Methods and Initial Results

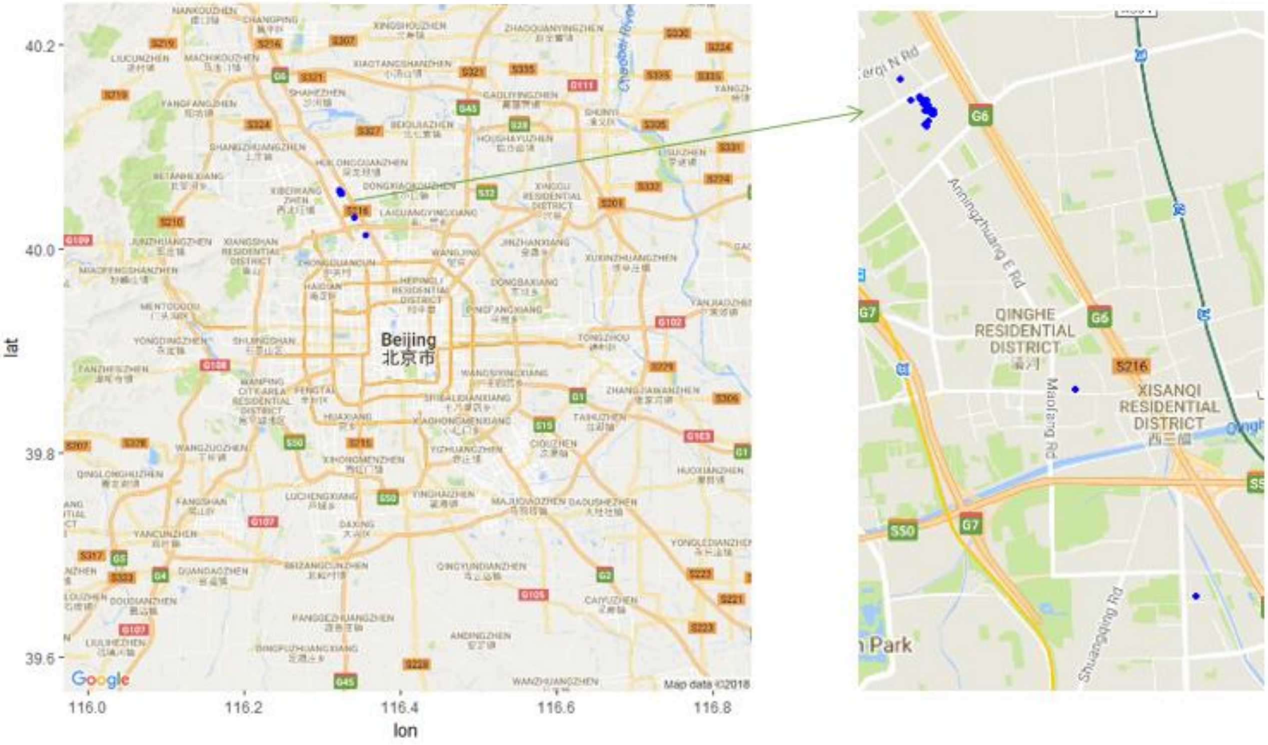
Collecting more information is one of the critical ways to improve predictive accuracy. In accident analysis, we can use computational methods to obtain three critical sets of always-on information from GPS data: individual-level human mobility pattern, moving behavior and latent socio-demographic characteristics.

In order to look at how GPS data can improve accident prediction that traditionally relies on self-report sociodemographic and vehicle-related characteristics, I use two datasets related to cars in Beijing to explore this question. The first dataset from a telematics company contains telematics information. Specifically, there are over 10,000 cars whose moving paths and patterns were recorded by in-vehicle telematics. In the dataset, each car has averagely over 10,000 GPS points from January 1st to March 15th, 2016. The second dataset is from a car insurance company. It contains various traditional self-report information for accident prediction: claim history, drivers' sociodemographic characteristics and vehicle-related characteristics. The claim history (claim loss) is an indicator of accident and hence can be used as the response of whether a car has accidents and how serious the accidents are, and the sociodemographic and vehicle-related characteristics are often used as the predictors in traditional insurance and accident analysis. These two datasets can be merged by vehicle identification number. Therefore, it is possible to use new features extracted from the GPS data to enrich the information for prediction.

