Understanding Drivers' Safety by Fusing Large Scale Vehicle Recorder Dataset and Heterogeneous Circumstantial Data

Daisaku Yokoyama[1(✉)], Masashi Toyoda[1], and Masaru Kitsuregawa[2]

[1] Institute of Industrial Science, The University of Tokyo, Meguro, Japan
{yokoyama,toyoda}@tkl.iis.u-tokyo.ac.jp
[2] Data Integration and Analysis System (DIAS), The University of Tokyo, Meguro, Japan
kitsure@tkl.iis.u-tokyo.ac.jp

Abstract. We present a method of analyzing the relationships between driver characteristics and driving behaviors on the basis of fusing heterogeneous datasources with large-scale vehicle recorder data. It can be used, for example, by fleet managers to classify drivers by their skill level, safety, physical/mental fatigue, aggressiveness, and so on. Previous studies relied on precise data obtained in only critical driving situations and did not consider their circumstances, such as road width and weather. In contrast, our approach takes into account not only a large-scale (over 100 fleet drivers) and long-term (one year's worth) records of driving operations, but also their circumstances. In this study, we focused on classifying drivers by their accident history and examined the correlation between having an accident and driving behavior. Our method was able to reliably predict whether a driver had recently experienced an accident (f-measure = 72%) by taking into account both circumstantial information and velocity at the same time. This level of performance cannot be achieved using only the drivers' demographic information or kinematic variables of operation records.

Keywords: Vehicle recorder · Fusing data from heterogeneous datasources · Driving safety · Accident history · Individual driving behavior

1 Introduction

Driver management has been an important issue for the transportation industry. Keeping drives safe and at the same time efficient is still a hard problem; transport companies typically manage their drivers by using demographic information to estimate their safety; however, such information overlooks the current condition and improvements in skill of the driver.

We have developed a method for analyzing the relationships between driver characteristics and driving behaviors on the basis of vehicle recorder data combined with other datasources such as weather reports and road maps. It can be

J. Kim et al. (Eds.): PAKDD 2017, Part II, LNAI 10235, pp. 734–746, 2017.
DOI: 10.1007/978-3-319-57529-2_57

used, for example, by fleet managers to classify drivers by their skill level, safety, physical/mental fatigue, aggressiveness, and so on. Our method manages drivers by not *who they are*, but rather *how do they drive*.

Several studies [1,3,5] have analyzed driving behaviors. They relied, however, on detailed and precise data on a small number of drivers, so it is difficult to extrapolate their results to the general driver population. Many transportation companies have introduced dashboard cameras (dashcams) and/or vehicle data recorders (which collect GPS, velocity, and acceleration data) into their fleets. Although the amount of data collected tends to be sparse due to storage limitations, data can be collected on a large number of drivers. Many kinds of transportation related information, such as weather, road structure, degree of traffic congestion, are also available nowadays. Utilizing such heterogeneous datasources would improve the preciseness of the management's understanding of each driver's characteristic.

Our method classifies drivers on the basis of long-term records of kinematic variables (maximum velocity, acceleration, etc.) related to their driving operations (braking, steering, etc.). It is based on the assumption that the distributions of these variables differs from driver to driver. Our method takes into account the factors of driving circumstances by fusing various heterogeneous datasources. We focused on classifying drivers who had recently been involved in accidents and examined the correlation between having an accident and driving behavior. Our findings are useful both for educating drivers and preventing accidents.

Many studies [4,13] have analyzed driving behaviors as a means of estimating driver risks. However, they only used driving operation information and tended to focus on extreme case of driving operation. Driver characteristics such as driving skill are reflected in all situations, not only in critical ones; for example, a skillful driver will brake smoothly on slippery roads during heavy rainfall. The previous studies thus overlooked the information to be obtained from operations performed in non-critical situations. By contrast, in this study, we used *all* driving information derived from many heterogeneous datasources to better estimate a driver's characteristics.

