VEHICLE CATEGORY CLASSIFICATION BASED ON GPS TRAJECTORY DATA

A Thesis

by

RUIHONG WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Nicholas G. Duffield
Committee Members,	Mark W. Burris
	Srinivas Shakkottai
Head of Department,	Miroslav Begovic

August 2020

Major Subject: Computer Engineering

ABSTRACT

Understanding the category of a vehicle is an essential study for transportation safety and operation. With the explosive number of GPS devices, there are massive vehicle GPS trajectory data sets whose sizes are beyond the traditional trajectory analysis method's capability. This study utilizes Apache Spark™ to build up a framework whose output data can be compatible with machine learning algorithms for vehicle category classification. Five types of features were extracted from the GPS trajectory data, namely driving habits statistics, trajectory sample quality statistics, geographical information statistics, origin and destination cluster statistics, and temporal statistics. The spatial clustering algorithm and spatial join are incorporated in the workflow, significantly broadening the number of features for the training data set. The results show that the five types of statistics extracted from the trajectory are adequate for distinguishing different vehicle categories by machine learning algorithms. The same accuracy rank sequence for the vehicle classes was observed across different types of features and algorithms, and the decision tree ensemble algorithms have better performance over the logistic regression and support vector machine algorithms.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Duffield, my committee members, Dr. Burris, Dr. Shakkottai, and Dr. Das, my mentor at Texas A&M Transportation Institute, and Mr. Turner, who helped me in acquiring the data, for their guidance and support throughout the course of this research.

Thanks also go to the technical support in HPRC, especially to Dr. Jian for his instruction on how to use spark engine on HPRC.

Finally, thanks to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis (or) dissertation committee consisting of Dr. Nicholas G. Duffield, and Dr. Srinivas Shakkottai from the Department of Electrical & Computer Engineering and Professor Mark W. Burris of the Department of Civil & Environmental Engineering. The Maryland GPS trajectory data was provided by Dr. Subasish Das from Texas A&M Transportation Institute. All other work conducted for the thesis (or) dissertation was completed by the student independently.

Funding Sources

This research was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology.

NOMENCLATURE

C	Consumer Vehicles
FD	Field Service / Local Delivery Fleets
T	For Hire / Private Trucking Fleets
TST	Taxi / Shuttle / Town Car Service Fleets
RF	Random Forest
SVM	Support Vector Machine
XGBoost	eXtreme Gradient Boosting
FHWA	Federal Highway Administration

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGEMENTS iii

CONTRIBUTORS AND FUNDING SOURCES iv

NOMENCLATURE v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES x

1. INTRODUCTION 1

2. LITERATURE REVIEW 3

3. DATA DESCRIPTION 5

3.1. INRIX® Maryland GPS trajectory and trip data 5
3.2. Roadway shapefile data 7

4. DATA PREPARATION 9

4.1. Driving habit statistics 12
4.1.1. Attributes description 12
4.1.2. Attributes distribution analysis 14
4.2. Trajectory sample quality statistics 17
4.2.1. Attributes description 17
4.2.2. Attributes distribution analysis 18
4.3. Geographical information statistics 20
4.3.1. Attributes description 21
4.3.2. Attributes distribution analysis 22
4.4. Origin and destination cluster statistics 25
4.4.1. Attributes description 27
4.4.2. Attributes distribution analysis 28
4.5. Temporal statistics 29

Page

4.5.1. Attributes description 30
4.5.2. Attributes distribution analysis 31

5. METHODOLOGY 34

5.1. Multinomial logistic regression 34
5.2. Support vector machine (SVM) 35
5.3. Random forest (RF) 36
5.4. XGBoost 36

6. EVALUATION 38

6.1. Classification for 4 classes 39
6.1.1. Multiclass classification baseline (four classes) 39
6.1.2. Improvement of baseline by other statistics individually 41
6.1.3. Study for highly compound models 43
6.2. Classification for 3 classes 46

7. LIMITATION AND FUTURE STUDY 50

8. CONCLUSION 51

REFERENCES 52

APPENDIX A PERFORMANCE DETAILS FOR THE MODELS MENTIONED IN THE THESIS 55

LIST OF FIGURES

Page

Figure 1 Device and trip number pie charts. 6

Figure 2 Road network in Maryland. 8

Figure 3 Data preparation procedure flow chart. 11

Figure 4 Gaussian density distribution for the numeric driving habit statistics (all vehicles, outliers removed). 15

Figure 5 Gaussian density distribution for the numeric driving habit statistics (vehicles with trip number > 10, outliers removed). 16

Figure 6 Gaussian density distribution for the trajectory sample quality statistics (all vehicles, outliers removed). 19

Figure 7 Gaussian density distribution for the trajectory sample quality statistics (vehicles with trip number > 10, outlier removed). 20

Figure 8 Common positions of GPS points on the roadway. 21

Figure 9 Gaussian density distribution for the geographical information statistics (all vehicles, outlier removed). 23

Figure 10 Gaussian density distribution for the geographical information statistics (vehicles with trip number > 10, outlier removed). 24

Figure 11 Origin destination plot (cluster size > 3). The triangles represent origins, and the plus symbols represent destinations. 26

Figure 12 Gaussian density distribution for the numeric OD cluster Statistics (vehicles with trip number > 10, outlier removed). 29

Figure 13 Gaussian density distribution for the temporal statistics (all vehicles, outlier removed). 32

Figure 14 Gaussian density distribution for the temporal statistics (vehicles with trip number > 10, outlier removed). 33

Figure 15 Micro average precision of test data set for baseline. 40

Page

Figure 16 Micro average accuracy of test data set for models with different combination of statistics. 42

Figure 17 Average F1 score of test data set across the algorithms for the four classes. .. 43

Figure 18 Micro accuracy of test data set for different combinations of statistics 44

Figure 19 F1 score across four vehicle classes for logistic regression and XGBoost..... 45

Figure 20 Macro average precision for different features and training data sizes 47

LIST OF TABLES

Page

Table 1 The test-set performance details for baseline models by vehicle category. 40

Table 2 Test set performance details for the 4-class model with all features (1948 trained, 487 tested).....48

Table 3 Test set performance details for the 3-class model with all features (57619 trained, 14405 tested).....49

Table 4 4-class classification, baseline (1948 trained, 487 tested). 55

Table 5 4-class classification, baseline + OD cluster statistics (1948 trained, 487 tested).....56

Table 6 4-class classification, baseline + trajectory sample quality statistics (1948 trained, 487 tested).....57

Table 7 4-class classification, baseline + temporal statistics (1948 trained, 487 tested). 58

Table 8 4-class classification, baseline + geographical information statistics (1948 trained, 487 tested).....59

Table 9 4-class classification, baseline + OD cluster + temporal + geographical information statistics (1948 trained, 487 tested).....60

Table 10 4-class classification, baseline + trajectory sample quality + OD + temporal + geographical information statistics (1948 trained, 487 tested).....61

Table 11 4-class classification, baseline + OD cluster + temporal + geographical information statistics and corresponding standard deviation (1948 trained, 487 tested).....62

Table 12 4-class classification, baseline +trajectory sample quality + OD cluster + temporal + geographical information statistics and corresponding standard deviation (1948 trained, 487 tested). 63

Table 13 3-class classification, baseline (57619 trained, 14405 tested). 64

Page

Table 14 3-class classification, baseline + OD cluster + temporal + geospatial information statistics and corresponding standard deviation (57619 trained, 14405 tested).....65

Table 15 3-class classification, baseline + trajectory sample quality + OD cluster + temporal + geospatial information and corresponding standard deviation (57619 trained, 14405 tested) 66

1. INTRODUCTION

Global Positioning System (GPS) trajectory analysis is a crucial study area within transportation as it can help manage and predict traffic flow. The trajectory data is composed of geographical locations (latitudes and longitudes) of a vehicle or human being with timestamps in sequence. The trajectory analysis used to be hard to conduct and mainly limited to a small data set because the trajectory can only be acquired through travel surveys. However, with the prevalence of GPS devices, many large GPS trajectory data sets are available for researches. Different categories of vehicles have a strong correlation with their vehicle weights, which will play an important role in transportation safety and operation. Besides, the prediction of a vehicle category through its trajectory has significant meaning for the data cleaning process for the GPS trajectory data. This study is mainly focusing on vehicle categories (vehicle types) classification based on the Maryland trajectory data. The trajectory data was collected by INRIX®, a company providing data analytics services to different public and private organizations. The data set contains vehicle type labels and is sufficient for the machine learning algorithm.

Since the raw GPS points cannot be applied directly to any machine learning method, the data preparation measures must be done in advance. There has already been some good trajectory data management study on extracting the Geographical information from the trajectory *(1-3)* and integrate heterogeneous data sources *(4)*. However, the INRIX®

trajectory data preparation requires efficient methods to process the data rapidly. Apache Spark™ is a powerful tool to handle big data analysis, with 10 to 100 times faster than MapReduce *(5).* Spark also provides SparkSQL *(6)* for developers to handle the large distributed relational database by high speed. This research will take advantage of Apache Spark™ in the process of data integration and aggregation. Some geographical attributes will be assigned to the GPS trajectory points through the spark engine *(7-8).* This study will mainly discuss the supervised learning method to distinguish the vehicle category by the vehicle trajectory data. This study will test several different popular models with combinations of features like logistic regression, support vector machine (SVM), and decision tree ensemble algorithms.

The study's main contributions are:

- The creation of a big data framework used to extract features from GPS trajectories that can be analyzed by machine learning algorithms.
- The creation of a big-data framework to join geographical information to GPS trajectory points.
- The comparison of several machine learning algorithms' ability to correctly classify vehicles based on geographical information and GPS trajectory points.

2. LITERATURE REVIEW

GPS trajectory pattern recognition covers a wide range of topics. Boukhechba M *(9)* clustered the human movement by the semantic location into different human activities. Siła-Nowicka K *(10)* introduced a framework to analyze the human mobility pattern from dynamic and static behavior. The result shows excellent performance in the stops and movement mode recognition contrast to the GPS survey. Other well studied GPS trajectory pattern recognition tasks include traffic rules detection *(11),* periodic travel pattern *(12),* and Disorientation detection *(13).* Nevertheless, the pattern recognition based on the vehicle category has not been widely investigated.

Many data cleaning and refinement methods have been proposed due to the discrepancy of sample frequency and accuracy by different GPS devices. Zair S *(14)* defined criterion with a particle filter to distinguishing the outliers to the inliers for GPS positions. The field study shows its robustness to the non-Gaussian noise. Patil V *(15)* brought up a secured method called ‘GeoSClean’ to detect the abnormal GPS point according to their distance, velocity, and acceleration.

The matching of GPS points to geographical information is another heated research area for trajectory analysis. Yin L *(16)* introduced an algorithm called ST-Matching, which matches a sequence of trajectory points with the road segments. The algorithm analyzes the road network structure and the spatial-temporal restrictions to select the right path

from all the candidates' segments. The algorithm outperformed the state of the art on both the synthetic data and real data. Nasri A and Fan J *(2-3)* estimated the vehicle mile traveled (VMT) for Maryland state based on the INRIX® Maryland GPS trajectory data. A method joining the trajectory data to the roadway data by spatial join is suggested in this research.

Machine learning has long dominated by the pattern recognition field since it was first brought up. There are varieties of models available to the researchers, such as logistic regression *(17)*, neural network, support vector machines *(18)*, decision trees (Random Forest) *(19)*, and boosting algorithms *(20)*.

The tree ensembled algorithms were tested to have outstanding performance over high dimensional data and were not sensitive to the data scale *(20-21)*. This research is not the first study to apply the machine learning algorithm into the trajectory data analysis. De Vries *(22)* conducted cluster and classification algorithms to the vessel trajectory data. Researchers first compressed the trajectory data by a segmentation method without detriment of its results. Cho S B. recognizes the user location by combining K-nearest neighbor and decision trees *(23)*. The model was tested on the real sampled data, achieving high accuracy.

3. DATA DESCRIPTION

There are three data sets utilized in this research. They are INRIX® Maryland trip data, INRIX® Maryland GPS trajectory data, and Maryland Roadway shapefile data. Our classification algorithms will be built on the first two data sets, and the latter two are the auxiliary data to broaden the model attributes.

3.1. INRIX® Maryland GPS trajectory and trip data

The data covers 1.5%-2% of the trips in Maryland during February, June, July and October of 2015 *(7)*. The GPS trajectory data contains 1,376,720,203 GPS points. Each GPS point contains a unique ID denoting which trip it belongs to. The GPS trajectory data is composed of locations, timestamps, and its sequence in the trip. There are 19,690,402 trips in the Maryland trip data. This data contains information about the pseudo location of the Origin and Destination (OD) as well as the timestamps for each trip. Due to the privacy policy, the ODs in the trip data have been insignificantly relocated. However, the distances between the true locations and the dummy locations of the ODs are usually 50 to 200 meters, which is acceptable for this study. Similar to the GPS trajectory data set, each trip contains a unique ID connecting the trip to a vehicle GPS device. For each trip, there is an attribute called ‘Provider Driving Profile’ which classifies the vehicles into four main categories, namely Consumer Vehicle (C), Field Service / Local Delivery Fleets (FD), For Hire / Private Trucking Fleets (T) and Taxi / Shuttle / Town Car Service Fleets (TST). The INRIX® collected this attribute from 148

data providers. Each data provider only supplied the trajectories of one vehicle category. In this research, the 'Provider Driving Profile' was considered as a credible attribute collected by the INRIX® and was set as the ground truth for the vehicle category classification. For the efficiency of reading, the paper will use C, FD, T, and TST to replace the four classes.

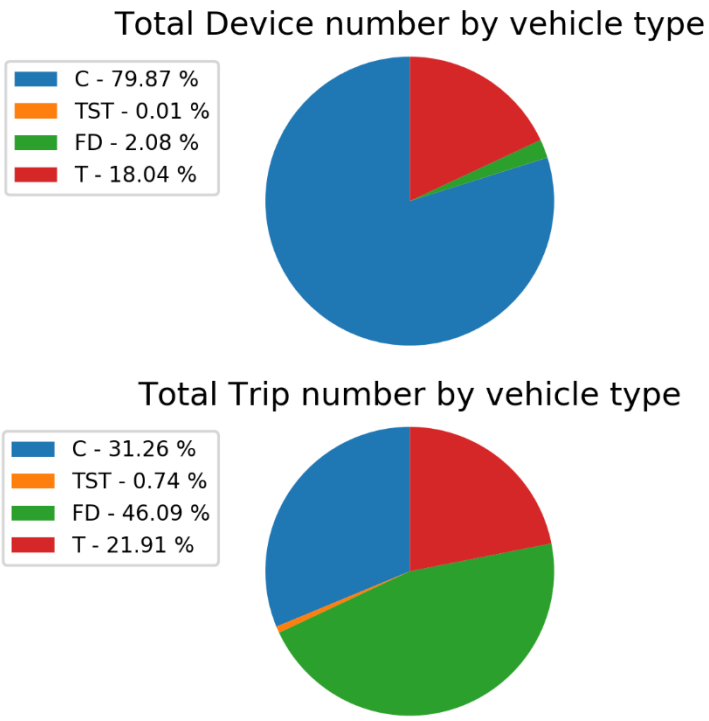

Figure 1 Device and trip number pie charts.

