SOA Research INSTITUTE

Telematics Car Driving Data Analytics

August 2021

Telematics Car Driving Data Analytics

AUTHOR

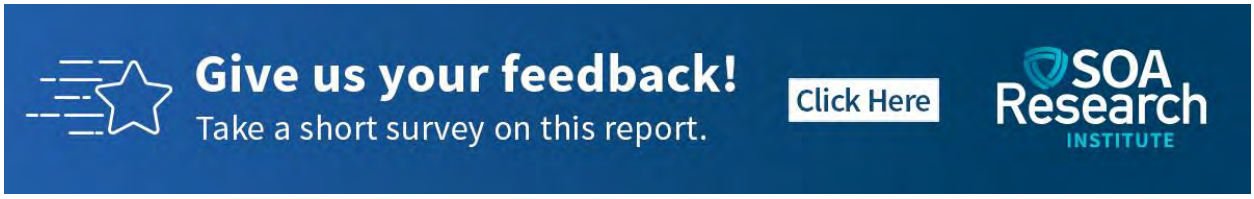
Guangyuan Gao
Center for Applied Statistics and School of Statistics, Renmin University of China, 100872 Beijing, China

Yanlin Shi
Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, Australia

He Wang
School of Finance, Southern University of Science and Technology,
Shenzhen, China.

SPONSOR

General Insurance Research Committee



CONTENTS

Telematics Car Driving Data Analytics

Executive Summary

The aim of this project has three aspects: (1) providing a comprehensive literature review on telematics car driving data analytics; (2) developing various visual tools and a data cleaning procedure for telematics data; and (3) extracting risk factors from individual trips and improving classic classification ratemaking models with those risk factors. In Section 1, we refer to literature in various fields including insurance, actuarial science, computer science, transportation, and machine learning. In Section 2, we firstly develop various visual tools, then impute missing telematics data and select among telematics variables, finally construct three formats of telematics data, time series, summary statistics and heatmaps. In Section 3, a one-dimensional convolutional neuron network is calibrated on binary labeled trips of selected archetypal drivers, which is then used to evaluate risk score of individual trips. By incorporating average risk score for each driver, the classic Poisson generalized linear model for claims frequency can be improved significantly.

Keywords: Claims frequency modelling, Telematics car driving data, Generalized linear model, One-dimensional convolutional neural network, Logistic regression.

Section 1: Introduction

With the advancements of telematics technology, insurers can collect detailed car driving information. Traditionally, insurers use classical actuarial risk factors to classify policyholders and determine premiums for different risk groups. Telematics car driving data are personalized, which encode differences in driving behavior not included in classical social-demographic risk factors, such as regions and ages. For example, some young drivers may have a cautious driving style, while some matured drivers may still demonstrate a wild and aggressive driving behavior. This work is among the first to explore how such information can be learned from telematics car driving data. Our ultimate goal is to improve the predictive accuracy of risk classification models by utilizing telematics car driving data.

Although there are concerns on using personalized data and algorithmic prediction (Mahapatra, 2019; Cevolini and Esposito, 2020), we prefer the statement by Cather (2020) that incorporating telematics data into auto insurance risk classification systems would "minimize insurance discrimination and increase cream skimming adverse selection", where asymmetric information favouring telematics-based insurers supports premium discounts that attract safer drivers, prompting an underpopulation of low-risk drivers among non-telematics insurers.

According to the actuarial literature, among others, there are three major approaches on learning *driving styles* from telematics car driving data: (a) Weidner et al. (2016, 2017) extract covariates from time series of telematics data using discrete Fourier transforms; (b) Huang and Meng (2019), Paefgen et al. (2014) and Verbelen et al. (2018) calculate and adopt summary statistics of telematics data; (c) Gao et al. (2020) apply convolutional neural networks to learn patterns in speed-acceleration heatmaps. Another stream of literature explores *risk exposure* of driving distances or driving duration; see Ayuso et al. (2016a,b), Boucher et al. (2017), and Lemaire et al. (2016).

Since telematics data are posterior experiences, Denuit et al. (2019) propose a credibility model to incorporate the posterior information of driving behavior. Guillen et al. (2020) study the association of telematics car driving data with near-miss events such as cornering, braking, and accelerating. Geyer et al. (2020) explores the effect of driving behavior on risk and insurance selection. Richman (2020) discusses and compares potential approaches to analyze telematics car driving data.

Our data are collected by on-board diagnostics (OBD) systems rather than smartphones. There are several works on smartphone-based telematics in the IEEE Transactions journals. Wahlström et al. (2017) review research on smartphone-based vehicle telematics such as vehicle navigation, driver classification, and road condition monitoring. Wahlström et al. (2018) study the fusion of OBD and global navigation satellite system (GNSS) measurements of speed. Wahlström et al. (2015) detect dangerous cornering using GNSS data. There are also a wide range of relevant studies on the IEEE conference papers. For instance, Savelonas et al. (2020), Girma et al. (2019) and Carvalho et al. (2020) employ recurrent neural networks to identify drivers or to learn different driving behaviors such as normal, moderate, aggressive, etc.

In transportation field, Joubert et al. (2016), Ma et al. (2018), and Hu et al. (2019) study driving behaviors and find certain driving behaviors are more related to at-fault accidents. They consider both telematics data and contextual data such as road conditions, traffic flow, speed limits. Ho et al. (2014), Hung et al. (2007), Kamble et al. (2009) adopt telematics data and driving cycles to understand vehicular emissions, energy consumption and impacts on traffic in different cities around the world.

Among existing literature, two telematics car driving data sets are publicly available: naturalistic driving study dataset and UAH-DriveSet (Romera, 2016). Some of the above literature is based on those two data sets. Gao and Wüthrich (2018) propose a v-a heatmap simulation machine. Our telematics data are privately owned and thus not publicly available.

In this project, we employ a one-dimensional convolutional neuron network (1D CNN) to classify trips into either potential risky or potential safe. 1D CNN can learn representations along time series of car driving trips. We demonstrate that those representations learned from such a classification task are related to claims frequency.

Convolutional neural network is one of the most critical deep learning algorithms. It is capable of learning representation and conducting shift-invariant classification. From 1980s to 1990s, LeCun invented the notion of convolutional network in his epoch-making LeNet5, which used convolution to extract similar spatial features at multiple locations with few parameters (LeCun et al., 2015). In his following studies, he applied convolutional neural networks to images, speech, and time series (LeCun and Bengio, 1995). Circumscribed by the unsatisfactory hash-rate in the 1990s, convolutional neural networks failed to deliver serviceable results in processing big data and stayed dormant for several years (Kai et al. 2013). After 20 years development and thanks to the recent breakthrough of computing speed, convolutional neural networks show outstanding performance in image identification and classification. Convolutional neural networks have been widely used in computer vision, recommending system, and natural language processing.

In car insurance field, Singh (2016) implemented two convolutional neural networks to classify pictures of drivers. From a similar perspective, Masood et al. (2018) propose a machine learning algorithm based on convolutional neural networks to not only detect distracted drivers but also identify the cause of distraction. The input images are obtained from camera module installed in cars. Albeit convolutional neural networks appear to be promising in various time series analysis, their application in car insurance is scarce. Gao and Wüthrich (2018) apply 1D CNN for trips identification.

The paper is structured as follows. In Section 2, we clean the telematics car driving data and construct three formats of telematics data, time series, summary statistics and heatmaps. In Section 3, the classic Poisson generalized linear model for claims frequency is improved either by boosting methods or by incorporating telematics variables. In Section 4, we conclude the paper with important findings.

Section 2: Telematics car driving data

Telematics car driving data analysis is far from being trivial because the challenges related to big data (the size is typically beyond terabytes), data recording error and others. In this section, we firstly describe the telematics variables collected. Then we conduct the data cleaning. Finally we extract three formats of telematics data: (1) time series; (2) summary statistics; and (3) heatmaps. Our focus is the time series since there are already several existing literatures on the last two formats of telematics data.

2.1 TELEMATIC VARIABLES

A trip is defined as the period from engine start to engine switch off. During a trip, the following 15 variables are recorded by seconds with detailed definitions listed below.

2.1.1 FIELD MASKS

1. *Field_Mask*. This is a hex code which can be converted into a binary vector indicating the validation of telematics variable 6. *GPS_Latitude* to 15. *Accel_Vertical* (using the function *BMS::hex2bin*). Note that the invalid data are recorded as zero, which need to be coded as *NA*.

2.1.2 VEHICLE IDENTIFICATION VARIABLES

2. *Device_ID*. This variable also appears in the *userlist* dataset. It is used to match telematics data with policy data.

3. *Detected_VIN*. Given the *Device_ID*, this variable is not very useful in the current study.

2.1.3 TIME

4. *Trip_Number*. This variable records the beginning time of a trip in the UTC format and is a constant throughout the entire trip. It can be converted to the Beijing time by *as.POSIXlt(,origin="1970-01-01",tz="Asia/Shanghai")*.

5. *Time_Stamp*. This variable is the UTC time of each record in a trip and increases by one in a trip file.

2.1.4 GPS VARIABLES

When the following GPS variables are invalid, they are recorded as zero in the raw dataset. We change those invalid zero values to *NA*.

6. *GPS_Latitude*. This variable is the GPS latitude in decimal degree multiplied by 10^7.

7. *GPS_Longitude. This variable is the GPS longitude in decimal degree multiplied by* 10^6

The spherical coordinates can be converted to the plane coordinates (x, y) using *SoDA::geoXY*. The plane coordinates are used to calculate the distance, speed $v^{(xy)}$, angle $\psi^{(xy)}$, acceleration $a^{(xy)}$ and angle change $\Delta^{(xy)}$. When GPS signal drift problems occur, the GPS coordinates (x, y) will be affected, so will variables dependent on them.

8. *GPS_Heading*. This variable is the approaching direction of the vehicle $\psi^{(gps)}$ in decimal degree multiplied by 10^2. It ranges from 0 to 360.

9. *GPS_Speed*. This variable is the vehicle speed $v^{(gps)}$ in km/h multiplied by 10. Note that $3.6km/h = 1m/s$.

10. *Positional_Quality*. This variable indicates the GPS signal quality. When it is zero, the vehicle cannot be located by the GPS satellite and the four GPS variables listed above are invalid. Note that when it is one, the GPS coordinates might still be zero in rare cases. **So it is important to validate the GPS data using both the GPS coordinates and this variable**.

2.1.5 INSTRUMENTAL PANEL VARIABLES

11. *VSS_Speed*. This variable is vehicle speed $v^{(vss)}$ shown in instrument panel. We are informed by the insurer that when valid, it is more reliable than the *GPS_Speed* $v^{(gps)}$.

12. *Engine_RPM*. This variable is the rotation of engine per minute shown in the instrument panel, which is not used in the study.

2.1.6 ACCELEROMETER VARIABLES

The following three variables are measured by a three-axis accelerometer.

1. *Accel_Lateral*. The lateral acceleration $a'^{(acc)}$ is perpendicular to the car approaching direction, which measures the change of car approaching direction.

2. *Accel_Longitudinal*. The longitudinal acceleration $a^{(acc)}$ is along the car approaching direction, which measures the change of speed values.

3. *Accel_Vertical*. The vertical acceleration is always at around the gravity $9.8m/s^2$.

There is calibration bias with these variables as we will explain in more details in the next section. Also, such bias is difficult to be removed.

2.2 TELEMATICS DATA CLEANING

Telematics car driving data have a large size and are deemed big data. Data cleaning for such big data is very challenging since the same procedure needs to be applied to all the trips of all cars. Thus, the data cleaning should be adequately flexible for this aim.

We firstly visualize several typical trips to demonstrate what data issues need to be addressed during the cleaning procedure. Then, a "naive" data cleaning procedure is designed, and its performance is monitored on the selected trips. Finally, a "universally" applied data cleaning procedure is further derived.

2.2.1 ORIGINAL TELEMATICS DATA

Three trips of three drivers are illustrated in the following figures. For each trip, 6 plots are presented:

1. Top-left: Time series of GPS signal quality, instrument panel signal quality, and accelerometer signal quality;

2. Top-right: Trajectory (x, y);

3. Middle-left: Time series of GPS speed $v^{(gps)}$, instrument panel speed (VSS speed) $v^{(vss)}$,

4. Middle-right: Time series of GPS heading $\psi^{(gps)}$;

5. Bottom-left: Time series of longitudinal acceleration $a^{(acc)}$;

6. Bottom-right: Time series of lateral acceleration $a'^{(acc)}$.

Figure 2.1 shows a very short trip of 2 minutes for driver 8. We have the following observations:

- The top-left plot shows that the instrument panel signal is missing in the last 40 seconds.
- The top-right and middle-right plots show that the vehicle starts from east-south and approaches to west-north. Note that there are jumps between $\psi^{(gps)} = 0$ and $\psi^{(gps)} = 360$.
- The middle-left plot demonstrates that GPS speeds match with VSS speed, and VSS speeds are missing for the last 40 seconds.
- The bottom two plots reveal that there is **calibration bias** with both the acceleration rates, which might be corrected by subtracting the median of acceleration rates.

Figure 2.2 shows another two trips of driver 8. We have the following observations:

- For the left trip, the bottom two plots capture the calibration bias of accelerometer again.
- For the right trip, there are several segments with a missing GPS signal. We need to interpolate the GPS coordinates (x, y), speed $v^{(gps)}$ and heading $\psi^{(gps)}$ when the GPS signal is missing.

Figure 2.3 displays three trips of driver 288. GPS signals seem stable for the three trips, and the accelerometer seems work well except for the lateral acceleration during the second trip. Figure 2.4 presents three trips of driver 1188. GPS signals are unstable in the first two trips, and instrument panel signals are unstable for the last two trips. There seems to be calibration bias of acceleration rates for all these trips.

In **summary**, we need to consider the following data cleaning issues:

- There are missing values in GPS coordinates (x, y), GPS speed $v^{(gps)}$, GPS heading $\psi^{(gps)}$, and VSS speed $v^{(vss)}$. **These missing values need to be interpolated.**
- There are frequent calibration issues with the accelerometer variables $a^{(acc)}, a'^{(acc)}$. It is difficult to remove the calibration bias since the timing and severity0 of such a bias are rather random. Consequently, it is better to consider other variables to describe the acceleration in two directions. **We will use the derived longitudinal acceleration rates $a^{(gps)}, a^{(vss)}, a^{(xy)}$, which are obtained from GPS speed $v^{(gps)}$, VSS speed $v^{(vss)}$, and GPS coordinates (x, y), respectively. And we will replace lateral acceleration $a'^{(acc)}$ by angle changes of heading directions $\Delta^{(gps)}, \Delta^{(xy)}$, which are derived from GPS heading $\psi^{(gps)}$ and GPS coordinates (x, y), respectively.**

