Derivation of the Shannon Entropy for Parameters that Captures Driving Tasks

Research Paper Business Analytics

Author: Marcoen Beeks (2581962)

Vrije Universiteit, Amsterdam, Netherlands
October 1, 2018

Supervisors: Drs. Emiliano Heyns,
Dr. Elenna R. Dugundji

Derivation of the Shannon Entropy for Parameters that Captures Driving Tasks

Marcoen Beeks

Research Paper

Vrije Universiteit Amsterdam
Faculty of Science
Business Analytics
De Boelelaan 1085

1081 HV Amsterdam

October 2018

Abstract

Nowadays, traffic congestion causes much problems in different fields and places. For that reason a well-flowing road network is vital to avoid extremely consequences. This paper will make usage of car sensor data to say something about the complexity of driver's behaviour and if these findings can be helpful to reduce the effects of traffic congestion. The driver behaviour is derived by the Shannon Entropy for time blocks of 30 seconds for different kind of parameters. For this research 3 datasets with several trips were available, however a single trip that was on the Dutch highway is investigated. The Shannon entropy from features in the category *Car Mode Settings* does not show very appealing results, since there is hardly any diversification in the usage during trips. Therefore the complexity of the driver behaviour can not be derived from these features. For the parameters in the Category *Vehicle Information* the mean Shannon entropy shows 11 timestamps where the complexity of the driver behaviour is clearly higher than the other timestamps. Half of these findings corresponds with a higher traffic density based on the data from the Nationale Databank Weggegevens.

Keywords: Traffic Congestion, Driver Behaviour, Shannon Entropy

Contents

1 Introduction

Many studies have shown that the effects of traffic congestion has major consequences in economics and the environment [1][2]. The Netherlands is a densely populated area where a well-flowing road network is essential. This will be confirmed in the coming years, especially since the Netherlands expects a grow of 38% over the next 5 years from the negative impact of traffic jams and incidents on travel times [3]. With these expectations in prospect VIA NOVA adds data quality research to "Talking Traffic" and this can be a possible solution for traffic congestion. Talking Traffic is a concept of car sharing data where it gets immense amounts of information about the infrastructure and it enables vehicles to communicate with the infrastructure directly and the other way around. However, the quality and its value of this data is not being used to the limit. Therefore, we investigate if the data contains enough information to tell something about the behaviour of drivers and if this insight can be helpful to reduce the effect of traffic congestion.

In this research paper, which is part of the VIA NOVA project, we analyse the behaviour of the driver from a single car that moves on the Dutch highways. Therefore, a measurement to derive the complexity of the driver's behaviour will be introduced. To chart this complexity we utilize the Shannon Entropy. The Shannon Entropy is a measure to get the amount of uncertainty in communication processes [4]. The entropy will be given for blocks of fixed time lengths of 30 seconds for each feature. When the complexity of the driver's behaviour can be derived from the Shannon Entropy, this may lead to more insights of the current traffic situation. Moreover, the results are perhaps helpful for predicting traffic congestion.

This paper starts with Section 2 where some related literature will be discussed. Section 3 mentions all the relevant information about the provided data. In the next Section 4 the utilized methodology will be explained. Afterwards, some of the results will be shown in Section 5 and there will be some room for discussion in Section 6.

2 Related work

The foundation in the field of information theory was introduced in the article "A Mathematical Theory of Communication" written by Claude E. Shannon and published in 1948 [5]. In this article the fundamentals of communication processes were discussed that led to an introduction of the concept information entropy and the term bit as unit of information. In the next couple of years it became clear that his findings has led to new applications in other science fields (e.g Economics, Psychology, Psychics and many more) [6]. With the rise of Big Data it is most likely that it will continue to be applied in even more sciences in the future.

Nowadays, data is coming from everywhere and this needs careful usage to make well-considered decisions. According to [7], the challenge with these huge amounts of data is to enable an efficient management that will results in quality outcomes and solutions. Especially the transportation sector is facing this challenge. An important issue here is to determine the behaviour of drivers on the basis of car sensor data.

Studies have shown a statistical analysis of the driver's behaviour before. In [8], a translation of the data into a time series has been done. They have used statistical measures such as the mean and variance for classifying the driver's behaviour. They came up with four different states of driver's behaviour: fast, relatively fast, slow and very slow. In [9], they also utilized time series analysis to identify different traffic states on highways. Only in this paper the described states include free-flow, synchronized and stop-and- go traffic. In this research paper the measure technique for determining the driver's behaviour will be much different.

