Driving Behavior Improvement and Driver Recognition Based on Real-Time Driving Information

Kexin Nie,[1] Luyan Wu,[1] and Jiafan Yu[1]
[1] *Stanford University, Stanford, CA 94305, USA*

(Dated: December 13, 2013)

Based on real-time driving information recording data, we raised a regression model for predicting gasoline consumption rate from speed, acceleration and heading degree information. We also developed a grouping method to sort high-frequency fragmented driving information to sets of trips, and then implemented supervised learning to illustrate the relationship between average speed and gasoline usage. The general MPG-MPH relationship gives useful suggestions of good driving speed. The ranking of different vehicles' MPG from real-time driving information are more close to practical data, compared with the data given by automakers, which is a useful criteria for new vehicle consumers. We also implemented both supervised learning (Naïve Bayes and Support Vector Machine) and unsuperivised learning (k-means clustering) to try to classify potential multi-driver vehicles. Though the ultimate goal is still not achievable, we can still classify driving behaviors to 2 patterns with different speeds, which is a hint for distinguishing different driving conditions, and driver's driving preference, which could be a good foundation for further work.

INTRODUCTION

A driver's gasoline usage is very sensitive to his/her driving habits. Since Americans spend about $500 Billion dollars on gasoline every year, a good driving habits which could spend less gasoline and less money could result in $100 Billion savings per year. In the meantime, a driver's driving behavior could be defined as a set of all real-time tractable driving information, including speed, gas usage, acceleration and heading degrees and it is a reasonable assumption that different drivers have different driving behavior patterns. Nowadays, we are able to record the high-frequency real-time driving information. This information is potential for us to find out more information, including driver's driving habit patterns and the effect to gasoline consumptions. This patterns in driver's driving behavior and gasoline consumption efficiency could give us better guidance to drive more efficiently. In the same time, the abundant information also gives us the potential to distinguish different drivers only by analyzing the real-time data by illustrating different drivers' driving behavior patterns. In this paper, we managed to find out the relation between gasoline usage and driving informations in order to give suggestions to better driving habits, and we made the initial trial to distinguish different drivers in the same vehicle.

DATA GROUPING AND INITIAL ANALYSIS

The original 500 MB data includes real-time driving information of 18 different vehicles. Every data sample only records transient speed, heading degrees, acceleration and gas usage. In the first stage, we sorted the fragmented data samples to long enough trips, which contain more systematically information and are beneficial for further analysis. We treat 2 original data points in the same trip if the time interval between 2 original data points ls less than a constant time. In practical, considering traffic lights, traffic congestion and passenger picking up, 2 minutes is a good threshold in our application. The separation of different trips is shown in FIG. 1.

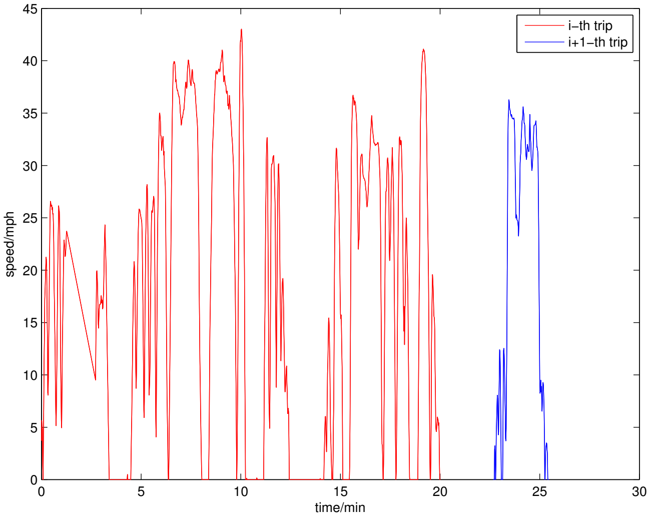
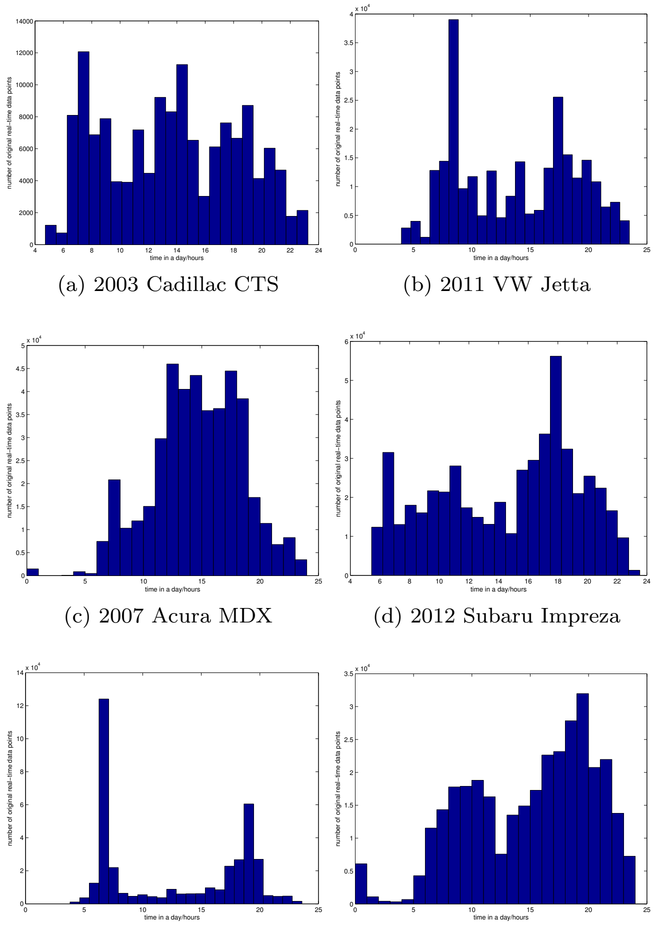