Figure 2.1

ONE TRIP OF DRIVER 8

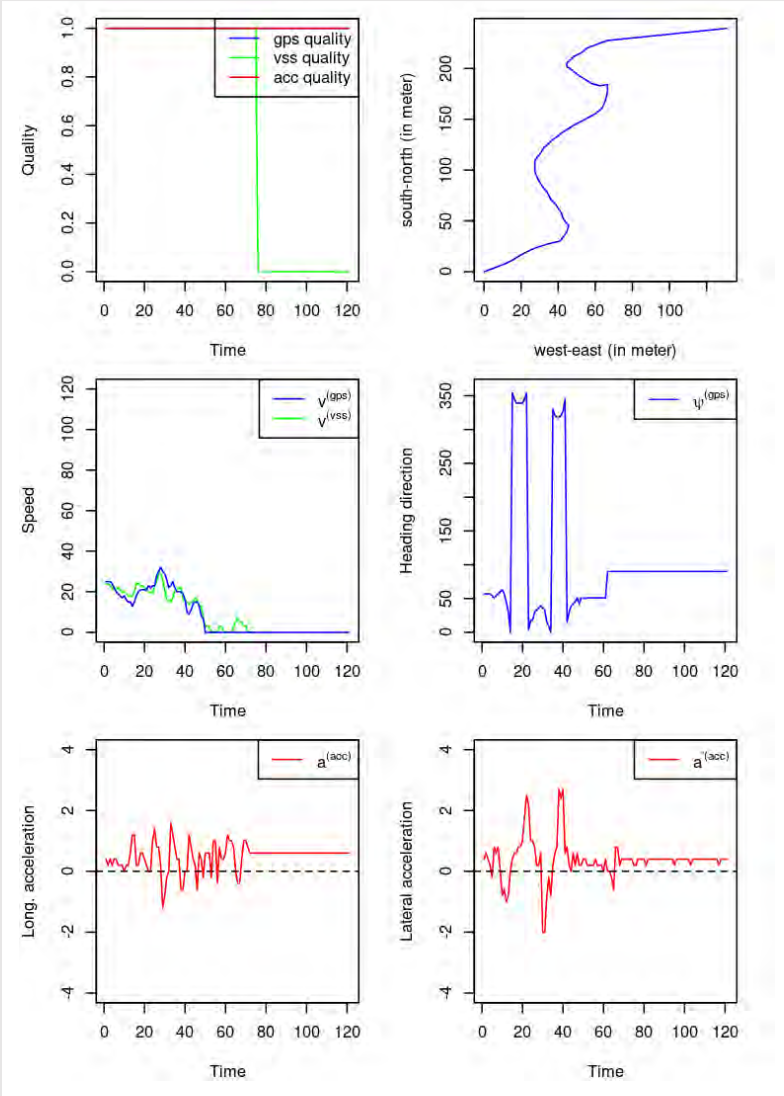


Figure 2.2

TWO TRIPS OF DRIVER 8

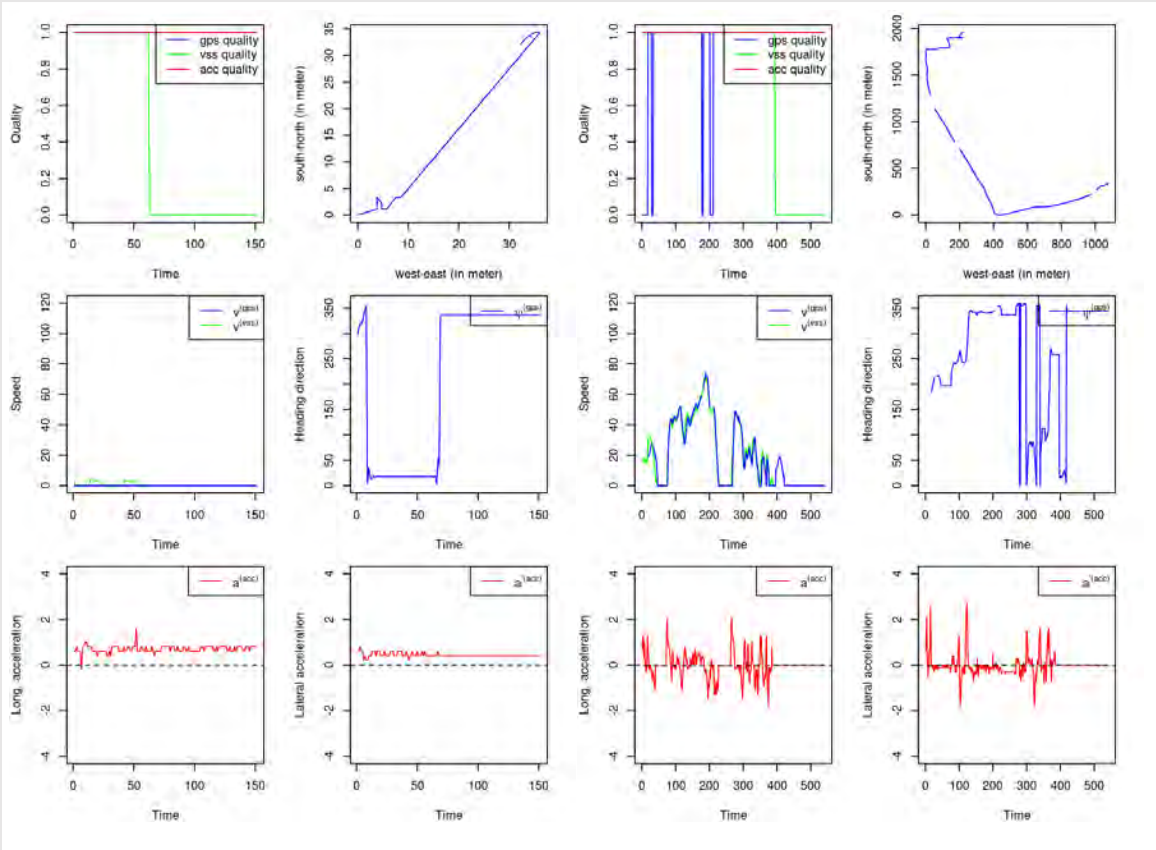


Figure 2.3
THREE TRIPS OF DRIVER 288

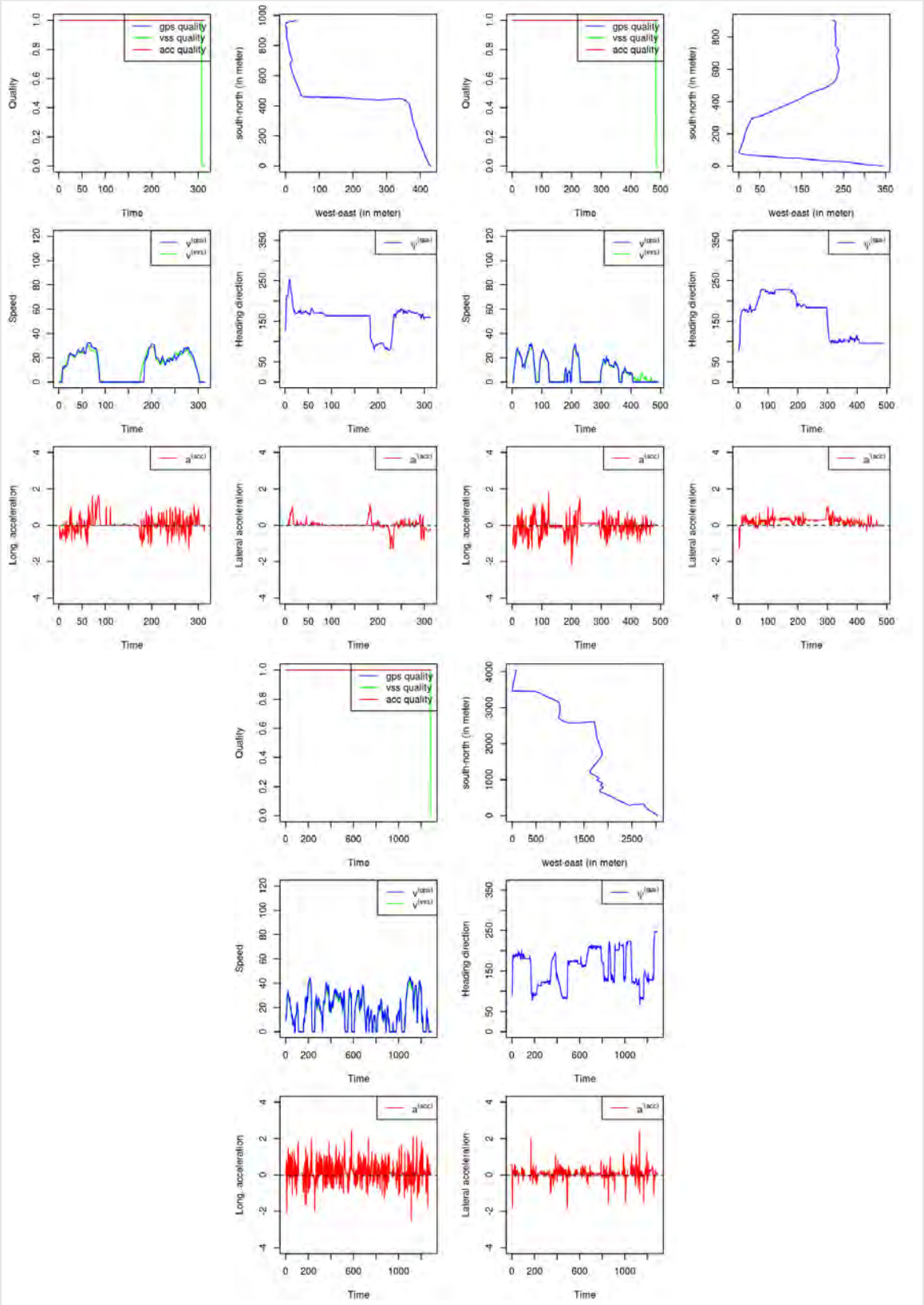
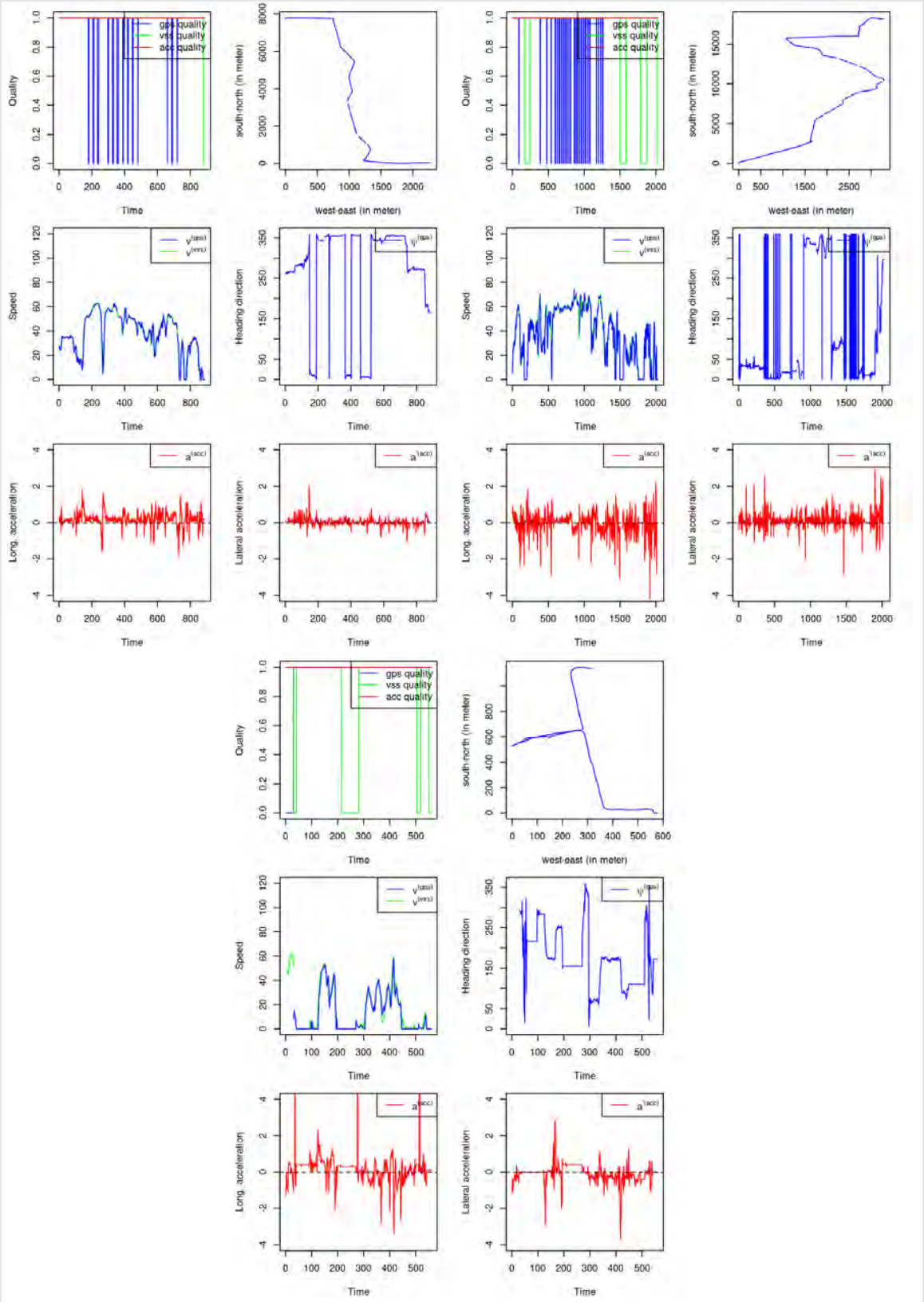


Figure 2.4
THREE TRIPS OF DRIVER 1188



2.2.2 DATA IMPUTATION

Data imputation is essential if time series of telematics variables are analyzed. On the contrary, if only the summary statistics of telematics variables are employed (e.g., mean of speed and standard deviation of acceleration), the interpolation of missing values is not necessary according to the law of large numbers.

More specifically, missing values of GPS coordinates (x, y), GPS speed $v^{(gps)}$, GPS heading $\psi^{(gps)}$, and VSS speed $v^{(vss)}$ are linearly interpolated, respectively. Note that missing values at boundaries of trips cannot be interpolated.

Remarks: The imputation of GPS heading $\psi^{(gps)}$ around 0 or 360 degree is challenging and may leads to biased GPS heading imputation. However, such biased imputation should have little effects on the derived angle change. Also, although GPS drifts may affect GPS speed, angle change and acceleration, GPS drifts do not occur frequently.

2.2.3 DERIVED ACCELERATION AND ANGLE CHANGE

Since the quality of accelerometer variables is relatively questionable, alternative measures are derived to measure the acceleration and direction changes. More specifically, acceleration rates are calculated by using GPS speed $v^{(gps)}$, VSS speed $v^{(vss)}$ and GPS coordinates (x, y), respectively:

$$\begin{aligned} a_t^{(gps)} &= v_t^{(gps)} - v_{t-1}^{(gps)} \\ a_t^{(vss)} &= v_t^{(vss)} - v_{t-1}^{(vss)} \\ a_t^{(xy)} &= v_t^{(xy)} - v_{t-1}^{(xy)} \end{aligned} \tag{2.1}$$

where

$$v_t^{(xy)} = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}.$$

Direction change is measured by angle change which can be obtained using either GPS heading $\psi^{(gps)}$ or GPS coordinates (x, y):

$$\begin{aligned} \Delta_t^{(gps)} &= \arcsin\left(\sin\left(\psi_t^{(gps)} - \psi_{t-1}^{(gps)}\right)\right) \\ \Delta_t^{(xy)} &= \arcsin\left(\sin\left(\psi_t^{(xy)} - \psi_{t-1}^{(xy)}\right)\right) \end{aligned} \tag{2.2}$$

where

$$\psi_t^{(xy)} = \mathbb{I}_{(-\infty,0)}\left(arctan2\left(\frac{x_t - x_{t-1}}{y_t - y_{t-1}}\right)\right) \times 2\pi + arctan2\left(\frac{x_t - x_{t-1}}{y_t - y_{t-1}}\right) \in [0, 2\pi].$$

Note that a positive value of Δ indicates a right turn whereas a negative one suggests a left turn. The jump between 0 and 360 is not an issue since we take sine values.

Remarks: Cares are needed when converting speed units from km/h to m/s, and transforming angle units from degree to radian. The original GPS coordinates are in spherical form, which needs to be converted to geodetic plane coordinates before calculating the acceleration and angle change.

2.2.4 SELECTION OF TELEMATICS VARIABLES

After data imputation and derivation of the corresponding telematics, we discuss the telematics variable selection among speed $v^{(gps)}, v^{(vss)}, v^{(xy)}$, acceleration $a^{(gps)}, a^{(vss)}, a^{(xy)}$, angle $\psi^{(gps)}, \psi^{(xy)}$, and angle change $\Delta^{(gps)}, \Delta^{(xy)}$. Those telematics variables are plotted in Figure 10 for the same trips as those in Figures 2.1, 2.2, 2.3 and 2.4. We have the following observations for Figure 2.5:

- The top two plots in Figure 2.5 are exactly the same as the top two in Figure 2.1.
- In the middle-left plot, we add $v^{(xy)}$ for comparison. Note that there is a GPS drift around 60 second causing a jump of $v^{(xy)}$.
- In the middle-right plot, we add $\psi^{(xy)}$ for comparison. The derived heading direction $\psi^{(xy)}$ is always zero when the vehicle stands still after 50 seconds.
- In the bottom plots we show the derived acceleration $a^{(gps)}, a^{(vss)}, a^{(xy)}$ and the derived angle change $\Delta^{(gps)}, \Delta^{(xy)}$.

We now investigate the GPS drift at around the 60th second. In Figure 2.6, time series of (x, y) coordinates, $v^{(xy)}$ and $\psi^{(xy)}$ from the 45th second to 65th second are plotted, respectively. A jump of (x, y) can be seen at the 62nd second, which leads to an extreme speed jump from 0 to more than 200 km/h and an unusual direction jump from 0 to more than 60 degree in one second.

We conclude that the derived variables using GPS coordinates (x, y) are unstable compared with those using GPS speed, GPS heading and VSS speed. This is due to the measurement error of GPS coordinates and its leverage effects on the acceleration and angle change.

We have the following observations for Figure 2.7:

- For the second trip of driver 8, we see a GPS drift at around the 70th second.
- For the third trip of driver 8, the imputation works very well for GPS coordinates, GPS speed and GPS heading. GPS drifts are often observed when there is a speed peak. The heading $\psi^{(xy)}$ derived from GPS coordinates is always incorrectly zero when the vehicle stops. Again, the acceleration and angle change derived from GPS coordinates are very unstable.

There are no new observations for Figures 2.8 and 2.9.

In summary, we argue that

- The linearly imputation works reasonably well.
- Due to GPS coordinates drift, we should avoid using variables derived from (x, y). Instead, the telematics variables $v^{(vss)}, v^{(gps)}, a^{(vss)}, a^{(gps)}, \psi^{(gps)}, \Delta^{(gps)}$ are more reliable.
- The distance should be derived using the speed variable rather than GPS coordinates (x, y).
- The three variables v, a, Δ are related to driving behavior, while heading direction $\psi^{(gps)}$ is irrelevant.
- Variables from the same sensor should be adopted for a certain study, i.e., we should use either $\left(v^{(vss)}, a^{(vss)}\right)$ or $\left(v^{(gps)}, a^{(gps)}, \Delta^{(gps)}\right)$ only.

