DISS. ETH NO. 26097

THREE ESSAYS ON THE ECONOMIC ASPECTS OF NEW INFORMATION TECHNOLOGIES

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

ANASTASIA SYCHEVA

M.Sc in Statistics ETH Zurich

born on 20.12.1989

citizen of

Russian Federation

accepted on the recommendation of

Prof. Dr. Wanda Mimra (IÉSEG School of Management), Examiner

Prof. Dr. Hans Gersbach (ETH Zurich), Co-Examiner

2019

Abstract

In the previous decade, we have observed a rapid growth in data volumes and in the processing power. New types of data and the algorithms to analyze them are known under the umbrella term “Big Data" technologies. They are adapted in various industries and are likely to become ubiquitous. This thesis comprises three empirical studies pertaining to Big Data technologies. Our results contribute to the analysis of their economic role and indicate promising avenues for future research in this area.

In the first two chapters of the thesis, we explore a particular use case of Big Data: *usage-based* insurance. Policyholders with such contracts install a drive recorder that continuously monitors the vehicle's motion. We have aggregated and combined these driving logs, generally referred to as “telematics data", with traditional contract and claim data. Our goal is twofold: a) to study accident risk factors b) to establish whether there are informational problems in the automobile insurance market. Telematics-based variables contain statistically significant predictors of subsequent accident involvement. This result prompt us to conclude that telematics data can be used to refine risk classification and inform the design of incentive contracts. Furthermore, the combined results with chosen insurance contracts indicate multiple and counteracting effects of private information.

In a second part of the thesis, we analyze the willingness to share personal data when this data is used for subsequent price discrimination. We have designed a laboratory experiment during which participants could sell a bundle of personal data. Participants were categorized based on the content of their personal data and received category-dependent payoffs in a subsequent stage. The experimental variations modified the category-dependent payoff structure. We find no effect of subsequent financial discrimination on the general willingness to sell personal data. A significant change in the data reservation price is only observed under strong negative discrimination. We observe important gender differences in the reservation price for private information and the role of underlying privacy concerns.

Zusammenfassung

In den letzten zehn Jahren haben wir ein schnelles Wachstum der Datenmengen und der Rechenleistung beobachtet. Neue Datentypen und die Algorithmen zu ihrer Analyse sind unter dem Oberbegriff "Big Data"-Technologien bekannt. Sie werden in verschiedenen Branchen eingesetzt und dürften allgegenwärtig werden. Diese Arbeit umfasst drei empirische Studien zu Big Data Technologien. Unsere Ergebnisse tragen zur Analyse ihrer wirtschaftlichen Rolle bei und zeigen vielversprechende Wege für die zukünftige Forschung in diesem Bereich auf.

In den ersten beiden Kapiteln der Dissertation untersuchten wir einen speziellen Anwendungsfall von Big Data: die nutzungsbasierte Versicherung. Versicherungsnehmer mit solchen Verträgen installieren einen Fahrtenschreiber, der die Bewegung des Fahrzeugs kontinuierlich überwacht. Wir aggregierten und kombinierten diese Fahrtenbücher, allgemein als "Telematikdaten" bezeichnet, mit traditionellen Vertrags- und Schadensdaten. Unser Ziel war es a) Unfallrisikofaktoren zu untersuchen b) festzustellen, ob es Informationsprobleme auf dem Kfz-Versicherungsmarkt gibt. Telematikbasierte Variablen enthielten statistisch signifikante Prädiktoren für die spätere Unfallbeteiligung. Aus diesem Ergebnis lässt sich schließen, dass Telematikdaten zur Verbesserung der Risikoklassifizierung und zur Information bei der Gestaltung von Anreiz-Verträgen verwendet werden können. Darüber hinaus zeigen die kombinierten Ergebnisse mit ausgewählten Versicherungsverträgen, dass die Auswirkungen von privaten Informationen mehrfach und entgegenwirkend sind.

In einem zweiten Teil der Arbeit analysierten wir die Bereitschaft, personenbezogene Daten weiterzugeben, wenn diese Daten zur späteren Preisdiskriminierung verwendet werden. Wir entwarfen ein Laborexperiment, bei dem die Teilnehmer ein Bündel persönlicher Daten verkaufen könnten. Die Teilnehmer wurden nach dem Inhalt ihrer persönlichen Daten kategorisiert und erhielten in einem weiteren Schritt kategorieabhängige Auszahlungen. Die experimentellen Variationen modifizierten die kategorieabhängige Ablaufstruktur. Wir finden keine Auswirkungen einer späteren finanziellen

Diskriminierung auf die allgemeine Bereitschaft, personenbezogene Daten zu verkaufen. Eine signifikante Veränderung des Datenreservierungspreises ist nur bei starker negativer Diskriminierung zu beobachten. Wir beobachten jedoch wichtige geschlechtsspezifische Unterschiede beim Reservierungspreis für private Informationen und die Rolle der zugrunde liegenden Datenschutzbedenken.

Acknowledgments

My first and foremost acknowledgments for my doctoral thesis are addressed to my advisor, Prof. Dr. Wanda Mimra, who has invested a lot of effort, time and understanding into supporting me in this endeavor. Our collaboration helped me to see the problems from a different perspective and at a different level of abstraction.