FIG. 1: Illustration of the saparation of different trips

For every trip we calculated calculated the average speed of the trip, total mileage, trip time and gas usage. In the meantime, we also calculated the speed variance and L-2 norm of the acceleration during the trip, which is sensitive to gas usage and driver's information patterns. The trip information is better organized, compared with the original data, for our analysis. At the same time, this grouping process could efficiently decrease the size of data samples so we could train the trip info in personal computer, which is crucial for the first stage implementation of the project. For example, There are 108692 data samples for a 2005 Volkswagen GTI 2-door hatchback, but after dividing them to different trips, there are only 1944 trips in total.



(a) 2003 Cadillac CTS

(b) 2011 VW Jetta

(c) 2007 Acura MDX

(d) 2012 Subaru Impreza

(e) 2011 MB GL 450

(f) 2003 Toyota Corolla

FIG. 2: Histograms of driving time of 6 vehicles

In addition, fundamental statistics of original data could be helpful for further analysis. It is a reasonable assumption that the number of original data points in some time range in a day is proportional to the total driving time in the time range in a day. We plotted the histogram of all original data's recording time in a day(FIG. 2), which clearly shows the difference of driving habit. For some vehicles, there are two main peaks in the morning and afternoon denoting the daily commuting. There are also some relatively even distribution in all day for some vehicles, which indicates a higher probability of commercial use in work time. This pattern is the basis for driver recognition in next section.

TABLE I: Model with response variable $\log(gas\_mpg + 1)$

Coefficients	Estimate
(intercept)	2.28
$speed\_mph$	3.35×10^2
$speed\_mph^2$	-2.72×10^2
$speed\_mph^3$	1.31×10^2
a_f	-3.49
a_f^2	1.47×10^1
a_f^3	6.25
a_u^3	4.34
$speed\_mph \times a_f$	-6.24×10^{-3}
$speed\_mph \times a_r$	1.28×10^{-4}

CONSIDERING THE EFFECT OF ACCELERATION IN REGRESSION MODEL TO PREDICT GAS USAGE EFFICIENCY

In the original datasets, we have variables a_x_gs, a_y_gs, and a_z_gs, which are accelerations at three orthogonal directions. However, since the acceleration measurement systems are different in different cars, it is hard to interpret these variables directly. Luckily, we have variables of $record\_date\_time$, $heading\_degrees$ and $speed\_mph$, from which we can obtain new variables a_f (forward acceleration, $a_f > 0$ stands for accelerating forwards) and a_r (centrifugal acceleration) according to mechanics. We can also obtain variable a_u (vertical acceleration, $a_u > 0$ stands for accelerating upwards) through deducting the gravity acceleration. Then we have gas_mpg as response variable and other variables as predictors to find an appropriate regression model for predicting gasoline usage efficiency.

The linear regression model was first implemented, due to the low adjusted R-squared criteria, linear model is not a good fit for predicting gas usage efficiency. When introducing polynomial fitting and interaction, the R-squared criteria got improved. We further introduced stepwise variable selection of polynomial fitting and got the lowest cross validation mean squared errors. Final improvement is done by performing a box-cox transformation to the response variable. The result shows that if we take $\log(gas\_mpg+1)$ instead of the gas_mpg as the response variable, we get better regression model, which is shown in table I.

