Covariate Selection from Telematics Car Driving Data

Mario V. Wüthrich[*][†]

December 19, 2016

Abstract

Car insurance companies have started to collect high-frequency GPS location data of their car drivers. This data provides detailed information about the driving habits and driving styles of individual car drivers. We illustrate how this data can be analyzed using techniques from pattern recognition and machine learning. In particular, we describe how driving styles can be categorized so that they can be used for a regression analysis in car insurance pricing.

Keywords. Telematics data, driving habits, driving styles, regression, categorical classes, pattern recognition, clustering, K-means clustering, unsupervised learning, machine learning.
.

1 Introduction

Nowadays, many cars are equipped with technology that transmits car driving information via telecommunication systems to central data warehouses. These data transmissions (called telematics data) comprise detailed car driving information. In the foreseeable future this information will be used for car insurance pricing because it displays driving habits and styles rather transparently, in fact, it is likely that this information will complement classical covariate information like age of driver, type and size of car, etc. These car driving transmissions may comprise rather different information (depending on the installed devices). This information may include high-frequency data about location (typically GPS data), speed, acceleration and braking, left- and right-turns. Moreover, it may include number of trips, total distance of trips, total duration of trips, day time of trips, road conditions, road types, as well as driver information. The main difficulty from a statistical point of view is to convert this high-dimensional and high-frequency data into useful covariate information.

In the actuarial literature there is only a very limited number of contributions that studies telematics data from a statistical point of view. Verbelen et al. [4] use generalized additive models to analyze the effect of telematics data in completion to classical covariates. Their study considers, however, rather classical information like number of trips, total distance driven, road type, and daytime and weekday of driving. It does not study speed, acceleration and braking,

[*] ETH Zurich, RiskLab, Department of Mathematics, 8092 Zurich, Switzerland

[†] Swiss Finance Institute Professor

left- and right-turns. Probably, the first contribution in the actuarial literature that considers this latter telematics data information is Weidner et al. [5]. These authors use Fourier analysis for pattern recognition to study driving behavior. In particular, they analyze the frequency spectrum obtained by single driving maneuvers and trips. We use a different approach here by studying simultaneously a speed and acceleration heatmap.

In this work we consider *classification* of different driving styles. We therefore use tools from the field of unsupervised learning, which is an area in machine learning that deals with cluster analysis and pattern recognition. For an introduction to machine learning we refer to Breiman et al. [1], Hastie et al. [2], James et al. [3], and in an actuarial context to Wüthrich–Buser [6].

Organization of this manuscript. In the next section we describe the available telematics data which in our situation is high-frequency GPS data. Firstly, we provide some simple statistics of this telematics data, and secondly, we introduce the v-a heatmap which is a two-dimensional (graphical) representation of the speed and acceleration behavior. In Section 3 we introduce the main tools of unsupervised learning. We first describe the dissimilarity of two different driving styles. This measure is used to classify different driving styles, which is done algorithmically by the so-called K-means clustering algorithm. Finally, in Section 4 we illustrate the algorithm on real high-frequency GPS data.

2 Description of the available telematics data

For this analysis we consider telematics data which comprises GPS location data second by second. We have this data of 1,753 individual car drivers. For each of these car drivers we have recorded 200 individual trips. For confidentiality reasons all individual trips are initialized to start at location $(0, 0)$, and they are randomly rotated (at this origin) and reflected at the coordinate axes. These transformations do not change the crucial driving characteristics like speed, absolute value of acceleration and braking, and intensity of turns. Therefore they are appropriate for our analysis. Unfortunately, we do not have any other information, like daytime of trip, type of car, etc., nor do we have any claims information. For the latter cause we cannot perform a risk assessment based on a supervised learning method, see Section 14.1 in Hastie et al. [2]. However, this telematics data allows us to classify the drivers according to their driving styles (unsupervised learning). This classification provides categorical classes which can be used (in a next step and having claims information) as covariates for insurance pricing, see Wüthrich–Buser [6].

In Figure 1 (lhs) we illustrate an example of a car driver doing 200 individual trips. Three of these individual trips are colored blue, red and green, respectively. Each of the individual trips consists of $(x_t, y_t)_t$ location data (in meters) second by second (denoted by $t \geq 0$). This location data allows us to calculate the average speed in each time interval $(t-1, t]$, for $t \geq 1$, given by

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

This average speed has unit m/s; for the illustrations we typically transform it to km/h. These

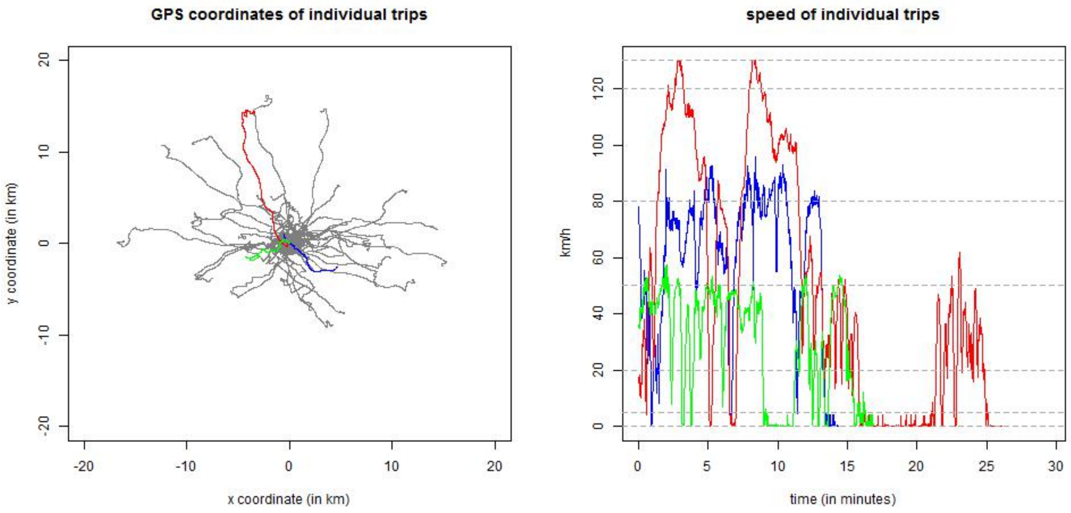

Figure 1: (lhs) 200 individual trips of a car driver with (rhs) resulting speed profiles of the three colored trips.

average speeds allow us to calculate an average acceleration/braking, for $t \geq 2$, defined by

$$a_t = (v_t - v_{t-1})/(t - (t-1)).$$

This average acceleration/braking has unit m/s^2. Note that the interpretation of this average acceleration/braking needs some care because it is determined from *average* speeds (and not from speeds). Therefore, it rather refers to the change in average speed. Nevertheless, we will call it average acceleration because it is the best approximation that we can get from our data, and considering the inertia of cars it is appropriate (i.e. sufficiently close to an instantaneous derivative). In Figure 1 (rhs) we plot the speed profiles of the three colored trips on the left-hand side of the figure. The green trip seems to be done in city area (speeds below 50 km/h), the blue trip is overland (with speeds around 80 km/h), and the red trip has some highway parts (where the maximal speed was even above 120 km/h). The red trip also shows a time period where the car has been standing for roughly 5 minutes (between minutes 16 and 21 of that trip).