Figure 2.5
ONE TRIP OF DRIVER 8

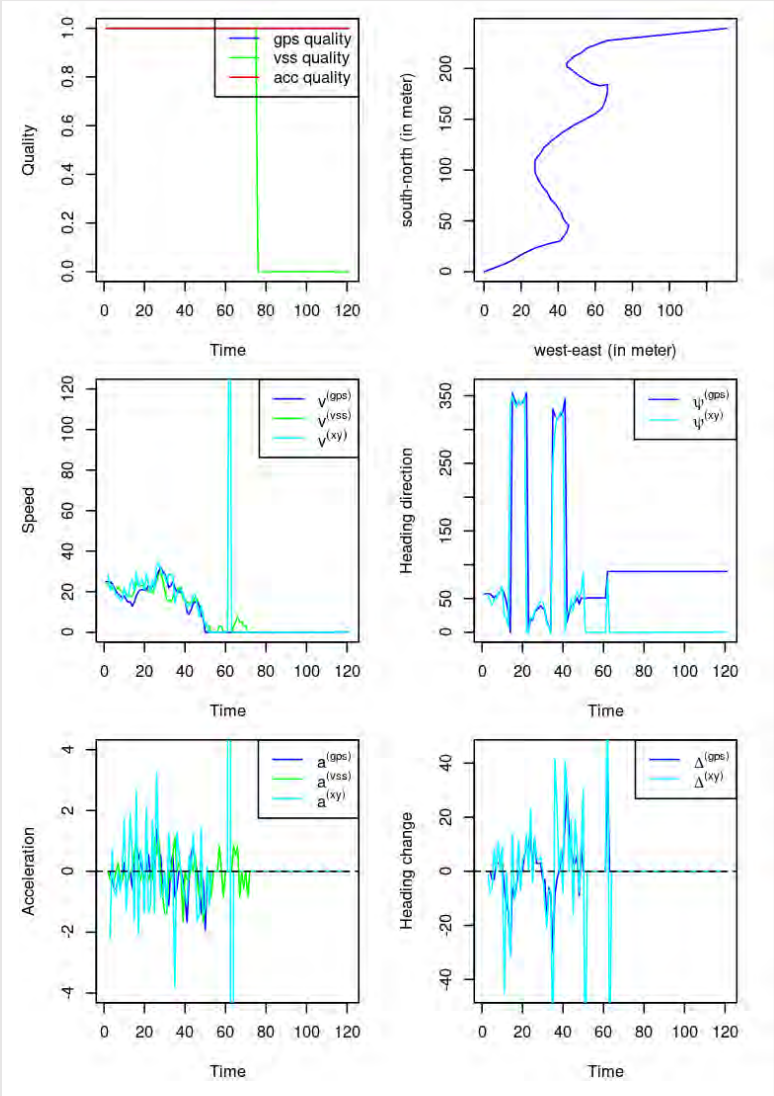


Figure 2.6
GPS DRIFT

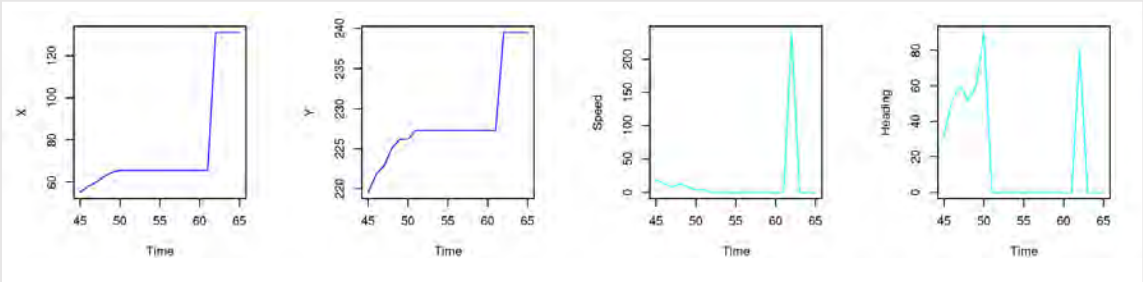


Figure 2.7

TWO TRIPS OF DRIVER 8

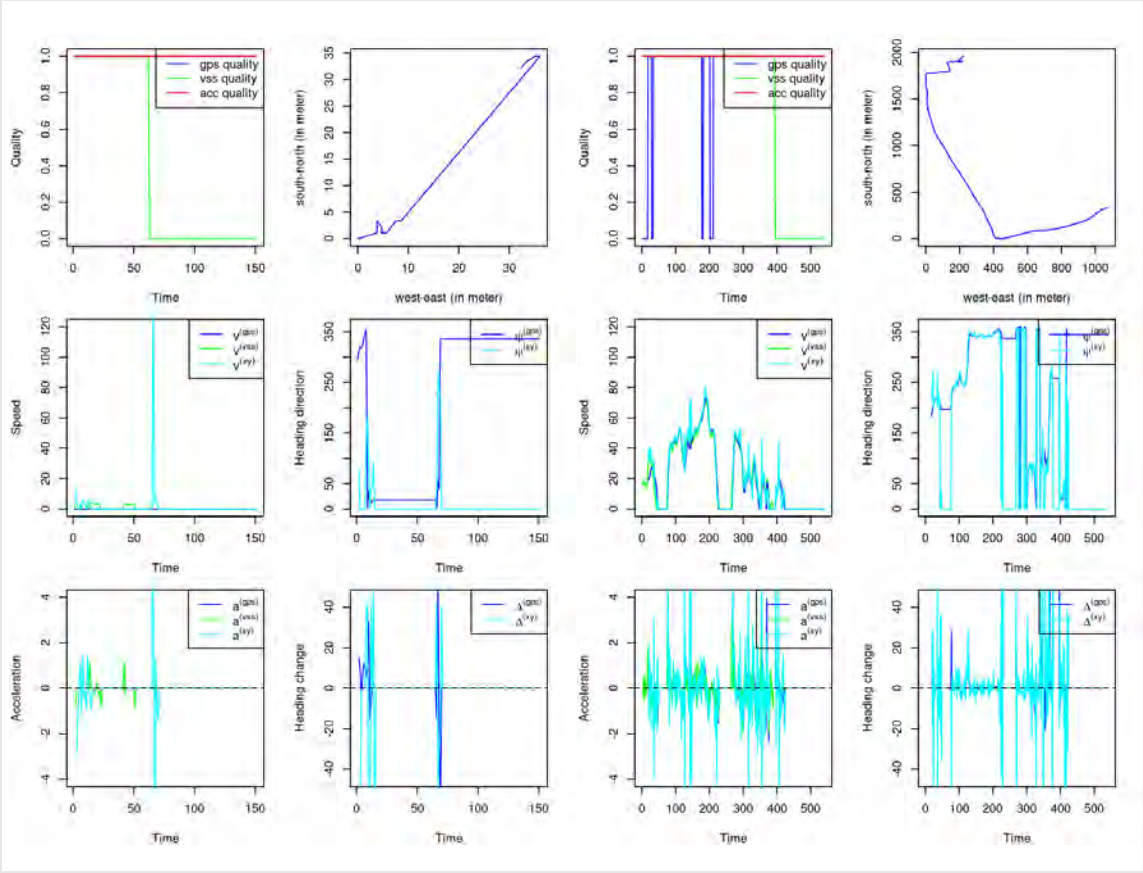


Figure 2.8

THREE TRIPS OF DRIVER 288

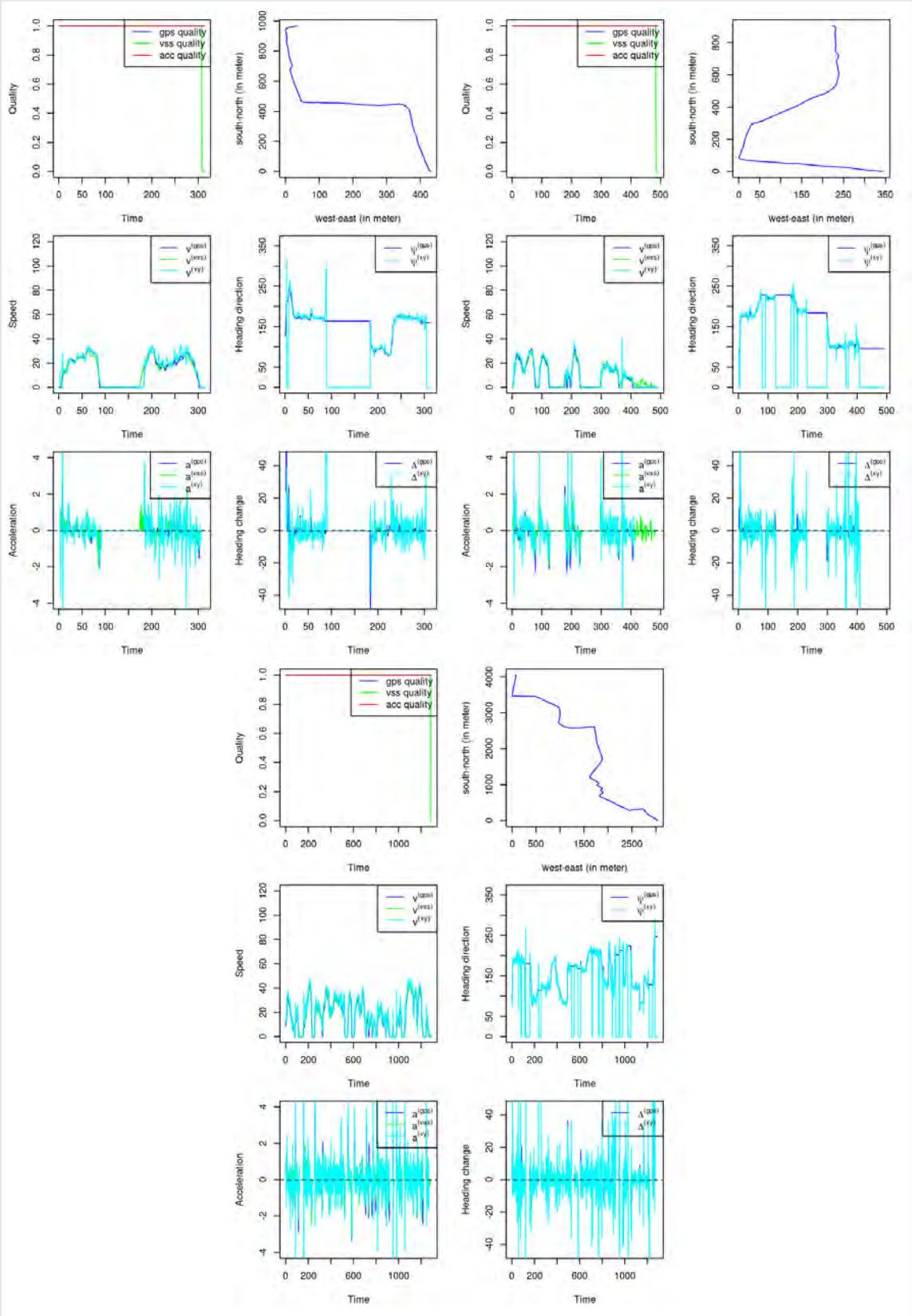
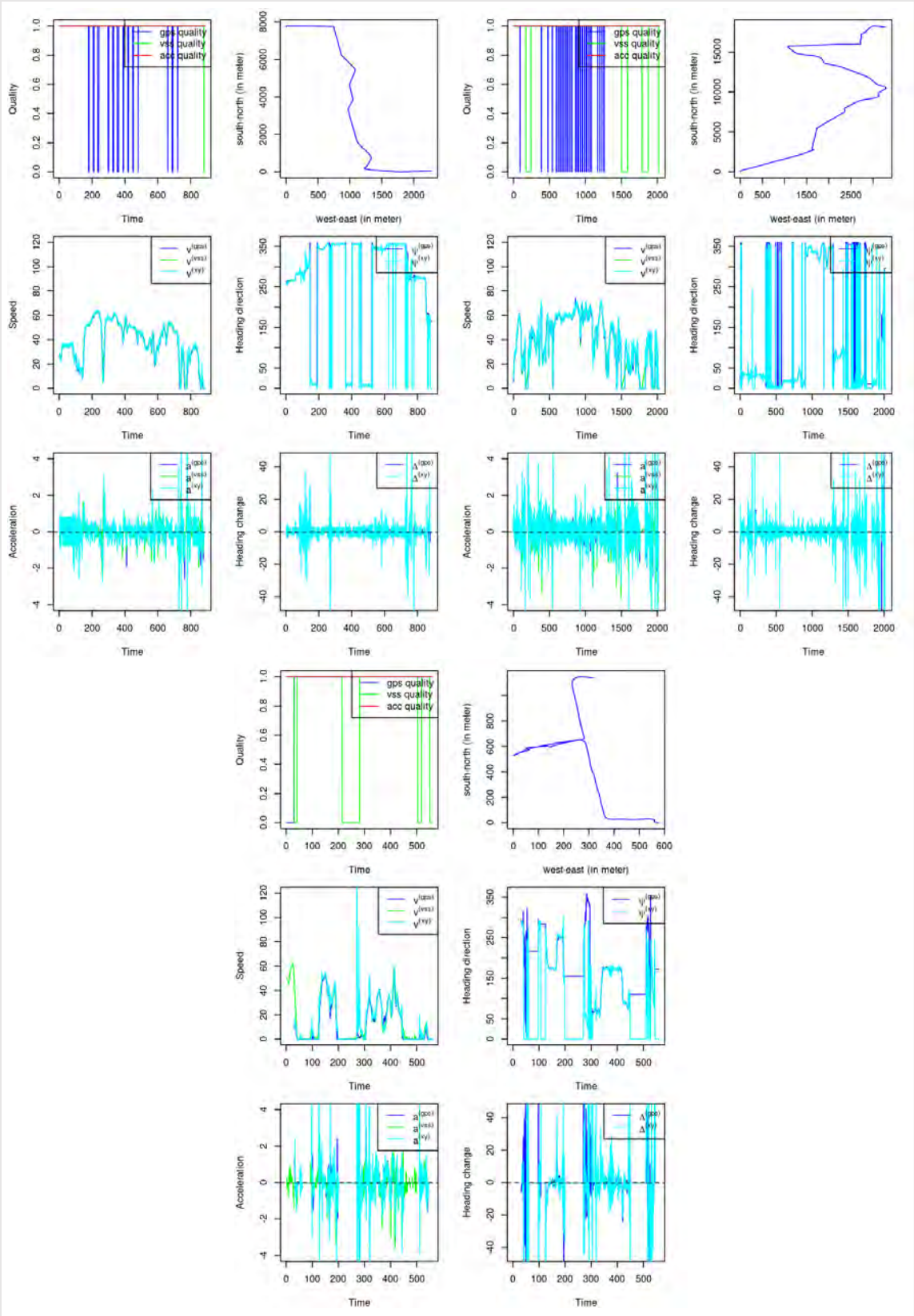


Figure 2.9

THREE TRIPS OF DRIVER 1188



2.2.5 ADOPTED TELEMATICS VARIABLES

For each trip, we extract the following (original) telematics variables which are actually adopted in the later study.

1. GPS speed $v^{(gps)}$
2. GPS heading $\psi^{(gps)}$
3. Validation of GPS signal
4. Instrument panel speed $v^{(vss)}$
5. Validation of instrument panel signal

Note that we can derive acceleration and angle change easily from speed and heading. Also note that if GPS signal is invalid, the corresponding variables are recorded as *NA*. So are instrumental panel variables.

We plot speed $v^{(gps)}$, heading direction $\psi^{(gps)}$, acceleration $a^{(gps)}$, angle change $\Delta^{(gps)}$ for the previously investigated trips of drivers 8 and 1188 in Figures 2.10 to 2.11. The timing of missing values is denoted by a horizontal line at the top of each plot. Note that we have capped the $a^{(gps)}$ between $(-4,4)$m/s^2, and $\Delta^{(gps)}$ between $(-45°, 45°)$. Finally, the missing $v^{(gps)}$ and $\psi^{(gps)}$ are interpolated linearly, and $a^{(gps)}$ and $\Delta^{(gps)}$ are obtained from the imputed values. Those telematics variables are shown in Figures 2.12 to 2.13.