The final model after variable selection is given as

$$\begin{aligned}&\log(gas\_mpg + 1) = 2.28 + 335 speed\_mph\\&-272 speed\_mph^2 + 131 speed\_mph^3 - 3.49 a_f + 14.7 a_f^2\\&+6.25 a_f^3 + 4.34 a_u^3 + 0.00624 speed\_mph\ a_f\\&+0.000128 speed\_mph\ a_r.\end{aligned}$$

We could use this model to predict MPG usage according to observed speed and acceleration information.

SPLINE FITTING OF GAS USAGE AND TRIP AVERAGE SPEED

We investigated the relationships between all trip informations and trip average gasoline usage. The relationship between trip average speed and trip average gasoline consumption rate is one of the most obvious pair, so we implemented several supervised learning algorithms to figure out the relationship, including polynomial fitting, exponential fitting and spline fitting. We chose spline fitting as our best estimation. Though for higher order polynomial provides smaller bias, the variance is enormous large. The same spline fitting parameters are used for all vehicles' trip information. The plot of spline fitting between trip average speed and trip average gas consumption rate is shown in FIG. 3.

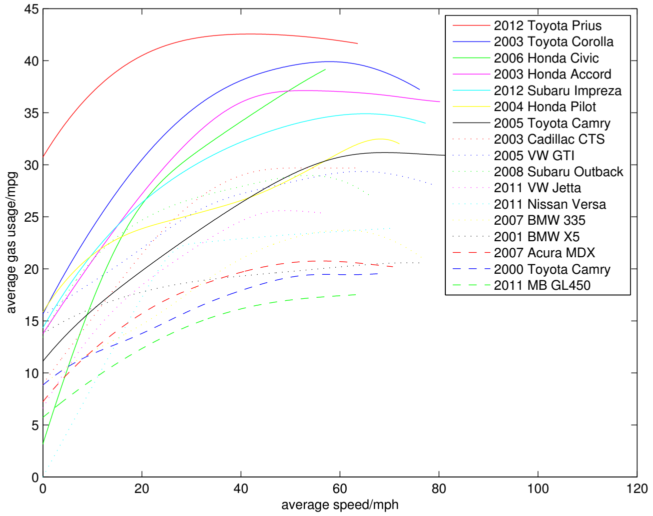


FIG. 3: Trip average mph-trip average MPG spline fitting

Regarding of different vehicles' curves, there are obvious difference in gas consumption rate between different vehicles. The gas consumption rate could be as high as 42 MPG for 2012 Toyota Prius, and could also be as low as 15 MPG for 2011 Mercedes-Benz GL 450, a difference of a factor of 3.

The championship of gasoline consumption efficiency competition is the only hybrid car in the list. The Toyota Prius could defeat all other vehicles in all speed range. The hybrid energy recycle system improves the gas consumption especially in lower speed range. The low speed MPG of Prius is higher than 30, at the same time, all other cars' low speed MPG is below 15. The curve is still lower than official labeled MPG, 51 city/48 highway, indicating that lower speed is still affecting the gas consumption performance even for hybrid cars.

The championship in regular vehicle division is also Japanese automaker, 2003 Toyota corolla. Actually, Japanese automakers dominates the energy efficiency competition. The 7 most energy efficient vehicles are Japanese brands, including a full-size SUV, Honda Pilot. Although we did not consider the effect of the difference of driver's driving behaviors, the result is sufficient to indicate that Japanese vehicles are more energy efficient than others.

The size of the vehicles is also one of the key factor affecting trip average MPG. most of the SUV models suffer. The contrast is more severe in high-speed range. Avoid full size SUV is a good way to save gasoline.

It is interesting to compare the MPG-MPH curve of 2 Toyota Camry vehicles. The 2000 Toyota Camry holds one of the lowest MPG. In comparison, the MPG of 2005 Toyota Camry is much higher than which of the 2000 one. It is reasonable to contribute this to the degradation of performance. Still, we ignore the difference of driver's habits.

Regarding of different speed regions, it is universal that the MPG is very low in very low speed region, saying, below 20 MPH. The gasoline consumption drops when the trip average speed increases, then the average MPG reaches a plateau, that MPG is relatively insensitive to average speed. The MPG value usually reaches its maximum from 40 MPH to 65 MPH, varying for different vehicles. Then MPG drops when the average speed goes to extremely high.

From the real-time data, low trip average speed indicates very frequently stop and start, which is one of the most inefficient driving mode. On the other hand, extremely high speed brings extremely high wind resistance and tire resistance, which increases the MPG.