2.1 Simple empirical statistics

In this section we consider three different drivers (called drivers A, B and C, respectively). Their 200 individual trips are shown in Figure 2. In orange color we plot the shorter 100 trips and in gray color the longer 100 trips. We see that driver A usually travels short distances (with a radius of less than 5 km) but he also drives a few longer trips; driver B is a long-distance traveler (with many trips longer than 10 km); and driver C only does shorter drives. In Table 1 we provide some simple empirical statistics of these three drivers, these include the total distance driven in these 200 trips and the total time therefore used. Then, we provide the average trip lengths (in distance and time), the average speed over all trips, and the median speed of the single trips.

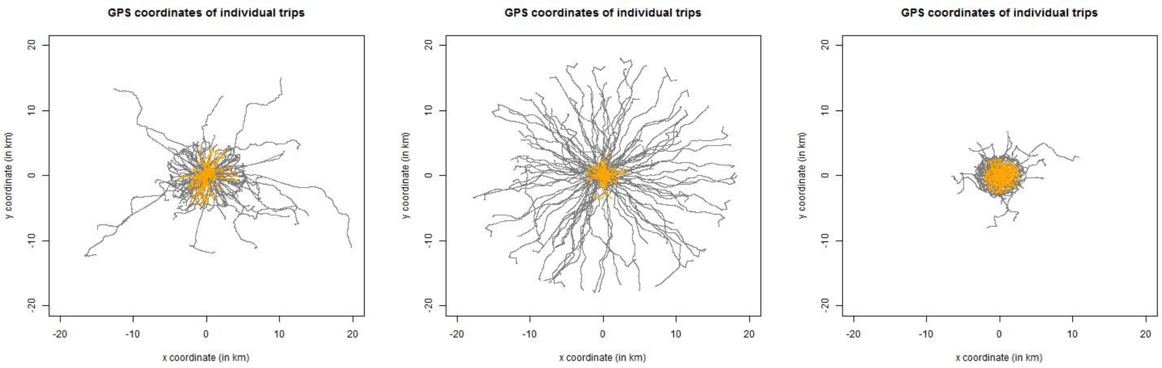

Figure 2: 200 individual trips of the three different car drivers A (left), B (middle) and C (right); in orange color are the shorter 100 trips and in gray color the longer 100 trips.

	driver A	driver B	driver C
total distance (in km)	1,235	1,808	1,001
average distance per trip (in km)	6.18	9.04	5.01
total time (in h)	40.84	51.27	39.43
average time per trip (in min)	12.15	15.38	11.83
average speed (in km/h)	30.25	35.27	25.39
median speed over trips (in km/h)	28.83	35.91	25.29

Table 1: Empirical statistics of the drivers A, B and C.

These statistics basically reflect the graphs in Figure 2. Next we analyze the amount of time spent in different speed buckets. We therefore consider the speed intervals (in km/h)

$[0]$	car stands still (and also does not accelerate),
$(0, 5]$	acceleration or braking phase (from/to speed 0),
$(5, 20]$	low speeds,
$(20, 50]$	urban area speeds,
$(50, 80]$	rural area speeds,
$(80, 130]$	highway speeds (we truncate speeds above 130 km/h).

In Figure 3 we show the amount of time spent in each of these speed buckets for the three drivers A, B and C. We observe that the distribution of the resulting speeds varies considerably over the speed buckets; not surprisingly driver B drives almost half of his time with speeds above 50 km/h, whereas the other two drivers mostly drive with speeds below 50 km/h. Also remarkable is that driver A stands still 28% of his total trip time of 40.84 hours (of the 200 trips), the corresponding figures for drivers B and C are 24% of 51.27 hours and 19% of 39.43 hours, respectively.

Up to now we have only been considering driving habits, i.e. whether we have a long-distance driver, an urban area driver, etc. We could calculate many more of these simple empirical statistics. These statistics can be transformed into covariate information which provides important complementary information to the classical pricing criteria. However, the statistics considered

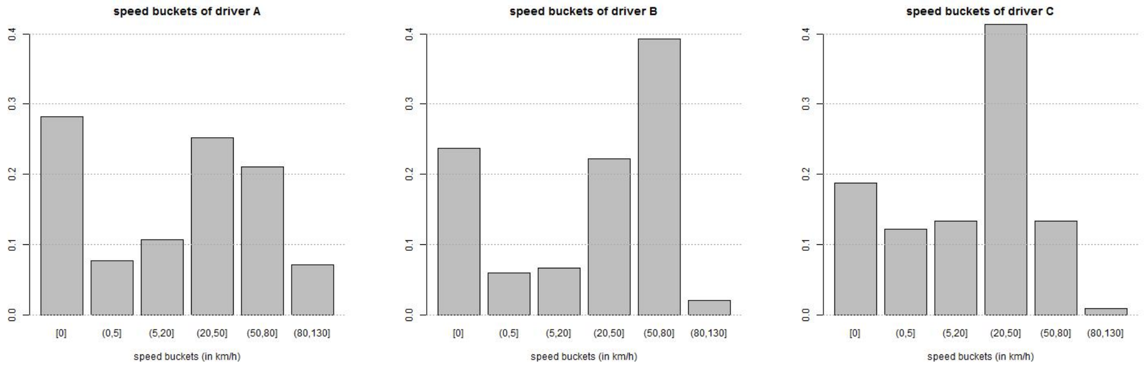

Figure 3: Speed bucket distribution (in driving time) of the three drivers A, B and C.

so far do not say much about driving styles. This analysis is our next aim.

2.2 The v-a heatmap

To analyze driving styles we consider so-called v-a heatmaps that display the average speed on the x-axis (in km/h) and the corresponding acceleration/braking pattern on the y-axis (in m/s^2). For this analysis we consider the same speed buckets as used in Figure 3. Speed buckets have the advantage that the different driving habits do not directly influence the v-a heatmap analysis, if we consider in each speed bucket the resulting empirical density (normalized to 1). These v-a heatmaps for the different speed buckets are provided in Figure 4: the column on the left-hand side shows driver A, the middle column gives driver B and the column on the right-hand side corresponds to driver C. The five different rows show the five non-zero speed buckets.

In the first row we have the acceleration and braking styles of the three drivers in speed bucket $(0, 5]$ km/h. We observe that these look quite differently. Driver B seems to accelerate more intensively than driver C and it seems that he also brakes more heavily than driver C. This may indicate that driver B is a more aggressive driver than driver C (because he needs to brake more abruptly, i.e. he approaches a red light at a higher speed). Driver A has intermediate braking and acceleration maxima, this indicates that he drives a manual gear car. The same consideration applies to the speed bucket $(5, 20]$ km/h in the second row of Figure 4, in particular, the v-a heatmaps of drivers B and C look much more smooth which is implied by driving an automatic gear car (compared to driver A).

