Overall descriptions:
Number of rows: 10,273,969 rows
Number of columns: 21
Number of trucks: 9 ('粤ABP691', '闽K59769', '粤ADS670', '粤ADW293', '闽K59938', '闽K59936', '粤ADP980', '粤ABW222', '闽K55572')
Collect date: 30 days, December of 2019 except December 5th

rowkey	STRING COMMENT 'rowkey',
value	STRING COMMENT '值',
ems_time	STRING COMMENT 'EMS upload time',
imei	STRING COMMENT 'imei code',
gpsno	INT COMMENT 'gps Number',
truckid	STRING COMMENT 'Vehicle ID',
item_id	INT COMMENT 'item_id',
model	STRING COMMENT ‘model',
truckno	STRING COMMENT ' plate number',
orgcode	STRING COMMENT 'institute ID',
lat	DOUBLE COMMENT 'lat',
lng	DOUBLE COMMENT 'lng',
course	INT COMMENT '360 degrees， 0 is north',
triggertime	TIMESTAMP COMMENT 'timestamp',
province	STRING COMMENT 'Province',
city	STRING COMMENT 'City',
county	STRING COMMENT 'country/district',
address	STRING COMMENT 'address',
r_name	STRING COMMENT 'Road Name， from Tencent‘’
r_level	STRING COMMENT 'road Level， from Tencent,'
date	

ID	Description	单位 unit
x7000	转速 Engine Speed	km/h
x7001	瞬时油耗 Fuel Rate	升/小时L/hr
x7002	总油耗 Total Fuel Consumption	0.5升L
x7003	刹车Brake	1：激活on 0：未激活off
x7004	EMS里程 EMS mileage	0.1公里KM

x7005	累计总里程 Total Mileage	mile
x7006	油门开度百分比 Throttle	%
x7007	冷却液温度 Temperature of coolant	1 摄氏度Celsius；从-100度开始计算。例如：0 表示-100 度，160 表示 60 度，60 表示-40 度
x006C	GPS车速 GPS speed	km/h
x7035	刹车状态（手刹） Brake	0：未激活off　1：激活 on
x000B	当前时间 Time	年月日时分秒
x7091	主油箱剩余油量百分比 Percentage of fuel left	0.1%；范围range：0.0%——100.0%
x70EB	未知 Unknown	

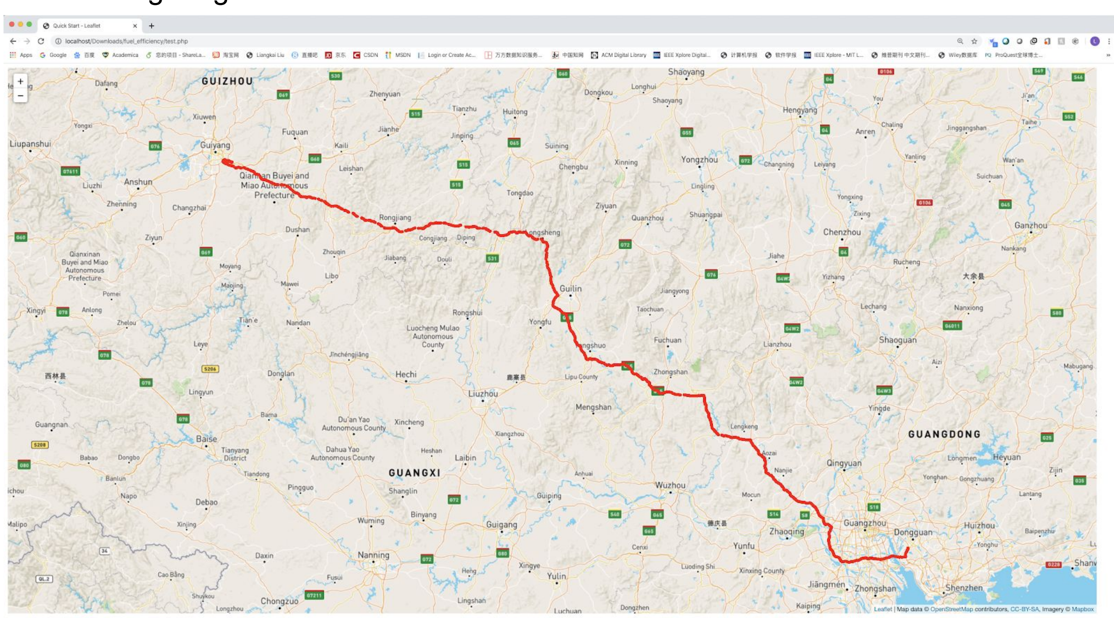
VVID	License Plate	Days	Date
3CCC005122531737E69EA3BA21324ECD	粤ABP691	30	20191201, 20191202, 20191203, 20191204, 20191206, 20191207, 20191208, 20191209, 20191210, 20191211, 20191212, 20191213, 20191214, 20191215, 20191216, 20191217, 20191218, 20191219, 20191220, 20191221, 20191222, 20191223, 20191224, 20191225, 20191226, 20191227, 20191228, 20191229, 20191230, 20191231

