Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Driver identification using in-vehicle digital data in the forensic context of a hit and run accident

Klara Dološ*, Conrad Meyer, Andreas Attenberger, Jessica Steinberger

Central Office for Information Technology in the Security Sector (ZITiS), Zamdorfer Str. 88, 81677, Munich, Germany

ARTICLE INFO

Article history:
Received 4 June 2020
Received in revised form
8 September 2020
Accepted 18 October 2020
Available online 30 October 2020

Keywords:
Natural driving behavior
Random forest
Validation
Time series
CAN-BUS data

ABSTRACT

One major focus in forensics is the identification of individuals based on different kinds of evidence found at a crime scene and in the digital domain. In the present study, we assessed the potential of using in-vehicle digital data to capture the natural driving behavior of individuals in order to identify them. Freely available data was used to classify drivers by their natural driving behavior. We formulated a forensic scenario of a hit and run car accident with three known suspects. Suggestions are provided for an understandable and useful reporting of model results in the light of the requirements in digital forensics. Specific aims of this study were 1) to develop a workflow for driver identification in digital forensics, 2) to apply a simple but sound method for model validation with time series data and 3) to transfer the model results to answers to the two forensic questions a) to which suspect does the evidence most likely belong to and b) how certain is the evidence claim. Based on freely available data (Kwak et al., 2017) the first question could be answered by unsupervised classification using a random forest model validated by random block splitting. To answer the second question we used model accuracy and false detection rate (FDR) which were 93% and 7%, respectively. Furthermore, we reported the random match probability (RMP) as well as the opportunity of a visual interpretation of the prediction on the time series for the evidence data in our hypothetical hit and run accident.

1. Introduction

One major task in forensics is the identification of individuals based on physical evidence found at a crime scene and also in the digital domain. We assessed the potential of using in-vehicle digital data to capture the natural driving behavior of individuals in order to identify them. As cars become increasingly reliant on sensors to perform everyday driving operations and companies as well as users begin to store their data either locally and in their cloud to e.g. optimize products, acquisition and analysis of these data is more likely to serve as evidence in court (Singleton et al. 2008; Wahab et al., 2009). In the field of digital forensics this could help to solve cases of hit and run accidents and also other activities where vehicles were involved. Enev et al. (2016), section 3.3 described a variety of interesting scenarios and thereby show the relevance of driver identification using data related to the individual's natural driving behavior. The forensic question to be answered is how likely one individual was the driver compared to all possible individuals (population). In DNA matching, camera identification or voice recognition (Goljan et al. 2009; Koehler et al. 1995; Campbell et al., 2009), relatively large databases are used to compare an observed pattern (a pattern characterizing the evidence data) to a large set of patterns and to calculate a likelihood for the individual in question compared to random individuals from the population. Experts evaluate matches among suspects and for example crime scene DNA evidence in terms of the probability of random matches across different reference populations. For forensic driver identification an analogous workflow could be established and analogous measures of evidence strength could be used (Thompson and Newman 2015).

User identification is also applied in other fields beside forensics such as e-commerce and insurance sectors (Yang 2010; Fugiglando et al., 2017; Hallac et al., 2016). This methodology can thus be transferred to a broader spectrum of digital evidence related to the behavior of individuals in the digital domain (smart home data, typing behavior). Most often machine learning is employed for such classification tasks (Lin et al., 2018; Chen et al. 2019; Fung et al., 2017). In digital forensics, practitioners can draw from the experiences in other research fields related to classification tasks especially from those based on structured data (spatially or temporally

* Corresponding author.
E-mail address: klara.dolos@zitis.bund.de (K. Dološ).

https://doi.org/10.1016/j.fsidi.2020.301090

auto-correlated data, clustered samples). Namely, not yet common in the machine learning literature on classification of individuals based on driving behavior is a validation procedure which accounts for temporal auto-correlation. Further, model results have been reported addressing preferences and needs of the data science community but not taking into account the context of forensic use cases.

In the present study we used freely available data to classify drivers by their natural driving behavior. We present a forensic scenario of a hit and run car accident with three known suspects. Suggestions are provided for an understandable and useful reporting of model results in the light of the needs in forensics. Specific aims of this study were 1) to develop a workflow for driver identification in digital forensics, 2) to apply a simple but sound method for model validation with time series data and 3) to transfer the model results to answers to the forensic questions a) to which suspect does the evidence most likely belong to and b) how certain is the evidence claim.

2. Methods

2.1. Forensic scenario: A hit and run accident

In order to show driver identification is relevant, we like to provide a clear use case. In a hypothetical hit and run accident law enforcement was able to identify the vehicle, which was involved, but not its driver. Only three known individuals C, E and I had access to the car and keys at the relevant time. In-vehicle digital data was available, e.g. provided by the insurance company as a result of a pay-as-you-drive car insurance contract. Such data provide information at high time frequency for vehicle speed, gas pedal positions, steering wheel positions, and changes of these variables. Thus, it could be possible to identify the actual driver out of the three suspects by his/her natural driving behavior calculated from electronic car data. In this forensic scenario we assume that one out of three suspects was the actual person of interest. This simplifies the calculation of the probability for the membership of the evidence pattern to the suspect classes since we know that there was no additional unknown person involved. For answering the question whether it was one of the three or an unknown individual, other methods need to be applied. In our scenario, driving samples were recorded after the criminal incident happened for each of the three suspects analogous to writing samples or voice samples and also fingerprints, approx. 40 min for each. These data were used to create a driver profile (using modeling, for example a random forest model). Using this driver profile, the person who drove the car before and after the point of time the (imagined) accident happened could be identified.

2.2. Workflow for forensic driver identification

The common workflow applied in machine learning studies aiming at classification comprises one data set which is split into training and test data. The training data subset is used for model calibration (model fitting, parametrization). The test data is considered statistically independent and is used for prediction based model quality assessment, the model validation. When this is done repeatedly (e.g. repeated data splitting, k-fold cross-validation, bootstrapping, Roberts et al., 2017) several predictions and model quality measures (e.g. accuracy, sensitivity, specificity) and their variability (e.g. CV of accuracy) can be calculated. In the end, there is a point estimate for model quality (e.g. median accuracy) and information about its variability. Additionally, for each data tuple for which a prediction is needed, all fitted models are applied and thereby a point estimate e.g. for the probability of a class membership as well as confidence intervals or the coefficient of variation can be provided (Dolos et al. 2016).

