Using GPS Data to Predict Accident Risk: Evidence from Beijing

Kanyao Han

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 socioeconomic characteristics. In this paper, I extract 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 and build machine learning model, the extraction method in this paper has a methodological implication.

1. Introduction

The development and popularization of digital technology over the last decade have ushered a new era for both data science and social science. Current digital data, which directly or indirectly embodies individual-level human behaviors and socio-demographic information, are being accumulated in 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 it can provide two critical sets of always-on information pertinent to accident analysis: individual-level moving behavior/pattern and socio-demographic information. More importantly, these sets of information are usually not included in traditional insurance survey (Jin, Deng, & Jiang, 2018), and hence they can provide supplementary information for accident analysis and prediction.

However, the use of GPS data is still insufficient in accident analysis for two reasons. First, current studies heavily focus on driving behaviors and patterns that can be directly obtained from vehicle telematics data, but socioeconomic information embodied in GPS data has not attracted enough attention. Second, machine learning models, which often outperform traditional statistical models in terms of prediction problems, are seldom used in accident prediction. In view of this, there are two goals in this paper. First, I design an algorithm to identify home address through GPS points and further mine useful features from it. Second, I combine socioeconomic and driving behavior features extracted from a GPS dataset with driver- and vehicle-related features from an insurance dataset, build several machine

learning models for accident analysis and prediction, and finally analyze model prediction accuracies and feature importance.

The final results of this paper provide four implications for accident analysis and prediction. First, it is possible to accurately extract useful socioeconomic information from GPS data for predictive purpose. However, the difficulty in this step is not only the design of an appropriate algorithm that can compute an accurate result, but the interpretation of GPS points based on space-time layout of a city before algorithm design. Second, 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. Third, 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 of the logistic model and lower the root of mean squared error via a gradient boosting tree model. Finally, driving behavior and mobility pattern can be more important than socioeconomic characteristics in accident prediction, but both sets of features have advantages and disadvantages when we comprehensively consider cost, accessibility and predictive power.

2. Possibility and usefulness: an empirical approach

What is the relationship between socioeconomic characteristics and behaviors? Which is more important in terms of knowing and explaining our society? Which is more important for pure prediction problems? Since these questions are closely related to feature extraction and selection, I make three assumptions in my project related to GPS data. First, both socioeconomic characteristics and behaviors from GPS data can provide useful information for accident prediction. Second, they can be both supplementary and substitutionary since

they share part and only part of information. Third, because the data we can obtain is always incomplete, taking advantage of both kinds of features is usually a better decision especially in prediction problems in which we care much more about information increase than causal inference. These three assumptions are based on previous empirical studies related to GPS data in social science research, which demonstrate the possibility of socioeconomic and behavioral information extraction, their usefulness for prediction and interpretation, and the possibility of predicting socioeconomic characteristics through behaviors (the obscure relationship between them).

The current use of GPS data in social science research can be summarized into three categories: 1) mapping human mobility patterns; 2) predicting socioeconomic or human behavioral characteristics; 3) and extracting human behavioral features from GPS data and use them for prediction or optimization.

Mobile phone data is the most widely used data in social sciences due to the popularization of smartphones in both developing and developed countries, and research based on this kind of data is usually highly relevant to population dynamics and mobility. Lu, Bengtsson, and Holme (2012) map the population movements over three months after the 2010 earthquake in Haiti solely relying on GPS data. Deville et al. (2014) demonstrate that, in Portugal and France, not only is the population mapping solely based on GPS data as accurate as that based on census, but it can measure population dynamics almost in real time. Jiang et al. (2016) adopt a more complex method for human mobility prediction. They first extract individual features from mobile phone GPS data and then build a mechanistic modeling framework to simulate individual daily mobility with fine resolution in Boston. In addition to the area of traditional social sciences, mobile phone GPS data has also been

used in public health for risk analysis. For example, three public health studies find that human aggregation or high human mobility can drive rubella transmission in Kenya (Wesolowski, Metcalf, et al., 2015), cholera transmission in Senegal (Finger et al., 2016) and dengue transmission in Pakistan (Wesolowski, Qureshi, et al., 2015).

Moreover, the combination of GPS data and mobile phone usage records or other totally disparate datasets are also used in socioeconomic and other human behavioral research. Blumenstock, Cadamuro, and On (2015) infer some socioeconomic features from mobile phone data and use them to reconstruct the distribution of wealth in Kigali. Pokhriyal and Jacques (2017) combine disparate data sources, including mobile phone data, environment data census data and poverty index, for improved poverty prediction and mapping in Senegal. It is worth noting that these studies simply implicate that GPS data embodies socioeconomic information, especially when it is combined with other existing data, rather than that we can use the predicted socioeconomic features to further predict accident risk. This is because a second-step prediction will lose or distort too much information. Therefore, at least for prediction problems, researchers should regard GPS data as latent socioeconomic information and put them into predictive models according to the social, political and economic layout of a city.

In addition to mobile phone GPS data, researchers also use the GPS data from vehicle telematics. However, most studies in this field heavily focus on the GPS data of public transportation and taxi rather than that of private cars. Therefore, the information from GPS data here reflects collective, rather than individual, socioeconomic and behavioral attributes. Besides, researchers in this field are more interested in optimization and pure prediction problems. They use human mobility patterns inferred from GPS data for traffic

planning (C. Chen et al., 2013; Z. Chen, Gong, & Xie, 2017; Kong et al., 2016; Rahmani, Jenelius, & Koutsopoulos, 2015; Tang, Liu, Wang, & Wang, 2015) and taxi service strategy (Xu, Zhou, Liu, Xu, & Zhao, 2015; Zhang et al., 2014), or combine GPS trajectory data with spatial socio-economic information, based on economic layouts of cities, to predict land-use (Pan, Qi, Wu, Zhang, & Li, 2013) or calibrate retail trading model (Yue et al., 2012).