A0A4A31F4C3509655240D8D7DB9CD389	闽K59769	30	20191201, 20191202, 20191203, 20191204, 20191206, 20191207, 20191208, 20191209, 20191210, 20191211, 20191212, 20191213, 20191214, 20191215, 20191216, 20191217, 20191218, 20191219, 20191220, 20191221, 20191222, 20191223, 20191224, 20191225, 20191226, 20191227, 20191228, 20191229, 20191230, 20191231
83E94E04A08895767DFE0D80A21A07D3	闽K59938	30	20191201, 20191202, 20191203, 20191204, 20191206, 20191207, 20191208, 20191209, 20191210, 20191211, 20191212, 20191213, 20191214, 20191215, 20191216, 20191217, 20191218, 20191219, 20191220, 20191221, 20191222, 20191223, 20191224, 20191225, 20191226, 20191227, 20191228, 20191229, 20191230, 20191231
B45F36E3944670E22C7EC735D833F709	粤ADW293	5	20191227, 20191228, 20191229, 20191230, 20191231
7571522FF0EBA036818CEACBA52D3B60	粤ADS670	4	20191228, 20191229, 20191230, 20191231
096B3BBA5216C10C7EDF72FD803ACFD7	粤ADP980	25	20191203, 20191204, 20191209, 20191210, 20191211, 20191212, 20191213, 20191214, 20191215, 20191216, 20191217, 20191218, 20191219, 20191220, 20191221, 20191222, 20191223, 20191224, 20191225, 20191226, 20191227, 20191228, 20191229, 20191230, 20191231
3EF3F915B5831AE8667B6FC54FBA89B7	闽K55572	30	20191201, 20191202, 20191203, 20191204, 20191206, 20191207, 20191208, 20191209, 20191210, 20191211, 20191212, 20191213, 20191214, 20191215, 20191216, 20191217, 20191218, 20191219, 20191220, 20191221, 20191222, 20191223, 20191224, 20191225, 20191226, 20191227, 20191228, 20191229, 20191230, 20191231
CABB5C23A2B5E1541DB6E75FD1D61E01	粤ABW222	30	20191201, 20191202, 20191203, 20191204, 20191206, 20191207, 20191208, 20191209, 20191210, 20191211, 20191212, 20191213, 20191214, 20191215, 20191216, 20191217, 20191218, 20191219, 20191220, 20191221, 20191222, 20191223, 20191224, 20191225, 20191226, 20191227, 20191228, 20191229, 20191230, 20191231
257C0D741E2CDDAFDA1A297FC5AC9964	闽K59936	30	20191201, 20191202, 20191203, 20191204, 20191206, 20191207, 20191208, 20191209, 20191210, 20191211, 20191212, 20191213, 20191214, 20191215, 20191216, 20191217, 20191218, 20191219, 20191220, 20191221, 20191222, 20191223, 20191224, 20191225, 20191226, 20191227, 20191228, 20191229, 20191230, 20191231

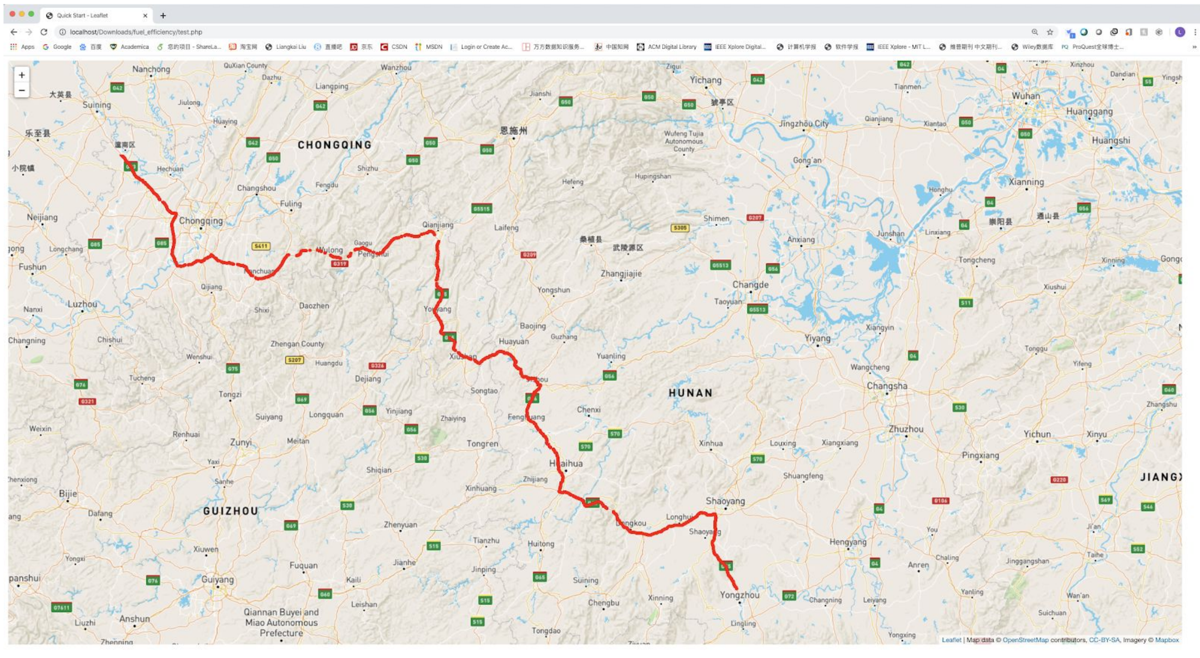
Trajectory of the route:

Truck ID: 83E94E04A08895767DFE0D80A21A07D3
License plate: 闽K59938　　Date: 20191201
From Guangdong to Hunan