Furthermore, many related work shows prediction methods for congestion related problems [10]. These methods are often based on the amount of information (Shannon Entropy) each feature has. For Instance, the Decision Tree is a commonly used technique which uses entropy for building the tree. In these papers they mainly focussed on the accuracy of the prediction models, however analysis of the underlying method is something that has not been deeply researched before. Therefore, we want to understand if the Shannon Entropy could give a possible advantages and insights for interpreting driver behaviour.

3 Data understanding

This section contains all the relevant information about the data and its analysis and applied aggregations. The provided data includes sensor measurements of cars driving through the Netherlands. Each row represents an observation for every 0.2 second. The features can be distinguished into the following categories: *Car Mode Settings*, *GPS Information*, *Vehicle Information* and *Road Information*.

3.1 Feature selection

To understand the behaviour of the driver only the features are selected where data is coming from the sensors. Table 1 shows the selection and its characteristics. It applies for the features into the category *Car Mode Settings* that the sensors give information once per second. However, the *Vehicle Information* features contain an observation for every 0.2 second. Furthermore, we assume that the features in the categories *GPS Information* and *Road Information* do not influence the drivers behaviour, so the Shannon Entropy will not be derived for these features.

Table 1: Features and its Characteristics

Feature Name	Category	Data type
Attitude_speed_km_h	Vehicle Information	Double
Attitude_yaw_deg_c	Vehicle Information	Double
Inputs_brakePedal_mm	Vehicle Information	Double
Inputs_steeringAngle_deg	Vehicle Information	Double
Inputs_throttlePedal_pct	Vehicle Information	Double
Inputs_ignitionState	Car Mode Settings	Boolean
Inputs_modeE	Car Mode Settings	Boolean
Inputs_modeL	Car Mode Settings	Boolean
Inputs_modeON	Car Mode Settings	Boolean
Inputs_shiftD	Car Mode Settings	Boolean
Inputs_shiftN	Car Mode Settings	Boolean
Inputs_shiftP	Car Mode Settings	Boolean
Inputs_statusAC	Car Mode Settings	Boolean

3.2 Trip selection

The provided data consists of individual trips of three cars that drive on the Dutch road network. This includes all kind types of roads. Since this paper is focused on the Dutch Highways a subset is created on the road types "Motorway" and "Motorway-link". This filters out the sensor data when the car is not driving on the Highways. Moreover, the cars are owned by companies, so they are not constantly driven by the same person. This means there is no deep explanation about the driving style of individuals.

The next step is to pick a trip that consists of enough information. When choosing a trip that exists only of sensor data when the car is driving a constant speed on the Dutch highway, it is hard to say something about the driver's behaviour. Therefore, a trip is needed where it is certain that it has some fluctuation in speed. Fig. 1 shows an acceptable fluctuation when focused on the car speed. For that reason we consider this visualization as a decent trip for measuring the Shannon Entropy for each feature. This particular trip took place at 6 April 2018 in the afternoon. Car number 1 was driving on the highways from "Gouda" to the German border nearby "Hengelo". The black line shows a time series of the car speed per second and the red line contains the maximum road speed. The remaining dataset consists of 25773 rows (sensor observations) and 28 features.

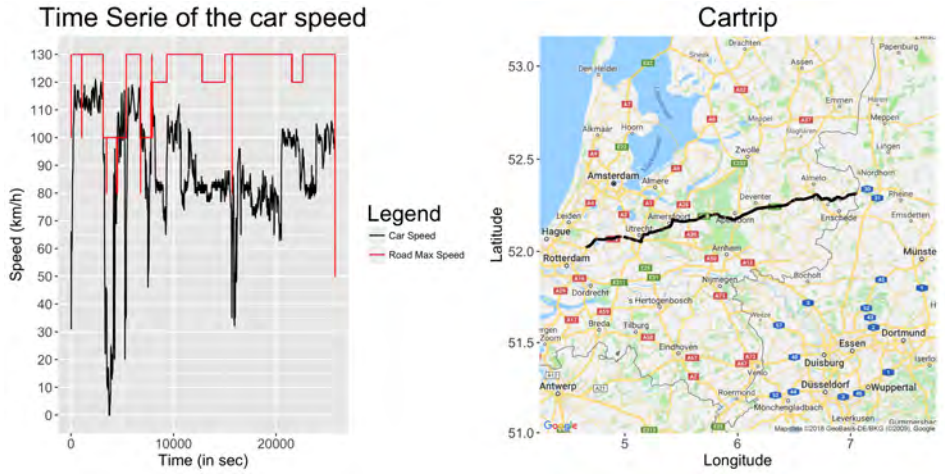

Figure 1: Selected Trip, **left**: Time Serie of the car speed, **right**: Roadmap