3. The use of vehicle telematics data in accident analysis

Although the studies mentioned above have demonstrated the huge potential of GPS data for prediction if various information can be extracted from it, the current accident and insurance analysis has not taken enough advantage of them.

In traditional car insurance industry, the accident prediction is based on some variables usually reported by the drivers. They typically contain annual mileage, drive-related variables such as age, gender, occupation and income, and vehicle-related variables such as vehicle age, type and price (Jin et al., 2018). Thanks to the 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 some studies that try to extract behavioral features from vehicle telematics data and add them to conventional baseline models for improving prediction of accident risk. These features include the rates of hard accelerations or hard brakes (Bagdadi & Várhelyi, 2011; Handel et al., 2014; J. Paefgen, Staake, & Fleisch, 2014; Weidner, Transchel, & Weidner, 2017), strategic driving behaviors such as road and time selection (Tselentis, Yannis, & Vlahogianni, 2017), and daily driving behaviors and mobility patterns such as nighttime driving and familiarity with

driving routes (Ayuso, Guillén, & Pérez-Marín, 2014, 2016; Jin et al., 2018) [1]. These features have also proven to be, in these studies, more useful than many self-report features in insurance survey in terms of prediction since the features in survey usually cannot reflect the real driver/drivers' behaviors and even socio-demographic information.

It is obvious that previous studies in accident analysis neither use socioeconomic features extracted from GPS data nor frequently build machine learning models for prediction. In this paper, I will focus on socioeconomic information extraction and machine learning models based on traditional insurance features, extracted socioeconomic features and driving behavior features.

4. Methodology

4.1 Data

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. 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 variables related to claim history are indicators of accident and hence can be used as

[1] It is worth noting that these researchers are apt to use the annual mileage values in their GPS data even though they are also contained in their insurance data. This is because the self-reported values in insurance survey are usually lower than the actual values (White, 1976).

the response of whether a car has accidents and how severe the accidents are. The driver- 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 dataset to enrich the information for prediction.

4.2 Features

The features used in this paper contain three parts: socioeconomic information from the GPS data, driving behaviors/patterns from GPS data and driver- and vehicle-related characteristics in the insurance data.

4.2.1 Home address and information mining

For the socioeconomic information, I identify home address through GPS points and further mine useful features from it. A typical approach to this task is to assume that most GPS data points at night can represent the home address, and then to 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 several outliers. Second, a small percentage of car owners have more than one residential addresses. Due to these two factors, mean longitudes and latitudes will usually not be in the resident areas.

I design an algorithm to solve these problems, which is similar to k-mean method. 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 the C_1 and d_{i1}. The first mean center of the car displayed on the map, C_1, will be far away from the home address represented by the purple area. In order to obtain a slightly better mean center, the point that has the largest distance is removed and then the mean center and distances are recalculated. This process will repeat again and again until N_j is 2. As the figure shows, the center will gradually approximate to an area that has most points with small variance of distance. Besides, I also remove all 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 points or that their GPS points cannot represent resident address.