In the workflow for forensic driver identification, the machine learning procedure mentioned above needs to be applied on a suspect and population sample gathered by the police and/or stored in a larger database. After model calculation including validation, the evidence data can be treated as another hold-out sample and probabilities for the suspect being the individual of interest can be calculated. This is the quantitative basis for answering the two forensic questions formulated in the study aims and addressed in more detail in the following two sections.

2.2.1. To which suspect does the evidence most likely belong to?

In many applications of supervised classification, predicted probabilities for class memberships are sufficient. This follows the “best guess” principle and attributes each data tuple to a class based on similarity. When it comes to forensic interpretation, results are usually calculated and presented differently. Using machine learning algorithms and their predictions we can calculate the probability for the membership of the evidence pattern to the suspect class(es). This probability can be related to those of the other classes. This is comparable to a random match probability (RMP) (Koehler et al. 1995; Thompson and Newman 2015) in a finite suspect group, not including the use case where an unknown suspect is involved.

2.2.2. How certain is the claim?

In remote sensing based mapping of landscapes a pixel is classified as road, forest or grassland according to the class for which the model provides the highest probability (Lopatin et al., 2019). The map is usually considered to be useful when the accuracy is around 0.8. Accuracy is traditionally used although it can be non-informative in cases of very unbalanced numbers of class members (1000 negatives and 2 positives will give a high accuracy although a model might not be able to find any of the 2 positives). Additionally to accuracy, in forensics especially the false detection rate (FDR) could be usefull $FDR = FP / (FP + TP)$ (FP: false positives, TP: true positives). Following forensic evidence for which FDR is high would lead to a high false conviction rate. If there would be only one single piece of evidence for calculating FDR, the value could directly be translated into the percentage of persons being convicted but are in fact innocent. For example a FDR of 0.05 statistically results in a false conviction rate of 5%.

Under the assumption that the suspect samples and data base are sufficiently similar to the evidence data, uncertainty can be transferred by providing confidence intervals for the overall model quality (e.g. accuracy, FDR) and also for each prediction.

The last step of evaluation of the uncertainty of the evidence belonging to one of the suspects (or other individuals in the population), is the interpretation of the probability (the model prediction for the evidence data to belong to the suspect) or probability odds given by the model, as mentioned above. The mere class probability is not sufficient since it does not include information on the random match probability (RMP). Additionally, non-statisticians are better in interpreting a statement such as “In 1 out of 20 this match could be positive just by chance” rather than “The subject class probability is 0.85, the mean probability for all others is 0.05”. For decisions in a forensic context, it is more appropriate to use the random match probability together with the classification results.

2.3. Electronic car data

For this study we used freely available data (http://ocslab.hksecurity.net/Datasets/driving-dataset) (Kwak et al. 2017;

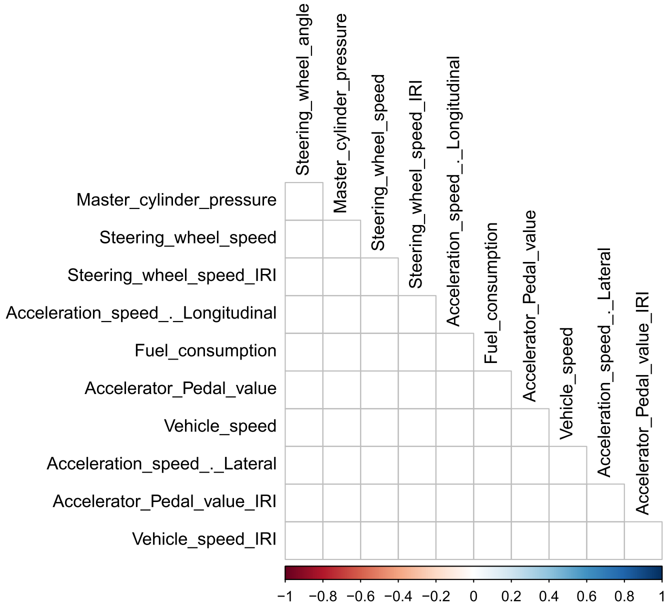

Fig. 1. Cross-correlation among features used in the model based on the time series data calculated as the maximum of all correlations for time lags up to 30 s. Some of the features were highly correlated, e.g. those related to vehicle speed (wheel velocity).

Martinelli et al., 2018). In total 10 drivers traveled between Korea University and the SANGAM World Cup Stadium in the surroundings of Seoul (South Korea) on three different road types. The number of features recorded was 51 in 1 s time intervals. Total driving time per individuals was between 121 and 184 min.

2.3.1. Driver subsets

The ten drivers were labeled from “A” to “J”. For the presented forensic scenario three drivers were considered as suspects. In order to get information on the variability of accuracy of driver identification we created data subsets for all combinations of three drivers. We completed the whole workflow for each of these 120 subsets of three out of ten drivers. Thereby it was possible to provide numbers on the variability of model quality depending on the variability of similarity among driver combinations (i.e. model quality will be lower for drivers with similar natural driving behavior than for a set with markedly different behavior).

2.3.2. Features

From all features gathered in the data (Kwak et al. 2017) some could be excluded prior to the actual feature selection due to their dependency on other factors than the driver himself, such as air temperature and the type of the car. After this step, 8 features remained for use in this study (Fig. 1).

In addition to the features in the data set that could be directly used in the model, we calculated three meta-features from the data that seemed to be related to driver behavior. Roughness was calculated from vehicle speed, namely the international roughness index (IRI, R-package rroad, IRI_COEF_250) (Simko and Laubis 2018). Analogously, we calculated the international roughness index for the features Accelerator_Pedal_value and Steering_wheel_speed.

The data was recorded in 1 s time intervals. This is very short compared to human reaction time and thus this time series data was highly auto-correlated. We expected that a reduction of the data to e.g. 3 s time intervals or even less by using median values could speed up model fitting and at the same time reduce the noise in the data. In order to test this, we created 8 data sets with 1 s intervals (the original), 2, 3, 4, 5, 10, 30, 60 s and compared model quality.