3.3 Time block selection

As discussed before the data holds observations for each 1 or 0.2 second. Each Time Block where the Shannon Entropy will be derived has a fixed duration of **30 seconds**. The logic behind the 30-second blocks is when the blocks are too long e.g. 5 minutes, it is too late to act on the found information to avoid congestion related traffic situations and when the blocks are too short the Shannon Entropy depends on too little information.

This means for the variables into category *Car Mode Settings* the Shannon Entropy will be calculated based on 30 observations. In the same way the Shannon Entropy will be calculated for the features *Vehicle Information*, but now it is based on 5*30= 150 observations.

4 Methodology

In this Section first some background knowledge will be given about the storage process from information. Afterwards we will convert this information into Discrete Markov Processes. Finally the Shannon Entropy will be introduced to quantify the amount of information.

4.1 Information storage

Information can be considered of variables that can take on different values. Computers store this information using bits. These bits represent "0" or "1". In case of boolean outcomes 1 bit is enough to store the data, since the boolean variable only has two possible outcomes (e.g. TRUE or FALSE, ON or OFF, 0 or 1). When a variable can take on more than 2 outcomes the amount of bits that is needed also increases. For instance, a variable which can take on 4 different outcomes need 2 bits for storage of all possible outcomes. In Eq. 1 the exponential relationship is shown. In the next subsection we will discuss how to deal with the corresponding probabilities of each possible outcome.

$$y = 2^n \tag{1}$$

Where,

y: is the amount of possible outcomes
n: is the amount of needed bits

4.2 Discrete Markov Process

In case of the *Car Mode Settings* features, the data will be given as boolean expressions. Briefly, this means the corresponding Car Mode Setting is always ON or OFF each time measurement. During the blocks of time the sensors keep up the frequency of the Car Mode Settings. For example, for a 60 seconds time block the sensor measures 45 times the corresponding Car Mode Setting is ON and 15 times it is OFF. Keep in mind, a measurement is once per second. The next step is to compute the interrelated probabilities. To gain a simple graphical representation we can demonstrate this in a discrete Markov process. In Fig. 2 this done for our example. So the probability that the process goes from any state to state ON is $45/60 = 0.75$ and it switches with a probability of $15/60 = 0.25$ to state OFF.

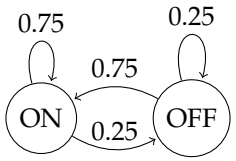

Figure 2: Markov Process with corresponding probabilities

4.3 Discretization of continuous features

Concerning the continuous features which include the *Vehicle Information* a discretization process is needed. This procedure normally consists of four steps: (i) sort all the continuous values of the feature to be discretized (ii) choose a cut point to split the continuous values into intervals. (iii) split or merge the intervals of continuous values (iv) choose the stopping criteria of the discretization process [11].

The most straightforward method for discretization is equal width interval binning. It sorts all values and breaks it into n equal sizes. Eq. 2 shows how it will be applied. There exists a lot of other discretizations method, but this will be out of of scope for this research paper.

$$\delta = \frac{x_{max} - x_{min}}{n} \tag{2}$$

4.4 Shannon Entropy

At this moment all the features can be recognized as Discrete Processes and all the probabilities of occurrences for each fixed time blocks are known. The next step is to quantify the information that is produced from the sensors. For each feature we have possible events in a set called T. In set T each event has its own probability of occurrence, say $p_1, p_2, .., p_n$. The quantification will be expressed in bits and therefore we need Shannon's formula for Entropy.

$$H(p) = -\sum_{i}^{n} p_i * \log_2 * p_i \tag{3}$$

To gain insight in this procedure an example will be shown in Table 2. Notice when the diversification in probabilities becomes lower the Shannon Entropy goes to 0. This means the uncertainty of the outcome will become lower.

Table 2: Probability of Occurrences and the corresponding Shannon Entropy

T	Low Speed	Normal Speed	High Speed	Shannon Entropy
19:00:00-19:00:30	0.333	0.333	0.333	1.58482
19:00:30-19:01:00	0.167	0.333	0.5	1.459481
19:01:00-19:01:30	0	0.333	0.667	0.9179621
19:01:30-19:02:00	0	0.167	0.833	0.6507958
19:02:00-19:02:30	0	0	1	0

5 Results

In this Section the results will be discussed for the categories separately. Finally the results of the features into Section 5.2 will be compared with the requested traffic density from the NDW (Nationale Databank Weggegevens) to find interesting correlations.

