modelling

MDPI

Article

Modelling the Energy Consumption of Driving Styles Based on Clustering of GPS Information †

Michael Breuß [1], Ali Sharifi Boroujerdi [2] and Ashkan Mansouri Yarahmadi [1,*]

[1] BTU Cottbus-Senftenberg, Institute for Mathematics, Platz der Deutschen Einheit 1, 03046 Cottbus, Germany
[2] Volkswagen Infotainment GmbH, 44799 Bochum, Germany
* Correspondence: yarahmadi@b-tu.de

† This paper is the extension of the conference paper: Breuß, M., Sharifi Boroujerdi, A., Mansouri Yarahmadi, A., Energy-Efficient Driving Model by Clustering of GPS Information, Operations Research Proceedings 2022, Oliver Grothe, Stefan Nickel, Steffen Rebennack and Oliver Stein (editors), Springer, submitted.

Abstract: This paper presents a novel approach to distinguishing driving styles with respect to their energy efficiency. A distinct property of our method is that it relies exclusively on the global positioning system (GPS) logs of drivers. This setting is highly relevant in practice as these data can easily be acquired. Relying on positional data alone means that all features derived from them will be correlated, so we strive to find a single quantity that allows us to perform the driving style analysis. To this end we consider a robust variation of the so-called "jerk" of a movement. We give a detailed analysis that shows how the feature relates to a useful model of energy consumption when driving cars. We show that our feature of choice outperforms other more commonly used jerk-based formulations for automated processing. Furthermore, we discuss the handling of noisy, inconsistent, and incomplete data, as this is a notorious problem when dealing with real-world GPS logs. Our solving strategy relies on an agglomerative hierarchical clustering combined with an L-term heuristic to determine the relevant number of clusters. It can easily be implemented and delivers a quick performance, even on very large, real-world datasets. We analyse the clustering procedure, making use of established quality criteria. Experiments show that our approach is robust against noise and able to discern different driving styles.

Keywords: clustering; energy efficiency; driving style analysis; jerk-based feature; GPS data

Citation: Breuß, M.; Sharifi Boroujerdi, A.; Mansouri Yarahmadi, A. Modelling the Energy Consumption of Driving Styles Based on Clustering of GPS Information. *Modelling* **2022**, *3*, 385–399. https://doi.org/10.3390/modelling3030025

Academic Editor: Alessandro Corsini

Received: 16 June 2022
Accepted: 22 August 2022
Published: 2 September 2022

1. Introduction

Driving style has a significant impact on the fuel consumption of a car. Intelligent hybrid cars can adapt to the driving style of the conductor to maximise their mileage. These optimisations could be manifold, and range from an efficient assistance of the electric engine to suggestions on energy-optimal routes [1,2]. In this work, we aim to provide a significant step towards integrating driving style as an additional constraint into this objective. We analyse driving style with respect to energy efficiency and provide an automated way to classify it. This results in a purely data-driven approach that robustly models driving styles concerning varying external environment factors that may have an impact on them, namely current traffic and weather conditions. In a sense, our proposed model efficiently judges the energy consumption pattern of a driver operating a car in a specific environmental condition, provided that it has already learnt the similar driver behaviours in terms of unseen log samples. This is indeed an advantage, as a unique optimal model can cover a wide range of environmental variations instead of having many tailored-made models at hand.

To allow the broad applicability of our method, we rely on data that can easily be obtained in any vehicle, namely global positioning system (GPS) data logs. Such setups are attractive due to their relatively low cost and the already-abundant availability of

GPS-enabled devices. However, these benefits come at a certain price. Insufficient accuracy yields noisy samples. Hardware failures may result in a partial or total loss of data. Furthermore, classifying the driving style without environmental information is a quite challenging task. Driving at a constant 120 $\mathrm{km\,h^{-1}}$ on a large motorway over lowland is surely an energy-efficient way to travel, but this may become less energy efficient when the motorway passes over rolling hills. Traffic jams on motorways lead to a stop-and-go motion in traffic, which is visible in the logs by frequent small variations in the acceleration and velocity, strongly resembling noise. However, the exact cause of such variations cannot always be deduced with absolute certainty. The driver may be looking for a free spot for parking, or their car may simply have broken down. An accurate and robust model must be able to handle these difficulties. As an advantage, we can equip our data-driven model with suitable data to validate all such conditions.

Let us also emphasis on the scaleability of our proposed method, specifically in the transportation context. Here, the collection of data in terms of videos and pictures, etc., becomes critical as a car is driven across different countries. In this case, affordable GPS hardware is the only required infrastructure for our method to collect a huge amount of logs which are independent of the driving location, with almost no law constraints needing to be enforced on the collected data. In this way, an abstract birds-eye view can be provided to the management layer when it comes to optimizing the transportation costs in terms of fuel [3].

Fortunately, notable advances in modelling and understanding driver behaviour have been made in recent years; however, these are mainly useful in the area of traffic safety engineering. We refer to [4–8] among the vast amount of literature in this highly active field of research. Although the objective in this field is completely different from that in our case, let us still review some works in more detail, so that we can identify—at a technical level—some aspects related to our approach.

In [4,8], the authors suggest the usage of sensors and simulators to identify driver movements and to predict their behaviour in the forthcoming seconds. These predictions can, for example, be used to prevent collisions. Both works take a probabilistic approach by analysing dynamic and hidden Markov models. The findings presented in [7,9] classify drivers according to their imminent risk to traffic. This classification is obtained by comparing characteristic features of a given driver with information gathered from other vehicles in the vicinity; while the authors of [9] use statistical measures, the authors of [7] combine a clustering algorithm with complex neuroscale and Bayesian factor analyses. Similar research is performed in [10], where cars are analysed for potentially dangerous behaviour. However, the focus of the latter work is more on the communication protocols between the traffic participants and not on their classification. All these approaches have commonalities in that they collect significant amounts of data from a heterogeneous pool of sources, and that they are designed to analyse the conductor during driving. Many works have also tried to model the conductor themself by analysing their reactions in various settings—see [11,12]. The authors of the latter work use a car with specialised sensors to measure various information about the drivers. These include the position and velocity of the car as well as the operational patterns of the gas and brake pedals. The analysis is performed by means of Gaussian mixture and optimal velocity models. In contrast to the first cited references, these latter works process the data only after collecting it.