I am greatly indebted to Prof. Dr. Hans Gersbach for his advice, support, availability as a co-examiner and for a very warm welcome at his Chair. I would also like to thank Prof. Dr. Sebastian Rausch for immediately accepting to chair my PhD defense.

I have greatly benefited from discussions with my colleague Dr. Christian Waibel. His work ethics and attention to details were a great inspiration for me. I am also particularly indebted to Dr. Margrit Buser for her constant assistance, advice and careful proofreading. I have profited a lot from the work of the student research assistants, with special gratitude to Bernhardt Mélanie, Boudemagh Emina, Julia Burri and Daniel Kun for accelerating my progress and for their valuable feedback.

I had a unique opportunity to interact and collaborate with brilliant motivated scholars from whom I could learn a lot. I would like to thank my current and former colleagues from the CER-ETH for constructive team work and a very pleasant atmosphere. Our discussions inspired me, helped to expand my horizons and I am looking forward to working together in the future! I have a lot of fond memories of my office mates: Fabio Wieser and Felix Gottschalk - our conversations always helped to lighten my spirits.

I am using this opportunity to express my gratitude to my family and friends. My family, especially my mother Alexandra Sycheva and my grandmother Svetlana Shevyakova, were a constant source of support, guidance and motivation. I am grateful to Johannes for taking care of me and patiently listening during the last stages of the thesis. I would like to thank Valya and Dima for sharing small beautiful moments of everyday life. I am truly indebted to Natasha, Anja, Beatrice, Sergio, Alina, Nastya and Sandra for being there for me in my darkest moments. I am also grateful to Sandro for helping me to clarify my priorities.

Contents

List of Figures

List of Tables

1. Introduction

> *There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.*
>
> *Eric Schmid*
> *Executive Chairman of Google*

Data creation rates in digital societies are steadily rising. Henke et al. (2016) project a 2x growth in the volume of data approximately every 3 years and note a 40x increase in the processing power between the fastest supercomputer of 2010 and 2016. Collected information pertains to different aspects of everyday life and individuals' activities. Companies have recognized the potential of new data sources to tackle existing challenges prompting them to adopt these innovations. A popular term, "Big Data Revolution" reflects the scale of the transition taking place in many industries and in society at large.

Big Data technologies can change the functioning of markets with information problems by reducing, or even reversing, information asymmetries between the different market participants. Information asymmetries lead to two major problems: *adverse selection* and *moral hazard*. *Adverse selection* arises when one of the parties has an informational advantage over the other prior to the transaction. A common example is the tendency of high-risk individuals to purchase a more comprehensive insurance protection. To better classify applicants' risk types, banks have complemented credit scoring by mining credit application texts, and insurance companies have offered life insurance policies predicated upon DNA test results.

Moral hazard refers to the situation when a counterparty engages in a riskier behavior knowing that somebody else (the insurer) will face the negative consequences of his actions. Collecting and contracting upon more verifiable information about an individual's behavior decreases the room for *moral hazard*. E.g. policyholders can be incentivized to lead a healthy lifestyle if the premiums for their health insurance policy depend on the activity records from a Fitbit device.

Since a long time data has played the central role in insurance pricing. Already in 1693 Edmond Halley created life tables based on demographic data and proposed to use them for life contingency calculation. Currently, predictability of the subsequent claims is one of the key characteristics of insurable risks. Predictions are obtained from statistical analysis of historical losses combined with various observable and verifiable covariates. To refine the risk models, insurance providers used numerous characteristics such as age, marital status, occupation and address. However , traditional sources of risk classification variables oftentimes could only yield imprecise proxies of underlying risk factors.

A more precise risk classification due to information gained using Big Data technologies can result in more tailored premiums and facilitate insurance portfolio optimization (Litman (1997)). The usage of Big Data technologies is also likely to affect market interaction and the efficiency of equilibria in the insurance industry. However, a lot more research is needed to understand how Big Data technologies can and should be used to improve market outcomes and societal welfare. In the first part of my thesis I examined a prominent use case of Big Data in the insurance industry: *usage-based* vehicle insurance. We analyzed a telematics dataset provided by a large Swiss insurer carrier to investigate the following:

Q1 Does telematics data provide evidence of *Asymmetric Information* in the automobile insurance market?

Q2 What further insights into accident risk factors can we extract from this data?

Q3 How can we use this information to reduce accident risk?

In Chapter 2, me (first author) and my co-authors Wanda Mimra and Christian Waibel created detailed driving profiles based on telematics data for young drivers. These profiles reflected where, when, how often and how long the policyholder drived. We augmented the driving logs with road speed limits obtained from an *Open Street Map* (OSM) database to evaluate the frequency and severity of speed violations. Furthermore, the device monitored information about the vehicle's acceleration and appended driving logs every time the values exceed a pre-specified threshold. The frequency and severity of these “elevated g-force events" reflect the policyholder's driving style.