There are 5,451,095 devices through the whole data set. According to Figure 1, Class FD tends to have more trips for each vehicle. Class C has fewer trips than others.

The features for machine learning algorithm will be summarized from the GPS point level to the device level. Due to the high volume of the data, the Apache Spark™ is used to do the data aggregation, as well as the roadway shapefile join.

3.2. Roadway shapefile data

The Roadway shapefile for Maryland, shown in Figure 2, is downloaded and utilized as our complement data set. The data set covers 124,509 road segments for Maryland state, including the Washington D.C. The road functional classes of the roadways were joined to the GPS trajectory data through Apache Spark™. The road functional class is a standard defined by the United States Federal Highway Administration (FHWA), including three main functional classes, namely arterial, collector, and local. The values for the functional class in Maryland roadway shapefile are Local, Collector, Principal Arterial and Minor Arterial. There are some vacant fields in the Maryland shapefile data, which is replaced by ‘None’.

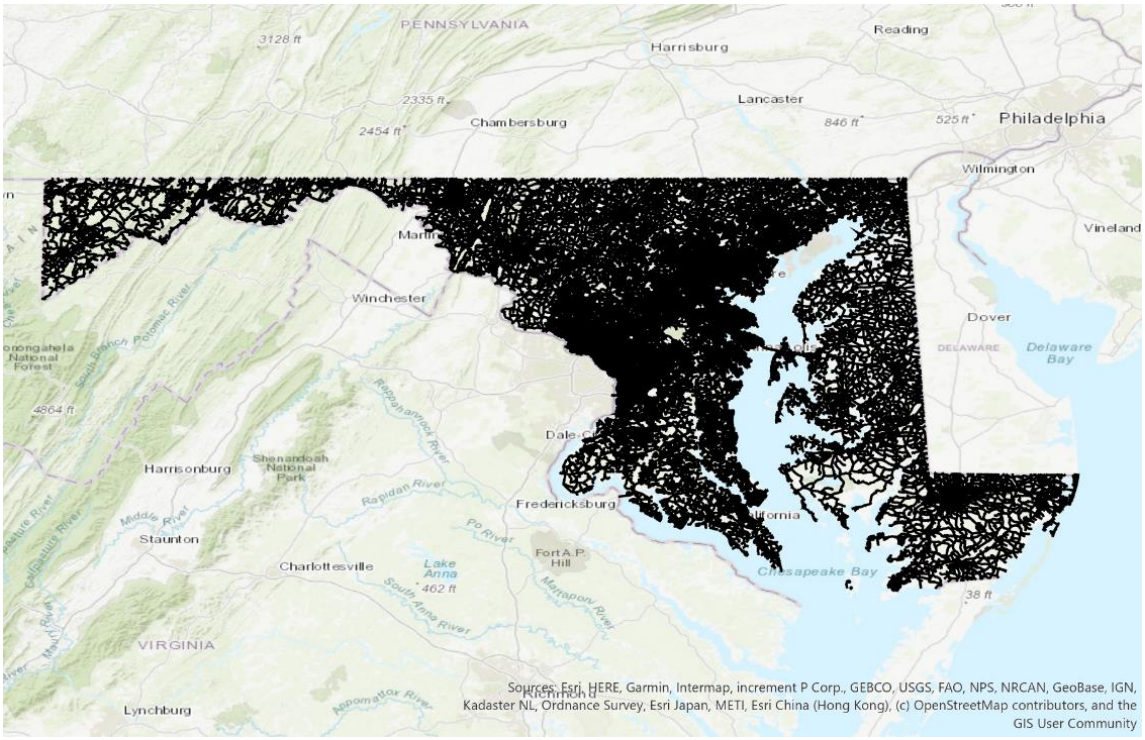

Figure 2 Road network in Maryland.

4. DATA PREPARATION

In order to utilizing machine learning algorithm, some features were collected from the raw data set. In this study, the researcher summarized some vehicle trajectory statistics from GPS trajectory data. Here, we separated the statistics into five types:

- Driving habit statistics
- Trajectory sample quality statistics
- Geographical information statistics
- Origin and destination cluster statistics
- Temporal statistics

Our method to prepare the data with a large amount rapidly was by utilizing the distributed computing schema Pyspark. The raw trajectory data was first read as the Spark Dataframe, on which SparkSQL command could be applied. The whole procedure could be separated into two levels (trajectory and trip).

As mentioned in Figure 3, driving habit statistics, and trajectory sample quality were prepared in the trajectory level. The distance and time difference between the consecutive GPS points were calculated by the window function, which are essential components for the trajectory sample quality and driving habit statistics. On the other hand, the extracted features can go beyond the trajectory itself. The geographical information around the GPS points could be useful for our vehicle types inference. For

example, the field service and local delivery fleets (FD) tend to travel on the local roads, while the truck fleets tends to travel on the arterial roads across the cities. In this study, the road function class of the nearest road segment for each GPS point was joined to the GPS trajectory data.

The information above were aggregated through different methods by the trip IDs. The numerical values were mostly collected by the mean function, while the category values were aggregated by the most frequent value or the value proportion to all the items in the trip.

In the trip level, the data set aggregated from the procedure above will first join the trip data. The OD locations and middle time of the trips can be clustered into different groups. Many of the vehicles may not have enough trips to support the OD cluster algorithm, so the researcher filtered the vehicles which have trips less than 10. The final usable vehicle records number is 140,952.

Let $\text{Traj}_{vt} = (p_0^{vt}, p_1^{vt}, \dots, p_{N_{vt}}^{vt})$ denotes the GPS trajectory points for vehicle v (v = 1, …, N) and its trip t (t = 1,…, N_v), where $p_i^{vt} = (x, y, T)$ is the i^{th} (i = 1,…, N_{vt}) GPS points for vehicle v's trip t, $T_{p_i^{vt}}$ represent the timestamp for GPS point p_i^{vt}.

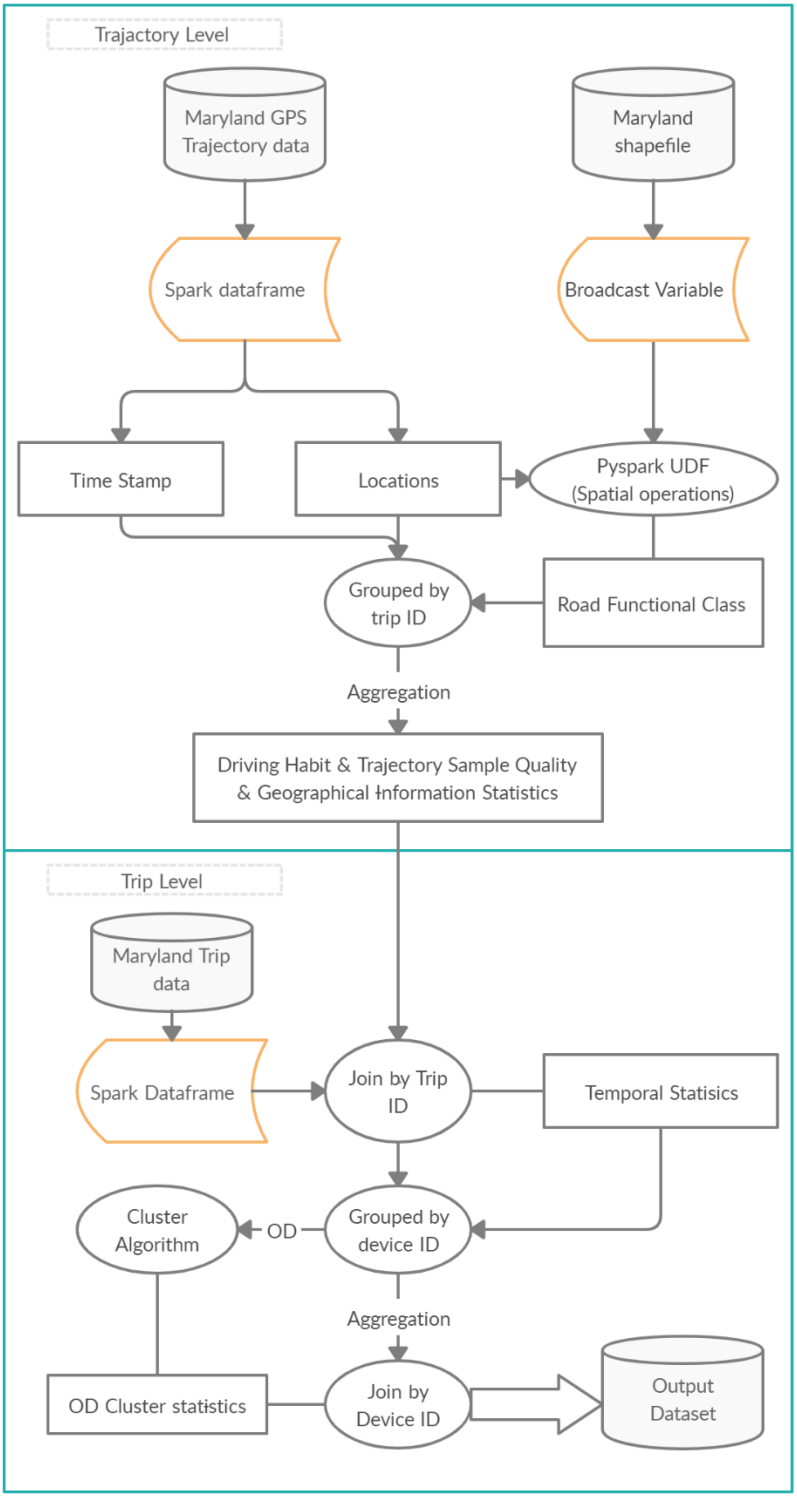

Figure 3 Data preparation procedure flow chart.

4.1. Driving habit statistics

This type of statistics can be extracted easily from the trajectory data by aggregating the spatial or temporal information from the trajectory. some of them have already be provided by the trip data set.

4.1.1. Attributes description

Average Speed (mile/sec): This attribute estimates the average speed over all the trips for one particular vehicle.

$$Average\ Speed(v) = \frac{\sum_{t}^{N_v} \frac{\sum_{i=0}^{N_{vt}-1} Distance(p_i^{vt}, p_{i+1}^{vt})}{T_{p_{N_{vt}}^{vt}} - T_{p_0^{vt}}}}{N_v}$$

Average Trip Time Duration (sec): This attribute estimates the average time duration over all the trips for one particular vehicle.

$$Average\ Trip\ Time\ Duration(v) = \frac{\sum_{t=0}^{N_v} T_{p_{N_{vt}}^{vt}} - T_{p_0^{vt}}}{N_v}$$

Average Trip Distance (mile): This attribute estimates the average trip distance over all the trips for one particular vehicle.

$$Average\ Trip\ Distance(v) = \frac{\sum_{t=0}^{N_v} \sum_{i=0}^{N_{vt}-1} Distance(p_i^{vt}, p_{i+1}^{vt})}{N_v}$$

Where $Distance(a, b)$ is the function to calculate the distance between point a and point b by their latitudes and longitudes.

Detour Metric: This attribute represents to what extent the real traveled distance is bigger than the geometric distance between the origin and destination. It illustrates whether the vehicle always follows the shortest path from the origin to the destination or always changes its direction.

$$Detour\ Metic(v) = \frac{\sum_{t=0}^{N_v} \frac{\sum_{i=0}^{N_{vt}-1} Distance(p_i^{vt}, p_{i+1}^{vt})}{Distance\left(p_0^{vt}, p_{N_{vt}}^{vt}\right)}}{N_v}$$

Average Number of Stop Locations: Stop Locations here represents a location where the vehicle stops or travels at very low speed. The way to detect that is checking whether a vehicle stays at the same latitude and longitude for over 2s. The threshold is set as 2 s because it is the least seconds of stopping the latitude and longitude accuracy can distinguish.

$$Average\ Number\ of\ Stop\ Locations(v) = \frac{\sum_{t=0}^{N_v} \frac{\sum_{i=0}^{N_{vt}-1} checkmove(p_i^{vt}, p_{i+1}^{vt})}{N_{vt}}}{N_v}$$

$$checkmove(p_1, p_2) = \begin{cases} 1 & if\ Distance(p_1, p_2) == 0\ and\ |T_{p_1} - T_{p_2}| > 2 \\ 0 & others \end{cases}$$

Most Frequent Geospatial Type: This attribute is extracted directly from the Trip data, there are four categories in this column, 'IE', 'EE', 'II', 'EI'. Letter 'I' means Internal of Maryland and letter 'E' means External for Maryland. For example, 'EI' means a trip starts from an external place of Maryland and then ends inside Maryland finally.

4.1.2. Attributes distribution analysis

The numeric variables above have different density distributions, whose Gaussian kernel density estimation plots and the density histograms are shown below. Since this research is mainly conducting the vehicle category classification on the vehicles with trip numbers larger than 10, two density distribution were plotted so that the researchers can validate the sample is greatly biased.

Figure 4 and Figure 5 shows that there is only a small distribution change for the five variables between the whole data set and the filtered data set. Nevertheless, there is a noticeable distribution change in the 'Detour Metric' for C and T. The C and TST classes can be easily classified by some values of the statistics. However, the T and FD classes always overlap in the density distribution plot, meaning hard to separate. Furthermore, we can also rank the importance of those variables for pattern recognition by separating the classes. The average speed and the detour metric can be the most important variables.