Our main contributions are:

- An intensive examination of large-scale vehicle recorder data covering all driving operations demonstrated the effectiveness of our method for analyzing the relationships between driver characteristics and driving behaviors. It was able to reliably predict whether a driver had recently experienced an accident (f-measure = 72%). This level of performance cannot be achieved by using only drivers' demographic information or kinematic variables of operation records.
- It showed that fusing heterogeneous data is essential to depicting driver behavior precisely. When we only used kinematic variables of driving records as the features of drivers, classification performance was poor (f-measure $< 66\%$).
- We found an appropriate way to combine circumstantial information. When we fused operation records and other non-kinematic information and took into account these information separately, the classification performance was almost same as using kinematic features. Performance improved after adding

features that took into account both velocity and circumstantial information at the same time.

In Sect. 2, we overview related work. In Sect. 3.1, we explain our analysis. We explain the driving operation dataset we used in Sect. 3.2 and describe other dataset to take into account driving circumstances in Sect. 3.3. In Sect. 3.4, we present our method for analyzing the relationships between driver characteristics and driving behaviors and evaluate its effectiveness. This article ends in Sect. 4 with a summary and a look at future work.

2 Related Work

There has been research on using vehicle recorded data, such as velocity and location, for various purposes [6,9,11]. The studies can be grouped into two categories: those that utilize large-scale vehicle location data [2] and those that investigate a small amount of driving operation data. We believe that ours is the first study to investigate both driving operation and its circumstances on a large-scale (more than 1000 drivers).

The 100-Car Naturalistic Driving Study [7] is one of the largest studies on the use of vehicle recorded data. It used many types of precise driving information and driver demographic data (age, gender, personality, etc.) and thoroughly analyzed the driver information statistically. Several studies have used the driving information in this archive to assess driver risk. For example, Guo et al. [4] reported an effective model for identifying high-risk drivers by using driver demographic information and the occurrence of critical-incident events. Their model mainly uses demographic information. Zheng et al. [13] collected data on naturalistic driving and analyzed the relationship between the kinematic information and driver risk-taking behavior. Their analysis focused on kinematic information for critical driving operations involving large accelerations. Yokoyama et al. [12] investigated the relationship between kinematic information and drivers' accident histories; however they did not utilize driving circumstances.

Some studies have tried to classify drivers on the basis of the aggressiveness of their driving behavior, with the aim of improving driving safety. Higgs et al. [5] analyzed the car-following behaviors of three drivers and identified the differences among them. Dang et al. [3] focused on the lane-changing behaviors of 12 drivers driving on a highway and found differences among them. Miyajima et al. [10] used data on 276 drivers and tried to identify drivers on the basis of their car-following behaviors and pedal operations. However, their data collection required the use of pedals with specially designed sensors. Their study and the other previous research relied on precise information on driving behavior, which is not always available.

3 Classification of Drivers' Accident Histories

3.1 Approach

Our research purpose is to identify the characteristics of drivers through their driving behaviors. In this study, we focused on classifying drivers as either safe

Table 1. Summary of vehicle recorder dataset

	All data	Driving days ≥ 20, driving hours ≥ 20
Number of drivers	1469	320
Driving duration in total	77,450 h	60,190 h

or unsafe on the basis of their driving records. Instead of using only critical operation records, we used a large amount of vehicle recorder data that included all driving operations and investigated how effective such data is for classifying drivers.

A driver performs various driving operations (braking, steering, etc.), each associated with several variables (maximum velocity, acceleration, etc.). A driver can be characterized by the distributions of these variables. We investigated ways to derive features from these variable distributions for use in classifying drivers as either safe or unsafe by using Support Vector Machine (SVM).

Each driving operation is affected by factors of the moment, such as the weather condition, road condition, degree of congestion, and time of day. We need to take into account the effects of these factors in order to derive good features from the operation records. These factors cannot be observed from the vehicle recorded operation records alone. Therefore, we should combine other datasources such as weather data to reconstruct other factors. Here, we focus on two circumstances: rainfall information and road width. We derived several features from the distributions of operation variables, taking into account the factors of the moment, and evaluated the effectiveness of our method.

