Insurance: Mathematics and Economics 88 (2019) 108–119

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Evaluation of driving risk at different speeds

Guangyuan Gao[a,*], Mario V. Wüthrich[b], Hanfang Yang[a]

[a] *Center for Applied Statistics and School of Statistics, Renmin University of China, 100872 Beijing, China*
[b] *Risk Lab, Department of Mathematics, ETH Zurich, 8092 Zurich, Switzerland*

ARTICLE INFO

Article history:
Received November 2018
Received in revised form April 2019
Accepted 11 June 2019
Available online 21 June 2019

Keywords:
Telematics data
v-a heatmap
K-medoids algorithm
Principal components analysis
Generalized additive model
Variable selection
Poisson regression
Claims frequency modeling

ABSTRACT

Telematics car driving data describes drivers' driving characteristics. This paper studies the driving characteristics at different speeds and their predictive power for claims frequency modeling. We first extract covariates from telematics car driving data using K-medoids clustering and principal components analysis. These telematics covariates are then used as explanatory variables for claims frequency modeling, in which we analyze their predictive power. Moreover, we use these telematics covariates to challenge the classical covariates usually used in practice.

1. Introduction

Telematics car driving data can be collected by internal sensors installed in cars or by drivers' mobile phones. While the telematics data from internal sensors directly describes drivers' driving habits and driving styles, the telematics data from drivers' mobile phones mixes driving phases with non-driving phases such as walking, etc. Our telematics data is collected from internal sensors installed by a Chinese insurance company. Drivers are encouraged to install such devices in return for a premium discount. However, telematics data is not yet allowed to be used in car insurance tariffication under the current regulation in China, and most relevant insurance products such as usage-based insurance (UBI) and pay-as-you-drive (PAYD) products are forbidden by the Chinese insurance regulatory authority. Nevertheless, almost all the car insurance companies and many technology companies in China have started to accumulate telematics data to evaluate driving risk. The use of driving risk factors will refine insurer's risk classification system, and therefore encourage insured to drive more safely.

There have been many studies on telematics data in both the actuarial area (e.g., Ayuso et al., 2016; Paefgen et al., 2014; Verbelen et al., 2018) and the transportation area (e.g., Hung et al., 2007; Wang et al., 2008a). In the following, we list some papers which are mostly related to our work. A series of papers studies Spanish car driving data which contains PAYD policies issued to young drivers (Ayuso et al., 2016, 2018; Boucher et al., 2017; Guillen et al., 2019; Denuit et al., 2019). Ayuso et al. (2016) and Boucher et al. (2017) study the risk exposure to total driving distances and its interaction with gender and policy duration. Ayuso et al. (2018) and Guillen et al. (2019) study the risk factors associated with the percentages of kilometers driven above speed limits, the distances in urban area, and night driving. Denuit et al. (2019) consider the incorporation of these posterior telematics covariates into a credibility model. Paefgen et al. (2014) and Verbelen et al. (2018) partition the total driving distances by road type and time slots. Verbelen et al. (2018) further use these compositional covariates and classical risk factors in claims frequency models. A second stream of literature directly analyzes driving styles by studying acceleration patterns. Weidner et al. (2016,b) investigate driving styles using the discrete Fourier transform. Gao et al. (2019) study the interaction between speed and acceleration rate, represented by so called v-a heatmaps, in the low speed bucket $[5, 20]$ km/h.

The v-a heatmaps of Gao et al. (2019) describe the speed–acceleration pattern as a two-dimensional functional illustrated by a pixel matrix. This functional reflects how drivers accelerate or brake at different speeds (called as *driving style*). The chosen speed bucket $[5, 20]$ km/h in Gao et al. (2019) is rather empirical, and it is desirable to analyze different speeds in a more systematic way. In this paper we study how a driver behaves at different speeds, and whether driving styles at low speeds are more related to accidents. In particular, we implement

[*] Corresponding author.
E-mail address: guangyuan.gao@ruc.edu.cn (G. Gao).

https://doi.org/10.1016/j.insmatheco.2019.06.004
0167-6687/© 2019 Elsevier B.V. All rights reserved.

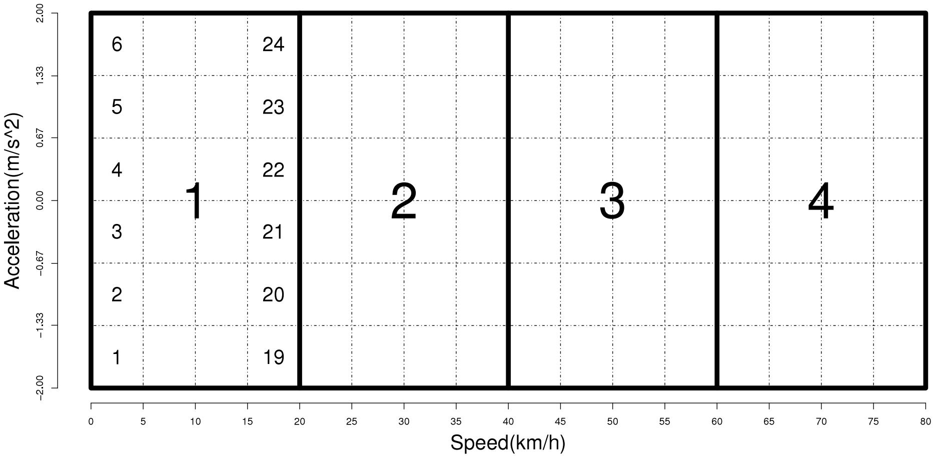

Fig. 1. The partition of $R = (0, 80] \times [-2, 2]$.

K-medoids clustering and principal components analysis to decompose the v-a heatmap functional in different speed buckets. We refer to Kaufman and Rousseeuw (1990) and Hastie et al. (2009) for the K-medoids clustering and its advantages/disadvantages compared with the K-means clustering; we refer to Hastie et al. (2009) for the principal components analysis. Moreover, we study the relative driving time at different speeds and the intensity of driving (called as *driving habit*), and we investigate whether certain driving habits are more related to accidents.

We use the driving style covariates and the driving habit covariates in generalized additive models (Wood, 2017) to describe the corresponding claims frequencies. Generalized additive models can address potential non-linear effects of these covariates. We then implement the backward elimination to select the variables. Three criteria are used to compare the models: the unbiased risk estimator (UBRE) (Section 4.5.4 of Wood, 2017), the AIC, and the average Poisson deviance loss. The UBRE and the AIC are measures of in-sample goodness-of-fits penalized by model complexity. We estimate the average Poisson deviance loss by 10-fold cross validation. Thus, the average Poisson deviance loss is a measure that mimics out-of-sample predictive performance; see Hastie et al. (2009).

The paper is structured as follows. Section 2 constructs the v-a heatmaps in different speed buckets, and it extracts the covariates using K-medoids clustering and principal components analysis. Section 3 establishes claims frequency models using both classical risk factors and the telematics covariates from Section 2. It then compares them with respect to UBRE, AIC and average Poisson deviance loss. Section 4 concludes the paper with our findings.

2. Telematics v-a heatmaps

Our telematics data collects the car's status second by second. That is, every second we receive the GPS location, the current speed and the acceleration in all directions from the internal sensor installed in the cars. We select the recorded speed and the recorded longitudinal acceleration to form the v-a heatmaps. In our analysis we consider the telematics data of $n = 973$ cars during three months of driving experience from 01/05/2016 to 31/07/2016. An assumption made here is that a driver's driving characteristics remain the same during his/her policy period, since we apply the same telematics covariates for all policies of a given driver.

2.1. Construction of v-a heatmaps in different speed buckets

We study the v-a rectangle denoted by $R = (0, 80]$ km/h $\times[-2, 2]$ m/s^2. Note that we have extended the previously analyzed speed interval [5, 20] km/h of Gao et al. (2019) to (0, 80] km/h because we want to study the driving styles both at low and at high speeds. Note that 95% of our drivers spend at least 79% of their total driving time in (0, 80] km/h. Also note that we cap the acceleration rate within $[-2, 2]$ m/s^2 because we do not have sufficiently many observations outside of this interval. We partition the speed interval (0, 80] km/h into four intervals at equal length, i.e., (0, 20] km/h, (20, 40] km/h, (40, 60] km/h, and (60, 80] km/h, see rectangles 1–4 in Fig. 1. This partition of the speed interval (0, 80] km/h to four sub-speed buckets is done because we would like to separate driving styles from driving habits. We consider normalized quantities on these sub-speed buckets which achieves to make drivers with different driving habits (e.g. commuting drivers vs. off-peak drivers) more comparable in driving styles. We choose the four speed buckets also because in China the speed limits in school zones, living areas, urban streets, and express ways are usually 20 km/h, 40 km/h, 60 km/h and 80 km/h, respectively.

2.1.1. Driving style and driving habit

For each speed bucket $m = 1, \dots, 4$, we further divide the v-axis (speed) into 4 intervals and the a-axis (acceleration) into 6 intervals, which results in 24 sub-rectangles $(R_{m,j})_{j=1:24}$ in each speed bucket m (see the numbers in speed bucket 1 in Fig. 1). For each driver i, we denote the amount of time spent in $R_{m,j}$ by $t_{i,m,j}$. Given a speed bucket m, for each driver i we calculate the relative amount (normalized amount) of time spent in $R_{m,j}$ as

$$z_{i,m,j} = \frac{t_{i,m,j}}{t_{i,m}} \geq 0, \tag{2.1}$$

where $t_{i,m} = \sum_{j=1}^{24} t_{i,m,j}$ is the total amount of time spent in speed bucket m by driver i. Eq. (2.1) induces an empirical discrete distribution $\boldsymbol{z}_{i,m} = (z_{i,m,1}, \dots, z_{i,m,24})'$ on speed bucket m, which lies in the $(24-1)$-unit simplex $\mathcal{Z} \subset \mathbb{R}_{+}^{24}$, i.e., has normalization $\sum_{j=1}^{24} z_{i,m,j} = 1$. The *driving style* of every car driver i is described by a J-vector $\boldsymbol{x}_i = (\boldsymbol{z}'_{i,1}, \dots, \boldsymbol{z}'_{i,4})' \in \mathbb{R}^J$ containing the four discrete distributions $\boldsymbol{z}_{i,m}$ on the rectangle $m = 1, \dots, 4$. This can be illustrated by four v-a heatmaps. Note that the dimension of $\boldsymbol{x}_i$ is $J = 24 \times 4 = 96$. Also note that we have the following relationship between elements in $\boldsymbol{z}_{i,m}$ and elements in $\boldsymbol{x}_i$:

$$z_{i,m,j} = x_{i,(m-1)\times 24+j}, \tag{2.2}$$

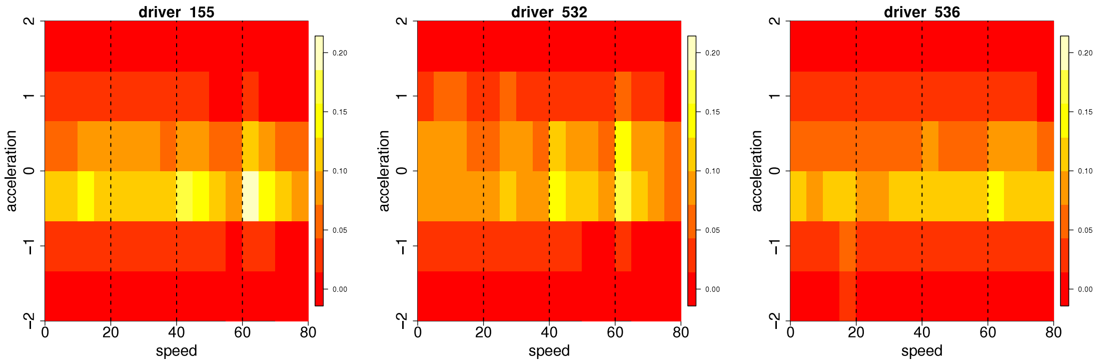

Fig. 2. v-a heatmaps of drivers 155, 532 and 536.

Table 1
The 90%, 95% and 99% quantiles of the minimal amount in the four speed buckets.

Speed bucket	Minimal amount					
	In days			In minutes		
	90%	95%	99%	90%	95%	99%
(0, 20] km/h	8	11	19	**163**	220	460
(20, 40] km/h	10	13	21	**194**	252	455
(40, 60] km/h	22	29	44	**292**	372	615
(60, 80] km/h	40	51	86	**366**	461	702

for $i = 1, \dots, 973$, $m = 1, \dots, 4$, $j = 1, \dots, 24$. We draw the four v-a heatmaps jointly for drivers 155, 532 and 536 in Fig. 2. It shows that the width of the level sets on the a-axis of driver 155 is more narrow than the ones of the other two drivers. This indicates a smoother acceleration and braking pattern of driver 155.

Driving habit of driver i is defined to be the relative amount of time spent in each speed bucket m:

$$h_{i,m} = \frac{t_{i,m}}{t_i}, \quad \text{for } m = 1, \dots, 4, \tag{2.3}$$

where $t_i = \sum_{m=1}^{4} t_{i,m}$ is the total amount of time spent in the entire speed interval (0, 80] km/h by driver i. Eq. (2.3) induces an empirical discrete distribution $\boldsymbol{h}_i = (h_{i,1}, \dots, h_{i,4})'$ on R, which lies in the $(4-1)$-unit simplex $\mathcal{H} \subset \mathbb{R}_+^4$, and has normalization $\sum_{m=1}^{4} h_{i,m} = 1$. Suppose that a commuting driver i and an off-peak driver i' had the same driving style, we would have $h_{i,1} > h_{i',1}, h_{i,4} < h_{i',4}$, but $\boldsymbol{x}_i = \boldsymbol{x}_{i'}$. Another driving habit covariate is the average driving hours in (0, 80] km/h per week, defined as

$$ave\_hours_i = \frac{t_i \times 7}{3600 \times 92},$$

which indicates the intensity of driving. In this paper, we consider the effects of both driving style and driving habit on claims frequencies.

The volume of telematics data is increasing with the length of the observation period, and the above procedure compresses this increasing telematics data to a J-vector $\boldsymbol{x}_i$, a four-vector $\boldsymbol{h}_i$ and a scalar ave_hours_i no matter how long the observation period is. We may directly use $\boldsymbol{h}_i$ and ave_hours_i in claims frequency models. But we should not directly use $\boldsymbol{x}_i$ in claims frequency models because the dimension of $\boldsymbol{x}_i$ is very large and there may be collinearity in $\boldsymbol{x}_i$ (i.e., most heatmaps have the largest values around the zero acceleration). Therefore, we analyze and pre-process $\boldsymbol{x}_i$.

For each speed bucket m, we stack the vectors $\boldsymbol{z}_{i,m}, i = 1, \dots, n$, to form the $n \times 24$ design matrix $\boldsymbol{X}_m \in \mathbb{R}^{n \times 24}$. For the four speed buckets altogether, we stack the vectors $\boldsymbol{x}_i, i = 1, \dots, n$, to form the $n \times J$ design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times J}$. Note that $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4)$. Denote the normalized design matrices by $(\boldsymbol{X}_m^0)_{m=1:4}$ and $\boldsymbol{X}^0$ (all column means are set to zero and variances are normalized to one). Denote the corresponding ith row by $(\boldsymbol{z}_{i,m}^0)_{m=1:4}$ and $\boldsymbol{x}_i^0$. In Sections 2.2 and 2.3 we aim at applying the K-medoids clustering and the principal components analysis to reduce the dimension of the normalized design matrices and extract the risk factors, still capturing explanatory power for claims frequency prediction.

2.1.2. Stability of v-a heatmaps

We need to determine the minimum data volumes required for stable v-a heatmaps. We fix a driver i and a speed bucket m, and split his/her telematics car driving data into different driving days. Denote by $\boldsymbol{z}^d$ the v-a heatmap derived from the telematics data of d driving days. When $d \to \infty$, denote the limiting heatmap by $\boldsymbol{z}^\infty$. The finite sample error between $\boldsymbol{z}^\infty$ and $\boldsymbol{z}^d$ is measured by the Kullback–Leibler (KL) divergence:

$$d_{\mathrm{KL}}(\boldsymbol{z}^\infty \parallel \boldsymbol{z}^d) = -\sum_{j=1}^{J} z_j^\infty \log \frac{z_j^d}{z_j^\infty}. \tag{2.4}$$

Appendix A of Gao et al. (2019) has derived an approximation of (2.4). Similar to Gao et al. (2019), we require the finite sample error (2.4) to be less than 0.05, which leads to the minimal days $d_{i,m}^*$ for each driver i and each speed bucket m. We multiply the minimal days with the average driving minutes in speed bucket m per day to get the minimal telematics data amount in minutes. Table 1 shows the 90%, 95% and 99% quantiles of the minimal amount in days and in minutes for a stable v-a heatmap in a specific speed bucket. Note that we need more telematics data in higher speed bucket to get a stable v-a heatmap in that speed bucket. If we require the driving time in the four speed buckets to be at least 160, 190, 290 and 360 min, respectively and simultaneously, we receive $n = 973$ cars in our database meeting this requirement.

2.2. K-medoids clustering

A clustering analysis aims at grouping the drivers with similar heatmaps into the same cluster. By applying the K-medoids clustering to $\boldsymbol{X}^0$, it effectively reduces the dimension J of $\boldsymbol{X}^0$ to the chosen number of clusters K. Compared to the K-means clustering, the K-medoids clustering is more robust against outliers and can be applied to any distance function such as Euclidean distance, Manhattan distance, Canberra distance, etc., though at the expense of more computational time (Hastie et al., 2009). Here, we consider the Euclidean distance.

Denote by $\mathcal{N} = \{1, \dots, n\}$ the driver labels. Denote by $\mathcal{K} = \{1, \dots, K\}$ the K cluster labels. Denote by $C = \{c_{K,1}, \dots, c_{K,K}\} \subset \mathcal{N}$ the increasing ordered K medoid driver labels (i.e., $c_{K,1} <$

$c_{K,2} < \cdots < c_{K,K}$). A classification structure is introduced by partitioning the set $\mathcal{N}$ into K disjoint clusters $\mathcal{N}_1, \ldots, \mathcal{N}_K$ satisfying

$$\bigcup_{k=1}^{K} \mathcal{N}_k = \mathcal{N} \quad \text{and} \quad \mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset \text{ for all } k \neq k'. \tag{2.5}$$

These K clusters define a classifier $\mathcal{C}$ on the set $\mathcal{N}$, given by

$$\mathcal{C} : \mathcal{N} \to \mathcal{K}, \qquad i \mapsto \mathcal{C}(i) = \sum_{k=1}^{K} k \mathbb{1}_{\{i \in \mathcal{N}_k\}}. \tag{2.6}$$

The within-cluster distance of the kth cluster is defined as

$$W_k(\mathcal{C}) = \sum_{i \in \mathcal{N}_k} d(\boldsymbol{x}_i^0, \boldsymbol{x}_{c_{K,k}}^0), \tag{2.7}$$

where we use the Euclidean distance $d(\boldsymbol{x}_i^0, \boldsymbol{x}_{i'}^0) = \|\boldsymbol{x}_i^0 - \boldsymbol{x}_{i'}^0\|_2$ for $\boldsymbol{x}_i^0, \boldsymbol{x}_{i'}^0 \in \mathbb{R}^J$. The medoid driver $c_{K,k}$ of cluster k satisfies

$$c_{K,k} = \arg\min_{j \in \mathcal{N}_k} \sum_{i \in \mathcal{N}_k} d(\boldsymbol{x}_i^0, \boldsymbol{x}_j^0).$$

Our goal is to find a classifier $\mathcal{C}$ that minimizes the total within-cluster distance given by

$$W(\mathcal{C}) = \sum_{k=1}^{K} W_k(\mathcal{C}). \tag{2.8}$$

The partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990) provides a K-medoid clustering. The PAM algorithm contains the following steps:

1. Randomly select K drivers $C^{(0)} = \{c_{K,1}^{(0)}, \ldots, c_{K,K}^{(0)}\}$ as medoids. Allocate each driver $i \in \mathcal{N}$ to its nearest medoid. This defines an initial classifier $\mathcal{C}^{(0)}$:

   $$i \mapsto \mathcal{C}^{(0)}(i) = \arg\min_{k \in \mathcal{K}} d(\boldsymbol{x}_i^0, \boldsymbol{x}_{c_{K,k}^{(0)}}^0).$$

2. Calculate the total within-cluster distance $W(\mathcal{C}^{(0)})$ according to (2.7) and (2.8).
3. Repeat the following steps for $l \geq 1$:

   (a) For each pair of a medoid in $C^{(l-1)}$ and a non-medoid driver in $\mathcal{N} \setminus C^{(l-1)}$, swap the medoid with the non-medoid driver, allocate each driver to its nearest medoid, and calculate the total within-cluster distance.

