# Machine Learning Methods for Driving Risk Prediction

Yibo Wang
School of Information
Renmin University of China
Beijing, 100872, China
wangyibo90@yeah.net

Wei Xu
School of Information, and
Smart City Research Center
Renmin University of China
Beijing, 100872, China
weixu@ruc.edu.cn

Yiqun Zhang
School of Information
Renmin University of China
Beijing, 100872, China
ruczyq@163.com

Yu Qin
School of Information
Renmin University of China
Beijing, 100872, China
qinyu.gemini@gmail.com

Wenping Zhang
School of Information
Renmin University of China
Beijing, 100872, China
wpzhang@ruc.edu.cn

Xue Wu
School of Information
Renmin University of China
Beijing, 100872, China
wuxue@ruc.edu.cn

## ABSTRACT

The development of technology makes the personalized analysis of driving behavior possible. A variety of attributes have been added to driving behavior analysis. Through in-depth analysis of driving behavior data, this paper uses machine learning methods to analyze and predict driving risk, thus laying a foundation for improving driver's driving behavior. Moreover, the experiment results shows that it is necessary to take detailed analysis into consideration.

## CCS CONCEPTS

•Information Systems → Information Systems Application

## Keywords

Driving behavior; machine learning; neural networks

## 1. INTRODUCTION

Traffic safety is an important component of emergency management. And Driving risk prediction can effectively improve traffic safety. Traditionally, driving risk prediction has been considered as a tough question because of the difficulty in prediction factors and methods. In some insurance companies, they measure the driving risk by a few factors, like dangerous driving history or individual's age. It has been thought as an inefficient way because they treat all customers in the same way. In actual situations, every customer is a unique individual, and they drive in different ways. Consequently, they get different possibilities of the vehicle accident. Treating all different customers in same way leads to poor results, for example, the traditional strategy encourages drives to pay less attention to safety because they always get covered when they meet accidents. Also, it makes drivers with good performance feel unfair because they have to pay almost the same money as drivers with poor driving behaviors even they never filed a single claim. Therefore, the drivers' behavior may worsen, which might even deteriorate transportation safety.

In order to overcome the problems that traditional strategy has, we developed a new method to predict driving risk. The method can distinguish every individual's driving risk, based on their driving behaviors. In detail, drivers' useful driving behavior can be recorded by vehicle hardware, such as the frequency of jerk acceleration, jerk deceleration, and jerk turns. These kinds of behaviors can imply driver's driving habits, and help to evaluate the possibility of a crash and near crash. More dangerous behavior leads to a higher possibility of the car crash, and these drivers should be treated specially.

By applying the new method of driving risk prediction, it is comprehensible that the new method can bring positive excitation and encourage every driver to pay more attention to their behavior, which has a significant influence on improvement of the whole society.

In America, the earliest risk factor is gender. They think male and female have different driving risk which needs detailed analysis. However, it was forbidden after a short time, because public believe that it involves gender discrimination. Then some research shows the influence of driving data like jerk acceleration and jerk deceleration. It is a great help for risk

EM-GIS'17, November 7–10, 2017, Redondo Beach, CA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5493-6/17/11…$15.00
https://doi.org/10.1145/3152465.3152476

control. Some countries and companies use these factors to analyze drivers' driving risk and improve the driving environment. For example, a study in the Netherlands showed that the applying of risk factors can reduce 5% of fatalities and 1000 less of the injured accident in Netherlands each year (Tselentis et al. 2017).

However, for the past few decades, limited by computing methods, researchers can only use basic models to analyze risk factors, which is helpful but not accurate enough. Traditional models cannot fully utilize the potential information of data and factors. Therefore, we consider applying machine learning methods in risk prediction, which helps us improve the accuracy of our model and pursue further latent risk factors to perfect the model.

In this paper, we are trying to build a practical model to evaluate the driving behaviors using machine learning methods. We select some representative features from behavior data and build a high-accuracy model to predict the possibility of vehicle violations. The model can be applied to help correct driving behavior for poor performance drivers, and make more contribution to the improvement of transportation environment.

This paper is organized as follows. In Section 2, we present a compendium of current work on driving risk prediction and related research. We give the framework of the whole research we did in this paper in Section 3. Section 4 shows the experimental processes and results in detail. In Section 5, we display our conclusion and suggest the future work of UBI research.