3.2 Dataset

Vehicle Recorder Dataset. In our experiments, we used a large number of actual driving records[1] collected by a parcel delivery service company (transport company). The data were for about 1450 drivers working in the Tokyo area and covered one year (from 21 July 2014). A multifunctional data recorder in each delivery vehicle recorded longitudinal accelerometer, lateral accelerometer, gyro compass, and GPS data.

Since we focused on long-term driving behavior, we eliminated the data of drivers who had driven on fewer than 20 days or for less than 20 h in total. A summary of the data is shown in Table 1. The driving duration does not include the time during which the engine was turned off.

The vehicle data recorder automatically detected four basic driving operations: braking, steering, turning, and stopping. Several variables, including maximum velocity and acceleration, during each operation were recorded. The operation variables are listed in Table 2. The numbers of recorded operations per driver are summarized in Table 3. As mentioned, our dataset contained data on all driving operations, while those used in previous studies contained data only on critical operations involving high acceleration.

[1] The vehicle recorder data was provided by Datatec Co., Ltd.

Table 2. Operation record variables

Operation	Variables
Braking	Velocity (V), longitudinal acceleration (Gx), and jerk (derivative of acceleration with respect to time, Jx)
Steering	V, yaw velocity (Yr), yaw acceleration, and lateral acceleration (Gy)
Turning	{Gx, V} before turn, {V, centrifugal force (CG), yaw acceleration} during turn, and {V, CG} after turn
Stopping	V, Gx, and stopping duration

Table 3. Operation record statistics

Operation	No. of records per driver		No. of records
	(min)	(max)	(total)
Braking	114	45,861	1,993,341
Steering	239	46,452	2,783,723
Turning	121	21,027	1,218,957
Stopping	418	40,625	2,221,166

Driver Histories. With the cooperation of the transport company, we accessed their drivers' histories, including the traffic violations they had received and the accidents in which they had been involved. We used their histories to define their *accident experience* and *driving experience*.

Accident experience. Drivers who had at least one accident during a certain time period were defined as an *accident* driver. Even though some accidents were only small ones without any responsibility being assigned, we treated all accidents the same.

Driving experience. To estimate how long a driver had been driving, we used the oldest record in the driver's history to estimate the minimum number of driving years.

Using these definitions and the estimates, we investigated the differences in driving operation between the accident and no-accident drivers. The no-accident drivers, however, are not necessarily safe drivers. For example, a reckless driver may simply have been lucky enough to avoid an accident over the course of a year. We therefore focused on drivers who had at least five years' worth of driving experience. We defined a driver who had at least five years' worth of driving experience without any accidents in the previous five years as safe and otherwise as unsafe. There were 82 safe drivers and 43 unsafe drivers.

3.3 Fusing the Driving Circumstances with Operation Records

To understand each driver's driving behavior, we focused on the distributions of variables for driving operations. Each driving operation is affected by factors of

the moment, such as the weather condition, road structure, and degree of traffic congestion. Therefore, we combined other datasources with the operation records to reflect the effect of these factors. Each operation record contains GPS data and time information; thus we could perform spatial- and temporal- matching with the other datasources.

To take into account driving circumstances, we created two different variable distributions: (a) splitting up operation records by circumstance and (b) splitting up operation records by the combination of two circumstances. We selected several factors that represent driving circumstances; they are as follows.

Velocity. Operation variables are correlated due to kinematic restrictions for both safe and unsafe drivers. For example, steering at a high velocity tends to cause a low yaw rate. We therefore treated velocity as the basic variable for each operation and split up the operation records according to their velocity values. For example, we divided the braking operation records into six bins on the basis of velocity and estimated the longitudinal jerk densities for each bin. We found that the shapes of the distributions differ among the velocity bins.

Time of Day. The degree of traffic congestion heavily affects driving behavior. To reflect this factor, we used the occurrence time of each operation record. We separated the operation records into several time ranges, and compared the variable distributions. We found the operation distributions in the morning and evening differ from at other time, which seems to be the result of traffic congestion. Time is not kinematic information; however, it surely affects kinematic variables of operations.