Rows three and four of Figure 4 are interpreted such that driver B has the smoothest driving style in these two speed buckets because the vertical diameter (level sets) of his heatmaps are smaller compared to the other two drivers. This may also be caused by the possibility that driver B drives a heavier car that is less agile than the other two drivers. Note also that the light upper and lower acceleration boundaries in these graphs are caused by the fact that we truncate (here) acceleration and braking at 2 m/s^2 and -2 m/s^2, respectively. Finally, the last row of Figure 4 shows the driving styles above speed 80 km/h. We observe that in this speed

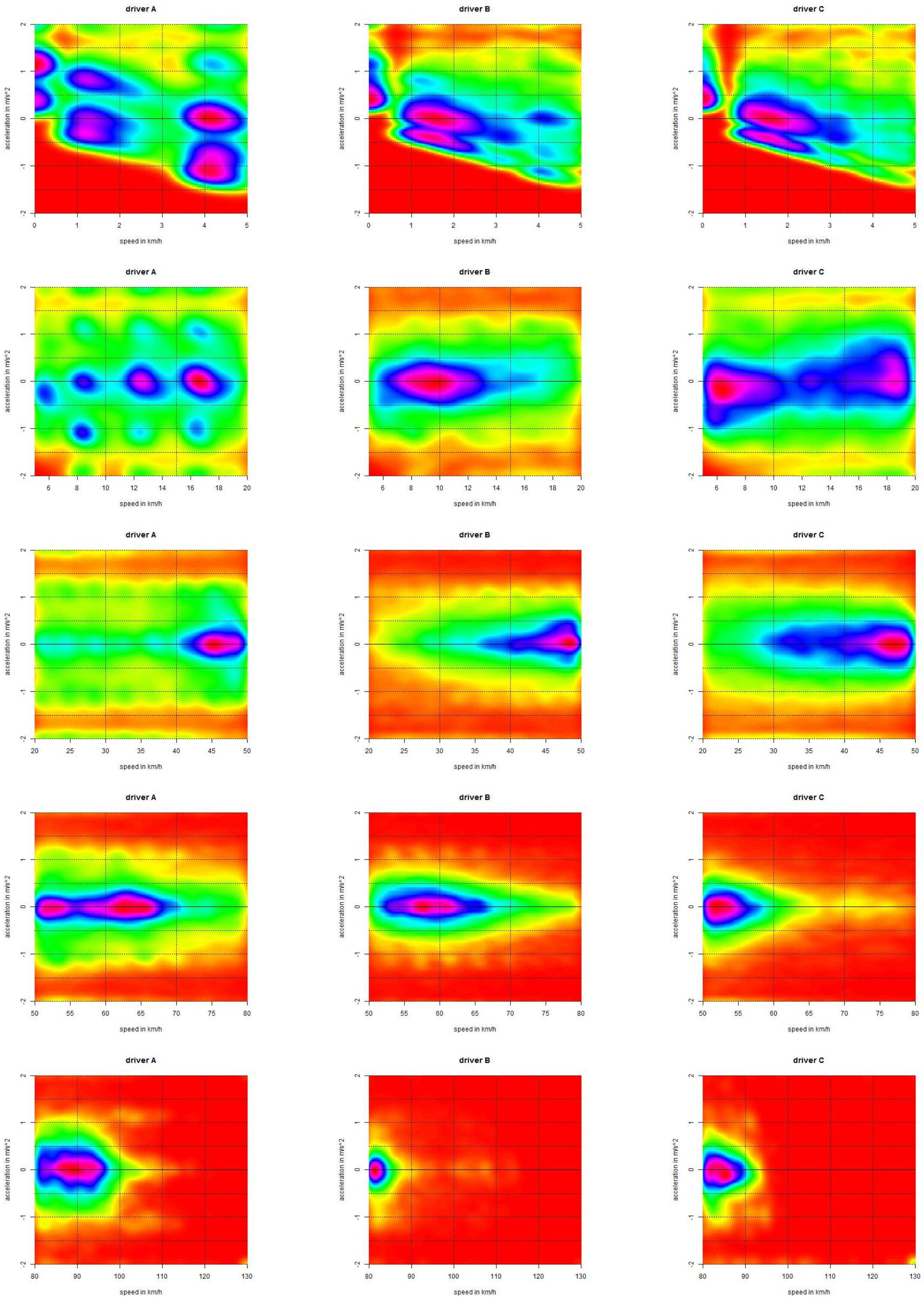

Figure 4: v-a heatmaps of car drivers A (left), B (middle) and C (right) in the different speed buckets $(0, 5]$, $(5, 20]$, $(20, 50]$, $(50, 80]$ and $(80, 130]$ km/h, respectively.

bucket we only have limited information because our three drivers only spend little time in this speed bucket.
Similar graphs could be provided for left- and right-turns (using the changes in angles obtained from the GPS data). We refrain from considering them here because the analysis of left- and right-turns can be done completely analogously to the one of acceleration and braking.

3 Cluster analysis

Our main goal is to classify the v-a heatmaps, i.e. we would like to identify the drivers that have similar v-a heatmaps. These drivers are then interpreted to have similar driving styles (according to their v-a heatmaps), and are therefore put into the same categorical classes for car insurance pricing. We do this for each speed bucket independently (and dependence is induced by regressing from these categorical classes simultaneously). Thus, we choose a fixed speed bucket and denote the resulting v-a rectangle by $R = (v_0, v_1] \times [a_0, a_1] \subset \mathbb{R}_+ \times \mathbb{R}$. In Figure 4 we choose maximal acceleration and braking $-a_0 = a_1 = 2$ m/s^2. Then, we consider the set of all probability distributions on this rectangle R, we use notation $\mathcal{P}(R)$ for this set of probability distributions. Since in the plots of Figure 4 we consider (normalized) empirical probability distributions (of each car driver in each plot), classification of the individual car drivers is achieved by looking at a classifier function

$$\mathcal{C} : \mathcal{P}(R) \to \mathcal{K}, \qquad F \mapsto \mathcal{C}(F), \tag{3.1}$$

with $\mathcal{K} = \{1, \ldots, K\}$ describing the K different categorical classes considered, and with $F \in \mathcal{P}(R)$ being the (empirical) probability distribution on R describing a particular car driver. The main difficulty is to find an appropriate classifier $\mathcal{C}$ that is able to cluster similar plots. For instance, on the first row of Figure 4, we would (typically) require that driver A falls into a different class than drivers B and C, i.e.,

$$\mathcal{C}(F_{\mathrm{A}}) \ \notin \ \{\mathcal{C}(F_{\mathrm{B}}), \mathcal{C}(F_{\mathrm{C}})\},$$

if $F_b \in \mathcal{P}(R)$ describes the empirical probability distributions of drivers $b =$ A, B, C on R.

3.1 Dissimilarity function

In order to construct a classifier $\mathcal{C}$ we start by describing the dissimilarity between two different probability distributions $F_b, F_c \in \mathcal{P}(R)$. We therefore partition the rectangle R into J different (non-empty) subsets $R_1, \ldots, R_J$ (typically also rectangles) with

$$\bigcup_{j=1}^{J} R_j = R \qquad \text{and} \qquad R_j \cap R_{j'} = \emptyset \quad \text{for all } j \neq j'.$$

We then consider the discretized probability distribution of $F \in \mathcal{P}(R)$ having probability weights

$$x_j = \int_{R_j} dF \ \geq 0, \qquad \text{for } j = 1, \ldots, J, \quad \text{satisfying} \quad \sum_{j=1}^{J} x_j = 1. \tag{3.2}$$

The probability weight x_j measures the total probability mass of F on subset R_j. We choose two probability distributions $F_b, F_c \in \mathcal{P}(R)$ with resulting discretized probability weights $\boldsymbol{x}_b = (x_{b,j})_j$ and $\boldsymbol{x}_c = (x_{c,j})_j$ on partition $(R_j)_j$. The dissimilarity between F_b and F_c is defined by

$$d(\boldsymbol{x}_b, \boldsymbol{x}_c) \;=\; d_w(\boldsymbol{x}_b, \boldsymbol{x}_c) \;=\; \frac{1}{2} \sum_{j=1}^{J} w_j \left(x_{b,j} - x_{c,j}\right)^2,$$

where $w_j \geq 0$ are predefined weights. Remark that we choose the (weighted) square distance because it has nice analytical properties (as we will see below), other choices are described in Section 14.3.2 of Hastie et al. [2]. The weights $w = (w_j)_j$ may allow us to emphasize dissimilarities on certain subsets R_j more than on others. Below, for simplicity, we only consider $w \equiv 1$.

