

Using GPS Data to Predict Accident Risk: Evidence from Beijing

Kanyao Han

University of Chicago

Abstract

The popularization of GPS technology over the last decade have ushered a new era for accident analysis since it can provide two critical sets of always-on information pertinent to accident analysis and prediction: individual-level moving behavior and sociodemographic characteristics. In this paper, I them from a car GPS dataset from a telematics company in Beijing, merge them with an insurance dataset from an insurance company, and test their role in accident analysis. I find the features extracted from the car GPS dataset can not only help us better analyze the probability of car accident, but improve the prediction of accident loss. Since previous studies in accident analysis seldom use sociodemographic features extracted from GPS data, the extraction method and the underlying theory in this project has a methodological and theoretical implication.

Objectives

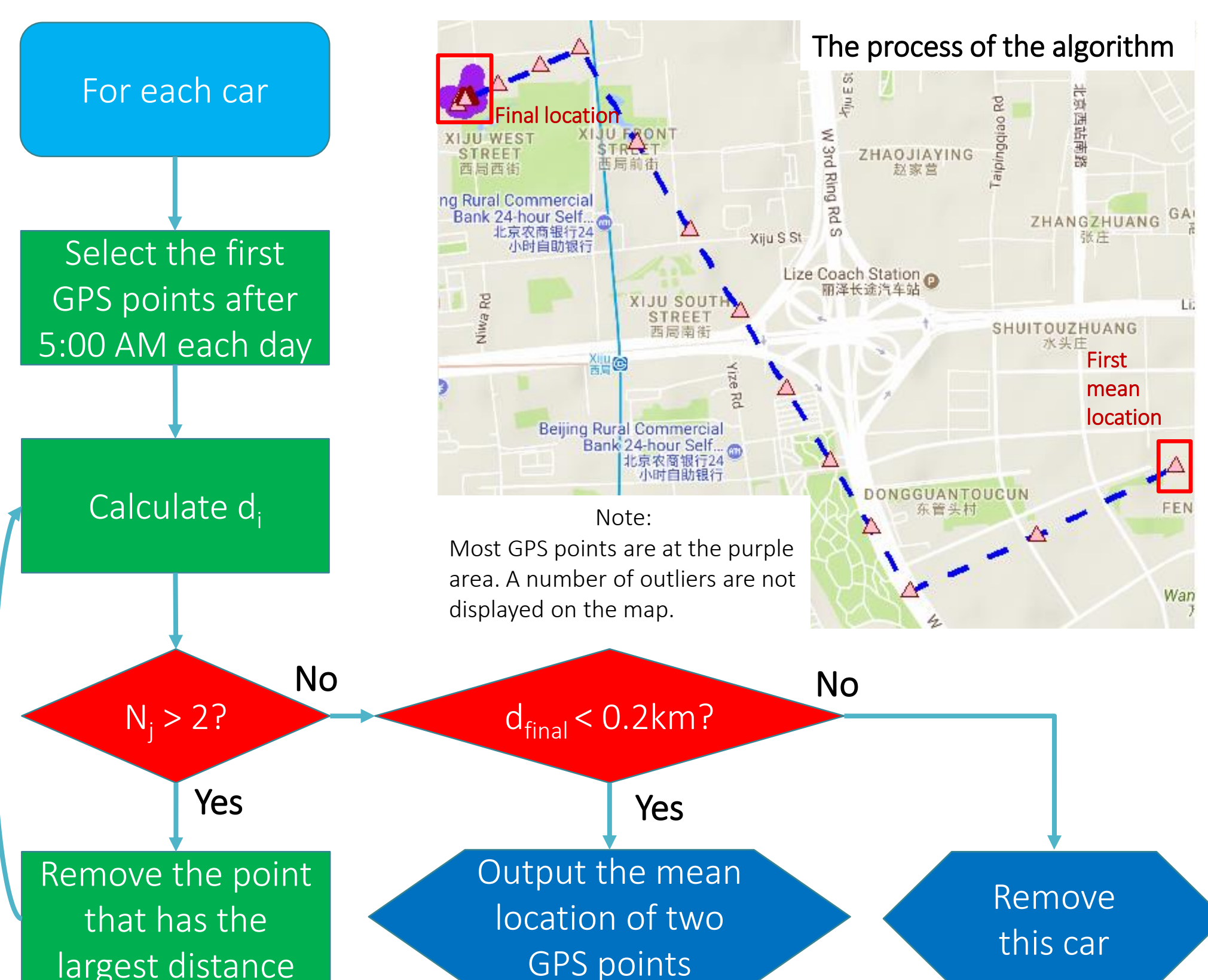
- Design a method to extract home address information from the GPS dataset.
- Further mine useful information from home address for prediction
- Extract driving behavior information from the GPS dataset
- Build predictive models for accident analysis and prediction