Let us now turn to the use of GPS data. Modern tracking devices, such as navigation systems, cell phones, or smartwatches, offer huge amounts of information that may be processed to distinguish driving styles. We now consider a few approaches based on the positional information collected by the tracking devices, which are used to analyse some aspects of driving, comprising a new and emerging field of work. The authors of [3] show that the transportation patterns of coal trucks is highly dependent on the daily routines of the drivers, which influences the effectiveness of travel time. This was revealed by studying the average speed of the trucks, as obtained based on their GPS logs during day and night on a field survey. Here, it is shown that transportation in the morning can be

more optimal than that at night. In addition, optimization scenarios are implemented by reducing the number of stopping time during trips, changing sleep and rest patterns, and adjusting workloads to reduce driver fatigue. However, this analysis of GPS logs does not consider driving style of conductors in higher detail. In [13], the authors aim to identify reckless taxi drivers by analysing the velocity of their cabs and the regions that they pass through. The authors of [14] have similar goals, but limit themselves to analysing the routes taken. Here, a taxi driver becomes suspicious when their route deviates strongly from those taken by the majority of their colleagues. The authors of [15] propose an investigation of the behaviour of drivers within a restricted area by considering three anomalies: an abrupt change in acceleration or deceleration, lane changes, and excessive speed. In this way, they were capable of classifying drivers as either careful, distracted, dangerous, or very dangerous. In [16], a limited set of statistical-based features—making use of jerk, steering wheel gradient, etc.—are extracted and ranked for each driver using the Boruta package [17] to perform identity recognition of the drivers. In all the above settings, the missing environmental data can, to a certain extent, be compensated by increasing the amount of positional information. Furthermore, offline processing of the data allows researchers to use more powerful hardware than that available inside a car.

Presently, and based on advancements in the machine learning field, modern vehicles are often supplied with the necessary driver monitoring systems for the purpose of providing further safety. Though such monitoring systems are not installed on many roadworthy vehicles that are still driven on roads, a huge amount of their trajectory data can be processed by the commodity hardware owned by drivers, via smart phone [18]. These collected data can be further investigated with adequate machine learning approaches to bring further safety to the road traffic, see [19] for a survey. As an example, we refer to [20] for an investigation of a sliding-window-based approach on collected logs by the accelerometer, linear acceleration, magnetometer, and gyroscope sensors of a cell phone. Here, it is reported that a random forest approach [21] performs best to classify drivers compared with support vector machine [22] and Bayesian network [23] approaches, as the window size increases over the gyroscope and accelerometer sensors.

Our contribution. We propose a novel method for the analysis of energy-efficient driving styles via GPS data collected by drivers. Similarly to the aforementioned studies concerned with traffic safety, we discern different driving styles, but we focus on energy efficiency. We delineate why our approach is well suited for this aim by providing a purely data-driven model with respect to the considered feature. Our computational approach resembles the approaches taken in the referenced studies which were concerned with GPS data. Our analysis is based on a set of GPS logs collected during driving. The processing of the data is performed in an offline post-processing step. Thus, one of our contributions is to carry over some ideas from traffic safety engineering and related areas to a novel field of application. Furthermore, we propose some dedicated techniques in order to meet the requirements of our setting, and provide a reliable solution in discerning energy-efficient driving styles.

Our ultimate goal is to classify drivers solely based on positional information, because this is relevant for the potential industrial applications of our approach. Let us stress that the use of GPS logs alone is therefore an important issue, and this also distinguishes our work from others that mark the current state of the art in the technically related, above-mentioned literature. Despite a high amount of noise in the existing real-world dataset [24] that we employed for demonstrating our method, we still obtain robust results. Our clustering algorithm employs an improved formulation of the jerk quantity introduced in [25], which those authors used for the analysis of driver's behaviour for traffic safety purposes.

Let us also stress that the sole use of positional information implies that it does not make much sense to employ a variety of features based on these data, as these will all be naturally correlated (see [26] for further details). With only the GPS data at hand, it is thus of primary interest to identify *one* reliable and robust feature to work with. Moreover, this

feature should be meaningful by relating to the energy efficiency of driving styles. One of the contributions of our paper is to provide this by means of a novel jerk-based feature.

As an advantage of our study, our approach has minimum amount of dependency on data, in a sense that the method does not require a huge amount of training data to judge the driver profile. Our formulated jerk-based feature requires only a finite and limited set of temporal information to label on a driver's style in a given environment.

In combination with an agglomerative hierarchical clustering algorithm, we achieve a reliable classification result with reasonable computational effort. Furthermore, we propose a strategy to determine a reasonable amount of clusters. Experimental results show that our approach is more robust than a straightforward adaptation of previous approaches.

Our proposed approach can be effectively adopted within the scope of the car insurance industry. In [27], a precise GPS positioning system was developed to aid traffic police and insurance personnel in remotely identifying driver behaviour while reviewing the events of an accident. The obtained results in [28] based on GPS logs show that males have riskier driving patterns than females. These gender differences and their impact on driving patterns result in different accident rates. This was discussed under the "no-gender" discrimination regulation and within the context of a pay-as-you-drive insurance scheme in [28].

In this paper, which is our significant extension to the short introduction we presented as a conference paper [29], we proceed as follows. First, we present a detailed discussion of modelling, informing the reader of our motivation in selecting the jerk as a meaningful feature for our application. Furthermore, we present some comparisons with their related possible feature choices. This is followed by a presentation of our clustering method for classifying drivers, which we analyse in this work in much higher detail than in [29] with respect to various quality indicators. In the final section, we discuss the application of our method using a real-world dataset [24], to demonstrate the viability of our approach. The paper is finished by a summary and our conclusions.

2. Motivation: On Energy Efficiency and Jerk

Let us now outline our motivation for selecting the jerk as the underlying feature of our investigation. We comment on two important aspects in pattern recognition—physical significance of the feature; useful invariances for our application.