In order to characterize correlations among features, cross-correlations among all features where calculated (Fig. 1). Since this is a time series with auto-correlation, strongest correlation can occur not only at the same time but also with a time lag. We calculated the correlation among feature pairs for time lags 1 up to 30 s (cross-correlogram) and displayed the maximum.

2.4. Modeling & validation

The random forest classification method from the R-package randomForest (Liaw and Wiener 2002) was used. Random forest is a supervised learning algorithm which is used for both classification as well as regression. This algorithm can be used to classify entities such as drivers depending on recorded features characterizing them. Once a relation is established for features and classes, for each data tuple class memberships can be calculated. This is also possible for data that the algorithm has not see before. These are considered as “predictions” which is a reason why this type of modeling is also called “predictive modeling” in contrast to descriptive statistics and exploratory analysis.

In order to decide for the number of trees necessary, we plotted the error against the number of trees. This showed that 20 to 50 trees were sufficient for our study. Since we did not check on every single model of the 120 driver subsets we used 100 trees to yield a safety margin. The models were evaluated by their accuracy and false detection rate (FDR). We showed both training and test data accuracies to demonstrate the relevance of a well designed validation strategy. We also compare these numbers for different temporal aggregation of the original time series data.

Variable importance was calculated for the non-aggregated data (1 s time intervals). This value shows the contribution of the feature to the classification of the drivers compared to other features. High values mean that including the feature in the model improves the separability of classes. Importance given by the function randomForest (R-package randomForest) was used. For each of the 120 models, decrease of accuracy was ranked. For each feature, the rank sum over all models was calculated. These sums of ranks were then divided by the maximum to standardize the values between zero and one.

Each time a model was calculated, it was trained with a training data set and accuracy was calculated based on a hold-out sample. In this study, we did not repeat this step using bootstrapping or similar methods to the advantage of calculating all possible combinations of 3 drivers out of 10. We assume that aggregating the test data accuracy from these 120 models provides a good measure on variability of model accuracy.

It is particularly important to consider the strong auto-correlation and redundancy in the time series data (Bergmeir et al. 2018) caused by high frequency of measurement (1 s time intervals) compared to human reaction time and frequency of changes in the traffic and road properties. To account for the temporal auto-correlation in the time series data, first the time series for each driver was split into 8 sequences of length 5 × 60 s (length in time, not data points, same time for all time-aggregated data sets). From these starting points a stratified random sample was drawn (4 starting points for each of the three drivers). All other data remained in the test data set. We did not do any repetitions at this point, because we did this for each driver combination and assume that this will provide us with information on variability already. We referred to the forensic scenario, thus the training data set was not the larger one as usual, but restricted to a realistic

40 min driving sample which could be gathered from a known suspect. The data split was 25% training and 75% test data (+/- 1%).

2.4.1. Forensic scenario

Predictions (as probabilities for class memberships) for one random example (drivers C, E and I) were provided for each driver for test data representing the forensic data (the evidence data). Additionally the random match probability RMP = mean(p(s2), p(s3))/p(s1) for each suspect being the individual of interest was calculated. In words, RMP represents the probability that the suspect was classified as driver during the accident just by chance.

In order to demonstrate the limitations of the presented approach, we also show model predictions on test data in which there were only drivers which were not present in the training data for the corresponding model.

Response curves for one example 3-driver subset (again drivers C, E and I) were calculated and displayed in decreasing order of importance. This helps to understand the impact of the variables on model quality. It also is a way to visualize differences of natural driving behavior of individuals represented by a random forest. Note that no interactions among features were displayed and discussed for now but are certainly present in the model.

3. Results

3.1. Model quality

Median train accuracy and median test data accuracy differed strongly as expected (Table 2). Realistic validation accuracy was around 0.78 and false detection rate (FDR) was 0.1 calculated on independent test data over all 120 models.

The effect of using no validation at all (columns “Accuracy train”, “FDR train” in Table 2) and using random split validation (columns “Accuracy random split”, “FDP random split”) compared to a simple but statistically sound random block-wise validation on model quality measures became obvious. The difference in this time series for accuracy was about 0.1 to 0.2. For the FDR the difference was even higher with 0.2.

3.2. Features

Correlations among variables were high for e.g. vehicle speed and current gear as well as fuel consumption and accelerator_pedal_value, as expected (Fig. 1). Variable importance summarized over all 120 models was high for all three roughness-features. Out of the original variables vehicle speed followed by accelerator pedal value and master cylinder pressure were most important.

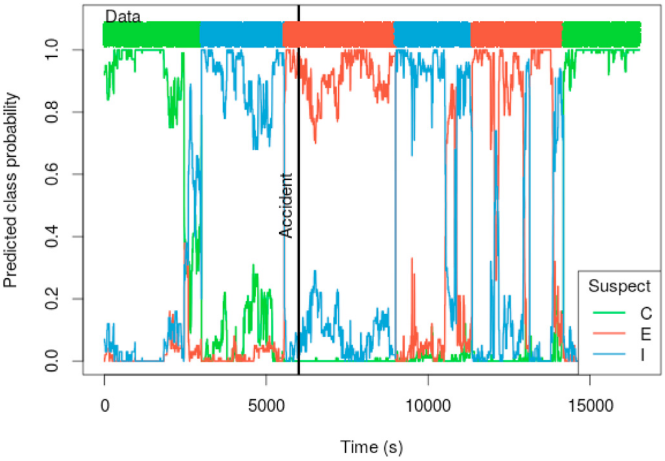

Fig. 2. Predictions on the validation data for class memberships. On the top the true class is given in the corresponding color (Data). On the bottom the predicted classes for each data point is provided (Model prediction). For most of the time series the attribution to one of the suspects is very clear. Note that the similarity between driver E and I appeared to be higher than for other pairs. Smoothed by running median with dt = 11 s. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.3. Forensic scenario

In order to provide not only summary statistics which are too abstract for a forensic interpretation, we present results for one example 3-driver combination, suspects C, E and I. The model was used to predict the probability for all three suspects for the test data (block-wise hold-out sample) which can be considered as the evidence data (Fig. 2). Most of the time there is a very clear attribution to one of the three classes. A comparison with the classes in the validation data, i.e. the truth (“Data” at top of Fig. 2) showed that this attribution was mainly correct. Model accuracy for this driver combination was 0.93, FDR was 0.07. FDRs for each suspect individually differed strongly. It was highest for C with (FDR = 0.13), and lower for E (FDR = 0.05) and I (FDR < 0.01). For E this means that 1 out of 20 data points classified as class E actually was a member of one of the other two classes.