In our classification problem we have I different car drivers labeled by $b \in \mathcal{I} = \{1, \ldots, I\}$. We denote their empirical probability distributions on the rectangle R by F_b, for $b \in \mathcal{I}$. These empirical probability distributions F_b correspond exactly to the v-a heatmaps (described by $\boldsymbol{x}_b$) on the chosen speed bucket. The *total dissimilarity* over all drivers is defined by

$$D(\mathcal{I}) \;=\; \frac{1}{I} \sum_{b,c=1}^{I} d(\boldsymbol{x}_b, \boldsymbol{x}_c).$$

Lemma 3.1 *The total dissimilarity over all drivers satisfies*

$$D(\mathcal{I}) \;=\; \sum_{j=1}^{J} w_j \sum_{b=1}^{I} \left(x_{b,j} - \bar{x}_j\right)^2,$$

with (empirical) means $\bar{x}_j = I^{-1} \sum_{b=1}^{I} x_{b,j}$, *for* $j = 1, \ldots, J$.

Proof of Lemma 3.1. A straightforward calculation provides

$$\begin{aligned} D(\mathcal{I}) &= \frac{1}{2I} \sum_{b,c=1}^{I} \sum_{j=1}^{J} w_j \left(x_{b,j} - x_{c,j}\right)^2 \;=\; \frac{1}{2I} \sum_{j=1}^{J} w_j \sum_{b,c=1}^{I} \left(x_{b,j} - \bar{x}_j + \bar{x}_j - x_{c,j}\right)^2 \\ &= \frac{1}{2I} \sum_{j=1}^{J} w_j \sum_{b,c=1}^{I} \left(x_{b,j} - \bar{x}_j\right)^2 + 2\left(x_{b,j} - \bar{x}_j\right)\left(\bar{x}_j - x_{c,j}\right) + \left(\bar{x}_j - x_{c,j}\right)^2 \\ &= \sum_{j=1}^{J} w_j \sum_{b=1}^{I} \left(x_{b,j} - \bar{x}_j\right)^2 + \frac{1}{I} \sum_{j=1}^{J} w_j \sum_{b,c=1}^{I} \left(x_{b,j} - \bar{x}_j\right)\left(\bar{x}_j - x_{c,j}\right). \end{aligned}$$

The second term is zero and the claim follows.

□

Lemma 3.1 has a nice interpretation: $\bar{x}_j$ is the empirical mean of all probability weights $x_{1,j}, \ldots, x_{I,j}$ of the I drivers on subset R_j and the dissimilarity in these probability weights is measured by the (empirical) variance defined by

$$s_j^2 \;=\; \frac{1}{I} \sum_{b=1}^{I} \left(x_{b,j} - \bar{x}_j\right)^2.$$

Thus, the total dissimilarity over all drivers is given by

$$D(\mathcal{I}) \;=\; I \sum_{j=1}^{J} w_j \; s_j^2.$$

The weights $w_j \geq 0$ can now be chosen such that different subsets R_j are weighted differently. For instance, if we are concerned with high acceleration, then we would give more weight to subsets R_j that cover the high acceleration region in R. In the examples below we will choose $w \equiv 1$. The following lemma is immediate.

Lemma 3.2 *The empirical means $\bar{x}_j$ are optimal in the sense that*

$$\bar{x}_j \;=\; \underset{m_j}{\operatorname{argmin}} \; \sum_{b=1}^{I} \left(x_{b,j} - m_j\right)^2.$$

Lemma 3.2 says that the total dissimilarity $D(\mathcal{I})$ is found by solving an optimization problem on each subset R_j. Moreover, this optimal solution $(\bar{x}_j)_{j=1,\ldots,J}$ is itself a probability distribution on R because it can be interpreted as a mixing distribution: all elements satisfy $\bar{x}_j \in [0,1]$ and

$$\sum_{j=1}^{J} \bar{x}_j \;=\; \frac{1}{I} \sum_{b=1}^{I} \sum_{j=1}^{J} x_{b,j} \;=\; 1.$$

Thus, $(\bar{x}_j)_{j=1,\ldots,J}$ is the discrete probability distribution that solves the optimization problems in Lemma 3.2 simultaneously for all $j = 1, \ldots, J$.

3.2 Classifier and clustering

In the previous section we have considered the total dissimilarity over all drivers given by

$$D(\mathcal{I}) \;=\; \frac{1}{2I} \sum_{b,c=1}^{I} \sum_{j=1}^{J} w_j \left(x_{b,j} - x_{c,j}\right)^2 \;=\; \sum_{j=1}^{J} w_j \sum_{b=1}^{I} \left(x_{b,j} - \bar{x}_j\right)^2.$$

This corresponds to the total (weighted) sum of squares (TSS), scaled with I^{-1}. This is the error measure if we do not assume any additional model structure.

We now introduce a regression structure by partitioning the set $\mathcal{I}$ of all drivers into K (non-empty) *clusters* $\mathcal{I}_1, \ldots, \mathcal{I}_K$ satisfying

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

These K clusters define a natural classifier $\mathcal{C} = \mathcal{C}_K$, restricted to the drivers $b \in \mathcal{I}$, given by

$$\mathcal{C} : \mathcal{I} \to \mathcal{K}, \qquad b \mapsto \mathcal{C}(b) = \sum_{k=1}^{K} k \; \mathbb{1}_{\{b \in \mathcal{I}_k\}}.$$

Note that we use a slight abuse of notation here: at the moment the classifier $\mathcal{C}$ is only defined on $\mathcal{I}$, this is in contrast to (3.1). Its extension to $\mathcal{P}(R)$ is provided in Remarks 3.4, below.

Under this regression approach on $\mathcal{I}$ we can decompose the TSS into two parts, the within-cluster sum of squares (WSS) and the between-cluster sum of squares (BSS), the latter being explained by the chosen regression structure. That is,

$$\begin{aligned} D(\mathcal{I}) &= \sum_{k=1}^{K}\sum_{j=1}^{J} w_j \sum_{b\in\mathcal{I}_k} (x_{b,j}-\bar{x}_j)^2 \qquad (3.3) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J} w_j \sum_{b\in\mathcal{I}_k} \left(x_{b,j}-\bar{x}_{j|k}\right)^2 \; + \; \sum_{k=1}^{K} I_k \sum_{j=1}^{J} w_j \left(\bar{x}_{j|k}-\bar{x}_j\right)^2, \end{aligned}$$

where $I_k = |\mathcal{I}_k|$ is the number of drivers in $\mathcal{I}_k$ and the empirical means on $\mathcal{I}_k$ are given by

$$\bar{x}_{j|k} = \frac{1}{I_k} \sum_{b\in\mathcal{I}_k} x_{b,j}. \qquad (3.4)$$

The last term on the right-hand side of (3.3) is interpreted as the *between-cluster dissimilarity* of classifier $\mathcal{C}$, defined by

$$B(\mathcal{C}) = \sum_{k=1}^{K} I_k \sum_{j=1}^{J} w_j \left(\bar{x}_{j|k}-\bar{x}_j\right)^2 \;\geq\; 0.$$

This is the term that can be explained by the regression structure induced by the partition $(\mathcal{I}_k)_k$ of $\mathcal{I}$. The first term on the right-hand side of (3.3) is the *aggregate within-cluster dissimilarity* of classifier $\mathcal{C}$, defined by

$$W(\mathcal{C}) \;=\; \sum_{k=1}^{K}\sum_{j=1}^{J} w_j \sum_{b\in\mathcal{I}_k} \left(x_{b,j}-\bar{x}_{j|k}\right)^2. \qquad (3.5)$$

Thus, we immediately get the following simple corollary.