Figure 1: The process of the algorithm

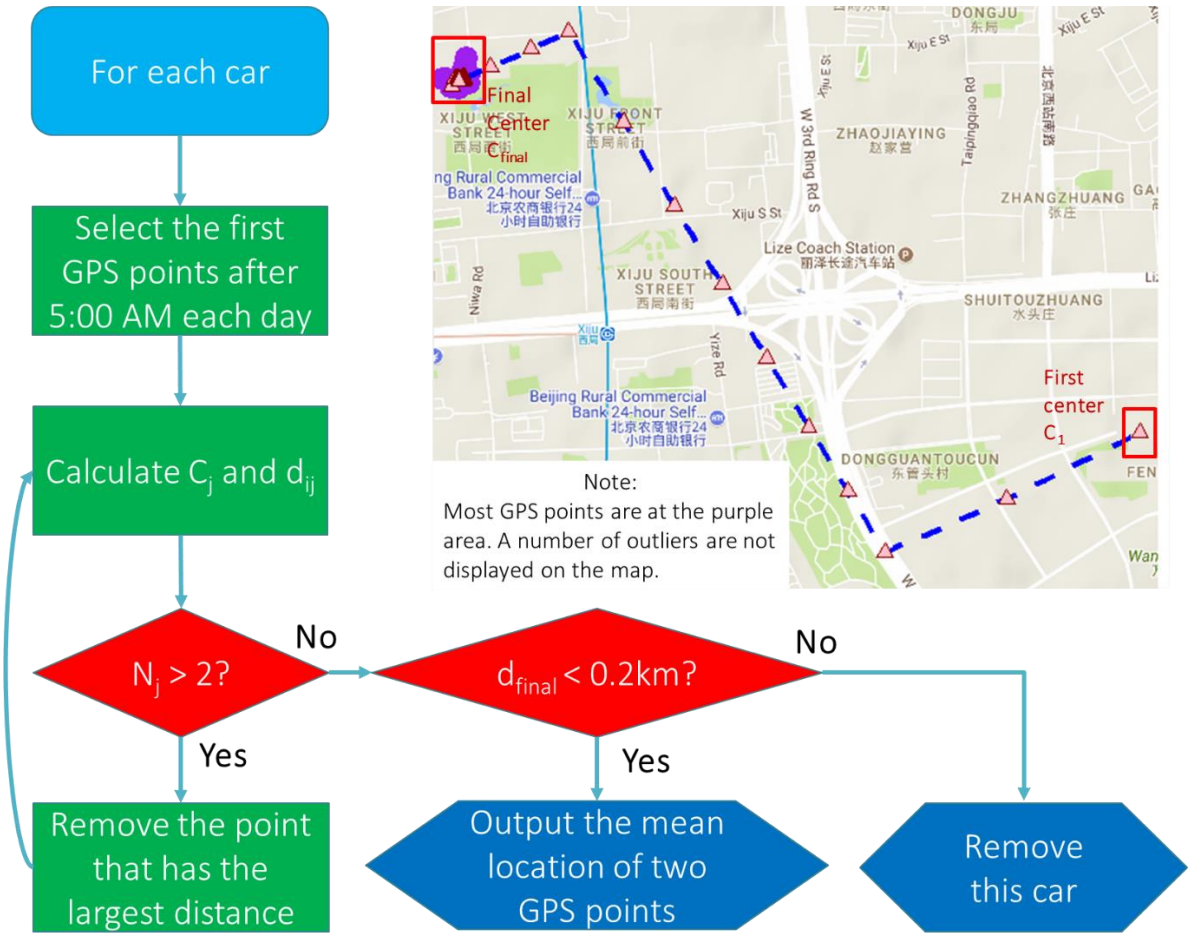


Figure 2: The distribution of home address

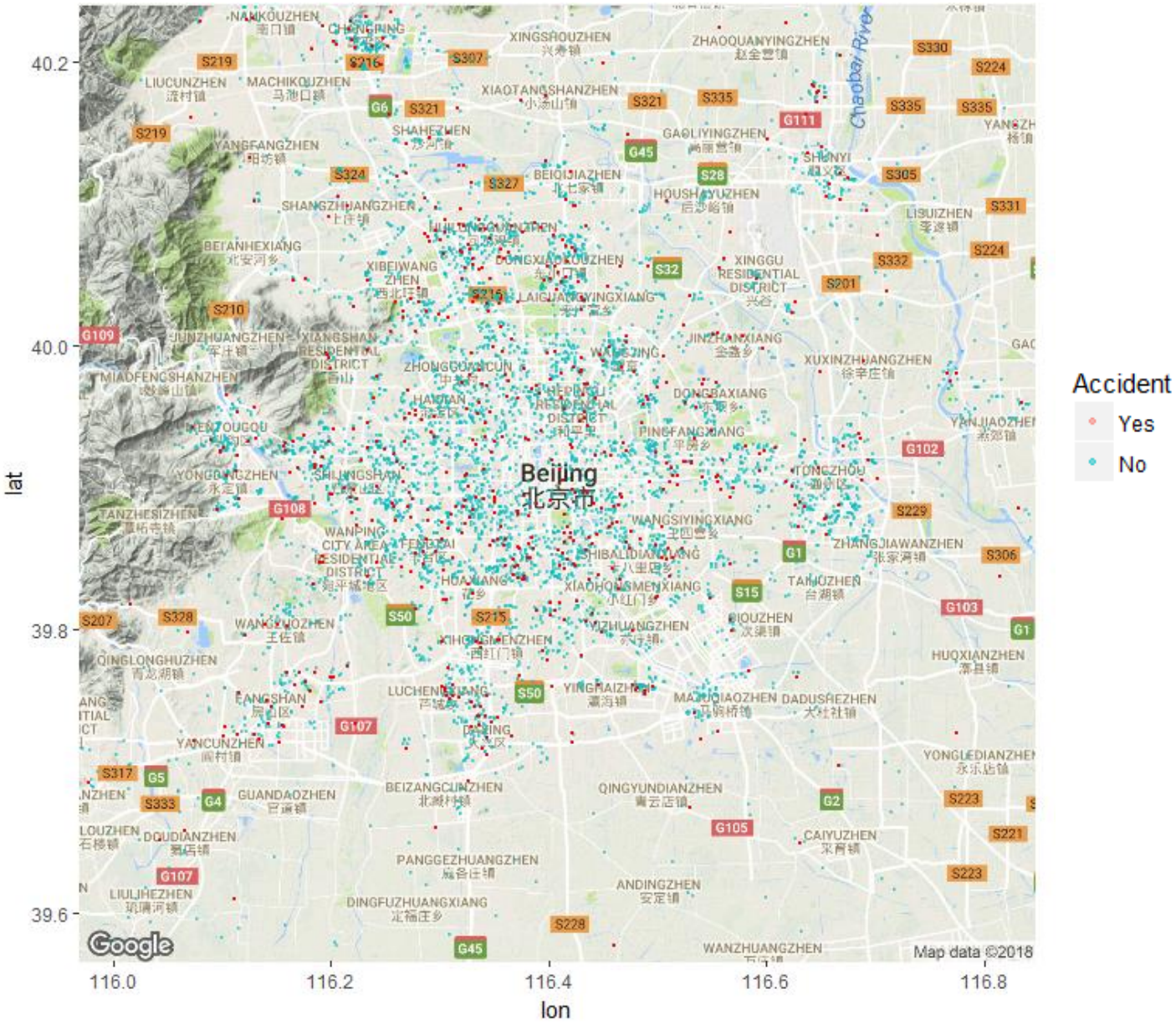
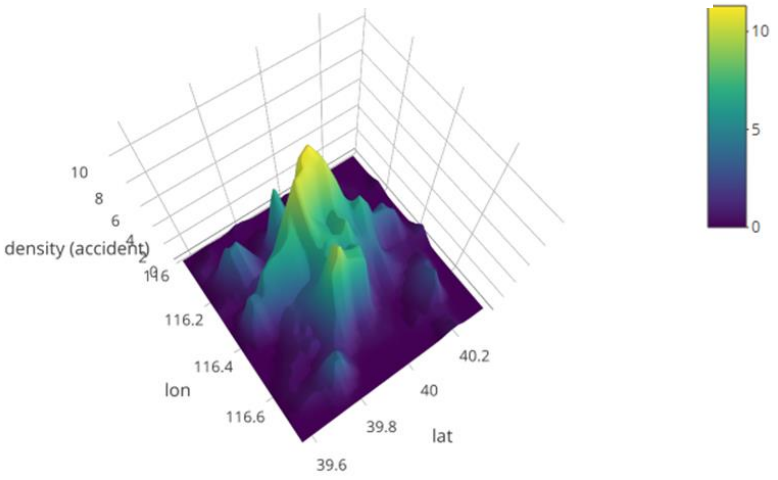
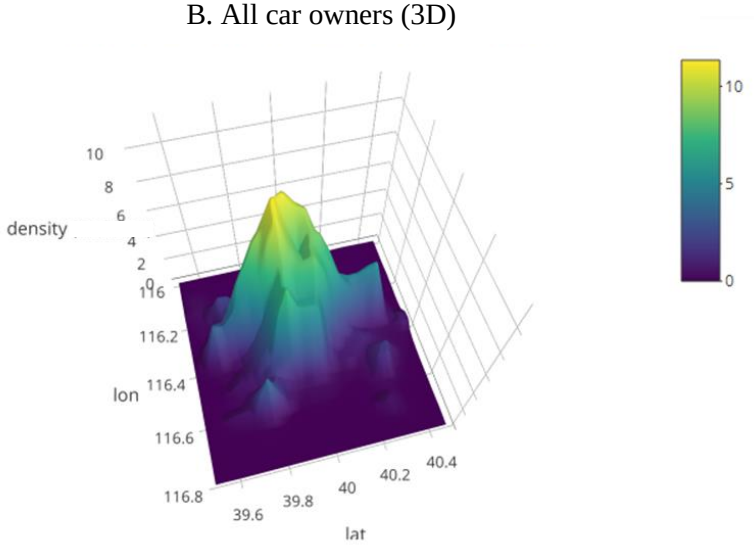