   (b) Choose the swap which leads to the smallest total within-cluster distance. This defines a proposed set of medoids $C^{(*)}$ and the corresponding classifier $\mathcal{C}^{(*)}$.

   (c) If $W(\mathcal{C}^{(*)}) - W(\mathcal{C}^{(l-1)}) < 0$, accept the proposed clustering and let $C^{(l)} = C^{(*)}, \mathcal{C}^{(l)} = \mathcal{C}^{(*)}, W(\mathcal{C}^{(l)}) = W(\mathcal{C}^{(*)})$. If $W(\mathcal{C}^{(*)}) - W(\mathcal{C}^{(l-1)}) \geq 0$, terminate the algorithm and accept $\mathcal{C}^{(l-1)}$ as the final clustering.

Note that the PAM algorithm finds a local minimum in the sense that no single switch of a medoid and a non-medoid will decrease the total within-cluster distance (2.8). Different initial sets of medoids $C^{(0)}$ may lead to different clustering. Reynolds et al. (1992) propose a build phase which looks for a good initial set of medoids following certain rules. The R function `pam` implements this build phase by default, and always returns the same clustering for a particular data set. Note that there may be ties in medoids, i.e., one of two identical points is selected as a medoid. For our data set containing $n = 973$ car drivers, it takes less than one second to get the result of 2-medoids clustering. For a larger portfolio containing millions of car drivers, K-means clustering might be a better choice.

The cluster covariate $\mathcal{C}(i) \in \mathcal{K}$ of each driver i could directly be used in claims frequency models as a categorical covariate with K levels. However, two drivers in the same cluster cannot be distinguished by their clusters. Here, we use the continuous covariate of distance between a driver and each medoid driver. The distance covariate can distinguish two drivers in the same cluster, i.e., the distance covariate provides more information than the cluster covariate.

For 2-medoids clustering, we have medoid drivers $c_{2,1} = 155, c_{2,2} = 820$. For 3-medoids clustering, we have $c_{3,1} = 155, c_{3,2} = 532, c_{3,3} = 536$ (see Fig. 2). For 4-medoids clustering, we have $c_{4,1} = 155, c_{4,2} = 165, c_{4,3} = 532, c_{4,4} = 536$. Hence, we find 5 different medoids $c_{2,1}, c_{2,2}, c_{4,2}, c_{4,3}, c_{4,4}$ from the three clusterings. In claims frequency modeling, we will use the distances between driver i and each of the 5 medoids, defined as

$$d_{i,k|K} = d(\boldsymbol{x}_i^0, \boldsymbol{x}_{c_{K,k}}^0) = \|\boldsymbol{x}_i^0 - \boldsymbol{x}_{c_{K,k}}^0\|_2. \tag{2.9}$$

Note that it is difficult to find the optimal value of medoids K leading to good fits in the claims frequency models because the K-medoids clustering is an unsupervised learning algorithm. We only consider $K = 2, 3, 4$ because we will also show that the principal components of $\boldsymbol{X}^0$ derived next always have a better predictive power than the K-medoids covariates (2.9).

2.3. Principal components analysis

We apply principal components analysis to directly analyze the normalized design matrix $\boldsymbol{X}^0 \in \mathbb{R}^{n \times J}$. Principal components analysis reduces the dimension of $\boldsymbol{X}^0$, if we only use the first few principal components in the claims frequency models. Principal components analysis can also deal with collinearity in $\boldsymbol{X}^0$.

Denote the J covariates in $\boldsymbol{X}^0$ by $(X_j^0)_{j=1:J}$, i.e., these are the columns of $\boldsymbol{X}^0$. Consider a direction $\boldsymbol{v}_1 = (v_{1,1}, \ldots, v_{J,1})'$ of the J-dimensional covariate space with normalization constraint

$$\sum_{j=1}^{J} v_{j,1}^2 = 1. \tag{2.10}$$

The first principal component of the covariates $(X_j^0)_{j=1:J}$ is their projected value onto the direction $\boldsymbol{v}_1$

$$P_1 = v_{1,1} X_1^0 + \cdots + v_{J,1} X_J^0,$$

which has the largest variance. The elements $v_{1,1}, \ldots, v_{J,1}$ are the loadings of the first principal components, and the vector $\boldsymbol{v}_1$ is the first principal component loading vector also called the first right-singular vector of the corresponding matrix $\boldsymbol{V}$ (which we are going to introduce below). The second principal component P_2 is the projected value of $(X_j^0)_{j=1:J}$ onto the direction $\boldsymbol{v}_2$ perpendicular to $\boldsymbol{v}_1$, which has the second largest variance; and so on.

Given the data $\boldsymbol{X}^0$ the first principal component loading vector can be derived by maximizing the sample variance of P_1:

$$\arg \max_{v_{1,1}, \ldots, v_{J,1}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{J} v_{j,1} x_{i,j}^0 \right)^2 \right], \tag{2.11}$$

under constraint (2.10). The above optimization problem can be solved via singular value decomposition as follows:

$$\boldsymbol{X}^0 = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{V}',$$

where $\boldsymbol{U}$ is an $n \times J$ orthogonal matrix, $\boldsymbol{V}$ is a $J \times J$ orthogonal matrix and $\boldsymbol{\Lambda} = \text{diag}(g_1, \ldots, g_J)$ is a $J \times J$ diagonal matrix with singular values $g_1 \geq \cdots \geq g_J \geq 0$. The wth column of the rotation

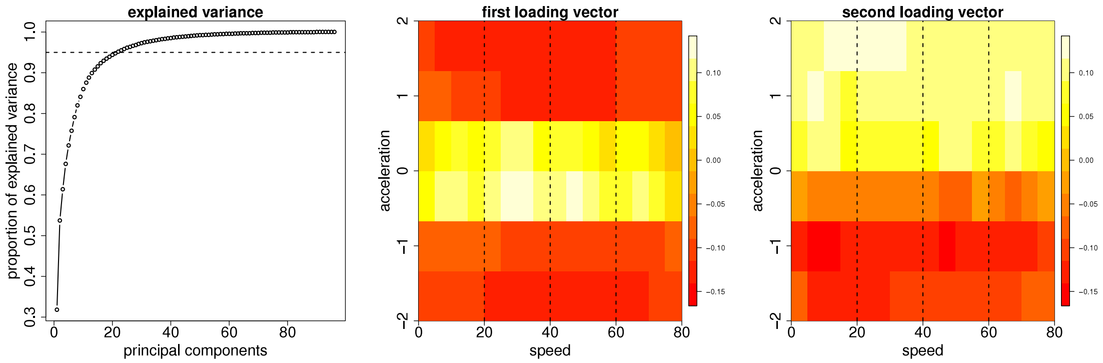

Fig. 3. The proportion of explained variance by the principal components (left). The first and second loading vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ (middle and right).

matrix $\boldsymbol{V}$ is the wth principal component loading vector (or right-singular vector) $\boldsymbol{v}_w = (v_{1,w}, \ldots, v_{J,w})'$, $w = 1, \ldots, J$. The wth principal component is the projected value of $(X_j^0)_{j=1:J}$ onto the direction $\boldsymbol{v}_w$, that is,

$$P_w = \sum_{j=1}^{J} v_{j,w} X_j^0. \tag{2.12}$$

The wth principal component of driver i is

$$p_{i,w} = \sum_{j=1}^{J} v_{j,w} x_{i,j}^0.$$

We illustrate the proportion of explained variance in $\boldsymbol{X}^0$ by the principal components in Fig. 3 (left). The first 20 principal components explain around 95% of the total variance in $\boldsymbol{X}^0$. Therefore, we only consider the first 20 principal components in claims frequency modeling. In Fig. 3 we show the first and second loading vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ in its corresponding sub-rectangle. The first principal component loadings describe abrupt braking in $[-2, -2/3)$ m/s^2 and strong acceleration in $(2/3, 2]$ m/s^2. Thus, the first principal component reflects the relative frequency of smooth acceleration/braking, i.e., the degree of concentration on the zero acceleration rate, see Fig. 3 (middle). The signs of the second principal component loadings switch around the acceleration rate zero, which indicates that the second principal component illustrates the difference in absolute value between acceleration and braking, see Fig. 3 (right).

Finally, we apply the principal component analysis to the matrices $(\boldsymbol{X}_m^0)_{m=1:4}$, respectively, to derive the principal components in each speed bucket. We denote by $p_{i,w}^m$, $w = 1, \ldots, 24$, $m = 1, \ldots, 4$, the wth principal component of driver i in speed bucket m. It shows that the first 7 principal components in each speed bucket m account for around 95% variance in $\boldsymbol{X}_m^0$, so we only consider the first 7 principal components of each speed bucket in claims frequency modeling. In Table 2, we calculate the coefficient of correlation among the first two principal components $p_{i,1}^m, p_{i,2}^m$. It shows that the driving characteristics in different speed buckets are quite similar in terms of the first two principal components.

3. Claims frequency modeling

We consider the compulsory third party policies purchased by these $n = 973$ cars (these policies have all the same coverage limit of CNY 122,000). We record the number of reported claims from 01/01/2014 to 29/06/2017 with effective exposures supported in the time interval from 01/01/2014 to 31/05/2017.

Table 2
The coefficients of correlation among the first two principal components $p_{i,1}^m, p_{i,2}^m$.

