Driver Behaviour Modelling: Travel Prediction using Probability Density Function

Alexey Uglanov[1], Kirill Kartashev[1], Felician Campean[1], Aleksandr Doikin[1], Amr Abdullatif[1], Emanuele Angiolini[2], Chunxing Lin[1], and Qichun Zhang[1]*

[1] Faculty of Engineering and informatics, University of Bradford, BD1 7DP, UK
[2] Jaguar Land Rover, UK
*q.zhang17@bradford.ac.uk

Abstract. This paper outlines the current challenges of driver behaviour modelling for real-world applications and presents the novel method to identify the pattern of usage to predict upcoming journeys in probability sense. The primary aim is to establish similarity between observed behaviour of drivers resulting in the ability to cluster them and deploy control strategies based on contextual intelligence and data-driven approach. The proposed approach uses the probability density function (PDF) driven by kernel density estimation (KDE) as a probabilistic approach to predict the type of the upcoming journey, expressed as duration and distance. Using the proposed method, the mathematical formulation and programming algorithm procedure have been indicated in detail, while the case study examples with the data visualisation are given for algorithm validation in simulation.

Keywords: driver behaviour modelling, probability density function, kernel density estimation, probabilistic predictions

1 Introduction

Driver behaviour modelling (DBM) becomes an important research topic in recent decades. The development of technology processes large volumes of information, which resulted in moving from mathematical models on paper like in [10] dissertation to profound computer-simulated ones [7] with large datasets [9].The digitalisation of the vehicle has made it possible to collect various performance indicators, including eco-score and others [1], which describe the type of journeys vehicle operator tends to do. The growing data availability provides the scope for characterising patterns of usage and driver similarity identification, which is a highly demanded research direction with a variety of applications starting from product design and development to system optimisation.

Interest in automotive data processing and journey patterns has been shown for a long time. Car manufacturers, such as Daimler AG [5], Ford [8] and Volvo Group [7], Jaguar Land Rover [3], taxi companies [11] and others are supporting the researchers with real-world data from the field. Public organisations are contributing to open source, providing large datasets (e.g., Taxi Trip Data used

in [2]). Nowadays the goal of DBM is getting the highest prediction rate where possible, as it is a key to successful vehicle control strategy optimisation. Collecting a large sample of the dataset is not enough by itself to solve the problem of DBM as the data set is always limited in practice. In our study, we follow ideas taken by this approach working with limited data provided by an automotive company.

From the practical application viewpoint considered in this work, a particular interest is to predict the type of the upcoming journey, which can be expressed as duration. Therefore, this paper outlines the approach based on the historical usage records to classify the type of the journey. To do so, the probability density function (PDF) driven by kernel density estimation (KDE) is used as a statistical prediction model. Therefore, the aim of the study is to develop the framework which allows predicting the duration (short / long), based on driver's trips history, for example, using the start time of a journey.

The paper is organised into the following sections. An overview of the related works on the same topic and approach applications is presented in Section 2. Section 3 provides the applied methodology, including the data used for predictions, the mathematical framework, and algorithms' details. In Section 4 real-data case examples are given, as well as an analysis of the model performance and results. Finally, Section 5 concludes the paper with a discussion of the significance of the results and future research questions.

2 Related Works

Most of the research carried out in DBM work nowadays with the large amounts of data, collected from the various sensors. [9] is conducting on-road data since 1999 and apply Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) and Bayesian Networks (BN) to it.

The earlier works were aimed at modelling driver behaviour in road traffic. [5] uses the road traffic sensors as input to the machine learning models. The paper covers two formulations of the problem: (1) predicting the speed of the vehicle at different parts of the route and (2) the transit time, with more emphasis on the second approach, using the k-nearest neighbour method. The research led to the conclusion that such features as the time and day of the journey are important for the prediction of the trip duration, as well as the rush hour effect.