At the time of the hypothetical accident around 6000 s, clearly driver E could be identified. This was not that clear at later points in time, e.g. from 11,000 s to 13,000 s. The model prediction on the evidence data did not allow the model for a clear and continuous separation of classes E and I during that time period. This also applied for drivers C and I around time period 3000 s.

Interpretation of the random match probability needed to be conducted together with the class prediction. When the evidence

Table 1
Importance of features summarized over all 120 models as described in the methods section. Clearly, roughness-features (IRI) were most important for separability of the drivers.

Feature	Importance	Description
Accelerator_Pedal_value_IRI	0.155	International roughness index (IRI) calculated based on the feature Accelerator_Pedal_value
Vehicle_speed_IRI	0.152	International roughness index (IRI) calculated based on Vehicle_speed
Steering_wheel_speed_IRI	0.148	International roughness index (IRI) calculated based on speed at which the driver turns the steering wheel
Vehicle_speed	0.112	Speed of the vehicle
Accelerator_Pedal_value	0.106	The degree to which the driver is depressing the accelerator pedal
Master_cylinder_pressure	0.095	The master cylinder pressure is related to the degree to which the driver is depressing the brake pedal
Fuel_consumption	0.073	The fuel efficiency of the engine
Acceleration_speed_-_Longitudinal	0.053	Longitudinal acceleration of the vehicle
Steering_wheel_angle	0.049	Angle up to which the steering wheel is turned
Acceleration_speed_-_Lateral	0.041	Lateral acceleration of the vehicle
Steering_wheel_speed	0.017	Speed at which the steering wheel is turned

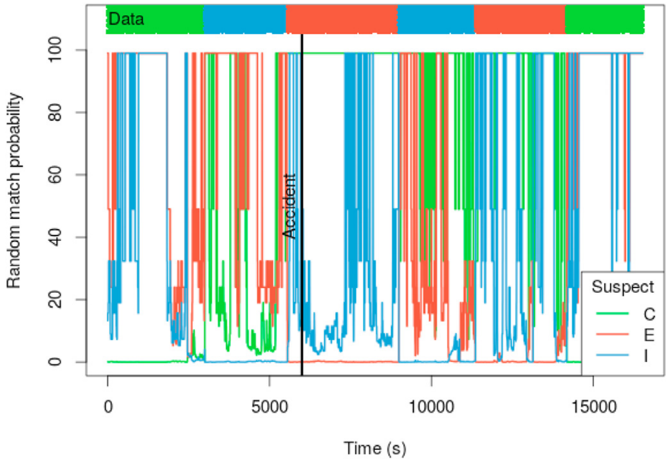

Fig. 3. Random match probability for each driver for our 3-known suspect example. The random match probability for one out of the suspect group can be calculated by the odds of the mean of probabilities predicted for the target class and each of the other classes mean(p(s2), p(s3))/p(s1). Numbers close to zero indicate a high certainty that the class attribution is not due to a random similarity between the data tuple at this point in time to the class characteristics. Smoothed by running median with dt = 11 s.

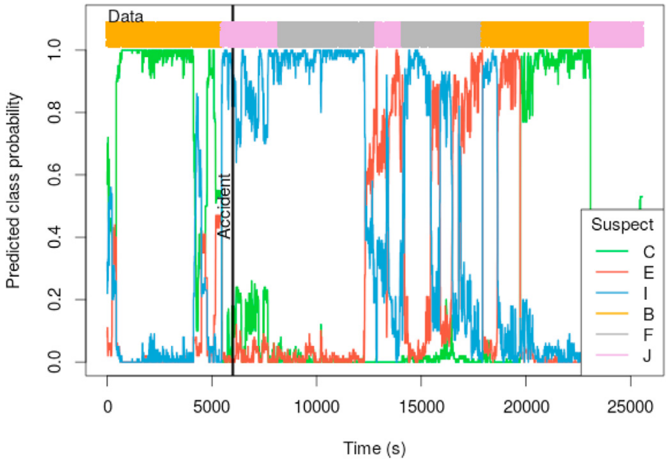

Fig. 4. Prediction with the model calibrated on C, E, I for three unknown individuals B, F, J. Random forest provides a class attribution for each data tuple to one of the classes used for calibration. This method cannot be used if the target individual is not in a group of known suspects. Smoothed by running median with dt = 11.

data at the time of the accident at time 6000 s is attributed to class E the RMP for E should be close to zero (Fig. 3) which was the case.

When the model was applied on data with drivers which were not in the calibration data (Caution: violation of model assumptions!), still high probabilities for class memberships were given by the model (but they are nonsense, *c.f.* discussion) (Fig. 4).

3.4. Response curves

For the same example (Drivers C, E, I) the response curves showed the natural driving behavior of the suspects (Fig. 5). For a classification of a data tuple as driver C, the probability was higher when vehicle speed roughness (Vehicle_speed_IRI) was higher. Also low values of the accelerator pedal value roughness (Accelerator_Pedal_value_IRI) increased the probability for class C. For class E it was the other way around. Low vehicle speed roughness combined with higher values of accelerator pedal value roughness increased the class probability for driver E.

For each feature interactions could be interpreted when aiming at describing the natural driving behavior. In this study, there was a focus on the identification, therefore we remain with this rather superficial interpretation of response curves for now.

4. Discussion

The aim of the study was to assess the possibility to identify drivers based on their natural driving behavior in the forensic context. The first question, “To which suspect does the evidence most likely belong to?“, could be answered by a supervised classification using a random forest model with 12 features (in our hit and run accident at time = 6000 s suspect E was identified as the driver). To answer the second question, “How certain is the claim?” we used model accuracy and false detection rate (FDR) which were 93% and 7%, respectively. Furthermore, we provided additional information, the random match probability (RMP), and the opportunity of a visual interpretation of the prediction on the time series hold-out sample which can be considered as the evidence data in our hypothetical hit and run accident. In the following, results are discussed in more detail and some limitations and thus interesting future research fields are identified.