	$p_{i,1}^1$	$p_{i,1}^2$	$p_{i,1}^3$	$p_{i,1}^4$	$p_{i,2}^1$	$p_{i,2}^2$	$p_{i,2}^3$	$p_{i,2}^4$
$p_{i,1}^1$	1.00	0.86	0.69	0.55	0	-1.2×10^{-2}	1.1×10^{-2}	3.0×10^{-2}
$p_{i,1}^2$	0.86	1.00	0.87	0.70	-2.2×10^{-2}	0	1.5×10^{-2}	3.9×10^{-2}
$p_{i,1}^3$	0.69	0.87	1.00	0.92	-8.3×10^{-2}	-4.6×10^{-2}	0	2.3×10^{-2}
$p_{i,1}^4$	0.55	0.70	0.92	1.00	-1.3×10^{-1}	-8.9×10^{-2}	-2.4×10^{-2}	0
$p_{i,2}^1$	⋯	⋯	⋯	⋯	1.00	0.95	0.91	0.86
$p_{i,2}^2$	⋯	⋯	⋯	⋯	0.95	1.00	0.96	0.89
$p_{i,2}^3$	⋯	⋯	⋯	⋯	0.91	0.96	1.00	0.93
$p_{i,4}^4$	⋯	⋯	⋯	⋯	0.86	0.89	0.93	1.00

The evaluation date is chosen as 31/05/2017 since a preliminary analysis has shown that more than 99% of all claims are reported with less than one month of reporting delay. For this reason, we do not expect a material influence on the claims frequency of claims with a reporting delay of more than one month.

Based on this data, we investigate three aspects in this section: (a) the predictive performance of driving habit covariates $(h_{i,m})_{m=1:4}$ and ave_hours_i; (b) the predictive performance of driving style covariates $(d_{i,k|K=2})_{k=1:2}$, $(d_{i,k|K=4})_{k=2:4}$ and $(p_{i,w})_{w=1:20}$; (c) the predictive performance of the covariates $(p_{i,w}^m)_{w=1:7}$ in each speed bucket m.

3.1. Description of variables

In the following, we describe the response variable of claims counts Y_i, the exposure of effective policy duration e_i (also called years-at-risk), the classical risk factors, the driving habit covariates and the driving style covariates. Note that we assume that the main driver of a car does not change and we aggregate all the policies of each car. For a car with several policies, we choose the median of the classical risk factors in these policies.

We show the aggregated years-at-risk for different claims counts Y_i in Fig. 4 (top-left); most policies do not suffer a claim. We show the number of cars for different exposures $e_i \in [1, 3.5]$ years-at-risk in Fig. 4 (top-middle); most cars have more than one year-at-risk. The average exposure is 2.24 years-at-risk. The average claims frequency is $\sum_1^n Y_i / \sum_1^n e_i = 0.24$ per year per car driver; this is consistent with the market benchmark in China but much higher than typically in Europe.

3.1.1. Description of classical risk factors

We consider four classical risk factors: regions ($region_i$), driver's gender ($gender_i$), driver's age ($driver\_age_i$), and car's age

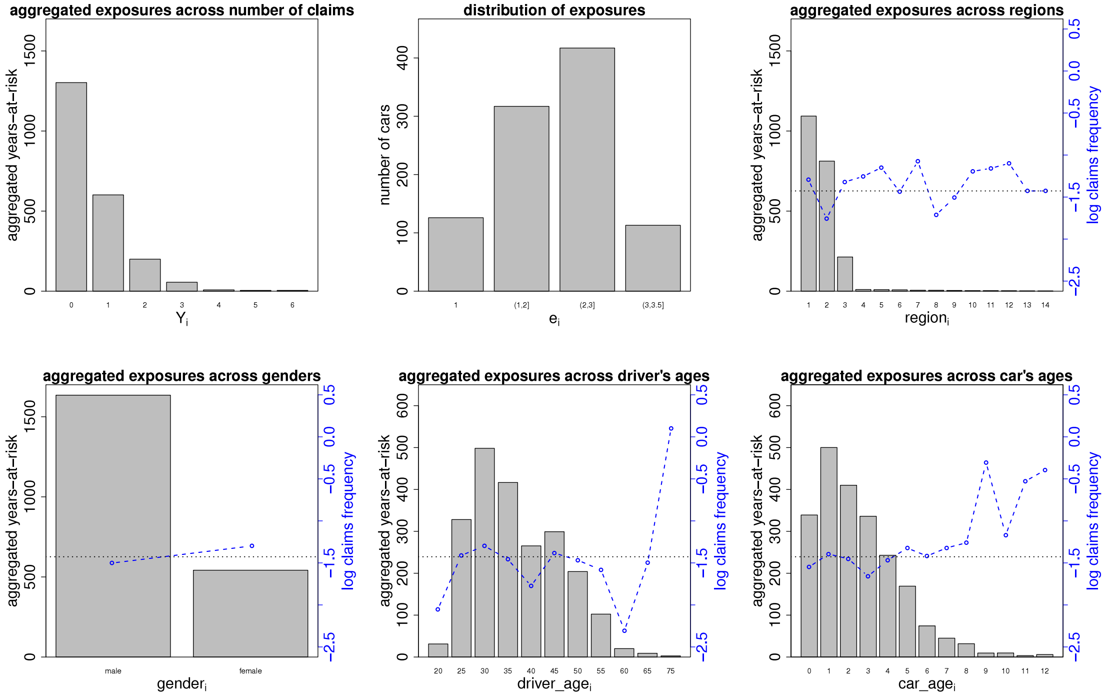

Fig. 4. Description of the claims counts (top-left) and the exposures (top-middle). Distribution of aggregated years-at-risk (left axis) and the corresponding logarithm of the empirical claims frequencies (right axis) across the four classical risk factors: regions (top-right), gender (bottom-left), driver's age (bottom-middle), and car's age (bottom-right).

(car_age_i) for each driver $i = 1, \dots, n$. We show the distribution of the aggregated years-at-risk (left axis) and the corresponding logarithm of the empirical claims frequencies (right axis) across the four variables in Fig. 4, respectively. We take the logarithm of the empirical claims frequencies because the empirical claims frequency for a small exposure might be quite large. Another reason is that we will use the log-link function for claims frequency modeling. If the claims count is zero for a covariate value/interval, we draw the portfolio average of log(0.24), see the black dotted lines in Fig. 4. We have the following observations:

- Regions: Regions 1, 2, 3 account for 97% of the total exposure. Regions 4–14 account for 15 years-at-risk in total. We cannot detect any significant difference in claims frequency among regions 4–14 based on such few exposures. Hence, we aggregate regions 4–14 as region 4. Note that one may merge each of the regions 4–14 with the nearest one of regions 1, 2, 3. Another option is to replace the region covariate by a continuous variable of the population density in that region. We refrain to do so since there are too few exposures in regions 4–14. We observe that the claims frequency in region 2 is lower than in regions 1 and 3, while the claims frequencies in regions 1 and 3 are similar.
- Gender: There are many more male drivers than female drivers, and female drivers tend to have a higher claims frequency.
- Driver's age: We bin the driver's ages into 11 groups of 5 years, i.e., $\{18 : 22\}, \{23 : 27\}, \dots, \{68 : 72\}, \{73 : 77\}$. This leads to a comparable y-scale with the other plots. Surprisingly, drivers at the ages of $\{18 : 22\}$ and $\{58 : 62\}$ have the lowest claims frequency, which is explained by small exposures at these ages and possibly extraordinary observations. Note that we will treat driver's age as a continuous variable in claims frequency modeling.
- Car's age: For cars used for less than 3 years, the claims frequencies are close. For cars used for more than 3 years, the claims frequency is increasing with the car's age. Note that the observed increasing pattern is based on relatively small exposure for older cars. Whether this observation is statistically significant is questionable.

3.1.2. Description of driving habit covariates and selected driving style covariates

We describe 5 driving habit covariates $(h_{i,m})_{m=1:4}$, ave_hours_i and 4 selected driving style covariates $d_{i,1|K=2}, d_{i,2|K=2}, p_{i,1}, p_{i,2}$. We show the distribution of the aggregated years-at-risk (left axis) and the corresponding logarithm of the empirical claims frequencies (right axis) across the 9 variables in Fig. 5, respectively. Note that all these covariates are continuous variables and we bin them appropriately to get a comparable y-scale as we did for driver's age. If the claims count is zero for a covariate interval, we draw the portfolio average of $\log 0.24$, see the black dotted lines in Fig. 5. We have the following observations:

- The relative time spent in speed bucket m: Most drivers spend less time in (40, 80] km/h than in (0, 40] km/h. There seems a slight decreasing pattern of claims frequency with the covariate $h_{i,3}$. We cannot observe any obvious patterns for the other 3 covariates.
- Average driving hours per week in (0, 80] km/h: Most drivers spend around 6 h in (0, 80] km/h per week. The logarithm of the claims frequencies increase with the intensity of driving, though there is higher volatility at larger covariate values due to the small exposures.
- The 2-medoids covariates: We cannot observe any obvious patterns in claims frequencies with the two covariates. And again there is higher volatility at larger covariate values.
- The first two principal components: The claims frequency decreases with the first principal component and fluctuates randomly with the second principal components. Note that the first principal component reflects the frequency of smooth acceleration/braking.