Different from many other datasets that directly contain useful variables for model building, the core task and methods of this project are to extract a set of features from the GPS data based on the physical, economic and political layout of Beijing as well as the time.

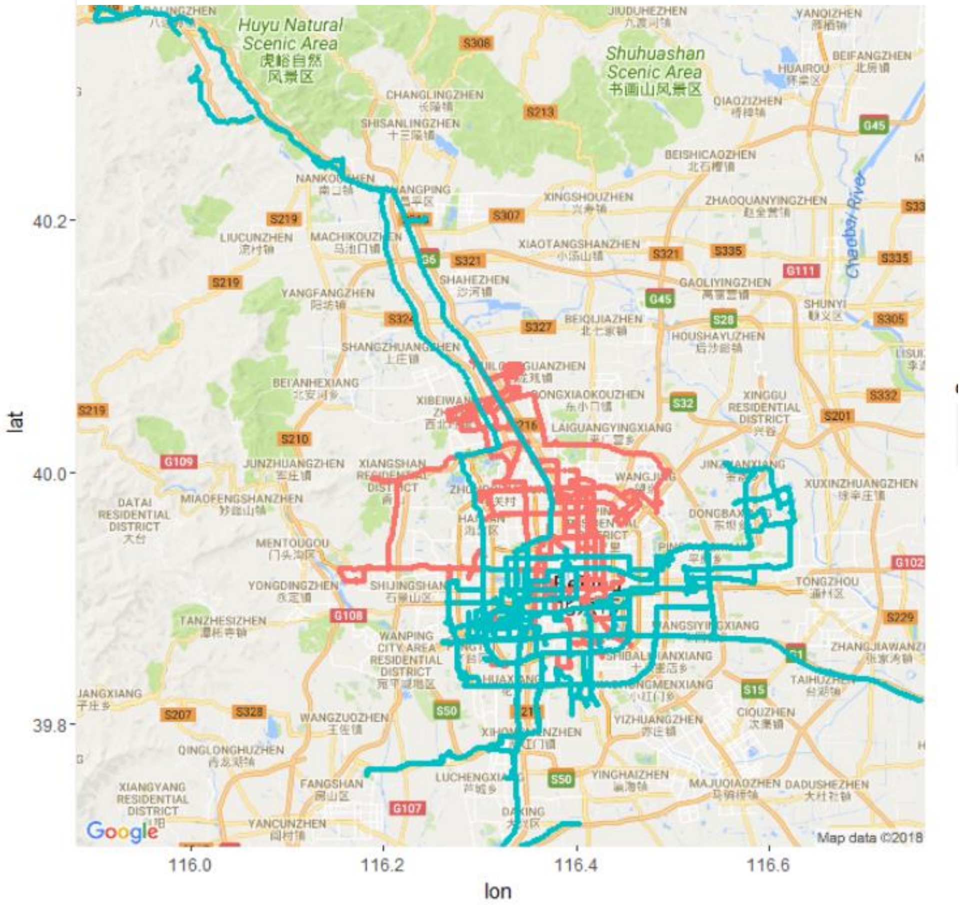
However, these features should not be reconstructed as explicit socioeconomic variables in fear of information loss and distortion. Therefore, the typical method is to use semi-socioeconomic features.

Figure 1: Dealing with stationary points



For example, Figure 1 (mapped by the ggmap package) shows that how I identify where a driver lives. I filter each driver’s first points of each day after 5 am, remove the outliers and then calculate the mean latitude and longitude. The mean location can provide the information about where they live. For the driver in the figure, his home is at suburb and I can speculate, according to the economic and cultural layout in Beijing, that he is not rich and also does not have a high social status. The mean location here embodies numerous latent socio-economic information and hence is a semi-sociodemographic feature. It is worth noting that we need not to expect that we should exactly know the socioeconomic meaning in this kind of mean location. Although it is possible to use a set of this kind of features to predict explicit socioeconomic characteristics, such as economic status, we should avoid to do this. This is because information loss is inevitable if we use the predicted explicit socioeconomic characteristics to predict accident. In other words, the second-step prediction cannot improve prediction. Therefore, the best way is to divide the layout of Beijing into several areas, find out in which areas the drivers live, and finally put this kind of information into predictive model. For instance, Beijing is politically, socially and economically divided by several ring roads. The area within the second ring road is highly political because there are a lot of government offices and dwellings for officers. Living there not only signifies the social status of most drivers, but also means the traffic administration in this area is stricter. In this case, the feature of mean location categorized by areas might improve the accident prediction.