5.1 Car Mode Settings

The results of the features in category *Car Mode Settings* are shown in Fig. 3. We notice only dark green ("Inputs_modeL") data points with a entropy higher than 0. The brown data points ("Inputs_modeE") are lying behind the dark green points. The reason for this is that these settings are contraries (e.g. when ModeL is on, modeE is switched off). For all other data points in the plot the entropy is 0. This means there is almost no uncertainty about the *Car Mode Settings* in each time block. Remarkable to mention is the gap of data points between 15:00 and 15:30. In this time block the car was not driving on the highway. In Summary, the results for this category are not very appealing. The most probable reason for this is that the settings barely change during trips. This means there is not really a complex driver behaviour based on this category. In the next subsection we will discuss the results from the features that captures *Vehicle Information*.

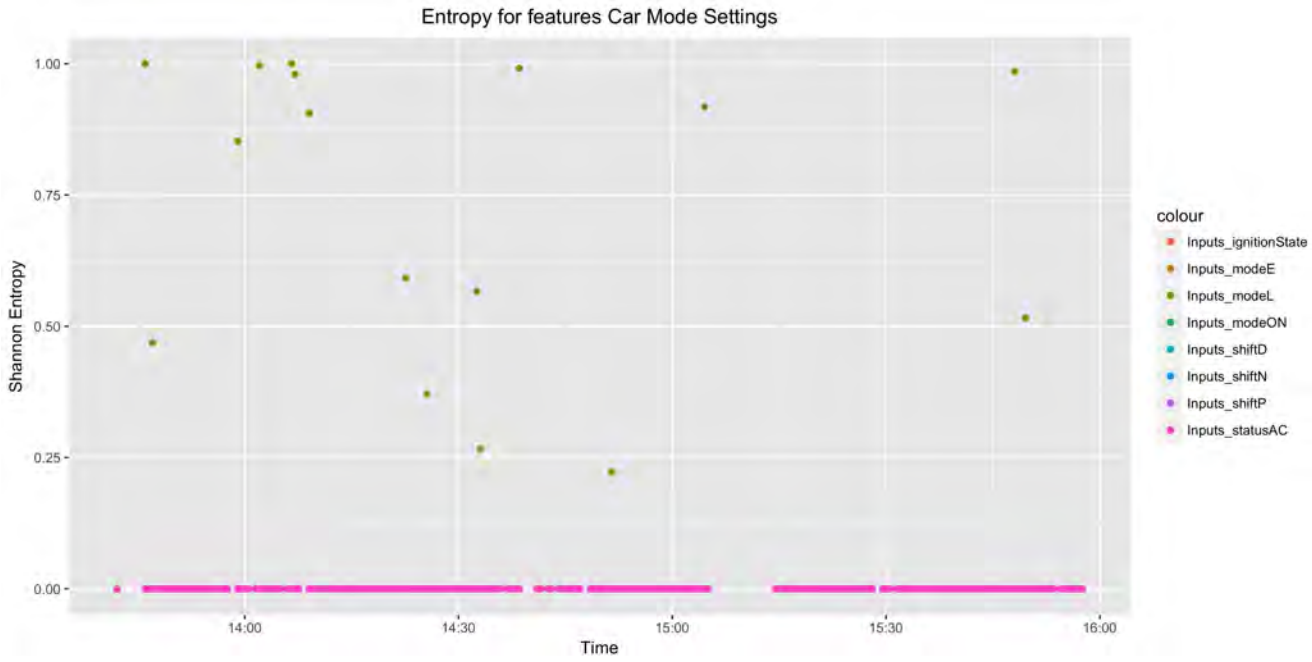

Figure 3: Entropy of features Car Mode Settings

5.2 Vehicle Information

Due to the fact that this features originally consist of continuous variables we observe more uncertainty in comparison with the features discussed in Section 5.1. For the discretization we have used $n = 4$ intervals. Important to mention is when the the amount of intervals increase (n becomes bigger) the uncertainty will also increase. In the first plot from Fig. 4 the results are presented from the features into category *Vehicle Information*. If we focus on the features Brake, Speed and Yaw we notice barely any diversification. The Throttle shows extremely diversification in usage in contrast to the feature Steering Wheel. The entropy from the feature Throttle indicates that the car does not make use of cruise control during this trip.