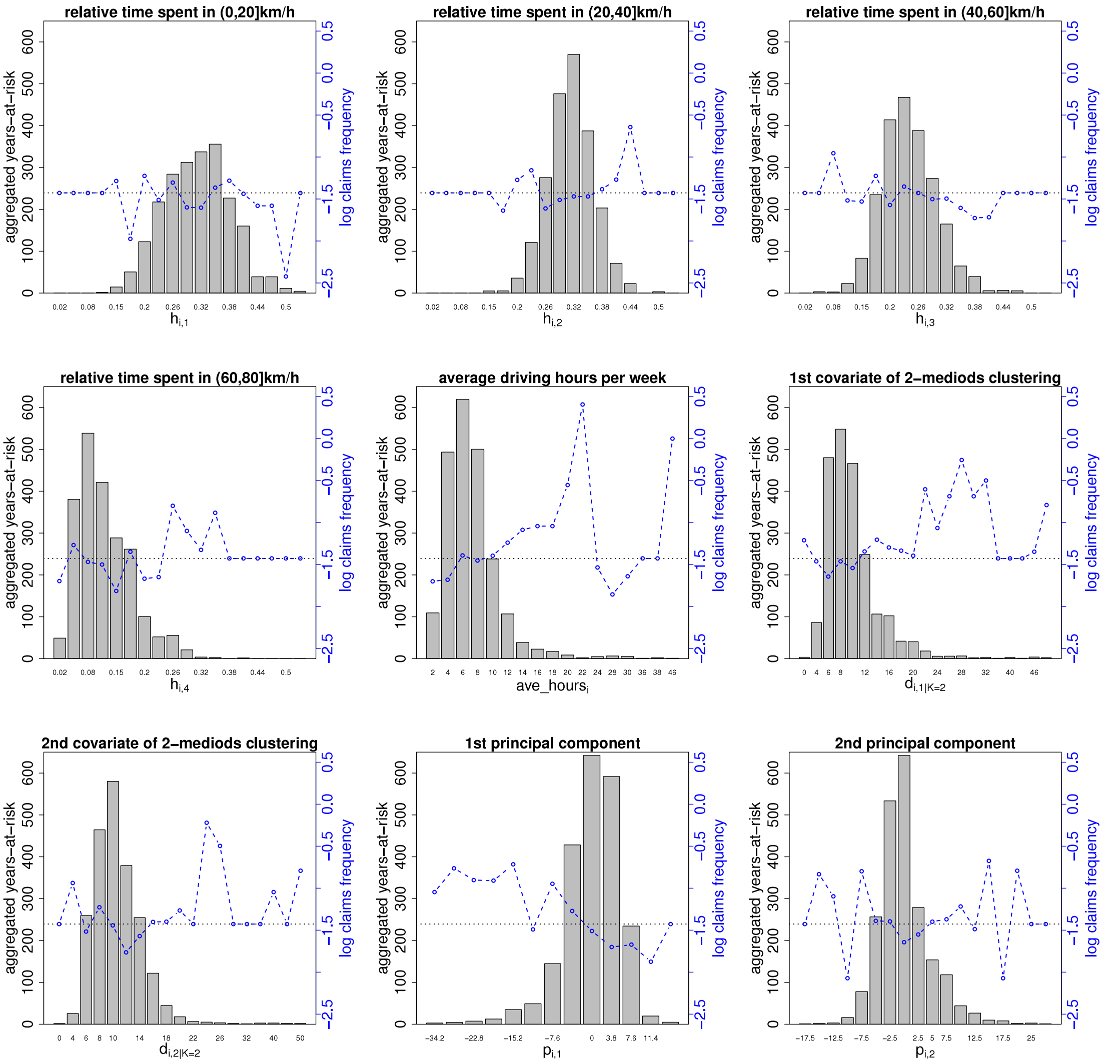

Fig. 5. Distribution of aggregated years-at-risk and the corresponding logarithm of the empirical claims frequencies across the driving habit covariates and the selected driving style covariates.

Note that the blue dashed lines in Figs. 4 and 5 show the marginal effect of each covariate on the claims frequency without considering interactions with other covariates. In claims frequency modeling, we will consider several covariates simultaneously, and they may have a different interacting effect on claims frequencies from the one shown in Figs. 4 and 5.

3.2. Poisson generalized additive models for claims frequency

In this section, we establish several Poisson generalized additive models for claims counts using different sets of covariates. In general, we assume that the number of claims Y_i of driver i follows a Poisson distribution with an underlying expected claims frequency of λ_i per year. The basic assumption is that this expected frequency λ_i has a multiplicative structure in the covariates. For example, the model would be

$$Y_i \overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_i e_i) \qquad \text{with}$$
$$\log \lambda_i = \beta_0 + \alpha_{u_i} + \beta_1 v_i + s(w_i; \boldsymbol{\beta}_2, \delta), \tag{3.1}$$

where $e_i \in [1, 3.5]$ years-at-risk is the total exposure of driver i measured by the effective policy duration to the evaluation date 31/05/2017, u_i is a categorical covariate, v_i is a continuous covariate having a linear effect (on the log-scale) and w_i is a continuous covariate having a non-linear effect (on the log-scale). The non-linear effect of w_i is described by a penalized thin plate regression spline s with regression parameters $\boldsymbol{\beta}_2$ and smoothing parameter δ. By using the penalized thin plate regression splines, we do not need to specify the knots (Section 4.1.5 of Wood, 2017).

UBRE. When the smoothing parameter δ is known, the regression parameters $\beta_0, \alpha_{u_i}, \beta_1, \boldsymbol{\beta}_2$ in (3.1) can be estimated by penalized iterative re-weighted least squares which extends the classical least squares method to incorporate a thin plate spline penalty for wiggliness and to incorporate an extra weight accounting for the variance of Y_i proportional to its mean. The optimal value of the smoothing parameter δ is determined by minimizing the expected deviance loss which leads to the unbiased risk estimator (UBRE) (Section 4.5.4 of Wood, 2017). The UBRE is known to have some tendency to overfitting on occasion. It has been suggested that an ad hoc way is to force each effective degree of freedom to count as 1.4 degrees of freedom in the UBRE, which leads to a larger smoothing parameter (Section 5.1.1 of Wood, 2017). Note that we will change a smooth term with less than 1.1 effective degrees of freedom to a linear term (on the log-scale). The UBRE plays a similar role as AIC and we prefer models with smaller UBRE.

Backward elimination. In the following claims frequency models, we always start with a full model containing all the considered covariates. Then we apply the backward elimination to select the covariates. That is, we sequentially drop the single covariate with the highest non-significant p-value from the model and refit the model until all the covariates are significant. Note that we need to remove one covariate at a time because two correlated covariates may be significant individually but both are non-significant when considered together. The test statistics for a non-smooth term follows approximately a normal distribution under the null hypothesis that the corresponding parameter equals zero. The test statistics for a smooth term follows approximately a $\mathcal{X}^2$-distribution under the null hypothesis that the smooth term equals a zero function.

Cross validation estimate of average Poisson deviance loss. Besides the UBRE and AIC, we also evaluate a model by estimating the average Poisson deviance loss using cross validation. We randomly partition the data of all cars $\mathcal{N}$ into 10 roughly equally-sized disjoint parts, denoted by $\mathcal{T}_1, \dots, \mathcal{T}_{10}$. We estimate the average Poisson deviance loss by 10-fold cross validation as

$$\widehat{D} = \frac{1}{10}\sum_{l=1}^{10} D(\mathcal{T}_l, \hat{\theta}_{-\mathcal{T}_l}), \tag{3.2}$$

where $D(\mathcal{T}_l, \hat{\theta}_{-\mathcal{T}_l})$ is the average Poisson deviance loss on the data $\mathcal{T}_l$ using the estimated claims frequencies $\lambda_i(\hat{\theta}_{-\mathcal{T}_l})$

$$D(\mathcal{T}_l, \hat{\theta}_{-\mathcal{T}_l}) = \frac{2}{|\mathcal{T}_l|}\sum_{i\in\mathcal{T}_l} Y_i\left[\frac{\lambda_i(\hat{\theta}_{-\mathcal{T}_l})e_i}{Y_i} - 1 - \log\left(\frac{\lambda_i(\hat{\theta}_{-\mathcal{T}_l})e_i}{Y_i}\right)\right]. \tag{3.3}$$

Note that the claims frequency $\lambda_i(\hat{\theta}_{-\mathcal{T}_l})$ depends on the regression parameter $\hat{\theta}_{-\mathcal{T}_l}$ which is estimated using all drivers except the ones in $\mathcal{T}_l$. Also note that the ith term on the right-hand side of (3.3) is set equal to $2\lambda_i(\hat{\theta}_{-\mathcal{T}_l})e_i$ if $Y_i = 0$. We prefer the Poisson deviance as our loss function because it is the natural choice for claims frequency modeling under the Poisson assumption.

We randomly partition all drivers $\mathcal{N}$ for 50 times, and calculate the cross validation estimate of the average Poisson deviance loss $\hat{D}_s$ for each partition $s = 1, \dots, 50$. We calculate the sample mean, the sample standard deviation, and the 5% and 95% quantiles of $(\hat{D}_s)_{s=1:50}$. Note that the cross validation estimate of the average Poisson deviance loss is sensitive to overfitting and we prefer models with smaller average Poisson deviance losses.

3.2.1. *Claims frequency modeling with the classical risk factors*

We start by assuming the underlying expected claims frequency λ_i be a multiplicative function of the classical covariates. That is, we start with the model

$$\begin{aligned}\log\lambda_i = \beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i; \boldsymbol{\beta}_1, \delta_1)\\ + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2),\end{aligned} \tag{3.4}$$

where s_1, s_2 are penalized thin plate regression splines with regression parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and smoothing parameters δ_1, δ_2 addressing potential non-linear effects of driver's age and car's age, respectively.

We show the UBRE, the AIC, the average Poisson deviance loss, the estimated coefficients (or effective degrees of freedom), and the corresponding significance tests in Table 3. Note that male drivers in region 1 are treated as reference class because it contains the largest exposures. At the 5% significance level, drivers in region 2 have a lower claims frequency than those in region 1, while drivers in regions 3 and 4 have the same claims frequency as those in region 1. At the 5% significance level, female drivers have the same claims frequency as male drivers. The optimal values of smoothing parameters δ_1 and δ_2 are determined as 18.65 and 9.40. The effective degrees of freedom of s_1 and s_2 are estimated as 1.44 and 1.81. At the 5% significance level, we cannot reject the null hypothesis that drivers at different ages have the same claims frequency. In addition, we estimate the dispersion parameter as 1.16 by assuming an over-dispersed Poisson structure, so there is no obvious evidence of over-dispersion.

Table 3
Model (3.4) with the UBRE of 0.0781, the AIC of 1851 and the average Poisson deviance loss of 1.0792 with standard error of 0.0034 and 90% interval of (1.0750, 1.0840).

Parameters	Estimate	Standard error	Test statistics	p-value
$\hat{\beta}_0$	−1.36	0.07	−20.31	0.00
$\hat{\alpha}_2$	−0.44	0.10	−4.25	0.00
$\hat{\alpha}_3$	0.01	0.15	0.07	0.95
$\hat{\alpha}_4$	−0.09	0.27	−0.35	0.73
$\hat{\gamma}_{female}$	0.17	0.10	1.70	0.09
$\hat{\boldsymbol{\beta}}_1$	1.44 (edf)	–	0.99	0.43
$\hat{\boldsymbol{\beta}}_2$	1.81 (edf)	–	6.98	0.04

Table 4
Model (3.5) with the UBRE of 0.0764, the AIC of 1851 and average Poisson deviance loss of 1.0769 with standard error of 0.0032 and 90% interval of (1.0724, 1.0817).