Figure 3: Kernel density estimation of the home address distribution

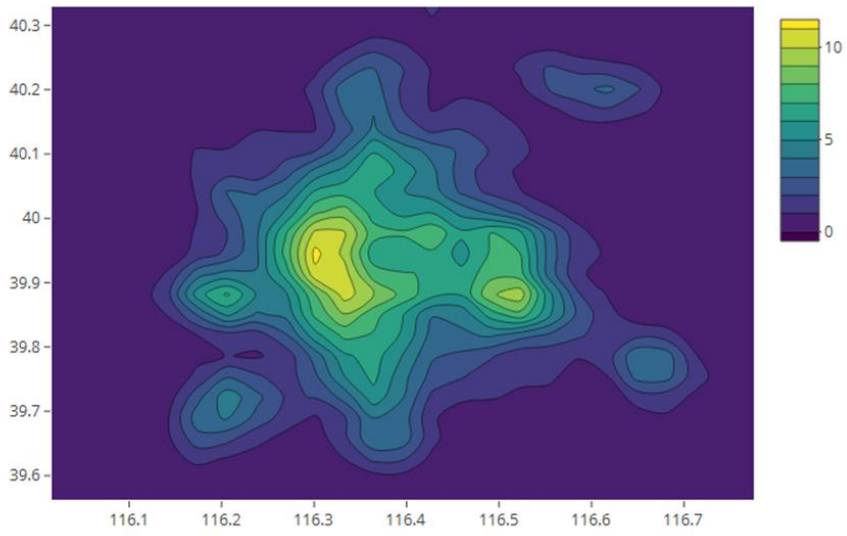
A. Car owners who report accidents (3D)



B. All car owners (3D)



C. Car owners who report accidents (2D)



D. All car owners (2D)

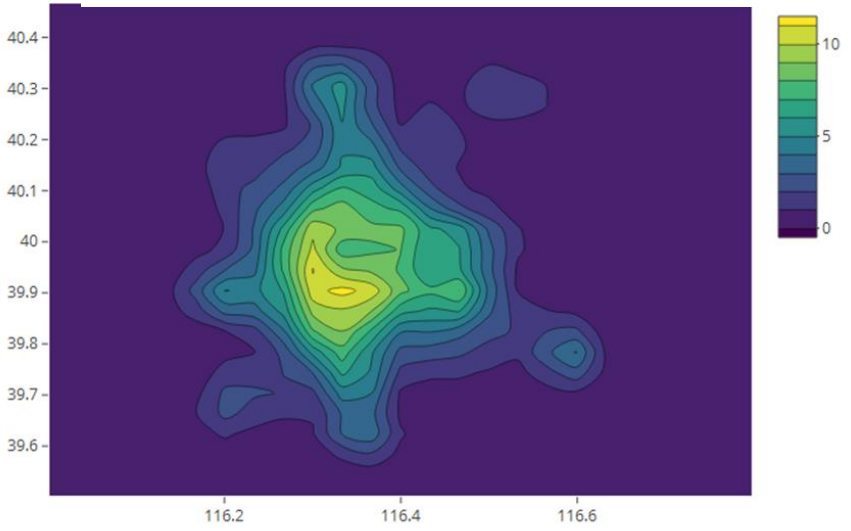


Figure 2 shows the result of the home address distribution. Actually, we cannot obtain enough useful information from this obscure figure. Therefore, I estimate the distribution densities by KDE method and plot them. In Figure 3, Facet A and C are the home address distribution density of those who report accidents and Facet B and D are all car owners'. I use the same vector of bandwidths for latitudes and longitudes when I estimate the densities of two groups, so that I can directly compare them. There is no doubt that the distributions of home address of these two groups are not identical, although they obviously have a general positive correlation. What we should care about is not the correlation since it is a common sense that a larger number of cars generally signifies a larger number of accidents in one area. On the contrary, the uncorrelation or the disproportion of some areas between two groups probably implies different probability of accident and we must pay more attention to.