To discover the complexity of the driver behaviour it makes sense to take all the features into consideration. Therefore we have measured the mean of the Shannon Entropy in each time block. The red line in the second plot from Fig 4 belongs to the results. We notice 11 peaks with a Shannon Entropy value of above 0.5. In Table 3 the corresponding time blocks are given. In these blocks the complexity of the driver's behaviour is at its highest point.

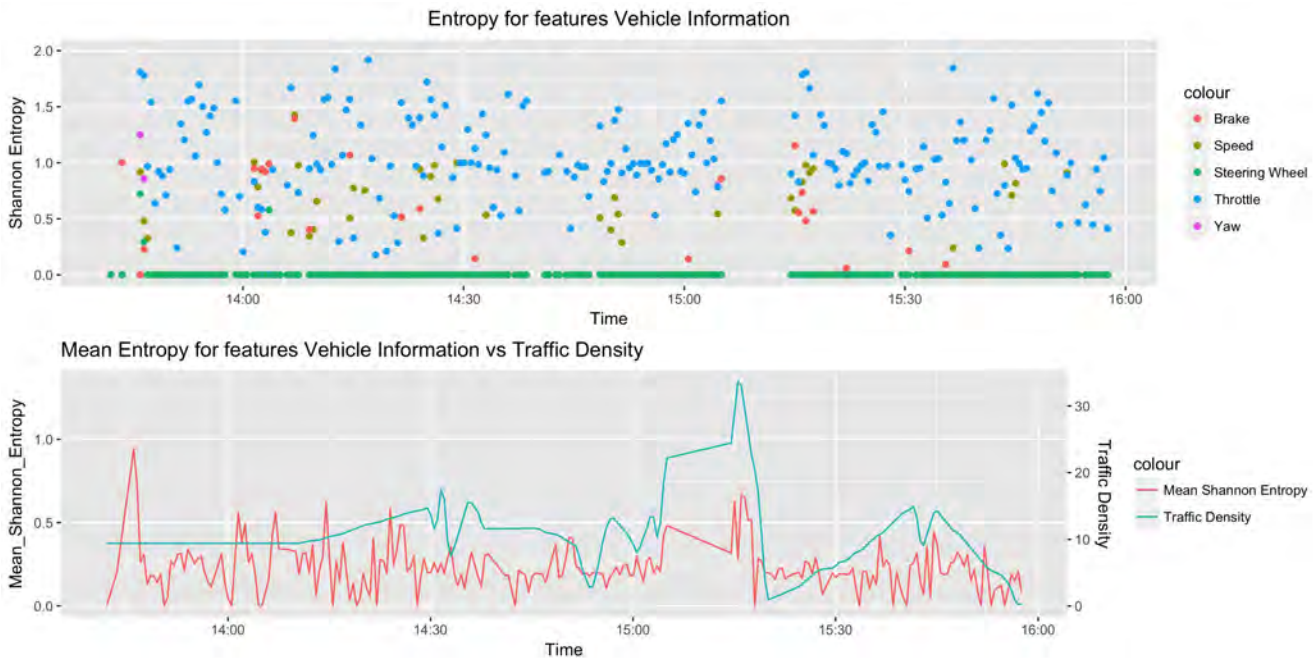

Figure 4: Entropy Vehicle Information

Finally, the mean of the Shannon Entropy is compared with the traffic densities from the NDW database. The blue line in the second plot from Fig. 4 shows this density. It is noticeable that there is peak around 15:15. In Table 3 all the timestamps with a Shannon entropy of above 0.5 and the corresponding densities are given. It is clear for the timestamps between 15:15 and 15:18 that the complexity of the driver matches with more traffic on the road. For the other timestamps there is no obvious explanation for the high entropy that accompanies the relative low density.