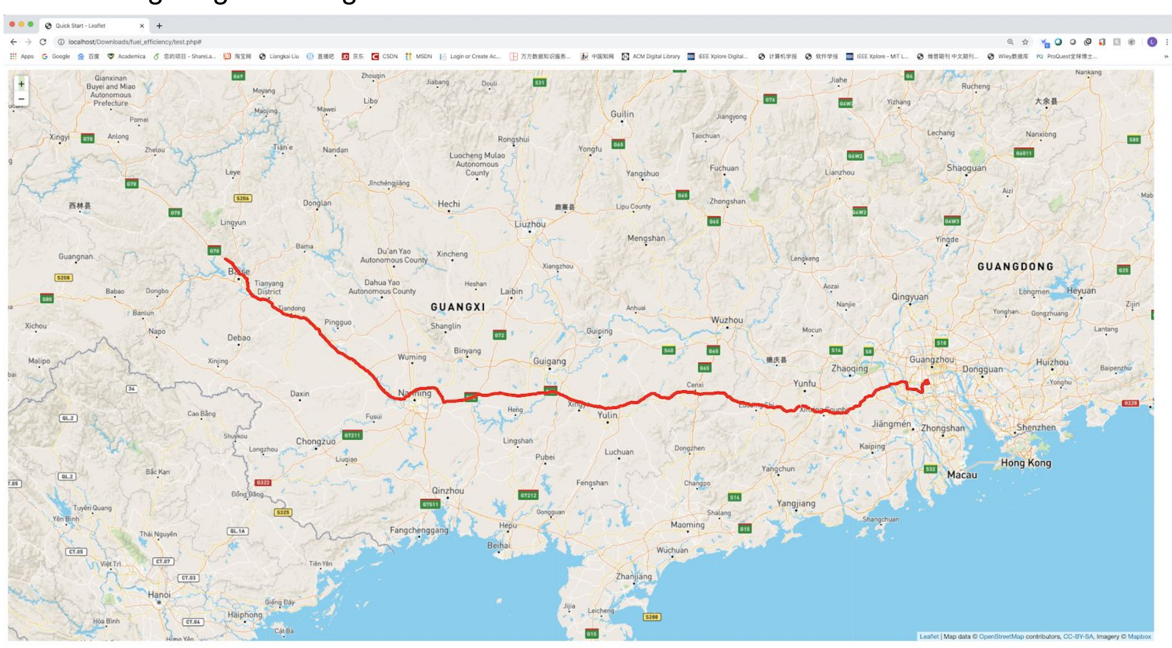
Truck ID: 3CCC005122531737E69EA3BA21324ECD
License plate: 粤ABP691　　Date: 20191201
From Guangdong to Guizhou



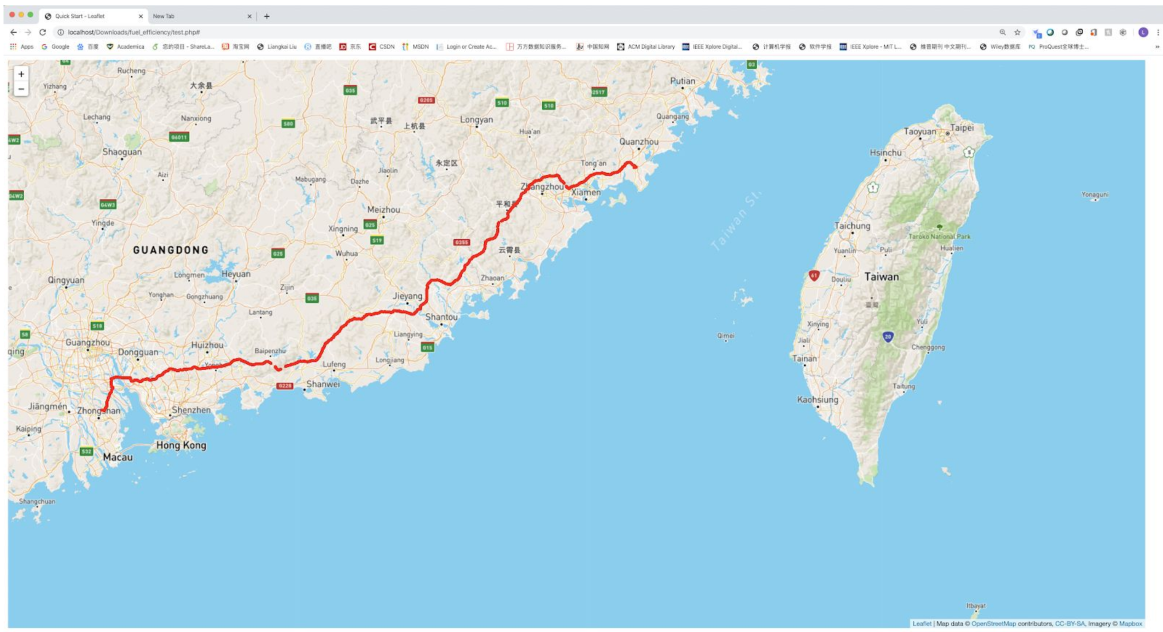
Truck ID: A0A4A31F4C3509655240D8D7DB9CD389
License plate: 闽K59769 Date: 20191201
From Hunan to Chongqing



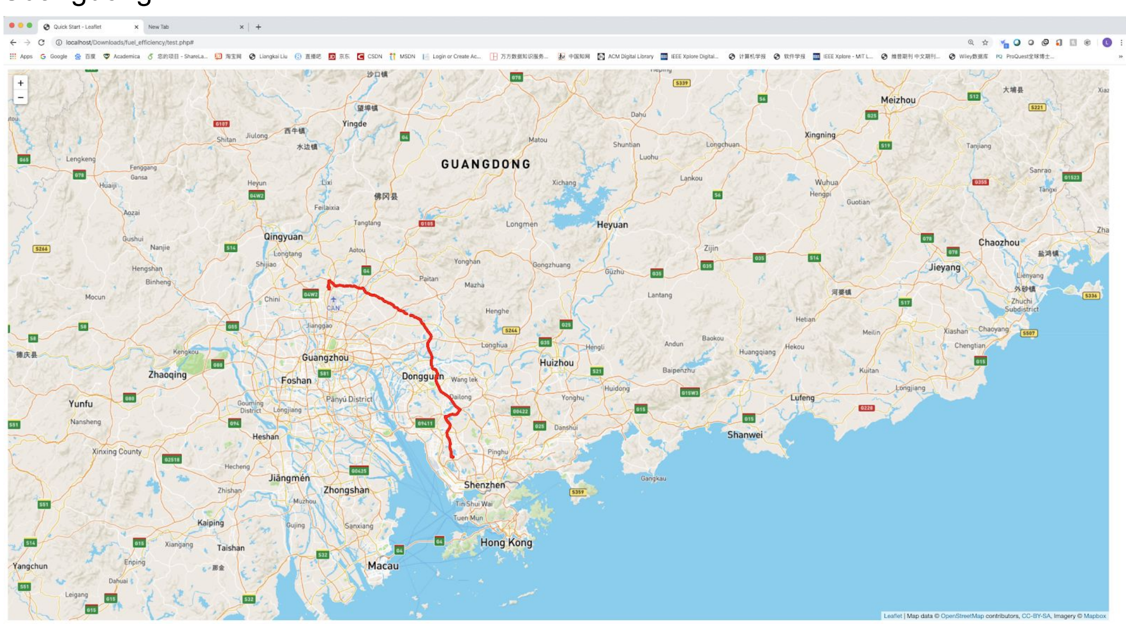
Truck ID: B45F36E3944670E22C7EC735D833F709
License plate: 粤ADW293 Date: 20191228
From Guangdong to Guangxi