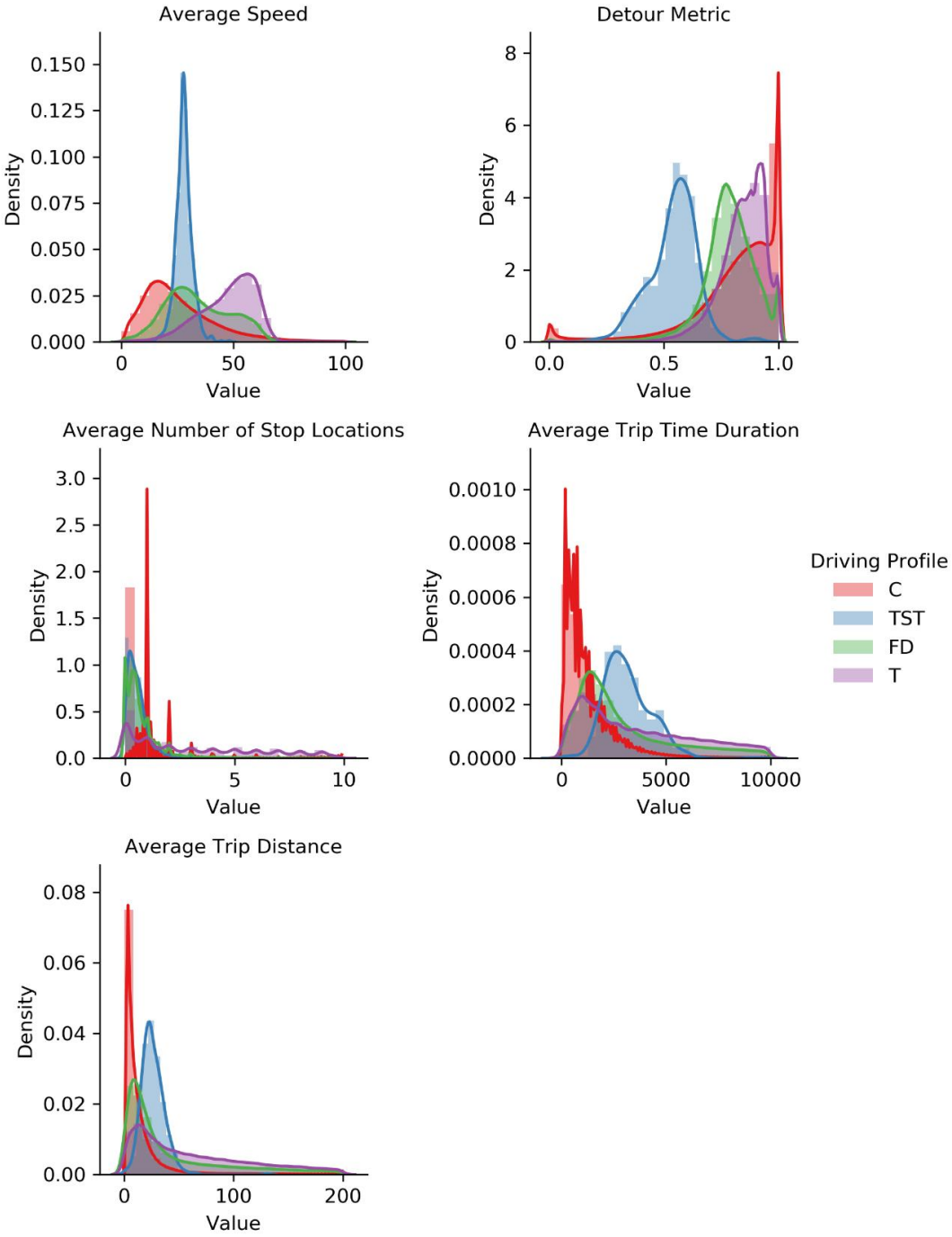

Figure 4 Gaussian density distribution for the numeric driving habit statistics (all vehicles, outliers removed).

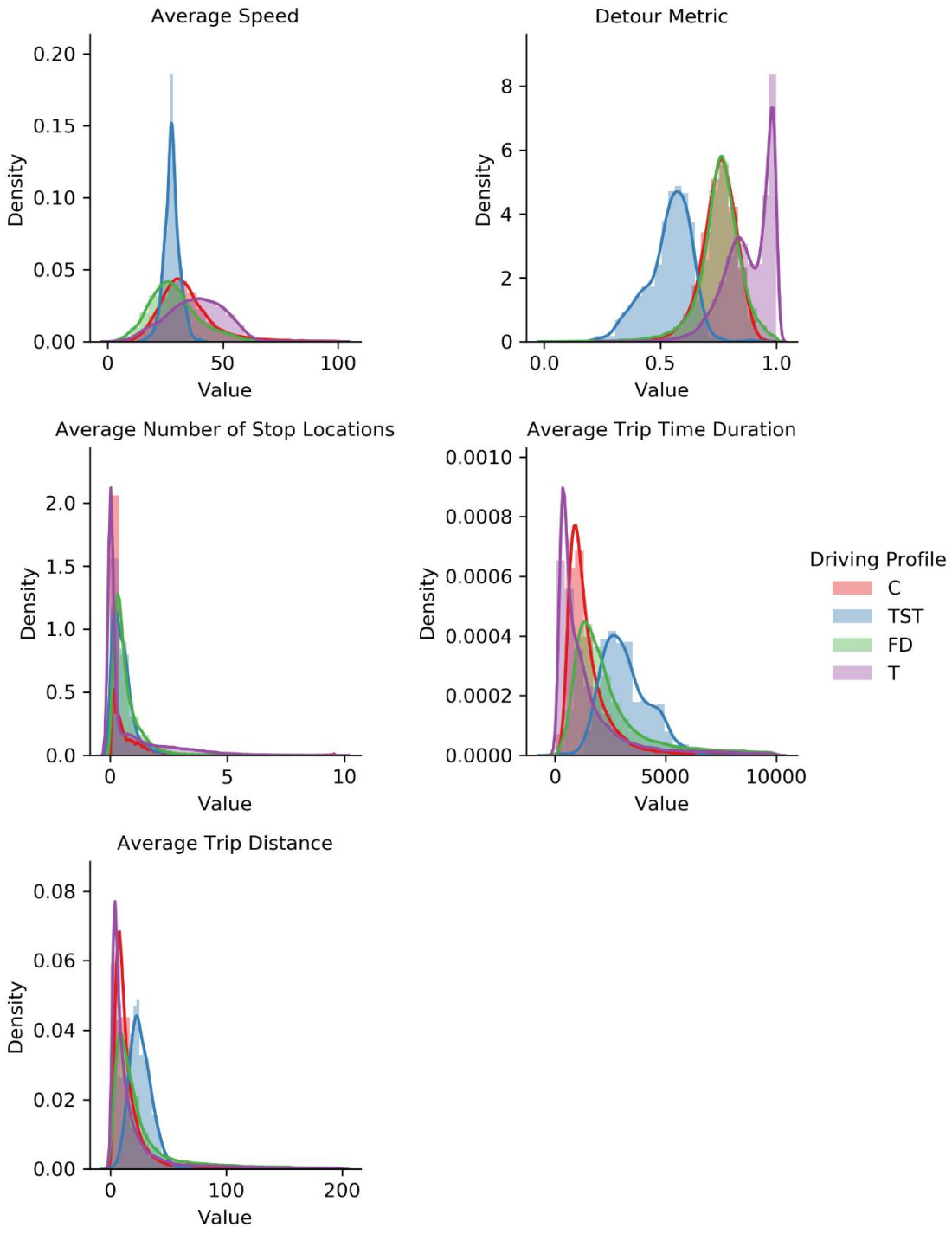

Figure 5 Gaussian density distribution for the numeric driving habit statistics (vehicles with trip number > 10, outliers removed).

4.2. Trajectory sample quality statistics

The trajectory sample quality statistics are extracted at the same time with the driving habit statistics, but it has a different impact on the vehicle category's classification task. Intuitively, there could be a strong relationship between GPS devices and the trajectory sample quality. GPS devices can be regarded as an indispensable factor for inferring a vehicle category. For example, the taxi and truck fleets tend to have high-precision embedded GPS devices on their vehicles. However, the precision for the GPS devices may differ by state and time, so its high precision may not be generalized in other states' data or the data in the future. Hence, two models with and without the trajectory sample quality statistics were tested in the later sections to address its generalization problem.

4.2.1. Attributes description

Average GPS Trajectory Point Number: This attribute shows how many trajectory points on average within each vehicle.

$$GPS\ Trajectory\ Point\ Number(v) = \frac{\sum_{t=0}^{N_v} N_{vt}}{N_v}$$

Average Time Granularity (s): This attribute measures the average time interval between two consecutive trajectory points.

$$Average\ Time\ Granularity(v) = \frac{\sum_{t=0}^{N_v} \frac{\sum_{i=0}^{N_{vt}-1} T_{p_{i+1}^{vt}} - T_{p_i^{vt}}}{N_{vt}}}{N_v}$$

Average Distance Granularity (mile): This attribute measures the average distance between two consecutive trajectory points.

$$Average\ Distance\ Granularity(v) = \frac{\sum_{t=0}^{N_v} \frac{\sum_{i=0}^{N_{vt}-1} Distance(p_i^{vt}, p_{i+1}^{vt})}{N_{vt}}}{N_v}$$

4.2.2. Attributes distribution analysis

According to Figure 6 and Figure 7, the four classes have different distributions in the three variables, especially for 'Average Time Granularity'. The filtering procedure does not greatly affect the distributions for the four vehicle classes. Only C and T for 'Average Time Granularity' are slightly affected by the filtering process.

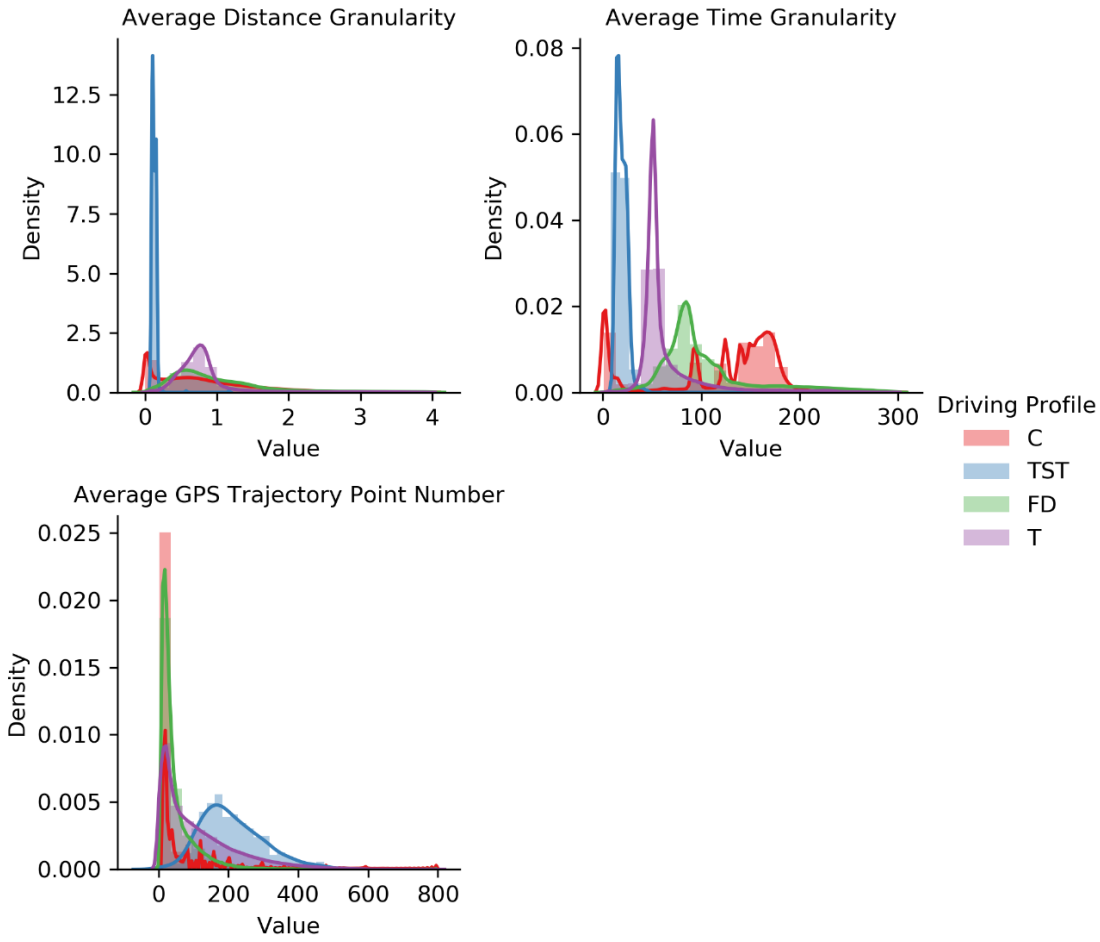

Figure 6 Gaussian density distribution for the trajectory sample quality statistics (all vehicles, outliers removed).

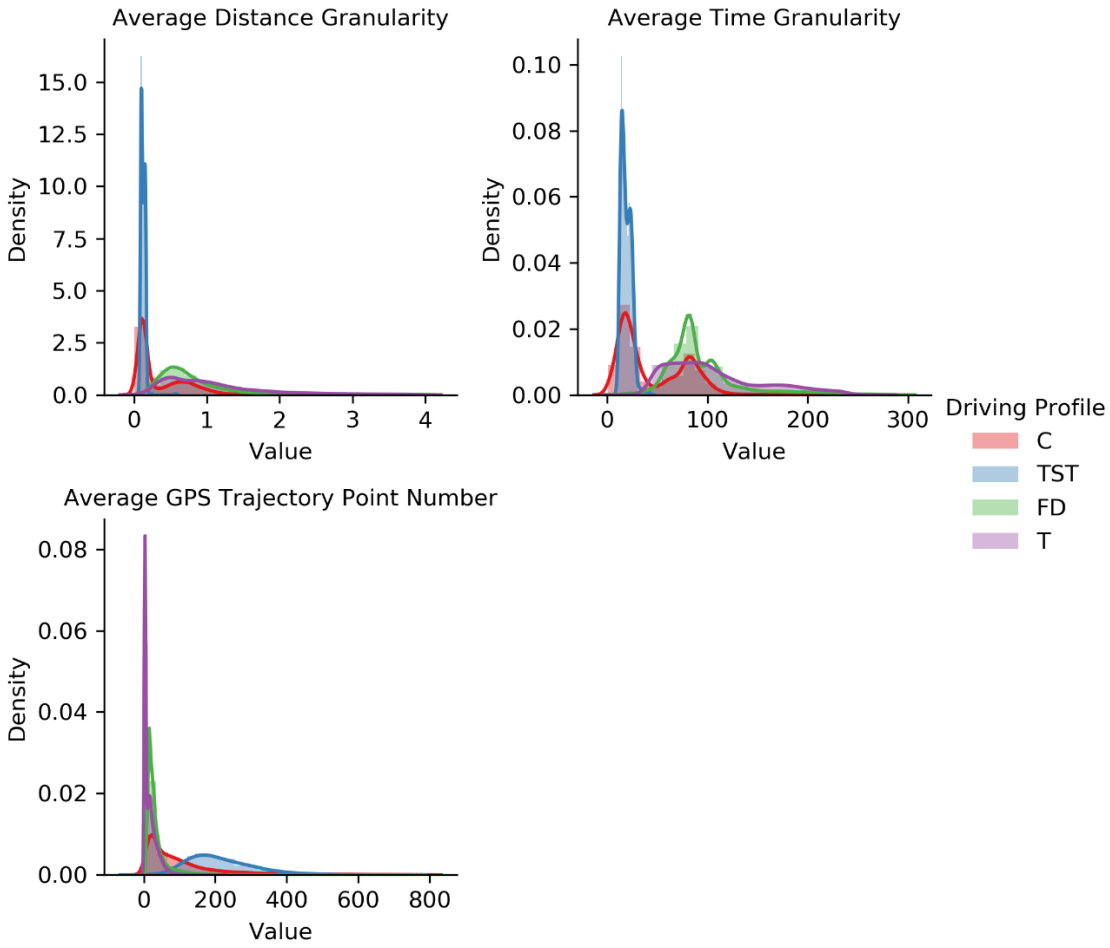

Figure 7 Gaussian density distribution for the trajectory sample quality statistics (vehicles with trip number > 10, outlier removed).