Figure 2: Overall moving patterns of two cars over three months

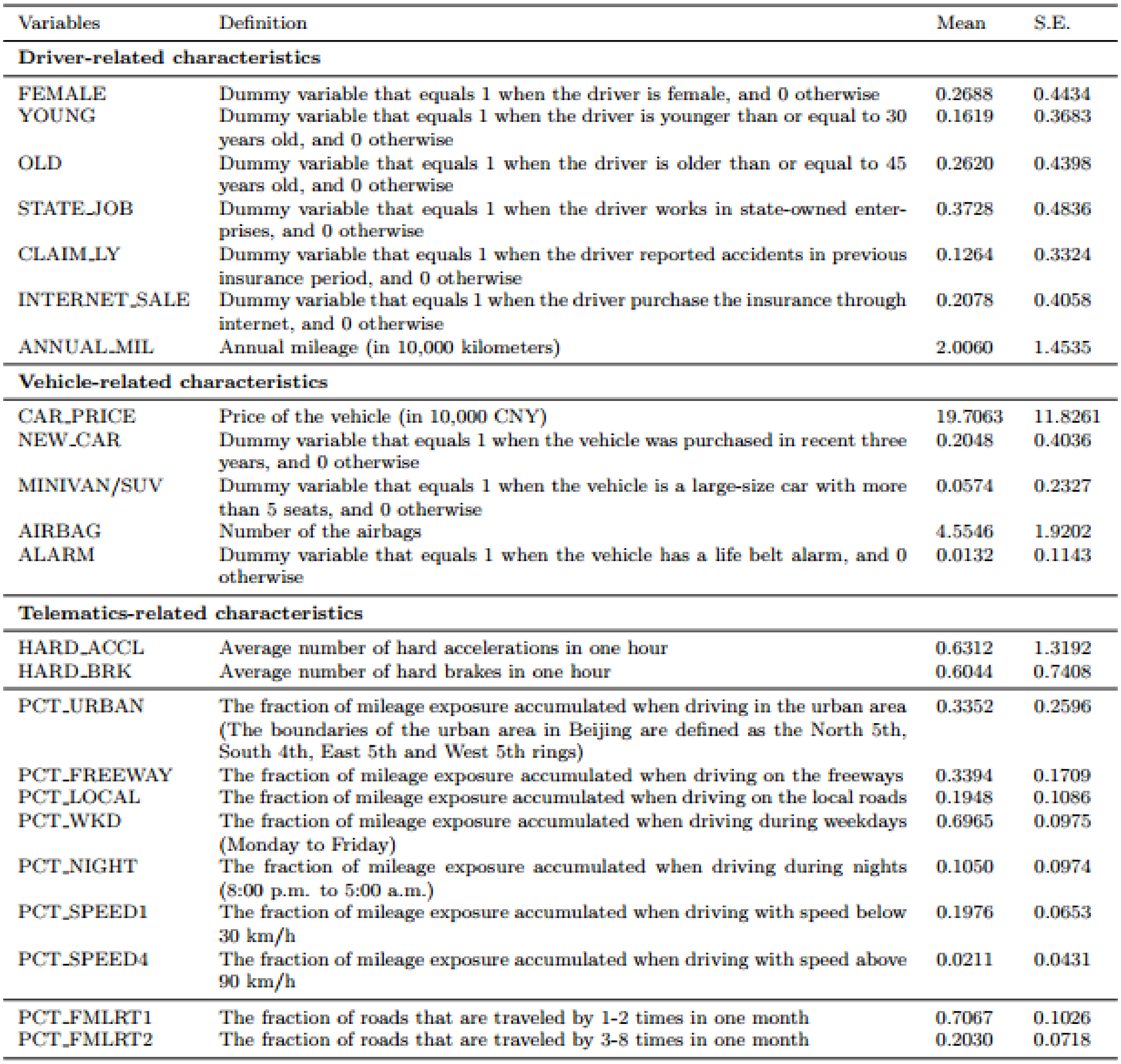


In addition to the stationary points, it is also possible and more important to extract moving patterns and paths from the GPS data. Figure 2 shows the overall moving patterns and paths of two cars. It is obvious that these two cars have different moving patterns. The overall moving patterns and paths can further break down based on time, city layout and more direct driving behavior, such as nighttime driving, urban driving and hard-acceleration behavior, which can be obtained or calculated from the GPS data. These features contain both latent sociodemographic information and direct driving behavior. As the same as the stationary point, we need not to clarify how they embody sociodemographic characteristics.