In our conclusion, in order to save gasoline consumption, you should replace super old car and buy or trade in a hybrid. If you want to keep your current car, try to pick up a good time to keep your average speed around 45 MPH and don't drive crazily fast even in a vacant highway, to get maximum energy efficiency.

DRIVING PATTERN/DRIVER RECOGNITION THROUGH SUPERVISED LEARNING

In this section we aim to detect if more than one driver has drove a certain car. Here we syntheses hypothesis testing method and classifiers to solve this problem. Based on the way data was collected, we found that for some car, such as the 2006 Honda Civic, the 2011 Nissan Versa and 2000 Toyota Camry are only driven in some certain time in a day, from the FIG. 2 shown in section 1. For this situation, we need only to test if the collections of data, which are recorded in different time periods, revealing different driving behaviors. If so, we may make the conclusion that this car was drove by several drivers. To test if different drivers have driven the car in different time periods, we can firstly pick two time periods and assume two drivers was driving the car. We test the following statement:

- H_0: The two time periods were droven by two drivers.

verses

- H_1: The two time periods were droven by one same driver.

To make the decision whether we should reject the null hypothesis H_0, we use classify methods, i.e. Naïve Bayes and Support Vector Machine to verify the test. We gain our result based on the following procedures.

1. Randomly choose 80% of the original data from the two time periods, use them as the training sample, and leaving the residual 20% to be the test dataset.

2. After labeling training data from the first time period as class one, and the other time period as class two, we implement Naïve Bayes and SVM on the test data to see which class they belong to. We can analysis on the outcomes to see whether the original two time periods are actually in the same class. For example, if the proportion of wrongly predicted samples is high in both time periods, we can draw the conclusion that there is actually no large difference in the two time periods, thus they should be in the same class.

3. If the two periods are in one class, then the classifier model (Naïve Bayes or SVM), which is constructed based on the training set will not be significant to class the test set. The predicted results will only based on the sample sizes of training data from two time periods. Let

$$N_1 = \#\{training\ sample\ of\ time\_period\_one\}$$
$$N_2 = \#\{training\ sample\ of\ time\_period\_two\}$$

   Since the two periods are actually in one class, then the testing data from *time_period_one* has the probability $p_1 = N_1/(N_1 + N_2)$ to be assigned to class one and has the probability $p_2 = N_2/(N_1 + N_2)$ to be assigned to class two. By comparing the predicting results with p_1 and p_2, we can draw the conclusion whether the two periods belong to the same class.

We perform this procedure on the 2006 Honda Civic. This dataset totally contains observations of 5 days, and has 10 time periods. To simplify the results, we only select time periods 3:00-5:00 in day2(data1), 21:00-24:00 in day3(data2), 23:00-23:50 in day4(data3) and 1:04-2:00 in day5(data4), 4 datasets in total to detect if any two time periods belong to different classes. By running Naïve Bayes and SVM, the results come as table II and table III.

TABLE II: Predict Error of Naïve Bayes

Predict Error^2(NB)	Data1	Data2	Data3	Data4
Data1	- - -	0.2920	0.0942	0.1804
Data2	0.5480	- - -	0.1184	0.1569
Data3	0.2365	0.3078	- - -	0.1157
Data4	0.2786	0.2706	0.0745	- - -

TABLE III: Predict Error of SVM

Predict Error^2(SVM)	Data1	Data2	Data3	Data4
Data1	- - -	0.3380	0.0461	0.0240
Data2	0.4760	- - -	0.0310	0.0667
Data3	0.2766	0.3661	- - -	0.0549
Data4	0.3166	0.3078	0.2196	- - -