Precise vehicle trajectory data is used by Toledo [10], who produced the model based on mandatory lane changing (MLC) and discretionary lane changing (DLC) used together. The paper covers DBM in traffic and in relation to other drivers.

With the growth of the data produced by the vehicle itself, the focus of the research has shifted to the production of the models for each particular driver based on his historical data, storing and using as many features as possible. [7] analyses the near-crash behaviour of heavy trucks in a developed simulation-based system. [11] and [2] work with taxi companies' data (first ones produce the individual model for each vehicle, the latter use Big Data Deep Learning

model), including the time of day, day of the week, GPS values. Day of the week influences the driving patterns of people using their vehicle to get to / from work, resulting in different behaviour on weekdays / weekend.

GPS data is also used in [6], who developed a Markov chain based on the dataset. The paper first suggests using the trip purpose derived from open-source data as an input of the destination prediction model.

3 Methodology

3.1 Case Study Data

The real-world dataset was provided by an automotive company, which contains information on car usage expressed as a statistical summary for each journey. The available data covers 3 months of usage and includes records of journey parameters (such as start time) and driver behaviour characteristics from over 300 vehicles. Data preparation for the modelling included an integrity check to satisfy data quality requirements. In this paper, the journey start time is considered as the main predictor to classify the type of the journey, assuming that this information is available at the beginning of each trip. Since this parameter is recorded in GMT format, it has been converted to decimal format to facilitate mathematical operations. Fig. 1 provides a visual example of statistical trip information collected for each vehicle.

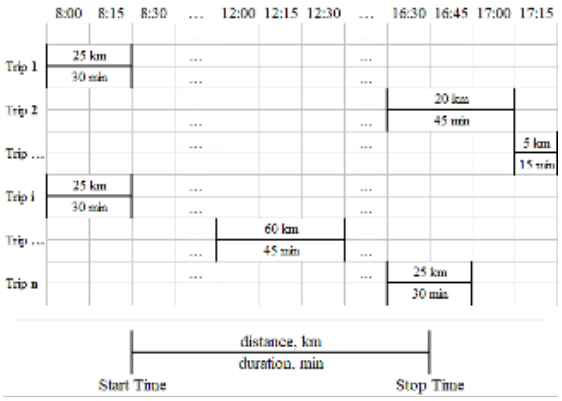

Fig. 1. Data stored for each vehicle

3.2 Mathematical Framework

Assuming that the behaviour of the driver tends to be repetitive (there should be some pattern like travelling to/from work, shop, lifting children to school or

even being a taxi driver), the chosen features should have correlation and their joint Probability Density Function (PDF) can be used to predict the future trips.

In this paper, the standard variable of human operation has been considered such as the speed, duration, eco score, journey duration, distance, trip number, etc. Suppose that the historical data set for these variables are available regarding to each individual driver. Therefore, the statistical historical recordings are available by averaging the values. Therefore, we can simply predict the probability of next trip subjected to many variables due to the fact that the driving habit is not time-variant.

In particular, the historical data can be used to generate the joint PDF to describe the relationship between the variables. Then the prediction in probability can be calculated using the generated joint PDF as follows:

$$Pr\left[a \leq x_d \leq b\right] = \int_a^b \gamma\left(x_d, x_s, x_e, x_t\right) dx_d \tag{1}$$

where $\gamma\left(\cdot\right)$ denotes the joint probability density function while x_d, x_s, x_e, x_t stand for the distance, speed, eco score and journey starting time, respectively. In this equation, a and b are the pre-specified interval for the investigated variable. Note that the formula can be used to predict all the other variables of the joint PDF.

The prediction problem has been converted as the estimation of joint PDF, where the KDE can be introduced into this approximation as follows.

$$\hat{\gamma}\left(\bar{x}\right) = \frac{1}{hn}\sum_{i=1}^{n} G\left(\frac{\bar{x} - x_i}{h}\right) \tag{2}$$

where $h > 0$ denotes bandwidth, $G\left(\cdot\right)$ denotes Gaussian kernel function and its dimension is equal to the dimension of the parameter vector $\bar{x}$, while $\bar{x} = [x_d, x_s, x_e, x_t]$ to simplify the expression. x_i denotes the historical data and n is the size of the data.