Parameters	Estimate	Standard error	Test statistics	p-value
$\hat{\beta}_0$	−1.30	0.06	−22.37	0.00
$\hat{\alpha}_2$	−0.46	0.10	−4.44	0.00
$\hat{\alpha}_3$	−0.03	0.14	−0.19	0.85
$\hat{\alpha}_4$	−0.05	0.26	−0.20	0.84
$\hat{\boldsymbol{\beta}}_2$	1.90 (edf)	–	6.62	0.05

We apply the backward elimination to model (3.4) to remove driver's age and gender sequentially. The resulting model is

$$\log\lambda_i = \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2). \tag{3.5}$$

We show the results of model (3.5) in Table 4. Models (3.4) and (3.5) have a similar predictive performance, though the second one is slightly better according to the UBRE and the average Poisson deviance loss. We remark that driver's age is usually an important risk factor in car insurance. However, since our portfolio is small and we do not have enough exposures for young and old ages, the model cannot detect the effect of driver's age at young and old ages on the claims frequency well.

Finally, we fit an intercept model:

$$\log\lambda_i = \beta_0. \tag{3.6}$$

The intercept model has the UBRE of 0.0940, the AIC of 1872 and the average Poisson deviance of 1.0938 with standard error of 0.0014 and 90% interval of (1.09241.0966). By including the region and car's age into the intercept model, we decrease the UBRE by 0.0176, the AIC by 21, and the average Poisson deviance by 0.0169. Therefore, we prefer the latter model.

In the following we add more covariates to model (3.4) to improve the predictive performance. Note that we include gender and driver's age in the model because the two covariates may become significant when we add other covariates.

3.2.2. *Claims frequency modeling with driving habit covariates*

Smooth terms of driving habit covariates. We add five smooth terms for the relative time spent in each speed bucket $(h_{i,m})_{m=1:4}$ and the average driving hours per week ave_hours_i to model

Table 5
The p-values during the backward elimination of model (3.7).

Step	$region_i$	$gender_i$	$driver\_age_i$	car_age_i	$h_{i,1}$	$h_{i,2}$	$h_{i,3}$	$h_{i,4}$	ave_hours_i
1	0.000	0.073	0.298	0.030	0.997	0.667	0.098	0.188	0.008
2	0.000	0.073	0.298	0.030	–	0.667	0.098	0.188	0.008
3	0.000	0.080	0.298	0.028	–	–	0.063	0.143	0.006
4	0.000	0.069	–	0.039	–	–	0.071	0.124	0.011
5	0.000	0.094	–	0.040	–	–	0.095	–	0.013
6	0.000	0.091	–	0.045	–	–	–	–	0.009
7	0.000	–	–	0.043	–	–	–	–	0.008

Table 6
Model (3.8) with the UBRE of 0.0702, the AIC of 1844 and average Poisson deviance loss of 1.0697 with standard error of 0.0036 and 90% interval of (1.06431.07645).

Parameters	Estimate	Standard error	Test statistics	p-value
$\hat{\beta}_0$	−1.31	0.06	−22.37	0.00
$\hat{\alpha}_2$	−0.46	0.10	−4.45	0.00
$\hat{\alpha}_3$	−0.04	0.14	−0.29	0.77
$\hat{\alpha}_4$	−0.07	0.26	−0.28	0.78
$\hat{\boldsymbol{\beta}}_2$	1.91 (edf)	–	6.93	0.04
$\hat{\boldsymbol{\beta}}_5^h$	1.50 (edf)	–	8.61	0.01

Table 7
Model (3.9) with the UBRE of 0.0690, the AIC of 1842 and average Poisson deviance of 1.0681 with standard error of 0.0038 and 90% interval of (1.0630, 1.0752).

Parameters	Estimate	Standard error	Test statistics	p-value
$\hat{\beta}_0$	−0.91	0.21	−4.29	0.00
$\hat{\alpha}_2$	−0.50	0.11	−4.78	0.00
$\hat{\alpha}_3$	−0.10	0.15	−0.65	0.51
$\hat{\alpha}_4$	−0.11	0.27	−0.41	0.68
$\hat{\beta}_3^h$	−1.55	0.80	−1.95	0.05
$\hat{\boldsymbol{\beta}}_2$	1.98 (edf)	–	7.52	0.03
$\hat{\boldsymbol{\beta}}_5^h$	1.51 (edf)	–	7.69	0.01

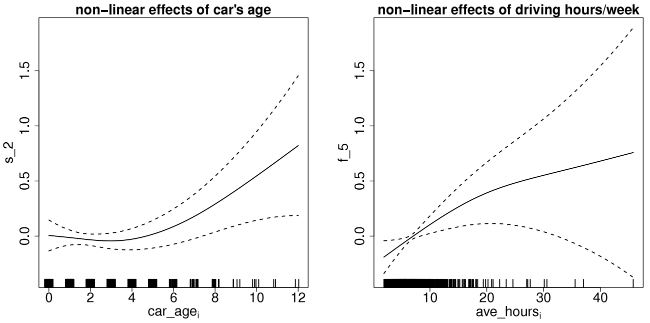

Fig. 6. The non-linear effects of car's age and average driving hours per week.

(3.4):

$$\begin{aligned}\log\lambda_i = &\beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i;\, \boldsymbol{\beta}_1, \delta_1)\\ &+ s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2) + f_1(h_{i,1};\, \boldsymbol{\beta}_1^h, \delta_1^h)\\ &+ f_2(h_{i,2};\, \boldsymbol{\beta}_2^h, \delta_2^h) + f_3(h_{i,3};\, \boldsymbol{\beta}_3^h, \delta_3^h) + f_4(h_{i,4};\, \boldsymbol{\beta}_4^h, \delta_4^h)\\ &+ f_5(ave\_hours_i;\, \boldsymbol{\beta}_5^h, \delta_5^h),\end{aligned} \tag{3.7}$$

where $f_1, \ldots, f_5$ are penalized thin plate regression splines. We apply the backward elimination to model (3.7). We show the p-values of each covariate during the backward elimination procedure in Table 5. Note that for the region we show the p-value of α_2. The last row of Table 5 indicates the following regression function:

$$\log\lambda_i = \beta_0 + \alpha_{region_i} + s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2) + f_5(ave\_hours_i;\, \boldsymbol{\beta}_5^h, \delta_5^h). \tag{3.8}$$

We show the results of model (3.8) in Table 6. By adding the average driving hours to model (3.5), the UBRE decreases by 0.0062, the AIC by 7, and the average Poisson deviance loss by 0.0072.

Linear terms of driving habit covariates. Another starting point of backward elimination is to include linear terms of $(h_{i,m})_{m=1:4}$ and the smooth term of ave_hours_i. That is, we start with the model

$$\begin{aligned}\log\lambda_i = &\beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i;\, \boldsymbol{\beta}_1, \delta_1)\\ &+ s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2)\\ &+ \beta_1^h h_{i,1} + \beta_2^h h_{i,2} + \beta_3^h h_{i,3} + f_5(ave\_hours_i;\, \boldsymbol{\beta}_5^h, \delta_5^h).\end{aligned} \tag{3.9}$$

Note that we have removed $h_{i,4}$ in the model because there is a constraint of $\sum_{m=1}^4 h_{i,m} = 1$ and most cars spend the least time in $(60, 80]$ km/h. This kind of compositional covariates is also studied in Verbelen et al. (2018). The backward elimination leads to the following regression function:

$$\begin{aligned}\log\lambda_i = &\beta_0 + \alpha_{region_i} + s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2) + \beta_3^h h_{i,3}\\ &+ f_5(ave\_hours_i;\, \boldsymbol{\beta}_5^h, \delta_5^h).\end{aligned} \tag{3.10}$$

We show the results of model (3.10) in Table 7. Comparing Tables 6 with 7, it seems that we should include $h_{i,3}$ into the model. The negative sign of $\hat{\beta}_3^h$ indicates that the speed bucket $(40, 60]$ km/h has a lower frequency than $(0, 20]$ km/h and $(20, 40]$ km/h, intuitively this is clear because more claims are caused at lower speeds. We show the estimated non-linear effects of car's age and average driving hours in Fig. 6. The claims frequencies of cars aged less than 4 years are close, while the claims frequencies of older cars increase log-linearly with the car's ages. This is consistent with the observed marginal effect in Fig. 4. The claims frequencies increase with the average driving hours per week and the slope decreases steadily. The dash lines of two standard errors become wider due to the small exposures at larger covariate values. By adding the relative time spent in $(40, 60]$ km/h and the average driving hours to model (3.5), the UBRE has decreased by 0.0074, the AIC by 9, and the average Poisson deviance loss by 0.0088.

3.2.3. Claims frequency modeling with driving habit and driving style covariates

The driving style covariates include 5 clustering covariates $d_{i,1|K=2}$, $d_{i,2|K=2}$, $d_{i,2|K=4}$, $d_{i,3|K=4}$, $d_{i,4|K=4}$, and 20 principal components $(p_{i,w})_{w=1:20}$. We include them in the model by considering either smooth terms or linear terms.

Smooth terms of driving habit and style covariates. We start with the following model:

$$\begin{aligned}\log\lambda_i = &\beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i;\, \boldsymbol{\beta}_1, \delta_1)\\ &+ s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2)\\ &+ f_1(h_{i,1};\, \boldsymbol{\beta}_1^h, \delta_1^h) + f_2(h_{i,2};\, \boldsymbol{\beta}_2^h, \delta_2^h) + f_3(h_{i,3};\, \boldsymbol{\beta}_3^h, \delta_3^h)\\ &+ f_4(h_{i,4};\, \boldsymbol{\beta}_4^h, \delta_4^h)\\ &+ f_5(ave\_hours_i;\, \boldsymbol{\beta}_5^h, \delta_5^h)\\ &+ g_1(d_{i,1|K=2};\, \boldsymbol{\beta}_1^c, \delta_1^c) + g_2(d_{i,2|K=2};\, \boldsymbol{\beta}_2^c, \delta_2^c)\\ &+ g_3(d_{i,2|K=4};\, \boldsymbol{\beta}_3^c, \delta_3^c) + g_4(d_{i,3|K=4};\, \boldsymbol{\beta}_4^c, \delta_4^c)\\ &+ g_5(d_{i,4|K=4};\, \boldsymbol{\beta}_5^c, \delta_5^c)\\ &+ r_1(p_{i,1};\, \boldsymbol{\beta}_1^p, \delta_1^p) + \cdots + r_{20}(p_{i,20};\, \boldsymbol{\beta}_{20}^p, \delta_{20}^p),\end{aligned} \tag{3.11}$$

where we have added 5 clustering covariates and the first 20 principal components to model (3.7). The backward elimination

Table 8
Model (3.12) with the UBRE of 0.0213, the AIC of 1784 and average Poisson deviance loss of 1.0190 with standard error of 0.0081 and 90% interval of (1.0087, 1.0326).