Based on the differences of densities shown on the map as well as the socioeconomic space-time layout of Beijing, I speculate Haidian District, the south part of main urban area and the northern resident 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. Haidian District is the education and technology center of Beijing, the south part of main urban area is the relatively underdeveloped urban area comparing to other areas, and the northern resident area have 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 socioeconomic features I extracted from the GPS data is living in Haidian District, living in the south part of main urban area and living in the northern resident area.

4.2.2 Driving behaviors and patterns

The feature extraction of driving behaviors and patterns is much easier than that of socioeconomic characteristics because we need not to make strong assumption about car owners' socioeconomic characteristics and can directly obtain them based on the combination of space-time layout of Beijing and telematics/GPS information. I extract two sets of driving behaviors from the GPS data. The first set is pure driving behaviors such as acceleration and brake behaviors. The second set is social driving behaviors such as night-time driving and urban driving, which actually imply obscure latent socioeconomic information. Table 1 is the full list of features extracted from GPS data.

Table 1: Summary statistics of driving behaviors/patterns

Features	Definition	Mean	S.D.
Pure driving behavior extracted from GPS data			
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
Social driving behavior extracted from GPS data			
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_WKD	The fraction of mileage of driving during weekdays	0.556	0.152
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 average times that roads are traveled by.	3.289	1.479
PCT_SPEED	The fraction of mileage of driving with speed below 90 km/h	0.931	0.075

4.2.3 Features and responses in the insurance data

The insurance dataset contains two responses I will use for model building: claim and claim loss. While whether a driver made a claim is the indicator of self-report accident, the claim amounts can measure the losses caused by accidents. Besides, the dataset also contains two sets of features for baseline model building. The first set is driver-related characteristics,

which mostly contains drivers' demographic features. The second set is vehicle-related characteristics: car price, type and equipment. Table 2 shows the full list of features and responses in the insurance data.

Table 2: 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	Loss caused by accident	295.6	2650
Driver-related characteristics in the insurance dataset			
FEMALE	Drivers are female	0.283	0.450
YOUNG	<= 30 years old	0.233	0.423
OLD	>= 45 years old	0.284	0.443
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	Car was purchased in recent three years	0.195	0.396
BIG_CAR	Car belongs to Minivan/SUV	0.060	0.237
AIRBAG	Number of airbags	4.539	1.939
ALARM	Equipped with a life belt alarm	0.987	0.115

4.3 Models

4.3.1 Accident/claim classification

For accident classification, I choose 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, accident is a very complicated process based on a set of physical, socioeconomic and interactional contexts. The information contained in variables are unlikely to greatly improve the accident prediction. For example, although Paefgen, Staake, and Thiesse (2013) find the prediction accuracies can be higher 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 an serious overestimation of the majority class and thus the classifier bias. In other words, the prediction precision of the accident-involved class can be extremely low. 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. Besides, since Paefgen et al. (2013) also find that, in accident prediction problem with driving behaviors, the performances of decision tree and neural networks models are only slightly better than logistic regression models, 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.

4.3.2 Accident amount prediction

For the response of accident amount, the difficulties of insufficient information and predictive performance mentioned in accident classification cannot be completely solved in accident amount prediction. However, they will not become big problems in this task. First, compared to accident classification, the response of claim amount has a higher variance and thus provides more information for model building. Second, even though the improvement of amount prediction is slight, it is still possible to make pricing decision based on the predicted claim amount. Finally, I select four machine learning algorithms that are usually used in continuous response prediction. They are random forest, gradient boosting trees, lasso and elastic net.

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

5. Results

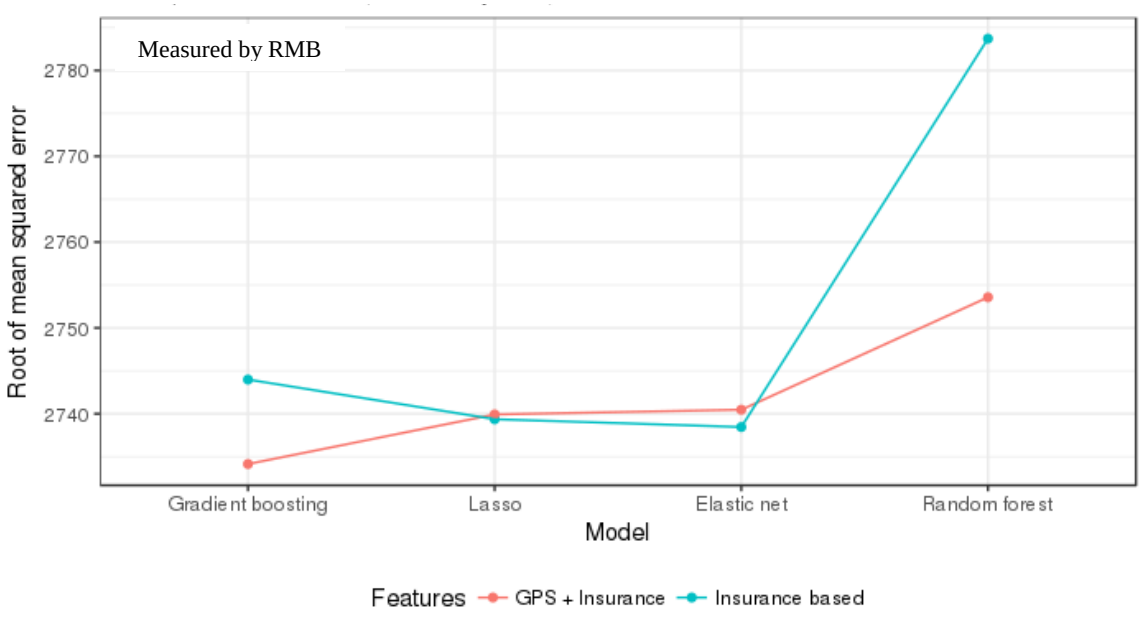
Table 3: Accident classification in binary logistic models