Data

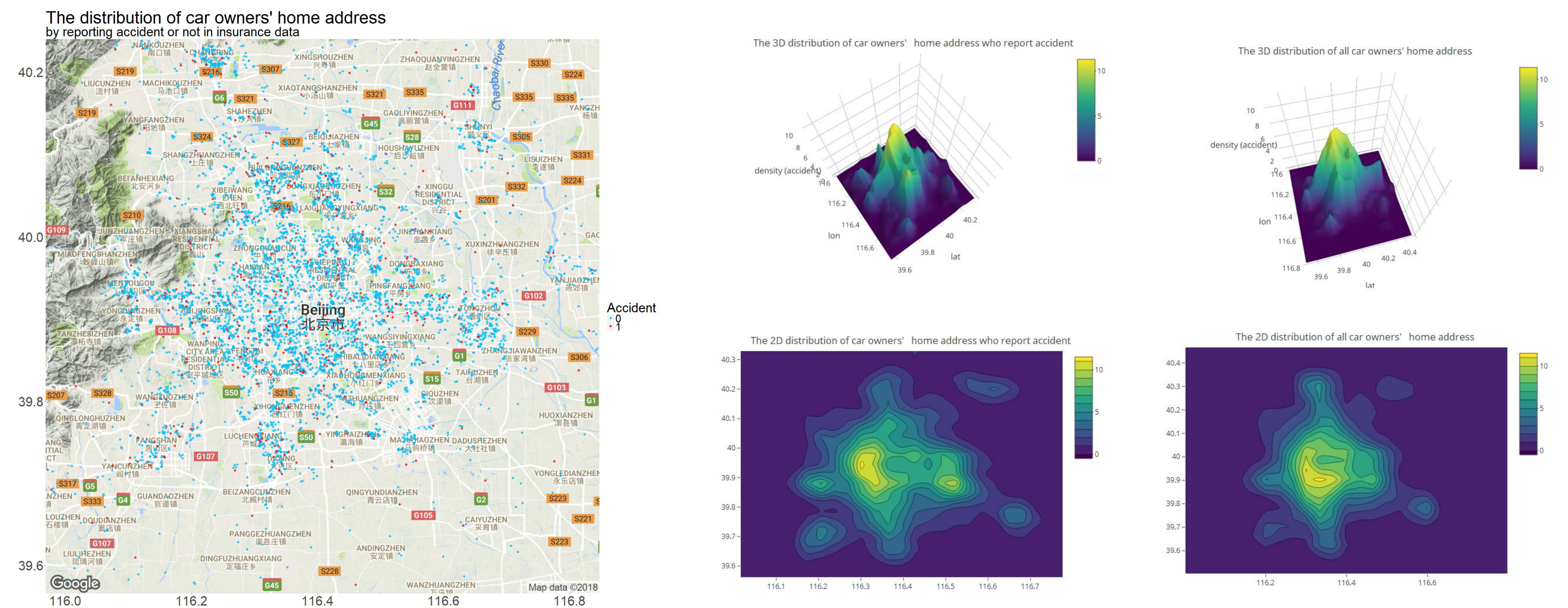
- **GPS DATA:** a three-month record of more than 10,000 cars, including GPS points and driving behavior records such as mileage and speed.
- **INSURANCE DATA:** claim history (used for measuring accident), insurance policy, a limited number of vehicle owners' demographic characteristics and some vehicle-related characteristics such as car price and type.