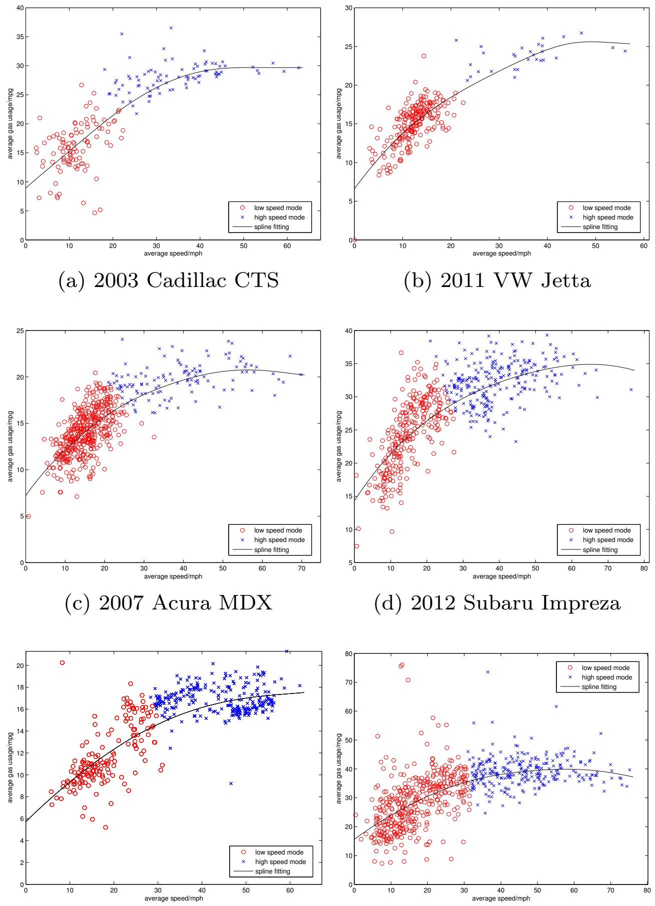
Since we have make the number of sample sizes to be same when testing two time period datasets, we only need to compare the Predict Error with 0.5. If the error is nearly 0.5, it means that no difference is show between the two class, thus they should belong to one class, and vice versa. Based on the results predicted by Naïve Bayes and Support Vector Machine, we might conclude that data1, data2 might be in one same class, due to the predict error is around 0.5, while data3 and data4 should not be in the same class, because the predicted errors are far less than 0.5. We may draw the conclusion that in the 2006 Honda Civic, drivers were changed after 4th day, and should be more than 2 drivers drove that car. We can adopt the same method to other cars and collect other driving behaviors on different time periods.

DRIVING PATTERN/DRIVER RECOGNITION THROUGH UNSUPERVISED LEARNING

Since the relationship between trip average MPG and trip average MPH is one of the most key characters of driving behavior, we also implemented k-means clustering algorithm to explore potential patterns/different drivers for one vehicle. FIG. 4 illustrates the k-means clustering results of 6 different vehicles.

We could clearly distinguish 2 different regions in MPH-MPG scattering plots, One region is mainly in low speed range and another is in high speed range, which is consistent with spline fitting analysis in previous section. For some vehicles, two different clusters are very close, which indicates the possibility of multiple drivers in the vehicle is low. But for some vehicles, two clusters are well separated (FIG. 4b and FIG. 4e), which is a sign of two drivers, or two very different gas consumption modes of the vehicle.

Besides of the main aim of driver/driving mode recognition, it is also interesting to observe that different vehicles/drivers have different cluster patterns, which shows



(a) 2003 Cadillac CTS (b) 2011 VW Jetta

(c) 2007 Acura MDX (d) 2012 Subaru Impreza

(e) 2011 MB GL 450 (f) 2003 Toyota Corolla

FIG. 4: K-means clustering and spline fitting of 6 vehicles

different daily driving habits or the driver(s). For example, the number of points in high speed region of the 2011 Mercedes-Benz GL 450 and the 2003 Toyota Corolla is almost the same as those in low speed region, respectively(FIG. 4e and FIG. 4f). It is reasonable to assume that points in high speed region represent highway driving, and points in low speed region represent local area driving. These two vehicles are equally driven in highway and local. In contrast, the 2011 VW Jetta are mostly driven in local area, which indicates the possibility that the distance between the driver's workplace and home is close, or the main use of the vehicle is not long distance commute. The analysis is beneficial for driver recognition and for future commercial opportunities, for example, precision advertising.

CONCLUSION AND FUTURE IMPROVEMENT

We sorted fragmental high-frequency real-time driving information to well-organized trip information. This information provided the first-hand practical relationship between trip average speed and trip average gasoline consumption rate. The spline regression provides highly visualizable MPG-MPH curves of different vehicles, which gave a very intuitive guidance to find an optimal driving speed and to choose a more energy efficient vehicle.

We also implemented both supervised learning method and unsupervised learning method to try to distinguish different driving patterns for different vehicles and (possible) different drivers in the same vehicle. Supervised learning requires some assertions but draw a powerful conclusion that 2006 Honda Civic may have 2 drivers. Unsupervised learning algorithms do not need any assumptions and could also provide useful hints.

There are still some open problems unsolved. Firstly, we still do not make full utilization of the data set. The acceleration and change of heading degree do not help us much in current stage. How to integrate the gas consumption with these parameters is an interesting topic. Furthermore, The lack of traffic information is a great loss. If we could know the traffic information and build a Markov chain, we could use dynamical programing to figure out the best action (throttle paddle, gear shift, etc) to get best gasoline consumption efficiency.

ACKNOWLEDGMENTS

We thank MetroMile, Inc. for providing real time driving information for our analysis.