4.3. Geographical information statistics

More statistics regarding the trajectories can be collected by combining the data set with other geographical information by the spatial operation.

Figure 8 Common positions of GPS points on the roadway.

As shown in Figure 8, the points are not exactly located on the roadway. The nearest neighbor algorithm could help to join the nearest road segment attributes to the corresponding trajectory points. Since the traditional spatial operation maintained in the Pyspark geo package does not apply the method to find the nearest neighbor, the Pyspark user-defined function is utilized to build up our own method. In the UDF, the Rtree is invoked to fast index the possible road segments and then find the nearest road segment among them.

4.3.1. Attributes description

Average Functional Type Proportion: In order to describe the preference to road function type for each vehicle type, the proportions of the functional types within each

trip are summarized for individual vehicles. There will be five variables will be extracted in this part, corresponding to five road functional type attributes. Besides, the trajectory outside Maryland will be assigned ‘None’ type by default.

4.3.2. Attributes distribution analysis

According to Figure 9 and Figure 10, the consumer value distribution has been changed by the filtering procedure. The vehicles in C with the number of trips smaller than 10 tend to have more trajectory located on the ‘None’ road function type, which may result from matching the road segment with missing value or locating out of Maryland state.
It seems there are overlapping peaks for the density distribution over the four classes, but their variable deviations are different. For example, TST tends to have more concentrated distribution for variable ‘Local’, while C tends to have more concentrated distribution for variable ‘None’.

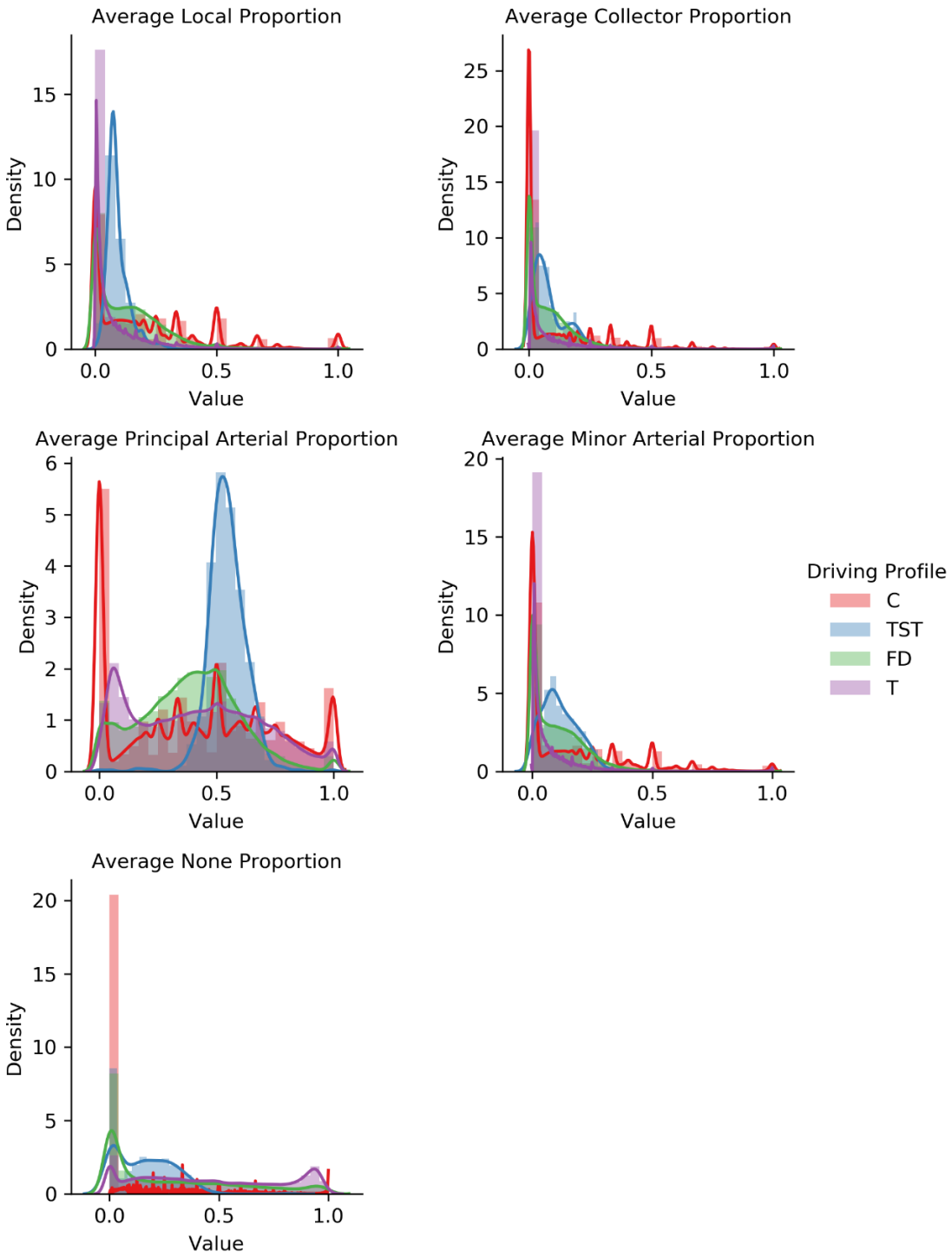

Figure 9 Gaussian density distribution for the geographical information statistics (all vehicles, outlier removed).

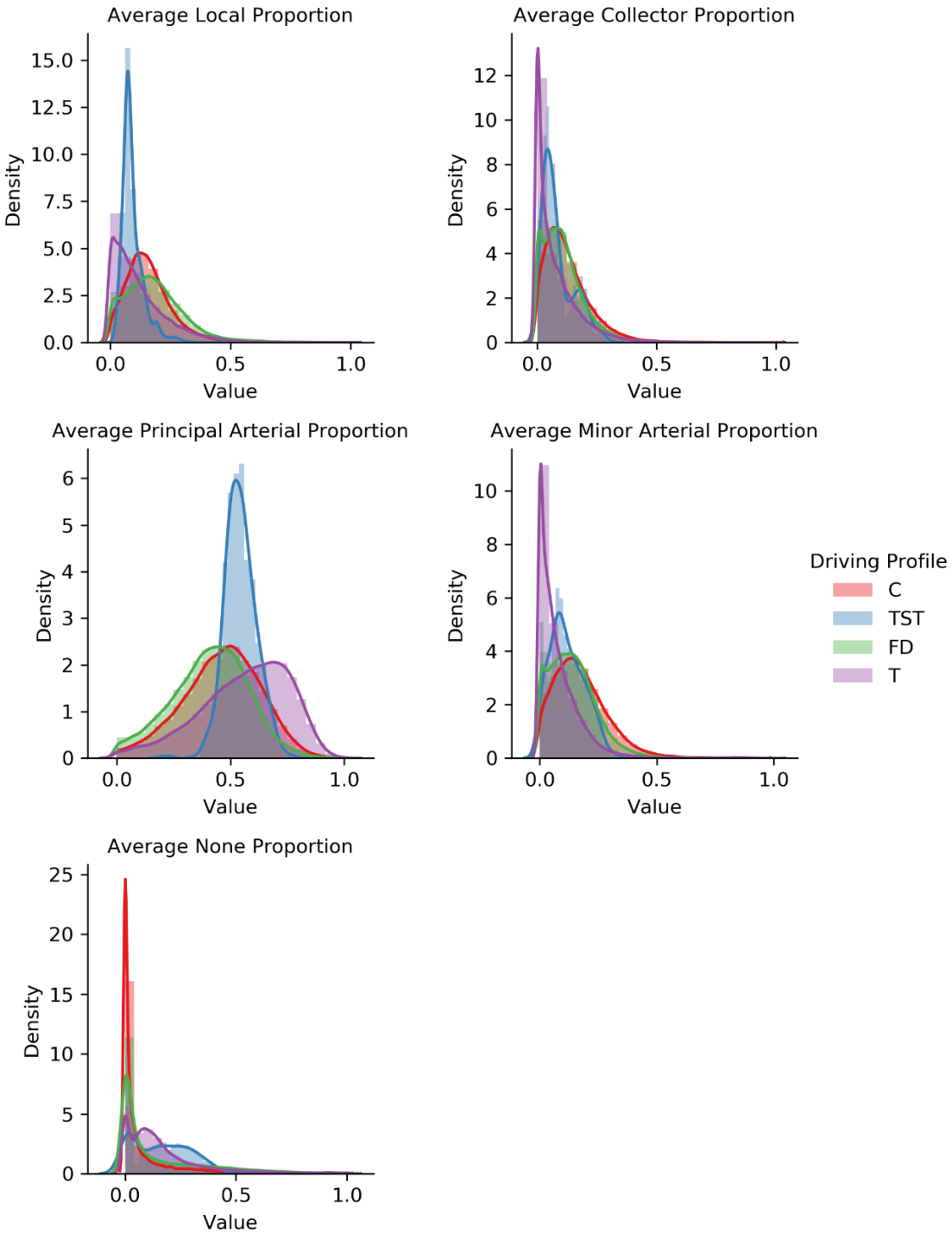

Figure 10 Gaussian density distribution for the geographical information statistics (vehicles with trip number > 10, outlier removed).

4.4. Origin and destination cluster statistics

The vehicles always contain some daily or weekly routines which start and end in similar locations. Here we developed an algorithm for discovering the frequent similar OD pairs. The approach to extract information for those pairs is by spatial cluster algorithm. The spatial cluster algorithms are widely applied to study the geographical distribution for a variety of topics. Kim J (24) studied the traffic patterns in the network by the GPS trajectory data. In this research, a well-defined spatial similarity measurement for trajectory data is introduced, taken as the distance of the DBSCAN algorithm. Guo D (25) utilized origin and destinations for the Taxi trajectory data to discovering spatial patterns by the k-means algorithm.

DBSCAN *(26)* is an efficient algorithm for the spatial cluster problem, which does not require a predefined cluster number. This study combined the origins and destinations to be 4-dimensional spatial points. Then the DBSCAN algorithm was applied to the spatial points with well-tuned hyperparameters. Figure 11 shows the result for the spatial cluster algorithm of one particular vehicle. The trips whose ODs are close was classified to the same cluster.

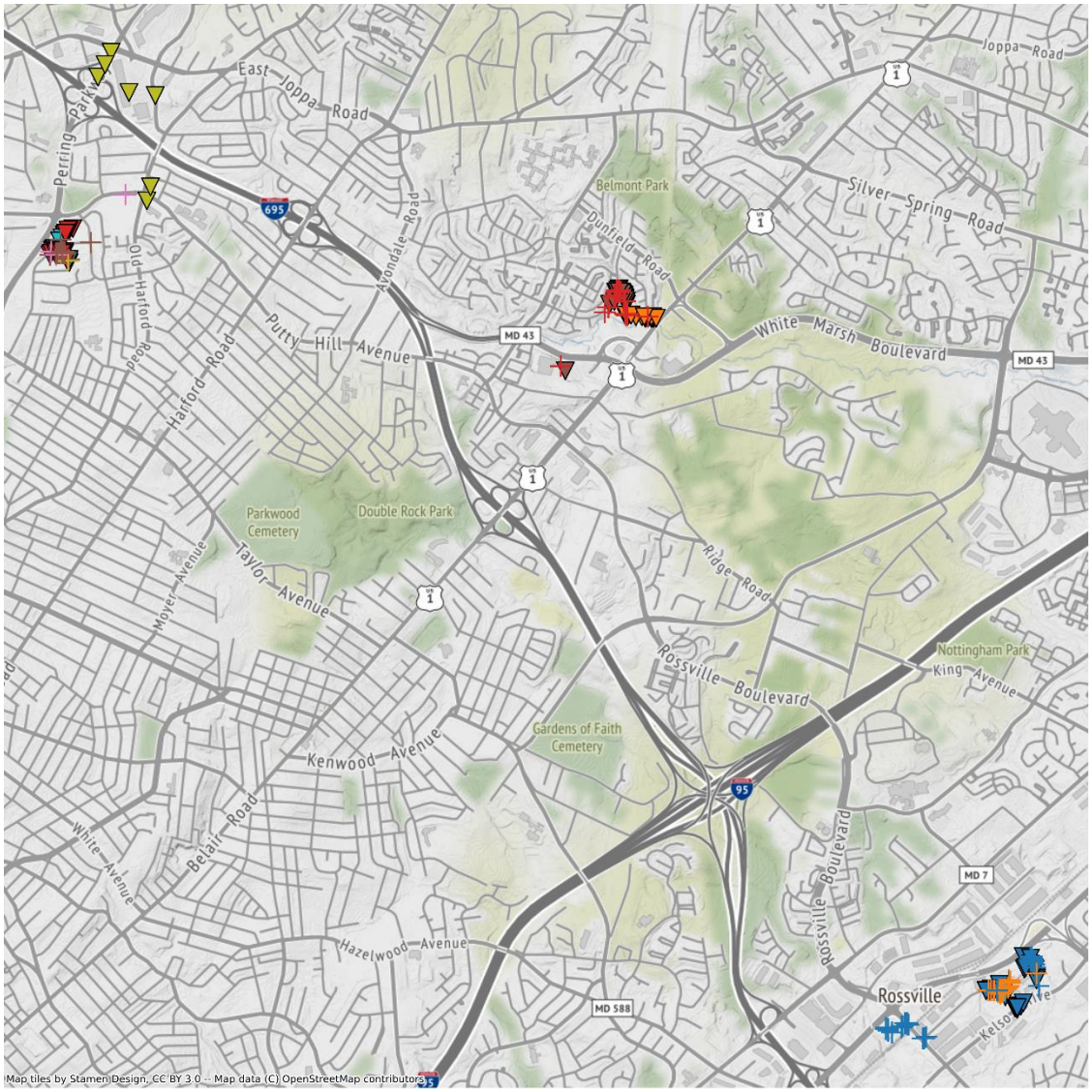

Figure 11 Origin destination plot (cluster size > 3). The triangles represent origins, and the plus symbols represent destinations.