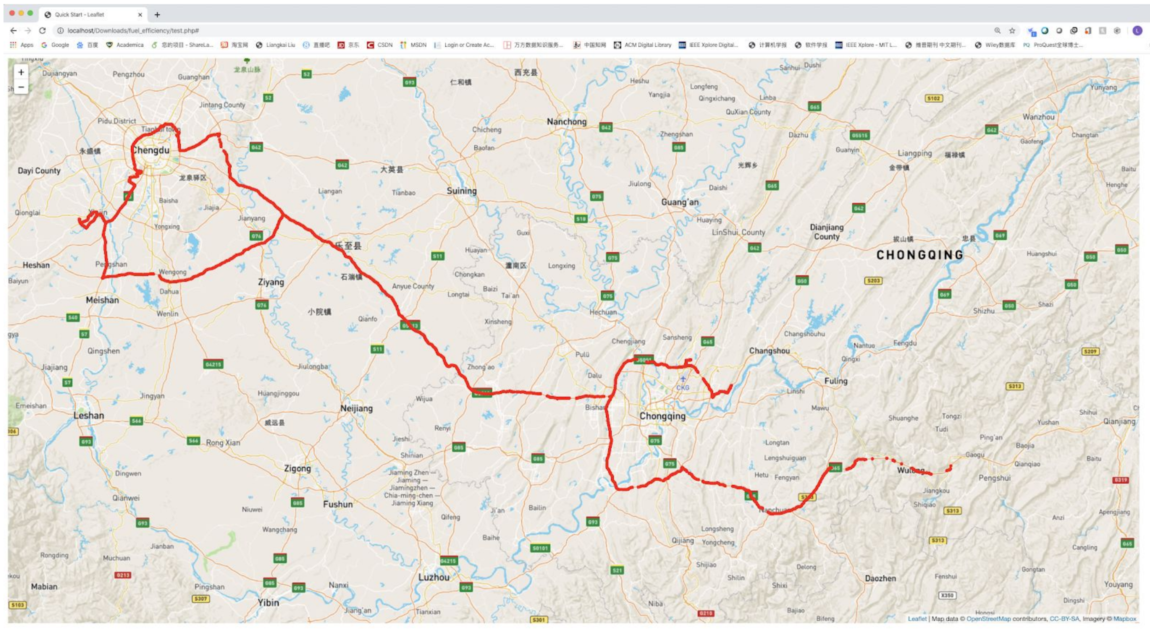
Truck ID: 7571522FF0EBA036818CEACBA52D3B60
License plate: 粤ADS670 Date: 20191230
From Guangdong to Fujian



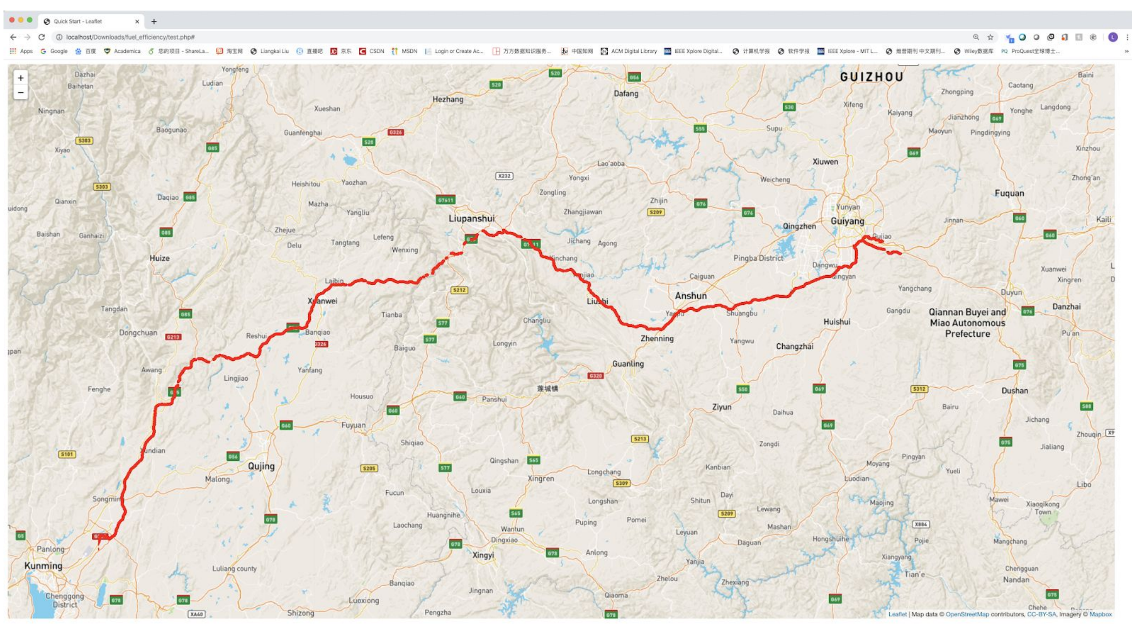
Truck ID: 096B3BBA5216C10C7EDF72FD803ACFD7
License plate: 粤ADP980 Date: 20191212
Guangdong