Parameters	Estimate	Standard error	Test statistics	p-value
$\hat{\beta}_0$	−0.081	0.54	−0.15	0.88
$\hat{\alpha}_2$	−0.40	0.11	−3.59	0.00
$\hat{\alpha}_3$	−0.10	0.16	−0.61	0.54
$\hat{\alpha}_4$	−0.18	0.27	−0.65	0.52
$\hat{\beta}^h_2$	−3.25	1.36	−2.39	0.02
$\hat{\beta}^h_3$	−2.18	0.94	−2.31	0.02
$\hat{\beta}^h_5$	0.03	0.01	3.24	0.00
$\hat{\beta}^p_1$	−0.04	0.01	−5.29	0.00
$\hat{\beta}^p_7$	0.11	0.03	3.38	0.00
$\hat{\beta}^p_{15}$	0.11	0.05	2.09	0.04
$\hat{\beta}^p_{16}$	−0.14	0.05	−2.76	0.01
$\hat{\boldsymbol{\beta}}^p_8$	3.23 (edf)	–	11.49	0.03
$\hat{\boldsymbol{\beta}}^p_{10}$	4.69 (edf)	–	15.81	0.01
$\hat{\boldsymbol{\beta}}^p_{12}$	4.79 (edf)	–	18.60	0.00

leads to the following model:

$$\begin{aligned}\log\lambda_i =& \beta_0 + \alpha_{region_i} + \beta^h_2 h_{i,2} + \beta^h_3 h_{i,3} + \beta^h_5 ave\_hours_i \\ &+ \beta^p_1 p_{i,1} + \beta^p_7 p_{i,7} + \beta^p_{15} p_{i,15} + \beta^p_{16} p_{i,16} \\ &+ r_8(p_{i,8}; \boldsymbol{\beta}^p_8, \delta^p_8) + r_{10}(p_{i,10}; \boldsymbol{\beta}^p_{10}, \delta^p_{10}) \\ &+ r_{12}(p_{i,12}; \boldsymbol{\beta}^p_{12}, \delta^p_{12}).\end{aligned} \tag{3.12}$$

Note that we have replaced a smooth term of less than 1.10 effective degrees of freedom by a linear term. The car's age is dropped out due to other more significant covariates. The results of model (3.12) are shown in Table 8. This model gives a big improvement compared to model (3.10). The UBRE has decreased by 0.0477, the AIC by 58, and the average Poisson deviance loss by 0.0491. The driving style covariates have much more predictive power than the driving habit covariates.

Linear terms of driving habit and style covariates. Another starting point of backward elimination is to include the linear terms for $(h_{i,m})_{m=1:4}$ and $(p_{i,w})_{w=1:20}$:

$$\begin{aligned}\log\lambda_i =& \beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i; \boldsymbol{\beta}_1, \delta_1) \\ &+ s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) \\ &+ \beta^h_1 h_{i,1} + \beta^h_2 h_{i,2} + \beta^h_3 h_{i,3} + f_5(ave\_hours_i; \boldsymbol{\beta}^h_5, \delta^h_5) \\ &+ \beta^c_1 d_{i,1|K=2} + \beta^c_2 d_{i,2|K=2} \\ &+ \beta^c_3 d_{i,2|K=4} + \beta^c_4 d_{i,3|K=4} + \beta^c_5 d_{i,4|K=4} \\ &+ \beta^p_1 p_{i,1} + \cdots + \beta^p_{20} p_{i,20}.\end{aligned} \tag{3.13}$$

The backward elimination leads to the following model:

$$\begin{aligned}\log\lambda_i =& \beta_0 + \alpha_{region_i} + \beta^h_3 h_{i,3} + \beta^h_5 ave\_hours_i \\ &+ \beta^p_1 p_{i,1} + \beta^p_3 p_{i,3} + \beta^p_7 p_{i,7} + \beta^p_{10} p_{i,10}.\end{aligned} \tag{3.14}$$

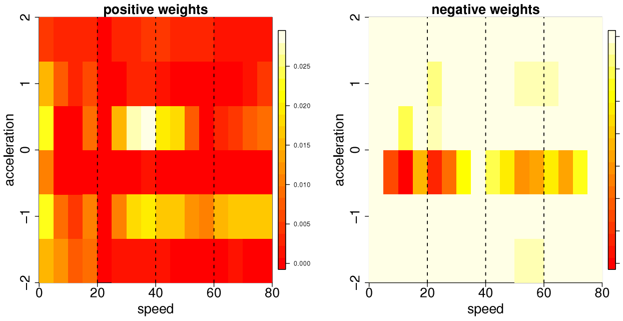

Fig. 7. The weights on the v-a rectangle in model (3.14).

Table 9
Model (3.14) with the UBRE of 0.0404, the AIC of 1813 and average Poisson deviance loss of 1.0365 with standard error of 0.0036 and 90% interval of (1.0311, 1.0422).

Parameters	Estimate	Standard error	Test statistics	p-value
$\hat{\beta}_0$	−0.95	0.25	−3.74	0.00
$\hat{\alpha}_2$	−0.35	0.11	−3.14	0.00
$\hat{\alpha}_3$	−0.07	0.15	−0.49	0.62
$\hat{\alpha}_4$	−0.18	0.27	−0.68	0.50
$\hat{\beta}^h_3$	−2.67	0.94	−2.84	0.00
$\hat{\beta}^h_5$	0.03	0.01	3.03	0.00
$\hat{\beta}^h_1$	−0.05	0.01	−5.70	0.00
$\hat{\beta}^h_3$	0.05	0.02	2.57	0.01
$\hat{\beta}^h_7$	0.06	0.03	2.11	0.03
$\hat{\beta}^h_{10}$	−0.08	0.03	−2.16	0.03

We calculate the weight for sub-rectangle j as $\hat{\beta}^p_1 v_{j,1} + \hat{\beta}^p_3 v_{j,3} + \hat{\beta}^p_7 v_{j,7} + \hat{\beta}^p_{10} v_{j,10}$ for $j = 1, \ldots, J$. We plot these weights in the v-a rectangle according to their signs in Fig. 7. Note that the acceleration zero is counted in the sub-rectangles under the acceleration zero line. Most sub-rectangles in $(0, 20]$ km/h are highlighted, indicating that $(0, 20]$ km/h is important in predicting claims frequency. Hard brake and acceleration have the positive effect on claims frequency, while smooth brake and acceleration have the negative effect on claims frequency. We show the results of model (3.14) in Table 9, which has a worse performance than model (3.12), but still much better than model (3.10).

3.2.4. Claims frequency modeling with driving style covariates in each speed bucket

In this section, we include the smooth terms of driving style covariates. Note that we do not need to consider the clustering covariates since they have linear relationship with principal components. For each speed bucket m, we either start with the model

$$\begin{aligned}\log\lambda_i =& \beta_0 + \alpha_{region_i} + \gamma_{gender_i} + s_1(driver\_age_i; \boldsymbol{\beta}_1, \delta_1) \\ &+ s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) \\ &+ f_1(h_{i,1}; \boldsymbol{\beta}^h_1, \delta^h_1) + f_2(h_{i,2}; \boldsymbol{\beta}^h_2, \delta^h_2) + f_3(h_{i,3}; \boldsymbol{\beta}^h_3, \delta^h_3) \\ &+ f_4(h_{i,4}; \boldsymbol{\beta}^h_4, \delta^h_4) \\ &+ f_5(ave\_hours_i; \boldsymbol{\beta}^h_5, \delta^h_5) \\ &+ r^m_1(p^m_{i,1}; \boldsymbol{\beta}^m_1, \delta^m_1) + \cdots + r^m_7(p^m_{i,7}; \boldsymbol{\beta}^m_7, \delta^m_7),\end{aligned} \tag{3.15}$$

or start with the model with only driving style covariates

$$\log\lambda_i = \beta_0 + r^m_1(p^m_{i,1}; \boldsymbol{\beta}^m_1, \delta^m_1) + \cdots + r^m_7(p^m_{i,7}; \boldsymbol{\beta}^m_7, \delta^m_7), \tag{3.16}$$

Note that in the following, we directly present the resulting models from backward elimination.

The first speed bucket $(0, 20]$ km/h. It turns out that only the first principal $p^1_{i,1}$ is significant among the 7 principal components and it has 1 effective degree of freedom. The model is as follows:

$$\begin{aligned}\log\lambda_i =& \beta_0 + \alpha_{region_i} + s_2(car\_age_i; \boldsymbol{\beta}_2, \delta_2) + \beta^h_3 h_{i,3} \\ &+ f_5(ave\_hours_i; \boldsymbol{\beta}^h_5, \delta^h_5) + \beta^1_1 p^1_{i,1}.\end{aligned} \tag{3.17}$$

Table 10
The comparison of predictive performance of different speed buckets.

Speed bucket	model	UBRE	AIC	Average Poisson deviance (sd.)	5%	95%
(0, 20] km/h	(3.17)	0.0399	1809	1.0332(0.0040)	1.0274	1.0407
(20, 40] km/h	(3.19)	0.0399	1809	1.0393(0.0061)	1.0320	1.0514
(40, 60] km/h	(3.21)	0.0493	1822	1.0471(0.0039)	1.0415	1.0538
(60, 80] km/h	(3.23)	0.0567	1829	1.0547(0.0038)	1.0499	1.0605
(0, 20] km/h	(3.18)	0.0564	1835	1.0561(0.0018)	1.0534	1.0593
(20, 40] km/h	(3.20)	0.0714	1850	1.0713(0.0019)	1.0684	1.0748
(40, 60] km/h	(3.22)	0.0682	1846	1.0683(0.0023)	1.0649	1.0718
(60, 80] km/h	(3.24)	0.0826	1861	1.0819(0.0017)	1.0794	1.0849

If only the driving style covariates $(p^1_{i,w})_{w=1:7}$ are considered, the backward elimination leads to the following model:

$$\log \lambda_i = \beta_0 + \beta^1_1 p^1_{i,1}. \tag{3.18}$$

We show the results of models (3.17) and (3.18) in Table 10.