We combined the telematics-based predictors with traditional risk classification variables to model *ex post* risk realization on the one hand and contract choice on the other. We found that the number of journeys per day is positively related to both risk and coverage, which would be consistent with the presence of asymmetric information given the existing risk classification (**Q1**). Further results suggested that there might be multiple dimensions of private information interacting. However, the interpretation of results has to be done with extreme caution since the sample consists of drivers that are younger than 26 years of age and thus are relatively inexperienced. In particular, the data did not allow to analyze how well the policyholders actually understand their risk.

Using the information from driving logs, we discovered the following accident risk factors: *average daily distance driven*, *average no. of journeys per day* and *average speeding* (**Q2**). Neither driving style, reflected by the frequency of elevated g-force events, nor percentage of urban driving are statistically significant predictors of accident involvement. Taken at face value, this suggests that pricing *usage-based* insurance policies based on the estimated quality of driving might be not that efficient for reducing accident risk. The absence of the link between driving style and accident involvement could also be attributed to *selection bias.* Policyholders in our sample were punished for frequent accelerations and braking, especially in an urban area or during the night. Under these circumstances it is plausible that (a) policyholders with reckless driving style did not purchase the policy (b) individuals holding such a policy exerted more effort and adjusted their driving behavior.

We stratified the sample by age to study whether main driving risk factors change in the course of time. Young and inexperienced drivers constitute a high-risk pool. Tailoring the risk classification scheme to account for age-specific risk factors could introduce stronger incentives for safe driving on the one hand and reduce the premium burden on lower-risk, older drivers on the other. The *average no. of journeys per day* and *average speeding* are important risk factors for both subsamples. The *average daily distance driven* is a significant risk factor for the older but not for the younger drivers. These results do not warrant the introduction of different telematics-based risk pricing schemes on these age strata.

In Chapter 3, I deepened the risk analysis of Chapter 2. Results in Chapter 2 in particular suggested that the *no. of journeys* is an important accident risk factor. To shed more light on the phenomenon, I aggregated the driving logs at the policyholder-journey level and augmented the driving profiles created in Chapter 2 with new predictors.

First, I studied the impact of route familiarity. Intuitively, driving on unknown or rarely visited roads could be associated with a higher hazard level. The scientific basis for this conjecture is provided by Posner et al. (2004): drivers on familiar roads perform this task automatically, therefore they have more free mental resources to react to unexpected hazards. In contrast, Burdett et al. (2019) and Yanko and Spalek (2013) argue that route familiarity breeds inattention and mind wandering with concomitant increase in accident risk. I found that driving on unfamiliar roads increases accident hazard, yielding empirical support to the first statement. This result can be incorporated in the risk classification scheme, for instance by putting higher weight on speed violations on unfamiliar roads **(Q3)**.

My further results indicate that neither driving under bad light conditions nor higher frequency of long journeys are associated with higher accident risk. There are two competing explanations: 1) high quality infrastructure offsets the risks stemming from impaired visibility and fatigue 2) the *usage-based* insurance policy succeeds in limiting companies' exposure to these risks. The data at our disposal was not sufficient to disentangle these causes.

The results in Chapter 2 and Chapter 3 show that these new sources of information can be used to refine risk classification and inform the design of incentive contracts. The ability of insurance providers to better tailor premiums to risk types and track and incentivize certain forms of behavior however gives rise to contentious debates. On the positive side, risk-based pricing can incentivize risk prevention, promote a healthier lifestyle and create desirable spill-over effects such as a decrease in traffic congestion and CO2 emissions. On the negative side, the premiums for high risk individuals could become unaffordable. Whether it is legal or acceptable that they bear the risk on their own is an ethical and social issue. From an ethical standpoint, the main distinction is made between controllable and uncontrollable risks. Vehicle accident risk falls in the former category, making risk-based pricing less controversial. The main discussion is centered on potential privacy infringements. Using DNA test results to adjust risk premiums and punishing individuals for factors outside their control is an entirely different matter.

Chapter 2 and Chapter 3 present evidence that new behavioral data has a certain economic value. A major factor preventing companies from collecting even more information are the individuals' privacy concerns. Their causes are the subject of various disciplines from psychology, sociology and anthropology to biology. Apart from emotional discomfort, with wider adoption of Big Data Technologies, sharing personal information might have long-lasting economic consequences. Consumer discrimination based on personal data is already prevalent: the Google search history determines what advertisements are displayed to the users, and digital footprints are progressively incorporated into pricing schemes.

In Chapter 4, which is joint work with Irina Gemmo and Wanda Mimra, we conducted a laboratory experiment to shed more light on privacy-related decision making. In particular, we analyzed participants' privacy concerns as well as their response to price discrimination based on the content of their personal data. The personal data that the participants could sell in the experiment consisted of the bundle of their height, weight, bank account balance information as well as a photo of their face. To implement price discrimination in the lab, participants were then categorized based on whether they had sold their data to the experimenters, and the content of their data, and received category-dependent payoffs.We added an extensive post-experimental survey to capture other privacy relevant information. Our setting allowed us to address the following questions:

Q4 Does data-based price discrimination affect the willingness to sell personal information?

Q5 What other factors influence the decision to disclose the data and the subsequent reservation price?