Road Properties. Driving operations are also affected by the road width. For example, turning onto a narrower road tends to require more deceleration than turning onto a wider road. We could match each operation location with a point on a digital road map[2]. We simply searched for the road segment nearest the operation location. If the nearest segment was more than 30 m away (due, for example, to being on a private site such as a factory or university), we considered that the location could not be matched to a point on the map and ignored that record. The road map contains information about the road width, represented in several ranks, and whether the road is bi-directional or not. If the road was bi-directional, we assumed that the width of the segment was one rank narrower. We used four road width ranges: >13 m, $13 > w > 5.5$, <5.5 m, and unknown.

Rainfall. Weather heavily affects road conditions and driving operations. When it is raining, for example, the accident rate is eight times higher when the weather is dry[3]. We used X-band Multi Parameter Radar information collected by the

[2] We used the "Advanced Digital Road Map Database" developed by Sumitomo Electric System Solutions Co., Ltd. The database was provided by the Center for Spatial Information Science at the University of Tokyo.

[3] From discussions with an Expressway company.

Ministry of Land, Infrastructure, Transport and Tourism[4]. It detects rainfall in a 250 m mesh every minute. This fine-grained weather radar can detect sudden rain showers that happen frequently in Japan. Since every operation record contained GPS data and time information, we could match each operation location with the rainfall information at that time.

3.4 Features

Derivation. We used all 17 dataset variables listed in Table 2 to derive the driver features. We also used driver demographic information known to be related to driving safety.

First, we created basic features that represent demographic characteristics or distributions of kinematic variables:

- Demographic features: We used the driver's age, gender, and time since obtaining a driver's license as three demographic features. This information is commonly used by insurance companies to set auto insurance rates.
- License feature: In Japan, a driver who has not had any accidents and has not been cited for a driving violation during the preceding five years is categorized as a "gold license" driver and is generally considered to be a safe driver. We thus defined a binary feature for whether a driver had a gold license or not. The license category is updated when one's license is renewed, and the renewal interval is three to five years. Therefore, a gold license does not always mean an accident-free driver; many drivers have had accidents in recent years and still hold a gold license. When we classified drivers as safe or unsafe by using their license category information alone, we achieved only a 35% precision, which is virtually the same performance as with a random classifier.
- Operation frequency features: We counted the number of instances for each of the four driving operations for each driver and normalized it by the driving duration.
- Variable distribution features: We defined the shapes of the variable distributions as features. Each variable value was binned into one of ten intervals; the maximum and minimum bin breakpoints were chosen by hand, and the other bins were defined to have the same width. Therefore, each variable distribution was represented by ten values. There were thus 170 variable distribution features (17 variables $\times$ 10 values).

Second, we consider the relationship between circumstances and basic kinematic variables (as described in Sect. 3.3, approach (a)):

- Variable distribution by velocity features: Driving operations are strongly affected by the vehicle's velocity. We therefore selected six velocity-related variables for use in separating the operation records, and combined them with other variables, as shown in Table 4. The operation records were separated by the corresponding velocity-related variable, and the distributions of the other

[4] XRAIN: http://www.river.go.jp/kawabou/ipXAreaMap.do.

Table 4. Combination patterns of operation variables

Operation	Velocity-related variable (number of bins)	Other variables combined with velocity-related variable
Braking	Velocity (6)	Gx, Jx
Steering	Velocity (5)	Yr, yaw acceleration, Gy
Turning	Velocity before turn (4)	Gx before turn
Turning	Velocity during turn (4)	CG, yaw acceleration during turn
Turning	Velocity after turn (5)	CG after turn
Stopping	Velocity (5)	Gx

variables were calculated separately. The velocity-related variables were digitized into b values by intervals with a constant width (5 km/h). The other variable distributions were digitized with ten intervals, so the feature of a variable is represented by $b \times 10$ values.