Following the aforementioned KED, there are two main challenges for journey prediction application. One of them is the high-dimensional explosion. In particular, the high-dimensional kernel function would lack of accuracy. In addition, the more variables we use the more information we obtain for the prediction. However, more variable will lead to high-dimension of the joint PDF. To solve this problem, the dimension reduction has to be considered. Basically, almost all these variables are not independent probability sense. Therefore, the correlation analysis can be done to describe the couplings between the data sets where Pearson correlation coefficient [4] for each two variables can be estimated using the collected historical data.

$$r_{x_d,x_t} = \frac{\sum_{i=1}^{n}\left(x_{d,i} - \bar{x}_d\right)\left(x_{t,i} - \bar{x}_t\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{d,i} - \bar{x}_d\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{t,i} - \bar{x}_t\right)^2}} \tag{3}$$

where $\bar{x}_d$ and $\bar{x}_t$ denote the mean value of journey distance and journey starting time. Similarly, a correlation matrix can be produced which is symmetric and the elements of the matrix is data-estimated Pearson correlation coefficient.

Another challenge is timeliness. The size n in Eq.(2) increase along collecting of the data which increases the computational complexity. Also, the historical data cannot promptly reflect the recent change of the driving habit. For example, the driving pattern will be partially changed if the driver change the job or location. Therefore, we cannot use the full historical data for the joint PDF estimation, the pre-defined sliding window can be adopted. Combining the idea of dimension reduction with KDE, Eq.(2) can be rewritten as follows.

$$\hat{\gamma}_{k_s}\left(\bar{x}_r\right) = \frac{1}{hk_s} \sum_{i=n-k_s+1}^{n} G\left(\frac{\bar{x}_r - x_{r,i}}{h}\right) \tag{4}$$

where k_s denotes the size of the sliding window, $\bar{x}_r$ and $\bar{x}_{r,i}$ stand for the dimension-reduced vector with selected dominant variables and their samples, respectively.

To highlight the recent pattern and distinguish the resent change from the historical data, a forgetting mechanism can be used for KDE by adding a forgetting factor $0 \leq \omega \leq 1$. Thus the joint PDF based prediction can be modelled by the following equations.

$$Pr\left[a \leq \bar{x}_{r,i} \leq b\right] = \int_a^b \hat{\gamma}_\omega\left(\bar{x}_r\right) d\bar{x}_{r,i}, i = 1, 2, \ldots \tag{5}$$

$$\hat{\gamma}_\omega\left(\bar{x}_r\right) = (1-\omega)\hat{\gamma}_{k_s}\left(\bar{x}_r\right) + \omega\hat{\gamma}\left(\bar{x}_r\right) \tag{6}$$

where Eq.(5) is the prediction equation and Eq.(6) is the information-updating equation.

3.3 Algorithm

The utilisation of the described algorithm work is illustrated in Fig. 2. The model takes as an input the train set, which contains the first 80% of the original dataset trips (sorted in chronological order from oldest to newest). The trips left (20% of the data) form the test set, on which the trained model will be verified. After the pre-processing step, data is prepared for a binary classification problem. The test sequence is transformed into binary one with output results where 1 is a "long" trip and 0 stands for a "short" trip based on a given m threshold. It is also used as an input to the trained model to get the binary sequence of the same size. These sequences are compared to determine the model's performance quality and construct the table of confusion for the subsequent analysis. The result is validated and, after threshold adjustment (if needed), is added to the database.

To summarise the presented concept and framework, the algorithms have been outlined as following Algorithms 1 and 2.