The analysis yielded several interesting insights. The general willingness to sell personal data is not significantly affected by data-based price discrimination. A marginally significant change in the reservation price is only observed when one data-based category implies a strong decrease in the subsequent payoff **(Q4)**. The general willingness to sell personal data is strongly affected by individuals' privacy concerns on the one hand and trust related to the context of the experiment on the other. Gender differences are also prominent: under all treatments in all data dependent categories women were less likely to agree to sell their data and when they did, stated a higher reservation price. Our model suggests however that the gender differences manifest themselves through the impact of general privacy concerns and trust towards the experimenter. The decision whether to sell the data is also affected by perceived control overs subsequent data usage and its sensitivity **(Q5)**.

Our results suggest three different approaches to increase the likelihood that individuals agree to share their data. First, to promote trust data subjects should be given more control over their data, and the subsequent usage should be transparent. Furthermore, Kehr et al. (2015a) have discovered that better design of a user interface promotes data sharing. Lastly, here is room for a compromise between extracting sensitive data and obtaining accurate insights out of it. As an example, insurance companies are interested in some aggregate statistics on driving behavior, whereas policyholders are primarily concerned about disclosing the exact coordinates and timestamps of visited locations. A possible middle ground could be to compute relevant statistics locally on the on-board device and share only that information.

2. Bigger Data and Risk Classification in Auto Insurance: Evidence from Telematics Contracts

We combine a large telematics based dataset on driving behavior with traditional contract and claims data in auto insurance to analyze accident risk factors and review the question of asymmetric information in auto insurance. We find that the telematics data can improve risk modelling and identify significant risk factors pertaining to both risk exposure and driving style. The combined results with chosen insurance contracts suggest that multiple dimensions of private information might be interacting. As the sample consists of only young drivers, conclusions regarding risk driven selection however cannot be drawn.

2.1 Introduction

Advances in GPS technologies allow new forms of auto insurance contracts. In particular, there now exist contracts for which insurance companies equip vehicles with a drive recorder, collecting detailed information about the vehicles motion. This type of datasets, referred to as telematics data, is a rich source of reliable information about driving patterns and style and could be used to alleviate information problems between insurers and policyholders in auto insurance. First, telematics data can further the detailed understanding of accident risk factors (Ayuso et al. (2016), Ayuso et al. (2014)) and could thereby refine risk classification schemes. Second, telematics data can be used in insurance contract design to incentivize the optimal accident prevention effort of drivers. On a societal level, aggregated telematics data can inform public policies aimed at the reduction of annual mileage, fuel consumption and CO2 emissions (Vickrey (1968), Litman (1997), Edlin (1999), Bordoff and Noel (2008)) or internalization of congestion externalities (Vickrey (1968)).

In this paper, we take advantage of a telematics dataset provided by a major swiss insurer to analyze accident risk factors as well as test for the presence of asymmetric information between policyholders and the insurer. Compared to earlier contributions regarding asymmetric information in auto insurance (e.g. Chiappori and Salanie (2000) and Cohen (2005)) the telematics data thereby provides much richer information about accident-relevant driving patterns and style. However, the telematics contract was only offered to drivers in Switzerland younger than 26 years of age, i.e. our sample consists of mostly inexperienced drivers that might be learning about risk themselves.

In the analysis, we first construct several measures of both risk exposure and driving style from the telematics raw data. The created driving indices e.g. measure the average distance driven per day, and the average speed difference after acceleration. We have furthermore extracted speed limits from the Open Street Map Database to combine them with the telematics data in order to have precise, street level measures of speeding behavior. Compared to other studies using telematics data (e.g. Kremslehner and Muermann (2016)), we can furthermore use information on so-called elevated g-force events about the vehile's accelerations and breaking.

Our main results pertain to the analysis of accident risk factors. We find that the information provided from telematics data is valuable to model risk, as measured by liability claim occurence. In particular, we find that both average distance per day as well as the number of journeys per day have a significant positive effect on risk. Both variables pertain to risk exposure, as opposed to driving style. With average speeding, we also observe a significant risk factor reflecting driving style.

Combining the results on risk factors with those of contract choice, we observe that a higher number of journeys per day is associated with a significantly higher propensity to purchase collision cover and BM protection as well as a significantly positive effect on liability claim submission. The number of journeys per day is thus positively related to both risk and coverage, which would be consistent with the presence of asymmetric information given the existing risk classification and contracting. We find the opposite result for average distance per day, which is significantly positively related to risk but negatively to the purchase of insurance coverage. One interpretation is that multiple dimensions of private information are interacting, as the latter result could be explained by overconfidence about driving skill for higher average distances.

However, the interpretation of results has to be done with extreme caution since the sample consists of drivers that are younger than 26 years of age and thus are relatively inexperienced. In particular, the data does not allow to analyze how well the policyholders actually understand their risk. Thus, while the telematics data is important to identify risk factors, the results for our dataset with young drivers do not allow the interpretation that the telematics data helps to identify relevant asymmetric information in the form of the policyholder having superior information about risk type or preferences that affects selection in a systematic way for insurer profits.

Related literature

Our paper is related to two strands of literature: The literature on motor vehicle accident risk factors and the literature on asymmetric information in insurance markets.