Corollary 3.3 *For any partition $(\mathcal{I}_k)_{k\in\mathcal{K}}$ of $\mathcal{I}$ we have the relationship*

$$D(\mathcal{I}) \;=\; W(\mathcal{C}) + B(\mathcal{C}) \;\geq\; W(\mathcal{C}).$$

This indicates that, in general, we try to find the partition $(\mathcal{I}_k)_{k\in\mathcal{K}}$ of $\mathcal{I}$ that gives a minimal aggregate within-cluster dissimilarity $W(\mathcal{C})$, because this partition $\mathcal{C} = \mathcal{C}_K$ (based on K parameters) maximally explains the observations. We consider the single terms of $W(\mathcal{C})$ and define the (individual) within-cluster dissimilarities by

$$D(\mathcal{I}_k) = \frac{1}{2I_k} \sum_{b,c\in\mathcal{I}_k} \sum_{j=1}^{J} w_j \left(x_{b,j}-x_{c,j}\right)^2 \;=\; \sum_{j=1}^{J} w_j \sum_{b\in\mathcal{I}_k} \left(x_{b,j}-\bar{x}_{j|k}\right)^2.$$

An easy consequence is

$$D(\mathcal{I}) \;=\; \sum_{j=1}^{J} w_j \sum_{b=1}^{I} (x_{b,j}-\bar{x}_j)^2 \;\geq\; W(\mathcal{C}) \;=\; \sum_{k=1}^{K} D(\mathcal{I}_k).$$

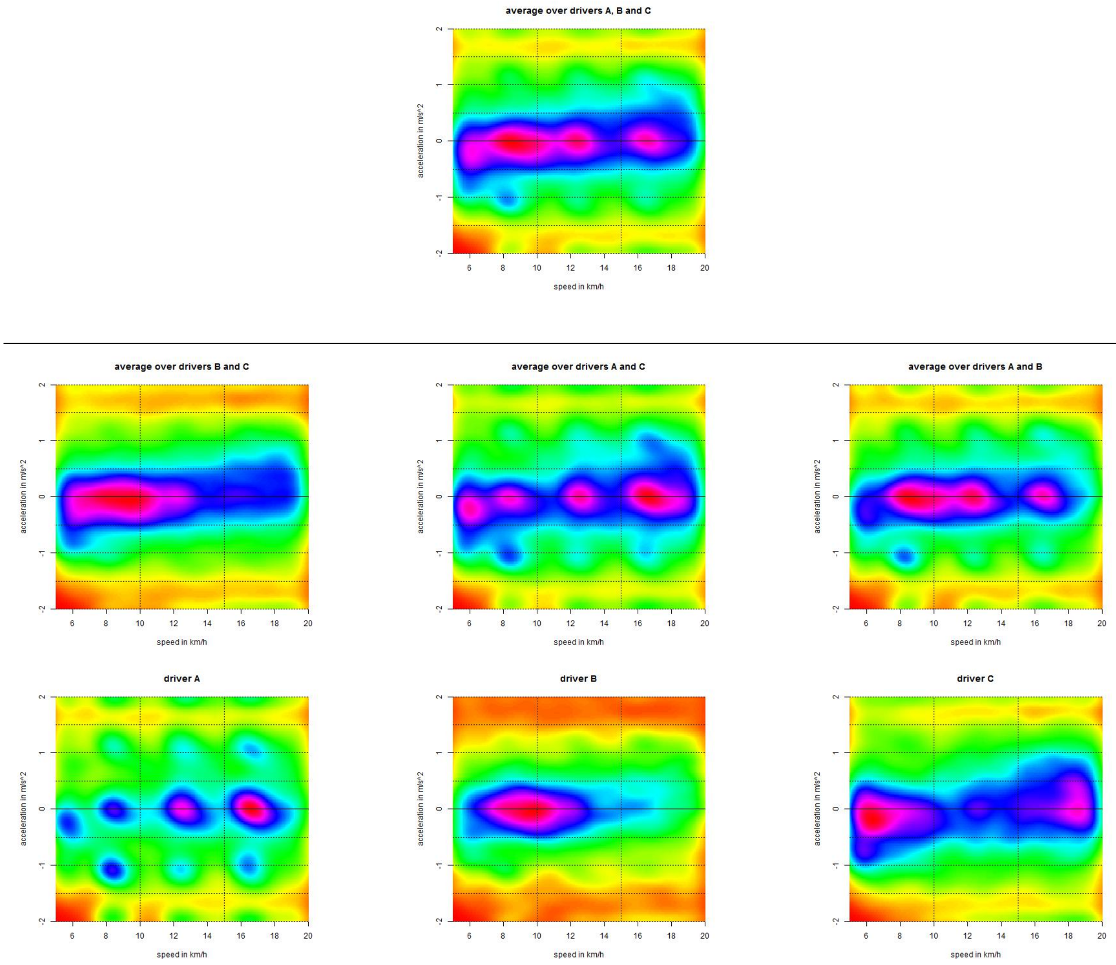

Figure 5: v-a heatmaps of car drivers A, B and C for speed bucket $(5, 20]$: empirical means $\bar{x}_j$ and $\bar{x}_{j|k}$ for $k = 1, 2$ with (1st row) no partition, i.e. on $\mathcal{I} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$; (2nd and 3rd rows) one partition with 1st column being $\mathcal{I}_1 = \{\mathrm{B}, \mathrm{C}\}$ and $\mathcal{I}_2 = \{\mathrm{A}\}$, 2nd column being $\mathcal{I}_1 = \{\mathrm{A}, \mathrm{C}\}$ and $\mathcal{I}_2 = \{\mathrm{B}\}$, and 3rd column being $\mathcal{I}_1 = \{\mathrm{A}, \mathrm{B}\}$ and $\mathcal{I}_2 = \{\mathrm{C}\}$.

Every partition of $\mathcal{I}$ reduces the aggregate within-cluster dissimilarity $W(\mathcal{C})$ induced by the chosen clusters $\mathcal{I}_k$, $k \in \mathcal{K}$. This explains the idea that we will pursue, namely, find the K-partition of $\mathcal{I}$ that leads to a minimal aggregate within-cluster dissimilarity $W(\mathcal{C})$ in (3.5).

We illustrate this for $K = 2$ on the three drivers A, B and C from Figure 4. Without loss of generality, we consider the speed bucket $(5, 20]$ which corresponds to the second row in Figure 4. On the first row of Figure 5 we show the empirical means $(\bar{x}_j)_j$ on the entire set $\mathcal{I} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. These empirical means are obtained on the partition of $R = (5, 20] \times [-2, 2]$ splitting this rectangle into equally 200^2 equally sized rectangles R_j. Thus, the first row of Figure 5 just shows the empirical mean of the three plots on the second row of Figure 4. The total dissimilarity on $\mathcal{I}$ is given by $D(\mathcal{I}) = 5.92 \cdot 10^{-6}$ (we choose weights $w_j \equiv 1$).