Variables	Definition	Models	
		Insurance	GPS
Driver-related characteristics in the insurance dataset			
FEMALE	Drivers are female	0.010	0.033
		(0.094)	(0.095)
YOUNG	<= 30 years old	0.076	0.098
		(0.105)	(0.107)
OLD	>= 45 years old	0.092	0.106
		(0.100)	(0.100)
STATE_JOB	Working for the state	-0.017	-0.010
		(0.090)	(0.091)
INTERNET_SALE	Buying insurance via internet	-0.321 **	-0.337 **
		(0.125)	(0.126)
ANNUAL_MIL	Annual driving mileage	0.160 ***	0.130 ***
		(0.025)	(0.028)
Vehicle-related characteristics in the insurance dataset			
CAR_PRICE	Car price	1.516e-07	-2.292e-09
		(4.3e-07)	(4.36e-07)
NEW_CAR	Car was purchased in recent three years	0.501 ***	0.498 ***
		(0.098)	(0.099)
BIG_CAR	Car belongs to Minivan/SUV	-0.187	-0.190
		(0.188)	(0.189)
AIRBAG	Number of airbags	0.003	0.005
		(0.028)	(0.028)
ALARM	Equipped with a life belt alarm	-0.0149	-0.140
		(0.349)	(0.354)
Home address extracted from GPS data			
HAIDIAN	Living at Haidian District		-0.413
			(0.279)
SOUTH	Living at the south of main urban area		0.278 ***
			(0.107)
NORTH	Living at the north resident area		0.265 *
			(0.139)
Pure driving behavior 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 ***
			(0.054)
Social driving behavior extracted from GPS data			
PCT_URBAN	The fraction of mileage of driving in the urban area		0.209
			(0.256)
PCT_FREEWAY	The fraction of mileage of driving on the freeways		-0.128
			(0.535)
PCT_LOCAL	The fraction of mileage of driving on the local roads		0.377
			(0.645)
PCT_WEEKEND	The fraction of mileage of driving during weekends		-0.360
			(0.382)
PCT_NIGHT	The fraction of mileage exposure of driving during nights		0.606 *
			(0.361)
MEAN_FMLRT	The average times that roads are traveled by.		-0.124 ***
			(0.036)
PCT_SPEED	The fraction of mileage of driving with speed below 90 km/h		-0.01
			(0.687)
Intercept		-2.476 ***	-2.327 **
		(0.359)	(0.915)
Log likelihood		-2039	-2015
AIC		4102	4078
BIC		4174	4231

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

Table 3 shows the logistic regression result[3]. Living at the south of main urban area, living at the north resident area, hard brake behavior, night driving behavior and familiarity with roads 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, we should also look at the overall model fit. I compare the baseline model with the GPS-related model by likelihood ratio test. The statistic is significant at 0.1% level. Besides, the value of Akaike information criterion (AIC) in the GPS-related model is also smaller than that in the baseline model. These two results indicate that the information extracted from the GPS data improves the model fit and thus are very likely to improve the accident prediction.

Figure 4: Performance of four models with different features



[3] Since the logistic models are mainly used for prediction purpose, I select significant level at 10% rather than at 5%.

Figure 4 displays the cross-validation results of claim amount prediction. Among four algorithms, the two models based on random forest have very high MSEs, although adding GPS-related features significantly reduce it. 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 into the insurance-based model, but also help me obtain the lowest MSE. 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 we choose an appropriate model.

However, we should also ask why and how they improve it. Figure 5 and Table 4 show the feature importance of four GPS + Insurance models. In tree-based models, behavior features have very large influences. Especially in the best performed gradient boosting model, nine of top ten features are driving behaviors, eight of them are extracted from the GPS data. The random forest model also relies on behavior features but gives to much importance to only two features. This might be the cause of 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 algorithm will not remove features from the model and thus rely on both sets of features. However, since lasso and elastic net have similar shrinkage processes, we can speculate the latter still relies more on the features in the insurance data. The difference of feature importance between tree-based models and 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 linear-based lasso and elastic net models heavily shrink the coefficients of behavior features and thus cannot improve claim amount prediction.