4.4.1. Attributes description

The Biggest Cluster Size Proportion: This attribute describes the total trip number of the biggest cluster. Through this variable, the model can understand whether there is one cluster dominated among the number of all trips. Since the total number of trips will affect the cluster size significantly, we extracted the proportion for the cluster number in the total number of trips for the vehicle instead.

$$The\ Biggest\ Cluster\ Size\ Metrics(v)\ =\ \frac{C_{1st}}{N_v}$$

The Second Biggest Cluster Size Proportion: Regarding that most of the daily commuting will have a similar number of trips with the reverse OD (trips from A to B and trips from B to A). Therefore, the second biggest of the cluster size is extracted as well.

$$The\ Second\ Biggest\ Cluster\ Size\ Metrics(v)\ =\ \frac{C_{2st}}{N_v}$$

Total Cluster Number: This attribute illustrates whether this vehicle tends to have constant OD routine, especially combined with the biggest and the second biggest cluster size.

$$Total\ Cluster\ number(v)\ =\ |C|$$

Where C_{1st} and C_{2st} are the first and second cluster numbers for a particular vehicle, C is the set of all the clusters.

4.4.2. Attributes distribution analysis

Due to the filtering process, only vehicles with trips more than 10 have the OD cluster features. As a result, there will only be one density distribution plot (Figure 12) for the vehicles with the number of trips < 10. Similar to the geographical information statistics, the distribution peaks for the variables are close but they tend to have varying deviation. Besides, when removing the outliers, there are more vehicles with no cluster in class T, resulting in 0 value in each variable.

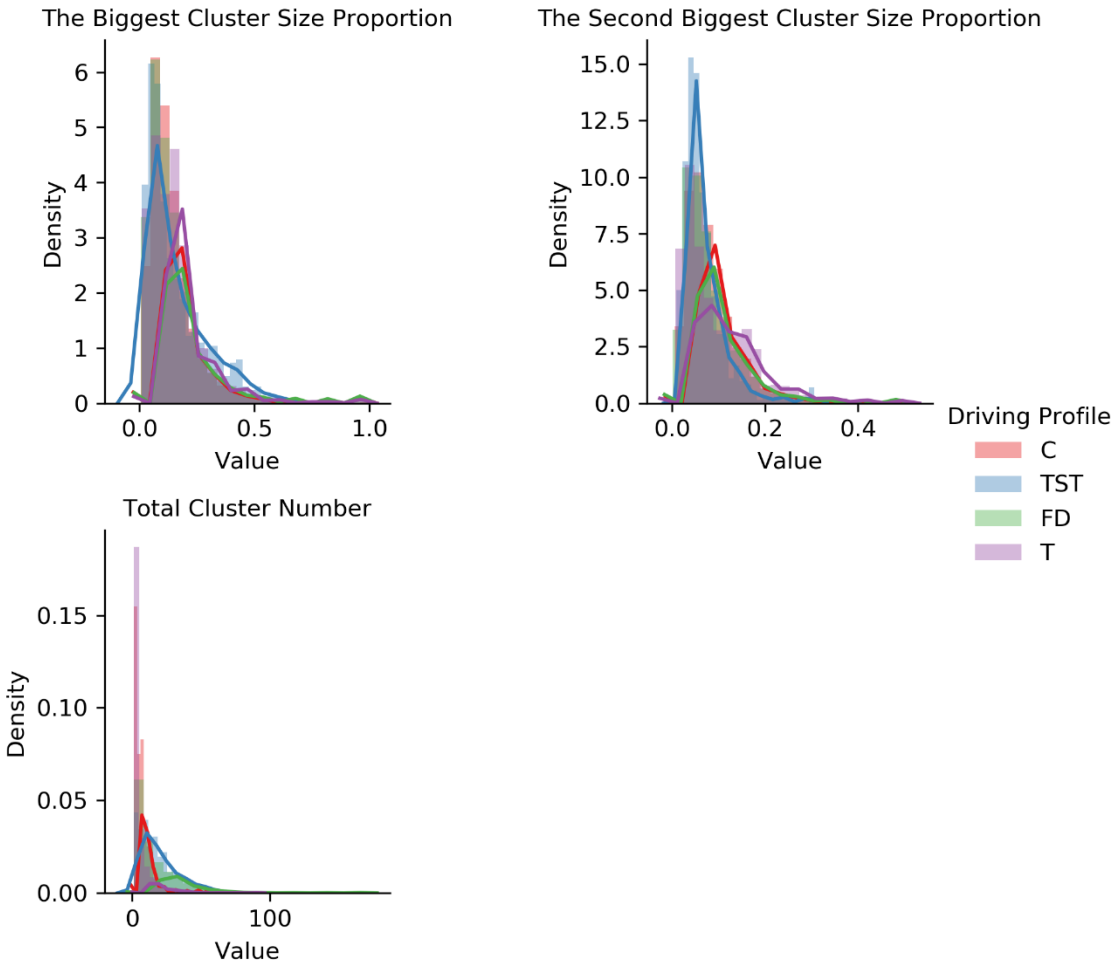

Figure 12 Gaussian density distribution for the numeric OD cluster Statistics (vehicles with trip number > 10, outlier removed).

4.5. Temporal statistics

Like the OD cluster algorithm, the timestamps can also be clustered, showing whether the vehicle has a regular travel routine. Here, the middle timestamp of their start and end will be assigned to different hour segments.

4.5.1. Attributes description

Most Frequent Hour: This attribute derives from the timestamp for the trip information. The trips for one vehicle may happen mainly into one time slot. Hence, the most frequent hours represents the time slot at which the vehicle is in service most frequently.

$$Most\ Frequent\ Hour(v) = max_{hour}(\sum_{t=0}^{N_v} CheckInside(\frac{T_{p_{N_{vt}}^{vt}} - T_{p_0^{vt}}}{2}, hour)$$

$$CheckInside(time, hour) = \begin{cases} 1 & Hour(time) == hour \\ 0 & others \end{cases}$$

Where Hour(timestamp) is the sharp hour for the timestamp.

Most Frequent Hour Proportion: This variable will further describe to what extent the vehicle will provide service within the 'Most Frequently Hour'.

$$Most\ Frequent\ Hour\ Proportion(v)$$

$$= \frac{\sum_{t=0}^{n_v} CheckInside(\frac{T_{p_{N_{vt}}^{vt}} - T_{p_0^{vt}}}{2},\ Most\ Frequent\ Hour(v))}{N_v}$$

Distinct Hours Number: Similar to the 'Most Frequent Proportion' statistics, this variable helps the model to determine whether the vehicle was in service in a fixed time slot every day.

4.5.2. Attributes distribution analysis

There is no big distribution difference among the four vehicle classes for ‘Most Frequent Hour’ and ‘Most Frequent Hour Proportion’ statistics. However, there is a significant discrepancy in the ‘Distinct Hours Number’ statistics’ distribution. Comparing Figure 13 with Figure 14, there are a large number of vehicles in class C having only one trip recorded, causing the discrepancy in ‘Distinct Hours Number’ and Most Frequent Hour Proportion.

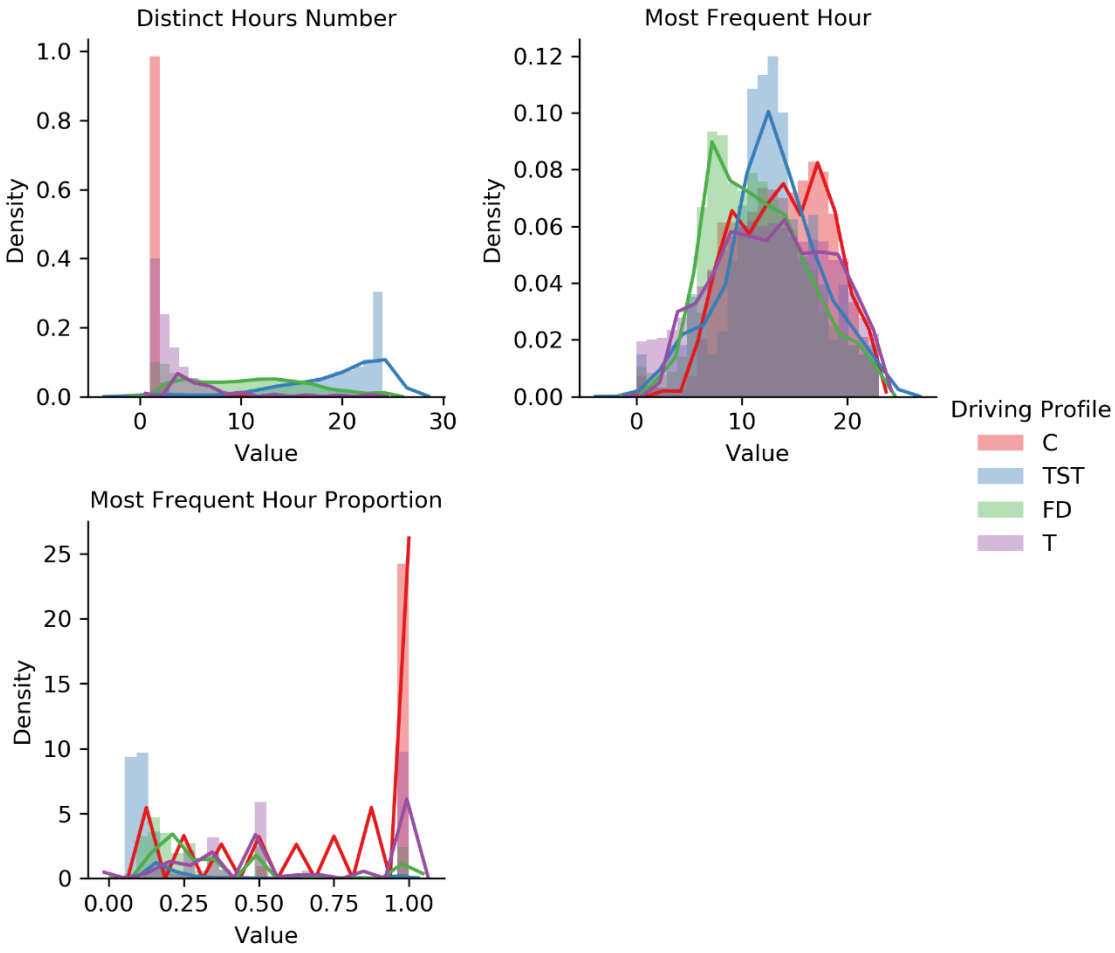

Figure 13 Gaussian density distribution for the temporal statistics (all vehicles, outlier removed).

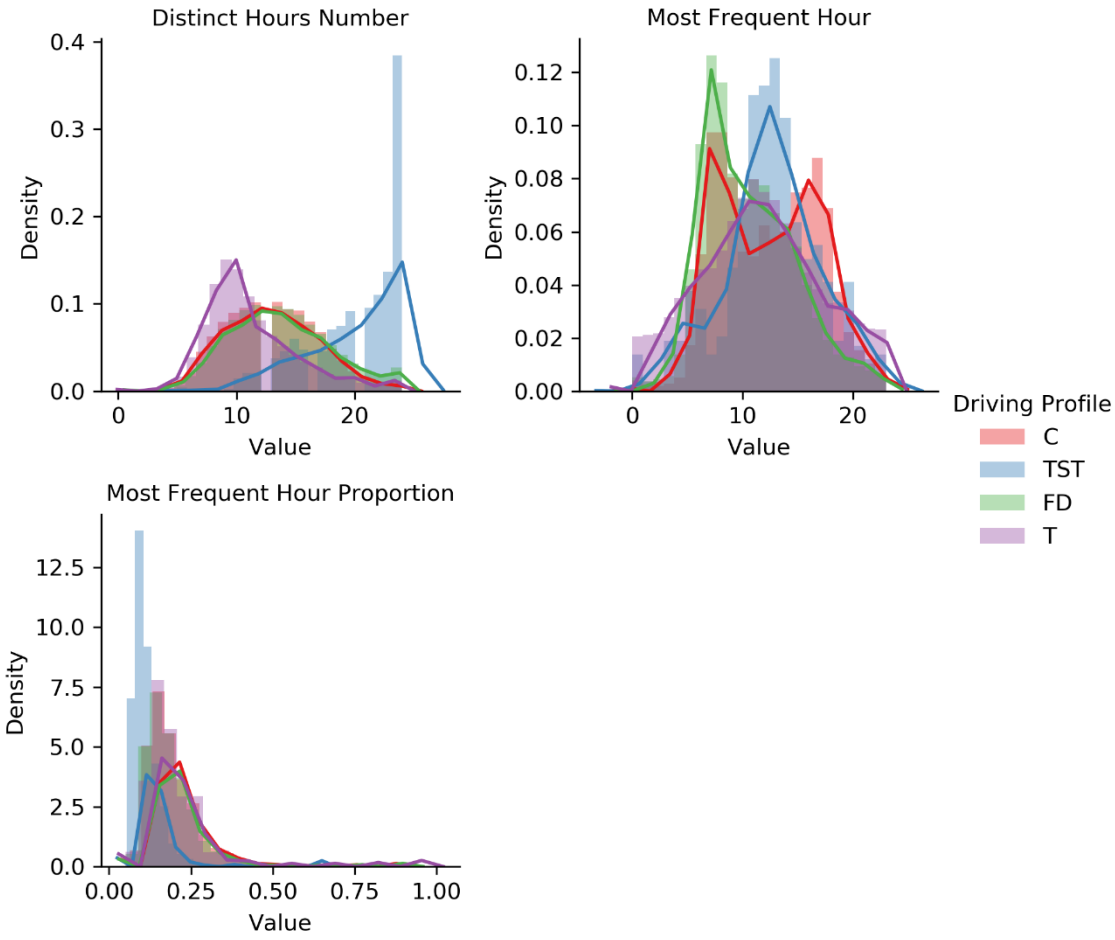

Figure 14 Gaussian density distribution for the temporal statistics (vehicles with trip number > 10, outlier removed).