2.1. Basic Physical Considerations

For modelling, we consider the movement of a car during a fixed time frame, which we parametrise via the time $t \in [0, T]$. For simplicity of notation, we assume that the car performs a 1D movement along a straight, horizontal path, S, in this time frame. We denote by s_0 and s_T the starting point and the end point of S, respectively, and we denote the length of the path with d. Concerning the time-dependent position of the car along S, we make use of a corresponding function, $x(t)$. We denote the velocity of the car by $v = \dot{x} := \dot{x}(t) = v(t)$ and the acceleration of the car by $a = \ddot{x} = \dot{v}$. We assume for simplicity that the mass, m_c, of the car is a constant over the considered time frame (neglecting, e.g., the mass loss due to consumed fuel). Another fundamental yet not-too-obvious assumption for our modelling is that the driver aims to travel the distance, d, of the path, S, during the considered time frame over the time interval, $[0, T]$. This underlying assumption is required because, otherwise, one may realise a fuel-saving driving style by performing a full break and shutting off the car. Let us note that this underlying assumption is in accordance to the content of the given data after our preprocessing, because GPS logs of cars standing still will be discarded.

Let us now assert that, in the engine, fuel is transformed into mechanical work, W. In order to obtain a measure for fuel consumption (denoted by fc), the transformation process has to be related to consumption over the considered time frame—or, equivalently,

to the distance, d, passed during the time frame. To this end, we opt to measure fc via the generated mechanical work over distance, d, and assume the following proportionality law:

$$\mathsf{fc} \sim \frac{W}{d} \quad \text{or} \quad \mathsf{fc} = p\frac{W}{d} \tag{1}$$

where the proportionality constant, p, is given by the combustion efficiency of the engine.

Let us note that, in terms of physical measurement units, we have the following by $[d] = m$:

$$[\mathsf{fc}] = \left[\frac{W}{d}\right] = \mathrm{kg\,m/s^2} = \mathrm{N} \tag{2}$$

that means that fuel consumption is proportional to the force, F, with $[F] = \mathrm{N}$ needed to move the car along S.

Now, by considering Newton's second law, we have the following for the 1D movement of the car:

$$W = F \cdot d \quad \Leftrightarrow \quad \frac{W}{d} = F = m_c \cdot \bar{a} \tag{3}$$

where $\bar{a}$ is the acceleration needed over $[0, T]$ in order to move the car from s_0 to s_T over distance, d, along S. As a consequence of (1)–(3), we obtain

$$\mathsf{fc} \sim \bar{a} \quad \text{or} \quad \mathsf{fc} = p \cdot m_c \cdot \bar{a} \tag{4}$$

Let us now consider the situation that the car enters S with a certain velocity $\bar{v} > 0$. Then, by basic physical principles, the kinetic energy, $E = \frac{1}{2}\mathrm{m}\bar{v}^2$, corresponds to the work, W, stored in the movement.

In an ideal, frictionless environment, the car moves on with constant speed, $\bar{v}$, and there is no need to install an additional acceleration to hold $\bar{v}$ as would be desired by a driver in order to travel along S during the considered time frame. It is clear that, in reality, where, e.g., friction is encountered, a certain acceleration is needed to allow the driver to travel the distance, d, along S, during $[0, T]$. Since $\mathsf{fc} \sim \bar{a}$, this also means that a certain amount of fuel has to be used up.

The the question arises: how can this acceleration be applied over $[0, T]$ in the most fuel-efficient manner? One may imagine here, e.g., that it could be beneficial to accelerate very strongly at the beginning over $[0, \epsilon]$, $0 < \epsilon \ll 1$, and to let the car roll over the remaining time frame, $[\epsilon, T]$, to arrive then at time, T, at point s_T. Let us note in this context that the acceleration, $\bar{a}$, just gives a total value over $[0, T]$, and does not reveal how this total value has to be realised.

In order to identify the most fuel-efficient way to accelerate, it is obvious by $\mathsf{fc} \sim \bar{a}$ that we have to minimise the total required acceleration, $\bar{a}$. To this end, we now consider the main forces acting on the car. We propose that *frictional forces* and *aerodynamic resistance* are the main sources of fuel consumption. Since forces are acting (by fundamental physical principles) in an additive and independent way on the car, we may discuss them separately.

First, however, we first have another look at the basic mechanism behind fuel consumption itself. We consider (1) again, and formulate the fuel consumption using the total transformed fuel over the considered time frame. For the computation, we make explicit that, by taking the absolute of W, braking does not generate fuel:

$$\mathsf{fc} \sim \int_0^T |W(t)|\,dt \overset{W=F\cdot d}{\sim} \int_0^T |a(t)|\,dt = \int_0^T |\ddot{x}(t)|\,dt \tag{5}$$

To minimise fuel consumption during the transformation process to mechanical work, it is therefore optimal to uphold a constant velocity, since the minimiser of the above expression is obtained for $(\ddot{x} = 0) \Leftrightarrow (v = \text{constant})$.

Let us turn to frictional forces, $F_{friction} = \eta \cdot F_n$, where η is the friction coefficient and $F_n = m_c g$ is the normal force described by the constant mass of the car, m_c, and the gravitational constant, g. In this context, let us note that we indirectly assume that the

car moves approximately horizontally. Assuming that η is constant over our time frame, this means that, to negate the frictional forces, implies that a constant acceleration, a, is upheld—compare (3).

Turning to the aerodynamic resistance of a car, this may be modelled by a force, $F_{air} \sim v^2 = \dot{x}^2$. In order to minimise the required acceleration, negating aerodynamic resistance, we thus have to find the minimiser of

$$\min_x \int_0^T (\dot{x}(t))^2 \, dt \tag{6}$$

which is subject to the boundary conditions $x(0) = s_0$ and $x(T) = s_T$. The corresponding optimality condition reads as $2\dot{v} = 0$. Therefore, it is optimal with regards to fuel consumption to uphold a constant velocity in order to negate the aerodynamic resistance. This can be realised by a constant acceleration added to the one we found to be required to negate frictional forces.

As a consequence of our investigation, it is optimal to negate the fuel-consuming forces by a constant acceleration, thereby maintaining a constant velocity of the car. Since a constant acceleration implies $\dot{a} = 0$; one may detect instances of potential fuel-wasting driving by evaluating $\dot{a} = \ddot{v}$, and the driving style appears to be energy-efficient if $|\ddot{v}s.|$ is kept low by a driver. This result is in accordance with the idea that drivers aiming to maintain an energy-efficient driving style usually perform smooth accelerations and braking, whereas fast, energy-inefficient drivers tend to have a more abrupt driving style. Furthermore, dense urban traffic with the typical changes in acceleration and deceleration—which is notorious for leading to high fuel consumption—is represented by strong instances of $|\ddot{v}s.|$.