- Variable distribution by road width features: We defined each of the four road width ranges as an indicator of a circumstance, and use it to split operation records.
- Variable distribution by rainfall features: We decided the raining condition to be when rainfall is larger than 5.0 mm/h. Thus we split up the operation records into three rainfall ranges: >5.0, ≤ 5.0, unknown (that is caused by the lack of observation).
- Variable distribution by time of day features: We defined five time ranges to capture the different traffic conditions of the operation records: [6:00–9:00], [9:00–12:00], [12:00–18:00], [18:00–21:00], [21:00–6:00].

Finally, we considered two of the above circumstances at the same time (as described in Sect. 3.3, approach (b)). In this study, we limited the number of sets of combination to three. Among the circumstance features, velocity has the largest possibility to restrict vehicle's motion. Thus we selected velocity as the fixed feature, and combined it with the other three circumstance features as follows:

- Variable distribution by velocity and road width features
- Variable distribution by velocity and rainfall features
- Variable distribution by velocity and time of day features

Increasing the number of combinations improved the accuracy of the depicted variable distribution for each driver. Although this helped to describe the difference between driving behaviors precisely, it may cause data sparsity because it reduces the number of operation occurrence in each bin, which means the features will be more strongly affected by noise.

Feature Expression. We tested two methods of expressing the variable distributions as features.

Table 5. Feature settings

Feature category (no.)	a	b	c	d	e	f
Demographic (3)	✓	✓		✓	✓	✓
License (1)		✓		✓	✓	✓
Operation frequency (4)			✓	✓	✓	✓
Variable distribution (170)					✓	✓
Variable distribution by velocity (540)						✓
Number of available features	3	4	4	8	178	718
Number of frequent features	2	3	4	7	172	601

Feature category (no.)	g	h	i	j	k	l	m
Features of setting f (718)	✓	✓	✓	✓	✓	✓	✓
Variable distribution by road width (680)	✓			✓	✓		✓
Variable distribution by rainfall (510)		✓		✓		✓	✓
Variable distribution by time of day (850)			✓		✓	✓	✓
Number of available features	1398	1228	1568	1908	2248	2078	2758
Number of frequent features	1160	1032	1337	1591	1896	1768	2327

Feature category (no.)	n	o	p	q
Features of setting m (2758)	✓	✓	✓	✓
Variable dist. by velocity and road width (2160)	✓			✓
Variable dist. by velocity and rainfall (1620)		✓		✓
Variable dist. by velocity and time of day (2700)			✓	✓
Number of available features	4918	4108	5008	8158
Number of frequent features	3550	3295	3986	6177

Probability method. We denoted each driver's frequency for each bin as N_i and computed each driver's occurrence probability P_i, which is N_i normalized by the number of operation instances for the driver. We used P_i itself as a feature.

KL divergence method. We described the difference between two distributions, P and Q. The KL divergence [8] is a representative definition of the distance between two distributions: $KL(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$.

We used $P_i \log \frac{P_i}{Q_i}$ of each bin as the feature.

Performance Evaluation. We tested 17 combinations of features, as shown in Table 5. The feature settings are categorized into four groups; (a) to (d) use only demographic and statistical information on the driver; (e) and (f) introduce the variable distributions of the driving operations; (g) to (m) introduce driving circumstance information from other datasources or non-kinematic information in the operation records; (n) to (q) take into account the effects of the combination of the velocity and other circumstantial information.

We evaluated the performance by 10-fold cross validation. Features that appeared in the driving records of less than 30 drivers were eliminated. The number of remaining features of each combination is shown in Table 5, as the "Number of frequent features". All remaining features were normalized beforehand. Three types of kernel functions (linear, polynomial, Gaussian) with hyperparameters (Table 6) were evaluated in a grid-search manner to achieve the best AUC (area under the ROC curve) value. We also used feature selection based on the χ^2 value. The best number of features was determined from the grid search.