5. METHODOLOGY

Four different algorithms are tested in this research, namely logistic regression, Support Vector Machine (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). The whole data set is separated into training, validation and testing setsby 3:4:8. The hyperparameters are tuned by the cross-validation on the training set. Since some algorithms take numeric variables only, the categorical variables need to be changed into one-hot encoding in advance. To better compare the performance, the precision, recall and the average accuracy for each class are calculated based on the test set.

5.1. Multinomial logistic regression

The logistic regression *(17)* derives from the linear regression by adding sigmoid function at the end of the model, representing the probability of binary classification. The algorithm can be described as the equation below.

$$\hat{y}_i = \frac{e^{W \cdot X_i + b}}{e^{W \cdot X_i + b} + 1}$$

$$Cost(W) = \frac{1}{n} \sum_{i=1}^{i=n} -y_i log(\hat{y}_i) - (1 - y_i) log(\hat{y}_i) + C|W|^2, y_i, \hat{y}_i \in \{0, 1\}$$

Where X_i represents the input vector, y_i represents the ground truth label, $\hat{y}_i$ represents the predicted label, W represents the weights vector, b represents the bias.

Here we use several simple Logistic Regression classifiers to classify multiple classes. If there are n different classes, there will be n classifiers that distinguish specifically on one class to the others, which is called one vs. rest strategy. Each classifier will be trained over the whole train data. The class which has the highest probability will be chosen as the result.

5.2. Support vector machine (SVM)

SVM *(18)* is a non-probability binary classification algorithm. This algorithm will find out a hyperplane that has the largest margin to both of the classes. In addition to linear classification, kernel function can be utilized to solve the non-linear classification problems.

The algorithm can be described as the equation below.

$$\hat{y}_i = \begin{cases} 1 & if\ W \cdot \varphi(X_i) + b > 0 \\ -1 & if\ W \cdot \varphi(X_i) + b < 0 \end{cases}$$

$$Cost(W) = \frac{1}{n}\left(\sum_{i=1}^{n} max(0, 1 - y_i\hat{y}_i)\right) + C|W|^2$$

Where X represents the input vector, y_i represents the ground-truth label, $\hat{y}_i$ represents the predicted label, W represents the weights vector, b represents the bias, $\varphi(X_i)$ determined by what kind of kernel function the model takes.

Polynomial, sigmoid, and Gaussian radial basis are tested as kernel function in our model. In order to do multilabel classification, one vs. one strategy is adopted in the algorithm. One vs. one algorithm will train $\frac{n(n+1)}{2}$ classifiers for n classes, because each pair of classes need one classifier. The class which is voted most for +1 will be selected as the final output.

5.3. Random forest (RF)

RF *(19)* is a bagging ensemble method that is made up of many randomly trained decision trees. The decision trees are trained by different samples of data and the randomly selected attributes. Decision trees can take both categorical and numerical input. The decision tree is trained by splitting the data space into two spaces, which has the biggest information gain.

The advantage of the tree ensemble algorithms is that they are suited for high dimensional data regardless of their correlations. Besides, there is no need for data normalization and it is easy to implement. The final output for the RF is the class voted by most of the decision trees in the RF.

5.4. XGBoost

XGBoost *(20)* is a well-known implementation of the gradient boosting algorithm. It has achieved good performance in a variety of tasks. The model is ensembled by a group of

decision trees. Unlike the RF algorithm, the trees are formed iteratively in order to minimize the loss function. In each iteration, the new trees will be built by fitting the residuals between the ground truth value and the predicted value in the last iteration. Supposing $\hat{y}_i^{(m)}$ and $Residual_i^{(m)}$ are the predicted value and the residual for the ith input in the mth iteration, T_m is the new added decision tree in the mth iteration.

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + T_m(X_i, \theta)$$

$$Residual_i^{(m)} = \left| y - \hat{y}_i^{(m-1)} \right|$$

$$\theta = \min_\theta \left(\left| Residual_i^{(m)} - T_m(X_i, \theta) \right| \right)$$

Where X denotes the input vector, y_i denotes the ground truth label, $\hat{y}_i$ denotes the predicted label, θ denotes the parameter for the decision tree T.

6. EVALUATION

To study the impact of each statistic on the classification model, we built a baseline model on the driving habit statistics. Beyond the baseline, the temporal, trajectory quality statistics, OD cluster statistics, and Geographical information were exploited to boost the model accuracy. Some discrepancies in the improvement will be discussed in this part.

There are 140,952 vehicles recorded in the final data set after filtering the vehicles which do not have adequate trips for the DBSCAN algorithm. However, there is a substantial data skew in the data set (C: 24,008, TST: 608, FD: 65,697, T: 50,639). In order to avoid this data skew badly affecting the accuracy of our model, mainly two approaches are selected. First, only subsets of individual vehicle classes are sampled and used to train the models. The sample number is set as the total number of trip in the ‘TST’ class, which is the class with the least vehicle number.

The advantage of this method is that all the four classes will be reserved and it is easy for the training procedure due to the less amount of data. The drawback is the accuracy will decrease to some extent due to the shrink of the training set (1945 totally).
The second approach is to discard the Taxi class from the whole data set. The class TST only occupies a tiny proportion of the vehicle devices. If we discard them and do the

classification based on the three remaining classes, more data (57619 totally) can be used as the train set for our models.

Four different models are tested in the experiment, namely, logistic regression, SVM, RF, and XGBoost. The whole data set is randomly separated into a training data set (80%) and a testing data set (20%). The hyperparameters are tuned by the 3-fold cross-validation on the training set. To better evaluate the performance, the precisions, recalls, average precisions and F1 scores are calculated based on the test set. The precision $(\frac{TP}{TP + FN})$ describes how many vehicles the classifier correctly classifies for one class compared with the ground truth. The recall $(\frac{TP}{TP + TN})$ describes how many vehicles the classifier correctly classified for one class compared with all the vehicles predicted by the same label. There are two ways to calculate the average precision, namely, micro and macro average. Macro average precision means the average across the precisions of all the classes, and micro average means average the precision weighted by the number of ground truth instances for that class. F1 score for one class is the harmonic mean of precision and recall for that class.

6.1. Classification for 4 classes

6.1.1. Multiclass classification baseline (four classes)

The driving habit statistics are utilized as the baseline input for the classification models. The four machine learning models discussed above fed by the driving habit statistics are tested as baselines.

Table 1 The test-set performance details for baseline models by vehicle category.

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.5079	0.8027	0.5443	0.7259	0.6632
	Recall	0.6095	0.8872	0.3333	0.8167	
SVM	Precision	0.7083	0.8571	0.5625	0.9	0.7495
	Recall	0.6476	0.9474	0.6279	0.75	
RF	Precision	0.8	0.9466	0.7163	0.8957	0.8378
	Recall	0.7619	0.9323	0.7829	0.8583	
XGBoost	Precision	0.7387	0.9462	0.7222	0.85	0.8172
	Recall	0.7619	0.9474	0.7519	0.7917	

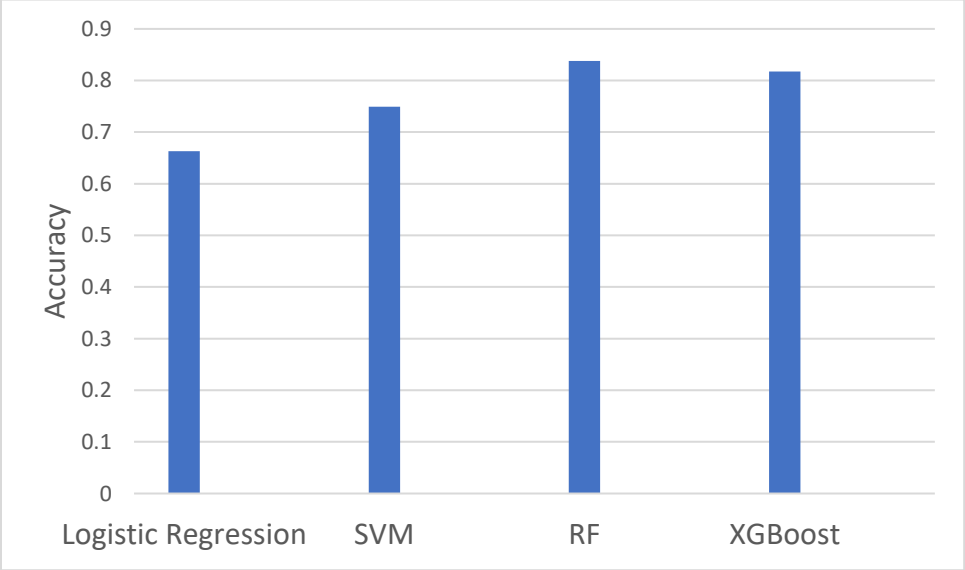

Figure 15 Micro average precision of test data set for baseline.

According to Figure 15 and Table 1, the RF and XGBoost algorithms perform far better than the other two among the four tested baseline algorithms. The TST class tends to have higher precision and recalls than the other three classes. Class C has the worst performance in the recall, while Class FD is at the rear of the recall.

6.1.2. Improvement of baseline by other statistics individually

The other four statistics will improve the performance of the model to varying degrees. Though the density distribution shows that the features can distinguish the four classes, it is necessary to test the performance of those features separately. As a result, each part of the statistics has been combined with the baseline model. The micro average of the Precision or Recall will be calculated on each new model. Besides, the average F1 scores in different algorithms are calculated so that the influence on the accuracy of each vehicle class can be studied.

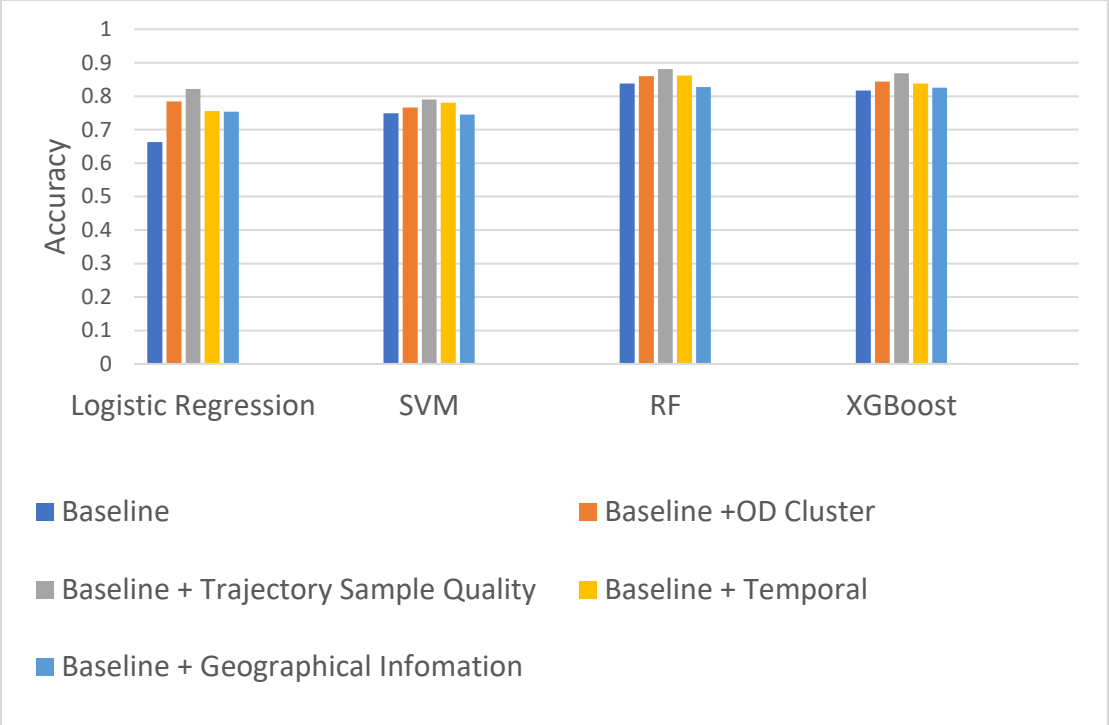

Figure 16 Micro average accuracy of test data set for models with different combination of statistics.

According to Figure 16, nearly all the types of the statistics increase the accuracy to some degrees, especially for the Logistic Regression model. Besides, the sample quality top the rank of accuracy and the other statistics will all improve the accuracy to some extent. What is also worthy of mentioning here is that though the OD Cluster statistics seem to have similar value distributions for the four classes, the boosting for the classification performance is pretty high. The geographical information statistic does not have satisfying performance-boosting, even causing damage to the accuracy of the SVM and RF models.