2.2. Invariances

As discussed, the quantity $\ddot{v} = \dot{a}$ describes the variation of the acceleration of a car. However, the positional GPS data gathered in our database allows us to formulate quantities that offer certain *invariances*. Our analysis should not depend on an absolute positioning and yield the same findings whether we analyse cars in Europe or in Asia. Such an invariance can be introduced by taking the derivatives of the movement. The velocity is independent of the exact location of a car. Furthermore, the acceleration of a car driving smoothly with $130\,\mathrm{km\,h^{-1}}$ is similar to a car driving with $50\,\mathrm{km\,h^{-1}}$. If the environmental circumstances are adequate, then both drivers should have the same classification.

This observation motivates the usage of a feature with a sufficiently large set of invariances. The quantity $\ddot{v}$ is invariant under affine transformations of the velocity and offers these benefits also.

2.3. Formalisation

We conclude our motivation by observing that the quantity $\ddot{v}$ is well known in physics as jerk. For convenience, we formalise this observation at hand of the following definition.

Definition 1 (Jerk function)**.** *The* jerk $j(t)$ *describes the third-order derivative of the position* $x(t)$ *with respect to time:* $j(t) := \frac{\mathrm{d}^3}{\mathrm{d}t^3} x(t)$.

Let us note that we refer here to jerk in terms of the derivative of a position x since our input data is given by positions.

We also remark that in [25], the authors use already the jerk function to classify drivers with respect to their aggressiveness, however, as we have shown, the jerk is also a reasonable physical quantity related to energy consumption during driving.

3. Driving Style Analysis by GPS Data

Relying exclusively on GPS data, as provided by navigation systems and tracking devices, comes with a certain number of hurdles that need to be overcome. Our goal is to

classify drivers with respect to their driving style. Currently, this analysis is performed offline. We first collect the data and store it in a database. The processing and classification is performed afterwards. To this end, we measure the drivers' spatio-temporal positioning.

Definition 2 below gives us a formal framework to work in. Positional data of the cars comes in the form of coordinate pairs—$(x_i, y_i) \in \mathbb{R}^2$. Each pair is accompanied by a label containing the time stamp, t_i, when the position has been recorded. This set of discrete samples gives us a complete description of the movement of a car in space and time. Yet, our focus lies on the analysis of moving cars. Therefore, we use a more specific structure to represent the displacement of a vehicle.

Definition 2 (Time frame; Movement pattern)**.** *We define a* time frame of size ℓ *a sorted vector* $T \in \mathbb{R}^\ell$ *containing* ℓ *individual time stamps in a non-decreasing order. We call the* movement pattern of size ℓ *a pair,* $\{T, P\}$*, consisting of a time frame,* T*, of size* ℓ *with the corresponding positional data* $P := \{(x_i, y_i) |\ i = 1, \ldots, \ell\}$*—if the car is not standing still during any moment within the complete time frame.*

Thus, we speak of a movement pattern if the car does not stop during a considered time interval. We elaborate a method to discard the idling objects from our dataset in the next section.

3.1. Data Preprocessing

In order to apply our method, we need to preprocess the given dataset of GPS logs. First of all, we discard corrupted or meaningless data. After the preprocessing, our data should only consist of movement patterns, as defined in Definition 2. To this end, we remove all logs for which the following conditions hold:

1. All GPS logs having the same longitude and latitude values as well as the same sampled time as their adjacent logs. These logs represent repeated GPS values being sampled at least twice due to hardware failure.
2. All GPS logs with same longitude and latitude and having different sampled time compared with their adjacent GPS logs. These GPS logs represent a car standing still.

Preprocessing is necessary to ensure reasonable findings. However, it also introduces holes in the displacement history of the analysed car. Certain records are contiguous over long time intervals, whereas others might only run for a few seconds before a gap occurs. Our forthcoming analysis is based on statistical measures of the movement pattern. To allow a meaningful and unbiased comparison, we additionally drop movement patterns that are too short and break down those that are too long into several smaller ones. In this work, we discard all movements patterns with less than 10 samples and split them if they exceed a length of 24. The motivation behind the chosen limits was to apply our standard-deviation-based proposed feature on a small window of GPS logs. Taking on a large number of GPS logs—namely, much more than 24—into account might result in the wrong results. This can be clearly explained by an example of a driver who completes a mix of slow and harsh driving styles. These styles may cancel each other out in a long trajectory, as our feature is constructed based on standard deviation as a quantity, expressing how much the members of a group differ from the mean value of the group. After the preprocessing, we obtained about 3225 movement patterns from the real-world dataset [24].

3.2. New Modelling and Algorithmic Adaptations

Let us implement our underlying considerations about jerk in the context of a typical example of GPS data.

As discussed, important physical quantities in the description of the behaviour of drivers are the position, velocity, and acceleration of their car. Figure 1 depicts an example of velocity, acceleration, and jerk obtained from a movement pattern of a calm driver, a racy driver, and one containing a noisy log. Even though a distinction among the three shown patterns could be deduced from the velocity and acceleration alone, the jerk has

the benefit of magnifying large, abrupt changes in the acceleration more than smaller ones; therefore, the noise detection is significantly simplified. This can clearly be observed in the noisy movement pattern shown in Figure 1. The previous arguments, together with our physical considerations, constitute the principal motivations to use the jerk as starting point in our work. Our exponentially based feature proposed in (7) takes advantage of the elaborated properties.

