15. An overview of telematics data

Guangyuan Gao

School of Statistics, Renmin University of China

Table of Contents

1. Data structure
   - Overview
   - Detailed explanation
2. A trip of a car
   - The distance and duration
   - The trajectory
   - Other information
3. The v-a heatmaps
   - Partition of v-a rectangle
   - Empirical discrete distribution on R
   - Design matrix

The vehicle mobility features are recorded from the start of engine to the shut down of engine, including:

- Bitmask data: Field_Mask
- Identification data: Device_ID; Detected_VIN
- Time data: Trip_Number;Time_Stamp
- GPS data: GPS_Latitude; GPS_Longitude; GPS_Heading; GPS_Speed; Positional_Quality
- Vehicle sensor data: VSS_Speed; Engine_RPM; Accel_Lateral; Accel_Longitudinal; Accel_Vertical

Bitmask data: Field_Mask

The bitmask in hexadecimal indicates the validity of ten fields: GPS_Latitude, GPS_Longitude, GPS_Heading, GPS_Speed, Positional_Quality,VSS_Speed, Engine_RPM, Accel_Lateral, Accel_Longitudinal, Accel_Vertical.

- 3FF in hexadecimal is 1111111111 in binary, indicating all the 10 fields have valid values.
  - This indicates a typical driving status.
- 1F in hexadecimal is 0000011111 in binary, indicating that the first five fields, from GPS_Latitude to Positional_Quality, have missing values.
  - This may indicate the status of the beginning of trip when the car is parked underground and the signals from the satellites are not received.

Identification data

1. Device_ID: Uniquely identifies the vehicle, similar to the vehicle identification number (VIN).
2. Detected_VIN: Vehicle identification number. If not detected or partially detected, the VIN is recorded as UNK.

Time data: Trip_Number

- The Coordinated Universal Time (UTC) of the beginning of the trip.
- The UTC is the time (in seconds) from 00:00:00, January 1, 1970.
- One can transfer UTC to Beijing time via a online tool at `http://tool.chinaz.com/Tools/unixtime.aspx`.
- Or use the R function

```
as.POSIXlt(UTC, origin=''1970-01-01'', tz=''Asia/Shanghai'')
```

Time data: Time_Stamp

- The UTC of each record.
- Time_Stamp is increasing by one, since the vehicle status is recorded every second.
- The time data are more or less related to the likelihood of an accident.
- For example, driving at midnight increases the claim severity but decreases the claim frequency.

GPS data

- GPS_Latitude:
  - Global position system latitude in decimal degrees, multiplied by 10^7.
  - The range of this record is $(-90 \times 10^7, 90 \times 10^7)$. Positive values indicate a location in the Northern Hemisphere.
  - All GPS_Latitude values are positive since all the cars recorded are in China.
- GPS_Longitude:
  - Global position system longitude in decimal degrees, multiplied by 10^6.
  - The range of this record is $(-180 \times 10^6, 180 \times 10^6)$.
  - One can locate the car in the Baidu map via a online tool at `http://api.map.baidu.com/lbsapi/getpoint/`.

GPS data

- GPS_Heading: The angle between the north and the vehicle heading in decimal degrees, multiplied by 10^2. The range of this record is $(0, 360 \times 10^2)$.
- GPS_Speed: The speed of GPS in km/h, multiplied by 10. This is the instantaneous velocity. It is closely related to driving style.
- Positional_Quality: The quality of GPS location. 1 indicates validity and 0 indicates invalidity.
  - For example, if this field is 0, then GPS longitude, latitude, heading, and speed should have missing value (recorded as zeros).

Vehicle sensor data

- VSS_Speed: Vehicle speed sensor speed in km/h, multiplied by 10. It should be close to GPS_Speed.
  - VSS_Speed can be recorded even without signals from GPS satellites.
  - If one drives in a tunnel, then GPS_Speed may be missed but VSS_Speed should be recorded.
- Engine_RPM: Engine revolutions per minute. From the viewpoint of insurers, this field is rarely related to the pricing.

Vehicle sensor data

- Accel_Lateral: The acceleration rate in the direction perpendicular to the vehicle heading and parallel to the road surface in m/s^2, multiplied by 10.
  - This field is closely related to the likelihood of an accident. Abruptly changing lanes is a major cause of pileup and scratch.
- Accel_Longitudinal: The acceleration rate in the direction parallel to the vehicle heading in m/s^2, multiplied by 10.
  - Similar to Accel_Lateral, this field is closely related to the likelihood of an accident.
- Accel_Vertical: The acceleration rate in the direction perpendicular to the road surface in m/s^2, multiplied by 10.
  - This record is comparable to the gravitational acceleration rate, 9.8 m/s^2.

