THE UNIVERSITY OF CHICAGO

Using GPS Data to Predict Accident Risk: A Data Mining and Machine Learning Approach

By

Kanyao Han

July 2018

A paper submitted in partial fulfillment of the requirements for the
Master of Arts degree in the
Master of Arts Program in the Social Sciences

Faculty Advisor: Muh-Chung Lin
Preceptor: Caterina Fugazzola

Abstract

The accumulation of digital data and the development of computation capacity have ushered in a new era for data exploration and analysis and thus bring new opportunities for accident analysis as well as social science research. However, conventional social science disciplines, including accident analysis, still seldom use powerful machine learning and data mining methods and frequently confuse explanatory power with predictive power. This paper 1) discusses the theoretical foundations of machine learning and data mining methods in the context of the digital era and also 2) applies these methods to a GPS dataset and an insurance dataset for accident prediction modeling and improvement.

1. Introduction

The development and popularization of digital technology over the last decade have ushered in a new era for both data science and social science. Current digital data, which directly or indirectly embodies individual-level human behaviors and socioeconomic information, are accumulated at an unprecedented speed. It thus brings many new opportunities to social science research. Among various types of data, Global Position System (GPS) data has proven to be one of the most promising types, especially in accident analysis and prediction for public policy and insurance strategy (Karapiperis et al. 2015). This is because we can directly obtain individual-level driving behavioral information from it, which are usually not included in conventional insurance surveys (Jin, Deng, and Jiang 2018). Besides, compared to survey data, digital observational data also has two distinctive advantages: always-on and non-reactive (Salganik 2018). For the first advantage, a telematics device can constantly collect data as long as a car is running. As to the second advantage, the device must provide unbiased records, while the information documented in insurance surveys often has a bias due to drivers' reaction to possible insurance plans. For example, White (1976) finds that the self-reported driving mileage in insurance surveys is usually lower than the actual mileage since drivers know that higher self-reported mileage can lead to a more expansive plan.

Because of the potential of GPS data mentioned above and the recent commercialization of the concept of Usage-Based Insurance (UBI), also known as Pay-As-You-Drive (PAYD), in the car insurance industry (Karapiperis et al. 2015), there have been several studies that try to extract behavioral features from GPS data and analyze their roles in accident risk. For example, actual driving mileage (Paefgen, Staake, and Fleisch 2014,

Lemaire, Park, and Wang 2016, Litman 2005, Elvik 2015), rates of hard accelerations or hard brakes (Weidner, Transchel, and Weidner 2017, Handel et al. 2014, Bagdadi and Várhelyi 2011, Paefgen, Staake, and Fleisch 2014), strategic driving behaviors such as road and time selection (Tselentis, Yannis, and Vlahogianni 2017), and daily driving behaviors and mobility patterns such as nighttime driving and familiarity with driving routes (Jin, Deng, and Jiang 2018, Ayuso, Guillén, and Pérez-Marín 2014, 2016, Paleti, Eluru, and Bhat 2010, Behnood, Roshandeh, and Mannering 2014, Eluru et al. 2012) have been used to build statistical models for accident analysis and prediction in specific cities.

However, although the vast majority of previous studies directly or indirectly claim that their ultimate goal is for prediction, the models they use are typically explanatory, instead of predictive (it will be discussed later). This phenomenon is nothing new. Even in information science, Shmueli and Koppius (2010) also find that, among 52 empirical papers with predictive claims in the two top-ranked information science journals between 1990 and 2006, only seven properly built and tested predictive models. As Shmueli (2010) indicates, statistical models in many conventional (social) scientific disciplines are used almost exclusive for causal explanation, there is a lack of understanding of the difference between building explanatory models versus creating predictive models, and nonstatistician researchers often incorrectly assume that a model with high explanatory power is inherently of high predictive power.

In view of the lack of a clear distinction between predictive and explanatory models within social sciences, this paper aims to introduce predictive models mainly represented by machine learning methods in accident analysis. This paper has two parts. The first part discusses the theoretical foundation of machine learning methods in terms of statistical and

social theories, especially in the context of accident analysis and of the digital era. The second part is an application of machine learning methods in a specific GPS dataset, which aims to show when and how we mine data and build machine learning models in accident analysis, as well as whether GPS data can improve accident prediction.

2. Two Cultures in Statistics

2.1 Two cultures

In fact, there are "two cultures" in statistics (Breiman 2001). The biggest distinction between these two cultures echoes the classic opposition about the primacy of theory versus data (Ollion and Boelaert 2018). The modeling process of the first culture is typically explanatory. The vast majority of the models in the first culture are causal theoretical models, in which a set of underlying factors are theoretically assumed to cause an underlying effect, and the pre-established theoretical hypotheses are usually tested by significance test (Shmueli 2010). Sometimes they also report R^2, adjusted R^2, log-likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC) as the goodness of fit or the evidence of predictive power (Shmueli 2010, Ollion and Boelaert 2018). For example, recently, Jin, Deng, and Jiang (2018) publish a paper in the top-ranked accident analysis journal, *Accident Analysis & Prevention*. They analyze the effects of various independent variables on the depend variables and conclude that their variables extracted from their GPS data can improve accident prediction, since the coefficients of some GPS-related variables are significant at 10% level in binary and/or latent-class-specific logistic models, and the log-likelihood of their insurance-based binary

logistic model is significantly lower than that of the GPS + insurance one. On the other hand, models in the second culture usually don't have pre-established conception, theory and even hypothesis and thus are much more empirical (Ollion and Boelaert 2018, Shmueli 2010). In addition, these models often don't aim to build theory or test hypothesis (Chollet and Allaire 2018) since their highly flexible form results in the fact that we cannot precisely interpret the relationship between predictors and responses (Ollion and Boelaert 2018) as well as the variables selection and balance process. Empirically trying a set of models and various GPS-related variables are the typical modeling process. For example, the only paper that properly builds and tests models in recent accident studies uses several sets of models as well as variables and simply reports the values related to predictive accuracy (Paefgen, Staake, and Thiesse 2013).

In a long history, the explanatory culture dominated in most academic and industrial departments. However, because of the growing computation capacity, the accumulation of big data and the success of machine learning methods in the recent years, the predictive culture gradually wins an increasing number of proponents, especially in the industry. They believe that the lack of pre-established conception, theory and hypothesis is the strength of their methods since the success of machine learning methods has proven that the proper use of intense computation and customized algorithms without strong theoretical restrictions fits data better and brings higher predictive power. For example, in the last two years, Kaggle, the biggest platform for predictive modeling and analytics competitions, was dominated by gradient boosting method and deep learning since these two black-box machine learning methods usually perform the best in prediction task (Chollet and Allaire 2018).

The distinction mentioned above brings several mathematical and methodological differences between these two cultures.

a) Mathematically

The expected prediction error (or expected measurement squared error) for a new observation with value x_0 is given by (James et al. 2013):

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon)$$

The goal of explanatory modeling is to minimize the squared bias, which is the result of the misspecification of the underlying optimal statistical model, for obtaining the most accurate representation of underlying theory (optimal model), while the goal of predictive modeling is to minimize the entire equation (Shmueli 2010, Hastie, Tibshirani, and Friedman 2009, James et al. 2013).

b) Variable selection

While in the first culture variable selection is based on the role of the selected variables in theoretical or hypothetical structures as well as on model operationalization (such as avoiding multicollinearity), in the second culture "there is no need to delve into the exact role of each variable in terms of an underlying causal structure" (Shmueli 2010). Additionally, multicollinearity is not a problem for predictive purpose in variable selection unless we are interested in the individual regression coefficients since it will not affect the predictive ability of a model (Vaughan and Berry 2005, Makridarkis, Wheelwright, and Hyndman 1998).

c) Evaluation

As mentioned previously, the first culture usually uses significance test, R^2, adjusted R^2, log-likelihood, AIC and BIC. All these statistical inferences are calculated from the data that is used to train the model. However, the second culture usually uses cross-validation methods to evaluate the predictive power of their models. Although there are many cross-validation methods and statistics, their principle is the same. Scholars split their data into more than one parts. Some are used to train models and some are used to test them. Specifically, the model trained from the training set will use the predictors (independent variables) in the testing set to do estimation, and then the estimated values are compared to the actual values.

2.2 Criticism and Defense

Although it has achieved huge success in the industry, the second culture still receives many harsh criticisms mainly because the methods in it usually don't have pre-established conception, theory and even hypothesis and thus are thoroughly empirical. The criticisms can be summarized into three sets of statements. First, the well-performed methods in the second culture are usually too flexible to be interpreted and thus we can know "whether", instead of "why" and "how" (Ollion and Boelaert 2018). Second, most machine learning models, whose mathematical foundations are not based on probability theory, cannot provide conventional statistics such as p-value so that they are not scientific enough. Third, this characteristic of the second culture increases the uncertainty of actual prediction even though the model we train in the training set have a high predictive accuracy in the testing set. Put differently, no matter what characteristics the testing set has,

it can simply represent the condition of the past and/or the known. It is unconvincing to use the known to test whether a model can predict the unknown when we do not understand the mechanism of the predictors. Salganik (2018) indicates that digital observational data has a bad characteristic, which is drifting. We cannot guarantee that there is no drifting in terms of generation, behavior and system in the future or in other locations. For example, assuming that 1 years later the transportation regulation in Beijing has a big change, individual driving behaviors and patterns probably become different and thus the predictive models tested by today's testing set are not useful.