Lemaire (2013) reviews standard risk classification variables used in different countries. Notwithstanding certain regional variations due in part to market regulations, there seems to be a consensus in the insurance industry regarding important risk factors. The list includes: vehicle characteristics, such as model, age, engine power, weight, and policyholders characteristics, for instance age, driving experience, occupation, number of active drivers and geographic area. In our analysis we observe an additional risk factor, a recent change of address.

Ayuso et al. (2014) use telematics data of spanish young drivers below the age of 30. Controlling for driving experience, driving patterns that increase the risk of an accident are the average number of kilometers driven per day, nighttime and urban driving as well as speed violations expressed as percentage of km driven above the mandatory speed limits. Furthermore, Ayuso et al. (2014) find that gender affects the hazard level. Male novice drivers have, on average, their first crash sooner than women. In addition, the authors find that driving at night reduces the time to the first accident for women, but it has no significant effect for men. Experienced drivers are likely to drive more kilometers until they are involved in their first accident, but excessive speeds reduce the time to the first accident more markedly for men than they do for women. Ayuso et al. (2010) find that exceeding the speed limit is the costliest traffic violation: it increases the expected loss of an accident by two-thirds, compared to accidents that do not involve any traffic violations. This result provides empirical evidence against the widespread hypothesis of independence between frequency and severity of the claim process. This paper complements Ayuso et al. (2016) and Ayuso et al. (2014) and differs in several aspects. First, the statistical approaches utilized are markedly different. Second, the created set of predictors is much richer.

Cohen and Siegelman (2010) present a survey of empirical studies of asymmetric information in insurance markets. With respect to auto insurance markets, Chiappori and Salanie (2000), Dionne et al. (2001) and Cohen (2005) do not find a statistically significant correlation between the level of insurance coverage and ex post realizations of risk given the existing risk classification of insurers. However, the positive correlation test may fail to detect asymmetric information if multiple dimensions of private information are interacting. De Meza and Webb (2001) show that if consumers differ not only in their risk characteristics but also in risk aversion, market equilibrium could feature zero or negative risk-cover correlation. The empirical evidence that both these factors affect insurance purchase decisions is presented in Cohen and Einav (2007). The authors use the data from an Israeli insurance company to fit a structural model that captures hidden heterogeneity in risk and risk preferences. They conclude that risk and risk aversion seem to be positively correlated and that unobservable heterogeneity in the latter is larger than unobserved heterogeneity in the former. This shows that the *positive correlation test* is not a universal approach to test for the presence of asymmetric information. In our dataset, the *positive correlation test* does not detect private information.

Finkelstein and Poterba (2014) introduce the *unused observables test* that can be used even with multidimensional private information. The logic is that if there exist a variable ("unused observable") that on the one hand is not used by the insurer for premium calculation (1) and on the other is a significant predictor for the choice of cover and ex-post risk realization (2), this indicates the presence of asymmetric information in the market. In the subsequent analysis, a rich telematics dataset allows us to construct a number of predictors that may satisfy both criteria.

The closest work to the present paper is Kremslehner and Muermann (2016). The authors take advantage of telematics data to test for the presence of asymmetric information and combine the *positive correlation test* with a slightly extended *unused observables* test. Kremslehner and Muermann (2016) find that the correlation between residual terms is not significantly different from zero, however the total number of car rides increases both the probability of purchasing more insurance cover and probability of ex post risk realization. Compared to Kremslehner and Muermann (2016), our analysis is based on a richer dataset covering both a larger sample of policyholders for a longer time period and a richer set of telematics-based created driving pattern variables. Similar to Kremslehner and Muermann (2016), we find that the number of journeys per day is positively related to risk and amount of insurance protection. A more detailed discussion of the comparison of the findings is relegated to Section 3.6.

The remainder of the chapter is organized as follows: In Section 2.2 the datasets and empirical approach are introduced, Section 3.6 presents and discusses the results, and Section 2.4 concludes.

2.2 Data and Empirical Approach

The data for the analysis was provided by a major Swiss insurer. The insurer offers a usage based automobile insurance policy under which the policyholder's driving behavior affects the premium paid. The relevant information on driving behavior is collected by a telematics device that records various aspects of a vehicle's motion. The telematics based policy is offered only to drivers younger than 26 years old in exchange for an initial premium discount. Our data analysis will therefore be based on drivers younger than 26 years of age. For the analysis, we combine the telematics dataset with traditional contract data and claims data.

In Switzerland, as in most countries, third party liability insurance is mandatory for all vehicle and motorcycle owners. Additionally, policyholders may purchase first party collision coverage.[1] This collision coverage is mandatory for drivers of leased vehicles, which make up 17.27% of contracts in the dataset.

The premium for insurance policies using telematics data has three components: First, the base premium, which is determined by traditional risk classification variables such as the policyholder's age and vehicle characteristics. The base premium is then adjusted according to the driver's Bonus Malus class as well as a telematics data based adjustment factor. Premium rates are recalculated annually and apply between January 1st and December 31st. The Telematics based factor is determined and the Bonus Malus class is adjusted based on driving behavior and claim submission before October 1st of the previous year.