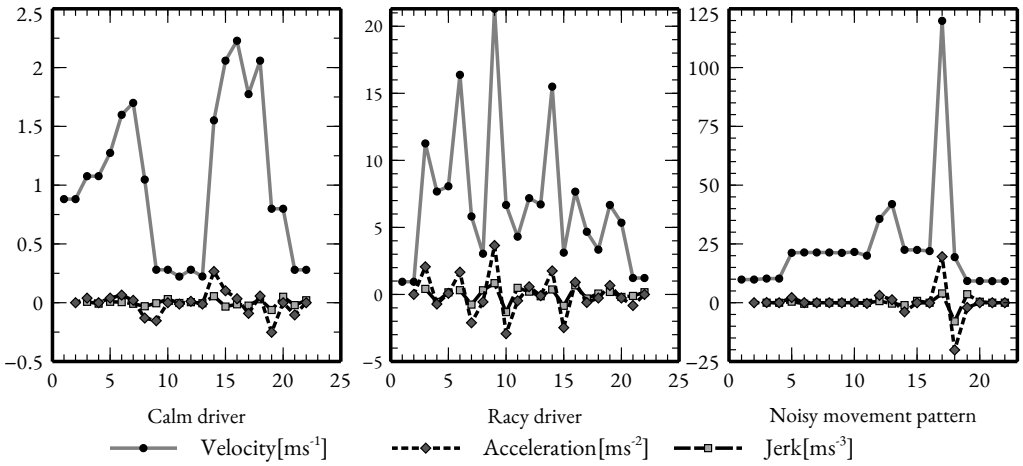

Figure 1. Velocity, acceleration, and jerk patterns corresponding to a calm and a racy driver in contrast to the same quantities of a movement pattern containing noisy GPS logs. Note that the y-axis in each plot has a completely different range. Either the acceleration or the jerk pattern could efficiently distinguish the calm driver from the racy driver. In case of the noisy movement pattern, either the acceleration or the jerk pattern could efficiently reveal the existence of a noisy GPS log. Our proposed exponentially based feature ω, cf. (7), shows high robustness with regard to noise. This helps to entirely accommodate noisy movement patterns in a separate cluster.

Modelling Details

Our proposed method seeks to classify driving style using an alternative formulation which also considers the jerk. We suggest the use of a clustering strategy exclusively based on the following feature:

$$\omega(\{T,P\}) := \exp\left(-\frac{1}{\sqrt{\sigma\left(j_{\{T,P\}}\right)}}\right) \tag{7}$$

Here, $j_{\{T,P\}}$ is the array of jerk values for each position in the movement pattern $\{T,P\}$. These are computed by means of standard finite difference schemes.

The function σ denotes the standard deviation. The motivation for proposing this feature is as follows. The exponential part of (7) discriminates those movement patterns with a very high fluctuation rate around their mean jerk values. They occur due to the existence of at least one noisy GPS log inside the movement pattern or a dubious driving style with a large number of strong accelerations and decelerations. Furthermore, in order to discern drivers with lower jerk fluctuation rates, we consider the square root of the standard deviation inside the exponential function. This facilitates the formation of a smooth elbow in our proposed feature ω, see Figure 2.

In our experiments, the modified jerk feature yielded the most intuitive and reliable clustering results. Small feature values correspond to those movement patterns representing drivers with fewer accelerations and decelerations in their driving patterns. An energy-saving driving style can therefore be identified by small feature values, whereas more racy drivers will usually exhibit larger feature values.

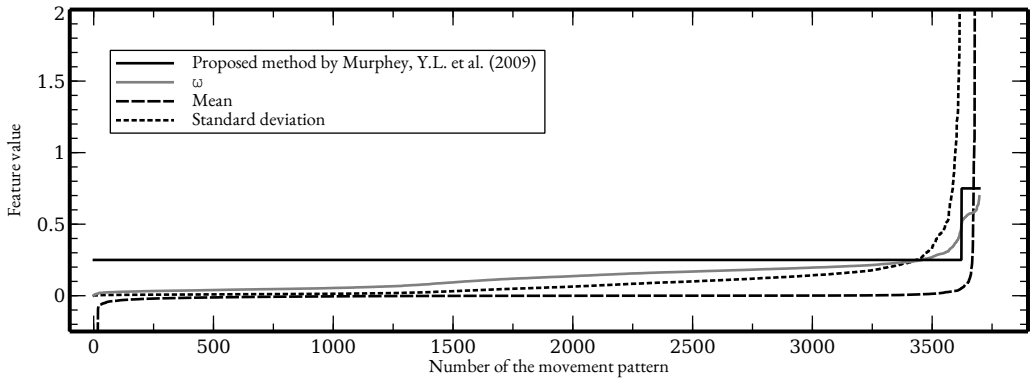

Figure 2. The sorted values of mean, standard deviation, the feature proposed by [25], and our proposed feature for all preprocessed movement patterns from [24]. The approach from [25] only yields two distinct classes and cannot detect noise. On the other hand, the curve representing the mean is almost flat, meaning it is very difficult to distinguish different clusters. The standard deviation depicts a similar behaviour to that of our proposed feature, but bears a less stable behaviour towards the end. Our feature yields a curve with a clear monotonic increase and a notable elbow at the end. Thus, we obtain the same segregation ability as the standard deviation but a clearer threshold for noise detection, which is given by the samples located in the elbow.

Within the context of driving style analysis—that boils down to unsupervised movement pattern clustering—one major obstacle is the varying size of the movement patterns recorded as GPS logs. Consequently, clustering of the movement patterns is impossible until they are all brought to a fixed-dimensional space or an efficient similarity measure is defined among them for the purpose of their comparisons [30]. Our proposed compact feature (7) resolves this length inconsistency problem and provides a meaningful semantic representation of the movement patterns with varying lengths. More precisely, (7) reduces the movement pattern dimensions from the range of $[10, 42]$, as already discussed in Section 3.1, to only one dimension, represented as ω, that later is used by an agglomerative clustering approach to label driver styles—as we will show in Section 4.

Algorithmic Details

Two classic and well-studied approaches to classify data are partitional and hierarchical clustering methods. The latter have the advantage to yield a complete-scale evolution for all possible numbers of clusters. Since the optimal number of clusters is hard to predict, we opt for a flexible approach using an agglomerative clustering method. Such a method starts by taking singleton clusters at the bottom level and continues merging two clusters at a time to build a bottom–up hierarchy of clusters.

We employ Ward's criterion for the merging strategy [31,32]. It makes use of the standard k-means squared error (SSE) to determine the distance between two clusters. For any two clusters, C_a and C_b, Ward's criterion is calculated by measuring the increase in the value of the SSE for the clustering obtained by merging them into a single cluster—$C_a \cup C_b$. The characteristic number of Ward's criterion is defined as follows:

$$\frac{|C_a||C_b|}{|C_a| + |C_b|} \sum_{\nu=1}^{\ell} \sum_{\eta=1}^{k} \left(c_{a\nu} - c_{b\eta} \right)^2 \tag{8}$$

where the cluster C_x has the individual components $c_{x\nu}$. The cardinality of a cluster is denoted by $|C_x|$.

We iteratively merge clusters in a bottom–up fashion. At the bottom level, each computed feature point is considered. Then, at each new level, we merge the pair of clusters that minimises (8). The algorithmic details of the agglomerative hierarchical clustering are also given in Algorithm 1.