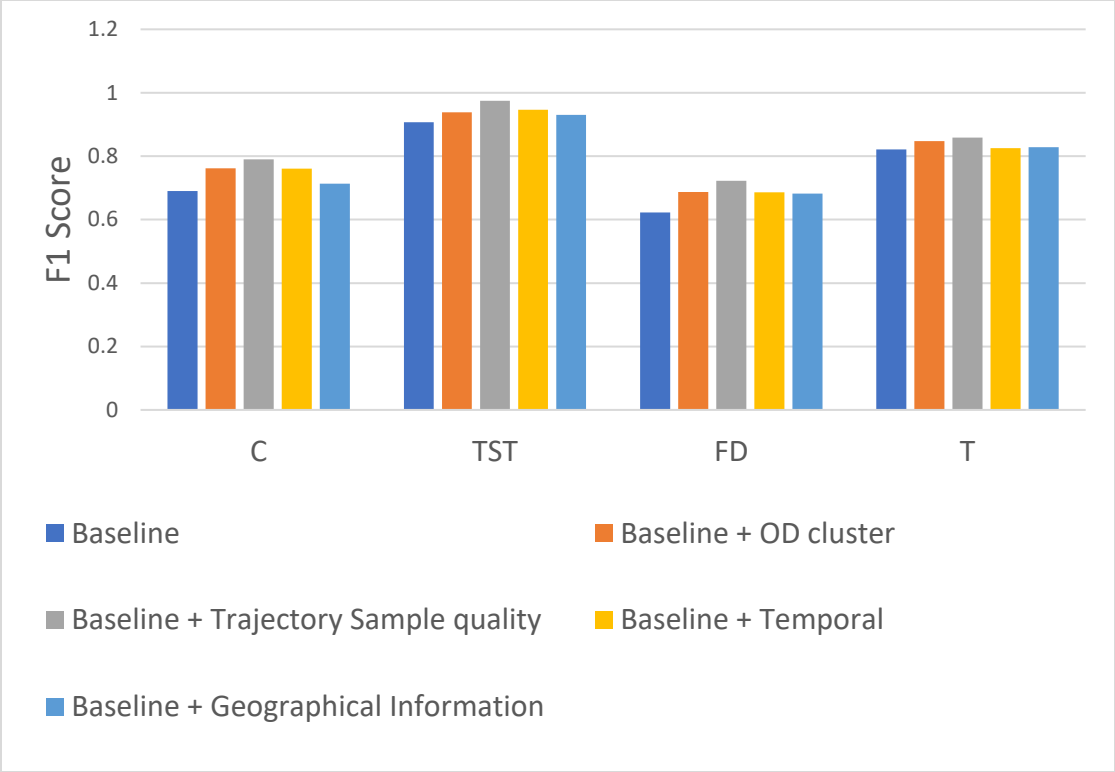

Figure 17 Average F1 score of test data set across the algorithms for the four classes.

Figure 17 illustrates which class contributes more to the accuracy by F1 score. It shows that every class contributes nearly the same to the total accuracy improvement, and the accuracy rank of the four classes does not change with different algorithm input variables.

6.1.3. Study for highly compound models

This section will test the compound models with features as much as possible. As we discussed above, the generalizability of the trajectory sample quality statistics is problematic. As a result, the highly compound model considers both the models with and

without the trajectory sample quality statistics. Besides, the variable amount can also be greatly enlarged further by adding variable deviations into the models. Thus, four models with different combinations of the statistics are tested in this section.

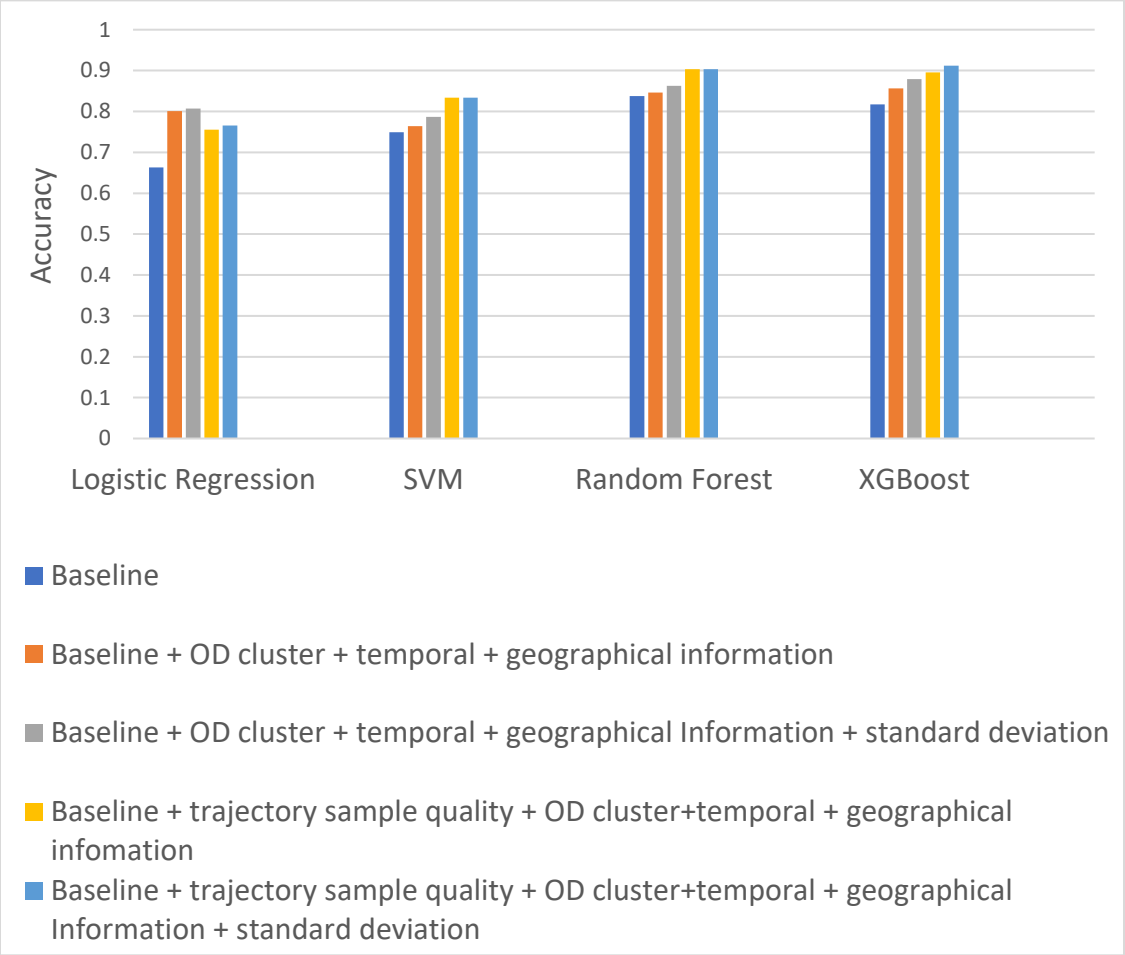

Figure 18 Micro accuracy of test data set for different combinations of statistics

According to Figure 18, combining all the features improves the accuracy to a great extent, but there is a decreasing trend in the bar for logistic regression when adding more

and more variables into the model. The standard deviations are also good features to distinguish the vehicle categories.

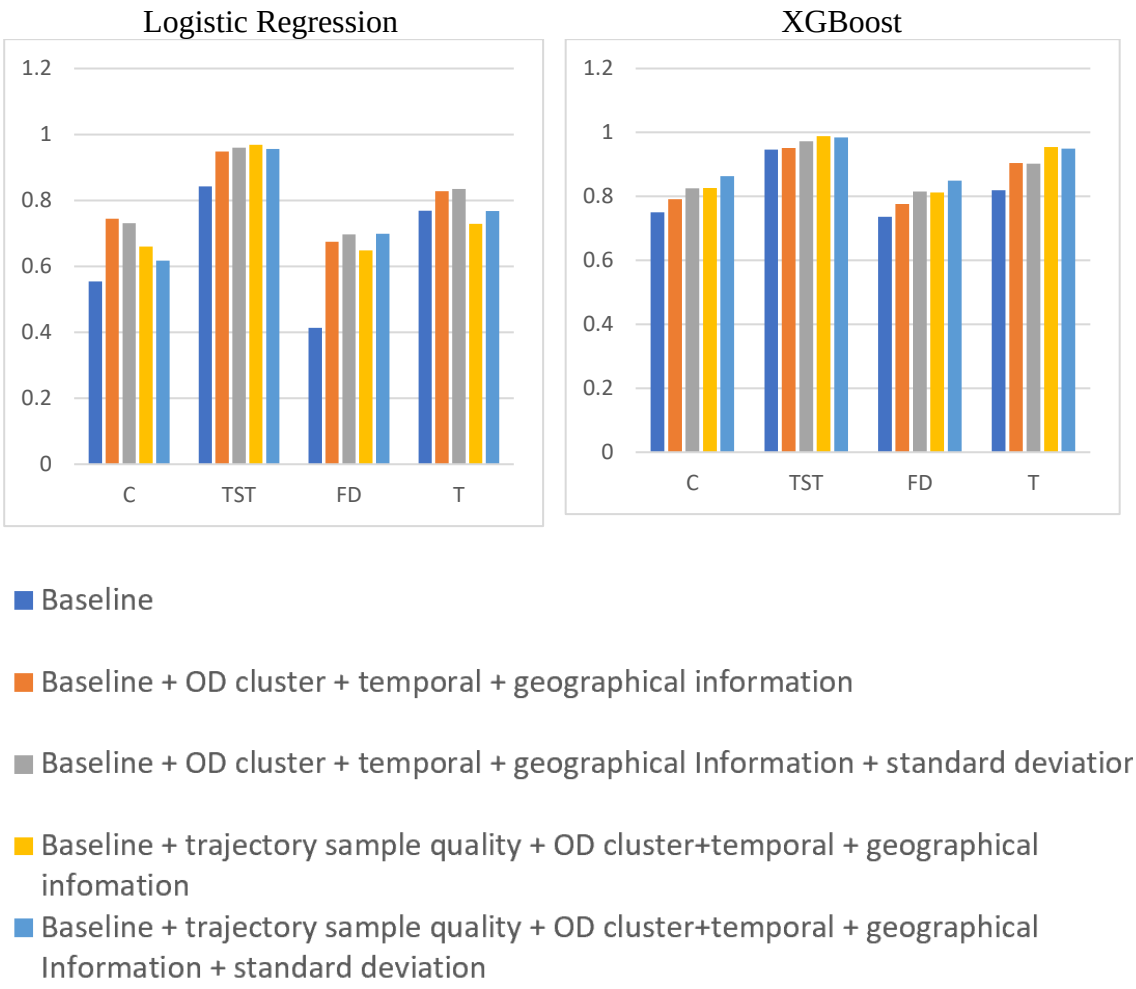

Figure 19 F1 score across four vehicle classes for logistic regression and XGBoost

The F1 score for different combinations of statistics by different vehicle classes are calculated to understand the abnormal declining trend for Logistic regression (shown in

Figure 19). The chart shows that the decreasing accuracy is mainly caused by C, FD and T.

6.2. Classification for 3 classes

Since the Taxi drops down the available amount to a great extent, the training data set can be greatly enlarged after dropping TST from the data set. Like the four-class classification task, the driving habit statistics are set as the baseline's input features. The driving habit combined with Temporal, OD cluster, Geographical Information and corresponding deviation statistics are set as the final model's input features. Since the SVC algorithm training complexity highly depends on the data size, it is abandoned in the 3-class classification task. The macro average of the precision for three classes (C, FD, T) is calculated to study the influence of the data size to the performance.

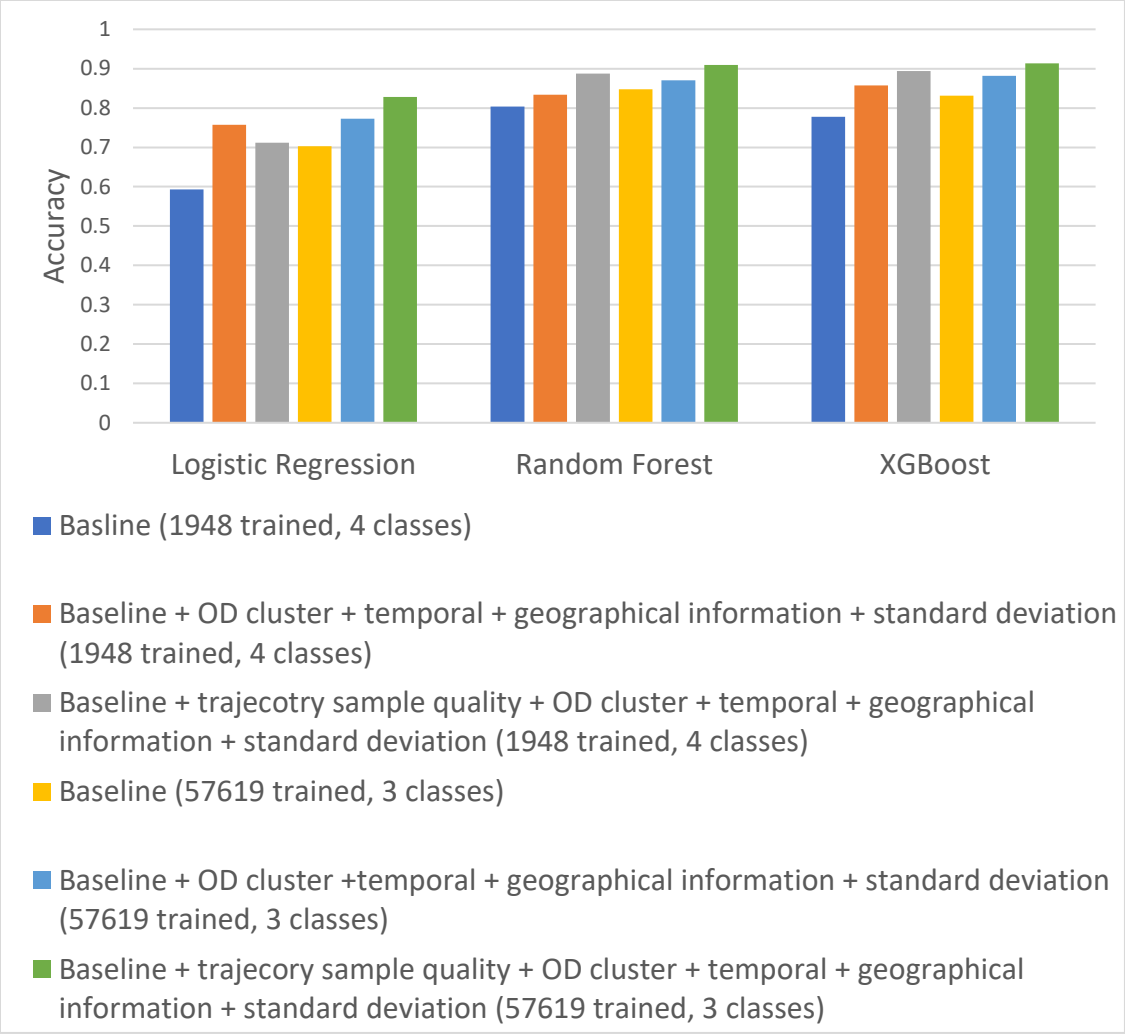

Figure 20 Macro average precision for different features and training data sizes

The result in Figure 20 illustrates that several thousand vehicle records are not adequate for the vehicle category classification, especially for logistic regression. There is an apparent increasing trend for the logistic regression algorithm when enlarging the training data size from two thousand to over fifty-thousand vehicles. The abnormal

decreasing trend in logistic regression has also diminished with the enlargement of the training data set.

The performance comparison for the 3-class and the 4-class models with all features is manifested in Table 2 and

Table 3. The TST and SVM are abandoned for the 4-class model because the 3class model does not have them.