Figure 2.10
THREE TRIPS OF DRIVER 8

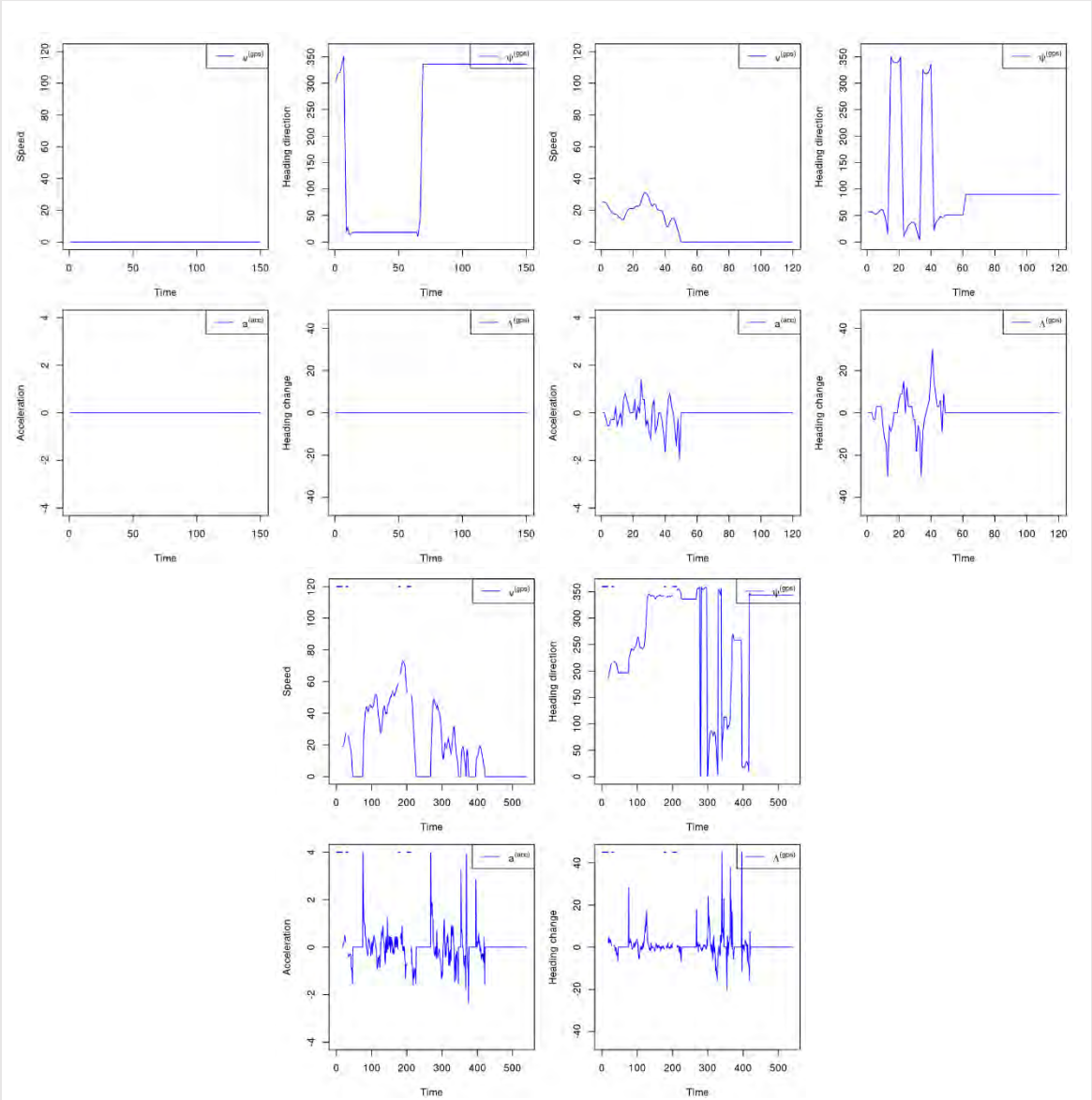


Figure 2.11
THREE TRIPS OF DRIVER 1188

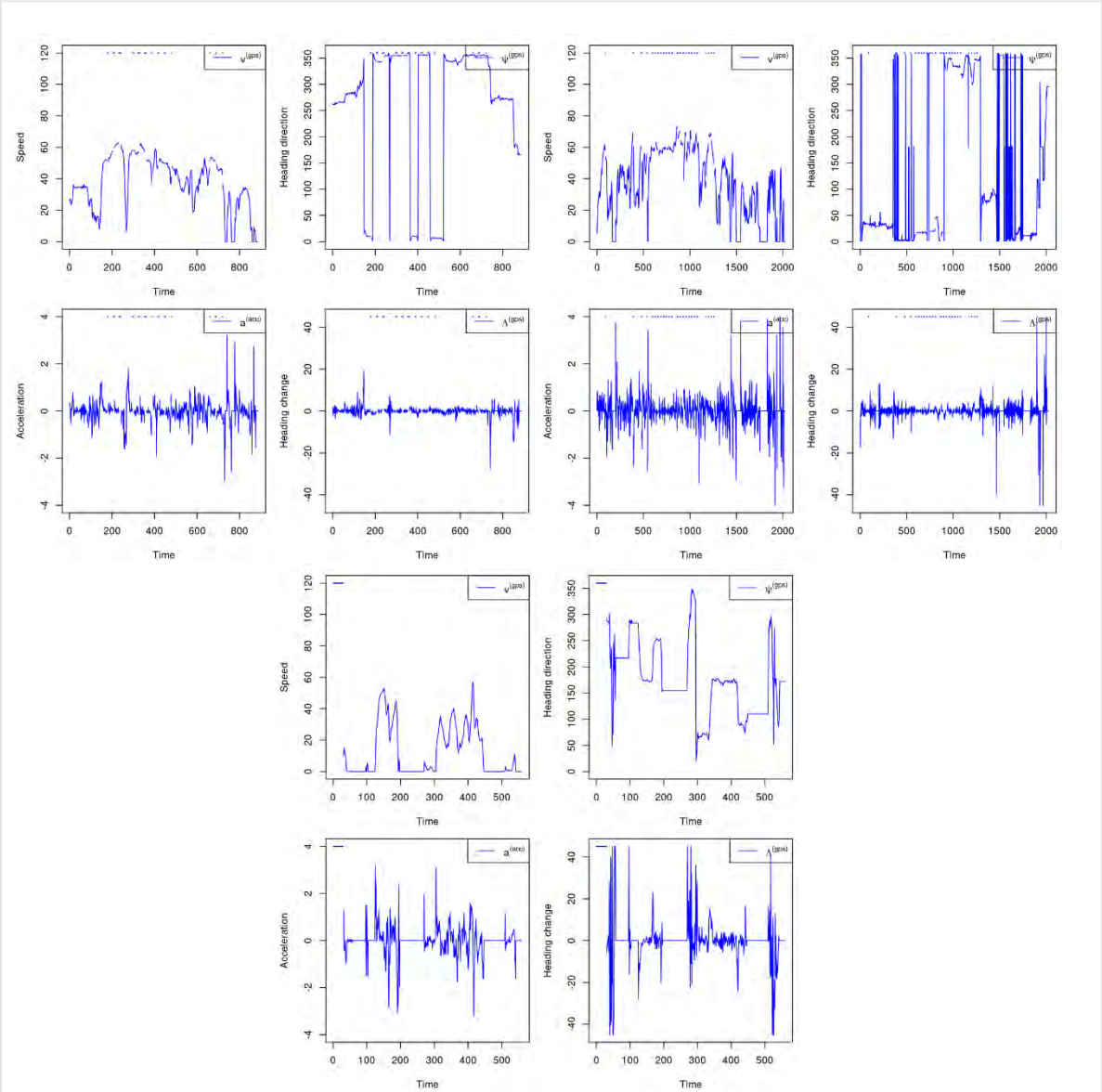


Figure 2.12
THREE TRIPS OF DRIVER 8

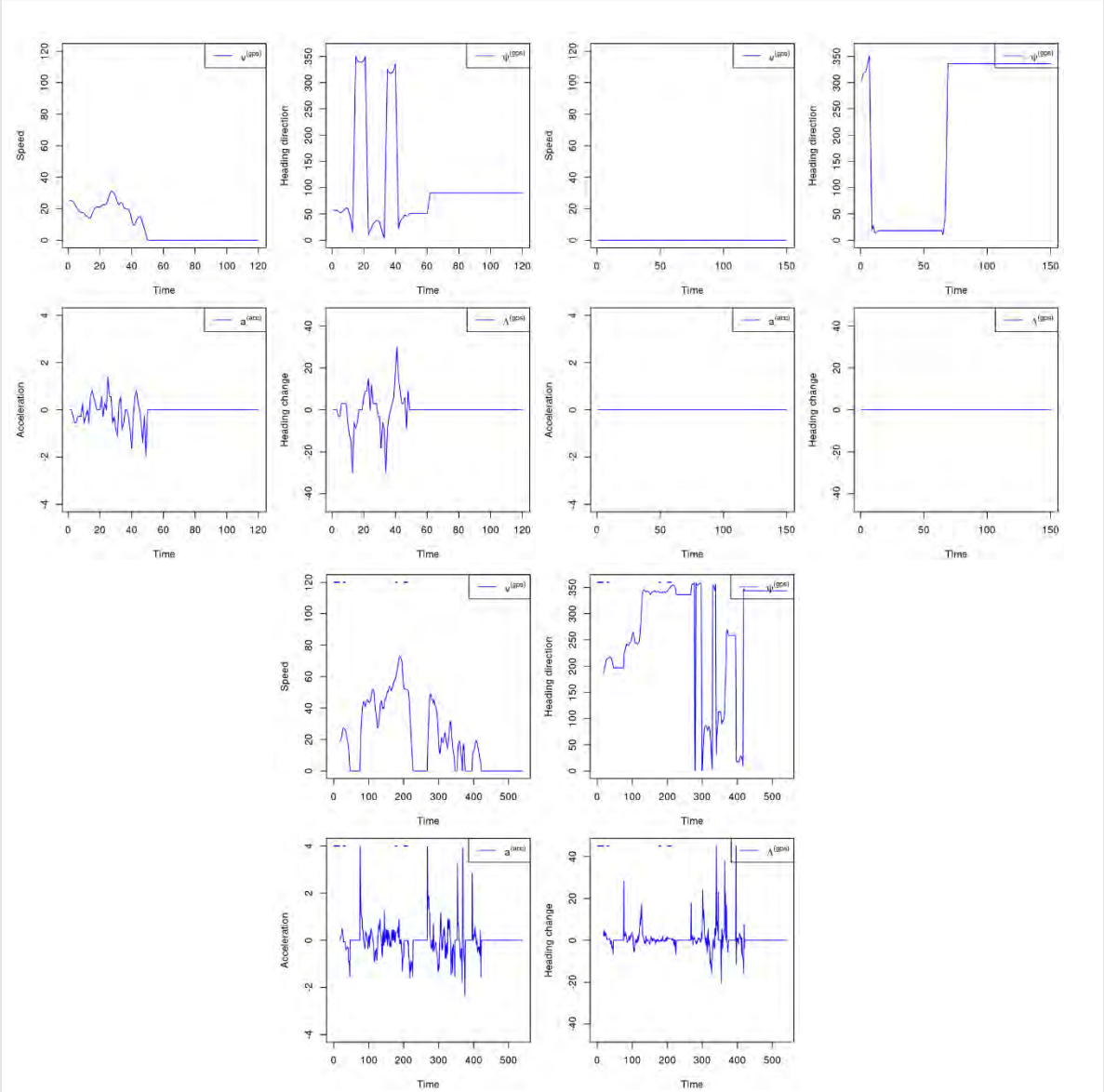
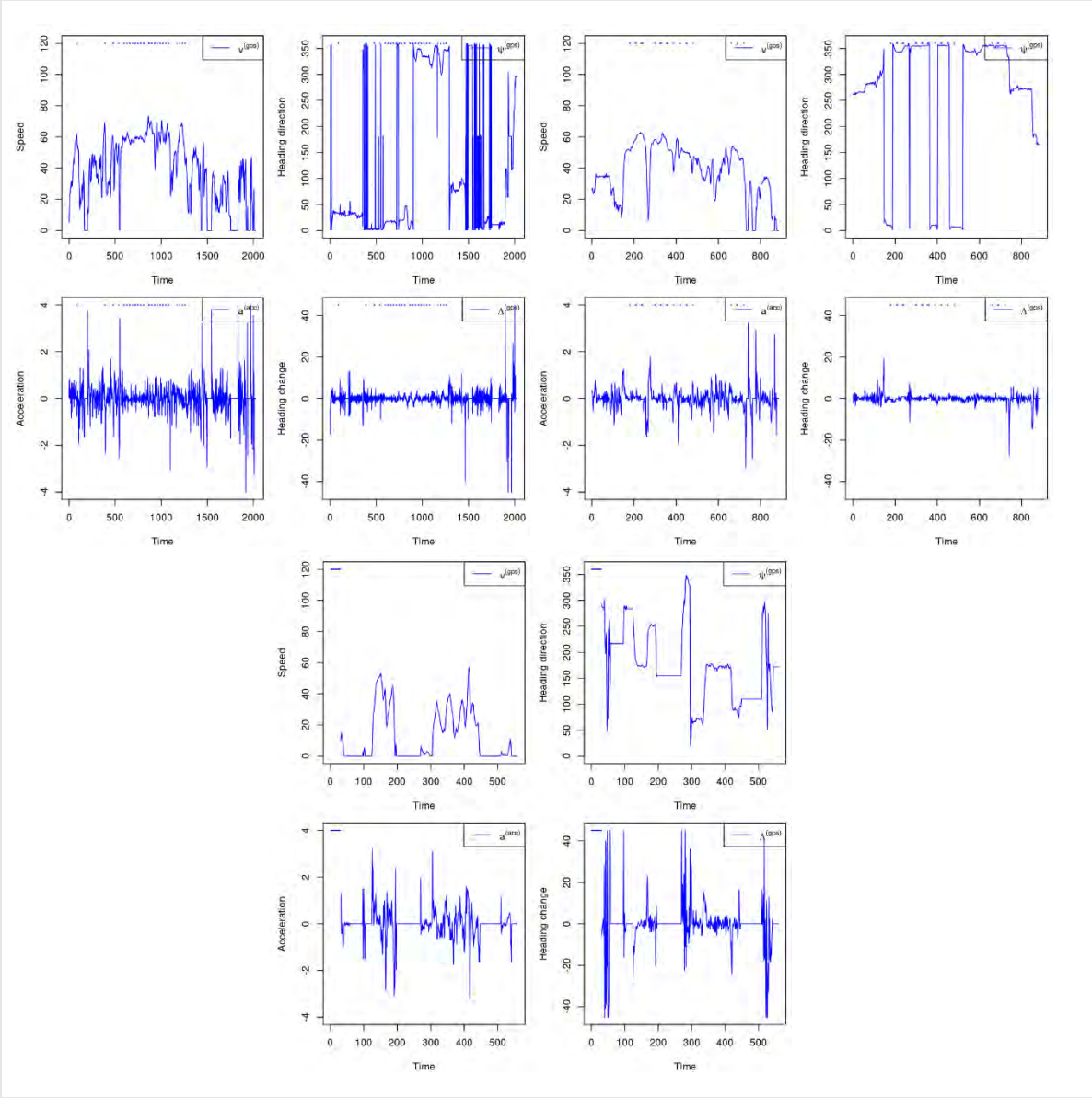


Figure 2.13
THREE TRIPS OF DRIVER 1188



2.2.6 TIME SERIES OF INDIVIDUAL TRIPS

We consider time series of individual trips, which will be used in the claims frequency modeling. In Appendix A, we consider the other two formats: summary statistics and heatmap. The data imputation is important for time series, rather than for the other two formats. Hence, a simpler data cleaning procedure may be performed to derive summary statistics and heatmap. For instance, only data with valid GPS variables are kept in this procedure.

In all cases, we must **specify particular speed intervals and truncate acceleration and angle change.** A particular speed interval makes the comparison among drivers more sensible, and truncation eliminates the leverage effects of outliers. The activation functions in neural networks play a similar role. Note that our focus is time series since there are already several existing literatures on the last two formats.

Recall from Figure 2.1, a trip may contain several standing still phases $\{t: v_t = 0, a_t = 0\}$. Those phases should be removed since they are irrelevant to driving behavior. Also, it is worth noting that the GPS variables tend to be more unreliable with a slow driving speed (e.g. parking). Thus, by focusing on a meaningful and usual speed interval $[10,60]$ km/h, we can make more sensible comparison among drivers.

We firstly extract partial trips in $[8,62]$ km/h and set the acceleration and angle change in the shifting parts $[8,10] \cup [60,62]$ km/h as zero. Further, we truncate the speed in the shifting parts $[8,10] \cup [60,62]$ km/h to 10 and 60 km/h, respectively. We save the first $\tau = 5 \times 60$ seconds of such concatenated trips for each trip. For the illustrative trips presented in the previous sections, we obtain three concatenated trips as shown in Figure 2.14. The other trips do not spend more than 5 minutes in speed interval $[10,60]$ km/h, and are discarded in the analysis.

We extract at most 500 trips for each car. Therefore, for each driver $i = 1, \dots, n$, we have $c_i \in [0,500]$ pre-processed trips. The distribution of $c_i \in (0,500)$ is shown in Figure 35. Altogether, there are 175 cars (8%) without pre-processed trips and 1721 cars (75%) cars with 500 pre-processed trips. For driver i with $c_i > 0$ trips, we stack its trips into a $c_i \times 300 \times 3$ array $\mathbf{z}_i = (z_{i,1}, \dots, z_{i,c_i})'$, where each trip $z_{i,j} \in [10,60]^{300} \times [-4,4]^{300} \times [-45°, 45°]^{300}$. Note that the un-bolded notation z_j refers to a trip j without considering which driver it belongs to.

Remark: One may choose the middle stage of a trip; however, it is difficult to define such stage. On one hand, if the middle 5 minutes were selected, it would be unfair for trips with varying durations. On the other hand, if the first several minutes (e.g. 10 minutes) were discarded and the next 5 minutes were selected, it would be not appropriate for short trips (e.g. shorter than 15 minutes). By extracting the first 5 minutes, we fairly compare driving behavior in the early stage of a trip.

Figure 2.14
THE THIRD TRIP OF DRIVER 288. THE FIRST AND SECOND TRIPS OF DRIVER 1188

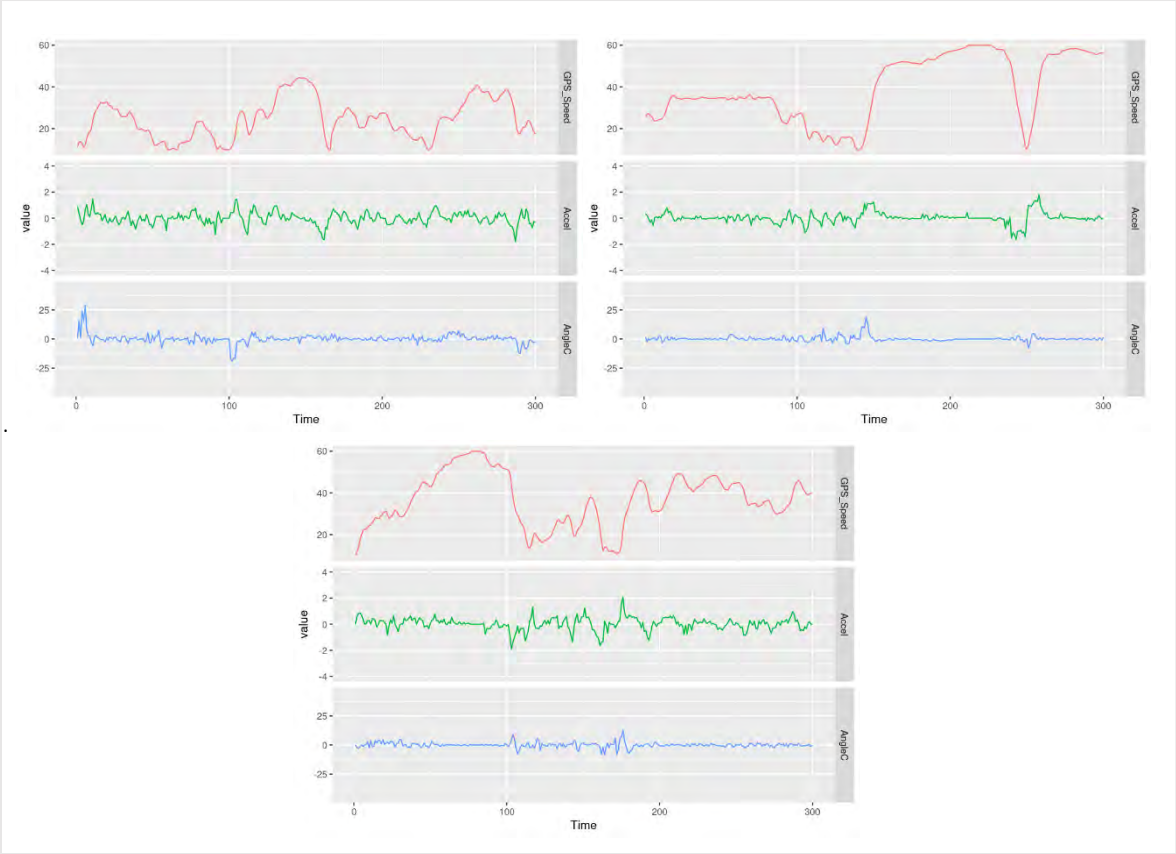
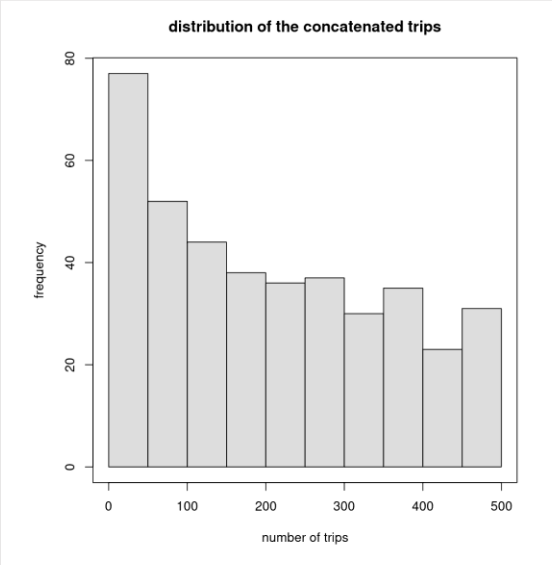