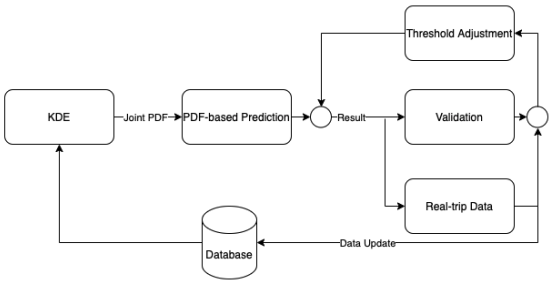

Fig. 2. Data processing flowchart

Algorithm 1 Pseudo code for implementing the presented PDF-based prediction

Require: The driver data is pre-stored in the database.
Input: The journey starting time
Output: The binary variable to describe the type of trip duration.

Initialisation: Setup the sliding window, forgetting factor, upper bound and lower bound of the investigated duration interval, threshold for probability
for $i = 1 \rightarrow n$ **do**
  if $i \geq n - k_s + 1$ **then**
    Generating the joint PDF with recent data
  end if
  Generating the joint PDF with the full data set
end for
Obtaining $\hat{\gamma}_\omega$ with for forgetting factor.
Obtaining the function by substituting the starting time to $\hat{\gamma}_\omega$.
Calculating the probability using upper bound and lower bound of the investigated duration interval.
if probability>threshold **then**
  prediction binary value =1 as long duration
else
  prediction binary value =0 as short duration
end if

Algorithm 2 Pseudo code for validating the presented PDF-based prediction

Require: The prediction developed by Algorithm 1
Input: Prediction binary value and the real time data can be accessed
Output: The updated threshold and updated database

```
Initialisation: Setup the accuracy requirement of the prediction.
if The real trip is 'long' duration then
else if prediction value = 1 then
  Insert the positive result into the validation statistical analysis.
else
  Insert the negative result into the validation statistic analysis.
end if
Updating the database by inserting the current real trip information.
if validation statistic meets the designed accuracy then
  decrease the threshold to release the conservativeness of the prediction.
else
  increase the threshold to make the prediction conservative
end if
```

4 Results and Analysis

4.1 Case Studies

The above-mentioned algorithms were applied to two vehicles' datasets A and B with identical structure (start time, duration of the journey). The available 339 and 798 respectively historical trips were separated into training and testing sets on proportion 80%/20% (270/69 vs. 637/161). KDE algorithm runs on train set and produces the smooth PDF (see Figures 3 and 4 for the given datasets).

In the case of a vehicle A (Fig. 3) the Smooth PDF is quite dispersed. There are two peaks at around 8pm with the length of the trips less than 10 and 20 minutes, several ones at 6am mostly lasting for 20-40 minutes and some trips of different duration close to midday.

Fig. 4 shows prominent peaks for the vehicle B at around 10am and 6pm with a duration of less than 15 minutes. There are almost no other trips which can be considered as noise. In this case, we expect the model to predict the short journeys at time close to that and lower and long journeys in other cases. The smooth PDF is then "cut into slices" for the whole day (24 hours) with a 30-minute gap to form a matrix for future predictions. For each of the 'time zones', the binary result is stored. Some of the time zone slices of vehicles A and B are presented in Figures 5 and 6.

There is an emerging pattern for the vehicle A peaks which were mentioned above. The trips in the morning (6am) tend to be long ones (above 20 minutes), while the evening journeys (at around 20pm) are shorter (below 20 minutes). This observation helps to identify the optimal threshold m=20 for the prediction model. It also provides the next hypothesis to test later: these are the trips from/to the same place (so the distance covered should be equal), and the duration of the trip in the morning / evening is affected by the rush hour. Vehicle

B has the highest peaks at 10pm and 18pm both with the duration of less than 10 minutes, and a smaller one close to 14pm longer than 15 minutes. Here the model's threshold separating the short / long journeys is m=15 minutes.