Identifying Home Address

For each car:
 d_i is the distance from GPS point i to the mean center of all points
 N_j is the number of GPS points in period j
 d_{final} is the distance between two GPS points when $N_j = 2$

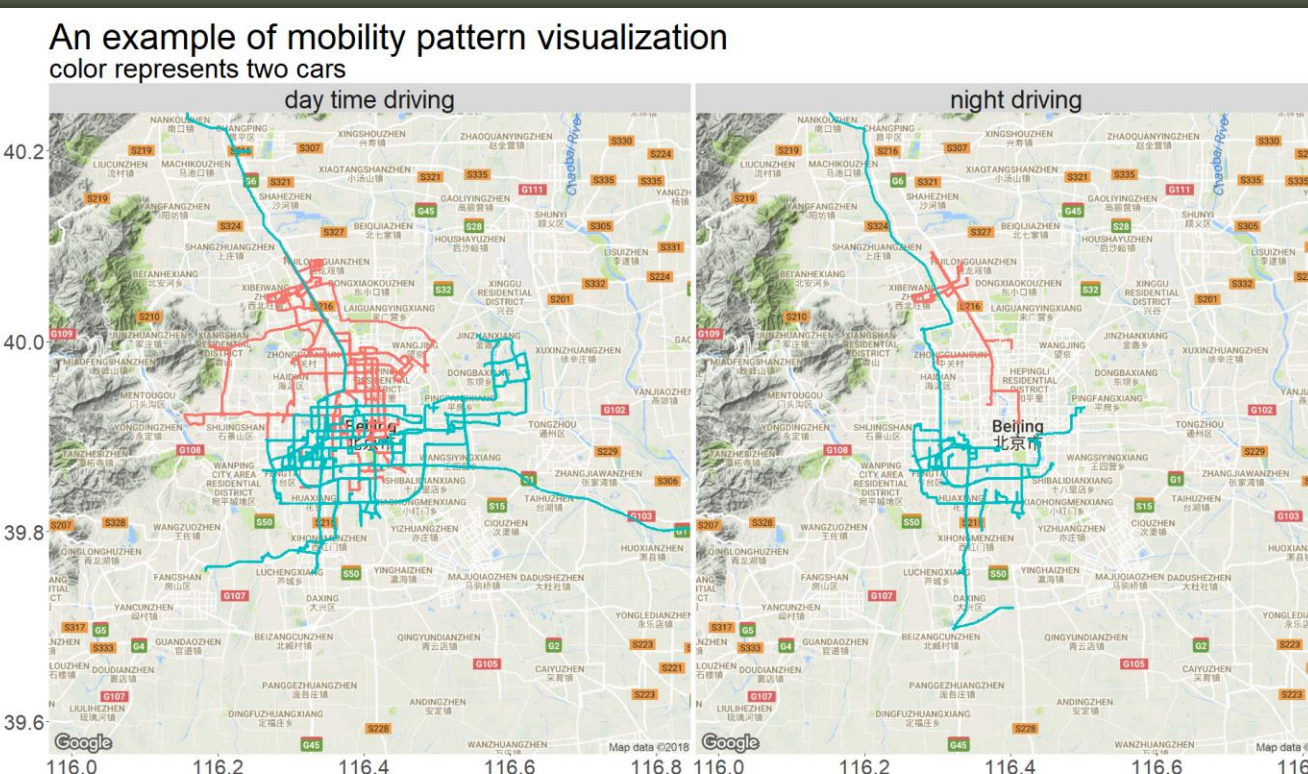


The Distribution of Car Owners' Home Address



The distributions of all car owners' home address and that of those who report accident are similar since more residents signifies more accidents in terms of quantity instead of proportion. In other words, the differences between them may imply different probability of accident. According to the maps and the socioeconomic layout in Beijing, I speculate that Haidian District, the south part of Beijing and northern resident area might provide some important information for accident prediction.