In a next step we consider the three possible partitions of $\mathcal{I}$ into two (non-trivial and disjoint) clusters $\mathcal{I}_1$ and $\mathcal{I}_2$. The empirical means $(\bar{x}_{j|1})_j$ and $(\bar{x}_{j|2})_j$ of these three possible partitions are shown in the three columns in the lower part of Figure 5, the second row corresponding to $(\bar{x}_{j|1})_j$ and the third row corresponding to $(\bar{x}_{j|2})_j$. The resulting aggregate within-cluster dissimilarities on these three partitions are given by

$$W(\mathcal{C}) = D(\mathcal{I}_1) + D(\mathcal{I}_2) = \begin{cases} 1.12 \cdot 10^{-6} & \text{for } \mathcal{I}_1 = \{\text{B}, \text{C}\} \text{ and } \mathcal{I}_2 = \{\text{A}\}, \\ 1.50 \cdot 10^{-6} & \text{for } \mathcal{I}_1 = \{\text{A}, \text{C}\} \text{ and } \mathcal{I}_2 = \{\text{B}\}, \\ 2.35 \cdot 10^{-6} & \text{for } \mathcal{I}_1 = \{\text{A}, \text{B}\} \text{ and } \mathcal{I}_2 = \{\text{C}\}. \end{cases}$$

Thus, if we aim to minimize the resulting aggregate within-cluster dissimilarity we choose partition $\mathcal{I}_1 = \{\text{B}, \text{C}\}$ and $\mathcal{I}_2 = \{\text{A}\}$ (which corresponds to the first column in the lower part of Figure 5).

In general, the main difficulty is to find the optimal (finite) partition into the clusters $(\mathcal{I}_k)_{k\in\mathcal{K}}$. In the previous example with $I = 3$ drivers we obtain three non-trivial partitions. In general, if we have $I \in \mathbb{N}$ drivers, there are $2^{I-1} - 1$ non-trivial binary splits. Unfortunately, it is not feasible to check all these binary splits for our $I = 1,753$ drivers and we need to find a different way to identify good partitions. One way to proceed is to apply the so-called K-means clustering algorithm, see Algorithm 10.1 in James et al. [3]. This is going to be described next.

3.3 K-means clustering algorithm

In this section we describe the K-means clustering algorithm, see Algorithm 10.1 in James et al. [3], which provides a classifier $\mathcal{C}$ on $\mathcal{I}$ for a fixed number K of categorical classes. This K-means clustering algorithm is very popular and it is probably the most used one for solving this kind of classification problem.

For a given classifier $\mathcal{C} : \mathcal{I} \to \mathcal{K}$ we consider the aggregate within-cluster dissimilarity given by, see (3.5),

$$W(\mathcal{C}) \;=\; \sum_{k=1}^{K} \sum_{j=1}^{J} w_j \sum_{b\in\mathcal{I}_k} \left(x_{b,j} - \bar{x}_{j|k}\right)^2,$$

with empirical means on $\mathcal{I}_k$ given by $\bar{x}_{j|k} = I_k^{-1} \sum_{b\in\mathcal{I}_k} x_{b,j}$.

The optimal classifier $\mathcal{C}^* : \mathcal{I} \to \mathcal{K}$ for given K is found by solving

$$\mathcal{C}^* \;=\; \underset{\mathcal{C}:\mathcal{I}\to\mathcal{K}}{\text{argmin}}\, W(\mathcal{C}) \;=\; \underset{\mathcal{C}:\mathcal{I}\to\mathcal{K}}{\text{argmin}} \; \min_{(m_{j|k})_{j,k}} \sum_{k=1}^{K} \sum_{j=1}^{J} w_j \sum_{\mathcal{C}(b)=k} \left(x_{b,j} - m_{j|k}\right)^2.$$

This optimal classifier $\mathcal{C}^*$ is approximated by alternating the two minimization steps.

K-Means Clustering Algorithm.

---

(0) Choose an initial classifier $\mathcal{C}^{(0)} : \mathcal{I} \to \mathcal{K}$ with corresponding empirical means $(\bar{x}^{(0)}_{j|k})_{j,k}$.

(1) Repeat for $\ell \geq 1$ until no changes are observed:

(a) given the present empirical means $(\bar{x}_{j|k}^{(\ell-1)})_{j,k}$ choose the classifier $\mathcal{C}^{(\ell)} : \mathcal{I} \to \mathcal{K}$ such that for each driver $b \in \mathcal{I}$ we have

$$\mathcal{C}^{(\ell)}(b) \;=\; \underset{k \in \mathcal{K}}{\operatorname{argmin}} \sum_{j=1}^{J} w_j \left(x_{b,j} - \bar{x}_{j|k}^{(\ell-1)} \right)^2 ;$$

(b) calculate the empirical means $(\bar{x}_{j|k}^{(\ell)})_{j,k}$ on the new partition induced by classifier $\mathcal{C}^{(\ell)}$.

Remarks 3.4

- The K-means clustering algorithm converges: Note that each iteration in (1) reduces the aggregate within-cluster dissimilarity, i.e., we have

$$W(\mathcal{C}^{(0)}) \;\geq\; \ldots \;\geq\; W(\mathcal{C}^{(\ell-1)}) \;\geq\; W(\mathcal{C}^{(\ell)}) \;\geq\; \ldots \;\geq\; 0.$$

  This provides convergence (in finite time because we have finitely many K-partitions). However, we may end up in a local minimum. Therefore, one may use different (random) initial classifiers $\mathcal{C}^{(0)}$ in step (0) of the algorithm to back-test the solution.

- Another issue is the choice of the constant K for the number of clusters considered. We may start with running the algorithm for $K = 2$ which leads to a binary partition $(\mathcal{I}_k)_k$, $k = 1, 2$, with empirical means $(\bar{x}_{j|k})_{j,k}$, $k = 1, 2$. For $K = 3$, we may then use these empirical means $(\bar{x}_{j|k})_{j,k=1,2}$ together with $(\bar{x}_j)_j$ as initial values for the K-means clustering algorithm with $K = 3$. This choice ensures that the resulting aggregate within-cluster dissimilarity is decreasing in K.

- For arbitrary driver $F \in \mathcal{P}(R)$ we calculate the corresponding probability weights $(x_j)_j$ on the partition $(R_j)_j$ of R, see (3.2). The classifier $\mathcal{C}$ extended to $\mathcal{P}(R)$ is defined by

$$\mathcal{C}(F) \;=\; \underset{k \in \mathcal{K}}{\operatorname{argmin}} \sum_{j=1}^{J} w_j \left(x_j - \bar{x}_{j|k} \right)^2 ,$$

  where $(\bar{x}_{j|k})_{j,k}$ are the empirical means obtained by the K-means clustering algorithm.

4 Example

We apply the K-means clustering algorithm to our $I = 1,753$ car drivers, for simplicity we only consider the speed bucket $(5, 20]$ which is subdivided into 200^2 rectangles R_j (this is analogous to the considerations in Section 3.2) and we set $w \equiv 1$.