I admit these criticisms are not completely wrong, but they ignore five critical facts. First, in most empirical and statistical cases, there is a tradeoff between the pursuit of "whether" is incompatible with that of "why and how". In statistical words, there is a bias-variance tradeoff (James et al. 2013). A model that have the minimized bias often cannot have the minimized sum of squared bias and variance. The method selection should be based on our goal.

Second, having more statistical inferences does not mean conventional explanatory models are more "scientific" than machine learning models. Although the conventional statistical models can be tested by a set of statistical inferences, many statisticians have indicated that some of them in social sciences are problematic (Ollion and Boelaert 2018, Wasserstein and Lazar 2016). First, many studies use adjusted R^2 to represent the predictive power of their regression models, which is given by:

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)\,(n-1)}{n-p-1} \right]$$

$$where \;\; R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

This statistic uses the number of variables and that of observations to calibrate R^2 in order to avoid overfitting. However, James et al. (2013) prove that this kind of statistic cannot avoid overfitting well. He uses different models that have the same numbers of variables and observations but the flexibility. The residual sum of squares gradually decreases with the increase of the flexibility. Since he uses the same simulated dataset and the same numbers of variables and observations, the adjusted R^2 will gradually increase. However, when he compares the actual model that was used to simulate the dataset, we can find that the adjusted R^2 cannot represent the actual model fit. Instead, cross-validation is more consistent with the actual model fit. Finally, James et al. (2013) state that most conventional statistics are solely based on the training set and cannot avoid overfitting well. Besides, the use of p-value, the quasi-golden standard in many academic disciplines, has been criticized by many statisticians. On one hand, American Statistical Association (ASA) published a statement in 2016, which highlights that p-value do not measure the probability of whether the studied hypothesis is true, and that there are many other approaches to measure models and coefficients (Wasserstein and Lazar 2016). On the other hand, in prediction tasks, significance or insignificance cannot measure whether the variables can or cannot improve prediction (Shmueli 2010). It is worth noting that the problems in conventional statistics do not mean that the measures in machine learning methods are

perfect. Rather, we should select proper measures and statistics according to specific datasets and goals.

Third, the characteristics of the data in the digital era are different from those of conventional data (Salganik 2018) and thus legitimate the efficiency of new prediction methods, but most criticisms still ignore them. As mentioned above, a very important characteristic of digital data is always-on (Salganik 2018). This means the predictive model can be dynamically calibrated by new data. Therefore, the goal of machine learning methods is continuous short-term prediction rather than long-term prediction. Although machine learning methods sometimes simply split a dataset related to past events into two, and then train in one and test in the other, we can still assume that the measurement squared error can measure its short-term predictive power. In other words, we only use our model trained and tested by the known to predict those which are close to the known in terms of time, location or other characteristics. An example of the importance of calibration and short-time prediction is Google Flu Trends (GFT). The researchers in Google achieve a huge success in flu trends prediction using past search engine query and official flue trends data (Ginsberg et al. 2009). However, several years later, Lazer et al. (2014) find that the actual prediction accuracy (measured by the actual flu trends when there are official flu reports) of their model trained and tested by past data would decrease over time since the researchers did not calibrate their model. A solution to this problem is dynamically to calibrate the model by new data, and the actual prediction accuracy can be constantly high (Lazer et al. 2014). Due to current computation capacity, it is easy to design an automatic algorithm for new data input and model calibration for short-term prediction. Additionally, how frequently we should calibrate previous models depends on many factors. For facial

recognition task such as using facial images to identify sexual orientation (Wang and Kosinski 2018), we need not calibrate models frequently since physical and biological characteristics will not change a lot in several years. On the other hand, for accident risk prediction, frequent calibration is desired since transportation policies and socioeconomic layouts of a city often change.

Actually, there is another criticism that is related to digital data instead of machine learning methods: despite the large size of observations or the large volume of information, digital data is often unrepresentative since it is collected by specific approach (Salganik 2018). For example, not all drivers will buy telematics device. Those who use it probably differ from those who don't use it. Similar to the fact that the second culture does not pursue long-term prediction, machine learning predictive models usually don't pursue external validity. In other words, machine learning models usually don't aim to generalize the result to other situations or to other groups of people. Instead, they focus on a small specific group of people and further apply the model in this kind of group. For example, in accident analysis related to GPS data, it is necessary to remove those who seldom use their cars since their actual driving behaviors cannot be detected well (Jin, Deng, and Jiang 2018). In this case, when the insurance companies make usage-based insurance plans, drivers without a minimum driving time should not be enrolled into plans. Put differently, a better way is to build another model or algorithm that focuses on those who seldom use their cars. In addition, there is no doubt that even though the data is collected from a specific approach, the characteristics of data might change. For example, the drifting of user of specific websites is a common phenomenon. In this case, model calibration based on new data is desired.

Finally, many conventional estimation methods do outperform machine learning models, but the goal of many predictive studies is not optimal prediction. Rather, they aim to balance cost, timeliness, accessibility and predicted results. For example, many developing countries cannot afford the cost of a census or a survey and thus an accurate estimation based on them. In this case, a problematic but cheap predictive model is better than nothing. For this purpose, Pokhriyal and Jacques (2017) and Blumenstock, Cadamuro, and On (2015) use mobile phone data to predict poverty respectively in Senegal and Kigali. The second kind of studies is used to balance the timeliness and accuracy. Google Flu Trends (Ginsberg et al. 2009) and population mapping in Portugal and France (Deville et al. 2016) aim to obtain information in real time, even though delayed official reports can provide more accurate information. The third kind of studies is to obtain the information that is not accessible for specific reasons. The event of Cambridge Analytics is an example. Cambridge Analytics uses an online game to collect users' private information, builds machine learning models that use Facebook records as predictors, and then further predict other users' private information through their Facebook records. Of course, this activity is unethical and illegal in many countries. However, their computation and modeling designs have been ethically and legally studied by two previous studies (Goel, Mason, and Watts 2010, Kosinski, Stillwell, and Graepel 2013). Therefore, as long as this kind of studies is carried out under regulation, it still provides a way to obtain and predict previously inaccessible information for ethical and legal usage. This paper will also use computation and modeling methods to extract information from my GPS data.

In summary, two cultures in statistics have their own advantages, research goals usages and target data types. We should not ignore, blame or even abandon either simply because they have unavoidable disadvantages.

3. Data and Method

3.1 Data and variables

I use two datasets related to cars in Beijing for the application of machine learning methods. 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 an average of 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 variables for accident prediction: claim history, drivers' sociodemographic characteristics and vehicle-related characteristics. The variables related to claim history are indicators of car accidents and hence can be used as the responses of whether a car has accidents and how severe an accident is. The driver-related and vehicle-related characteristics can also be used to build insurance-based baseline models. These two datasets can be merged by vehicle identification number at the individual level. Therefore, it is possible to use new features extracted from the GPS dataset to enrich the information for prediction. Since vehicle identification number belongs to identifying information, this study has been reviewed and approved by the IRB at the University of Chicago.

3.2 Variables

Although most machine learning methods do not require variables selection based on pre-established theory and hypothesis before modeling since they are selected during computationally intense modeling, I still make sure that the vast majority of variables have been proven that they can be used for explanation since I also build models when machine learning methods are not suitable.

Conventional car insurance often uses driver-related and vehicle-related characteristics in insurance surveys, including, for example, drivers' gender, age, and occupation, as well as age, price and type of vehicle. (Weidner, Transchel, and Weidner 2017, Tselentis, Yannis, and Vlahogianni 2017, Jin, Deng, and Jiang 2018). Usually, they can be correlated with accident risk, and the detailed results vary according to specific cities. For example, Jin, Deng, and Jiang (2018) find that the variable of new car can expect a higher accident probability and big car can expect low probability. In addition to this basic kind of information, annual mileage is usually a very important variable in both conventional insurance modeling and GPS-related insurance modeling. Almost all previous studies find that it has a positive correlation with accident risk (Elvik 2015, Lemaire, Park, and Wang 2016, Litman 2005, Paefgen, Staake, and Fleisch 2014).

Table 1 shows the variables in my insurance dataset. There are two responses (or independent variables), claim and claim amount, that can be used to measure self-report accident and its severity. In addition to the basic variables mentioned above, there are three more variables: number of airbag, safe belt alarm and internet sale. For the first two variables, it is safe to assume that the equipment of safe belt alarm and airbag can reduce the probability of accident and/or its severity. However, for the third variables, although

Jin, Deng, and Jiang (2018) find that it is negatively correlated with accident risk, we cannot provide a solid interpretation since we don't have enough socioeconomic information about those who purchase the insurance on the Internet. Therefore, I simply regard it as a control variable in conventional modeling.