## 2. LITERATUE REVIEW

Driving risk prediction has undergone the development process from unified risk assessment to personalized risk prediction. At the very beginning, the driving risk is evaluated according to the fixed information, such as the type of vehicles. However, under this mechanism, quality customers is not rewarded while non-quality customers is not penalized. Therefore, some researchers and practitioners began to develop in a non-fixed way in which driving risk is more dependent on human factors.

Human factors play a crucial role in vehicle accidents (Machin & Sankey, 2008). Both researchers and practitioners have invested great efforts in searching for a reasonable explanation of the relationship between human factors and accidents. In the beginning, the exploration for this area is focused on demographic factors, such as gender and age. For instance, research of Guo and Fang (2013) showed that young and old drivers were more likely to have an accident compared to others. However, demographic characteristics are not the essential causes of accidents. Driving behaviors are more likely to be responsible for the accidents (Ulleberg & Rundmo, 2003). Shaout and Adam (2011) analyzed the influence of speed, acceleration and fuel consumption on driving efficiency and safety. Johannes et al. (2013) showed that mileage was the strongest factor for predicting accidents. In addition to the data associated with the speed, some studies have also added vehicle engine data, such as engine speed, engine load and throttle position (Chen et al. 2015). Moreover, Shi et al. (2015) proposed that driving behaviors can be represented by throttle position, brake pressure because they cause the change of vehicle speed. Wang et al. (2015) were more concerned with instantaneous changes in driving behavior. They analyzed drivers' behaviors such as acceleration followed by deceleration, increasing accelerations or decelerations and decreasing accelerations or decelerations. Johnson and Trivedi (2011) highlighted the importance of turning in driving behavior research. What's more, Mercedes et al. (2016) used GPS to study the influence of location information for accidents.

At the same time, many research focused on algorithms in analyzing driving behaviors. Constantinescu et al. (2010) used principal component analysis to generate new variables and put them into Ward's hierarchical clustering algorithm for data analysis. McCall (2007) employed Bayesian learning method to predict braking behavior. Mudgalet al. (2014) employed hierarchical Bayesian regression to model speed of drivers. Augustynowicz (2009) classified drivers' driving risk into mild, neutral and aggressive using neural networks. Vaitkus et al. (2014) employed k-nearest neighbors to classify driving styles. Moreover, Ly et al. (2013) come up with a method using k-means and support vector machines to differentiate driving behaviors. Paefgen et al. (2013) employed logistic regression, neural networks and decision trees to predict accident risk. Their experiment results illustrated that neural networks outperformed logistic regression and decision trees in terms of accuracy while logistic regression is the most suitable classifier for its interpretability.

However, previous studies on driving behavior data were not deep enough. In order to get a better understand of the problem, this paper uses driving behaviors derived from real-world data to personalize drivers to help predict driving risk and help drivers improve their driving behaviors.

## 3. FRAMEWORK

Since driving risk prediction is a personalized strategy for users, it is necessary to make a detailed analysis of the user's driving behavior. We start from the driving data, and the data is preprocessed and combined to get the route data. Then from the route data, feature engineering is employed to get more features. Finally, the features are put into the classifier to build the model based on vehicles' records whether it violates traffic rules.