A policyholder's Bonus Malus class is determined by his claim history with the insurer. There are 18 different Bonus Malus classes, each Bonus Malus class corresponds to a different scaling factor of the relevant base premium, ranging from 30% to 150 %. New policyholders are assigned to the 14th class with the scaling factor 100%. A year without claims leads to a 1 level decrease of the Bonus Malus class, every relevant claim however results in 4 level increase. Bonus Malus scores for collision and liability claims are calculated separately, which implies that a liability claim submission won't affect the Bonus Malus score for collision claims and vice versa. Most contracts allow a policyholder to purchase bonus protection for liability and collision claims: Under this option, the first relevant claim does not lead to a Bonus Malus reclassification.[2] Besides the purchase of collision cover, we will consider the purchase of protection against Bonus Malus reclassification as a second insurance choice in our analysis.

All telematics contracts in our dataset include an initial premium discount compared to non-telematics based insurance policies. The subsequent telematics data based adjustment factor is determined by how the policyholder drives, measured by the oc-

[1] Collision coverage pays indemnity primarily in case when the policyholder causes an accident. If the policyholder is not at fault his losses might also be covered if the insurance company of the at fault driver does not compensate in time.

[2] For policyholders that purchase collision cover, bonus protection for both types of claims can only be purchased simultaneously.

currence and the number of certain events such as acceleration, harsh braking or cornering.[3] These events will be summarily referred to as elevated g-force events.[4] The corresponding driving patterns of the policyholders are summarized into driving scores. These scores are used to determine a policyholder's rank compared to other drivers, which in turn determines the premium discount for the subsequent insurance year. Even policyholders with the lowest rank still receive a discount compared to their counterparts without a telematics usage based contract.

2.2.1 Contract Data

Each policyholder in the dataset is uniquely identified by a policy number. The policy number is used to link information from the three datasets: the contract data, the claims data and the telematics data. The contract data we have access to covers the period between 01.01.2014 and 03.11.2017 and consists of a total of 27'998 contract entries for 9'244 policyholders.[5]

For the analysis, we will concentrate on data from the year 2016 for 5690 policyholders, since this is the period and policyholders for which we have both enough telematics observations as well as full claims data. A more detailed discussion of this can be found in *Section 2.2.4*.

The available information can be grouped into the following categories: a policyholder's personal information (age, gender, nationality, address,...), properties of the insured vehicle (e.g. age, price and horsepower), details about contract features and the premia paid. Table 2.1 displays the summary statistics of policyholders' and vehicles' characteristics.

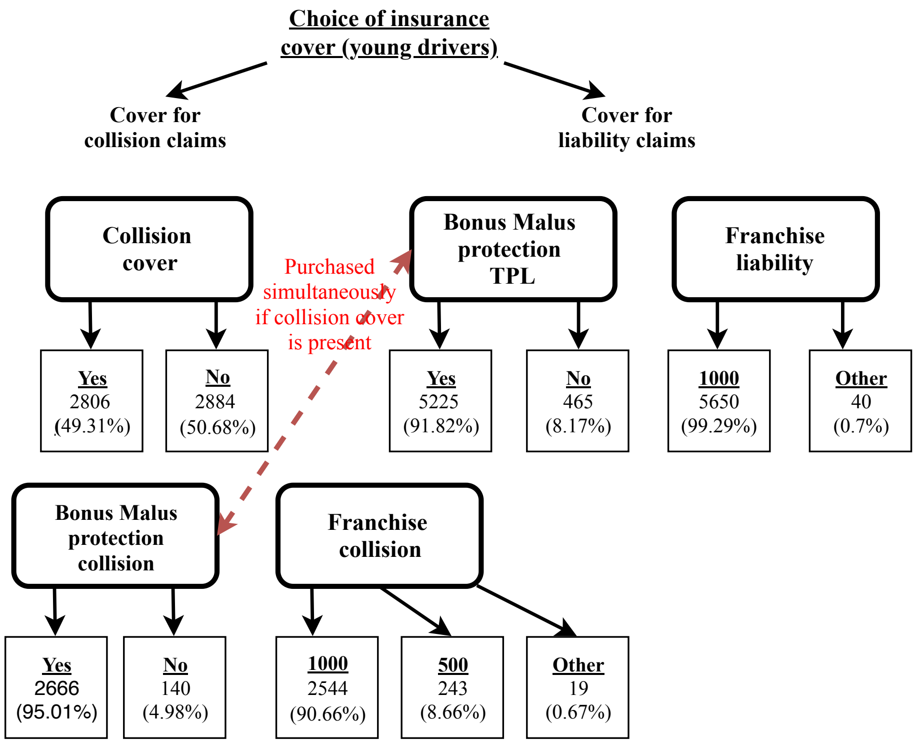
Policyholders have various options with respect to insurance coverage. The coverage options and insurance choices for the 5690 policyholders analyzed with 2016 contracts

[3] Another approach in pricing based on telematics data are pay-as-you-drive-policies in which the price is primarily determined by the distance driven. Some modifications of this approach also account for *where* (e.g. urban, motorway) and *when* (e.g. night, rush hour) the policyholders drives.