The total dissimilarity on all drivers is $D(\mathcal{I}) = 0.002404$. The resulting empirical means $(\bar{x}_j)_j = (\bar{x}_{j|k})_{j,k}$, for $k = K = 1$, are plotted on the first row of Figure 6.

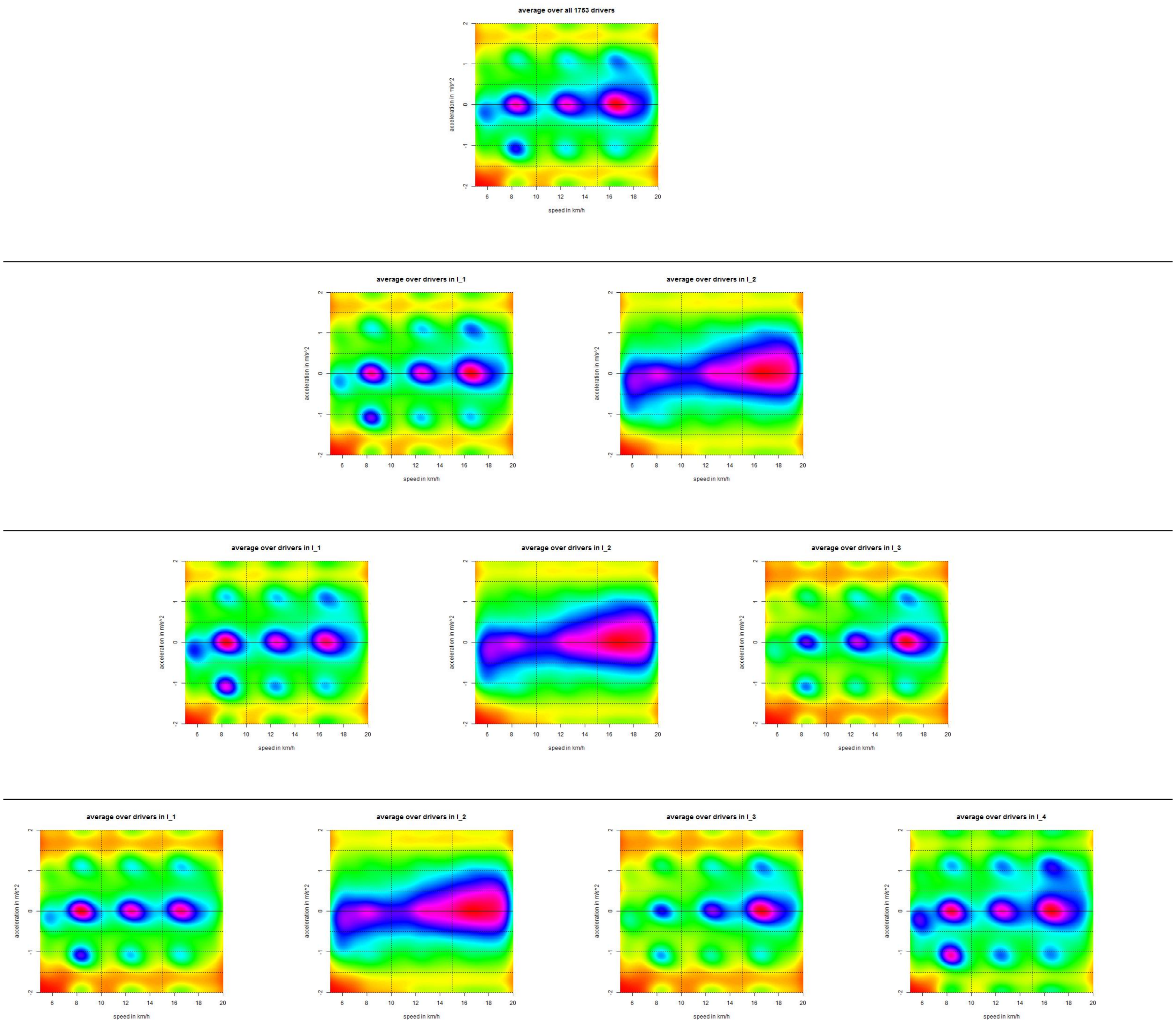

Figure 6: K-means clustering algorithm: resulting empirical means $(\bar{x}_{j|k})_{j,k}$. $k = 1, \ldots, K$ for the different constants $K = 1, 2, 3, 4$ corresponding to rows 1-4.

We now apply the K-means clustering algorithm. We use the R package `kmeans` which exactly performs this iterative optimization. For $K = 2$ we obtain the binary classifier $\mathcal{C} = \mathcal{C}_2$ with aggregate within-cluster dissimilarity

$$W(\mathcal{C}_2) = D(\mathcal{I}_1) + D(\mathcal{I}_2) = 0.001284 + 0.000635 = 0.001919 \; < \; 0.002404 = D(\mathcal{I}).$$

The resulting empirical means $(\bar{x}_{j|k})_{j,k}$, $k = 1, 2$, for $K = 2$ are plotted on the second row of Figure 6. We observe that this (first binary) partition mainly aims at separating manual gear cars from automatic ones.

Next we apply the algorithm to $K = 3$ with starting values as described in Remarks 3.4. The aggregate within-cluster dissimilarity obtained is $W(\mathcal{C}_3) = 0.001660$ for this classifier $\mathcal{C} = \mathcal{C}_3$. The resulting empirical means $(\bar{x}_{j|k})_{j,k}$, $k = 1, 2, 3$, for $K = 3$ are plotted on the third row of Figure 6. We observe that this additional partition is mainly used to split the manual gear cars because they have a bigger within-cluster dissimilarity than the automatic ones. Similarly the fourth row of Figure 6 shows the results for $K = 4$. We iterate this procedure for $K = 1, \ldots, 20$. This provides aggregate within-cluster dissimilarities $W(\mathcal{C}_K)$ for the different K's. In Figure 7 we plot the numerical results. We observe a steep drop of $W(\mathcal{C}_K)$ for small values of K and for K between 10 and 20 the negative slope starts to get smaller. This suggests that an appropriate choice of the number of disjoint classes should be in the range of 10 and 20. In Figure 8, below, we show the results for $K = 10$. The first column provides the empirical means $(\bar{x}_{j|k})_{j,k}$ for the classes $k = 1, \ldots, K$ with $K = 10$. The columns 2-5 show four different drivers $b \in \mathcal{I}_k$ that fall into the corresponding classes $\mathcal{I}_k$. From this plot it may seem that the granularity of the chosen classes is not sufficiently fine, yet. Moreover, it seems that the automatic gear cars should be considered individually for more granular classes. If we only consider automatic gear cars (corresponding to the car drivers on the right-hand side of the 2nd row in Figure 6), we obtain for $K = 10$ the categorical classes shown in Figure 9. We see that we receive a much finer distinction now, and we believe that in this particular example it is worth to consider manual and automatic gear drivers independently.

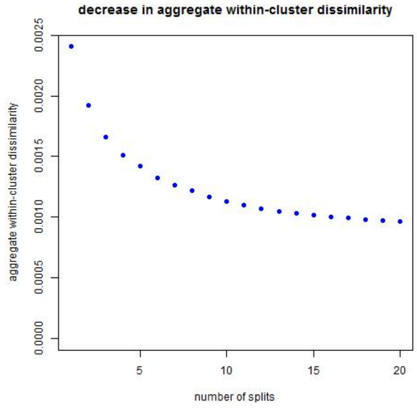

Figure 7: Aggregate within-cluster dissimilarities $W(\mathcal{C}_K)$ as a function of $K = 1, \ldots, 20$.