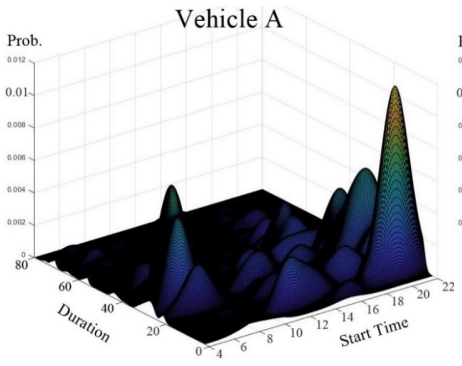

Fig. 3. Vehicle A Smooth PDF (Start Time & Trip Duration)

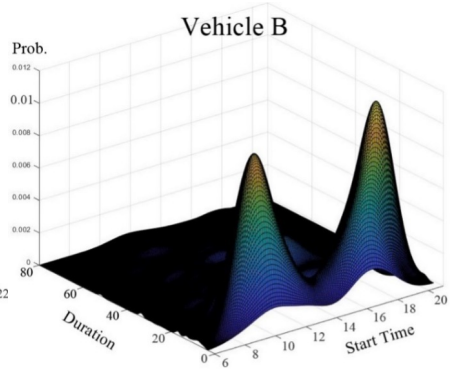

Fig. 4. Vehicle B Smooth PDF (Start Time & Trip Duration)

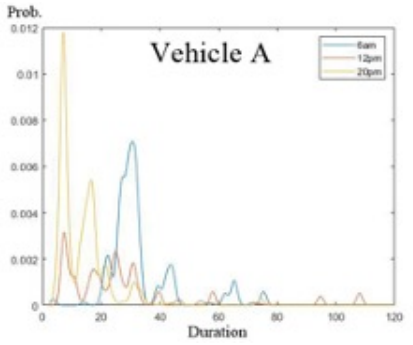

Fig. 5. Vehicle A Smooth PDF Time Slices (Trip Duration)

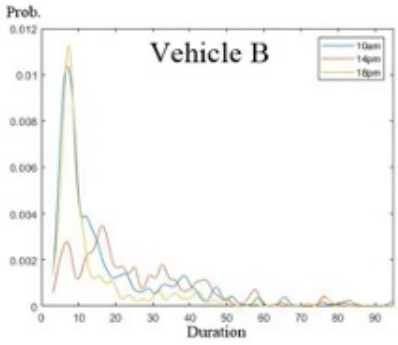

Fig. 6. Vehicle B Smooth PDF Time Slices (Trip Duration)

The time zones matrix for each vehicle is used to predict the duration of the trips in the test set. The test set is converted into binary sequences twice. The first sequence stores the real values transformed to zeroes and ones based only on m threshold of the durations of the trips. Given their start time, predicted values are generated using the time zones matrix and saved in the second sequence. Comparison of the model results with the real values is presented in the confusion matrix produced based on these two sequences (see Figures 7 and 8 for

the Vehicles A and B). For both observed vehicles prediction models are using the threshold p=0.4. The second parameter m has different values (20 minutes for Vehicle A and 15 minutes for Vehicle B) based on the features of drivers discussed earlier in this paper. Nozari Zarmehri and Soares (2015) use a similar approach in their work: "Given the difference between the data collected by different taxis, the best model for each one can be obtained with different algorithms and/or parameter settings". The testing sample for Vehicle A is relatively small and holds 69 trips to work with (20% of the whole dataset), 65 of which are longer than 20 minutes, and only 4 - shorter. The trained model predicts 54 long journeys and 15 short ones, having 11 trips as type II error. The overall accuracy of the model for this vehicle is 84,06%. Other calculated parameters are presented in Table I. Table II shows the metrics for the Vehicle B, which has more than twice as many trips for the same period, so the train and test sets are bigger. Model accuracy is 71,43% due to the high False Negative Rate (FNR).