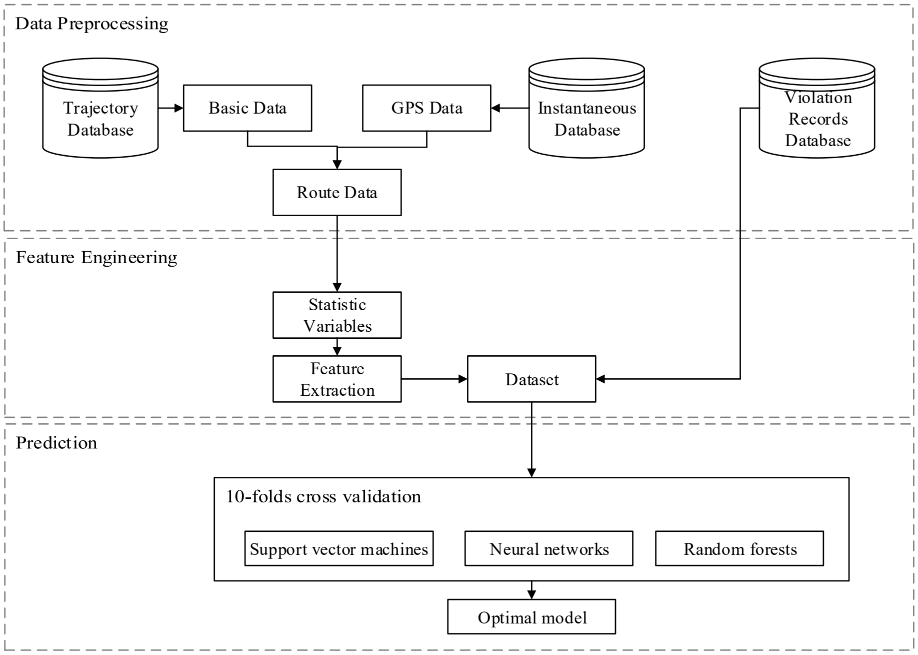


Figure 1. Framework of our proposed method

## 3.1 Data Preprocessing

The data we used is derived from multiple sources. A large amount of data will be generated by the driver in the process of driving the vehicle. In order to conduct a comprehensive evaluation of the driving behavior of drivers, we need to integrate the data from different sources. First, we do data cleaning on both trajectory data and instantaneous data and then we fill the missing values. In the data cleaning process, we mainly delete some outliers and check the consistency. Meanwhile, we uses mean imputation to fill the missing data. Cleaning data and filling in missing values can provide good data stability and foundation for subsequent model building. Then we integrate the above two kinds of data to get route data.

## 3.2 Feature Engineering

After the data preprocessing, we carried out the feature engineering on the route data. First, we make a statistical analysis on the route data. For example, we calculate the maximum, minimum and average speed of the driver. Then, we have a combination of basic attributes to facilitate the feature extraction, such as speed and night driving combination. We calculate drivers' maximum, minimum and average speed at night driving. Previous studies focused on research on basic attributes, such as speed, mileage and so on. These basic attributes can reflect the user's driving behavior and driving risk to some extent, but the lack of more detailed analysis makes the evaluation of driving risk of drivers not comprehensive. For example, if two drivers drive different miles in the day and night, they are likely to be different in terms of driving risk. Separate statistical analysis can make the description of driving behaviors more accurate.

## 3.3 Prediction

After feature engineering, we put data into classifiers to build prediction models. In this paper, we choose a variety of classifiers for comparison. Alternative classifiers include support vector machines (SVM), random forests (RF) and neural networks (NN). SVM uses the kernel function to map attributes to high-dimensional space, and then use the optimal hyperplane to maximize the distinction between drivers whether violates traffic rules. RF will integrate the results of multiple decision trees to classify the driving behaviors. Moreover, NN classify the driving behaviors by mapping the features in a non-linear way.

# 4. EMPIRICAL ANALYSIS

## 4.1 Data Description

This paper uses real-world data to validate the proposed method. The raw data consists of two parts: instantaneous data and trajectory data. Instantaneous data records data such as the time, instantaneous speed and distance traveled in the process of vehicle moving. The trajectory data records the time and GPS information during the vehicle moving. GPS data allows us to locate the vehicle's trajectory in real time, which is important in calculating the turning data that cannot be calculated only when the instantaneous data is available. In this paper, the data of vehicles' violation records is used as a criterion to judge the driving risk. Finally, we got 260 violation data and 359 non-violation data. The data attributes we eventually use are shown in Table 1. Each attribute is calculated twice by day and night