Table 1: Summary statistics of features/response in the insurance data

Features/response	Definition	Mean	S.D.
Responses			
CLAIM	Self-report accident	0.105	0.306
CLAIM AMOUNT	Claim Amount (Chinese Yuan)	295.6	2650
Driver-related characteristics in the insurance dataset			
FEMALE	Drivers are female	0.283	0.450
AGE	Drivers' age	39.07	9.869
STATE_JOB	Working for the state	0.370	0.483
INTERNET_SALE	Buying insurance via internet	0.215	0.411
ANNUAL_MIL	Annual driving mileage	1.989	1.413
Vehicle-related characteristics in the insurance dataset			
CAR_PRICE	Car price	199001	123540
NEW_CAR	Cars were purchased in recent three years[1]	0.195	0.396
BIG_CAR	Cars belong to Minivan/SUV	0.060	0.237
AIRBAG	Number of airbags	4.539	1.939
ALARM	Equipped with a safe belt alarm	0.987	0.115

For the driving behaviors and patterns extracted from GPS dataset, there are two kinds of variables. The first type is pure driving behavior information. Number of hard brakes and accelerations, driving speed and familiarity with roads have been studied in recent research. Handel et al. (2014) find that the rate of hard brake is very useful for insurance pricing. Jin, Deng, and Jiang (2018) also have two findings. First, they divided the fraction of mileage of driving speed into several classes and find that 90 km/h is a good cutoff for modeling. Second, they confirm that drivers' familiarity with road has a negative correlation with accident risk. The second kind of variables is defined as exposure

[1] The definition of new car is provided by the insurance company.

variables that are used to measure driving time and region (Paefgen, Staake, and Fleisch 2014, Paefgen, Staake, and Thiesse 2013). Night driving, weekend driving, local driving, freeway driving and urban driving are all proven to be useful for insurance pricing or significantly correlated with accident risk (Jun, Guensler, and Ogle 2011, Ayuso, Guillén, and Pérez-Marín 2016, 2014, Jin, Deng, and Jiang 2018). Therefore, I extracted these variables from the GPS dataset. Table 2 is a list of these variables.

Table 2: Summary statistics of driving behaviors/patterns

Features	Definition	Mean	S.D.
HARD_ACCL	Average number of hard accelerations in one hour	0.684	1.490
HARD_BRK	Average number of hard brakes in one hour	0.636	0.814
PCT_SPEED	The fraction of mileage of driving with speed below 90 km/h	0.931	0.075
PCT_URBAN	The fraction of mileage of driving in the urban area	0.510	0.296
PCT_FREEWAY	The fraction of mileage of driving on the freeways	0.461	0.184
PCT_LOCAL	The fraction of mileage of driving on the local roads	0.381	0.119
PCT_WEEKEND	The fraction of mileage of driving during weekends	0.444	0.152
PCT_NIGHT	The fraction of mileage exposure of driving during nights	0.157	0.122
MEAN_FMLRT	The frequency in the same roads.	3.289	1.479

It is worth noting that drivers' living region has also been considered in previous accident research (Jin, Deng, and Jiang 2018). In many social science disciplines, living regions at various scales are often used as area-based measures of socioeconomic status, including race/ethnicity, education, income, poverty and property values, and then scholars use them to explore the relationship between their target events and area-based features (Drewnowski, Rehm, and Solet 2007, Krieger et al. 2002, Gordon-Larsen et al. 2006, Krieger et al. 2003). For accident prediction, this implies that variables related to regions might provide marginal socioeconomic information since the insurance dataset does not contain direct demographic and socioeconomic variables other than gender, age and whether a driver works for the state. However, the relationship between the area-based

measures of socioeconomic status and accident are seldom discussed in accident analysis, and the few uses are also relatively empirical. Besides, even if we assume that we already have a knowledge of it, identifying the socioeconomic segregation patterns based on conventional methods is still extremely difficult in Beijing because of the house demolition system and the chaotic house-rental market. On one hand, many poor-educated and low-income "native Beijingers", who was born and grew up in Beijing, obtained a large amount of money or several expensive houses from the government as the compensation of house demolition in gentrification period and areas, where there are also many high-income and well-educated immigrants. Second, due to the lack of regulation of the house-rental market in Beijing, Rental agencies often illegally change house structures to cater to market demand. For example, they often divide a house into many extremely small rooms and thus low-income people can rent a room in some neighborhoods with high real estimate prices. These two facts result in the difficulty for the pattern recognition of socioeconomic segregation in terms of living region.

In view of this, a substitution method is to directly explore area-based accident pattern. Because conventional GIS methods should be based on underlying theories or experiences about the appropriate geographical unit of analysis (Grubesic and Matisziw 2006), they are not suitable in my data and analysis which lack a clear knowledge of the appropriate unit. In this case, I use kernel density estimation (KDE) to directly explore the hidden area-based accident patterns. KDE is a non-parametric way to estimate the probability density function of variables, which does not require a clear knowledge of the appropriate geographical unit and we can use mathematical methods to look for hidden data patterns. Although this method cannot provide the knowledge of the relationship

between area-based socioeconomic variables and accident risk, it can directly imply which areas can be used as variables for accident analysis and/or prediction.

3.3 Area-based feature extraction

My insurance dataset does not contain drivers' home addresses. Even if it contains them, it is not safe to directly use them. In Beijing, it is very difficult for residents to get new car plates due to the strict regulation of car plate application, and thus there is a large market for long-term car-rental. This fact means we cannot guarantee that the home addresses in the insurance data are car-users'. Therefore, identifying where car-users live via GPS data is a better choice. A typical approach to this task is to assume that most GPS data points at night can represent the home address, and then calculate their mean longitudes and latitudes. However, this method has two problems that lead to the unrepresentativeness of mean longitudes and latitudes. First, for most cars, there are outliers which usually result in a severe error of home address estimation. Second, a small percentage of car owners have more than one residential addresses. Due to these two factors, the points of mean longitudes and latitudes will not represent the home address.

I use an algorithm to solve these problems. Figure 1 shows an example of the process of the algorithm. For each car, N_j is the number of GPS points in period j, C_j is the mean center of GPS points in period j, d_{ij} is the distance from GPS point i to the mean center C_j, and d_{final} is the distance between two GPS points when N_j is 2. I select the first GPS points after 5:00 a.m. each day for each car and calculate C_1 and d_{i1}. The first mean center (C_1) of a car in Figure 1 is far away from the home address represented by the purple area. In order to obtain a slightly better mean center in every period, the point that has the

largest distance is removed, and then a renewed mean center and renewed distances are recalculated. This process is repeated again and again until $N_j = 2$. As the figure shows, the center will gradually approximate to an area that has the most points with a small variance of distance. Besides, I also remove around 1% of cars when the distance between the final two points of each car is too large since this result implies either that these cars don't have enough GPS records or that their GPS points cannot represent home addresses.