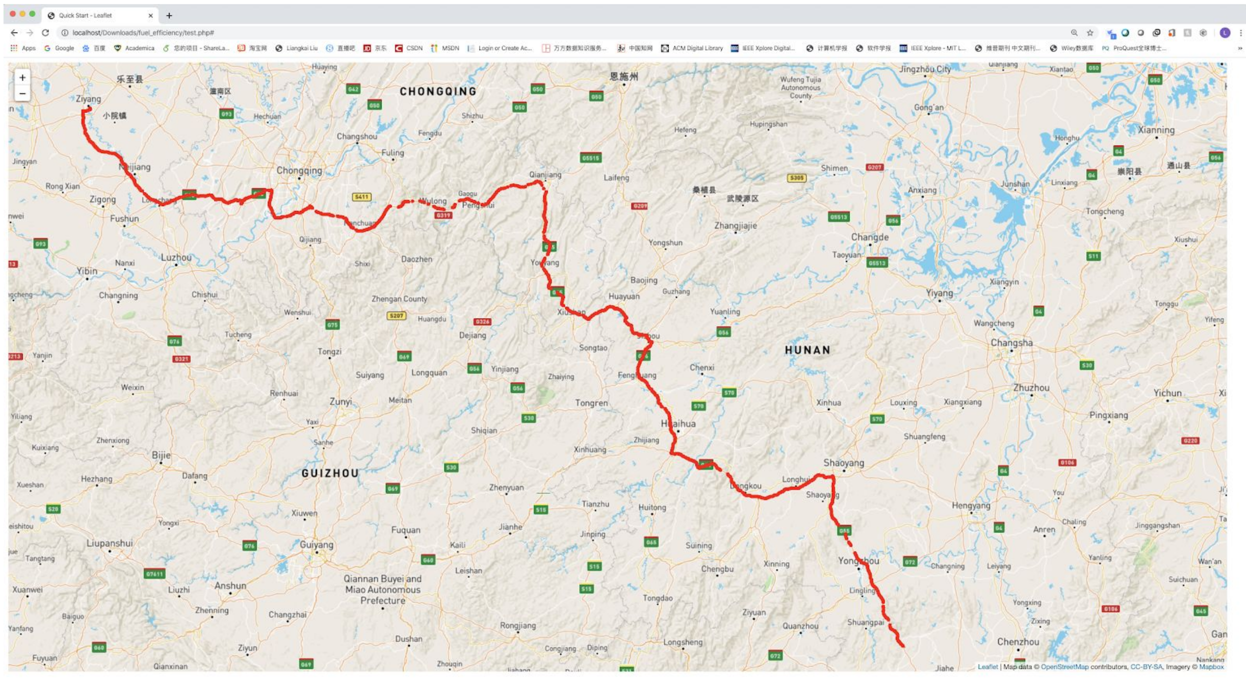
Truck ID: 3EF3F915B5831AE8667B6FC54FBA89B7
License plate: 闽K55572 Date: 20191201
From Chongqing to Sichuan



Truck ID: CABB5C23A2B5E1541DB6E75FD1D61E01
License plate: 粤ABW222 Date: 20191201
From Yunnan to Guizhou



Truck ID: 257C0D741E2CDDAFDA1A297FC5AC9964
License plate: 闽K59936   Date: 20191201
From Guangdong to Sichuan



Questions on the data:

1. The data is not sorted for different vehicle, timestamp, and date
2. Some data metrics are missing, especially the ems data in column 2
3. For some date, the truck is just parked there and not driving (like 20191227 for 粤ADW293, 20191210 for 粤ADP980)

Fuel Efficiency Analysis (*Specific vehicle on a specific day)*
VVID: 257C0D741E2CDDAFDA1A297FC5AC9964
License Plate: 闽K59936
Date: 20191201

Linear regression based approach:

Input: d['x7000'],**d['x7002']**,d['x7003'],d['x7004'],d['x7005'],d['x7006'],d['x7007'],d['x006C'],d['x7035'],d['x7091'], lat, lng
Output: **d['x7001']**
Train: 64252
Test: 10000
Coefficients:
[[3.17465326e-03 -4.83697138e-06 6.42187046e+00 8.13119386e-02
1.87307609e-02 8.09133162e-01 1.70884482e-01 -3.55271368e-15
1.30340511e+01 9.47522911e-02 2.02923432e+00 -2.21454512e+00]]
Mean squared error (MSE): **38.89**
Coefficient of determination (R2): **0.91**

Input: d['x7000'],d['x7003'],d['x7004'],d['x7005'],d['x7006'],d['x7007'],d['x006C'],d['x7035'],d['x7091'], lat, lng
Output: **d['x7001']**
Train: 64252
Test: 10000
Coefficients:
[[3.17400533e-03 6.42252792e+00 8.13104329e-02 1.87990254e-02
8.09141491e-01 1.70914629e-01 1.95399252e-14 1.30347769e+01
9.38992513e-02 2.03293073e+00 -2.22533441e+00]]
Mean squared error (MSE): **38.89**
Coefficient of determination (R2): **0.91**

Input: d['x7000'],d['x7003'],d['x7004'],d['x7005'],d['x7006'],d['x7007'],d['x006C'],d['x7035'],d['x7091'], lat, lng
Output: **d['x7002']**
Train: 64252
Test: 10000
Coefficients:
[8.15209464e-04 -7.58280654e-01 -1.69451224e-02 5.08176126e-01
-7.81788000e-03 8.65023507e-02 -1.37223566e-13 2.20035076e+00
9.48877311e-03 1.18874090e+01 -1.15637862e+01]
Mean squared error: **12.08**
Coefficient of determination: **0.97**

MLP based approach:

Input: d['x7000'],d['x7003'],d['x7004'],d['x7005'],d['x7006'],d['x7007'],d['x006C'],d['x7035'],d['x7091'], lat, lng
Output: d['x7001']
10-fold cross-validation
Epoch: 30
Batch size: 50
Number of instances: 74252

Two layer: [11, 1]
Activation: relu
Optimizer: adam
Without standardization MSE: **7.73** R2: **0.896**
With standardization MSE: **6.16** R2: **0.934**

Three-layer: [11, 6, 1]
Activation: relu
Optimizer: adam
With standardization MSE: **5.91** R2: **0.948**

+++

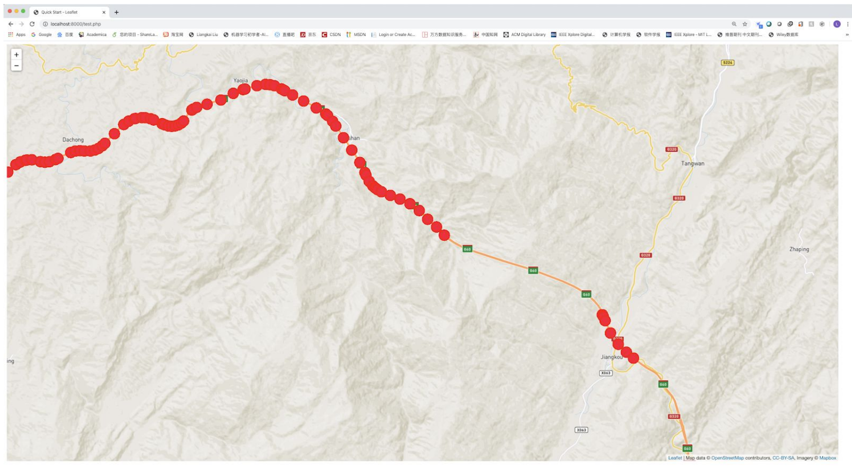
Input: d['x7000'],d['x7003'],d['x7004'],d['x7005'],d['x7006'],d['x7007'],d['x006C'],d['x7035'],d['x7091'], lat, lng
Output: d['x7002']
10-fold cross-validation
Epoch: 30
Batch size: 50
Number of instances: 74252

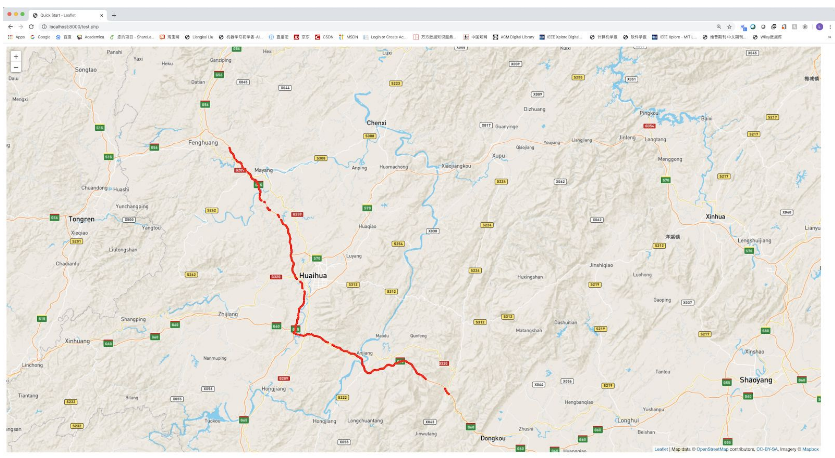
Two layer: [11, 1]
Activation: relu
Optimizer: adam
Without standardization MSE: **14.015** R2: **-0.150**
With standardization MSE: **14603.62** R2: **- 2146518.09**

Three-layer: [11, 6, 1]
Activation: relu
Optimizer: adam
With standardization MSE: **867.84** R2: **-4725.70**

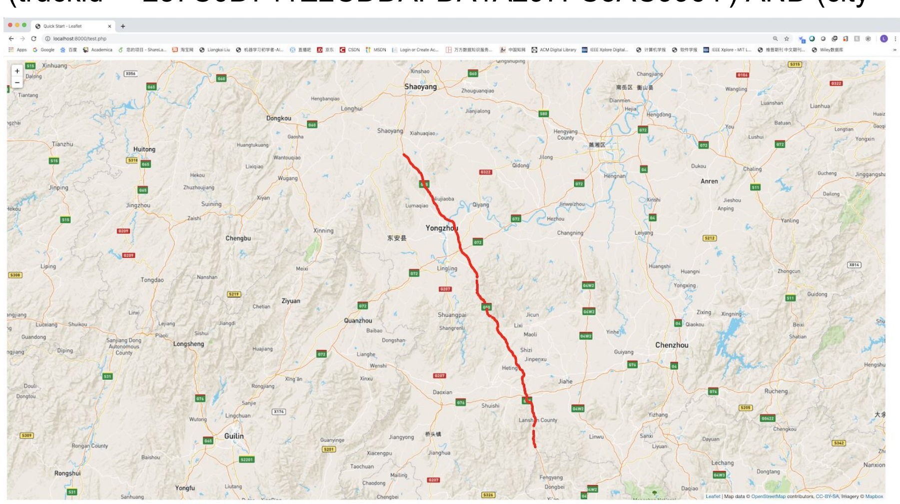
LSTM-based approach:

To be added





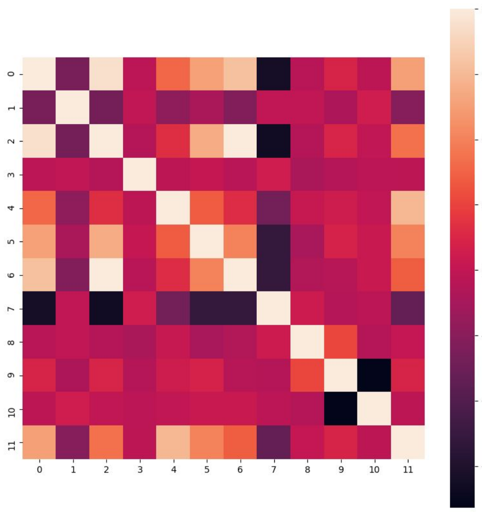
(truckid = '257C0D741E2CDDAFDA1A297FC5AC9964') AND (city = '永州市')