Algorithm 1: Agglomerative hierarchical clustering of movement patterns

Data: A set of movement patterns.
Result: A clustering of the input movement patterns.
Compute the feature value ω from (7) for each movement pattern.
Place each feature value in its own cluster.
repeat
    Find pair of clusters that minimises (8).
    Merge this pair into a single cluster.
until *a single cluster is left*
return the smallest number of clusters, that yielded a decrease in (9) below a threshold T.

The optimal number of clusters is obtained through an L-curve heuristic [33,34]. We compare the number of clusters against the within-cluster sum of squares (WCSS) and consider the smallest number of clusters that yields a decrease in the WCSS below a given threshold. The WCSS is given by:

$$\sum_{a=1}^{M} \sum_{x \in C_a} \|x - c_a\|_2^2 \tag{9}$$

where M is the total number of clusters and c_a is the centroid of cluster C_a.

In our WCSS-based experiments, the optimal number of clusters was usually found to be four. In what follows, the silhouette index SI_i (10) is computed for members of each four established clusters:

$$SI_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{10}$$

Here, a_i is the mean distance between any cluster member ω_i to all other members in same cluster and b_i is the mean distance among the cluster member ω_i and all other members in next nearest cluster. The computed SI_i corresponding to ω_i takes negative or positive values in range of $[-1, +1]$ as being mostly dissimilar or similar to other members in same cluster. Well established clusters require silhouette indices to be positive over their members.

Our best clustering result is analysed with respect to its labelling in Figure 3b. The negative silhouette indices inside the 4th cluster represent the compactly accommodated noisy ω values. This observation validates our approach for identifying the noisy values. In addition, only a few members in the 2nd cluster are wrongly located making the final clustering result meaningful.

Let us emphasise that an alternative strategy would also be possible with the k-means algorithm. However, k-means needs to be restarted for each new amount of clusters and thus adds a significant computational burden if the number of clusters is not predetermined exactly.

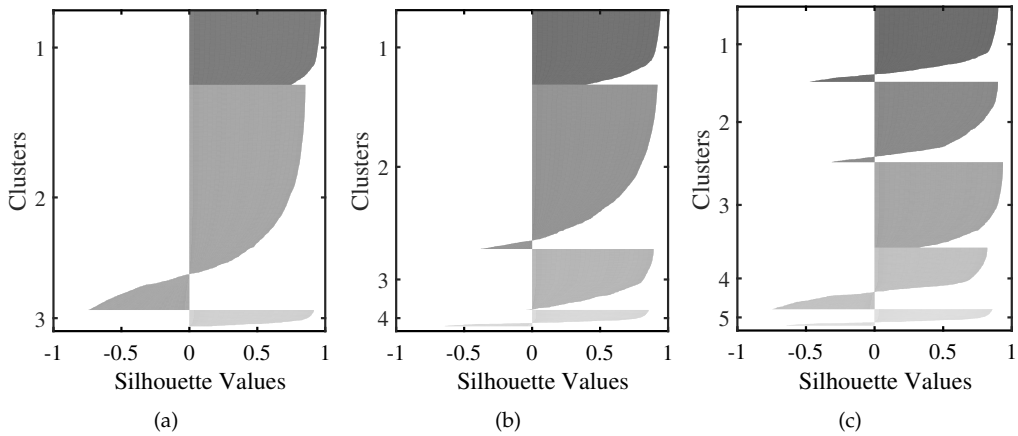

Figure 3. The silhouette indices of a clustering setup establishing (**a**) three, (**b**) four, and (**c**) five distinct groups of ω values of the central area drivers in Beijing. Inside each cluster, the more positive a silhouette index of a feature, the more similar is the feature to its group mates, which indicates a good clustering. (**a**) Here, the number of clusters is considered to be three. In this case, a group of drivers in the 2nd cluster is probably not optimally labelled. It is not possible to decide whether they belong to average-fuel-consuming drivers or to noisy patterns. Only the energy-saving drivers could be clearly located in the 1st cluster. (**b**) A minimal number of drivers with a negative silhouette index is found with 4 clusters. This clustering seems to have the best labelling. Noisy driver patterns are likely to be found in the 4th cluster. All energy-saving drivers are located in the 1st cluster and average-fuel-consuming drivers expand across two classes in the 2nd and 3rd cluster, respectively. (**c**) By increasing the number of clusters to five, drivers with negative silhouette indices appear even in the 1st cluster with small ω values. This depicts a clustering result which may not be considered as optimal, since almost all clusters accommodate dissimilar drivers inside them.

4. Experimental Results and Validation

In order to investigate the efficiency of our proposed feature, we consider a database with GPS logs from two distinct parts of Beijing city [24]. The whole database contains the trajectories of more than 10,000 taxis collected during the period of 2–8 February 2008. In total there are more than 17 million GPS logs that cover a total distance of around 9 million kilometres. Since the dataset corresponds to various taxi drivers, we expect to find many different driving styles in this dataset. The regions that we consider contain a vast area of Beijing city centre and a freeway running from east to west. We preprocess the logs as described in Section 3.1 and correspondingly obtain 3225 patterns for the city centre and 106 movement patterns for the freeway. All our feature values are computed according to (7).

Next, we evaluate our proposed approach by applying the agglomerative clustering detailed in Algorithm 1. The distribution of all feature values among the different clusters is visualised in Figure 4. As we can see, the feature values from the first cluster and the fourth cluster are clearly different from those in the second and third cluster. This confirms our expectations. The first cluster contains the energy-saving drivers, whereas the fourth cluster is supposed to contain noise as well as the high-fuel-consuming drivers. Finally, the majority of the drivers lies within the bounds of the second and third cluster and present drivers with average fuel consumption.