Figure 1: The process of the algorithm

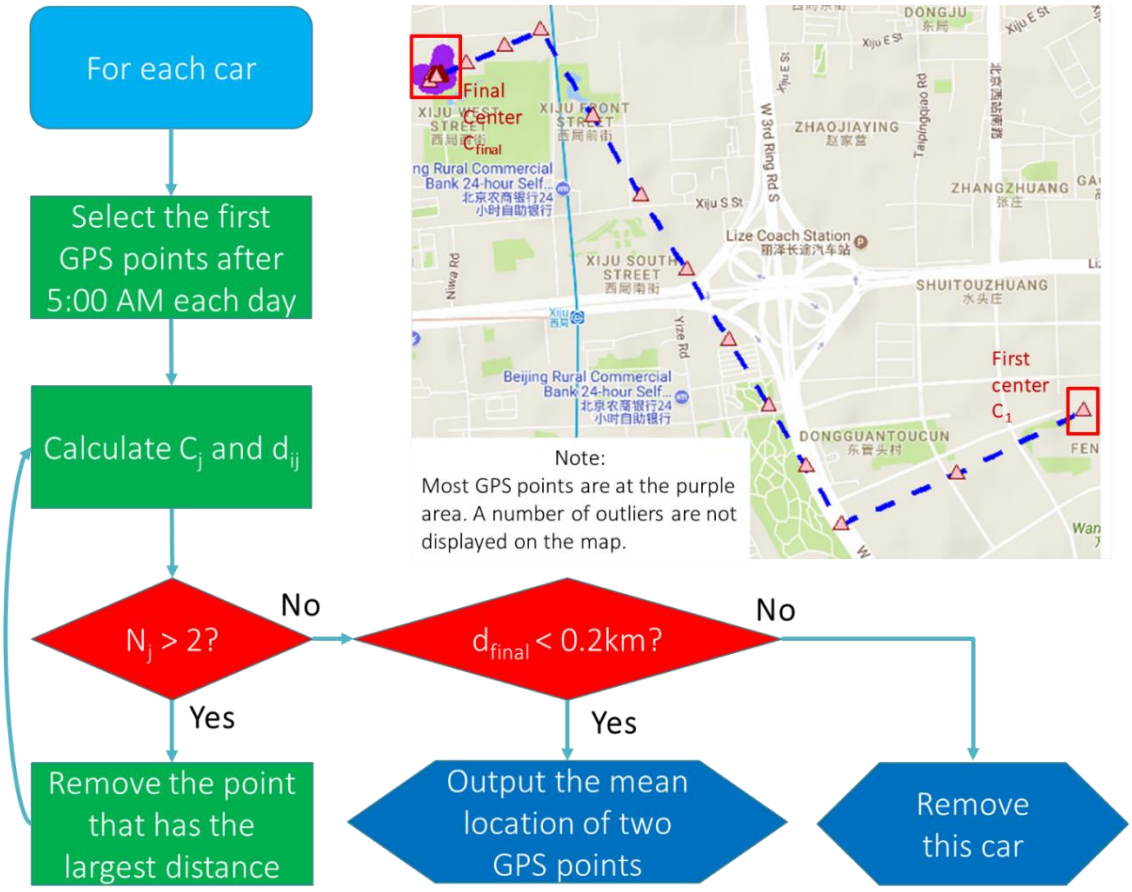


Figure 2: The distribution of home address

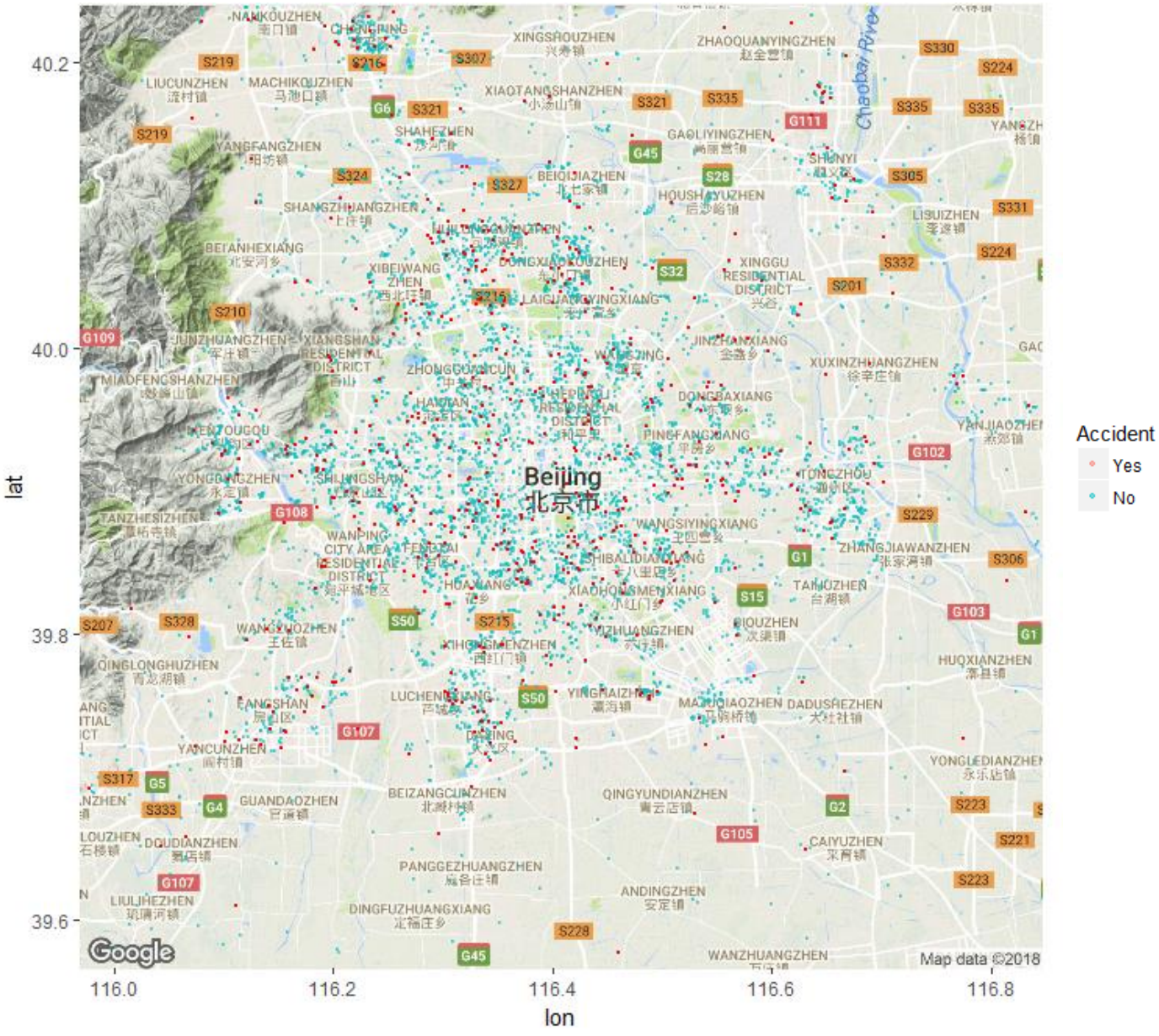


Figure 2 shows the result of the estimated home address. Actually, we cannot obtain enough useful information from this obscure figure. Therefore, I calculate the distribution densities through kernel density estimation and visualize the results. The kernel density estimate is given by

$$\hat{f}_H(X) = \frac{1}{n}\sum_{i=1}^{n} K_H(X - X_i)$$

and the kernel function I choose is

$$K(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^T \mu}{2}}$$

such that

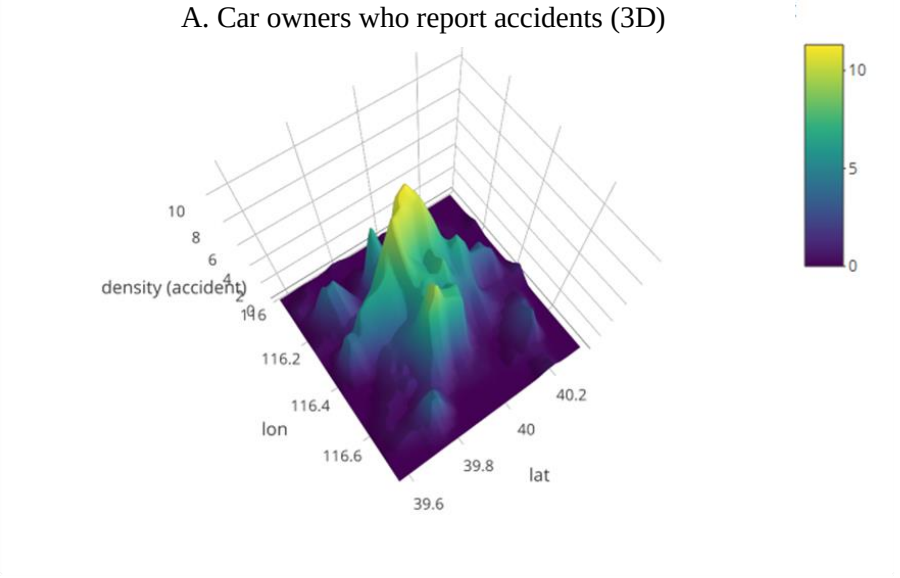
$$K_H(X) = \frac{1}{2\pi} |H|^{-\frac{1}{2}} e^{-\frac{1}{2} X^T H^{-1} X}$$

In the estimator, $X = (X_1, X_2)^T$ contains two vectors: $X_1 = (X_{11}, X_{21}, ..., X_{i1})$ and $X_2 = (X_{12}, X_{22}, ..., X_{i2})$. They are respectively the longitudes and the latitudes in the data.

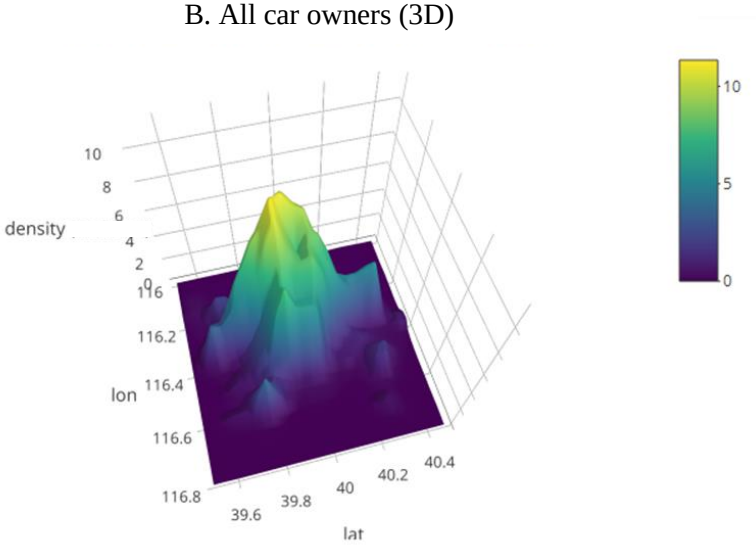
It is worth noting that the bandwidth H, a 2×2 symmetric matrix, always has a big influence on the estimation and visualization, we should find its value that can bring a bias-variance balance of the kernel density estimate and assume that this kind of estimate will reflect an appropriate distribution pattern. In other words, the visualization of the distributions should be smooth enough so that we can recognize their patterns, but we should also simultaneously avoid the problem of oversmoothing that results in a high bias.

Figure 3: Kernel density estimates of the home address distribution

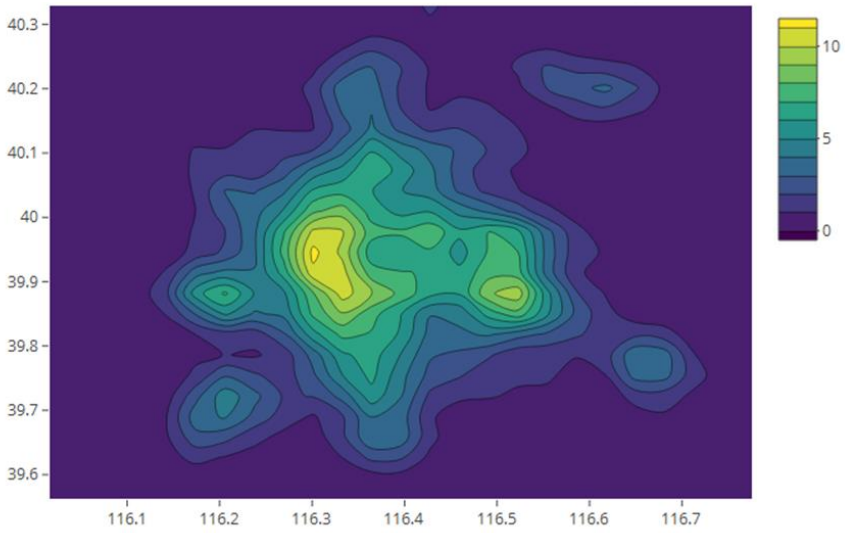
A. Car owners who report accidents (3D)