Table 1. Attributes used in the experiments

| Attributes | Description |
|---|---|
| Speed_max | Maximum speed of a vehicle |
| Speed_min | Minimum speed of a vehicle |
| Speed_avg | Average speed of a vehicle |
| Mileage_max | Maximum number of miles for trips of a vehicle |
| Mileage_min | Minimum number of miles for trips of a vehicle |
| Mileage_avg | Average number of miles for trips of a vehicle |
| Jerk_acceleration_max | Maximum jerk acceleration times a vehicle undergoes |
| Jerk_acceleration_min | Maximum jerk acceleration times a vehicle undergoes |
| Jerk_acceleration_avg | Average jerk acceleration times a vehicle undergoes |
| Jerk_deceleration_max | Maximum jerk deceleration times a vehicle undergoes |
| Jerk_deceleration_min | Minimum jerk deceleration times a vehicle undergoes |
| Jerk_deceleration_avg | Average jerk deceleration times a vehicle undergoes |
| Jerk_turns_max | Maximum jerk turn times a vehicle undergoes |
| Jerk_turns_ min | Minimum jerk turn times a vehicle undergoes |
| Jerk_turns_ avg | Average jerk turn times a vehicle undergoes |
| Fuel_consumption_max | The maximum amount of fuel consumed during vehicle running |
| Fuel_consumption_min | The minimum amount of fuel consumed during vehicle running |
| Fuel_consumption_avg | The average amount of fuel consumed during vehicle running |

## 4.2 Evaluation criteria

To evaluate the performance of the proposed method, four different criteria are applied to measure different aspects of the experiment results. TP rate (Recall) refers to the ratio of the number of violation vehicles identified by the model and the number of violation vehicles. Precision is the ratio of the number of real violation vehicles identified by the model and the number of violation vehicles judged by the model including the misjudgments, which measures the pertinence of the model. However, precision and recall are contradictory for a model to some extent, so we introduce F-measure. F-measure is the weighted harmonic average of precision and recall, which can better measure the performance of the model in a more comprehensive manner. Moreover, FP Rate measures the misjudgment of the model. The confusion matrix is shown in Table 3, and the four criteria are defined below:

$$TP\ Rate = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{1}$$

$$FP\ Rate = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{2}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{3}$$

$$F1 = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision+Recall}} \tag{4}$$

Table 2. The Confusion Matrix

| | | Predicted Class | |
|---|---|---|---|
| | | Violation | Non-violation |
| Actual Class | Violation | TP | FN |
| | Non-violation | FP | TN |

## 4.3 Experimental Results

We compare the performance of support vector machines with radial basis function as the kernel, random forests and neural networks on the dataset. In this experiment, we use data with attributes calculated separately in day and night.

Table 3 shows the results of our experiments. NN performs better than SVM and RF in all four evaluation criteria. The TP Rate of NN is 6.5% higher than that of SVM and RF, which means that NN can find more violation vehicles than the other two classifiers. In terms of FP Rate, NN outperforms SVM and RF with a lower value of 0.045, which means that the misjudgment of NN is low. The precision of NN is 0.969, which is 5.2% higher than that of SVM and 6.5% higher than that of RF. At last, as the weighted harmonic average of precision and recall, F-Measure of NN is higher than that of SVM and RF.

Table 3. Experiment results of three classifiers

| Classifiers | TP Rate | FP Rate | Precision | F-Measure |
|---|---|---|---|---|
| SVM | 0.903 | 0.134 | 0.917 | 0.901 |
| RF | 0.903 | 0.113 | 0.904 | 0.903 |
| NN | 0.968 | 0.045 | 0.969 | 0.968 |

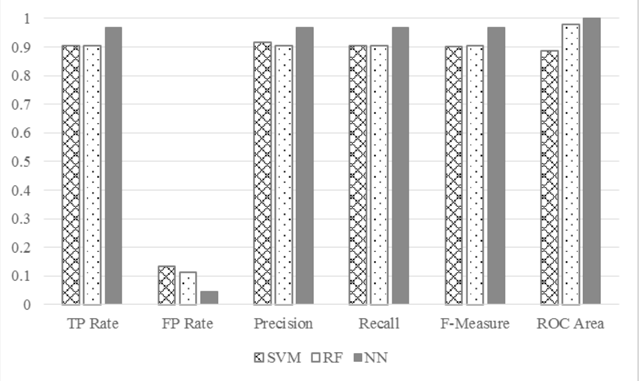


Figure 2. Comparison of experiment results of three classifiers

Therefore, from the experimental results, NN is the most suitable algorithm for the study of driving behavior in the three classifiers.