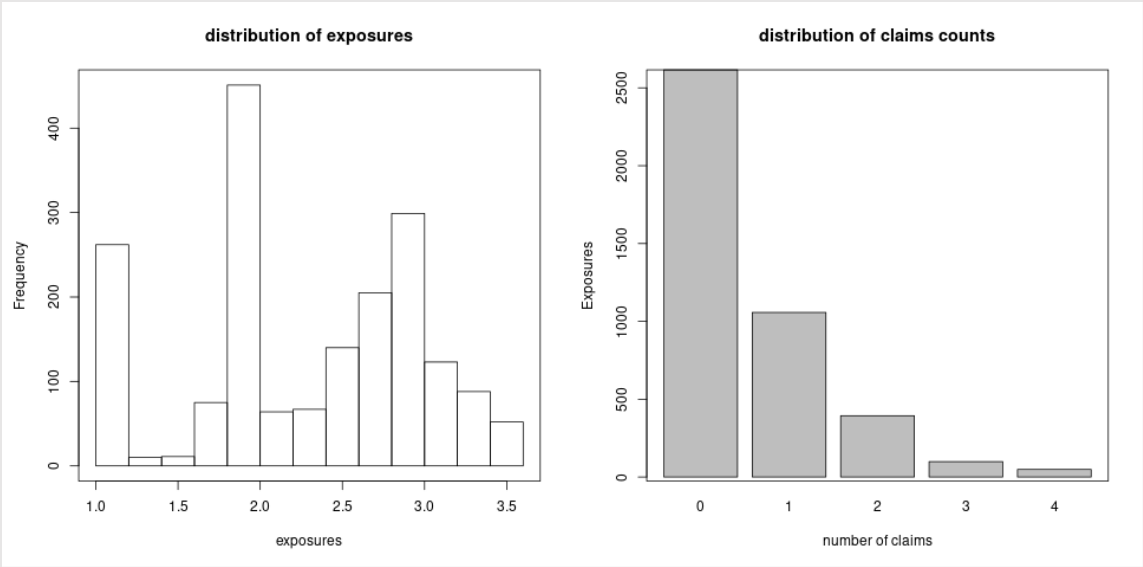
Figure 2.15
THE DISTRIBUTION OF PRE-PROCESSED TRIPS IN (0, 500)



Section 3: Claims frequency modeling

We consider the claims history of $i = 1, \ldots, n = 1847$ cars whose exposure e_i is ranging from 1 to 3.48 years. Those cars have over 100 and up to 500 concatenated trips of 5 minutes (see the previous section). Altogether, there are 1598 car drivers with 500 such telematics trips. We show the distribution of claims number N_i and exposures e_i in Figure 38. We observe that most drivers do not have a claim. The total claims number is $\sum_{i=1}^{n} N_i = 933$ and the total exposures is $\sum_{i=1}^{n} e_i = 4215$. The empirical claims frequency is $\sum_{i=1}^{n} N_i \,/\, \sum_{i=1}^{n} e_i = 22.14\%$ per driver per year.

Figure 3.1
THE DISTRIBUTIONS OF CLAIMS NUMBER AND EXPOSURES



Remark: Our preliminary data cleaning ensures that the main driver of a car does not change over the entire observation period, and we concatenate policy renewals of the same driver over this observation period. Thus, we can follow the same driver for at most 3 years and 5 months from 01/01/2014 to 31/05/2017.

Remark: We follow insurance policies over multiple years, but only for the most recent periods there is telematics data available. For this reason, we typically have a longer observation period of claims history on insurance policies than that of corresponding telematics data. An implicit assumption is that the driving habits and styles in the most recent periods are good representations for the entire observation period of insurance exposure.

Generally we assume that the claims number N_i follows a Poisson distribution with mean of $e_i\lambda(\mathbf{x}_i)$

$$N_i \overset{ind}{\sim} \text{Poi}\big(e_i\lambda(\mathbf{x}_i)\big),$$

where e_i is the exposure and $\lambda(\mathbf{x}_i)$ is the claims frequency per driver per year given the risk factors $\mathbf{x}_i \in \mathcal{X}$. The function λ is a mapping from risk factors to claims frequency:

$$\lambda: \mathcal{X} \to \mathbb{R}_+, \quad \mathbf{x} \mapsto \lambda(\mathbf{x}).$$

We firstly establish the base line generalized linear model (GLM) for claim counts with λ as a log-linear function. Then we improve it by either relaxing the log-linear constraint or introducing telematics covariates to expand the covariate space $\mathcal{X}$.

We will compare models in terms of out-of-sample prediction. Stratified split (w.r.t. claims numbers) is used to partition the data into train $\mathcal{D}_{\text{train}}$, validation $\mathcal{D}_{\text{validation}}$, and test $\mathcal{D}_{\text{test}}$ data sets $(0.6: 0.2: 0.2)$. The exposures, claims counts, and claims frequency of each data set are listed in Table 1. We denote the index of data sets $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{validation}}, \mathcal{D}_{\text{test}}$ by $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{validation}}$ and $\mathcal{I}_{\text{test}}$, respectively. The index of learning data set $\mathcal{D}_{\text{learn}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{validation}}$ is denoted by $\mathcal{I}_{\text{learn}} = \mathcal{I}_{\text{train}} \cup \mathcal{I}_{\text{validation}}$.

Table 3.1
DATA PARTITIONS

data	cars		exposure	claims	frequency
train	1107		2535	557	0.2197
validation	372		843	192	0.2279
test	368		837	184	0.2198

It is natural to use the Poisson deviance loss to compare different models. The out-of-sample Poisson deviance loss on the data $\mathcal{D}_{\text{test}}$ is defined as:

$$\mathcal{L}\left(\hat{\lambda}, \mathcal{D}_{\text{test}}\right) = \frac{2}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \left(e_i \hat{\lambda}(\mathbf{x}_i) - N_i - N_i \ln e_i \hat{\lambda}(\mathbf{x}_i) + N_i \ln N_i\right). \tag{3.1}$$

Note that the mapping $\hat{\lambda}$ contains estimated parameters using the training data $\mathcal{D}_{\text{train}}$ (or the learning data $\mathcal{D}_{\text{learn}}$ in GLM).

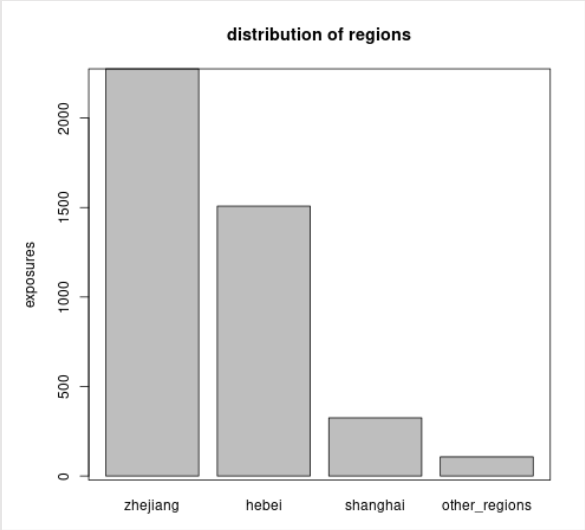
3.1 COVARIATES PREPROCESS

The available actuarial risk factors are *regions*, *driver's age*, *driver's gender*, *car brand*, *car's age*, *seat count*, *car's price* and *average daily distance*. We preprocess them as follows:

- *regions*: There are three main regions, Hebei Province, Zhejiang Province and Shanghai City, which have accounted to 97.67% of total cars. Hence, we merge the remaining regions as other regions. The distribution of exposures across those four regions is shown in Figure 3.2.

  Note that one may create a continuous variable of population density in each region. Population density should be related to claims frequency, since it is a proxy of traffic density.

Figure 3.2
THE DISTRIBUTION OF EXPOSURES ACROSS FOUR REGIONS.

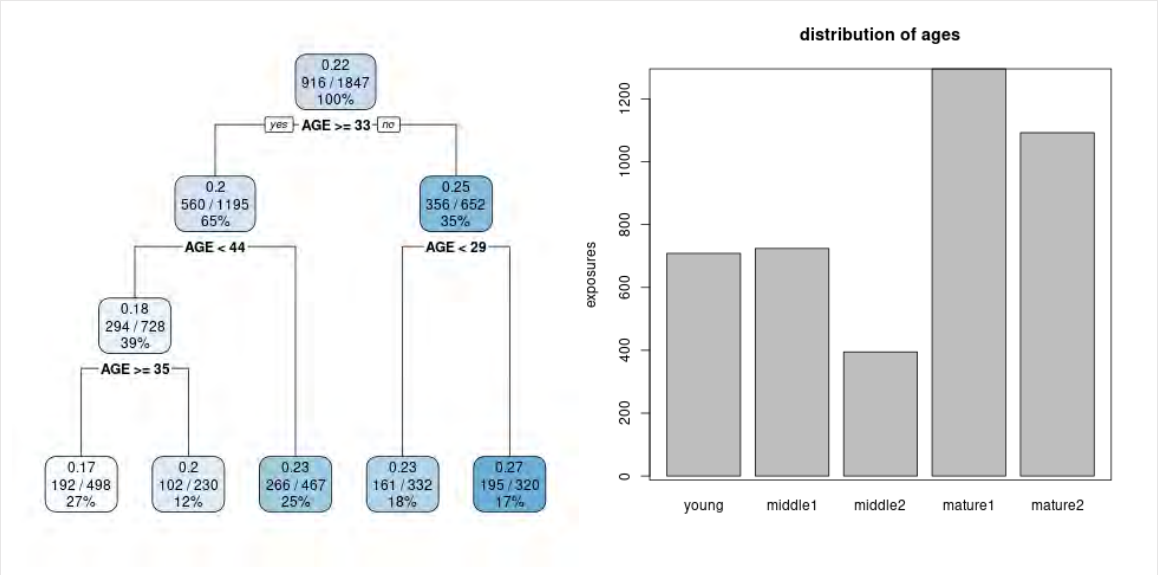


- *driver's age*: It is known that age has a non-linear relationship with claims frequency. Thus, we discretize driver's age into five groups using a marginal Poisson regression tree model. The cut-off values of age are 29, 33, 35 and 44. Note that we try to obtain a fine grouping of age here, and merge age groups during GLM variable selection. Figure 3.3 shows the marginal Poisson tree and the distribution of exposures across the age groups.

Table 3.2
AGE GROUPS

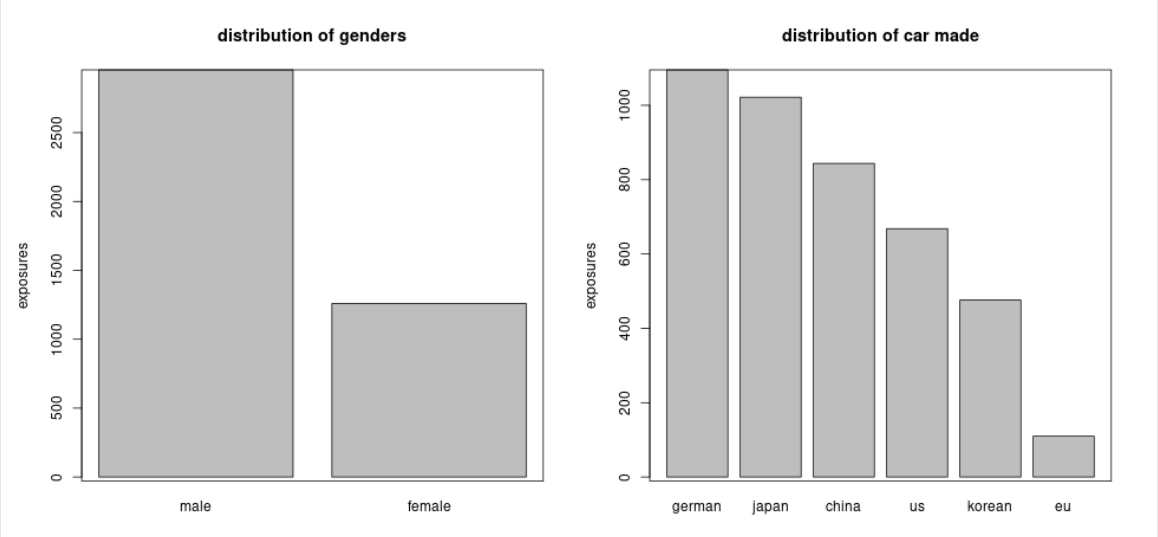
age group	age interval
young	[18,29)
middle1	[29,33)
middle2	[33,35)
mature1	[35,44)
mature2	[45,100)

Figure 3.3
THE MARGINAL POISSON TREE AND THE DISTRIBUTION OF EXPOSURES ACROSS AGE GROUPS



- *gender*: Male drivers are almost twice the size of female drivers as shown in Figure 3.4.
- *car brand*: There are 66 different car brands. Due to the small sample sizes of individual car brands, we create a new categorical variable of car made country with 6 levels: Germany, Japan, China, US, Korean and Europe (excluding Germany). Figure 3.4 displays the distribution of exposures across car made countries.

Figure 3.4
THE DISTRIBUTION OF EXPOSURES ACROSS GENDERS. THE DISTRIBUTION OF EXPOSURE ACROSS CAR MADE COUNTRIES

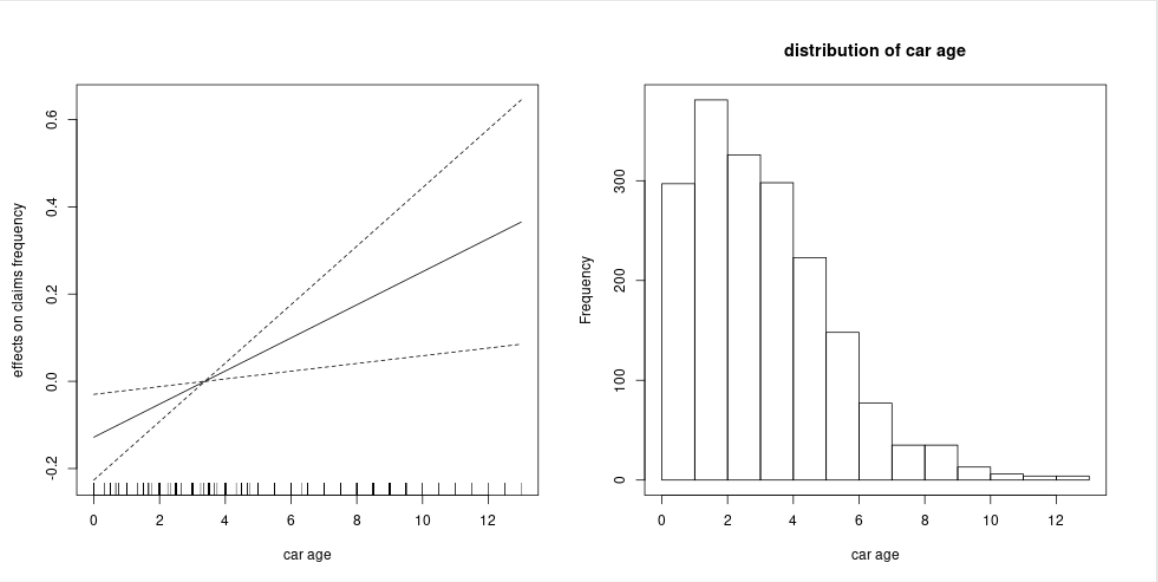


- *car's age*: The distribution of car's age is plotted in Figure 3.5. A marginal generalized additive model (GAM) for claims frequency indicates that claims frequency is related to car's age in a log-

linear fashion (due to logarithm link function) as shown in Figure 3.5. Thus, discretization is not needed for this variable.

Figure 3.5

HISTOGRAM OF CAR'S AGE. THE EFFECT OF CAR'S AGE ON CLAIMS FREQUENCY (FROM A MARGINAL GAM)

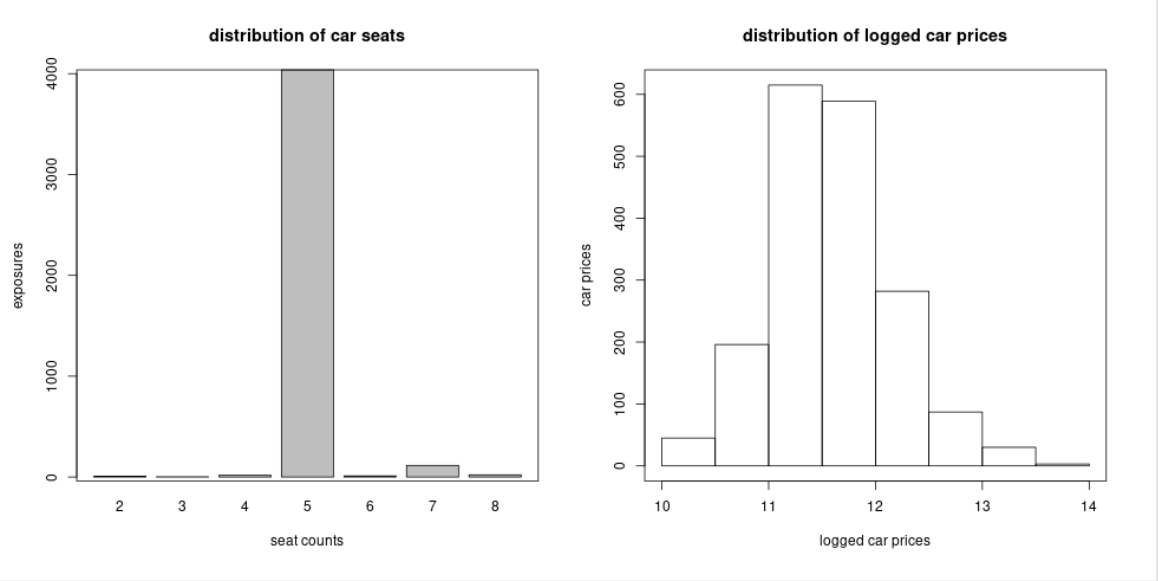


- *seat count*: Around 95.67% cars have 5 seats as shown in Figure 3.6. So, this variable is not quite useful for claims frequency prediction.