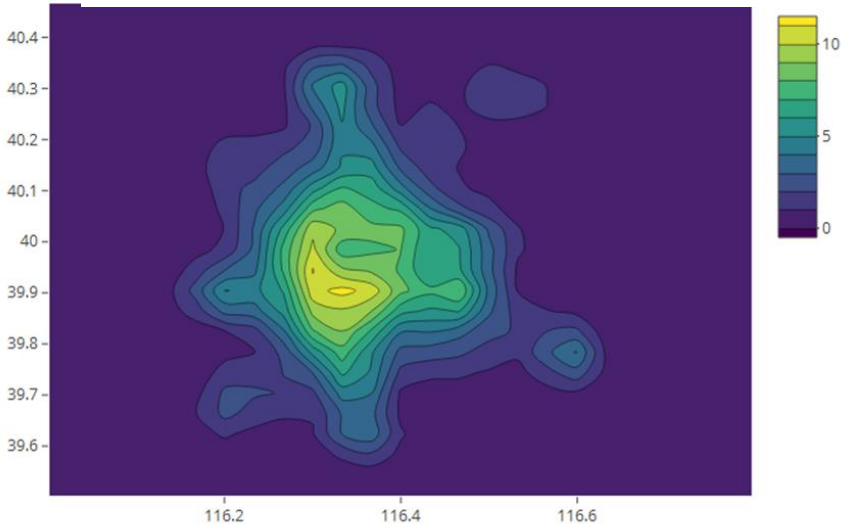
B. All car owners (3D)



C. Car owners who report accidents (2D)



D. All car owners (2D)



I adopt the modified cross-validation method proposed by Stute (1992) to determine the bandwidth H of the estimation equation. Since my purpose is to compare the overall home address distribution with the accident-involved drivers' home address distribution and also because only when the density estimation has the same bandwidth H can we directly compare them, I directly apply the values from the former to the latter.

The final result is shown in Figure 3. Facet A and C are the home address distribution density of those who report accidents and Facet B and D are all car owners'. There is no doubt that the distributions of the home address of these two groups are visibly not identical. This fact means the effects of some latent factors in terms of accident probability can be well represented by area-based information. Mathematically, if area-based information cannot represent the latent effects well, the probability densities of these two groups will be too similar to be recognized. Therefore, the areas with different densities can be used as the area-based indicators of accident probability.

In addition to the differences in density distributions shown on the map, I also take the socioeconomic space-time layout of Beijing into account. This is because the results of the KDE method, which heavily relies on the selection of the bandwidth H, are a bit arbitrary. Finally, I believe Haidian District, which is the northwestern part of the inner city, the southern part of the main urban area, and the northern residential area might provide important information for accident analysis and prediction. These three areas not only show a higher or lower probability of accident on the map but also have distinctive socioeconomic characteristics. a) Haidian District is the education and technology center of Beijing; 2) the southern part of the main urban area is the relatively underdeveloped urban area comparing to other areas, and 3) the northern residential area has a large number

of residents with a high proportion of tenants. Therefore, it is safe to assume that these three areas can be used for accident analysis and prediction. Finally, three area-based dummy variables I extracted from the GPS data is living in Haidian District, living at the southern part of the main urban area and living in the northern resident area.

4. Models, validation and testing

4.1 Accident/claim classification

For accident classification, I choose conventional binary logistic regression models with statistical inference analysis instead of machine learning classifiers with cross-validation. There are several reasons for this choice. First and most importantly, accidents are very complicated processes based on a set of physical, socioeconomic and interactional contexts. The information contained in variables are unlikely to be sufficient and greatly improve the accident prediction. For example, although Paefgen, Staake, and Thiesse (2013) find the prediction accuracies can be higher than 80 % in logistic regression, decision tree and neural networks models with driving behavior features, they also admit that the high accuracies do not mean any exciting result since the typical imbalanced response in accident prediction sample[2] results in a serious overestimation of the majority class and thus a heavy classifier bias. Specifically, when the accident-involved and accident-free vehicles are respectively considered as the positive class and the negative class, the accuracy is calculated as

[2] The ratio between the accident-free class and the accident-involved class is usually 1 to 9.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ [3]

and the precisions of positive and negative classes are calculated as

$$Precision_{involved} = \frac{TP}{TP + FP} \quad and \quad Precision_{free} = \frac{TN}{TN + FN}$$

In typical car-accident classification with imbalanced classes, the precision of the accident-involved class is very likely to be extremely low, while that of the accident-free class is thus very high. This is because models often classify almost all observations as accident-free. For instance, assuming that the ratio between the accident-free class and the accident-involved class in the testing set is 1 to 9, the predictive accuracy is 90% if all observations in the testing set are classified as accident-free. Obviously, this is a very bad model since all accident-involved cars are misclassified as accident-free cars. Therefore, the superficially high accuracy cannot provide any implication for insurance pricing strategy and public policy making. In this case, models with statistical inference become a better choice even though they cannot provide direct measures of accident prediction. Besides, since Paefgen, Staake, and Thiesse (2013) also find that the performances of decision tree and neural networks models are only slightly better than logistic regression models in accident prediction problem with driving behaviors, we need not to worry much about the real prediction performance of logistic regression models when we don't use cross-validation to evaluate them. Additionally, I also build a logistic model using the AIC-stepwise method which selects variables one by one and is based on conventional significance tests as well as AIC.

[3] TP, TN, FP and FN are respectively true positive, true negative, false positive and false negative.

4.2 Claim amount prediction

In the task of claim amount prediction, the difficulties in terms of insufficient information and bad predictive performance mentioned in accident classification cannot be completely solved. However, they will not become big problems in this section. First, compared to accident classification, the response of accident amount has a higher variance and thus provides more information for model building. Second, even though the improvement of the amount prediction is slight, it is still possible to make pricing decision or public policy evaluation based on the specific predicted accident amount. Therefore, in this section, I select four machine learning algorithms: least absolute shrinkage and selection operator (lasso), elastic net, random forest and gradient boosting tree. The first two algorithms are regularized regressions and the last two are tree-based methods.

Both lasso and elastic net algorithms are based on ordinary least squares (OLS). However, different from OLS that tries to obtain unbiased estimates, regularized regressions belong to biased estimation methods that aim to largely decrease the variance of estimates at the cost of a slight increase in squared bias. The coefficient solution of lasso is given by

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where p is the number of variables and the second term is a penalty term in which the tuning parameter λ determines how strong the penalty is. Elastic net is similar to lasso but has a more complicated penalty term. It can be written as

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \left(\sum_{j=1}^{p} [\alpha |\beta_j| + (1-\alpha) \beta_j^2] \right)$$

where the parameter λ determines the degree of penalty and the parameter α determines its type.[4] For these two methods, the most important thing is to find the optimal λ and α. For each method, I split the data into a training set and a testing set, build a set of models with different parameters based on the training set, adopt 10-fold cross-validation to calculate the mean squared error (MSE) of each model and then select the parameters that can result in the lowest MSE. In my dataset, the best λ of the lasso model is 57.4, while the best α and λ of the elastic net model are 0[5] and 3865.[6] I finally test the predictive powers of the optimal lasso and elastic net models based on the testing set.

Different from lasso and elastic net that are based on OLS method, random forest and gradient boosting tree are based on decision tree that takes advantage of a method called recursive binary splitting and can deal with any kind of highly non-linear relationship. Specifically, in the first step, decision tree selects a predictor X_j and a cut-point such that it splits the predictor space R into two regions: $R_1(j, s) = \{X \mid X_j < s\}$ and $R_2(j, s) = \{X \mid X_j >= s\}$ and minimize the residual sum of squares (RSS) which is given by

$$RSS = \sum_{x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2$$

$$\left(\hat{y}_{R_1} = \overline{y}_{R_1} \; and \; \hat{y}_{R_1} = \overline{y}_{R_2}\right)$$

After the first step, R_1 and R_2 can further be respectively split into two sub-regions by this method. The process can be repeated again and again until we believe further splitting will result in the overfitting problem. Finally, the whole repeated process is seen

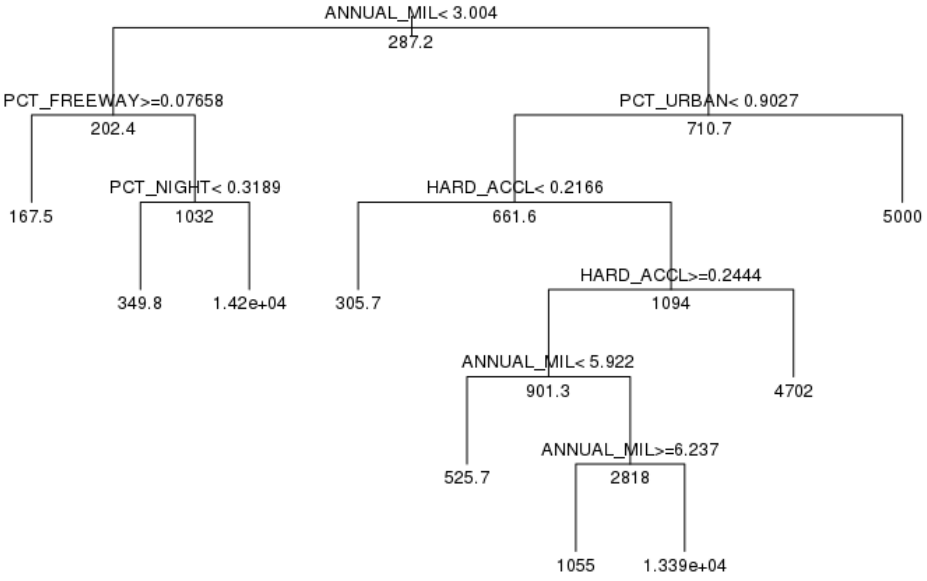
[4] Although in theoretical studies lasso can be seen as a special form of elastic net when α = 1, in practice they are different and lasso often outperforms elastic net. This is because elastic net is more likely to result in the problem of overfitting.