Table 1. Vehicle A prediction model metrics

Total: 69		**Real values**	
		Long trips: 65	***Short trips: 4***
Predicted	***Long trips: 54***	54	0
	Short trips: 15	11	4
Accuracy: 84,06%		TPR: 83,08%	FPR: 0,00%
F1-score: 0,91		FNR: 16,92%	TNR: 100,00%

Table 2. Vehicle B prediction model metrics

Total: 161		Real values	
		Long trips: 114	Short trips: 47
Predicted	Long trips: 78	73	5
	Short trips: 83	41	42
Accuracy: 64,04%		TPR: 64,04%	FPR: 10,64%
F1-score: 0,76		FNR: 35,96%	TNR: 89,36%

High F1-score, which is calculated based on precision (ratio of TP to all positives predicted) and recall (ratio of TP to all real positives), also shows that method is efficient for predicting the long trips in imbalanced data sets. Both models have the FNR higher than FPR (long trips predicted as short ones). It should be taken into consideration while improving the model mechanics in the future. One of the possible solutions is to improve the algorithm working with the calculated probability Pr from formula (6) by having the probability thresholds for both short and long journeys. Journeys with uncertain values (both probabilities lower than the thresholds) should be processed in another way to improve the model's accuracy.

4.2 Model Results

131 vehicles from the provided dataset (consists of more than 300 vehicles) show an accuracy higher than 80% (see Fig. 7), 52 are in the range 70%-80%, 41 are with the accuracy between 60% and 70%. Taking into consideration that some of the drivers had almost no driving history (the model did not work with 12 vehicles due to lack of information) and having the hypothesis on improving the model's mechanics, we can state that the further development of the presented approach is relevant. In addition, the model can be used to generate the data regarding the driver behaviours which can be further used for other algorithms' validation.

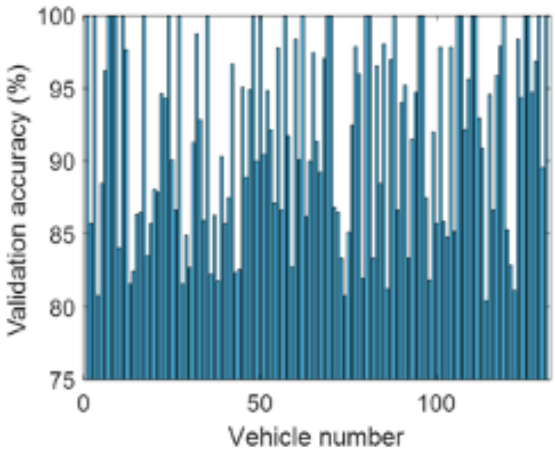

Fig. 7. Illustration of cluster with highly predictable drivers

4.3 Distance Smooth PDF Hypothesis

This section slightly covers the idea brought by the analysis of Vehicle A smooth PDF in Fig. 3 and its time slices in Fig. 5. The difference in the morning and evening trips' durations on the peaks can be caused not only by the different destination points but also due to rush hours. In this case, the smooth PDF produced for the distance instead of duration should have peaks: (1) on the same start times (as the observed trips are the same); (2) with equal distance values (being at the same point of distance axis); (3) of higher probability (as the multiple peaks with different duration, but the same distance will be accumulated there). The graph produced is presented in Fig. 8 and satisfies all the above assumptions. It means that the journeys from / to the same places (so the trips are of the same distance) can have a different duration depending on the time of the day.

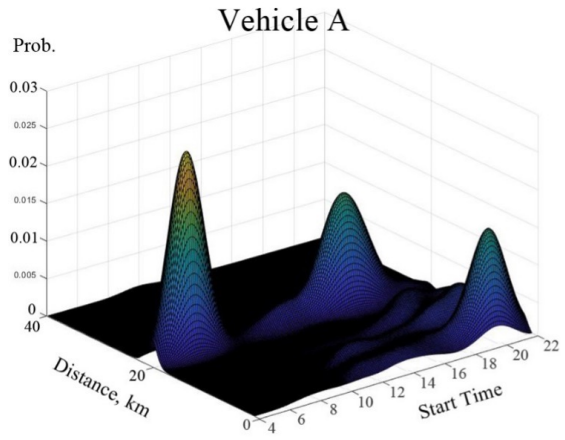

Fig. 8. Vehicle A Smooth PDF (Start Time & Trip Distance)