- We import the telematics data of a trip recorded by Device 863158020753697.
- The trip beginning time is 1441148723 UTC or 2015-09-01 23:05:23 GMT or 2015-09-02 07:05:23 CST, calculated from Trip_Number
- As shown in Figure 1,
    - At the beginning, car is parked underground and there is no GPS signal.
    - During the trip, both GPS data and VSS data are collected.
    - At the end of trip, VSS data are missed.

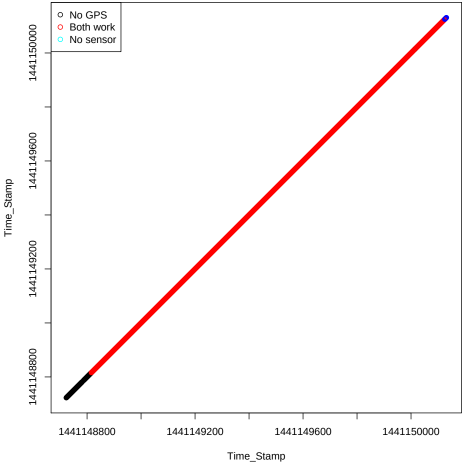

Figure 1: A trip

- Distance: transfer degree measure to radian measure, then calculate the spherical distance as 10 km.
- Duration: using Time_Stamp, the duration can be determined as 23 minutes.

- Transfer GPS coordinates to geodetic coordinates.
  `geoXY(GPS_Latitude, GPS_Longitude,unit=1000)`
- The trajectory is shown in Figure 2. According to the GPS headings in Figure 3, the car was traveling from northwest.
- The speeds in the trajectory can be viewed via a 3D plot. `plot3d`.
- Combining GPS data with time data, we might know the road condition at that moment, which might be a risk factor.

The distance is roughly 10 km from eyeballing the following figure.

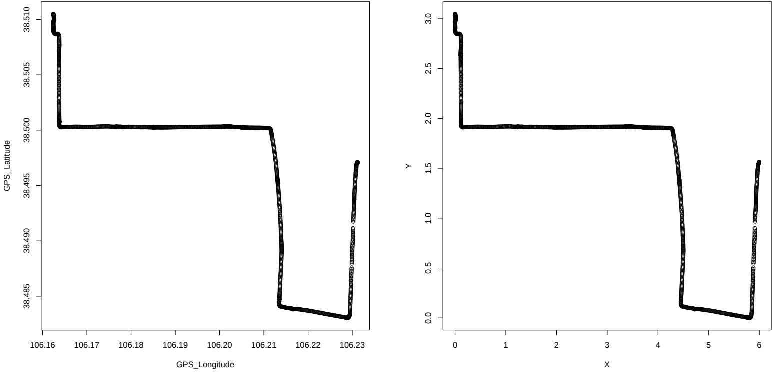

Figure 2: The trajectory in GPS and geodetic coordinates

GPS Headings

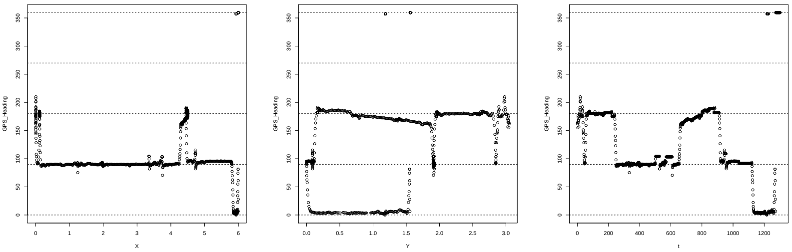

Figure 3: The GPS headings vs x or y or time

Cumulative distance and speed

The derivative of the line in the first graph is the speed.

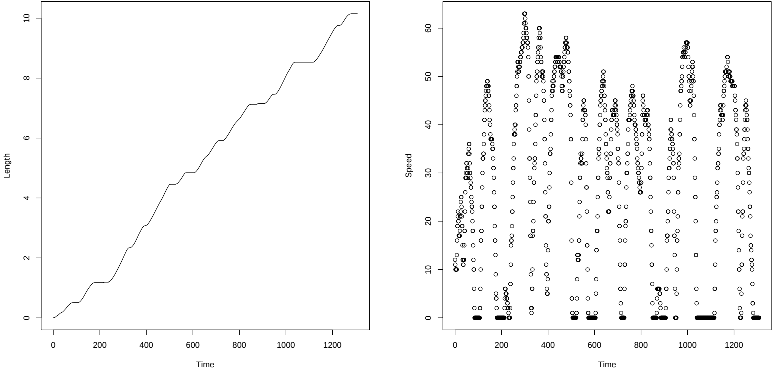

Figure 4: Cumulative distance and speed vs time

VSS_Speed v.s. GPS_Speed

According to the previous analysis, VSS speed is more reliable than GPS speed.