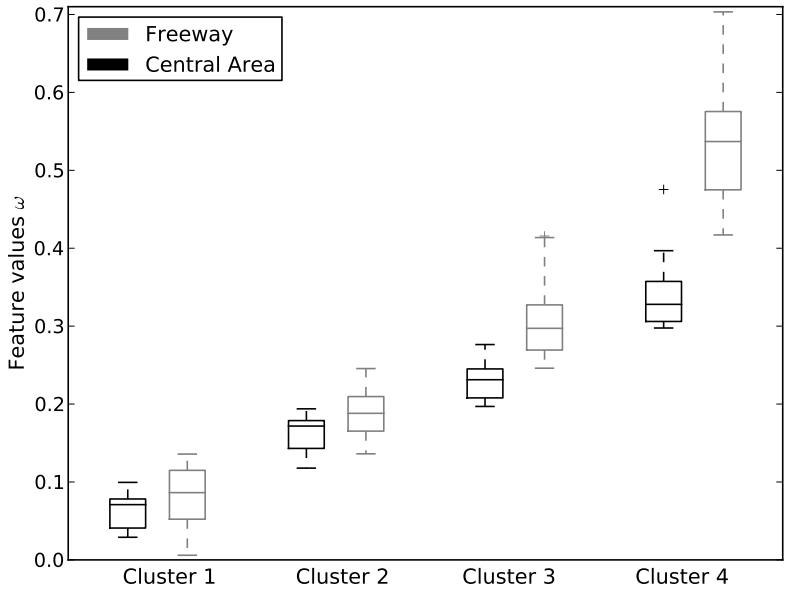

Figure 4. A whisker plot showing the distribution of the feature values among the different clusters for our two experimental setups. The box marks the boundaries of the first and third quartiles, while the bars indicate the full extent of the feature values in that cluster. The horizontal line indicates the median. As we can see, the first cluster and—on the other hand—the second and third cluster are well separated in terms of feature values. There is a continuous overlap from the second to the third cluster. These two clusters represent the majority of the drivers with an average-fuel-consuming driving style, with a tendency towards being more (cluster 2) or less (cluster 3) energy efficient. The first cluster represents the energy-saving drivers and the last cluster represents the drivers with high fuel consumption, also containing noisy driver patterns. Thereby, the lower the feature values are, the more energy saving the corresponding driving style is.

4.1. Findings for Beijing City Centre Area

The first, second, and third clusters are totally noise-free and could be adopted as an accurate driver's behavioural model. The within-cluster sum of squares (WCSS) index does not show any remarkable amount of reduction by adding a fifth cluster or more. Hence, according to our L-term heuristic, we should set the final number of clusters to four. The noisy patterns are located fully in the fourth cluster.

Let us stress that our algorithmic proceeding that results in setting the number of clusters to four is confirmed in an independent way by evaluating the silhouette indices as shown in Figure 3. This is exemplarily shown here for the data of the Beijing city centre area, but also holds for the other analysed data.

4.2. Findings for the Freeway Area

The first and second clusters are noise free. The L-term strategy suggests the use of four clusters in this case. Since we only have very few feature values in this area, our algorithm does not yield a cluster solely containing noise. However, the fourth cluster partially contains the noisy movement patterns. Such a cluster was present for the Beijing centre area, where the fourth cluster was filled with corrupted patterns only. This observation also shows that our classification algorithm benefits from having an extremely large amount of samples. This is a realistic scenario in an industrial application where large amounts of data can be collected, e.g., via navigation systems. The more samples we can process, the more clearly we can distinguish noise and classify individual drivers.

Summary on Comparison with Other Possible Approaches

We compare the efficiency of our proposed feature against the work proposed in [25] by applying both of them on GPS logs of Beijing city centre provided in [24]. In addition,

we visualize our proposed feature applied on GPS logs obtained from the Beijing city centre to other commonly used feature choices. The findings are visualised in Figure 2. We compare our suggested feature against the mean value and standard deviation as well as the jerk-based feature from [25]. The approach from [25] (proposed for the purpose of analysing driver behaviour, not for energy efficiency) yields clear-cut results. It segregates our drivers into two categories, namely defensive and aggressive drivers—by the context of the work in [25]. For this classification, we used the parameter suggestions from [25]. We selected a window size of 10 s and set the parameters $\text{norm}_{\text{threshold}}$ and $\text{agg}_{\text{threshold}}$ to 0.5 and 1, respectively. This method yields a very clear threshold that discerns the two classes of drivers. However, it is unable to detect noisy data. If we consider the average value and standard deviation as features, then we are able to detect the corrupted data, but a discrimination of the remaining drivers becomes difficult. All the feature values that we compare against our proposed feature are very close to each other. Finally, our proposed feature yields a larger range of values as well as a clear jump at the end of the curve. Thus, we are in a position to distinguish different driving styles as well as filter out noisy patterns.

5. Summary and Conclusions

In this paper, we show that it is possible to discern different driving style patterns with respect to their energy consumption from their GPS logs alone. To this end, we use a dedicated variation of the jerk feature and combine it with a hierarchical clustering approach.

Our model is quite simple but still capable of classifying drivers into different categories and filtering out noisy data logs. Let us note again that the use of just the GPS logs may potentially be highly relevant for industrial use of our results in the context of navigation systems.

In the future, we aim to combine the driver model information with other optimisation tools concerned with energy-efficient routing and motor control in order to improve the energy efficiency of hybrid cars.

Author Contributions: The conceptualization, funding acquisition, project administration, supervision and writing the original draft are all performed by M.B. Data curation is conducted by A.M.Y. and A.S.B. The tasks of formal analysis, investigation, methodology, software, validation, visualization, review and editing are performed by Ashkan Mansouri Yarahmadi. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the German Federal Ministry of Education and Research (BMBF), funding code 05M13ICC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used GPS data are collected as part of T-drive that is a smart driving direction services based on GPS trajectories of a large number of taxis. The service helps its user to practically find out quickest possible path to a destination. We used the real-world trajectory dataset generated by 30,000 taxis in Beijing in a period of 3 months. Further details and download link can be found in reference [24] of the current paper or under the https://www.microsoft.com/en-us/research/project/t-drive-driving-directions-based-on-taxi-traces.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Merting, S.; Schwan, C.; Strehler, M. Routing of Electric Vehicles: Constrained Shortest Path Problems with Resource Recovering Nodes. In Proceedings of the 15th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems, Patras, Greece, 17 September 2015; Volume 48, pp. 29–41.
2. Pickenhain, S.; Burtchen, A. Optimal Energy Control of Hybrid Vehicles. In Proceedings of the 6th International Conference on High Performance Scientific Computing, March 16–20. Hanoi, Vietnam, 2015;