[5] α = 0 means it is actually a ridge regression model, which is also a special form of elastic net.

[6] I only report the parameters of the GPS + Insurance models. Those of the insurance-based models are different.

as a tree. Figure 4 shows an example of a medium decision tree fit from my data. The feature of annual mileage is obviously non-linear because it is used several times in the model.

Figure 4: A decision tree with a depth of 6



Random forest is a method of combining hundreds or thousands of trees to obtain a robust result. If the number of variables we want to use is p and that of observations is n, we randomly select $m \approx \sqrt{p}$ variables and a bootstrapped sample with n observations to build a decision tree. Next, we use this decision tree model to predict the observations that are not in the bootstrapped sample. By using this method, we can build q trees and obtain q sets of results. The final predicted result is the average of them and thus has a low variance. Gradient boosting tree also aims to combine trees, but it uses residuals in every step as the response for tree growth. The process of this method can be written as

$$Initial\ setting:$$
$$\hat{f}(x) = 0\ and\ r_i = y_i$$
$$Iterating\ with\ for\ loop:$$
$$for\ tree\ in\ 1\ to\ q$$
$$\{\hat{f}(x_i) = \hat{f}(x_i) + \lambda \hat{f}^{tree}(x_i)$$
$$r_i = r_i - \lambda \hat{f}^{tree}(x_i)\}$$

where the parameter λ is used to control the learning rate. The control parameter λ will not have a big influence on the final result as long as it is small enough.[7] As for the number of trees that does not matter in random forest models in terms of overfitting, too many trees in gradient boosting models are very like to result in this problem.[8] I use cross-validation method, in training set, to select the optimal number of trees. Specifically, I calculate the MSE for each tree and find that 1320 trees (λ = 0.001) can bring the lowest MSE. Additionally, the depth of trees in my model, another parameter that controls the size of each tree to avoid overfitting, is 4. This parameter will be very important when I evaluate different types of feature in Section 5.2.

Of course, as the final procedure of lasso and elastic models, I finally test the predictive powers of the optimal random forest and gradient boosting models via the testing set.

[7] However, the control parameter determines how many trees have to be built. Its decrease signifies that we need more trees.
[8] The standard of "too many" is determined by the value of λ.

5. Results

5.1 Accident classification

Table 3 shows the logistic regression results whose dummy dependent variable is self-report accident. We can first take a look at the model performances. Generally, likelihood ratio test, AIC and BIC are widely used. Although the vast majority of scholars think they are in competition, they measure different dimensions of model performance (Sober 2002). Specifically, AIC is an indicator of predictive power and the other two are measures of explanatory power. Since the GPS-related model is respectively nested with the other two, we can respectively test them in terms of likelihood ratio test. The result shows that there is no significant difference between the GPS-related model and the stepwise model, but the GPS-related model has a significantly bigger likelihood than the insurance-based model. In other words, the insurance-based model has the lowest explanatory power. However, when we look at BIC values, the GPS-related model become the worst model with the lowest BIC value. This fact means the GPS-related model has a higher overfitting risk since BIC is a better overfitting detector than likelihood ratio test. While the stepwise model seems the best explanatory model, it also has a problem. Even though stepwise method often helps us build a statistically optimal explanatory model, it is not advisable unless we can guarantee that the variables we remove do not have actual effects despite bad statistical inference results (Jaccard 2001).

Table 3: Accident classification in binary logistic models

Variables	Definition	Models		
		Insurance	GPS	Stepwise
Driver-related characteristics in the insurance dataset				
FEMALE	Drivers are female	0.008	0.030	
		(0.094)	(0.095)	
AGE	Drivers' age	-0.001	0.001	
		(0.004)	(0.107)	
STATE_JOB	Working for the state	-0.016	-0.008	
		(0.090)	(0.091)	
INTERNET_SALE	Buying insurance via internet	-0.323 ***	-0.339 ***	-0.324 ***
		(0.125)	(0.126)	(0.120)
ANNUAL_MIL	Annual driving mileage	0.160 ***	0.130 ***	0.133 ***
		(0.025)	(0.028)	(0.028)
Vehicle-related characteristics in the insurance dataset				
CAR_PRICE	Car price	1.578e-07	-1.629e-09	
		(4.3e-07)	(4.37e-07)	
NEW_CAR	Car was purchased in recent three years	0.503 ***	0.498 ***	0.501 ***
		(0.098)	(0.099)	(0.098)
BIG_CAR	Car belongs to Minivan/SUV	-0.190	-0.194	
		(0.188)	(0.189)	
AIRBAG	Number of airbags	0.003	0.005	
		(0.028)	(0.028)	
ALARM	Equipped with a safe belt alarm	-0.0143	-0.132	
		(0.349)	(0.354)	
Home address extracted from GPS data				
HAIDIAN	Living at Haidian District		-0.409	-0.417
			(0.279)	(0.275)
SOUTH	Living at the south of the main urban area		0.278 ***	0.262 **
			(0.107)	(0.103)
NORTH	Living in the northern resident area		0.255 *	0.269 **
			(0.139)	(0.133)
Driving behaviors and patterns extracted from GPS data				
HARD_ACCL	Average number of hard accelerations in one hour		-0.025	
			(0.032)	
HARD_BRK	Average number of hard brakes in one hour		0.187 ***	0.159 ***
			(0.054)	(0.042)
PCT_SPEED	The fraction of mileage of driving with speed below 90 km/h		-0.028	
			(0.687)	
PCT_URBAN	The fraction of mileage of driving in the urban area		0.205	
			(0.256)	
PCT_FREEWAY	The fraction of mileage of driving on the freeways		-0.148	
			(0.535)	
PCT_LOCAL	The fraction of mileage of driving on the local roads		0.360	
			(0.645)	
PCT_WEEKEND	The fraction of mileage of driving during weekends		-0.334	
			(0.382)	
PCT_NIGHT	The fraction of mileage exposure of driving during nights		0.611 *	0.621 *
			(0.361)	(0.349)
MEAN_FMLRT	The frequency in the same roads.		-0.123 ***	-0.121 ***
			(0.036)	(0.035)
Intercept		-2.403 ***	-2.225**	-2.390 ***
		(0.392)	(0.921)	(0.152)
Log-likelihood		-2039.3	-2015.5	-2017.7
AIC		4100.6	4077	4055
BIC		4175	4232	4123

standard errors are in parentheses. $***p \leqslant 0.01$, $** p \leqslant 0.05$, $* p \leqslant 0.1$

Although we cannot tell which model is the best one, it is still possible to use these models for explanation. When we consider the coefficients, we can find that the results are highly robust in terms of the coefficients, their standard errors as well as the significance test results. Therefore, despite the difference in statistical inferences, the actual explanatory powers of the last two models are similar. To some extents, the second model is even better than the third one since it provides the information related to more variables that are empirically or theoretically important to accident risk in previous studies.

As for the explanatory coefficient results, we can obtain two conclusions from it. First, my models statistically confirm many findings in previous studies that annual mileage, new car, hard brake and night driving are positively correlated with self-report accident while familiarity with roads is negatively correlated with roads. Second, two of the tree area-based dummy variables extracted from the GPS data have a robust significant result in the models: living in the southern part of the main urban area and living in the northern residential area expect a higher probability of self-report accident. Since the variables are derived from KDE, a data-driven substitution method when we lack the knowledge of geographical segregation in Beijing, the results demonstrate that it is possible to use computational pattern recognition methods to mine useful information when we don’t have a solid theory about the relationships between the new recognized information and the target event. What’s more, according to the characteristics of these two areas where there is a moderately larger proportion of residents, tenants and migrant workers with middle or low income and education who usually don't have good family background, we can inductively speculate that accident risk can be closely correlated with the socioeconomic layout of a city as well as the socioeconomic status of a driver. It’s true that

this kind of interpretation is unsolid and highly speculative due to the lack of detailed data and systematic empirical studies in China, but this kind of data-driven pattern recognitions do bring a new research question or direction: why can these areas be correlated with self-report accident, and what are the latent factors and effects behind them?