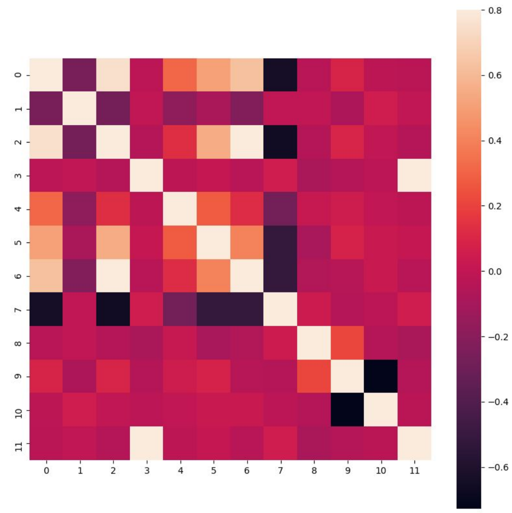
Correlation analysis

1E01 all the data

0-10: 'x7000', 'x7003', 'x7004', 'x7005', ''x7006', 'x7007', 'x006C', 'x7035', 'x7091', lat, lng
11: x7001 or x7002

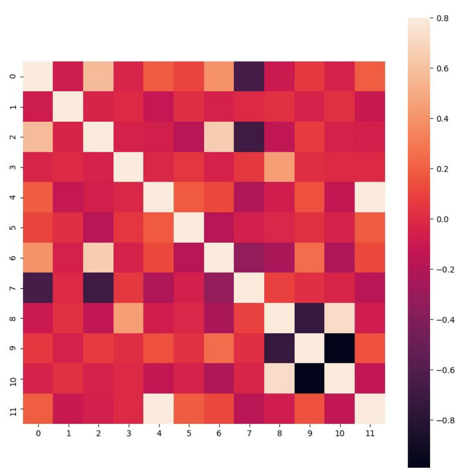


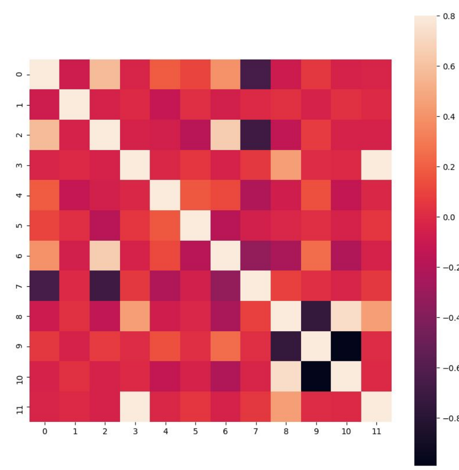
x7001 (fuel rate)



x7002 (Total fuel consumption)

9964 '永州市' same routes





Same Route Analysis:

Route matching algorithm:

1. Merge data instance together based on city and vvid to get the date group;
2. For each date in the **date** group, calculate the total milerages for that day and get the start and end points' position (latitude, longitude);
3. Compare the start and end point of different routes to determine the direction. Compare the mileage of different routes to get **the most frequent length** of the routes L_1. The routes with length L_1 and the same direction will be grouped together. There will be two groups: each direction has one;
4. For each group, the route can be represented as a (2,X), where X is the number of instances for each route. Then do the **sampling** on each route to make each route can be represented as a **fixed length** two-dimensional matrix;
5. Calculate the **distance matrix** of the routes (difference of the fixed-length two-dimensional matrix). If the sum of the absolute values of the matrix is less than a predefined threshold, then these two routes are the same route.

Two examples: (blue as one group of same routes, red as another)

1E01 桂林市

	Start Position	End Position	Milerage	Instances
20191202	25.852 109.743	24.517 111.021	259.2	11339
20191204	24.517 111.02	25.852 109.743	258.6	10740
20191206	25.851 109.743	24.517 111.021	386.7	11769
20191208	24.517 111.02	25.851 109.743	258.5	10981
20191211	25.852 109.742	24.517 111.02	259.7	11328
20191213	24.517 111.02	25.853 109.742	259.7	11732
20191217	24.517 111.02	25.853 109.742	259.0	12010
20191219	25.852 109.742	24.517 111.02	259.5	11326
20191220	24.517 111.021	25.852 109.743	259.0	10684
20191226	24.517 111.02	25.853 109.743	259.0	9958
20191228	25.852 109.742	24.517 111.02	259.8	10741
20191229	24.517 111.02	25.851 109.744	258.6	10192

9964 永州市

	Start Position	End Position	Mileage	Instances
20191201	26.861 111.355	25.487 112.12	187.4	10186
20191202	25.487 112.12	25.222 112.175	33.8	3314
20191204	25.22 112.175	26.862 111.354	227.6	10095
20191207	26.86 111.355	25.22 112.175	221.6	11859
20191209	25.22 112.175	26.861 111.355	227.2	10740
20191213	26.862 111.354	25.221 112.175	221.7	13130
20191214	25.223 112.174	26.12 111.813	125.4	5772
20191217	26.856 111.359	25.257 112.166	219.8	11015
20191219	25.22 112.175	26.861 111.355	227.7	11397
20191221	26.86 111.355	26.45 111.672	60.4	2944
20191222	26.45 111.672	25.221 112.175	161.1	9174
20191224	25.22 112.175	26.86 111.356	227.7	10410
20191228	25.283 112.167	26.862 111.354	227.5	10530
20191231	26.862 111.354	26.455 111.671	59.4	2879