3. Yuniar, D.; Djakfar, L.; Wicaksono, A.; Efendi, A. Truck driver behavior and travel time effectiveness using smart GPS. *Civil Eng. J.* **2020**, *6*, 724–732. [CrossRef]
4. Pentland, A.; Liu, A. Modeling and Prediction of Human Behaviour. *Neural Comput.* **1999**, *11*, 229–242. [CrossRef] [PubMed]
5. Quintero, M.C.G.; Lopez, J.O.; Pinilla, A.C.C. Driver behavior classification model based on an intelligent driving diagnosis system. In Proceedings of the 15th International IEEE Conference Intelligent Transportation Systems, Anchorage, AK, USA, 16–19 September 2012; pp. 894–899.
6. Quintero, M.G.; López, J.A.O.; Rúa, J.M.P. Intelligent erratic driving diagnosis based on artificial neural networks. In Proceedings of the IEEE ANDESCON Conference, Bogota, Colombia, 15–17 September 2010; pp. 1–6.
7. Rigolli, M.; Williams, Q.; Gooding, M.J.; Brady, M. Driver behavioural classification from trajectory data. In Proceedings of the International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria, 16 September 2005; Volume 6, pp. 889–894.
8. Sathyanarayana, A.; Boyraz, P.; Hansen, J.H.L. Driver behavior Analysis and route recognition by Hidden Markov Models. In Proceedings of the IEEE International Conference on Vehicular Electronics and Safety, Columbus, OH, USA, 22–24 September 2008; pp. 276–281.
9. Imamura, T.; Yamashita, H.; bin Othman, M.R.; Zhang, Z.; Miyake, T. Driving behavior classification river sensing based on vehicle steering wheel operations. In Proceedings of the SICE Annual Conference, Chofu, Japan, 20–22 August 2008; pp. 2714–2718.
10. Verroios, V.; Vicente, C.M.; Delis, A. Alerting for vehicles demonstrating hazardous driving behavior. In Proceedings of the 11th ACM International Workshop on Data Engineering for Wireless and Mobile Access, Arizona, Scottsdale, 20 May 2012; pp. 894–899.
11. Michon, J. A Critical View of Driver Behavior Models: What Do We Know, What Should We Do? In *Human Behavior and Traffic Safety*; Springer: New York, NY, USA, 1985; pp. 485–524.
12. 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* **2007**, *95*, 427–437. [CrossRef]
13. Liao, Z.; Yu, Y.; Chen, B. Anomaly detection in GPS data based on visual analytics. In Proceedings of the IEEE Symposium on Visual Analytics and Science Technology, Salt Lake City, UT, USA, 25–26 October 2010; pp. 51–58.
14. Zhang, D.; Li, N.; Zhou, Z.H.; Chen, C.; Sun, L.; Li, S. iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces. In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 17–21 September 2011; pp. 99–108.
15. Carboni, E.; Bogorny, V. Inferring Drivers Behavior through Trajectory Analysis. In *Intelligent Systems' 2014*; Springer: Cham, Switzerland, 2015; pp. 837–848.
16. Banerjee, T.; Chowdhury, A.; Chakravarty, T.; Ghose, A. Driver authentication by quantifying driving style using GPS only. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; pp. 1–6.
17. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
18. Kashevnik, A.; Lashkov, I.; Ponomarev, A.; Teslya, N.; Gurtov, A. Cloud-based driver monitoring system using a smartphone. *IEEE Sensors J.* **2020**, *20*, 6701–6715. [CrossRef]
19. Lindow, F.; Kaiser, C.; Kashevnik, A.; Stocker, A. Ai-based driving data analysis for behavior recognition in vehicle cabin. In Proceedings of the 2020 27th Conference of Open Innovations Association (FRUCT), Trento, Italy, 7–9 September 2020; pp. 116–125.
20. Vlachogiannis, D.M.; Vlahogianni, E.I.; Golias, J. A reinforcement learning model for personalized driving policies identification. *Int. J. Transp. Sci. Technol.* **2020**, *9*, 299–308. [CrossRef]
21. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
22. Cortes, C.; Vapnik, V. Support Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
23. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian Network Classifiers. *Mach. Learn.* **1997**, *29*, 131–163. [CrossRef]
24. Yuan, J.; Zheng, Y.; Zhang, C.; Xie, W.; Xie, X.; Huang, Y. T-Drive: Driving Directions Based on Taxi Trajectories. In Proceedings of the ACM Sigspatial GIS 2010, San Jose, CA, USA, 2–5 November 2010; pp. 99–108.
25. Murphey, Y.L.; Milton, R.; Kiliaris, L. Driver's style classification using jerk analysis. In Proceedings of the IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems, Nashville, TN, USA, 30 March–2 April 2009; pp. 23–28.
26. Ross, S.M. Chapter 2—Descriptive statistics. In *Introduction to Probability and Statistics for Engineers and Scientists*, 6th ed.; Ross, S.M., Ed.; Academic Press: Cambridge, MA, USA, 2021; pp. 11–61. [CrossRef]
27. Zuo, W.; Guo, C.; Liu, J.; Peng, X.; Yang, M. A Police and Insurance Joint Management System Based on High Precision BDS/GPS Positioning. *Sensors* **2018**, *18*, 169. doi: [CrossRef] [PubMed]
28. Ayuso, M.; Guillen, M.; Pérez-Marín, A. Using GPS data to analyze the distance traveled to the first accident at fault in pay-as-you-drive insurance. *Transp. Res. Part C Emerg. Technol.* **2016**, *68*, 160–167. doi: [CrossRef]
29. Breuß, M.; Sharifi Boroujerdi, A.; Mansouri Yarahmadi, A. Energy-Efficient Driving Model by Clustering of GPS Information. In Proceedings of the Manuscript submitted for publication to International Conference on Operations Research, Xi'an, China, 27–29 May 2022.
30. Morris, B.T.; Trivedi, M.M. Understanding vehicular traffic behavior from video: A survey of unsupervised approaches. *J. Electron. Imaging* **2013**, *22*, 041113. doi: [CrossRef]
31. Ward, J. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
32. Wishart, D. An algorithm for hierarchical classifications. *Biometrics* **1969**, *25*, 165–170. [CrossRef]

33. Hansen, P.C. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev.* **1992**, *34*, 561–580. [CrossRef]
34. Hansen, P.C.; O'Leary, D.P. The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems. *SIAM J. Sci. Comput.* **1993**, *14*, 1487–1503. [CrossRef]