Mobility Pattern and Driving Behavior



The mobility pattern and driving behavior can be easily obtained because the GPS dataset also directly contains driving parameters. Generally, they can be divided into three groups:

Pure driving behavior	Social driving behavior
Acceleration and brake behavior, driving speed and familiar road driving	Driving record based on time and place such as night driving and urban driving

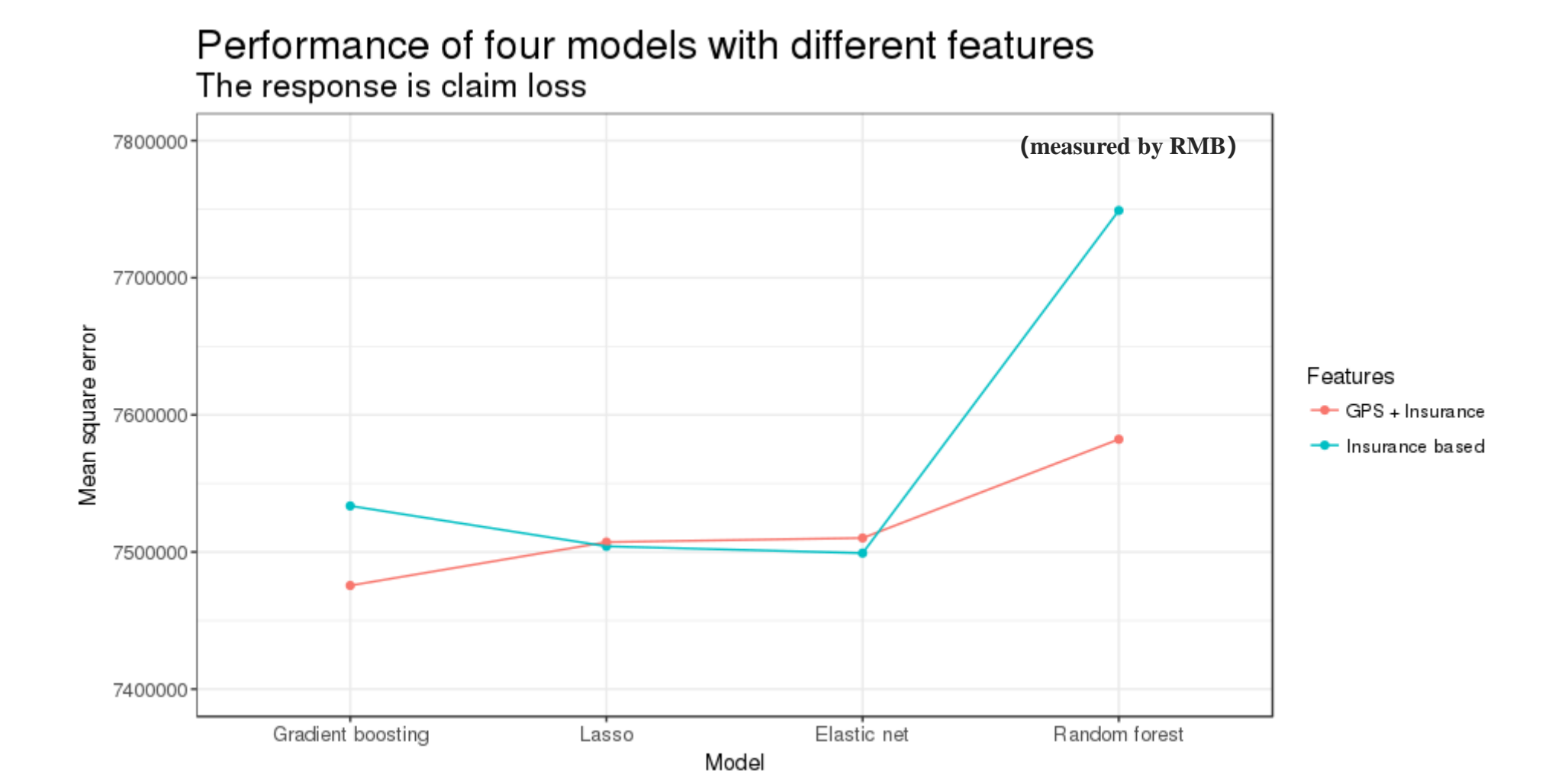
Binary Logistic Regression for Accident Analysis