Table 2 Test set performance details for the 4-class model with all features (1948 trained, 487 tested)

Model		C	FD	T	Macro average
Logistic Regression	Precision	0.7412	0.6597	0.7344	0.7118
	Recall	0.5294	0.7422	0.8034	0.6917
RF	Precision	0.9556	0.7931	0.9134	0.88740.8874
	Recall	0.7227	0.8984	0.9915	0.8709
XGBoost	Precision	0.9074	0.8397	0.9339	0.8937
	Recall	0.8235	0.8594	0.9658	0.8829

Table 3 Test set performance details for the 3-class model with all features (57619 trained, 14405 tested)

Model		C	FD	T	Macro average
Logistic Regression	Precision	0.9013	0.7147	0.8684	0.8281
	Recall	0.7663	0.8090	0.8828	0.8194
RF	Precision	0.9637	0.8035	0.9603	0.9092
	Recall	0.7867	0.9343	0.9789	0.9
XGBoost	Precision	0.9341	0.8383	0.9671	0.9132
	Recall	0.8374	0.9111	0.9823	0.9103

7. LIMITATION AND FUTURE STUDY

Although this research has achieved accuracy over 90% to classify the vehicle category, there are still parts worthy of further study. The author spent much effort to design the Geographical Information part, expecting this variable could bring about huge performance improvement. Nevertheless, there are still two problems with that part. The accuracy of combining the Geographical Information Statistics with the driving habit Statistics bottomed the four combined models' ranks, which is disappointing. In addition, there is only one type of geographical information extracted from the shapefile. It is possible that other attributes like the population and job number of the areas in which the vehicle passed can have better performance to infer the vehicle category.

Besides, more studies can be conducted on how the variables improve accuracy. The data preparation part mentions that the density distribution for OD cluster variables is similar. However, the result shows that the OD cluster variables can improve the accuracy greatly. More exploration of the association between the variables and model performances can be examined in future studies.

8. CONCLUSION

The vehicle category classification based on GPS trajectory data can be well conducted by extracting statistics from the trajectory data, and a variety of machine learning algorithms can be utilized. In this research, the Tree Ensemble algorithms (RF and XGBoost) have been proved to be the best algorithm over logistic regression and SVM irrespective of the variables. The varying accuracy of the four classes shows a discrepancy for the task complexity as well. Generally, TST usually tops the list of four classes in terms of accuracy, followed by T. More attributes and more training data set will increase the accuracy for vehicle classification.

Besides, different variables show different power to improve the accuracy over the baseline model. The GPS trajectory sample quality statistics are the best features to improve accuracy, but more studies are needed to prove the generalizability of this feature. Though Geospatial Information Statistics do not increase the performance like the other statistics, it provides us a framework to enrich the attributes by joining geographical information with GPS trajectory data through spark.

REFERENCES

[1] Tishkin V F, Yashina M V, Moseva M S, et al. Method of the GPS tracking analysis for extraction of geometrical properties[C]//2018 IEEE International Conference" Quality management, transport and information security, information technologies"(IT&QM&IS). IEEE, 2018: 266-270.

[2] Nasri A, Zhang L, Fan J, et al. Advanced vehicle miles traveled estimation methods for non-federal aid system roadways using GPS vehicle trajectory data and statistical power analysis[J]. Transportation research record, 2019: 0361198119850790.

[3] Fan J, Fu C, Stewart K, et al. Using big GPS trajectory data analytics for vehicle miles traveled estimation[J]. Transportation research part C: emerging technologies, 2019, 103: 298-307.

[4] Zhang W, Qi Y, Zhou Z, et al. Method of speed data fusion based on Bayesian combination algorithm and high-order multi-variable Markov model[J]. IET Intelligent Transport Systems, 2018, 12(10): 1312-1321.

[5] Zaharia M, Xin R S, Wendell P, et al. Apache spark: a unified engine for big data processing[J]. Communications of the ACM, 2016, 59(11): 56-65.

[6] Armbrust M, Xin R S, Lian C, et al. Spark sql: Relational data processing in spark[C]//Proceedings of the 2015 ACM SIGMOD international conference on management of data. 2015: 1383-1394.

[7] Fan J, Fu C, Stewart K, et al. Using big GPS trajectory data analytics for vehicle miles traveled estimation[J]. Transportation research part C: emerging technologies, 2019, 103: 298-307.

[8] Nasri A, Zhang L, Fan J, et al. Advanced vehicle miles traveled estimation methods for non-federal aid system roadways using GPS vehicle trajectory data and statistical power analysis[J]. Transportation research record, 2019: 0361198119850790.

[9] Boukhechba M, Bouzouane A, Bouchard B, et al. Online recognition of people's activities from raw GPS data: Semantic Trajectory Data Analysis[C]//Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments. 2015: 1-8.

[10] Siła-Nowicka K, Vandrol J, Oshan T, et al. Analysis of human mobility patterns from GPS trajectories and contextual information[J]. International Journal of Geographical Information Science, 2016, 30(5): 881-906.

[11] Wang J, Wang C, Song X, et al. Automatic intersection and traffic rule detection by mining motor-vehicle GPS trajectories[J]. Computers, Environment and Urban Systems, 2017, 64: 19-29.

[12] Zhang D, Lee K, Lee I. Mining hierarchical semantic periodic patterns from GPS-collected spatio-temporal trajectories[J]. Expert Systems with Applications, 2019, 122: 85-101.

[13] Lin Q, Zhang D, Connelly K, et al. Disorientation detection by mining GPS trajectories for cognitively-impaired elders[J]. Pervasive and mobile computing, 2015, 19: 71-85.

[14] Zair S, Le Hégarat-Mascle S, Seignez E. Coupling outlier detection with particle filter for GPS-based localization[C]//2015 IEEE 18th International Conference on Intelligent Transportation Systems. IEEE, 2015: 2518-2524.

[15] Patil V, Singh P, Parikh S, et al. Geosclean: Secure cleaning of gps trajectory data using anomaly detection[C]//2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2018: 166-169.

[16] Lou Y, Zhang C, Zheng Y, et al. Map-matching for low-sampling-rate GPS trajectories[C]//Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems. 2009: 352-361.

[17] Wright R E. Logistic regression[J]. 1995.

[18] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.

[19] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[20] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[21] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[22] De Vries G K D, Van Someren M. Machine learning for vessel trajectories using compression, alignments and domain knowledge[J]. Expert Systems with Applications, 2012, 39(18): 13426-13439.

[23] Cho S B. Exploiting machine learning techniques for location recognition and prediction with smartphone logs[J]. Neurocomputing, 2016, 176: 98-106.

[24] Kim J, Mahmassani H S. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories[J]. Transportation Research Procedia, 2015, 9: 164-184.

[25] Guo D, Zhu X, Jin H, et al. Discovering spatial patterns in origin-destination mobility data[J]. Transactions in GIS, 2012, 16(3): 411-429.

[26] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Kdd. 1996, 96(34): 226-231.

APPENDIX A

PERFORMANCE DETAILS FOR THE MODELS MENTIONED IN THE THESIS

Table 4 4-class classification, baseline (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.5079	0.8027	0.5443	0.7259	0.6632
	Recall	0.6095	0.8872	0.3333	0.8167	
SVM	Precision	0.7083	0.8571	0.5625	0.9000	0.7495
	Recall	0.6476	0.9474	0.6279	0.7500	
RF	Precision	0.8000	0.9466	0.7163	0.8957	0.8378
	Recall	0.7619	0.9323	0.7829	0.8583	
XGBoost	Precision	0.7387	0.9462	0.7222	0.8500	0.8172
	Recall	0.7619	0.9474	0.7519	0.7917	

Table 5 4-class classification, baseline + OD cluster statistics (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.6718	0.9474	0.7071	0.7903	0.7844
	Recall	0.8381	0.9474	0.5426	0.8167	
SVM	Precision	0.6639	0.8921	0.6260	0.8774	0.7659
	Recall	0.7524	0.9323	0.5969	0.7750	
RF	Precision	0.7909	0.9545	0.7903	0.8760	0.8604
	Recall	0.8286	0.9474	0.7597	0.8833	
XGBoost	Precision	0.7500	0.9403	0.7750	0.8974	0.8439
	Recall	0.8286	0.9474	0.7209	0.8750	

Table 6 **4-class classification, baseline + trajectory sample quality statistics (1948 trained, 487 tested).**

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.8021	0.9562	0.6875	0.8254	0.8214
	Recall	0.7333	0.9850	0.6822	0.8667	
SVM	Precision	0.8068	0.9357	0.6439	0.7717	0.7906
	Recall	0.6762	0.9850	0.6589	0.8167	
RF	Precision	0.8600	0.9924	0.7891	0.8740	0.8809
	Recall	0.8190	0.9850,	0.7829	0.925	
XGBoost	Precision	0.8131	0.9848	0.7717	0.8926	0.8686
	Recall	0.8286	0.9774	0.7597	0.9000	

Table 7 4-class classification, baseline + temporal statistics (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.6612	0.9137	0.6465	0.7578	0.7556
	Recall	0.7619	0.9549	0.4961	0.8083	
SVM	Precision	0.7353,	0.8888	0.6402	0.8627	0.7803
	Recall	0.7143	0.9624,	0.6899	0.7333	
RF	Precision	0.8224	0.9701	0.7481	0.9099	0.8624
	Recall	0.8381	0.9774	0.7829	0.8417	
XGBoost	Precision	0.7736	0.9545	0.7440	0.8629	0.8378
	Recall	0.7905	0.9549	0.7674	0.8417	

Table 8 4-class classification, baseline + geographical information statistics (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.6723	0.9070	0.6364	0.7881	0.7536
	Recall	0.6723	0.9512	0.6016	0.7949	
SVM	Precision	0.7283	0.8276	0.6563	0.7541	0.7454
	Recall	0.6723	0.9512	0.6016	0.7949	
RF	Precision	0.8333	0.9291	0.6954	0.8850	0.8275
	Recall	0.6723	0.9593	0.8203	0.8547	
XGBoost	Precision	0.7589	0.9597	0.7059	0.8870	0.8255
	Recall	0.7143	0.9675	0.7500	0.8718	

Table 9 4-class classification, baseline + OD cluster + temporal + geographical information statistics (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.7417	0.9297	0.7232	0.7953	0.8008
	Recall	0.7479	0.9675	0.6328	0.8632	
SVM	Precision	0.6818	0.9015	0.6538	0.8087	0.7639
	Recall	0.6303	0.9675	0.6641	0.7949	
RF	Precision	0.8969	0.9444	0.7609	0.8571	0.8460
	Recall	0.7311	0.9675	0.8203	0.9231	
XGBoost	Precision	0.8198	0.9440	0.7795	0.8790	0.8563
	Recall	0.7647	0.9593	0.7734	0.9316	

Table 10 4-class classification, baseline + trajectory sample quality + OD + temporal + geographical information statistics (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.8750	0.9457	0.6343	0.6447	0.7556
	Recall	0.5294	0.9919	0.6641	0.8376	
SVM	Precision	0.7818	0.9683	0.7209	0.8607	0.8337
	Recall	0.7227	0.9919	0.7266	0.8974	
RF	Precision	0.9468	0.9919	0.8042	0.8976	0.9035
	Recall	0.7479	0.9919	0.8984	0.9744	
XGBoost	Precision	0.8774	0.9762	0.7970	0.9344	0.8953
	Recall	0.7815	1.000	0.8281	0.9744	

Table 11 4-class classification, baseline + OD cluster + temporal + geographical information statistics and corresponding standard deviation (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.7311	0.9449	0.7328	0.8080	0.8070
	Recall	0.7311	0.9756	0.6641	0.8632	
SVM	Precision	0.7551	0.9154	0.6471	0.8585	0.7864
	Recall	0.6218	0.9675	0.7734	0.7778	
RF	Precision	0.8750	0.9603	0.7698	0.8571	0.8624
	Recall	0.7059	0.9837	0.8359	0.9231	
XGBoost	Precision	0.8846	0.9531	0.7883	0.8983	0.8789
	Recall	0.7731	0.9919	0.8438	0.9060	

Table 12 4-class classification, baseline +trajectory sample quality + OD cluster + temporal + geographical information statistics and corresponding standard deviation (1948 trained, 487 tested).

Model		C	TST	FD	T	Micro Average
Logistic Regression	Precision	0.7412	0.9308	0.6597	0.7344	0.7659
	Recall	0.5294	0.9837	0.7422	0.8034	
SVM	Precision	0.7736	0.9606	0.7164	0.8833	0.8337
	Recall	0.6891	0.9919	0.7500	0.9060	
RF	Precision	0.9556	0.9840	0.7931	0.9134	0.9035
	Recall	0.7227	1	0.8984	0.9915	
XGBoost	Precision	0.9074	0.9685	0.8397	0.9339	0.9117
	Recall	0.8235	1	0.8594	0.9658	

Table 13 3-class classification, baseline (57619 trained, 14405 tested).

Model		C	FD	T	Micro Average
Logistic Regression	Precision	0.6995	0.6234	0.7861	0.7054
	Recall	0.7477	0.5785	0.7888	
RF	Precision	0.9092	0.7465	0.8878	0.8392
	Recall	0.7454	0.8669	0.9033	
XGBoost	Precision	0.8897	0.7295	0.8742	0.8219
	Recall	0.7349	0.8594	0.8699	

Table 14 3-class classification, baseline + OD cluster + temporal + geospatial information statistics and corresponding standard deviation (57619 trained, 14405 tested).

Model		C	FD	T	Micro Average
Logistic Regression	Precision	0.7921	0.7040	0.8212	0.7721
	Recall	0.7774	0.7186	0.8194	
RF	Precision	0.8948	0.7857	0.9316	0.8668
	Recall	0.7785	0.8803	0.9397	
XGBoost	Precision	0.8922,	0.8095	0.9433	0.8802
	Recall	0.8153	0.8695	0.9542	

Table 15 3-class classification, baseline + trajectory sample quality + OD cluster + temporal + geospatial information and corresponding standard deviation (57619 trained, 14405 tested)

Model		C	FD	T	Micro Average
Logistic Regression	Precision	0.9013	0.7147	0.8684	0.8199
	Recall	0.7663	0.8090	0.8828	
RF	Precision	0.9637	0.8035	0.9603	0.9007
	Recall	0.7867	0.9343	0.9789	
XGBoost	Precision	0.9341	0.8383	0.9671	0.9109
	Recall	0.8374	0.9111	0.9823	