- *car's price*: The distribution of car's price in logarithm is shown in Figure 3.6. A preliminary analysis (using marginal tree, GAM, and GLM) shows that this variable doesn't have a close relationship with claims frequency.

Figure 3.6

THE DISTRIBUTION OF EXPOSURE ACROSS SEAT COUNT. HISTOGRAM OF LOGGED CAR'S PRICE



- *average daily distance*: We fit a marginal GAM to investigate the non-linear effect of average daily distance on claims frequency. The left plot in Figure 3.7 suggests that the logged daily distance are mainly between 2.5 and 4.5. We truncate the logged daily distance at 2.5 and 4.5, and then fit another marginal GAM with the truncated variable. The right plot in Figure 3.7 shows that claims frequency is related to truncated logged daily distance in a log-linear fashion (due to logarithm link function). Figure 49 compares the distributions of the original and truncated variables.

Figure 3.7
THE EFFECT OF LOGGED DAILY DISTANCE ON CLAIMS FREQUENCY (FROM TWO MARGINAL GAMS)

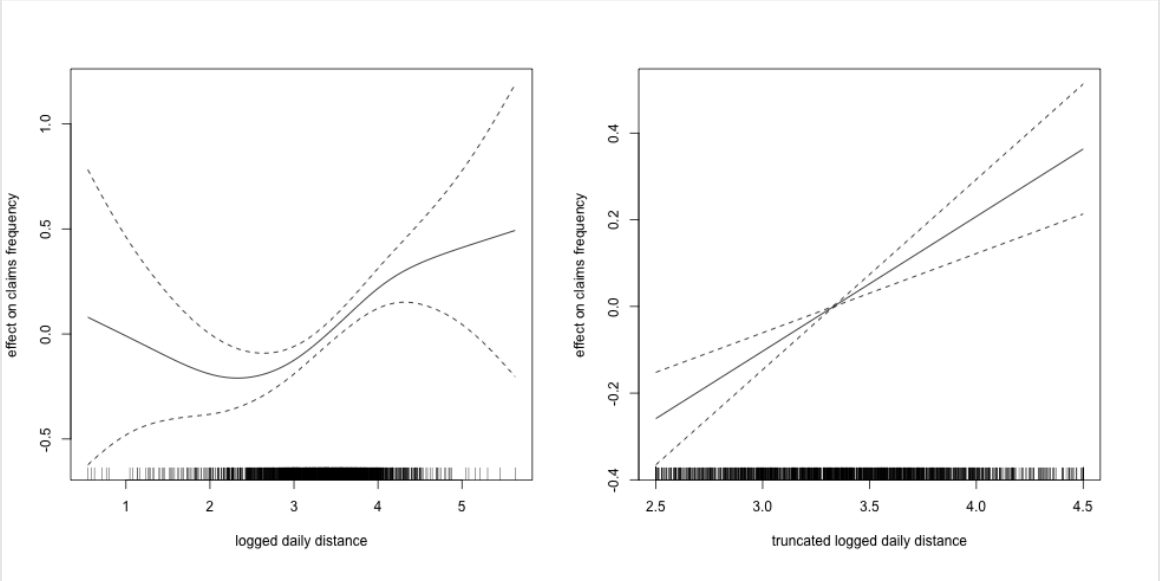
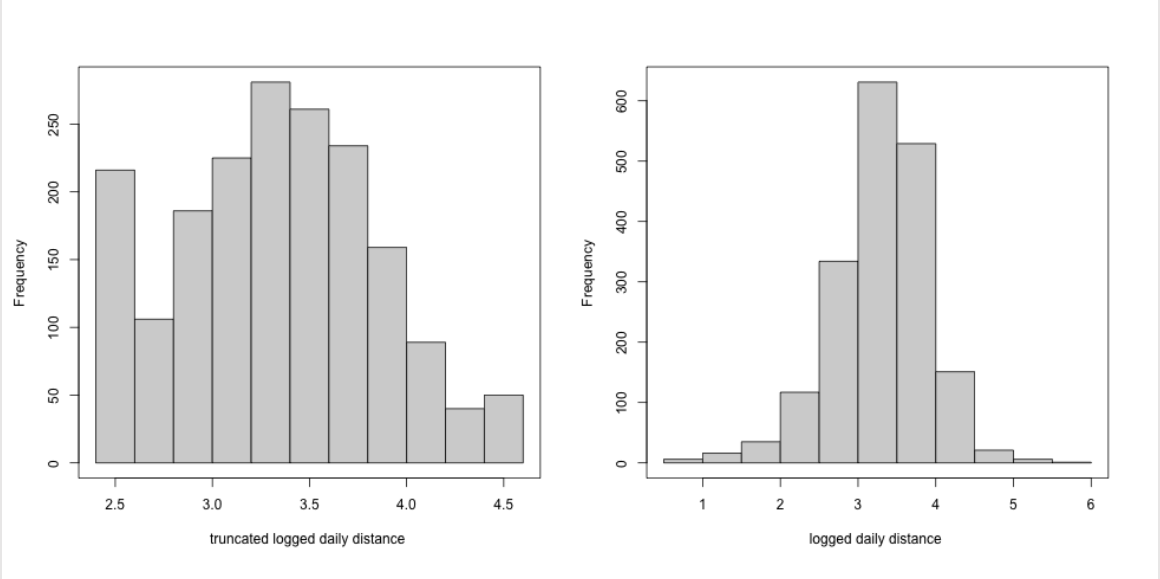


Figure 3.8
THE DISTRIBUTION OF LOGGED DAILY DISTANCE



3.2 GENERALIZED LINEAR MODEL FOR CLAIMS FREQUENCY

We begin with the classic generalized linear model (GLM), which assumes the following log-linear mapping:

$$\ln\lambda^{(\mathrm{GLM})}(\mathbf{x}) = \beta_0 + \alpha_{\mathrm{region}} + \gamma_{\mathrm{age\_group}} + \zeta_{\mathrm{female}} + \delta_{\mathrm{car\_made}} + \beta_1 \mathrm{car\_age} + \beta_2 \mathrm{trun\_logged\_distance}, \tag{3.2}$$

where we assume Zhejiang Province, middle age 1, female, Germany made car as reference levels. The coefficients are estimated using the learning data $\mathcal{D}_{\mathrm{learn}}$. We then perform a step-wise variable selection according to the Akaike information criterion (AIC). The final model is selected as follows:

$$\ln\lambda^{(\mathrm{GLM})}(\mathbf{x}) = \beta_0 + \alpha_{\mathrm{hebei}} + \gamma_{\mathrm{young}} + \gamma_{\mathrm{middle1}} + \gamma_{\mathrm{mature2}} + \delta_{\mathrm{china}} + \delta_{\mathrm{eu}} + \beta_1 \mathrm{car\_age} + \beta_2 \mathrm{trun\_logged\_distance}, \tag{3.3}$$

Hence, we have merged Shanghai Province and other regions with Zhejiang Province, middle age 2 with mature age 1, and car made in all countries except those in China and Europe. We have removed the gender and the car's price from the full model (3.2). The estimated coefficients are listed in Table 3.3. The out-of-sample Poisson deviance loss (3.1) for model (3.3) is $\mathbf{1.0306}$, compared with $\mathbf{1.1003}$ for a homogeneous model

$$\ln\lambda^{(\mathrm{Homo})}(\mathbf{x}) = \beta_0. \tag{3.4}$$

Table 3.3
THE ESTIMATED COEFFICIENTS IN THE GLM (3.3)

coefficients	estimate	std. error	z-value	p-value
β_0	-2.7864	0.2741	-10.1648	0.0000
α_{hebei}	-0.5320	0.0864	-6.1591	0.0000
γ_{young}	0.2764	0.1111	2.4874	0.0129
$\gamma_{\mathrm{middle1}}$	0.3719	0.1023	3.6371	0.0003
$\gamma_{\mathrm{mature2}}$	0.2656	0.0939	2.8277	0.0047
δ_{china}	0.1411	0.0905	1.5589	0.1190
δ_{eu}	0.3885	0.2053	1.8920	0.0585
β_1	0.0295	0.0170	1.7344	0.0829
β_2	0.3283	0.0734	4.4721	0.0000

3.3 IMPROVED GLM WITH BOOSTING METHODS

Next we explore the possibility of improving GLM using either *generalized boosted regression model* (GBM) or *XGBoost*. The mapping λ from actuarial risk factors to claims frequency is assumed as follows:

$$\ln\lambda(\mathbf{x}) = \ln\lambda^{(\mathrm{GLM})}(\mathbf{x}) + \ln\lambda^{(\mathrm{BST})}(\mathbf{x}), \tag{3.5}$$

where we keep the GLM estimate $\lambda^{(\mathrm{GLM})}(\mathbf{x})$ during the calibration of boosting model $\lambda^{(\mathrm{BST})}(\mathbf{x})$. The boosting modification factor $\lambda^{(\mathrm{BST})}(\mathbf{x})$ captures the non-linear effects and interaction among the covariates $\mathbf{x}$, which are omitted in the GLM (3.3). We include *region*, *driver's age* (continuous variable), *age group*, *gender*, *car made*, *car's age*, (logged) *car's price* and (logged) *average daily distance* into the boosting model $\lambda^{(\mathrm{BST})}$.

3.3.1 GENERALIZED BOOSTED REGRESSION MODELING

R code of the generalized boosted regression model is listed as follows:

Listing 1

R CODE FOR GENERALIZED BOOSTED REGRESSION.

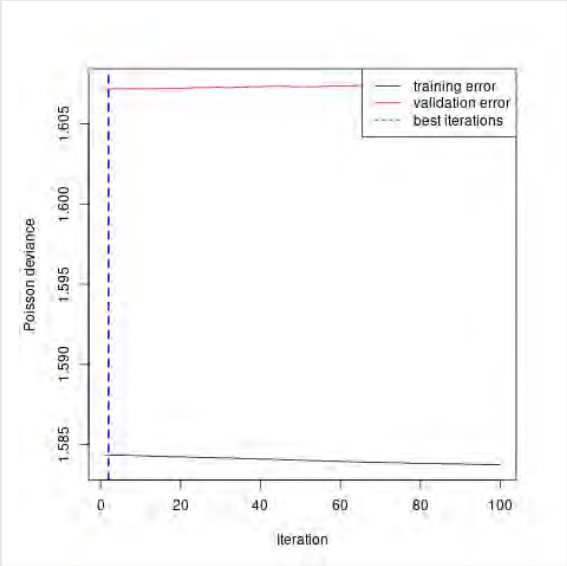
```
gbm1 <-
       gbm(Claim_Count ~  BRANCHNAME +  AGE + AGE_G + SEX +
                      Car_Made + USEYEARS + Price_log +
                      Daily_log + offset(log(Fit_GLM)),
              data = rbind(train_data, valid_data),
              distribution = "poisson",
              n.trees = 100,
              shrinkage = 0.001,
              interaction.depth = 1,
              n.minobsinnode = 100,
              bag.fraction = 0.5,
              train.fraction = nrow(train_data) / nrow(learn_data),
              cv.folds = 0,
              verbose = T
       )
(best.iter <- gbm.perf(gbm1, method = "test"))
gbm1$valid.error[best.iter]
```

Most of the code is self-explanatory. Some illustrations of important arguments are further described as follows.

- *offset (log(Fit_GLM))* indicates that the GBM starts boosting from the GLM prediction $\ln e\hat{\lambda}^{(\mathrm{GLM})}$. Hence we learn from the area not explored by the GLM.
- *N.trees* is the number of iterations (trees) we tend to boost. *shrinkage* is the learning step size. Normally, these two variables are inversely related. It is suggested that using a small learning setp size and a large amount of iterations will lead to a better out-of-sample performance
- *Interaction.depth* is the depth of weak learner of tree. Depth of 1 implies that we do not consider the interaction term. This variable needs to be tuned using validation error.
- *N.minobsinnode* is the minimal observations in a leaf node.
- *Bag.fraction* is the proportion of training data used to grow the trees.
- *Train.fraction* indicates that the first proportion of *data* are used as training data and the rest as validation data.

We tune the parameters ***shrinkage, interaction.depth*** by the changes in validation error. It turns out that ***interaction.depth=1*** leads to the minimal validation error and ***shrinkage*** does not affect the results too much. The calibration is displayed in Figure 3.9, which implies that there is little improvement by employing the GBM. The out-of-sample Poisson deviance loss for GBM is 1.0306, almost identical to that 1.0306 for GLM.

Figure 3.9
CALIBRATION OF GRADIENT BOOSTING MODEL



3.3.2 XGBOOST

Alternatively, we apply the XGBoost to improve the GLM prediction. R code is provided as follows:

Listing 2

R CODE FOR XGBOOST.

```
bst<-
       xgb.train(
              data = dtrain,
              watchlist = list(train = dtrain, test = dvalid),
              objective = "count:poisson",
              nrounds = 1000,
              eta = 0.001,
              max_depth = 2,
              min_child_weight = 100,
              subsample = 1,
              early_stopping_rounds = 5,
              nthread = 4,
              verbose = F
       )
bst$best_ntreelimit
bst$best_msg
```

Most of the code is self-explanatory. Some illustrations of important arguments are further described as follows.

- *nrounds, eta, max_depth, min_child_weight, subsample* play similar roles as *n.trees*, *shrinkage*, *interaction.depth*, *n.mnobsinnode*, *bag.fraction* in GBM.
- *early_stopping_rounds = 5* indicates that if the validation error on *dvalid* does not improve for 5 iterations the calibration will stop.

Again, there is no obvious improvement by employing XGBoost (with a test error of 1.0308). **We conclude that** the GLM (3.3) has sufficiently captured the prediction power of actuarial risk factors. It may be attributed to the appropriateness of our pre-processing of variables and the non-existence of significant interaction effects. Remark that our data only contains $n = 1847$ cars with 4215 years exposure, so we may not discover potential non-linear effects and interaction effects based on such a small portfolio.

3.4 IMPROVED GLM WITH RISK SCORES OF INDIVIDUAL TRIPS

In this section, we employ one-dimensional convolutional neural networks (1D CNNs) to evaluate driving risk associated with individual trips. 1D CNNs are suitable to learn patterns from time series data. Gao and Wüthrich (2019) have studied the usefulness of 1D CNNs for driver identification rather than driving risk evaluation of individual trips.

Our proposed method includes three steps: First, in Section 3.4.1 we select **archetypal drivers** and label their trips as either **potential risky or potential safe**. Second, in Section 3.4.2 we calibrate a 1D CNN to classify those labeled trips of archetypal drivers. The sigmoid probability of output neuron is defined as **risk score** of a trip. Third, in Section 3.4.3 we employ the calibrated 1D CNN to evaluate risk score of all trips and calculate **average risk score** for each driver. We improve the GLM (3.3) with average risk score.

3.4.1 SELECTION OF ARCHETYPAL DRIVERS

Recall that our purpose is to improve the GLM prediction using individual trips risk scores. **Ideally, average risk score of each driver should explain some variations in residuals of the GLM (3.3).** This motivates how we select archetypal drivers and label their trips.

We calculate the deviance residuals of the Poisson GLM (3.3) as follows:

$$r_i = \text{sign}\left(N_i - e_i\hat{\lambda}^{(\text{GLM})}(\mathbf{x}_i)\right)\sqrt{2e_i\hat{\lambda}^{(\text{GLM})}(\mathbf{x}_i) - 2N_i - 2N_i\ln\left(e_i\hat{\lambda}^{(\text{GLM})}\right) + 2N_i\ln N_i}.$$

We draw the histogram of $r_i, i = 1. \ldots, n$ in Figure 3.10.

We select 10 drivers with the largest deviance residuals in the learning data $\mathcal{D}_{\text{learn}}$ as **archetypal risky drivers**, and label their 5000 trips as potential risky trips (coded as 1). For those archetypal risky drivers, their exposure, number of claims, region, age, deviance residuals, and estimated claims frequency are listed in Table 3.4.

We select 10 drivers with no claim and the largest exposure in the learning data $\mathcal{D}_{\text{learn}}$ as **archetypal safe drivers**, and label their 5000 trips as potential safe trips (coded as 0). For those archetypal safe drivers, their exposure, number of claims, region, age, deviance residuals, and estimated claims frequency are listed in Table 5. We denote the index set of the archetypal drivers by $\mathcal{I}_{\text{sel}}$.