The second speed bucket (20, 40] km/h. It turns out that the first and seventh principal components $p^2_{i,1}, p^2_{i,7}$ are significant among the 7 principal components, and the smooth term for the seventh principal component has more than 1.1 effective degrees of freedom. The model is as follows:

$$\begin{aligned}\log \lambda_i =&\beta_0 + \alpha_{region_i} + s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2) + \beta^h_3 h_{i,3}\\ &+ f_5(ave\_hours_i;\, \boldsymbol{\beta}^h_5, \delta^h_5) + \beta^2_1 p^2_{i,1} + r^2_7(p^2_{i,7};\, \boldsymbol{\beta}^2_7, \delta^2_7).\end{aligned} \tag{3.19}$$

If only the driving style covariates $(p^2_{i,w})_{w=1:7}$ are considered, the backward elimination leads to the following model:

$$\log \lambda_i = \beta_0 + \beta^2_1 p^2_{i,1}. \tag{3.20}$$

We show the results of models (3.19) and (3.20) in Table 10.

The third speed bucket (40, 60] km/h. It turns out that only the first principal $p^3_{i,1}$ is significant among the 7 principal components, and it has 1 effective degree of freedom. The model is as follows:

$$\begin{aligned}\log \lambda_i =&\beta_0 + \alpha_{region_i} + s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2) + \beta^h_3 h_{i,3}\\ &+ f_5(ave\_hours_i;\, \boldsymbol{\beta}^h_5, \delta^h_5) + \beta^3_1 p^3_{i,1}.\end{aligned} \tag{3.21}$$

If only the driving style covariates $(p^3_{i,w})_{w=1:7}$ are considered, the backward elimination leads to the following model:

$$\log \lambda_i = \beta_0 + \beta^3_1 p^3_{i,1} + \beta^3_4 p^3_{i,4}. \tag{3.22}$$

We show the results of models (3.21) and (3.22) in Table 10.

The fourth speed bucket (60, 80] *km/h*. It turns out that only the first principal $p^4_{i,1}$ is significant among the 7 principal components, and it has 1 effective degree of freedom. The model is as follows:

$$\begin{aligned}\log \lambda_i =&\ \beta_0 + \alpha_{region_i} + s_2(car\_age_i;\, \boldsymbol{\beta}_2, \delta_2) + \beta^h_3 h_{i,3}\\ &+ f_5(ave\_hours_i;\, \boldsymbol{\beta}^h_5, \delta^h_5) + \beta^4_1 p^4_{i,1}.\end{aligned} \tag{3.23}$$

If only the driving style covariates $(p^4_{i,w})_{w=1:7}$ are considered, the backward elimination leads to the following model:

$$\log \lambda_i = \beta_0 + \beta^4_1 p^4_{i,1}. \tag{3.24}$$

We show the results of models (3.23) and (3.24) in Table 10.

From Table 10, we conclude that the driving style covariates in the low speed buckets have a better predictive performance than the driving style covariates in the high speed buckets.

Table 11
The selected representative models.

Model index	Covariates in the model	Equation
1	No covariates	(3.6)
2	Classical	(3.5)
3	Classical, driving habit	(3.10)
4	Classical, driving habit, driving style (in smooth terms)	(3.12)
5	Classical, driving habit, driving style (in linear terms)	(3.14)
6	Classical, driving habit, driving style of (0, 20] km/h	(3.17)
7	Classical, driving habit, driving style of (20, 40] km/h	(3.19)
8	Classical, driving habit, driving style of (40, 60] km/h	(3.21)
9	Classical, driving habit, driving style of (60, 80] km/h	(3.23)
10	Driving style of (0, 20] km/h	(3.18)
11	Driving style of (20, 40] km/h	(3.20)
12	Driving style of (40, 60] km/h	(3.22)
13	Driving style of (60, 80] km/h	(3.24)

3.2.5. *Model comparison*

We select several representative models for comparison, which has been listed in Table 11. We plot the UBRE, the AIC and the average Poisson deviance loss with 90% interval for these selected models in Fig. 8. According to UBRE, AIC and average Poisson deviance loss, Model 4 has the best predictive performance. The 90% interval of average Poisson deviance loss for Model 4 is very wide because there are three smooth terms in Model 4. Note that Model 4 has only one classical risk factor for region. The decreasing pattern from Model 1 to 4 indicates that driving style covariates have a much better predictive power than driving habit covariates. Comparing Models 4–6, we conclude that the driving style covariates from (0, 20] km/h already have the similar predictive power to the driving style covariates from the whole speed range (0, 80] km/h. The increasing pattern from Model 6 to 13 indicates that the driving style covariates in low speed buckets have a better predictive power than those in high speed buckets. Finally, comparing Models 3 and 10, we conclude that the driving style covariates from (0, 20] km/h have a better predictive power than the classical risk factors and driving habit covariates. This answers the three questions at the beginning of Section 3.

4. Conclusions

We have studied three months telematics car driving data. We construct four v-a heatmaps in four speed buckets for each car. Driving habit covariates and driving style covariates are extracted from the v-a heatmaps using unsupervised learning algorithms. We then investigate the predictive power of the classical risk factors, the driving habit covariates and the driving style covariates for claims frequency modeling. We implement the backward elimination to select variables and determine the final models. We compare the models in terms of UBRE, AIC and average Poisson deviance loss. The main results have been summarized in Table 11 and Fig. 8. We list several important findings as follows:

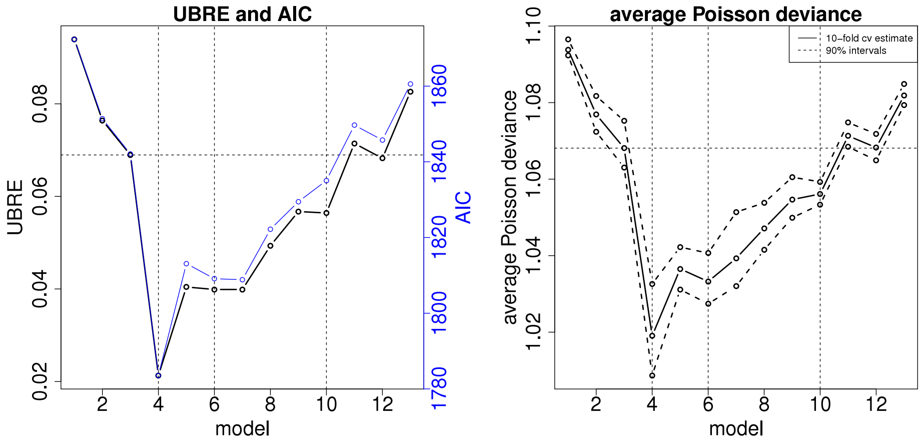

Fig. 8. The UBRE, the AIC and the average Poisson deviance loss with 90% interval for the models in Table 11.

- Driving style is much more related to claims frequency than driving habit.
- The average driving hours per week have a positive effect on claims frequency, and this effect is decreasing with the average driving hours per week.
- Driving at the speeds (20, 60] km/h is safer for claims frequencies than at the speeds (0, 20] km/h or (60, 80] km/h.
- The driving style in (0, 20] km/h is the most related covariate to claims frequencies among the four speed buckets, and it also reflects the driving style at other speeds.

We realize that our analysis is conducted on a comparably small portfolio of 973 cars. So the predictive power of v-a heatmaps may not be fully explored. Note that driver's age is not significant in our analysis probably due to the small portfolio. We implement the unsupervised learning algorithms, clustering analysis and principal components analysis, to extract covariates from v-a heatmaps, and fortunately the first principal component is the most important principal component to predict claims frequencies. Another option is to implement a supervised principal components analysis, which first finds the most important subrectangles in Fig. 1 to predict claims frequencies (Bair et al., 2006).

Acknowledgments

We gratefully acknowledge support from the Forschungsinstitut für Mathematik (FIM) during the research stay of the first author at ETH Zurich. We also thank the anonymous referees for their useful comments that have helped to improve our manuscript. Guangyuan Gao gratefully acknowledges financial support from the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (19XNF023).

References

Ayuso, M., Guillen, M., Nielsen, J.P., 2018. Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. Transportation http://dx.doi.org/10.1007/s11116-018-9890-7.

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

Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. J. Amer. Statist. Assoc. 101 (473), 119–137.

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

Denuit, M., Guillen, M., Trufin, J., 2019. Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. Ann. Actuar. Sci. http://dx.doi.org/10.1017/S1748499518000349.

Gao, G., Meng, S., Wüthrich, M.V., 2019. Claims frequency modeling using telematics car driving data. Scand. Actuar. J. 2019 (2), 143–162.

Guillen, M., Nielsen, J.P., Ayuso, M., Pérez-Marín, A.M., 2019. The use of telematics devices to improve automobile insurance rates. Risk Anal. 39 (3), 662–672.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, second ed. Springer-Verlag, New York.

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. Transp. Res. D 12 (2), 115–128.

Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Paefgen, J., Staake, T., Fleisch, E., 2014. Multivariate exposure modeling of accident risk: insights from pay-as-you-drive insurance data. Transp. Res. A 61, 27–40.

Reynolds, A., Richards, G., de la Iglesia, B., Rayward-Smith, V., 1992. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. J. Math. Model. Algorithms 5 (4), 475–504.

Verbelen, R., Antonio, K., Claeskens, G., 2018. Unraveling the predictive power of telematics data in car insurance pricing. J. R. Stat. Soc. Ser. C. Appl. Stat. 67 (5), 1275–1304.

Wang, Q., Huo, H., He, K., Yao, Z., Zhang, Q., 2008a. Characterization of vehicle driving patterns and development of driving cycles in chinese cities. Transp. Res. D 13 (5), 289–297.

Weidner, W., Transchel, F.W.G., Weidner, R., 2016. Classification of scale-sensitive telematic observables for riskindividual pricing. Eur. Actuar. J.urnal 6 (1), 3–24.

Weidner, W., Transchel, F.W.G., Weidner, R., 2016b. Telematic driving profile classification in car insurance pricing. Ann. Actuar. Sci. 11 (2), 213–236.

Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, second ed. Chapman & Hall, New York.