4.1. Feature selection

From the 11 features we used, all three calculated roughness-features (international roughness index, IRI) were found to be most important for separability among classes over all 120 models. The roughness index considered that we dealt with time series data and that we could use not only the current value of a feature but also its change during time. Actually, changes of feature values during time appeared to better characterize natural driving behavior than absolute values. This indicated that there is still high potential to improve predictive ability of models by searching for suitable features representing the natural driving behavior not only in forensics but also in the field of individual based pricing in car insurance. Furthermore, machine learning methods especially suitable for time series or spatial data need to be developed and deployed in an easy to use manner (similar to random forest, SVMs etc.).

In contrast to other studies using the same data, we excluded clearly environment-dependent features in advance. Instead, we focused on features for which we could assume a relation to the driving behavior (see Supplement 1). Most of the features in the original data were strongly influenced by the car model and were also environment dependent due to the sampling design (for a full list of features see Martinelli et al. (2018), and Supplement 1 of our study). An example for features that were strongly related to weather conditions, i.e. outside air temperature, is intake air temperature. Another example for a feature we did not use was Engine_soacking_time, which is the duration a vehicle's engine is at rest prior to being started. Such features will not contribute to the classification of the drivers and a characterization of their natural driving behavior.

Good examples for rather person than environment related features are lateral acceleration and steering wheel speed (Table 1). There are persons with “faster” driving styles negotiating curves with higher lateral acceleration and less security margin than others (Van Winsum and Godthelp 1996; Reymond et al., 2001). The feature Master_cylinder_pressure (related to the force applied to the brake pedal) can be assumed to be strongly related to individual reaction time (the earlier the need for speed reduction is seen, the less force needs to be applied to the brake pedal) and also physical strength of the driver.

One can still argue, that weather conditions (wet/dry road)

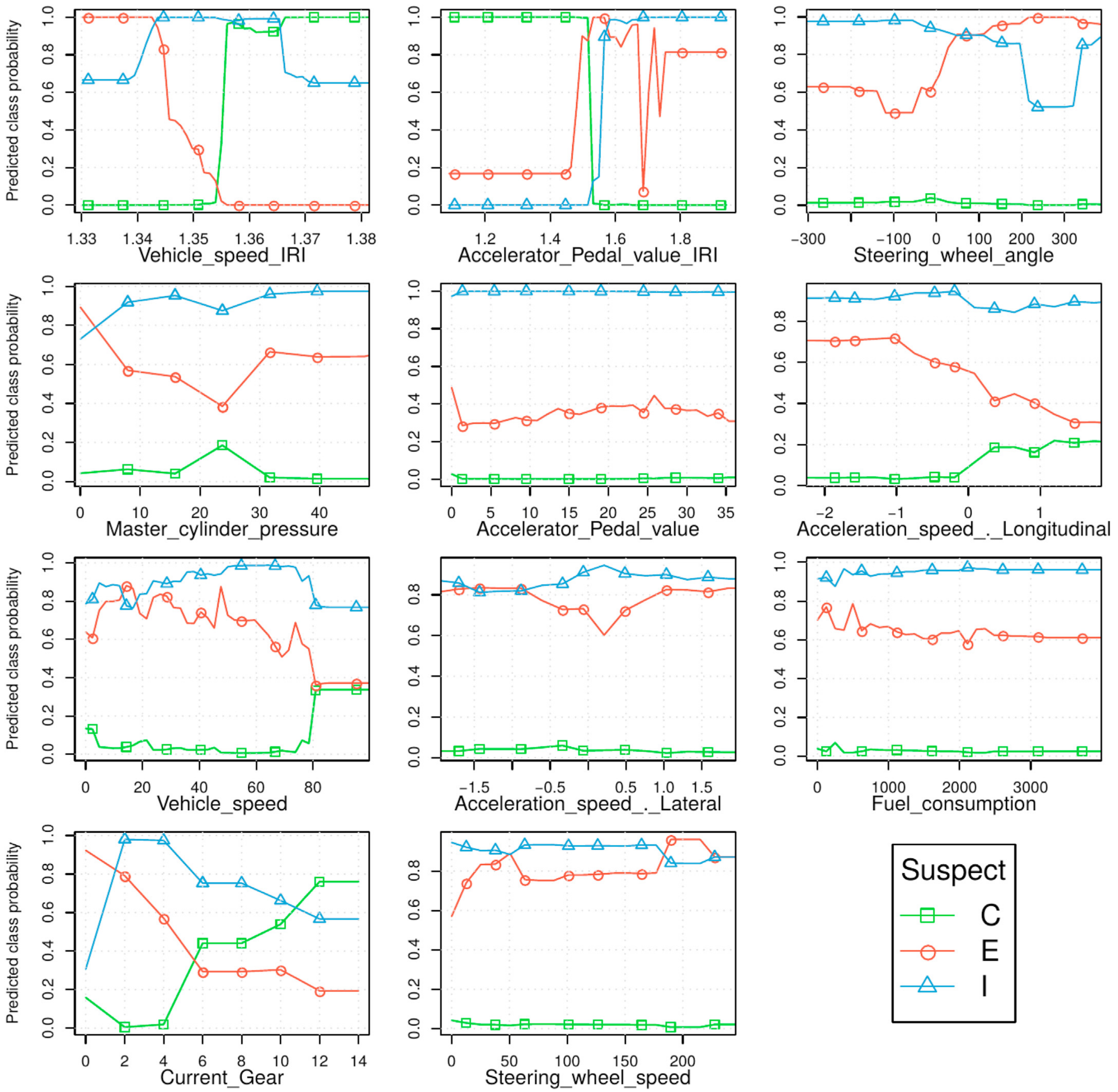

Fig. 5. Response curves calculated using partialPlot (R-package randomForest) for the example subset characterizing the natural driving behavior. Features are ordered by decreasing importance. The curves show the relative contribution of each feature to each of the three class probabilities. Note, that no figures for interactions are shown.

influence the driving style. In order to trust our model, we need to assume that a wet road compared to a dry road changes the drivers reaction time and security margin less than in comparison to another driver. Under extreme weather conditions this is maybe not the case. With a mere statistical approach it is not possible to solve these issues completely with the existing data. But it can be minimized by selecting presumably driver-dependent explaining features and a sound sampling design for the training data. Coming back to forensics, the weather conditions during the incident will be mostly known and could therefore be considered when gathering the training data.