5 Conclusion and outlook

We have discussed how high-frequency GPS location data of different car drivers can be analyzed and classified. The classification was done by considering the dissimilarity between the corresponding v-a heatmaps of the individual car drivers, and the K-means clustering algorithm was used to allocate these different driving styles to the different categories. This allocation technique belongs to the unsupervised learning methods, which was employed because we do not have car accident information for this GPS data. That is, we have only analyzed the different driving styles, but we are not able to associate these driving styles with the riskiness of the underlying drivers.

There are two main directions that should be pursued. The first direction considers other classification methods. We can therefore consider other statistics, other methods and other algorithms. For instance, we could analyze v-a heatmaps with self-organized maps which aim at bending topological spaces to the observations.

The second direction should consider the effectiveness of the classification. Therefore, we need more complete data, unfortunately. For instance, we should consider the question how the classification is related related to other tariff criteria like age of driver, size of car, etc. This would involve, for instance, an analysis similar to the classification problem studied in Section 5.3 of Wüthrich–Buser [6]. Finally, and most importantly, these classes need to be related to corresponding claims and one should study to which extent telematics information improves the quality of prediction.

References

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth Statistics/Probability Series.

[2] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* 2nd edition. Springer Series in Statistics.

[3] James, G., Witten, D., Hastie, T., Tibshirani, R. (2015). *An Introduction to Statistical Learning. With Applications in R.* Corrected 6th printing. Springer Texts in Statistics.

[4] Verbelen, R., Antonio, K., Claeskens, G. (2016). Unraveling the predictive power of telematics data in car insurance pricing. *SSRN Manuscript* ID 2872112.

[5] 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.

[6] Wüthrich, M.V., Buser, C. (2016). Data analytics for non-life insurance pricing. *SSRN Manuscript* ID 2870308. Version November 15, 2016.

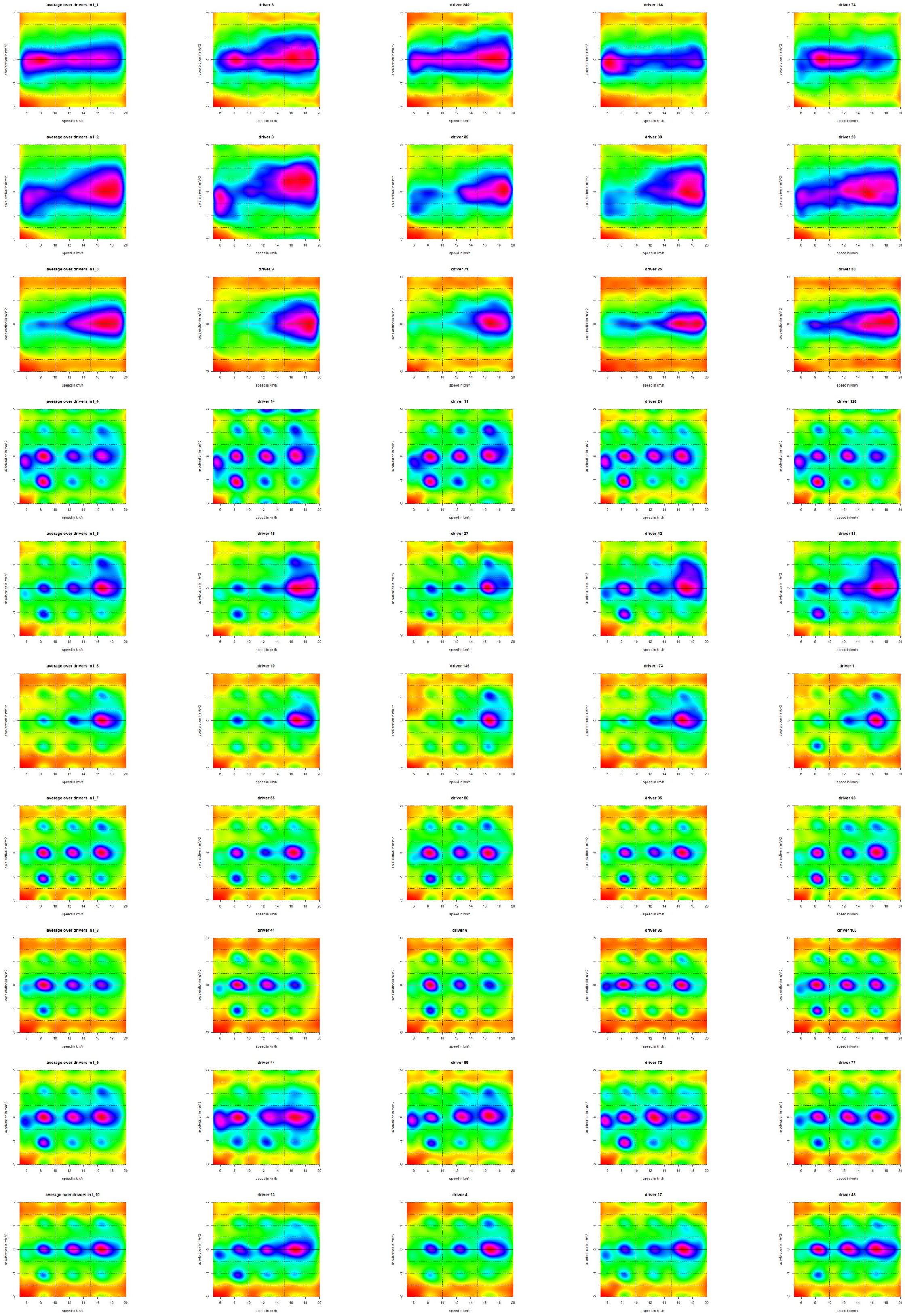

Figure 8: Partition results for $K = 10$: column 1 shows the empirical means $(\bar{x}_{j|k})_{j,k}$; columns 2-5 are four drivers $b \in \mathcal{I}_k$ that belong to the corresponding classes for $k \in \mathcal{K}$.

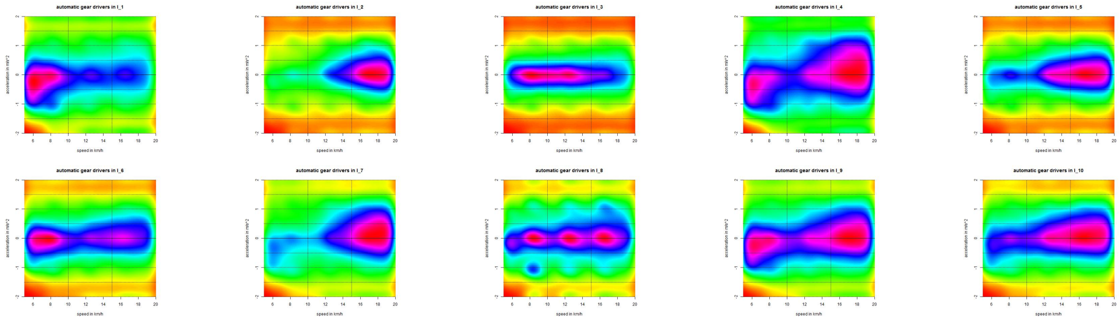

Figure 9: Empirical means $(\bar{x}_{j|k})_{j,k}$ for $K = 10$ if we only consider automatic gear drivers.