Table 1 shows three sets of variables. The first two sets are from the insurance datasets and the last one is from the GPS dataset. **It is worth noting that I don't list all variables I extract from the GPS data (including the variables I mentioned above) because I still need more time to handle them. In this assignment, I just use part of variables, but the other will be used in the final paper.** The first set is the driver-related characteristics that contains self-report demographic characteristics, insurance purchase channel, previous claim and self-report annual mileage. The second set is the vehicle-related characteristics that contains car price, type and equipment. The final set is the characteristics extracted from the GPS data based on time and city layout as well as the driving behavior information recorded by in-vehicle telematics device. The response I choose here is whether a drive has self-report claim or not.

For binary classification problems, many algorithms, such as tree-based method, SVM and neural networks, outperform binary logistic regression. However, since binary logistic regression can provide interpretive results, it is helpful to explore its results before building more predictively powerful models. In this assignment I mainly present the results from the logistic regression with latent class model. There are two reasons for this. First, logistic regression can help select predictors when I build other models. Second, some predictors for other models have not been full prepared yet and the models themselves have to be further adjusted. Rather than presenting defective models, I choose to present solid results. But other models (neural networks and tree-based models for classification and lasso for regression) will be presented in the poster and the final paper.

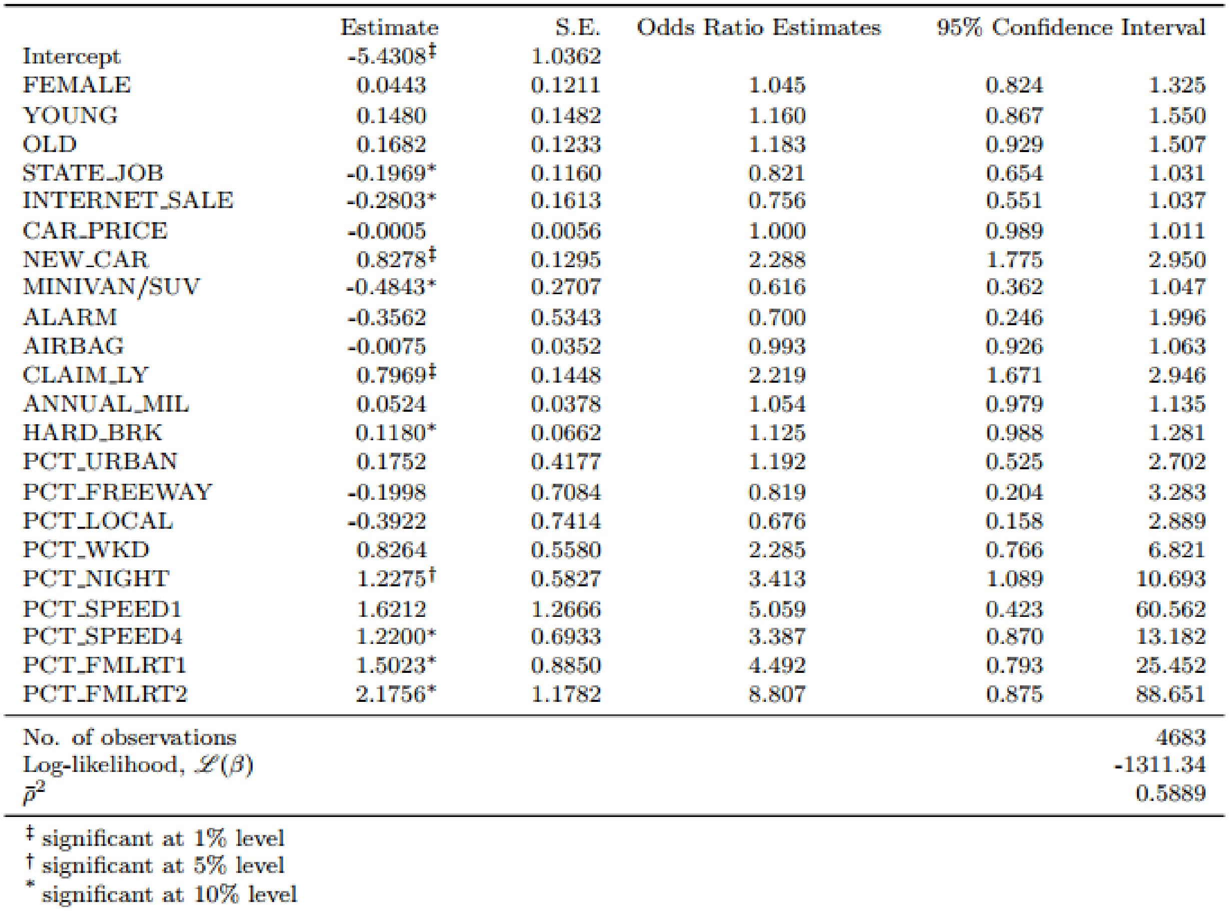
Table 1: Summary statistics of independent variables