Table 6. Parameters for grid search

Kernel	Hyperparameter
Linear	$C : [2^{-5}, ..., 2^{10}],\ w_{accident} : \{1, 2, 3, 5, 10\}$
Polynomial	$C : [2^{-5}, ..., 2^{10}],\ \gamma : [2^{-10}, ..., 2^{3}],$ $degree : \{2, 3\},\ w_{accident} : \{1, 2, 3, 5, 10\}$
Gaussian	$C : [2^{-5}, ..., 2^{10}],\ \gamma : [2^{-10}, ..., 2^{3}], w_{accident} : \{1, 2, 3, 5, 10\}$

Table 7. Classification performance

Setting	Method	No. of selected features	Precision	Recall	F-measure	AUC
a	-	2	0.36	1.00	0.52	0.57
b	-	3	0.43	0.93	0.58	0.64
c	-	4	0.36	1.00	0.53	0.45
d	-	5	0.38	0.88	0.53	0.59
e	p	50	0.47	0.88	0.62	0.71
f	p	20	0.57	0.79	0.66	0.79
g	KL	20	0.67	0.67	0.67	0.80
h	KL	40	0.53	0.77	0.63	0.76
i	KL	30	0.70	0.74	0.72	0.81
j	KL	50	0.55	0.84	0.66	0.77
k	KL	30	0.70	0.70	0.70	0.80
l	KL	50	0.59	0.77	0.67	0.81
m	KL	60	0.56	0.81	0.67	0.81
n	KL	80	0.59	0.79	0.67	0.83
o	KL	80	0.57	0.93	0.71	0.81
p	KL	90	0.72	0.77	0.74	0.85
q	KL	40	0.60	0.88	0.72	0.85
Random classifier			0.37			0.50

Figures 1 and 2 show the best f-measure and AUC for each setting, respectively. Representative results are shown in Table 7. The random classifier was used as a baseline; it had a precision of 37% (= 43/125).

The demographic information was not so helpful in classifying drivers, although it was slightly better than the random classifier: the AUC values for settings (a) and (b) were greater than 0.5. Since all the drivers were well-trained professionals, the demographic information may not have reflected their driving skills so well.

The use of the kinematic information obtained from vehicle recorders was helpful in classifying the drivers, as we can see from the results for setting (e). When we took into account the velocity at which the operation was performed,

performance improved slightly (see results for (e) and (f)). Adding circumstantial information (road width, rainfall, and time of day) to the kinematic information resulted in almost same performance ((g) to (i)). This circumstantial information was of much help when it was combined with the velocity ((n) to (q)).

Figure 3 shows the ROC curves of representative results. Taking into account the velocity of driving operations improved performance ((e) and (f)). Adding circumstantial information improved performance; it was not so helpful when we combined it with simple variable distributions (m); however, it greatly improved performance when it was combined with both variable distributions and velocity (q).