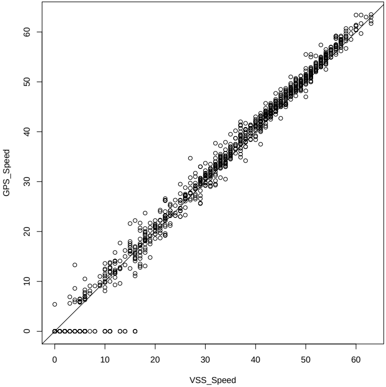

Figure 5: The relationship between VSS speed and GPS speed

- The telematics data is available from 01/01/2014 to 29/06/2017.
- The data is roughly 1 GB per day, which amounts totally in 1.2 TB of data over the whole observation period.
- For computing time consideration, we use the data from 01/05/2016 to 31/07/2016.
- We use v-a heatmap to visualize the speed and acceleration rate data.
- v-a heatmap also compresses the original data dramatically.

- We consider the low speed bucket [5,20] km/h.
- The vehicle sensor speed (VSS) only takes values in integers.
- We partition the v-a rectangle $R = [5, 20] \times [-2, 2]$ by dividing the v-axis (speed) into 16 intervals and the a-axis (acceleration) into 20 intervals.
- The resulting sub-rectangles are denoted by $(R_j)_{j=1:J}$ with $J = 320$, these are illustrated in Figure 6.

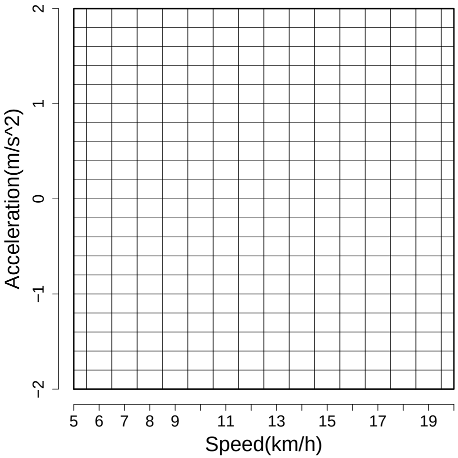

Figure 6: The considered partition of sub-rectangles $(R_j)_{j=1:J}$ of $R = [5, 20] \times [-2, 2]$

- For each car driver $i = 1, \ldots, n$, we denote the relative amount of time spent in sub-rectangle $R_j \subset R$ by $x_{i,j} \geq 0$.
- The induced empirical discrete distribution of driver i on the v-a rectangle R is denoted by $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,J})'$, which lies in the $(J-1)$-unit simplex $\mathcal{X} \subset \mathbb{R}^J$, i.e. has normalization $\sum_{j=1}^J x_{i,j} = 1$.
- Every car driver $i = 1, \ldots, n$ is characterized by a discrete distribution $\boldsymbol{x}_i \in \mathcal{X}$; and $\mathcal{X}$ represents all possible car driver's discrete distributions on $\bigcup_{j=1}^J R_j = R = [5, 20] \times [-2, 2]$.

- In Figure 7, we plot the resulting v-a heatmaps $\boldsymbol{x}_i \in \mathcal{X}$ of the selected car drivers $i = 72, 608$ and 718.
- Driver 72 tends to accelerate and brake less frequently than the other two drivers, while driver 718 tends to accelerate and brake most frequently among the three drivers.

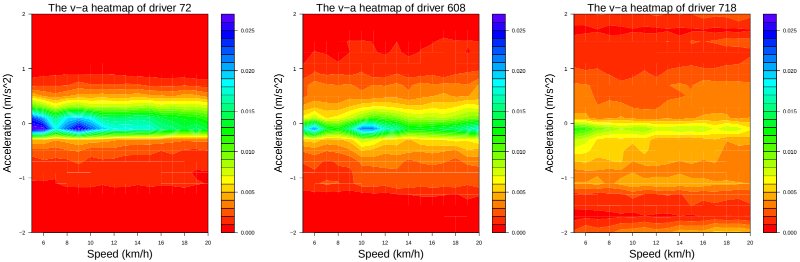

Figure 7: The v-a heatmaps $\boldsymbol{x}_i \in \mathcal{X}$ of the selected car drivers $i = 72, 608$ and 718.

- The v-a heatmaps describe a driver's driving habit and is a summary of high frequency GPS location data.
- We denote by $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times J}$ the $n \times J$ design matrix that contains $\boldsymbol{x}_i$ of all $n = 1,478$ car drivers.
- Directly using the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times J}$ as covariates in the claims frequency regression model would lead to over-parametrization (and over-fitting).

Open questions (will not be graded).

- List three points with regard to the value of telematics data to insurance companies, Baidu map, or other related companies.
- List three ways to reduce the dimension of the design matrix.