Variables	Definition	Mean	S.E.
Driver-related characteristics			
FEMALE	Dummy variable that equals 1 when the driver is female, and 0 otherwise	0.2688	0.4434
YOUNG	Dummy variable that equals 1 when the driver is younger than or equal to 30 years old, and 0 otherwise	0.1619	0.3683
OLD	Dummy variable that equals 1 when the driver is older than or equal to 45 years old, and 0 otherwise	0.2620	0.4398
STATE_JOB	Dummy variable that equals 1 when the driver works in state-owned enterprises, and 0 otherwise	0.3728	0.4836
CLAIM_LY	Dummy variable that equals 1 when the driver reported accidents in previous insurance period, and 0 otherwise	0.1264	0.3324
INTERNET_SALE	Dummy variable that equals 1 when the driver purchase the insurance through internet, and 0 otherwise	0.2078	0.4058
ANNUAL_MIL	Annual mileage (in 10,000 kilometers)	2.0060	1.4535
Vehicle-related characteristics			
CAR_PRICE	Price of the vehicle (in 10,000 CNY)	19.7063	11.8261
NEW_CAR	Dummy variable that equals 1 when the vehicle was purchased in recent three years, and 0 otherwise	0.2048	0.4036
MINIVAN/SUV	Dummy variable that equals 1 when the vehicle is a large-size car with more than 5 seats, and 0 otherwise	0.0574	0.2327
AIRBAG	Number of the airbags	4.5546	1.9202
ALARM	Dummy variable that equals 1 when the vehicle has a life belt alarm, and 0 otherwise	0.0132	0.1143
Telematics-related characteristics			
HARD_ACCL	Average number of hard accelerations in one hour	0.6312	1.3192
HARD_BRK	Average number of hard brakes in one hour	0.6044	0.7408
PCT_URBAN	The fraction of mileage exposure accumulated when driving in the urban area (The boundaries of the urban area in Beijing are defined as the North 5th, South 4th, East 5th and West 5th rings)	0.3352	0.2596
PCT_FREEWAY	The fraction of mileage exposure accumulated when driving on the freeways	0.3394	0.1709
PCT_LOCAL	The fraction of mileage exposure accumulated when driving on the local roads	0.1948	0.1086
PCT_WKD	The fraction of mileage exposure accumulated when driving during weekdays (Monday to Friday)	0.6965	0.0975
PCT_NIGHT	The fraction of mileage exposure accumulated when driving during nights (8:00 p.m. to 5:00 a.m.)	0.1050	0.0974
PCT_SPEED1	The fraction of mileage exposure accumulated when driving with speed below 30 km/h	0.1976	0.0653
PCT_SPEED4	The fraction of mileage exposure accumulated when driving with speed above 90 km/h	0.0211	0.0431
PCT_FMLRT1	The fraction of roads that are traveled by 1-2 times in one month	0.7067	0.1026
PCT_FMLRT2	The fraction of roads that are traveled by 3-8 times in one month	0.2030	0.0718

Before build logistic regression model, I remove the cars which traveled less than 14 days per month. There is no doubt that less travel can also be a useful predictor of accident risk in many models. However, less travel will result in the difficulty of extracting information from GPS data and hence might reduce their utility.