5 Conclusion and Future Work

This paper presented a study of driver behaviour modelling with an advanced smooth PDF-based approach for predicting expected usage. The distinguishing feature of the proposed approach is the application of kernel density estimation of the joint PDF for improving operational management during real-world usage. From a practical application viewpoint, of particular interest is to predict the length of the journey, which provides the scope for contextual intelligence insight for vehicle performance optimisation. The introduced approach proved to be efficient to predict trip durations defined as "short" or "long" based on the pre-defined threshold. It was also investigated that some drivers are more predictable than others. In these terms, the study may be extended to comparing the performance of the presented approach with a variety of other classification models. Given that drivers complete different types of cycles, which reflect on the usage distribution, other modelling approaches may be more efficient for certain groups of drivers, where the PDF approach can be employed as the model of choice.

Another advantage of the PDF approach is that it can be used as a similarity metric for further clustering. In further work, the Kolmogorov-Smirnov 2D statistic, presented in [12], is considered to compare PDFs from different vehicles. The interest here is represented to identify clusters with similar driver behaviour, which can support the optimal choice of the ML model to deliver further benefit and performance optimisation on a system level.

Acknowledgment

Alexey Uglanov and Kirill Kartashev acknowledge the support from the Erasmus+ Programme for their research placement with the Advanced Automotive

Analytics research laboratory at the University of Bradford. They also acknowledge the support of their home institution Plekhanov Russian University of Economics, Moscow. This research was supported by aiR-FORCE project, IDE, UK.

References

1. AbuAli, N., Abou-zeid, H.: Driver behavior modeling: Developments and future directions. International journal of vehicular technology **2016** (2016)
2. de Araujo, A.C., Etemad, A.: Deep neural networks for predicting vehicle travel times. In: 2019 IEEE SENSORS, pp. 1–4. IEEE (2019)
3. Byrne, T.J., Campean, F., Neagu, D., et al.: Towards a framework for engineering big data: An automotive systems perspective. In: DS 92: Proceedings of the DESIGN 2018 15th International Design Conference, pp. 1511–1522 (2018)
4. Grimmett, G.S., et al.: Probability and random processes. Oxford university press (2020)
5. Handley, S., Langley, P., Rauscher, F.A.: Learning to predict the duration of an automobile trip. In: KDD, pp. 219–223. Citeseer (1998)
6. Krause, C.M., Zhang, L.: Short-term travel behavior prediction with gps, land use, and point of interest data. Transportation Research Part B: Methodological **123**, 349–361 (2019)
7. Markkula, G.: Driver behavior models for evaluating automotive active safety: From neural dynamics to vehicle dynamics. Chalmers University of Technology (2015)
8. Matuszyk, T.I., Cardew-Hall, M.J., Rolfe, B.F.: The kernel density estimate/point distribution model (kde-pdm) for statistical shape modeling of automotive stampings and assemblies. Robotics and Computer-Integrated Manufacturing **26**(4), 370–380 (2010)
9. Miyajima, C., Takeda, K.: Driver-behavior modeling using on-road driving data: A new application for behavior signal processing. IEEE Signal Processing Magazine **33**(6), 14–21 (2016)
10. Toledo, T.: Integrated driving behavior modeling. Northeastern University (2002)
11. Zarmehri, M.N., Soares, C.: Using metalearning for prediction of taxi trip duration using different granularity levels. In: International Symposium on Intelligent Data Analysis, pp. 205–216. Springer (2015)
12. Zhang, B., Chen, R.: Nonlinear time series clustering based on kolmogorov-smirnov 2d statistic. Journal of Classification **35**(3), 394–421 (2018)