However, why exactly did we decide to use only a small number of available feature? To understand our reasoning, three main issues need to be discussed when it comes to statistical modeling and feature selection: spurious correlation, over-fitting and multi-colinearity (Chandrashekar and Sahin 2014). All these three issues are related to the nature and the number of features chosen to remain in the model.

Using too many features and especially also environment dependent features, result in a *spurious correlation problem* which is augmented by the presence of *multi-colinearity* (Dormann et al., 2013). Under such circumstances instead of the causal explaining feature (independent variable) another correlated feature is used in the model to separate classes. This might result in seemingly good

Table 2
Overview on model accuracy at different aggregation levels (rolling mean). Accuracy/FDR train: Accuracy/FDR calculated on the training data; accuracy/FDR random split: Accuracy/FDR calculated on random split test data; accuracy/FDR random blocks: Accuracy/FDR considering temporal auto-correlation in the time series data. We decided to proceed with the non-aggregated original data because aggregation did not enhance results significantly.

dt for rolling mean (in s)	Accuracy train	Accuracy random split	Accuracy random blocks	FDR train	FDR random split	FDR random blocks
1	1	0.99	**0.78**	0	0.22	**0.01**
2	1	0.98	**0.8**	0	0.2	**0.02**
3	1	0.98	**0.785**	0	0.215	**0.02**
4	1	0.98	**0.785**	0	0.215	**0.02**
5	0.995	0.97	**0.8**	0.005	0.2	**0.03**
8	0.93	0.88	**0.78**	0.07	0.22	**0.12**
10	0.99	0.96	**0.8**	0.01	0.2	**0.04**
30	0.96	0.92	**0.8**	0.04	0.2	**0.08**

predictions as long as the correlation structure within training and test data remains the same. For example, for the training data, driver A samples were always taken at sunny and warm days whereas driver B samples were taken at rainy and cold days, just by chance. As long as the test data (incidence/evidence data) also remain with this pattern, the spurious correlation between e.g. intake air temperature and the driver will provide good results.

Furthermore, using too many features inevitably result in model *over-fitting* (Hawkins 2004) especially when using a flexible algorithm such as random forest (shapes of response curves are not restricted). A model is over-fitted, if its prediction corresponds very closely or even exactly to particular data. Therefore, it will fail to predict on independent test data. Since in real world studies there are often no completely independent test data available (data splitting procedures do usually not result in independent training and test data), it is sometimes hard to detect over-fitting directly. Therefore it is useful to define that a model is over-fitted if its predictions are no better than those of another simpler model (principle of parsimony). Still 11 features which are not completely statistically independent (shown by their correlation), potentially cause over-fitting.

The model could be improved regarding spurious correlation, over-fitting and issues with multi-colinearity, by applying a feature selection on the remaining 11 driver related features (e.g. full factorial design or forward/backward selection) (Chandrashekar and Sahin 2014). At the same time, such a study could analyze to which degree digital vehicle features are related to the driver compared to the environment (e.g. using multivariate variance partitioning). However, for this study we focused on the transfer of model results in the forensic context considering validation and model quality measures and thereby neglected this step. In a final routine on how to apply machine learning models on digital time-series user data, we recommend to include this step.

4.2. Validation

Other studies on driver classification (Martinelli et al., 2018; Chen et al. 2019) seemed to find better model accuracy up to values as high as 0.99. Unlike their work, we accounted for auto-correlation in the data and calculated the validation accuracy on test data from random block-splitting and not on random split data to account for temporal auto-correlation (Bergmeir and Benítez 2012). Accuracy and FDR need to be calculated on statistically independent validation (or “test”) data (Roberts et al., 2017; Bergmeir et al. 2018). This is mostly done by data splitting techniques, i.e. using a larger part of the data for model training and a hold-out sample for the calculation of model quality measures. In the field of driver identification a simple random data splitting is inappropriate. Non-independence of hold-out data from the training data erroneously makes models appear more reliable than they are. This is of high practical relevance when model accuracy and related model quality measures are to be used to estimate the reliability of digital forensic evidence because an over-estimation of the reliability of the model will lead to a higher rate of false convictions.

It was suggested to aggregate the data to a 3 s time period because model accuracy increased at this aggregation level. We found no indication for this. These results could still be interesting when developing methods to monitor CAN-BUS data and to save storage as well as computation time. We suggest that if computing resources are scarce data for 3 s time intervals could be sufficient for the purpose of forensic driver identification.

4.3. Forensic scenario

For the application of classification in the context of forensics, we suggested to use the false detection rate (FDR, also called fall-out) as model quality measure in addition to accuracy because it can be directly translated into the false conviction rate. Especially the individual FDR for the suspect in question (e.g. suspect E) could provide valuable information for the decision makers on the risk of false conviction. Providing this measure for each suspect individually and not a single number for the entire model, helps forensic staff without a background in statistics to evaluate the degree of uncertainty of the classification.

Furthermore, we presented the results of driver identification in terms of a time series of predicted class probabilities. We added the random match probability because it is an established way of presenting e.g. matching DNA, shoe print evidences and finger print pattern (Thompson and Newman 2015), thus decision makers might already be familiar with its interpretation. By presenting the time series prediction, we made use of the nature of the data to improve our ability to evaluate the reliability of the class prediction for a certain point in time. It is highly unlikely that the driver changed when there is a clear attribution during 10 min. In contrast, when the predicted probability and the class with highest probabilities changes every 2 min or even randomly with high frequency, the model should not be considered reliable.

However, it is still necessary to develop a sound method for combining model uncertainty of different levels. The overall model accuracy and FDR need to be considered when interpreting point estimates of probabilities together with the point estimate of the random match probability. Additionally, there is a need to calculate confidence intervals for each point estimate which could be calculated using bootstrapping. Our study represents a groundwork for developing a quantitative workflow for the application of machine learning and the interpretation of results in digital forensics.