[4] This requires data from additional sensors, such as accelerometer.

[5] A change in e.g. the policyholder's address or a contractual feature results in a new contract entry. Furthermore, since premium rates for collision and liability cover are recalculated annually, policyholders are expected to have a separate entry at least once per year. Policyholders have on average 3.03 and up to 14 corresponding entries.

Figure 2.1: Available options and contract choices insurance cover against collision and liability claims (2016)



are summarized in Figure 2.1. Different deductible levels are offered for (mandatory) liability coverage. However, a look into the data reveals that over 99% of contracts in our dataset feature the default deductible of 1000 CHF. Therefore, the extent of liability coverage is not an informative variable regarding insurance choice. Collision cover is optional and is purchased in 49.31% of contracts. Various levels of deductibles for collision claims are available but, similar to liability coverage, over 90% of contract with collision coverage have the default deductible (1000 CHF).

As described above, premiums paid for collision and liability protection are subject to experience rating. Interestingly, a coverage against reclassification risk is available: A Bonus Malus reclassification following a first accident resulting in a claim can be avoided if the policyholder purchases so-called ‘bonus protection’. If the policyholder chooses collision coverage, bonus protection can only be bought for both collision and liability claims jointly. 8.17% of contracts for liability claims and 4.98% of contracts with collision cover do not have bonus protection. In the following analysis, we will

Table 2.1: Summary statistics: Policyholder's and vehicle's characteristics.

	mean	median	std	Sample quantiles					
				min	5%	25%	75%	95%	97.5%
Policyholder's characteristics									
Age of driver (years)	21.78	22	2.30	17	18	20	24	25.00	26.00
Bonus Malus score for liability claims	13.41	13	1.17	12	12	13	14	15.00	17.00
No. of years without first contract	2.84	3	2.29	0	0	1	5	7.00	7.00
No. of previous mobility claims	0.11	0	0.43	0	0	0	0	1.00	1.00
Vehicle's characteristics									
Price (CHF)	31351.02	28400	17142.01	0	0	22000	39600	60900.00	70300.00
Age (years)	7.01	6	5.59	0	0	2	11	17.00	19.00
Horsepower	130.63	115	56.63	45	68	90	150	240.55	291.55
Weight (kg)	1278.85	1250	227.59	785	940	1115	1410	1665.55	1787.00
Mileage	80.55	73	68.39	1	1	21	123	200.00	220.00

use collision coverage as the primary variable of insurance choice and bonus protection as an additional variable of the level of insurance.

2.2.2 Claims Data

Among 5690 policyholders we analyze, 429 have submitted at least 1 liability claim in 2016. [6] Figure 2.2 shows the histogram of liability claim sizes. One can observe that in 13.5% of cases the compensation did not exceed 10 CHF. We have decided not to include them in our main analysis.[7] Thus, there are in total 371 liability claims in 2016 for policyholders that actively use the drive recorder which are used in the analysis.

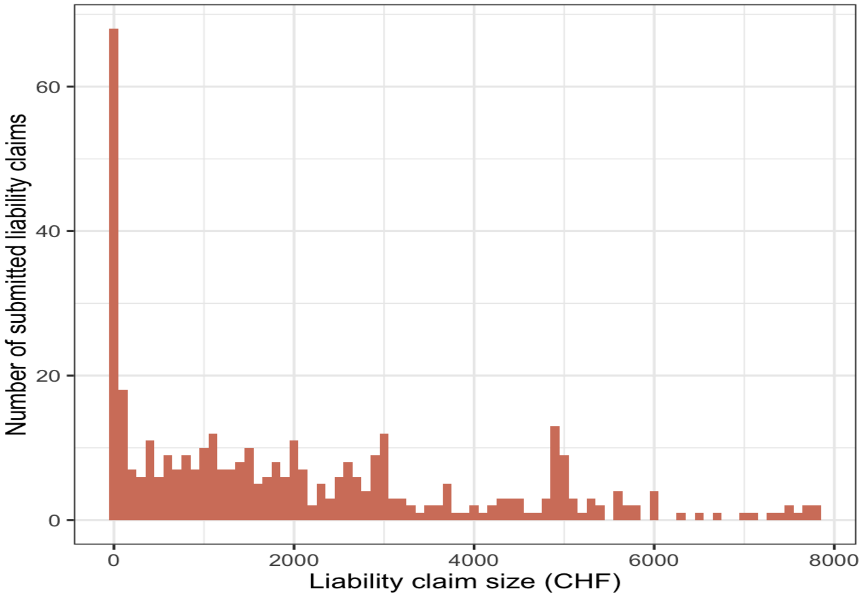
The summary statistics for liability claims grouped by insurance coverage are reported in Table 2.2. In Table 2.2, we can observe that the liability claim rates are lower for policyholders with collision cover compared to policyholders without collision cover.

As noted in Section 2.2, liability claims are experience rated. Information about an individual's claim history is summarized by the Bonus Malus score. The Bonus Malus

[6] In 353 cases, driving logs corresponding to stated accident date were available. Absence of driving logs for accident date can be explained as follows: (1) reported accident date slightly deviates from actual (2) policyholder decided to switch to a *usage based* policy following the accident involvement

[7] Robustness of our results with respect to the claim size threshold is explored in Section 2.3.3.