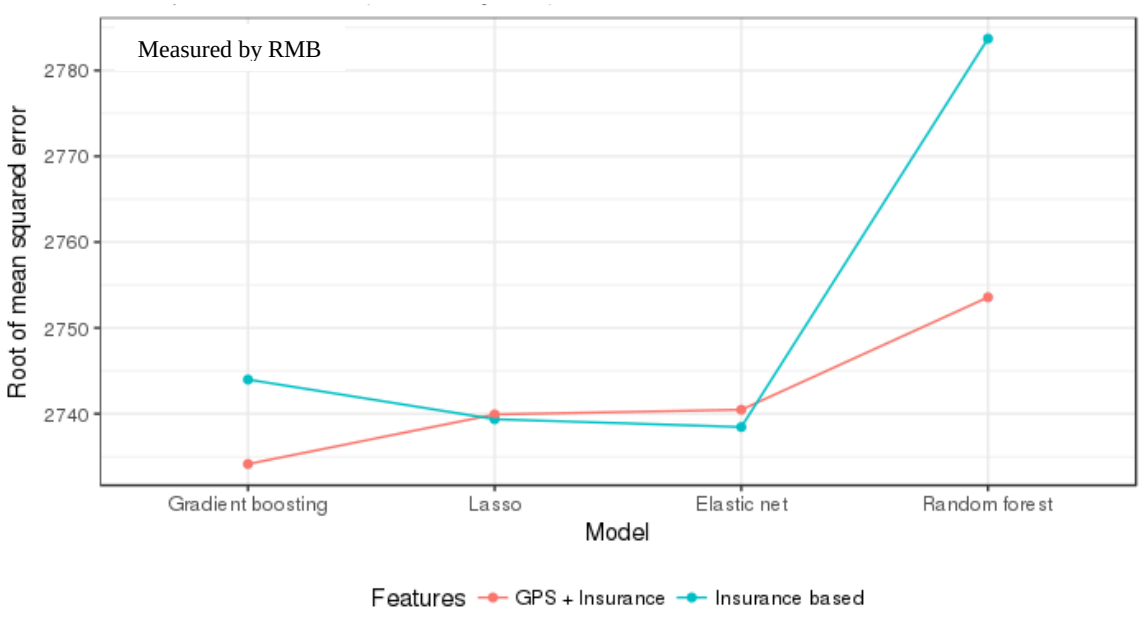
Different from the evaluation of explanatory power, prediction power evaluation is much easier. The only statistic we need to concern here is AIC. Obviously, the third model has the best performance while the first model has the worst. From this result we can speculate that adding GPS-related into conventional models can improve accident prediction. Besides, since this stepwise regression model excludes all variables that cannot improve AIC when they are respectively added into the model one by one, we can also know that all three area-based variables, as well as three driving behavior variables (hard brake, night driving and familiarity with roads), do improve AIC. Of course, we need to admit that AIC is actually not a good predictive metric since it solely based on the training set and thus cannot avoid overfitting well. However, when it is not suitable to build models via machine learning methods, a problematic predictive indicator is better than nothing.

5.2 Claim amount prediction

Figure 5 displays the final cross-validation results of the claim amount prediction in the testing set. Among these four algorithms, two models based on random forest have very high MSEs, although adding GPS-related features significantly reduces the MSEs when we compare it with the insurance-based model. Another tree-based algorithm, gradient boosting tree, has a much better performance. It not only reduces the MSE when I add the features extracted from the GPS data, but also helps us obtain the lowest MSE

among eight models. Besides, the models based on lasso and elastic net have very similar MSEs. It means, for these two algorithms, whether I add GPS-related features almost has no influence. According to these results, we can conclude that the features extracted from the GPS data can improve accident amount prediction if and only if we choose an appropriate model.

Figure 5: Performance of four models with different features

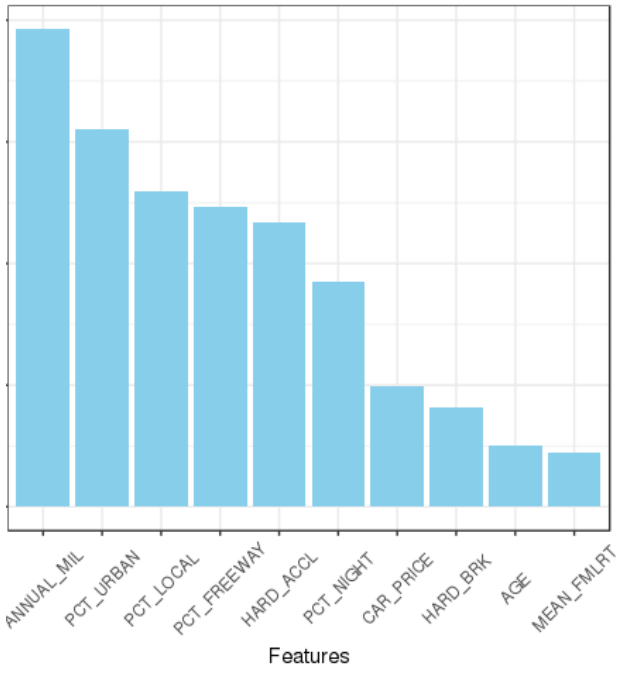


Additionally, we should also ask why and how the features and models work. Figure 6 and Table 4 show the feature importance of four GPS + Insurance models. In tree-based models, behavioral features have very large influences. Especially in the best-performed gradient boosting model, eight of the top ten features are driving behaviors, six of them are extracted from the GPS data. The random forest model also relies on behavioral features but gives too much importance to only two features. This might be the reason for its poor performance. As for the coefficients in Table 4, the lasso model removes all driving behavior features so that the result of the baseline model is almost the same as that of the

GPS + Insurance model. The elastic net model, which is actually a ridge regression model because the parameter α is 0, does not remove any feature from the model and thus relies on both sets of features. However, since lasso and ridge regression have very similar shrinkage processes, it is safe to believe that the latter still relies more on the features in the insurance data. This is why the predictive powers of the lasso and elastic net models are very similar no matter whether we add GPS-related features or not. Besides, the difference of feature importance between tree-based models and the other two models also provides us a very important information: the relationship between driving behavior features and the response are very likely to be non-linear. This is why the linear-based lasso and elastic net models heavily shrink the coefficients of behavior features and thus cannot improve claim amount prediction, and the tree-based models that can deal with highly non-linear relationships give much more importance to the behavioral features in prediction.

Figure 6: Most important features in tree-based models

A. Gradient boosting tree



B. Random Forest

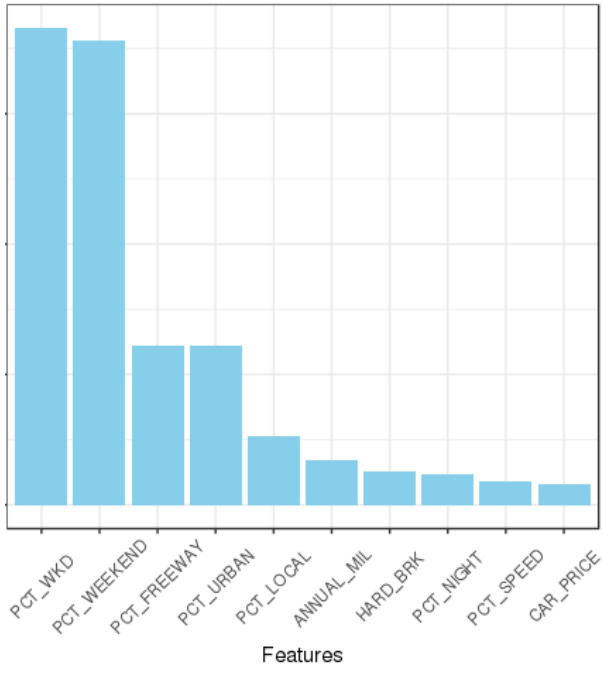


Table 4: Coefficients in lasso and elastic net models

Features	Lasso	Elastic net
(Intercept)	124.86195	238.1374285
FEMALE	0.00000	-39.6983812
AGE	0.00000	-1.4078986
STATE_JOB	0.00000	-5.3195166
INTERNET_SALE	-43.16971	-75.3376973
ANNUAL_MIL	75.75039	42.5669875
CAR_PRICE	0.00000	0.0000799
NEW_CAR	94.68740	101.6758343
BIG_CAR	0.00000	-53.5523879
AIRBAG	0.00000	5.1872406
ALARM	0.00000	55.0133765
HARD_ACCL	0.00000	11.9825361
HARD_BRK	0.00000	2.0905299
PCT_URBAN	0.00000	4.3814444
PCT_FREEWAY	0.00000	-128.2260988
PCT_LOCAL	0.00000	-209.5152897
PCT_WKD	0.00000	74.8755466
PCT_WEEKEND	0.00000	-74.7806586
PCT_NIGHT	0.00000	225.8326604
PCT_SPEED	0.00000	75.3444686
MEAN_FMLRT	0.00000	-11.8825541
SOUTH	0.00000	-21.4410552
HAIDIAN	0.00000	-93.5227628
NORTH	0.00000	-23.5936012

Additionally, the interpretation of the feature importance/influence must be very cautious. Although the behavioral features are much more important than other features in the best-performed model, it does not mean the former is much more useful for prediction

than the latter. The most persuasive evidence in my dataset is that the best-performed gradient boosting model performs only slightly better than all lasso and elastic net models and moderately better than the insurance-based gradient boosting model. Actually, the values of feature importance are influenced not only by the real usefulness of the features in the model but also by the working mechanism of the algorithms as well as the parameters I choose. The feature importance in each tree of gradient method is determined by the number of times a feature is selected, weighted by the improvement it brings in each split, and the final importance is the average of all trees. Because the decision trees give priority to features with a larger proportion of pure information related the response in the model when they select splitting features, we can speculate that socioeconomic features contain a larger proportion of noisy information than behavioral features. Besides, a tree has contained more mutual information when it has selected more features, and thus the marginal information new features can provide is less when most features have mutual information with each other as we assumed. In theory, if socioeconomic features can still provide marginal information, they still have chances to be selected even though they are at the very bottom of a tree. However, the very small value of tree depth I choose in Section 4.2 largely reduces the probability. Therefore, although the socioeconomic features can provide only slightly less useful information for prediction than the combination of socioeconomic and behavioral features, the feature importance in the best-performed gradient boosting model still shows a vast difference. In view of this, a more appropriate conclusion in this section is that the behavioral features extracted from the GPS dataset do improve accident risk prediction, but their role should not be overstated.