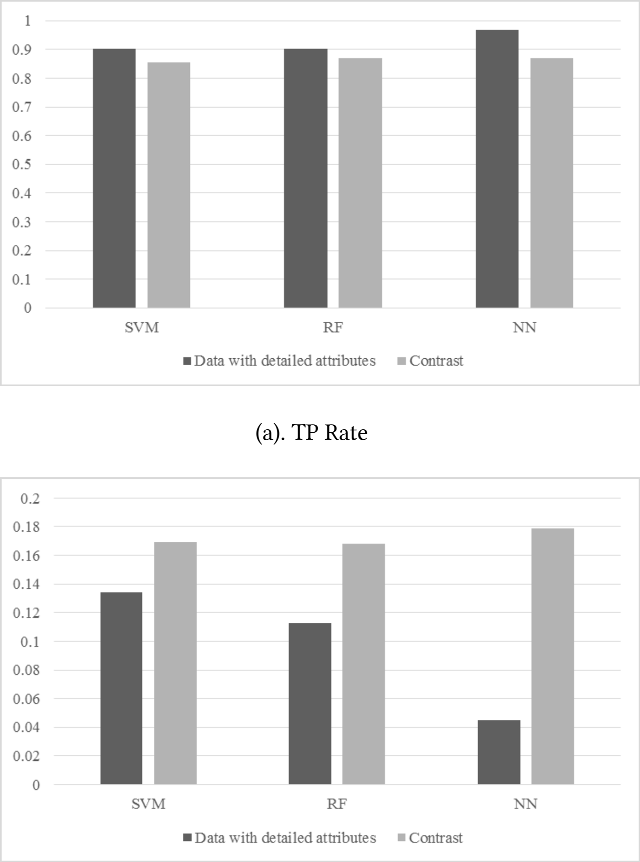
## 4.4 Comparative experiment results

In order to further validate the effectiveness of our proposed model, we conducted a comparative experiment on a different dataset. The experiment compares that in the analysis of variables, whether detailed attributes can increase accuracy of the model. Data with detailed attributes refers that we calculate the attributes separately, for instance, max speed in day and max speed in the evening are two different attributes. The contrast experiment was performed on datasets that did not distinguish driving behaviors by driving time. Table 4 shows the results of the contrast experiment.

Table 4. Experiment results of the contrast experiment

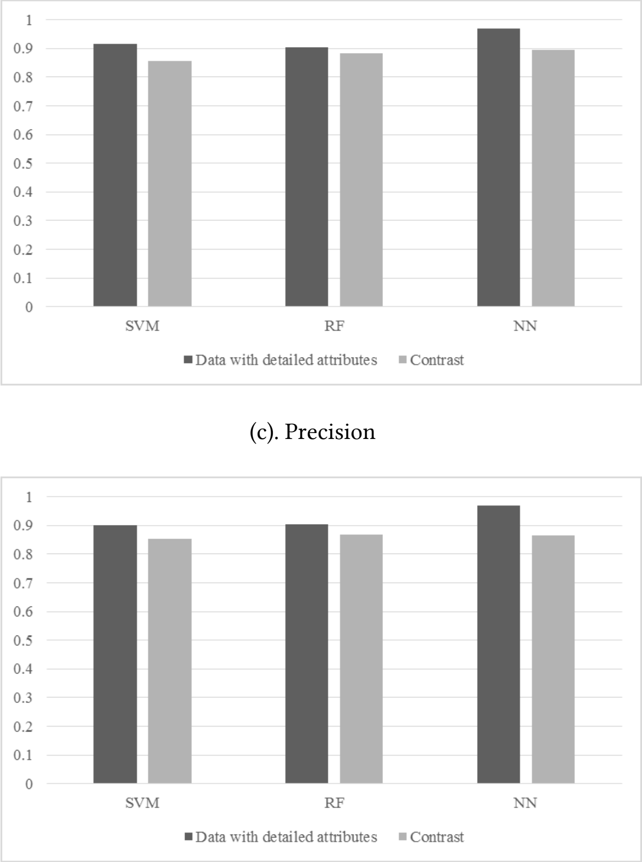
| Classifiers | TP Rate | FP Rate | Precision | F-Measure |
|---|---|---|---|---|
| SVM | 0.903 | 0.134 | 0.917 | 0.901 |
| RF | 0.903 | 0.113 | 0.904 | 0.903 |
| NN | 0.968 | 0.045 | 0.969 | 0.968 |

Figure 3 shows the comparison of results between two experiments.