Table 3: All 30 seconds Time Blocks with a mean Shannon Entropy above 0.5 and the corresponding traffic densities

	Timestamp	Mean Shannon entropy	Traffic density
3	2018-04-06 13:46:02	0.94	9.39
4	2018-04-06 13:46:32	0.73	9.39
31	2018-04-06 14:01:32	0.56	9.39
41	2018-04-06 14:07:02	0.56	9.39
54	2018-04-06 14:14:32	0.63	10.24
73	2018-04-06 14:24:02	0.59	13.15
148	2018-04-06 15:15:02	0.63	29.00
150	2018-04-06 15:16:02	0.67	33.03
151	2018-04-06 15:16:32	0.65	29.58
152	2018-04-06 15:17:02	0.52	26.14
153	2018-04-06 15:17:32	0.52	22.71

6 Discussion

To interpret the behaviour from car drivers some complications have to discussed. In the first place this research paper was focussed on a single trip. A larger sample of trips will give an better approximation of the driver's behaviour, since persons can have different driving styles. For example, one driver can be very aggressive in using the throttle and brake and someone else barely use both of them. With a larger sample we will get a more general usage of the car and this reduce exceptional behaviour. However, the additional costs for putting sensors on more cars will increase. The challenge is to cancel out these costs with the reduction of the financial consequences from traffic congestion.

Another point to mention is about the results from the features in the category *Car Mode Settings*. Unfortunately these results were not very appealing, however with the growing number of new settings in cars this may change in the future. For example, a lot of new cars have settings like Eco, Sports and Off-road mode, which can be turned on or off during a trip. The more settings the new cars having the more likely it is that there will be more diversification in these settings. So for the future the results for this category may become more attractive.

References

[1] R Robinson. Problems in the urban environment: traffic congestion and its effects. *Wollongong Studies in Geography*, 14, 1984.

[2] Amal S Kumarage. Urban traffic congestion: the problem and solutions. 2004.

[3] Ivo Schrijer. VIA NOVA – Car in the Cloud. Technical report, Hogeschool van Arnhem en Nijmegen HAN University of Applied Sciences.

[4] Alan R Rogers. Shannon Uncertainty and Information. 2018.

[5] C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[6] C.E. Shannon. The bandwagon (Edtl.). *IRE Transactions on Information Theory*, 2(1):3–3, March 1956.

[7] H. Gilbert Miller and Peter Mork. From Data to Decisions: A Value Chain for Big Data. *IT Professional*, 15(1):57–59, January 2013.

[8] M. E. Fouladvand and A. H. Darooneh. Statistical analysis of floating-car data: an empirical study. *The European Physical Journal B*, 47(2):319–328, September 2005.

[9] L. Neubert, L. Santen, A. Schadschneider, and M. Schreckenberg. Single-vehicle data of highway traffic - a statistical analysis. *Physical Review E*, 60(6):6480–6490, December 1999.

[10] Gys Meiring and Hermanus Myburgh. A Review of Intelligent Driving Style Analysis Systems and Related Artificial Intelligence Algorithms. *Sensors*, 15(12):30653–30682, December 2015.

[11] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and Unsupervised Discretization of Continuous Features. In *Machine Learning Proceedings 1995*, pages 194–202. Elsevier, 1995.

Appendix

Table 4: Description of the car sensor data

Variable name	Description
Car	ID
Trip	ID (start timestamp)
Timestamp	Date and time of measurement
Attitude_latitude_g	Lateral acceleration in g
Attitude_longitude_g	Longitudinal acceleration in g
Attitude_speed_km_h	Vehicle speed in km/h as measured by speedometer
Attitude_yaw_deg_c	Yaw rate in degrees per second
GPS_height_m	Height in meters
GPS_latitude_deg	Latitude in degrees
GPS_longitude_deg	Longitude in degrees
GPS_numSatellites	Number of satellites found at time of establishing lat/lon
GPS_speed_km_h	Vehicle speed in km/h as measured by GPS
Inputs_brakePedal_mm	Amount of brake pedal input in mm
Inputs_ignitionState	Boolean whether ignition is on or off
Inputs_modeE	Boolean whether car is in "economy" mode
Inputs_modeL	Boolean whether car is in "low gear" mode
Inputs_modeON	Boolean whether car is in "drive" mode
Inputs_shiftD	Boolean whether shift is in "drive" mode
Inputs_shiftN	Boolean whether shift is in "neutral" mode
Inputs_shiftP	Boolean whether shift is in "park" mode
Inputs_shiftR	Boolean whether shift is in "reverse" mode
Inputs_statusAC	Boolean whether air conditioning is on or off
Inputs_steeringAngle_deg	Steering wheel angular position in degrees
Inputs_throttlePedal_pct	Amount of throttle pedal input in percentage of max input
Power_accumOpTime_s	Accumulative operation time in seconds
road_lanes	Number of lanes on the road
road_max_speed	Maximum speed at latitude/longitude
Road_type	Road type at latitude/longitude as classified by Open Street Map (OSM)