Table 2: Logistic regression

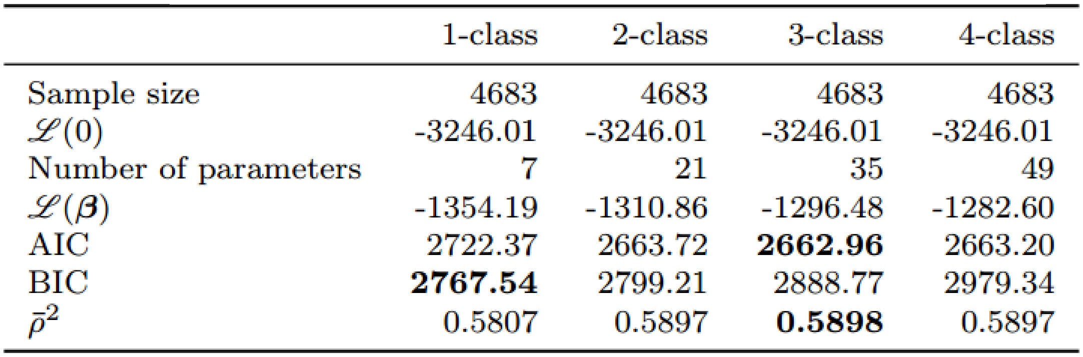


	Estimate	S.E.	Odds Ratio Estimates	95% Confidence Interval	
Intercept	-5.4308‡	1.0362			
FEMALE	0.0443	0.1211	1.045	0.824	1.325
YOUNG	0.1480	0.1482	1.160	0.867	1.550
OLD	0.1682	0.1233	1.183	0.929	1.507
STATE_JOB	-0.1969*	0.1160	0.821	0.654	1.031
INTERNET_SALE	-0.2803*	0.1613	0.756	0.551	1.037
CAR_PRICE	-0.0005	0.0056	1.000	0.989	1.011
NEW_CAR	0.8278‡	0.1295	2.288	1.775	2.950
MINIVAN/SUV	-0.4843*	0.2707	0.616	0.362	1.047
ALARM	-0.3562	0.5343	0.700	0.246	1.996
AIRBAG	-0.0075	0.0352	0.993	0.926	1.063
CLAIM_LY	0.7969‡	0.1448	2.219	1.671	2.946
ANNUAL_MIL	0.0524	0.0378	1.054	0.979	1.135
HARD_BRK	0.1180*	0.0662	1.125	0.988	1.281
PCT_URBAN	0.1752	0.4177	1.192	0.525	2.702
PCT_FREEWAY	-0.1998	0.7084	0.819	0.204	3.283
PCT_LOCAL	-0.3922	0.7414	0.676	0.158	2.889
PCT_WKD	0.8264	0.5580	2.285	0.766	6.821
PCT_NIGHT	1.2275†	0.5827	3.413	1.089	10.693
PCT_SPEED1	1.6212	1.2666	5.059	0.423	60.562
PCT_SPEED4	1.2200*	0.6933	3.387	0.870	13.182
PCT_FMLRT1	1.5023*	0.8850	4.492	0.793	25.452
PCT_FMLRT2	2.1756*	1.1782	8.807	0.875	88.651
No. of observations					4683
Log-likelihood, $\mathscr{L}(\beta)$					-1311.34
$\bar{\rho}^2$					0.5889

‡ significant at 1% level
† significant at 5% level
* significant at 10% level

Table 2 shows the logistic regression result. Since the logistic model is still a prediction model, I choose significant level at 10% rather than at 5%. It is obviously that 5 predictors from the GPS data are significant and hence have the potential to improve the risk prediction. It is also worth noting that, for prediction model, insignificant predictors do not mean that they cannot improve the prediction, and significant predictors are not necessary to improve it. Therefore, I also test the overall model fit of the model. I build a logistic regression model without telematics information and compare it with the model with telematics information by testing the likelihood ratio according to Chi-squared distribution. The statistic is significant at 5% level. The result means the information extracted from the GPS data are very likely to improve the accident prediction. Besides, although I can also calculate the mean square error and AUC-ROC through cross-validation, I don’t do it because my intention of using logistic regression is to find hints for further predictive model building and logistic regression is overall not a good predictive model.