4.4. Natural driving behavior

Models used to characterize the natural driving behavior of a person should also be evaluated based on the plausibility of the response curves, which represent the individual's driving styles. These response curves also reveal limitations of the model approach. In this study, the way of data splitting (separation of training and test data) needed to be considered. Each full time series was split into blocks. From these blocks, four were randomly drawn and used as training, others as test data. If for one driver highway was missing in the training set, this driver would be characterized as a slow driver compared to the others. Thus, response curves in this particular study design needed to be interpreted together with the distribution of speed limits during the journey (which we did not have). This effect could be reduced by a statistically informed study design. Since this is not feasible in the field, additional information on e.g. speed limits (e.g. using geo-data), could be recorded to avoid such artifacts caused by unbalanced data.

Response curves could also be used together with the evidence data to further describe why the data of a the time period of the accident was attributed to a certain class, for example which features were important. This could strengthen the credibility of the method.

4.5. Limitations in the forensic context

Except all typical limitations of data driven modeling, two main limitations for application of predictive modeling in the forensic context need to be mentioned. First, predictions of a random forest algorithm are only meaningful when all suspects were represented in the training data. Otherwise the random forest classifier will predict high probabilities for class memberships because of its nature. Each tree will still suggest one class. Probabilities for each class memberships are then calculated by the odds of votes for all trees. Methods for use cases with unknown suspect need to be developed (Kang et al. 2019).

Secondly, in forensics training data might not be valid. The driver could be in a different physical or mental state during recording the training data compared to the time the accident happened. This could be caused by an agitated mental state, a medical incident, fatigue or alcohol as well as deliberate pretending. Especially stressed drivers could be over-represented in accidents, since their mental state may be the cause of the accident. Thereby, more aggressive drivers are in danger of having to stand in for “normal” drivers who were agitated and therefore caused the accident.

Variation in driving behavior of one person in different mental states is a serious challenge for the application of driver classification in the forensic context. Therefore, we think that a study on the influence of confounding conditions on the behavior and driving style of one person is a necessary next step. One could record in-vehicle digital data together with e.g. data provided by a fitness watch. The drivers could be “stressed” by noise during driving or physical exercises before the driving sample is taken. Although this can only approximate reality such experiment could be used to learn more about the variability in specific features and driving maneuvers such as turning and stopping (Fung et al., 2017). There is the chance that the way a person takes curves can be recognized also under exceptional conditions.

5. Conclusion

As the amount of digital evidence related to behavior is likely to increase in future, methods for processing this data need to be developed. On the one hand, forensic methods which lack sound statistical foundation are not acceptable (Arshad, Jantan, and Abiodun 2018). Therefore, we need to test suggested methods systematically. On the other hand, we need to acknowledge that digital forensic science is a very young discipline compared to other natural science disciplines.

In this study we developed a workflow analogous to established practices for identification of individuals based on evidences such as DNA and finger prints for driver identification. We used an appropriate model quality measure to represent the false conviction rate (false detection rate, FDR) and made the effort to calculate the random match probability (RMP) already used in the forensic expert community. We also pointed out, that a statistically sound validation procedure, especially when it comes to the evaluation of evidence, needs to be employed. A block-wise random sampling of training and test data accounting for auto-correlation in the time series data is highly recommended. This study has shown that it is possible to adjust the reporting of model results and their quantitative evaluation to the special needs of forensics.

When we discussed our results we still found some obstacles in the way of using statistical driver identification directly as evidence in court such as the physical-mental state of the individual during the accident and within-subject variability in driving style. These two factors need to be studied before a final decision on how to use driver identification in the forensic context. Furthermore, the presented model clearly pointed to the need to assess the contribution of available features to the classification. Without a deeper understanding of the influence of the driving behavior on each feature it will not be possible to build trustworthy models. With a smart experimental design, data could be generated which could help to separate the effect of the environment (traffic, weather) and the driver and at the same time assess the within-subject variability in natural driving behavior.

Despite of all these challenges typical for a young science and the development of novel methodology, we see the chances in the field of identification of subjects using time series data on digital behavior in using vehicles, smart homes and computer keyboards. Especially in light of increasing numbers of crimes in the digital domain, we need to develop methods that enable prosecution.

Declaration of competing interest

The authors declare that they have no conflicts of interests.

Acknowledgments

The authors like to thank Byung Il Kwak, Ji Young Woo, and Huy Kang Ki for releasing the driving data for academic research (Kwak et al. 2017). Furthermore, I like to thank the organizers and members of the EU COST Action CA17124 “Digital forensics: evidence analysis via intelligent systems and practices” (DigForASP) for supporting this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fsidi.2020.301090.

References

Arshad, Humaira, Aman Bin Jantan, Oludare Isaac Abiodun, 2018. Digital forensics: review of issues in scientific validation of digital evidence. Journal of Information Processing Systems 14 (2), 346–376. https://doi.org/10.3745/JIPS.03.0095.

Bergmeir, Christoph, Benítez, José M., 2012. On the use of cross-validation for time series predictor evaluation. Inf. Sci. 191 (May), 192–213. https://doi.org/10.1016/

j.ins.2011.12.028.

Bergmeir, Christoph, Hyndman, Rob J., Koo, Bonsoo, 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput. Stat. Data Anal. 120 (April), 70–83. https://doi.org/10.1016/j.csda.2017.11.003.

Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.-F., Matrouf, D., 2009. Forensic speaker recognition. IEEE Signal Process. Mag. 26 (2), 95–103. https://doi.org/10.1109/MSP.2008.931100.

Chandrashekar, Girish, Sahin, Ferat, 2014. A survey on feature selection methods. Comput. Electr. Eng. 40 (1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chen, W., Lin, Y., Chen, W., 2019. Comparisons of machine learning algorithms for driving behavior recognition using in-vehicle CAN bus data. In: In 2019 Joint 8th International Conference on Informatics, Electronics Vision (ICIEV) and 2019 3rd International Conference on Imaging. Vision Pattern Recognition (IcIVPR). https://doi.org/10.1109/ICIEV.2019.8858531, 268–73.