Fuel Prediction

***Dataset:* 9964 '永州市' same routes**
Number of rows: 53171

Based on the correlation analysis results, we divide the features into three feature groups.

x7001 (fuel rate)
Feature groups:
F: 'x7006'
N: 'x7000', 'x7006'
S: 'x7000', 'x7006', 'x006C'
T: 'x7000', 'x7006', 'x006C', 'lat', 'lng'
For LSTM and CNN-LSTM, the input includes the history 'x7001', size is 10

5 fold cross-validation R2 value:

R2 value	LR	PR (N=5)	MLP	LSTM	CNN	CNN-LSTM
F	0.92		0.925	0.92	0.924	0.915
N	0.884	0.951	0.952	0.901	0.951	0.91
S	0.91		0.951	0.91	0.950	0.913
T	0.91		0.929	0.905	0.942	0.913

MLP network:
Model: "sequential_5"

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_33 (Dense)             (None, 100)               700
_________________________________________________________________
dense_34 (Dense)             (None, 100)               10100
_________________________________________________________________
dense_35 (Dense)             (None, 100)               10100
_________________________________________________________________
dense_36 (Dense)             (None, 100)               10100
_________________________________________________________________
```

dense_37 (Dense)	(None, 50)	5050
dense_38 (Dense)	(None, 50)	2550
dense_39 (Dense)	(None, 50)	2550
dense_40 (Dense)	(None, 1)	51

Total params: 41,201
Trainable params: 41,201
Non-trainable params: 0

LSTM:
Model: "sequential_5"

Layer (type)	Output Shape	Param #
lstm_9 (LSTM)	(None, 4, 100)	44400
lstm_10 (LSTM)	(None, 50)	30200
dense_5 (Dense)	(None, 1)	51

Total params: 74,651
Trainable params: 74,651
Non-trainable params: 0

CNN:
Model: "sequential_5"

Layer (type)	Output Shape	Param #
dense_33 (Dense)	(None, 100)	200
dense_34 (Dense)	(None, 100)	10100
dense_35 (Dense)	(None, 100)	10100
dense_36 (Dense)	(None, 100)	10100
reshape_5 (Reshape)	(None, 5, 5, 4)	0
time_distributed_13 (TimeDis	(None, 5, 2, 128)	2176

```
_________________________________________________________________
time_distributed_14 (TimeDis (None, 5, 1, 128)       0
_________________________________________________________________
time_distributed_15 (TimeDis (None, 5, 128)          0
_________________________________________________________________
flatten_10 (Flatten)     (None, 640)              0
_________________________________________________________________
dense_37 (Dense)              (None, 50)              32050
_________________________________________________________________
dense_38 (Dense)              (None, 50)              2550
_________________________________________________________________
dense_39 (Dense)              (None, 50)              2550
_________________________________________________________________
dense_40 (Dense)              (None, 1)               51
=================================================================
Total params: 69,877
Trainable params: 69,877
Non-trainable params: 0
```

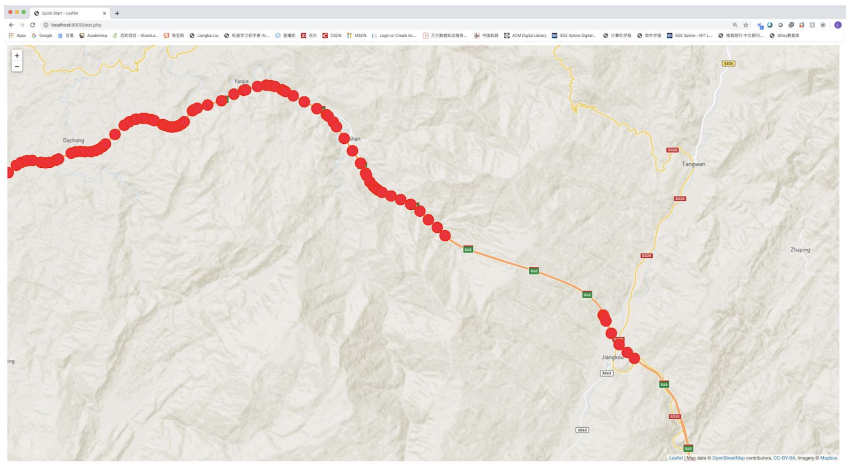
CNN-LSTM:

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)              Output Shape            Param #
=================================================================
time_distributed_10 (TimeDis (None, 6, 2, 128)       1152
_________________________________________________________________
time_distributed_11 (TimeDis (None, 6, 1, 128)       0
_________________________________________________________________
time_distributed_12 (TimeDis (None, 6, 128)          0
_________________________________________________________________
dense_16 (Dense)              (None, 6, 200)          25800
_________________________________________________________________
dense_17 (Dense)              (None, 6, 100)          20100
_________________________________________________________________
dense_18 (Dense)              (None, 6, 100)          10100
_________________________________________________________________
dense_19 (Dense)              (None, 6, 50)           5050
_________________________________________________________________
flatten_8 (Flatten)      (None, 300)              0
_________________________________________________________________
dense_20 (Dense)              (None, 1)               301
=================================================================
Total params: 62,503
```

Trainable params: 62,503
Non-trainable params: 0

Question:

1. There are some gaps in GPS data, maybe when the truck is going through the tunnel. EMS data is generated every second.



2. MLP shows better performance than LSTM, the highest R2 is 0.951, which is still far from the objective. Advice for optimization?

R2 value	**LR**	**PR (N=5)**	**MLP**	**LSTM**	**CNN**	**CNN-LSTM**
F	0.92	-	0.925	0.92	0.924	0.915
N	**0.884**	**0.951**	**0.952**	**0.901**	**0.951**	**0.91**
N-PR	**-**	**0.951**	**0.951**	**0.90**	**0.950**	**0.950**
S	0.91	-	0.951	0.91	0.950	0.913
T	0.91	-	0.929	0.905	0.942	0.913