Accident Classification in Binary Logistic Models

	Model	
	Insurance	GPS + Insurance
<i>Driver-related characteristics in the insurance dataset</i>		
FEMALE	0.0010 (0.009)	0.0032 (0.009)
YOUNG	0.0063 (0.010)	0.0094 (0.010)
OLD	0.0085 (0.009)	0.0070 (0.009)
STATE_JOB	-0.0015 (0.009)	-0.0006 (0.009)
INTERNET_SALE	-0.0244** (0.010)	-0.0251** (0.010)
ANNUAL_MIL	0.0184*** (0.003)	0.0156*** (0.003)
<i>Vehicle-related characteristics in the insurance dataset</i>		
CAR_PRICE	1.278e-08 (4.06e-08)	-6.401e-09 (4.12e-08)
NEW_CAR	0.0560*** (0.010)	0.0557*** (0.010)
BIG_CAR	-0.0179 (0.017)	-0.0169 (0.017)
AIRBAG	0.0002 (0.003)	0.0003 (0.003)
ALARM	-0.0148 (0.034)	-0.0161 (0.034)
<i>Home address extracted from GPS data</i>		
HAIDIAN		-0.0289 (0.021)
SOUTH		0.0270*** (0.010)
NORTH		0.0262* (0.014)
<i>Pure driving behavior extracted from GPS data</i>		
HARD_ACCL		-0.0020 (0.003)
HARD_BRK		0.0202*** (0.006)
<i>Social driving behavior extracted from GPS data</i>		
PCT_URBAN		0.0222 (0.023)
PCT_FREEWAY		-0.0108 (0.046)
PCT_LOCAL		0.0375 (0.056)
PCT_WKD		0.0323 (0.032)
PCT_WEEKEND		0.0026 (0.031)
PCT_NIGHT		0.0619* (0.035)
MEAN_FMLRT		-0.0090*** (0.003)
PCT_SPEED		0.0078 (0.062)
Intercept	0.0714** (0.035)	0.0349 (0.054)
Log likelihood	-1412.5	-1388.4
AIC	2849	2825
BIC	2930	2986