Dolos, Klara, Tobias, Mette, Wellstein, Camilla, 2016. Silvicultural climatic turning point for European beech and sessile oak in western europe derived from national forest inventories. For. Ecol. Manag. 373 (August), 128–137. https://doi.org/10.1016/j.foreco.2016.04.018.

Dormann, Carsten F., Jane Elith, Bacher, Sven, Buchmann, Carsten, Carl, Gudrun, Gabriel, Carré, Jaime, R., Marquéz, García, et al., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Enev, Miro, Takakuwa, Alex, Koscher, Karl, Kohno, Tadayoshi, 2016. Automobile driver fingerprinting. Proceedings on Privacy Enhancing Technologies 2016 (1), 34–50. https://doi.org/10.1515/popets-2015-0029.

Fugiglando, Umberto, Massaro, Emanuele, Santi, Paolo, Milardo, Sebastiano, Abida, Kacem, Stahlmann, Rainer, Netter, Florian, Ratti, Carlo, 2017. Driving behavior analysis through CAN bus data in an uncontrolled environment. ArXiv: 1710.04133 [Physics], October. http://arxiv.org/abs/1710.04133.

Fung, Nathanael C., Wallace, Bruce, Chan, Adrian D.C., Goubran, Rafik, Porter, Michelle M., Marshall, Shawn, Frank, Knoefel, 2017. Driver identification using vehicle acceleration and deceleration events from naturalistic driving of older drivers. In: In 2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 33–38. IEEE, Rochester, MN, USA. https://doi.org/10.1109/MeMeA.2017.7985845.

Goljan, Miroslav, Fridrich, Jessica, Filler, Tomáš, 2009. Large scale test of sensor fingerprint camera identification. In: Edward J. Delp III, Jana Dittmann, Nasir D. Memon, and Ping Wah Wong, 72540I. https://doi.org/10.1117/12.805701. San Jose, CA.

Hallac, David, Sharang, Abhijit, Stahlmann, Rainer, Lamprecht, Andreas, Huber, Markus, Roehder, Martin, Sosic, Rok, Leskovec, Jure, 2016. Driver identification using automobile sensor data from a single turn. In: In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, Rio de Janeiro, Brazil. https://doi.org/10.1109/ITSC.2016.7795670, 953–58.

Hawkins, Douglas M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44 (1), 1–12. https://doi.org/10.1021/ci0342472.

Kang, Yong Goo, Park, Kyung Ho, Kim, Huy Kang, 2019. Automobile theft detection by clustering owner driver data. ArXiv:1909.08929 [Cs, Stat]. https://doi.org/10.13154/294-6675.

Koehler, Jonathan J., Chia, Audrey, Samuel, Lindsey, 1995. The random match probability (RMP) in DNA evidence: irrelevant and prejudicial? Jurimetrics 35, 201–219.

Kwak, Byung Il, Woo, JiYoung, Kim, Huy Kang, 2017. Know your master: driver profiling-based anti-theft method. ArXiv:1704.05223 [Cs], April. http://arxiv.org/abs/1704.05223.

Liaw, Andy, Wiener, Matthew, 2002. Classification and regression by RandomForest. R. News 2 (3), 18–22.

Lin, Xin, Zhang, Kai, Cao, Wangjing, Zhang, Lin, 2018. Driver evaluation and identification based on driving behavior data. In: In 2018 5th International Conference on Information Science and Control Engineering (ICISCE). Zhengzhou: IEEE. https://doi.org/10.1109/ICISCE.2018.00154, 718–22.

Lopatin, Javier, Dolos, Klara, Teja Kattenborn, Fabian, E., Fassnacht, 2019. “How canopy shadow affects invasive plant species classification in high spatial resolution remote sensing.” edited by ned horning and dolors armenteras. Remote Sensing in Ecology and Conservation. https://doi.org/10.1002/rse2.109. February.

Martinelli, Fabio, Mercaldo, Francesco, Orlando, Albina, Nardone, Vittoria, Santone, Antonella, Kumar Sangaiah, Arun, 2018. Human Behavior Characterization for Driving Style Recognition in Vehicle System. Computers & Electrical Engineering, January. https://doi.org/10.1016/j.compeleceng.2017.12.050. S0045790617329531.

Reymond, Gilles, Kemeny, Andras, Droulez, Jacques, Berthoz, Alain, 2001. Role of lateral acceleration in curve driving: driver model and experiments on a real vehicle and a driving simulator. Hum. Factors: The Journal of the Human Factors and Ergonomics Society 43 (3), 483–495. https://doi.org/10.1518/001872001775898188.

Roberts, David R., Bahn, Volker, Ciuti, Simone, Boyce, Mark S., Elith, Jane, Guillera-Arroita, Gurutzeta, Hauenstein, Severin, et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929. https://doi.org/10.1111/ecog.02881.

Simko, Viliam, Laubis, Kevin, 2018. Rroad: road condition analysis. https://CRAN.R-project.org/package=rroad.

Singleton, Nathan, Jeremy Daily, Manes, Gavin, 2008. Automobile event data recorder forensics. In: In Advances in Digital Forensics IV, vol. 285. Springer US, Boston, MA, pp. 261–272. https://doi.org/10.1007/978-0-387-84927-0_21. Indrajit Ray and Sujeet Shenoi.

Thompson, William C., Newman, Eryn J., 2015. Lay understanding of forensic statistics: evaluation of random match probabilities, likelihood ratios, and verbal equivalents. Law Hum. Behav. 39 (4), 332–349. https://doi.org/10.1037/lhb0000134.

Van Winsum, Wim, Godthelp, Hans, 1996. Speed choice and steering behavior in curve driving. Hum. Factors: The Journal of the Human Factors and Ergonomics Society 38 (3), 434–441. https://doi.org/10.1518/001872096778701926.

Wahab, A., Chai, Quek, Tan, Chin Keong, Takeda, K., 2009. Driving profile modeling and recognition based on soft computing approach. IEEE Trans. Neural Network. 20 (4), 563–582. https://doi.org/10.1109/TNN.2008.2007906.

Yang, Yinghui (Catherine), 2010. Web user behavioral profiling for user identification. Decis. Support Syst. 49 (3), 261–271. https://doi.org/10.1016/j.dss.2010.03.001.