteristics of weekday and weekend crashes. *Journal of Safety Research*, 46, 91–97.
- Yu, R., & Abdel-Aty, M. (2013b). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis and Prevention*, 51, 252–259.
- Yu, R., & Abdel-Aty, M. (2014). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science*, 63, 50–56.
- Zhang, W., Zou, Y., Tang, J., Ash, J. & Wang, Y. (2016). Short-term prediction of vehicle waiting queue at ferry terminal based on machine learning method. *Journal of Marine Science and Technology*, 21(4), 729–741.
- Zhao, Y. (2013). *R and data mining: Examples and case studies*. San Diego, CA: Academic Press.
- Zhou, M., & Sisiopiku, V. (1997). Relationship between volume-to-capacity ratios and accident rates. *Transportation Research Record*, 1581, 47–52.
- Zhiqing, Y., Du, X., Guo, Z., & Liu, B. (2007). Modeling freeway real-time operating safety evaluation based on fault tolerance. In *Proceedings of the ICTCT Extra Workshop* (pp. 285–297), Beijing, China.