Table 3: Model fit statistics for latent class analysis

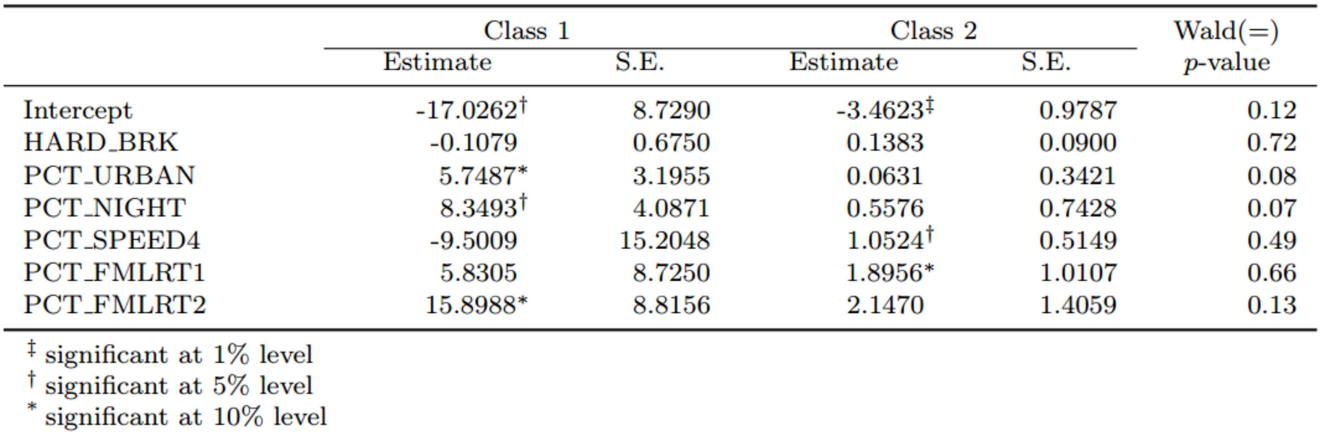


	1-class	2-class	3-class	4-class
Sample size	4683	4683	4683	4683
$\mathscr{L}(0)$	-3246.01	-3246.01	-3246.01	-3246.01
Number of parameters	7	21	35	49
$\mathscr{L}(\boldsymbol{\beta})$	-1354.19	-1310.86	-1296.48	-1282.60
AIC	2722.37	2663.72	**2662.96**	2663.20
BIC	**2767.54**	2799.21	2888.77	2979.34
$\bar{\rho}^2$	0.5807	0.5897	**0.5898**	0.5897

In order to capture the heterogeneity among drivers, I also use latent class model to divide the drivers into different groups. Table 3 shows the model fit statistics for latent class analysis. Although the 1-class and 3-class models outperform the 2-class model in terms of one or two statistics, the 2-class model is still more suitable for further analysis since its all statistics are sub-optimal.

Based on the result in latent class analysis I also build a class-specific binary logistic model with the information from the GPS data. Table 4 shows that different variables play different roles in different classes. This result signifies that predetermine different classes might help us further build a more efficient predictive model.

Table 4: Class-specific binary logistic model



	Class 1		Class 2		Wald(=)
	Estimate	S.E.	Estimate	S.E.	*p*-value
Intercept	-17.0262†	8.7290	-3.4623‡	0.9787	0.12
HARD_BRK	-0.1079	0.6750	0.1383	0.0900	0.72
PCT_URBAN	5.7487*	3.1955	0.0631	0.3421	0.08
PCT_NIGHT	8.3493†	4.0871	0.5576	0.7428	0.07
PCT_SPEED4	-9.5009	15.2048	1.0524†	0.5149	0.49
PCT_FMLRT1	5.8305	8.7250	1.8956*	1.0107	0.66
PCT_FMLRT2	15.8988*	8.8156	2.1470	1.4059	0.13

‡ significant at 1% level
† significant at 5% level
* significant at 10% level