6. Concluding Discussion

The accumulation of digital data and the development of computation capacity have ushered in a new era for data exploration and analysis and thus bring new opportunities for accident analysis as well as social science research. However, due to the indiscrimination of two cultures in statistics, conventional social science disciplines still seldom use many powerful machine learning and data mining methods. This paper discusses the theoretical foundations of machine learning and data mining methods and also applies these methods in a GPS dataset.

The application of machine learning and data mining for predictive purpose should be understood in the new era where the characteristics of data have drastically changed. In the culture of prediction, most researchers will never think their specific models and estimations have long-term usefulness or can be generalized to other groups or datasets. Rather, they continuously calibrate their black-box models with new data and wish that every model can be useful in a very small field and in a very short time period but that the overall process is sustainable. My attitude is the same as many predictive researchers in terms of prediction. In fact, the models only tested by the testing set in the last section are far from being useful despite the result of a low MSE. I still need more data to calibrate and test it. Every predictive model is an always unfinished business but a testing result from a holdout set is always the first step.

The word “black-box” mentioned above only refers to that machine learning methods often have a very complex and repeated process of variables selection, balance and regularization and thus obtaining parameters or statistical inferences is almost impossible. In fact, we can evaluate and interpret how variables work based on the

mathematical foundation, the initial parameters and the characteristic of data. This is why this paper talks a lot about technique details.

However, as I mentioned in the modeling section about the predictive accuracy trap, machine learning can use and often uses beautiful but useless results to deceive other people. Therefore, it is very important to use conventional statistical models when machine learning is not suitable even for prediction. Even though most conventional statistical models and inferences is not good enough for predictive modeling, an imperfect exploration is better than doing nothing. For example, the logistic models in this paper obviously don't have high predictive powers, although they can obtain a very high but deceptive prediction accuracies. However, the logistic models still provide some information. They 1) validate several correlations in many previous studies, 2) test the relationship between the area-based variables, which is recognized from the GPS data, and self-report accident, and 3) tell us through AIC values in a stepwise model that driving behaviors and area-based variables at least are likely to improve prediction accuracy.

It is worth noting that pattern recognition methods such as KDE should also be valued in social science studies. In my project, although we can also use conventional GIS methods to recognize patterns, it is difficult to determine the scale of geographic unit for variable projection when we still lack a clear understanding of a city or an event. It is true that the pattern recognition without pre-established theory or conception is dangerous, but its results, as well as the unclear relationships discovered by machine learning models, can be regarded as latent hypotheses or suggestion. They can further be tested and interpreted in explanatory models after proper datasets become available.

Reference

Ayuso, Mercedes, Montserrat Guillén, and Ana María Pérez-Marín. 2014. "Time and Distance to First Accident and Driving Patterns of Young Drivers with Pay-as-you-Drive Insurance." *Accident Analysis & Prevention* 73:125 - 131.

Ayuso, Mercedes, Montserrat Guillén, and Ana María Pérez-Marín. 2016. "Using GPS Data to Analyse the Distance Travelled to the First Accident at Fault in Pay-as-You-Drive Insurance." *Transportation Research Part C: Emerging Technologies* 68:160 - 167.

Bagdadi, Omar, and András Várhelyi. 2011. "Jerky driving—An indicator of accident proneness?" *Accident Analysis & Prevention* 43 (4):1359 - 1363.

Behnood, Ali, Arash M. Roshandeh, and Fred L. Mannering. 2014. "Latent Class Analysis of the Effects of Age, Gender, and Alcohol Consumption on Driver-Injury Severities." *Analytic Methods in Accident Research* 3 (4):56 - 91.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264):1073 - 1076.

Breiman. 2001. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical Science* 16 (3):199 - 231.

Chollet, and Allaire. 2018. *Deep Learning with R*. New York: Manning Publications Co.

Deville, Pierre, Chaoming Song, Nathan Eagle, Vincent D. Blondel, Albert-László Barabási, and Dashun Wang. 2016. "Scaling Identity Connects Human Mobility and Social Interactions." *PNAS* 113 (26):7047 – 7052.

Drewnowski, Rehm, and Solet. 2007. "Disparities in obesity rates: Analysis by ZIP code area." *Social Science & Medicine* 65 (12):2458 - 2463.

Eluru, Bagheri, Miranda-Moreno, and Fu. 2012. "A latent class modeling approach for identifying vehicle driver injury severity factors at hightway-railway crossing." *Accident Analysis & Prevention* 47:119 - 127.

Elvik. 2015. "Some implications of an event-based definition of exposure to the risk of road accident." *Accident Analysis & Prevention* 76:15-24.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (1012 - 1015).

Goel, Mason, and Watts. 2010. "Real and Perceived Attitude Agreement in Social Networks." *Journal of Personality and Social Psychology* 99 (4):611 - 621.

Gordon-Larsen, Nelson, Page, and Popkin. 2006. "Inequality in the Built Environment Underlies Key Health Disparities in Physical Activity and Obesity." *Pediatrics* 117 (2).

Grubesic, and Matisziw. 2006. "On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data." *International Journal of Health Geographics* 5 (58).

Handel, Peter, Isaac Skog, Johan Wahlstrom, Farid Bonawiede, Richard Welch, Jens Ohlsson, and Martin Ohlsson. 2014. "Insurance Telematics: Opportunities and Challenges with the Smartphone Solution." *IEEE Intelligent Transportation Systems Magazine* 6:57 - 70.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Jaccard. 2001. *Interaction Effects in Logistic Regression*. CA: SAGE.

James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Jin, Wen, Yinglu Deng, and Hai Jiang. 2018. "Latent Class Analysis of Accident Risks in Usage-Based Insurance: Evidence from Beijing." *Accident Analysis & Prevention* 115:79-88.

Jun, Guensler, and Ogle. 2011. "Differences in observed speed patterns between crash-involved and crash-not-involved drivers: Application of in-vehicle monitoring technology." *Transportation Research Part C: Emerging Technologies* 19:569 - 578.

Karapiperis, Dimitris, Birny Birnbaum, Aaron Brandenburg, Sandra Castagna, Allen Greenberg, Robin Harbage, and Anne Obersteadt, eds. 2015. *Usage-Based Insurance and Vehicle Telematics: Insurance Market and Regulatory Implications, CIPR Study Series 1*.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. "Private Traits and Attributes are Predictable from Digital Records of Human Behavior." *PNAS* 110 (15):5802 – 5805.

Krieger, Chen, Waterman, and Soobader. 2002. "Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project." *American Journal of Epidemiology* 156 (5):471 - 482.

Krieger, Chen, Waterman, Soobader, Subramanian, and Carson. 2003. "Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US)." *J Epidemiol Community Health* 2003 (57):186 - 199.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176):1203 - 1205.

Lemaire, Park, and Wang. 2016. "The use of annual mileage as a rating variable." *Astin Bulletin* 46:39-69.

Litman. 2005. "Pay-as-you-drive pricing and insurance regulatory objectives." *Journal of Insurance Regulation* 23:35.

Makridarkis, Wheelwright, and Hyndman. 1998. *Forecasting: Methods and Application, 3rd ed.* New York: Wiley.

Ollion, and Boelaert. 2018. "The Great Regression. Machine learning, Econometrics, and the Future of Quantification." *Revue Française de Sociologie* forthcoming.

Paefgen, Staake, and Fleisch. 2014. "Multivariate Exposure Modeling of Accident Risk: Insights from Pay-as-You-Drive Insurance Data." *Transportation Research Part A* 61:27 - 40.

Paefgen, Staake, and Thiesse. 2013. "Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach." *Decision Support Systems* 56:192 - 201.

Paleti, Eluru, and Bhat. 2010. "Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes." *Accident Analysis & Prevention* 42:1839 - 1854.

Pokhriyal, Neeti, and Damien Christophe Jacques. 2017. "Combining Disparate Data Sources for Improved Poverty Prediction and Mapping." *PNAS* Publish online October 31, 2017:E9783–E9792.

Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age*. New Jersey: Princeton University Press.

Shmueli. 2010. "To Explain or to Predict?" *Statistical Science* 25:289 - 310.

Sober. 2002. "Instrumentalism, parsimony, and the Akaike framework." *Philos. Sci* 69:S112 - S123.

Stute, W. 1992. "Modified Cross Validation in Density Estimation." *Journal of Statistical Planning and Inference* 30:293 - 305.

Tselentis, D. I., G. Yannis, and E. I. Vlahogianni. 2017. "Innovative Motor Insurance Schemes: A Review of Current Practices and Emerging Challenges." *Accident Analysis & Prevention* 98:139 - 148.

Vaughan, and Berry. 2005. "Using Monte Carlo Techniques to Demonstrate the Meaning and Implications of Multicollinearity." *Statistical Education* 13.

Wang, Yilun, and Michal Kosinski. 2018. "Deep Neural Networks are more Accurate than Humans at Detecting Sexual Orientation from Facial Images." *Journal of Personality and Social Psychology* 114 (2):246 - 257.

Wasserstein, and Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70 (2):129 - 133.

Weidner, Wiltrud, Fabian W.G. Transchel, and Robert Weidner. 2017. "Telematic Driving Profile Classification in Car Insurance Pricing." *Annals of Actuarial Science* 11 (2):213 - 236.

White, S. B. 1976. "On the Use of Annual Vehicle Miles of Travel Estimates from Vehicle Owners." *Accident Analysis & Prevention* 8 (4):257-261.