Figure 5: Most important features in tree-based models

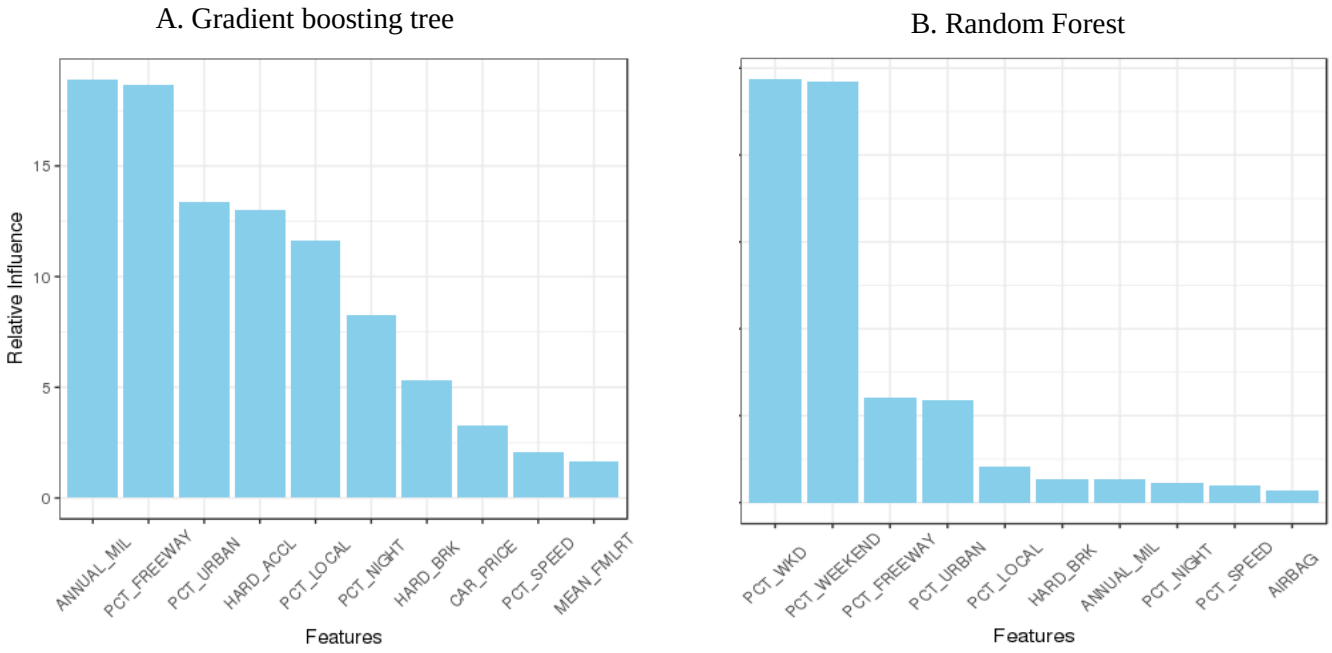


Table 4: Coefficients in lasso and elastic net models

Features	Lasso	Elastic net
(Intercept)	104.40640	145.9751410
FEMALE	0.00000	-40.9657286
YOUNG	63.29623	82.5514520
OLD	0.00000	20.2865867
STATE_JOB	0.00000	-6.7003830
INTERNET_SALE	-53.05628	-78.2793726
ANNUAL_MIL	78.79185	44.9593110
CAR_PRICE	0.00000	0.0000878
BIG_CAR	102.20539	105.3582429
Bigcar	0.00000	-54.5922592
AIRBAG	0.00000	5.4798248
ALARM	0.00000	56.9537935
HARD_ACCL	0.00000	12.4458047
HARD_BRK	0.00000	1.7914637
PCT_URBAN	0.00000	7.9627252
PCT_FREEWAY	0.00000	-135.3897806
PCT_LOCAL	0.00000	-227.9265286
PCT_WKD	0.00000	81.4530760
PCT_WEEKEND	0.00000	-81.3322563
PCT_NIGHT	0.00000	227.2951522
PCT_SPEED	0.00000	89.2196131
MEAN_FMLRT	0.00000	-12.5385013
South	0.00000	-22.8120987
Haidian	0.00000	-100.2764192
North_residence	0.00000	-22.1062539

It is worth noting that the socioeconomic features extracted from the GPS data are less important than the behavior features in the best gradient boosting tree model. However, even though the socioeconomic features are usually less important, it still has its distinctive advantage. Actually, when we consider more practical factors such as cost and accessibility of GPS data, we can find the socioeconomic features extracted from the GPS data may be more useful and feasible in many cases. Since socioeconomic features are more stable features compared to behavior features, we can obtain them from a set of GPS data and use them in a broad range of studies even though we are not able to get more GPS data. On the contrary, behavior features are more ephemeral, especially in accident analysis. The data that contains it are expensive, sensitive and sometimes inaccessible.

6. Conclusion and discussion

In this paper, I use a set of methods to extract socioeconomic features and driving behavior features from a GPS dataset and test their role in several models. There are three implications that can be obtained from this study.

First, it is possible to accurately extract useful socioeconomic information from GPS data for predictive purpose. However, the difficulty in this step is not only the design of an appropriate algorithm that can compute an accurate result, but the interpretation of GPS points based on space-time layout of a city before algorithm design. I provide an example of how we can identify car owners' home address and speculate some areas that might provide information for accident prediction. The logistic regression results also demonstrate the usefulness of this method.

Second, the combination of socioeconomic and driving behavior features extracted from GPS data can improve the prediction of both accident and/or its loss. At least in my study, they can increase the log likelihood in the logistic model and lower the mean squared error via gradient boosting model. How much contribution these features can make also depends on what predictive models we build. My result shows the relationship between driving behaviors and claim amount are very likely to be non-linear so that adding GPS-related features into linear-based models cannot make a difference.