Note that we do not select drivers with the smallest residuals. The reason is that the corresponding drivers to this criterion will be those with high estimated claims frequency but no claims. Such drivers can be potentially risky drivers, since even the claims frequency is high, the chance of making no claim is also high. **Our failed trials show that if we chose drivers with the smallest residuals, we would not calibrate a useful 1D CNN for the ultimate goal of improving the GLM.**

Figure 3.10
HISTOGRAM OF DEVIANCE RESIDUALS

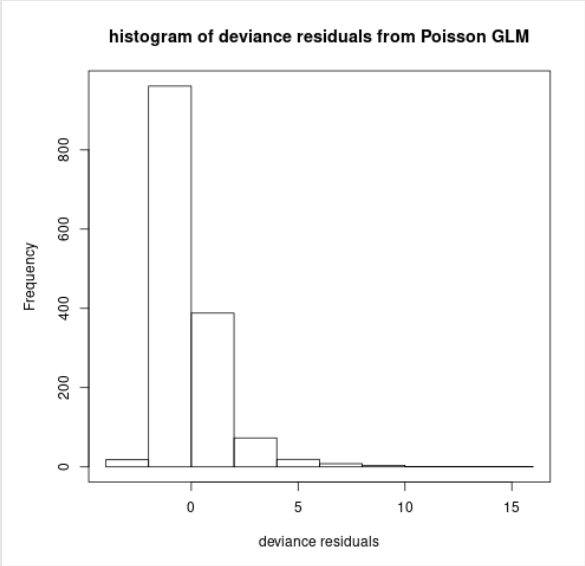


Table 3.4
10 ARCHETYPAL RISKY DRIVERS

e_i	N_i	region_i	age_i	r_i	$\hat{\lambda}(\mathbf{x}_i)^{(\text{GLM})}$
3.1836	4	other_regions	33.5	7.7461	0.2070
2.9096	4	hebei	28.0	8.8773	0.1917
2.0000	4	zhejiang	45.0	9.0390	0.2724
2.7918	3	hebei	56.0	6.7977	0.1458
2.2055	4	zhejiang	29.0	7.7946	0.2966
2.5288	3	hebei	50.0	8.3549	0.1200
3.0000	4	zhejiang	34.0	10.0270	0.1576
1.0000	2	shanghai	40.0	5.2395	0.2218
2.9507	3	hebei	29.0	5.5946	0.1748
2.0000	3	shanghai	36.0	6.4458	0.2179

Table 3.5
10 ARCHETYPAL SAFE DRIVERS

e_i	N_i	region$_i$	age$_i$	r_i	$\hat{\lambda}(\mathbf{x}_i)^{(GLM)}$
3.4438	0	hebei	32.5	-1.4907	0.2164
3.4795	0	hebei	39.5	-0.8265	0.1188
3.4658	0	zhejiang	29.5	-2.1375	0.3084
3.4466	0	zhejiang	62.0	-1.6836	0.2442
3.4466	0	hebei	33.5	-0.7848	0.1139
3.4329	0	shanghai	44.5	-1.5932	0.2321
3.4438	0	hebei	30.5	-1.2430	0.1805
3.4795	0	hebei	35.5	-0.7124	0.1024
3.4274	0	hebei	43.5	-1.2207	0.1781
3.4603	0	hebei	23.5	-0.9866	0.1426

3.4.2 ONE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK

We add two more telematics variables, squared acceleration rates a^2 and squared angle changes Δ^2 to time series of individual trips. We denote the jth trip of driver $i \in \mathcal{I}_{\text{sel}}$ by $z_{i,j} \in [-1,1]^{300\times 5}$ for $j = 1, \ldots, 500$. Note that we have normalized telematics variables $v, a, \Delta, a^2, \Delta^2$, using the min-max normalization. We label the trips of 10 archetypal risky drivers as 1 and those of archetypal safe drivers as 0. We split all trips $(z_{i,j})_{i\in\mathcal{I}_{sel}, j=1:500}$ into training data $(z_{i,j})_{i\in\mathcal{I}_{sel}, j=1:300}$, validation data $(z_{i,j})_{i\in\mathcal{I}_{sel}, j=301:400}$, and test data $(z_{i,j})_{i\in\mathcal{I}_{sel}, j=401:500}$.

Using R interface to *keras* and *tensorflow*, we calibrate a 1D CNN ϕ to classify the trips of selected archetypal drivers:

$$\phi: [-1,1]^{300\times 5} \to (0,1), \quad z \mapsto \phi(z). \tag{3.6}$$

The 1D CNN is constructed using the *keras* code shown in Appendix 5 with *q1=32, q2=16, q3=8, L=300, d=5*. The structure of 1D CNN is as follows:

Listing 3

THE STRUCTURE OF 1D CNN.

```
Layer (type)                 Output Shape              Param #
=================================================================
trips (InputLayer)           [(None, 300, 5)]          0
_________________________________________________________________
cov1 (Conv1D)                (None, 294, 32)           1152
_________________________________________________________________
ave1 (AveragePooling1D)      (None, 58, 32)            0
_________________________________________________________________
cov2 (Conv1D)                (None, 52, 16)            3600
_________________________________________________________________
ave2 (GlobalAveragePooling1D)  (None, 16)              0
_________________________________________________________________
dropout (Dropout)            (None, 16)                0
_________________________________________________________________
dense1 (Dense)               (None, 8)                 136
_________________________________________________________________
```

```
dropout_1 (Dropout)              (None, 8)                   0
_________________________________________________________________
dense2 (Dense)                   (None, 1)                   9
=================================================================
Total params: 4,897
Trainable params: 4,897
Non-trainable params: 0
_________________________________________________________________
```

Layer *cov1* extracts 32 features from a sliding window of 7 seconds. Layer *ave1* calculates average of each feature for every 5 seconds without overlapping. Layer *cov2* further extracts 16 features from a sliding window of 7 seconds. Layer *ave2* calculates average of each feature. Layers *dense1* and *dense2* explore interaction among 16 features and return the output in $(0,1)$ by using a sigmoid activation function in layer *dense2*. Dropout layers are inserted to reduce overfitting. The sigmoid probability of output neuron $\phi(z)$ is defined as **risk score** of trip z. If output neuron $\phi(z)$ is closer to 1, then this trip receives a higher risk score.

The calibration is shown in Figure 3.11. The hyperparameters in Listing 3 are tuned by monitoring validation loss on $\left(z_{i,j}\right)_{i\in\mathcal{I}_{sel},j=301:400}$. The network weights resulting in the lowest validation loss are then saved for the rest of modelling. The validation accuracy is around 70% and the test accuracy is at the same level. The calibrated 1D CNN is employed to evaluate risk score of test trips $\left(z_{i,j}\right)_{i\in\mathcal{I}_{sel},j=401:500}$. We compare the distribution of risk scores for test potential risky trips with that for test potential safe trips in Figure 3.12. The test potential safe trips tend to have lower risk scores than those risky ones.

Figure 3.11
CALIBRATIONS OF THE 1D CNN

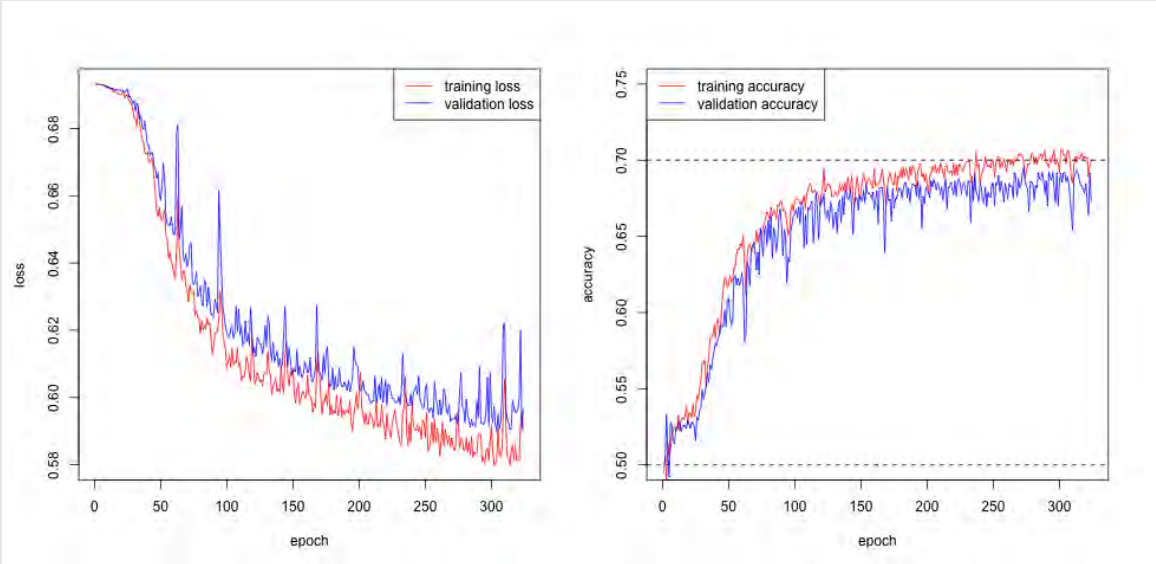
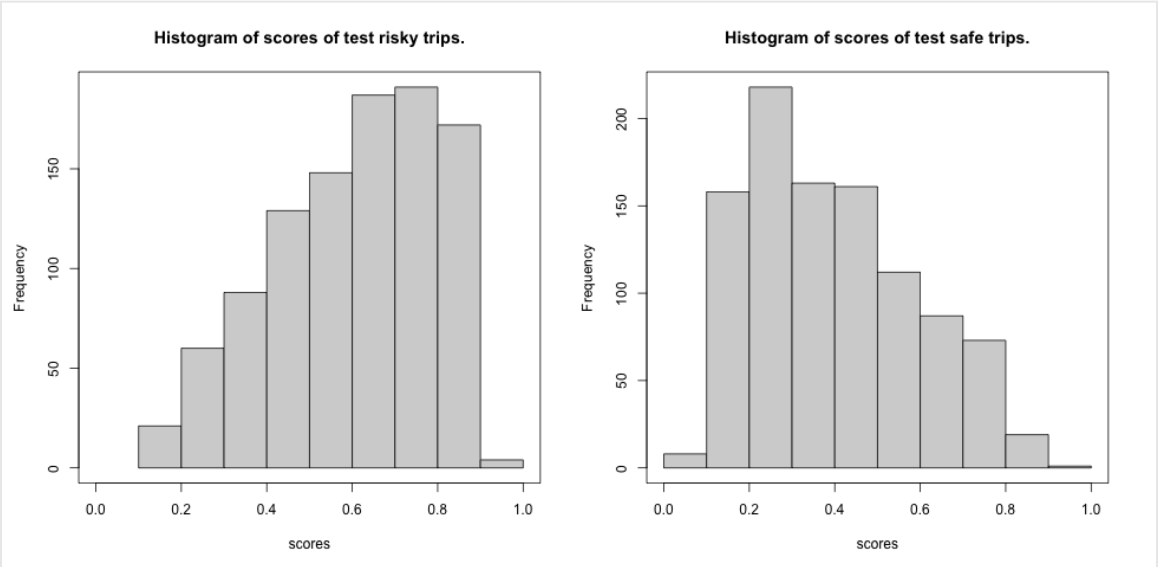


Figure 3.12
RISK SCORES OF TEST TRIPS $\left(z_{i,j}\right)_{i\in\mathcal{I}_{sel},j=401:500}$.



Remark: In failed trials, we find that hyperbolic tangent activation is better than relu activation function, and average pooling is better than max pooling. Including squared acceleration rates and squared angle changes can improve prediction accuracy of 1D CNN. The disadvantage of 1D CNN is that it is a black box and we do not know how it performs feature engineering.

Remark: One may choose the entire trips of 2 risky drivers and 2 safe drivers as test data. **According to our failed trials, the neural network cannot be calibrated on such a partition, i.e., the validation error cannot be reduced.** The reason may be that some safe drivers could still have many risky trips and vice versa.

3.4.3 IMPROVED GLM WITH AVERAGE RISK SCORES

The calibrated 1D CNN is now employed to evaluate risk scores of up to 200 individual trips for each driver. We calculate average risk scores (*ave_risk_score*) for each driver. We improve the GLM (3.3) with average risk score as follows:

$$\ln\lambda(\mathbf{x}) = \ln\lambda^{(\text{GLM})}(\mathbf{x}) + \ln\lambda^{(\text{TEL})}(\mathbf{x}) \tag{3.7}$$

where the (logged) **telematics modification factor** is

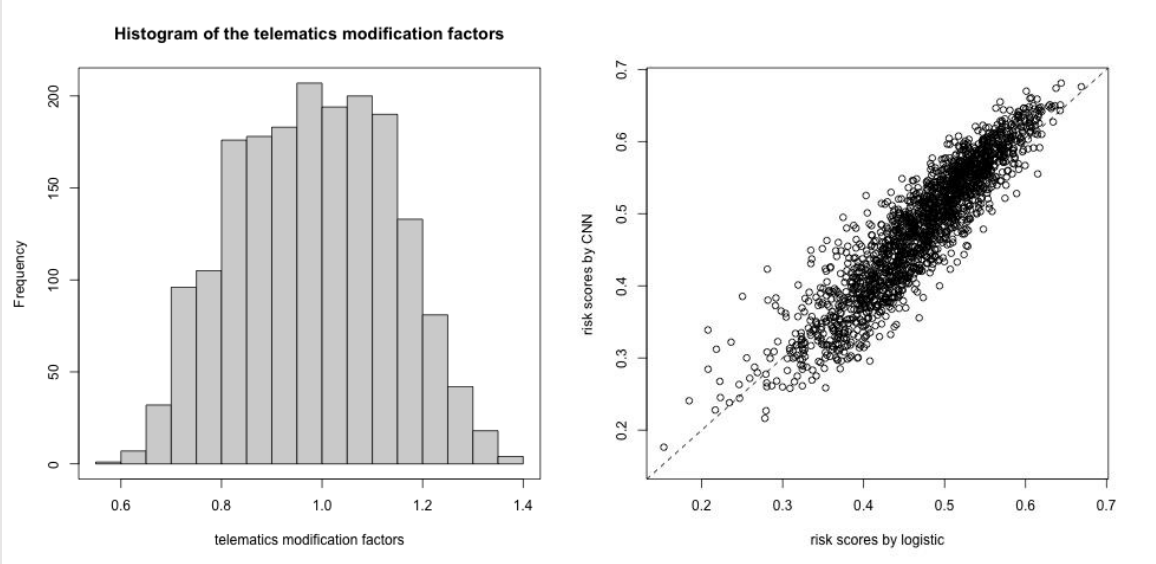
$$\ln\lambda^{(\text{TEL})}(\mathbf{x}) = \beta_3 + \beta_4 \text{ave\_risk\_score}.$$

Note that we have expanded the covariate space $\mathcal{X}$ to include average risk score. It turns out that the out-of-sample Poisson deviance loss for model (3.7) is 1.0286, comparing with 1.0306 for the GLM (3.3). The estimated telematics modification factor $\lambda^{(\text{TEL})}$ is

$$\exp\left(\hat{\beta}_3 + \hat{\beta}_4 \text{ave\_risk\_score}\right) = \exp(-0.6878 + 1.3712 \times \text{ave\_risk\_score})$$

Our calculated average risk scores for test drivers $\mathcal{I}_{\text{test}}$ are in $(0.1511, 0.7837)$, so the telematics modification factors fall in $(0.6184, 1.4722)$. We plot the histogram of telematics modification factors in Figure 3.13.

Figure 3.13
THE HISTOGRAM OF TELEMATICS MODI_CATION FACTORS FOR TEST DRIVERS $\mathcal{I}_{\text{TEST}}$. COMPARISON OF AVERAGE RISK SCORES FOR TEST DRIVERS $\mathcal{I}_{\text{TEST}}$ OBTAINED FROM 1D CNN AND LOGISTIC REGRESSION



Remark: The selection of archetypal drivers is the key to successfully calibrate a 1D CNN. Our selection discussed above is the only viable one among all the investigated selections. For instance, one may consider to calibrate a Poisson 1D CNN on all drivers' trips with claims counts as response variable. However, the calibration fails when using either individual trips or connected individual trips.

3.5 ALTERNATIVE METHODS

In this section, we discuss two alternative methods to the proposed method: (1) using telematics summary statistics; and (2) replacing 1D CNN by a logistic regression. An important finding is that given sufficient manually engineered features the second alternative method has a similar out-of-sample prediction performance.

3.5.1 TRIPS CLASSIFICATION BY A LOGISTIC MODEL

We replace the 1D CNN by a logistic regression. The specification and calibration of logistic model are much simpler than that of 1D CNN. The manually engineered features (covariates) denoted by $\mathbf{s}_{i,j} \in \mathbb{R}^{35}$ include the mean, median, minimum, maximum, 10% and 90% quantiles and standard deviation of speed v, acceleration a, angle change Δ, squared acceleration a^2, and squared angle change Δ^2 of each trip $\mathbf{z}_{i,j}$. The test accuracy is 66.7%, which is not far from that of 1D CNN 70.5%.

It may be concluded that the logistic regression also captures the major difference between potential risky trips and potential safe trips by using sufficient summary statistics. **This might be explained by the fact that the chronological property in individual trips does not play a critical role in trips classification.**

3.5.2 TWO ALTERNATIVE METHODS

We consider two alternative methods, both of which need sufficient manually engineered features.

Alternative 1: Averaged telematics summary statistics. We calculate average of above derived summary statistics $\boldsymbol{s}_{i,j}$ over all the trips j for each driver i:

$$\boldsymbol{s}_i = \frac{1}{c_i}\sum_{j=1}^{c_i} \boldsymbol{s}_{i,j}.$$

Those averaged telematics summary statistics $\mathbf{s}_{\mathrm{i}}$ are then used as the covariates in the Poisson GLM for claims frequency. We conduct a step-wise variable selection according to the AIC. The test Poisson deviance loss is 1.0379 which is even worse than that of the GLM (1.0306).

Alternative 2: Risk scoring with a logistic regression. This alternative method follows the same procedure as the proposed method, but replaces the 1D CNN by a logistic regression as discussed in Section 3.5.1. This alternative method uses manually engineered features $\boldsymbol{s}_{i,j}$, whereas our proposed method uses automatically engineered features by the neural network. The test Poisson deviance loss is 1.0287 for this alternative method, which is close to that of the proposed method (1.0286). Figure 3.13 compares average risk scores for test drivers $\mathcal{I}_{\text{test}}$ from the logistic regression and the 1D CNN. Their correlation is 0.9181.