(a). TP Rate

(b). FP Rate



(c). Precision

(d). F-Measure

Figure 3. Comparison of experiment results between data with detailed attribute and the contrast

All evaluation criteria show that the model with detailed attributes outperforms model without detailed attributes. Therefore, detailed analysis can make the evaluation of driving behaviors more accurate.

## 5. CONCLUSIONS AND FUTURE WORK

Our research has implied that the possibility of vehicle violation is intensely relevant to driver's behavior, and it is worth to notice that outstanding machine learning algorithms play an important role in model building. We compared three different classifiers and pick the best model NN as the most suitable prediction algorithm. With the help of our high-accuracy predict model, we are able to evaluate a driver's behavior and help identify driving risk. Under the current circumstances, it is an enormous promote for society and surely has significant influence. Moreover, it is necessary to take a detailed analysis into consideration.

For the future work, we think the diversity of data source is necessary. More data can be helpful to improve the model. Furthermore, it is important to take more machine learning methods into consideration.

## 6. ACKNOWLEDGMENTS

This research work was partly supported by National Natural Science Foundation of China (Grant No. 71771212), Hebei Social Science Fund (HB13GL021), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075, 15YJA630068), and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (No. 15XNLQ08).

## 7. REFERENCES

[1] Augustynowicz, A. (2009). Preliminary classification of driving style with objective rank method. International Journal of Automotive Technology, 10(5), 607-610.

[2] Ayuso, M., Guillén, M., and Marín, A. M. P. (2016). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. Transportation Research Part C: Emerging Technologies, 68, 160-167.

[3] Chen, S. H., Pan, J. S., and Lu, K. (2015). Driving behavior analysis based on vehicle OBD information and adaboost algorithms. Paper presented at the Proceedings of the International MultiConference of Engineers and Computer Scientists.

[4] Constantinescu, Z., Marinoiu, C., and Vladoiu, M. (2010). Driving style analysis using data mining techniques. International Journal of Computers Communications & Control, V(5), 654-663.

[5] Guo, F., & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. Accident Analysis & Prevention, 61, 3-9.

[6] Johnson, D. A., and Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. Paper presented at the 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC).

[7] Ly, M. V., Martin, S., and Trivedi, M. M. (2013). Driver classification and driving style recognition using inertial sensors. Paper presented at the 2013 IEEE Intelligent Vehicles Symposium (IV).

[8] Machin, M. A., and Sankey, K. S. (2008). Relationships between young drivers' personality characteristics, risk perceptions, and driving behaviour. Accident Analysis & Prevention, 40(2), 541-547.

[9] McCall, J. C., and Trivedi, M. M. (2007). Driver behavior and situation aware brake assistance for intelligent vehicles. Proceedings of the IEEE, 95(2), 374-387.

[10] Mudgal, A., Hallmark, S., Carriquiry, A., and Gkritza, K. (2014). Driving behavior at a roundabout: A hierarchical Bayesian regression analysis. Transportation Research Part D: Transport and Environment, 26(Supplement C), 20-26.

[11] Paefgen, J., Staake, T., and Thiesse, F. (2013). Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. Decision Support Systems, 56, 192-201.

[12] Shaout, A., and Bodenmiller, A. E. (2011). A mobile application for monitoring inefficient and unsafe driving behaviour. Paper presented at the International Arab Conference on Information Technology.

[13] Shi, B., Xu, L., Hu, J., Tang, Y., Jiang, H., Meng, W., and Liu, H. (2015). Evaluating driving styles by normalizing driving behavior based on personalized driver modeling. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(12), 1502-1508.

[14] Tselentis, D. I., Yannis, G., and Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis & Prevention, 98(Supplement C), 139-148.

[15] Ulleberg, P., and Rundmo, T. (2003). Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. Safety Science, 41(5), 427-443.

[16] Vaitkus, V., Lengvenis, P., and Žylius, G. (2014). Driving style classification using long-term accelerometer information. Paper presented at the 2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR).

[17] Wang, X., Khattak, A. J., Liu, J., Masghati-Amoli, G., and Son, S. (2015). What is the level of volatility in instantaneous driving decisions? Transportation Research Part C: Emerging Technologies, 58, Part B, 413-427.