Third, driving behavior features seem more important than socioeconomic characteristics in accident prediction. However, they have their own distinctive advantages and disadvantages. For example, driving behavior features are actually expansive and sometimes inaccessible. Therefore, extracting stable socioeconomic features from GPS data, such as home address and the socioeconomic characteristics embodied in it, can be helpful even though we cannot get more GPS data in the future.

Reference

Ayuso, M., Guillén, M., & Pérez-Marín, A. M. (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, M., Guillén, M., & Pérez-Marín, A. M. (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, O., & Várhelyi, A. (2011). Jerky driving—An indicator of accident proneness? *Accident Analysis & Prevention, 43*(4), 1359 - 1363.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting Poverty and Wealth from Mobile Phone Metadata. *Science, 350*(6264), 1073 - 1076.

Chen, C., Zhang, D., Zhou, Z.-H., Li, N., Atmaca, T., & Li, S. (2013). *B-Planner: Night bus route planning using large-scale taxi GPS traces*. Paper presented at the IEEE International Conference on Pervasive Computing and Communications (PerCom), San Diego.

Chen, Z., Gong, X., & Xie, Z. (2017). An Analysis of Movement Patterns between Zones Using Taxi GPS Data. *Transactions in GIS, 21*(6), 1341 – 1363.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., . . . Tatem, A. J. (2014). Dynamic Population Mapping Using Mobile Phone Data. *PNAS, 111*(45), 15888 - 15893.

Finger, F., Genolet, T., Mari, L., Magny, G. C. d., Mang, N. M., Rinaldo, A., & Bertuzzo, a. E. (2016). Mobile Phone Data Highlights The Role of Mass Gatherings in the Spreading of Cholera Outbreaks. *PNAS, 113*(23), 6421 – 6426.

Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J., & Ohlsson, M. (2014). Insurance Telematics: Opportunities and Challenges with the Smartphone Solution. *IEEE Intelligent Transportation Systems Magazine, 6*, 57 - 70.

Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., & González, M. C. (2016). The TimeGeo Modeling Framework for Urban Mobility without Travel Surveys. *PNAS, Published online*, E5370–E5378.

Jin, W., Deng, Y., & Jiang, H. (2018). Latent Class Analysis of Accident Risks in Usage-Based Insurance: Evidence from Beijing. *Accident Analysis & Prevention, 115*, 79-88.

Karapiperis, D., Birnbaum, B., Brandenburg, A., Castagna, S., Greenberg, A., Harbage, R., & Obersteadt, A. (Eds.). (2015). *Usage-Based Insurance and Vehicle Telematics: Insurance Market and Regulatory Implications*.

Kong, X., Xu, Z., Shen, G., Wang, J., Yang, Q., & Zhang, B. (2016). Urban Traffic Congestion Estimation and Prediction Based on Floating Car Trajectory Data. *Future Generation Computer Systems, 61*, 97 - 107.

Lu, X., Bengtsson, L., & Holme, a. P. (2012). Predictability of Population Displacement After the 2010 Haiti Earthquake. *PNAS, 109*(29), 11576 – 11581.

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

Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate Exposure Modeling of Accident Risk: Insights from Pay-as-You-Drive Insurance Data. *Transportation Research Part A, 61*, 27 - 40.

Pan, G., Qi, G., Wu, Z., Zhang, D., & Li, S. (2013). Land-Use Classification Using Taxi GPS Traces. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 14*(1), 113 - 123.

Pokhriyal, N., & Jacques, D. C. (2017). Combining Disparate Data Sources for Improved Poverty Prediction and Mapping. *PNAS, Publish online October 31, 2017*, E9783–E9792.

Rahmani, M., Jenelius, E., & Koutsopoulos, H. N. (2015). Non-Parametric Estimation of Route Travel Time Distributions from Low-Frequency Floating Car Data. *Transportation Research Part C: Emerging Technologies, 58*, 343 - 362.

Tang, J., Liu, F., Wang, Y., & Wang, H. (2015). Uncovering Urban Human Mobility from Large Scale Taxi GPS Data. *Physica A: Statistical Mechanics and its Applications, 438*, 140 - 153.

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

Weidner, W., Transchel, F. W. G., & Weidner, R. (2017). Telematic Driving Profile Classification in Car Insurance Pricing. *Annals of Actuarial Science, 11*(2), 213 - 236.

Wesolowski, A., Metcalf, C. J. E., Nathan Eaglea, g., Janeth Kombich, Grenfell, B. T., Bjørnstad, O. N., Lessler, J., . . . Buckee, C. O. (2015). Quantifying Seasonal Population Fluxes Driving Rubella Transmission Dynamics Using Mobile Phone Data. *PNAS, 112*(35), 11114 – 11119.

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., . . . Buckee, C. O. (2015). Impact of Human Mobility on the Emergence of Dengue Epidemics in Pakistan. *PNAS, 112*(38), 11887 – 11892.

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

Xu, X., Zhou, J., Liu, Y., Xu, Z., & Zhao, X. (2015). Taxi-RS: Taxi-Hunting Recommendation System Based on Taxi GPS Data. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 16*(4), 1716 - 1727.

Yue, Y., Wang, H.-d., Hu, B., Li, Q.-q., Li, Y.-g., & Yeh, A. G. O. (2012). Exploratory Calibration of a Spatial Interaction Model Using Taxi GPS Trajectories. *Computers, Environment and Urban Systems, 36*(2), 140-153.

Zhang, D., Sun, L., Li, B., Chen, C., Pan, G., Li, S., & Wu, Z. (2014). Understanding Taxi Service Strategies From Taxi GPS Traces. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 16*(1), 123 - 135.