3.6 SENSITIVITY ANALYSIS

Since our portfolio size is small, we conduct a sensitivity analysis to test the robustness of our results. Similar to cross-validation, we repeat the above analysis for 5 times and evaluate test Poisson deviance loss on 5 mutually exclusive test data for the GLM (3.3), homogeneous model 3.4, proposed method 3.7, and two alternative methods, respectively. The results are shown in Table 6. Note that the above analysis corresponds to test index 2.

Table 3.6
TEST POISSON DEVIANCE LOSS FOR DIFFERENT DATA PARTITIONS

test index	homo.	glm	proposed	alt_1	alt_2
1	1.1095	1.0981	**1.0918**	1.1071	1.0952
2	1.1003	1.0306	**1.0286**	1.0379	1.0287
3	1.0949	1.0641	1.0439	**1.0190**	1.0373
4	1.0952	1.0721	1.0654	1.0889	**1.0617**
5	1.0996	1.0318	1.0268	1.0975	**1.0262**

We conclude that alternative 2 performs as well as our proposed method, and alternative 1 is the least robust. Despite its simple specification and calibration, alternative 2 relies on manually engineered features which requires thorough considerations. In contrast, our proposed method learns the representations automatically from time series of individual trips.

Section 4: Conclusions

We investigate telematics variables from three independent sensors: GPS, instrumental panel and accelerometer. Several visual tools are developed to reveal data issues. In our data, accelerometer has calibration bias, which leads to systematic data error. We distinguish between idle phase and moving phase, and idle phase should be removed in the analysis. We consider three formats of telematics data: time series, summary statistics and heatmaps. Data imputation is not crucial for the last two formats. In call cases, we must specify particular speed intervals and truncate acceleration and angle change. A particular speed interval makes the comparison among drivers more sensible, and truncation eliminates the leverage effects of outliers.

In claims frequency modeling, we propose a risk scoring 1D CNN for individual trips. The 1D CNN is calibrated on binary labeled trips of selected archetypal risky and safe drivers. The classic Poisson GLM for claims frequency is considerably improved by incorporate average risk score. Comparison with two alternative methods indicates that the chronological property in individual trips does not play a critical role in trips classification. As 1D CNN, the logistic regression has also captured important features from individual trips related to claims frequency. Nevertheless, with 1D CNN, we do not need to perform a thorough manually feature engineering as alternative methods, which is an advantage of the proposed method.

Future research may relax the constraint of fixed length individual trips by employing recurrent neural network. Another limit is that we assign the same weight to average risk scores without considering number of individual trips.

Acknowledgments

This research was supported by a grant from the Society of Actuaries. We are grateful to the supervision from the Project Oversight Group including Rob Montgomery, Emil Valdez, Rasa McKean, Syed Danish Ali, Xuefeng Wang, Andrea Marcovici, Yazhan Lin, Xin Tang. We are grateful to Mario Wüthrich for suggestion of archetypal drivers. Guangyuan Gao gratefully acknowledges financial support from the National Natural Science Foundation of China (71901207).

References

Ayuso, M., Guillen, M., Pérez-Marín, A.M. 2016a. Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks* 4/2, article 10.

Ayuso, M., Guillen, M., Pérez-Marín, A.M. 2016b. Using GPS data to analyse the distance traveled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies* 68, 160-167.

Boucher, J.-P., Côté, S., Guillen, M. 2017. Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks* 5/4, article 54.

Carvalho, E. 2017. Exploiting the use of recurrent neural networks for driver behavior profiling. *2017 International Joint Conference on Neural Networks*, 3016-3021.

Cather, D.A. 2020. Reconsidering insurance discrimination and adverse selection in an era of data analytics. *Geneva Papers on Risk and Insurance - Issues and Practice* 45, 426–456.

Cevolini, A., Esposito, E. 2020. From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data and Society* 7/2.

Denuit, M., Guillen, M., Trufin, J. 2019. Multivariate credibility modelling for usage-based motor insurance pricing with behavioral data. *Annals of Actuarial Science* 13/2, 378-399.

Gao, G., Wang, H, Wüthrich, M.V. 2021. Boosting Poisson regression models with telematics car driving data. *Machine Learning*.

Gao, G., Wüthrich, M.V. 2018. Feature extraction from telematics car driving heatmaps. *European Actuarial Journal* 8/2, 383-406.

Geyer, A., Kremslehner, D., Muermann, A. 2020. Asymmetric information in automobile insurance: Evidence from driving behavior. *Journal of Risk and Insurance* 87/4, 969-995.

Girma, A., Yan, X., Homaifar, A. 2019. Driver identification based on vehicle telematics data using LSTM-recurrent neural network. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence*, 894-902.

Guillen, M., Nielsen, J.P., Pérez-Marín, A.M., Elpidorou, V. 2020. Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal* 24/1, 141-152.

Ho, S.-H., Wong, Y.-D., Chang, V. W.-C. 2014. Developing Singapore driving cycle for passenger cars to estimate fuel consumption and vehicular emissions. *Atmospheric Environment* 97, 353-362.

Hu, X., Zhu, X., Ma, Y.-L., Chiu, Y.-C., Tang, Q. 2019. Advancing usage-based insurance – a contextual driving risk modelling and analysis approach. *IET Intelligent Transport Systems* 13/3, 453-460.

Huang, Y., Meng, S. 2019. Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems* 127.

Hung, W.T., Tong, H.Y., Lee, C.P., Ha, K., Pao, L.Y. 2007. Development of practical driving cycle construction methodology: a case study in Hong Kong. *Transportation Research Part D: Transport and Environment* 12/2, 115-128.

Joubert, J. W., de Beer, D., de Koker, N. 2016. Combining accelerometer data and contextual variables to evaluate the risk of driver behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour* 41/A, 80-96.

Kamble, S.H., Mathew, T.V., Sharma, G.K. 2009. Development of real-world driving cycle: case study of Pune, India. *Transportation Research Part D: Transport and Environment* 14/2, 132-140.

Lemaire, J., Park, S.C., Wang, K. 2016. The use of annual mileage as a rating variable. *ASTIN Bulletin* 46, 39-69.

Ma, Y.-L., Zhu, X., Hu, X., Chiu, Y.-C. 2018. The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice* 113, 243-258.

Mahapatra, A. 2019. Privacy: A growing risk in the insurance industry. *Creative Components* 410.

Paefgen, J., Staake, T., Fleisch, E. 2014. Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice* 61, 27-40.

Richman, R. 2020. AI in actuarial science - a review of recent advances. *Annals of Actuarial Science*.

Romera, E., Bergasa, L.M., Arroyo, R. 2016. Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. *2016 IEEE 19th International Conference on Intelligent Transportation Systems*, 387-392.

Savelonas, M. et al. 2020. Hybrid time-series representation for the classification of driving behaviour. *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization*, 1-6.

Verbelen, R., Antonio, K., Claeskens, G. 2018. Unraveling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C Applied Statistics* 67, 1275-1304.

Wahlström, J., Skog, I., Händel, P. 2017. Smartphone-Based Vehicle Telematics: A Ten-Year Anniversary. *IEEE Transactions on Intelligent Transportation Systems* 18/10, 2802-2825.

Wahlström, J., Skog, I., Händel, P. 2015. Detection of Dangerous Cornering inGNSS-Data-Driven Insurance Telematics. *IEEE Transactions on Intelligent Transportation Systems* 16/6, 3073-3083.

Wahlström, J., Skog, I., Nordström, R.L., Händel, P. 2018. Fusion of OBD and GNSS Measurements of Speed. *IEEE Transactions on Instrumentation and Measurement* 67/7, 1659-1667.

Weidner, W., Transchel, F.W.G., Weidner, R. 2016. Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal* 6/1, 3-24.

Weidner, W., Transchel, F.W.G., Weidner, R. 2017. Telematic driving profile classification in car insurance pricing. *Annals of Actuarial Science* 11/2, 213-236.

Appendix A: Other formats of telematics data

We have considered time series of individual trips. Now we consider the other two formats of telematics data: **summary statistics** and **heatmap.** The data imputation is important for time series, rather than for the other two formats. Hence, a simpler data cleaning procedure may be performed to derive summary statistics and heatmap. For instance, only data with valid GPS variables are kept in this procedure.

In all cases, we must **specify particular speed intervals and truncate acceleration and angle change.** A particular speed interval makes the comparison among drivers more sensible, and truncation eliminates the leverage effects of outliers. The activation functions in neural networks play a similar role. Note that our focus is time series since there are already several existing literatures on the last two formats.

A.1 Summary statistics of telematics variables

We extract several summary statistics for each trip, which are categorized into two types: **exposure** and **driving style**. Two statuses of driving are defined: **idle phase and moving phase**. In idle phase, the speed, acceleration and angle change are simultaneously equal to zero. In moving phase, at least one of three variables is not equal to zero. We calculate the following summary statistics for 2296−165 = 2131 cars since 165 cars do not have telematics information.

Exposures

1. Total distance in km (1 variable).

2. Total driving time of moving phase in second (1 variable).

3. Total standing time of idle phase in second (1 variable).

4. The day of trip from Monday to Sunday (1 variable).

5. Driving distance split into morning peak (7-9), day (9-17), afternoon peak (17-19), evening

(19-22), and night (22-7) (5 variables).

6. Driving time of moving phase split into morning peak (7-9), day (9-17), afternoon peak

(17-19), evening (19-22), and night (22-7) (5 variables).

In Figure A.1, we plot the distribution of observation period, number of days with trips, driving hours per day and average daily driving distance.

Driving styles

1. Average/median/standard deviation/extreme values (5% and 95% quantiles) of speed/acceleration/ angle change during moving phase (15 variables).

2. Driving time of moving phase split into in the speed interval 0-10/10-20/20-30/. . . /110-120/>120 (13 variables).

3. Driving time of moving phase split into the acceleration intervals (-4)-(-3.5)/(-3.5)-(-2.5)/. . . /2.5-3.5/3.5-4 (9 variables).

4. Driving time of moving phase split into the angle change intervals (-45)-(-35)/(-35)-(-25)/. . . /25-35/35-45 (9 variables).

A.2 Heatmaps

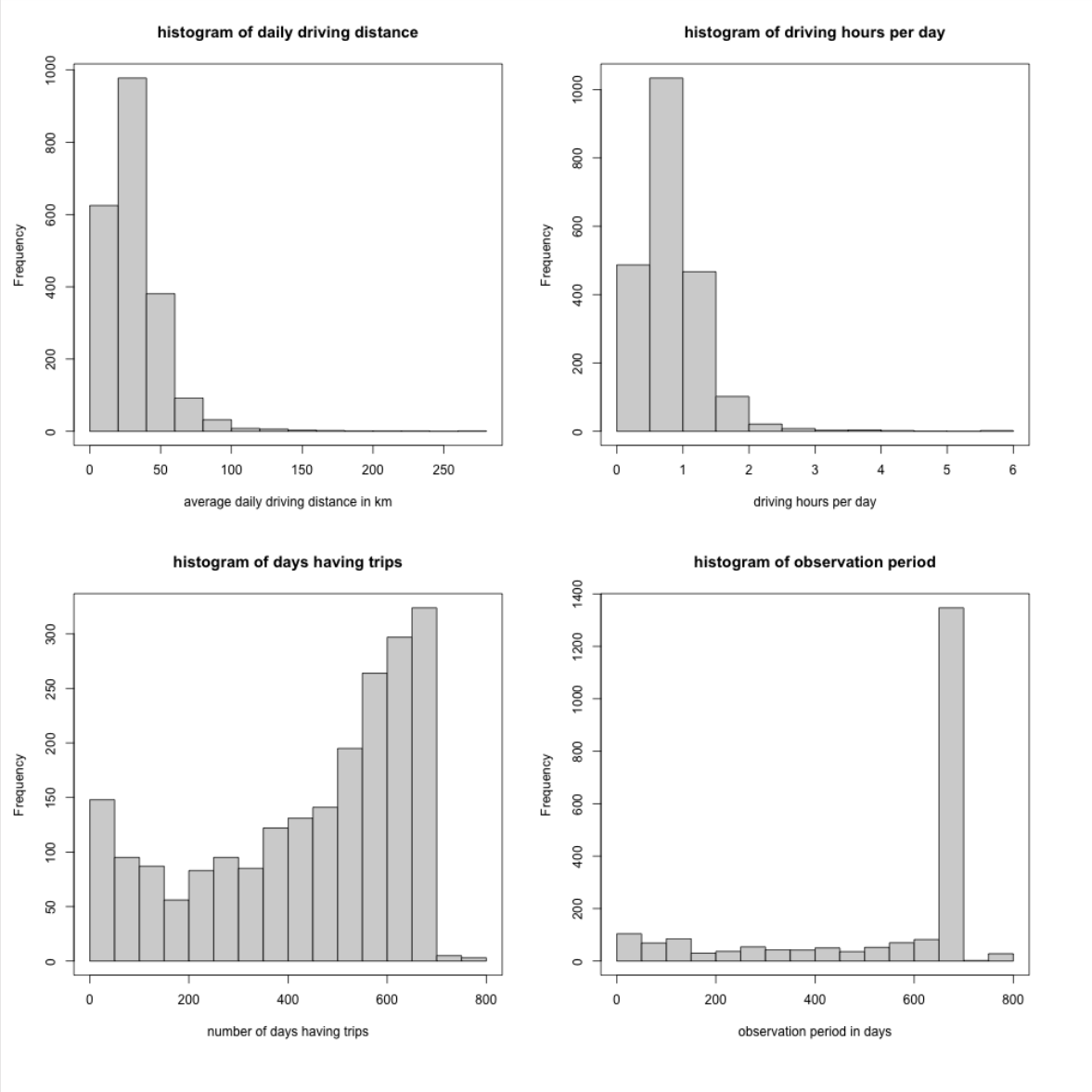
Heatmaps reflect the interaction among speed, acceleration, and angle change. For each car, we construct three heatmaps: **speed-acceleration, speed-angle, and acceleration-angle heatmaps.** We use telematics data in moving phase to construct heatmaps.

In heatmaps, the speed, acceleration and angle change intervals are divided into speed subintervals at 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120 and 200 km/h, and acceleration sub-intervals at -4, -3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5 and 4 m/s2, and angle change subintervals at -45, -35, -25, -15, -5, 5, 15, 25, 35, 45 degrees, respectively.

We typically truncate the speed interval and cap the acceleration and angle change intervals. The considered intervals should contain sufficient observations to receive stable heatmaps. Moreover, we need to normalize the heatmap either in each sub-interval or in several consecutive subintervals.

Figure A.1

THE DISTRIBUTION OF TELEMATICS EXPOSURE



Appendix B: Keras code for 1D CNN

Listing 4

KERAS CODE FOR 1D CNN IN LISTING 3

```
build_model_cnn2 <- function(q1, q2, q3, L, d) {
        ### input layer
        trips <-
        layer_input(shape = c(L, d),
        dtype = "float32",
        name = "trips")
        ### convolutional neural network
        trips_score = trips %>%
        layer_conv_1d(
        filters = q1,
        kernel_size = 7,
        activation = "tanh",
        name = "cov1"
        ) %>%
        layer_average_pooling_1d(pool_size = 5, name = "ave1") %>%
        layer_conv_1d(
        filters = q2,
        kernel_size = 7,
        activation = "tanh",
        name = "cov2"
        ) %>%
        layer_global_average_pooling_1d(name = "ave2") %>%
        layer_dropout(rate = 0.5) %>%
        # layer_dropout(rate = 0.2) %>%
        layer_dense(
        units = q3,
        activation = "tanh",
        name = "dense1"
        ) %>%
        layer_dropout(rate = 0.2) %>%
        layer_dense(
        units = 1,
        activation = "sigmoid",
        weights = list(array(c(0), dim = c(q3, 1)), array(0, dim = c(1))),
        # units = 10,
        # activation = "softmax",
        # weights = list(array(c(0), dim = c(q3, 10)), array(0, dim = c(10))),
        name = "dense2"
        )

        ### compile model
        model <- keras_model(inputs = trips, outputs = trips_score)
        model %>% compile(
        optimizer = optimizer_adam(),
        loss = "binary_crossentropy",
        # loss = "categorical_crossentropy",
        metrics = c("accuracy")
        )
        model
}
```

About The Society of Actuaries

With roots dating back to 1889, the Society of Actuaries (SOA) is the world's largest actuarial professional organization with more than 31,000 members. Through research and education, the SOA's mission is to advance actuarial knowledge and to enhance the ability of actuaries to provide expert advice and relevant solutions for financial, business and societal challenges. The SOA's vision is for actuaries to be the leading professionals in the measurement and management of risk.

The SOA supports actuaries and advances knowledge through research and education. As part of its work, the SOA seeks to inform public policy development and public understanding through research. The SOA aspires to be a trusted source of objective, data-driven research and analysis with an actuarial perspective for its members, industry, policymakers and the public. This distinct perspective comes from the SOA as an association of actuaries, who have a rigorous formal education and direct experience as practitioners as they perform applied research. The SOA also welcomes the opportunity to partner with other organizations in our work where appropriate.

The SOA has a history of working with public policymakers and regulators in developing historical experience studies and projection techniques as well as individual reports on health care, retirement and other topics. The SOA's research is intended to aid the work of policymakers and regulators and follow certain core principles:

Objectivity: The SOA's research informs and provides analysis that can be relied upon by other individuals or organizations involved in public policy discussions. The SOA does not take advocacy positions or lobby specific policy proposals.

Quality: The SOA aspires to the highest ethical and quality standards in all of its research and analysis. Our research process is overseen by experienced actuaries and nonactuaries from a range of industry sectors and organizations. A rigorous peer-review process ensures the quality and integrity of our work.

Relevance: The SOA provides timely research on public policy issues. Our research advances actuarial knowledge while providing critical insights on key policy issues, and thereby provides value to stakeholders and decision makers.

Quantification: The SOA leverages the diverse skill sets of actuaries to provide research and findings that are driven by the best available data and methods. Actuaries use detailed modeling to analyze financial risk and provide distinct insight and quantification. Further, actuarial standards require transparency and the disclosure of the assumptions and analytic approach underlying the work.

Society of Actuaries
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org