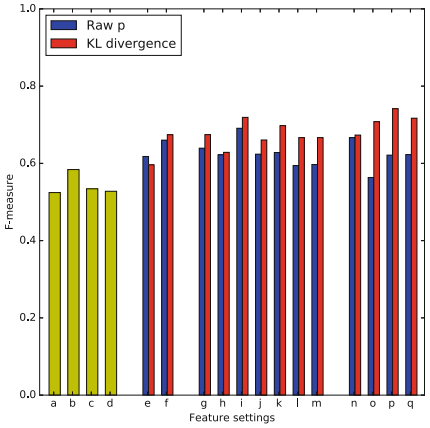

Fig. 1. F-measure for different feature settings

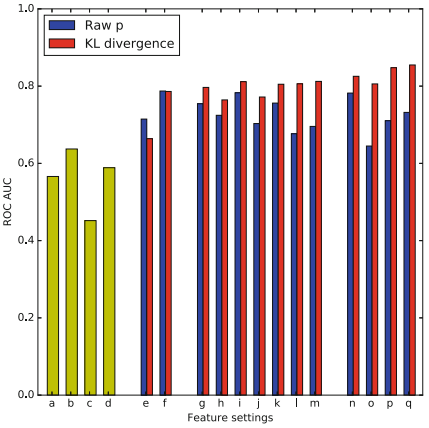

Fig. 2. AUC under the ROC curve for different feature settings

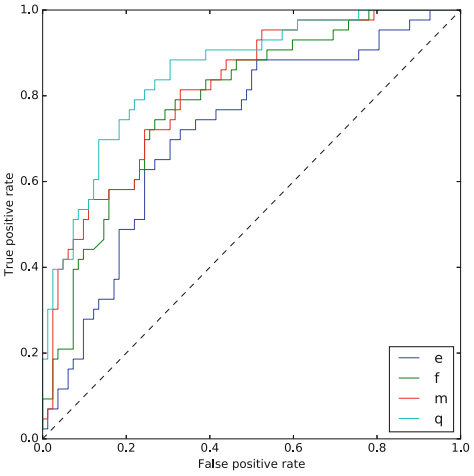

Fig. 3. ROC curves of representative results

4 Conclusion

We thoroughly examined a large-scale archive of recorded vehicle data in order to clarify the relationship between safety and driver behavior. We used multiple datasources to compensate for driving circumstances in operation records and successfully classified drivers as either safe or unsafe (f-measure = 72%). Methods that use only driver demographic information or kinematic variables of operation records have not achieved this level of performance.

This is the first step toward a better understanding of the relationship between safe driving and driver behavior. Although this study considered only past accidents, the knowledge acquired will be helpful in investigating driver safety and preventing future accidents. We thus plan to apply our method to predicting accidents. Our findings on the characteristics of drivers through their driving behaviors will be helpful in educating drivers.

References

1. Castignani, G., Frank, R., Engel, T.: Driver behavior profiling using smartphones. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), pp. 552–557, October 2013
2. Castro, P.S., Zhang, D., Chen, C., Li, S., Pan, G.: From taxi GPS traces to social and community dynamics: a survey. ACM Comput. Surv. **46**(2), 17:1–17:34 (2013)
3. Dang, R., Zhang, F., Wang, J., Yi, S., Li, K.: Analysis of Chinese driver's lane change characteristic based on real vehicle tests in highway. In: ITSC 2013, pp. 1917–1922, October 2013
4. Guo, F., Fang, Y.: Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. **61**, 3–9 (2013). http://www.sciencedirect.com/science/article/pii/S0001457512002382
5. Higgs, B., Abbas, M.: A two-step segmentation algorithm for behavioral clustering of naturalistic driving styles. In: ITSC 2013, pp. 857–862, October 2013
6. Johnson, D., Trivedi, M.: Driving style recognition using a smartphone as a sensor platform. In: ITSC 2011, pp. 1609–1615 (2011)
7. Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J.: The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. Technical report DOT HS 810 594, National Highway Traffic Safety Administration (2006)
8. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Statist. **22**(1), 79–86 (1951). doi:10.1214/aoms/1177729694
9. Liu, W., Zheng, Y., Chawla, S., Yuan, J., Xie, X.: Discovering spatio-temporal causal interactions in traffic data streams. In: SIGKDD 2011, August 2011
10. Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., Itakura, F.: Driver modeling based on driving behavior and its evaluation in driver identification. Proc. IEEE **95**(2), 427–437 (2007)
11. Wu, W., Ng, W.S., Krishnaswamy, S., Sinha, A.: To taxi or not to taxi? - enabling personalised and real-time transportation decisions for mobile users. In: IEEE 13th International Conference on Mobile Data Management, pp. 320–323, July 2012

12. Yokoyama, D., Toyoda, M.: Do drivers' behaviors reflect their past driving histories? - large scale examination of vehicle recorder data. In: 2016 IEEE International Congress on Big Data, pp. 361–368. IEEE (2016). doi:10.1109/BigDataCongress.2016.58
13. Zheng, Y., Wang, J., Li, X., Yu, C., Kodaka, K., Li, K.: Driving risk assessment using cluster analysis based on naturalistic driving data. In: ITSC 2014, pp. 2584–2589. IEEE (2014)