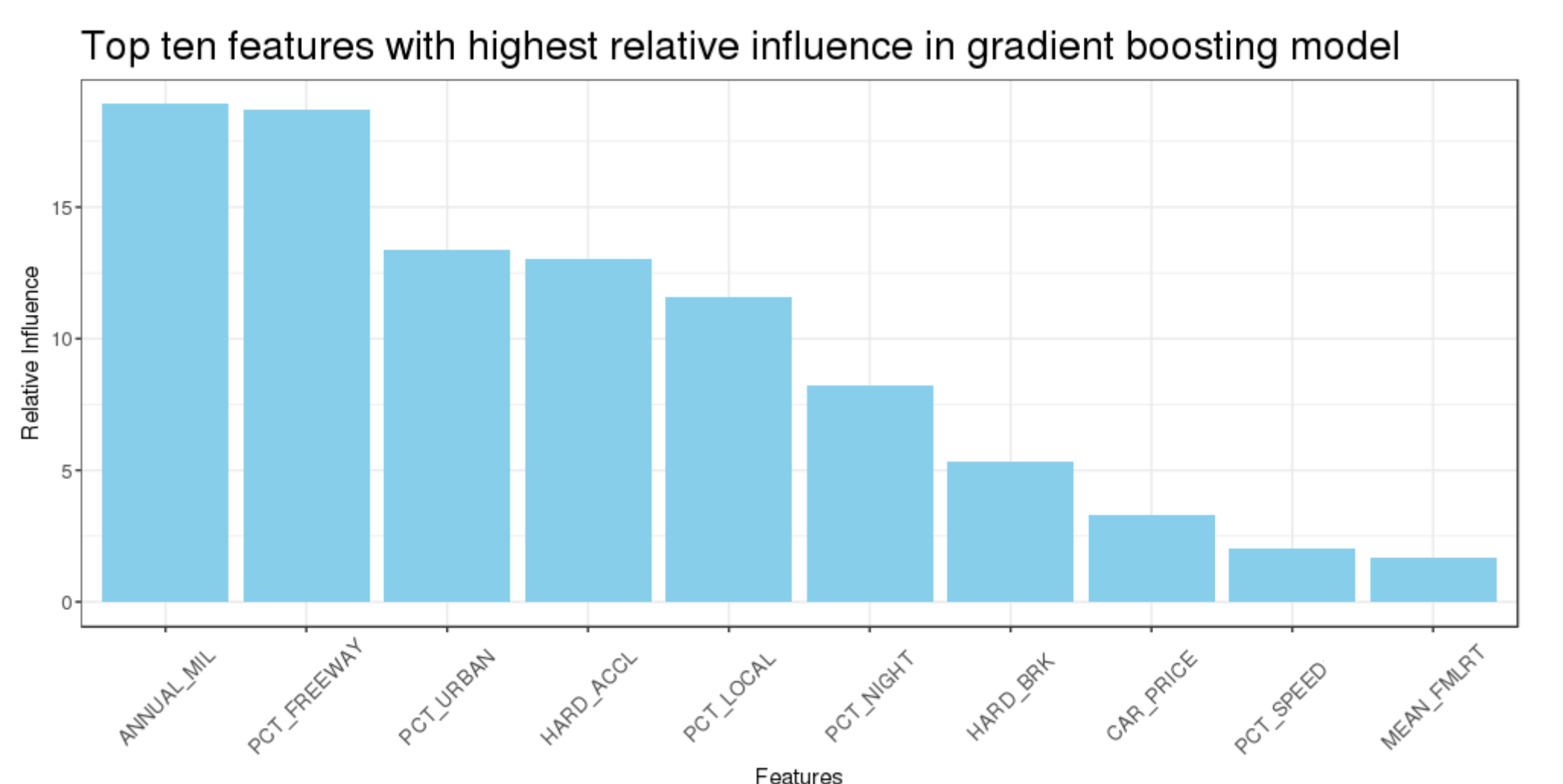
standard errors are in parentheses. ***p ≤ 0.01, **p ≤ 0.05, *p ≤ 0.1

Accident Loss prediction



The features extracted from GPS data do improve the prediction of accident loss (claim loss). The result shown in the graph is robust.

Feature Influence / Importance



The gradient boosting model with features from both GPS and insurance data heavily relies on driving behavior and mobility pattern.

Conclusion and Discussion

- It is possible to extract useful socioeconomic information from GPS data for predictive purpose. However, the difficulty in this step is not only to design an appropriate method/algorithm, but to interpret the GPS points and assign the social meaning to it.
- I identify car owners' home address and speculate some areas that might provide information for prediction. The logistic regression results show that this method is useful since the variables are significant or quasi-significant. .
- The combination of socioeconomic and driving behavior features extracted from GPS data can improve the prediction of both accident and its loss, since they increase the log likelihood in the logistic model and lower the mean squared error via gradient boosting model.
- Driving behavior is more important than socioeconomic characteristics in accident prediction.