Figure 2.2: Histogram of the size of liability claims submitted by policyholders with active drive recorders in 2016



†33 claims that exceeded 8000 CHF and are not depicted

scores of policyholders were not provided in our dataset by the insurer, however we estimate them with the information available about the claim history in the claims data applying the Bonus Malus rules described in Section 2.2.[8] The validity of the resulting variable Bonus Malus class is based on the assumption that the dataset contains the complete claim history of policyholders from our subject pool. This assumption might be violated due to several reasons. First, due to the experience rating, individuals have an incentive to under-report small claims. Second, the dataset does not contain information on accidents that could have taken place before 2010. Since our analysis is based on data for drivers younger than 26 years of age, this however does not seem to be an important consideration. Third, neither claim history nor Bonus Malus score of policyholders previously insured by another insurer are available.

[8] For every policyholder, for every insurance year with active policy the number of relevant claims are counted. If bonus protection for a corresponding year is present, the first claim is discarded. The number of remaining claims is multiplied by 5 (4 levels up + 1 to account for the fact that there is no annual decrease in the score) added to 14 and the number of insurance years with the company is extracted from the sum. The result is substituted by 18 (highest bm score) if it exceeds this number. An additional input parameter of our algorithm is the *claim threshold* that allows us to filter out claims not exceeding certain thresholds.

Table 2.2: Summary statistics: Liability claims grouped by the amount of insurance cover purchased.

	Total			Collision cover			
		Bonus protection TPL		No	Yes	Bonus protection collision	
		No	Yes			No	Yes
Total no. of liability claims	371	32	339	216	155	7	148
Annual liability claim rate (%)	6.4	6.7	6.3	7.3	5.4	4.9	5.4
Average liability claim size (CHF)	3852	3151	3919	3497	4347	4513	4339

In the claims data, we also observe mobility claims. Mobility insurance covers transportation and towing fees of the vehicle after a breakdown. Frequently, following a collision, policyholders without collision insurance protection can cover some losses using the mobility policy. The number of mobility claims may in general arise from several factors: collision, technical problems of the vehicle or possibly reckless behavior of the driver. Overall, information about the number of previous mobility claims may be valuable for predicting subsequent accident occurrence.

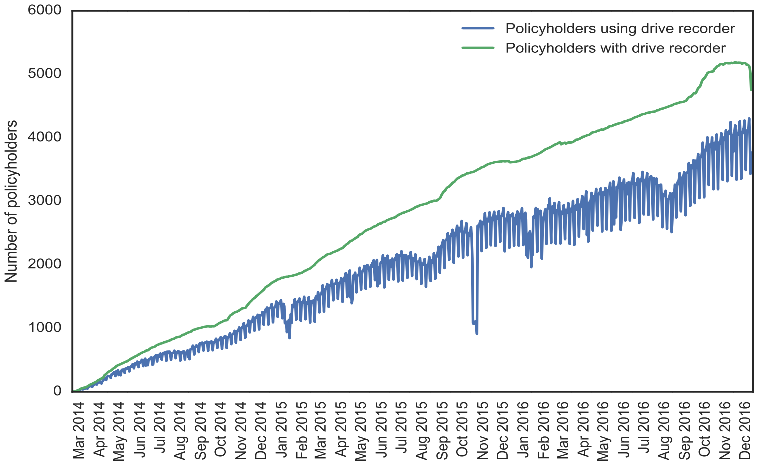
2.2.3 Telematics Data

Description of the Data

Policyholders with the telematics *usage based* insurance policy are expected to install a drive recorder in their vehicle. The device is developed by an independent telematics company and collects detailed information about the vehicle's motion. Under normal conditions, it stores data on the vehicle's location, time of the signal, distance traveled, average speed, trip identification number, engine status and quality of the GPS signal approximately every two kilometers. The device also constantly monitors information from the vehicle's accelerometers to detect elevated g-force events such as harsh brakings, accelerations and cornerings. If certain pre-defined thresholds are exceeded, information about the vehicle's motion and accelerations along three principal axes is recorded.

The telematics dataset contains approximately 390 million data points corresponding to about 2.1 million days of observations for 6'773 policyholders, spanning the

Figure 2.3: Usage of drive recorders by policyholders



period between February 7, 2014 to December 26, 2016. To provide more background information on drive recorder usage, Figure 2.3 displays time series data for drive recorder usage patterns. In *Figure 2.3*, the green line represents the total number of policyholders actively using the drive recorder. We thereby speak of active usage at a certain date x if there are available driving logs from some date prior to x and some date after x. The blue line shows the number of individuals with driving logs at that date. The green line indicates that the adoption level of *usage-based* insurance policies has been steadily increasing throughout the period March 2014 - End 2016. High-frequency variations of the blue line can be attributed to weekly usage patterns of the vehicle. We also observe temporal decreases during vacation seasons. Statistics for the number of daily observations for 2016 are shown in Table 2.3.

The telematics dataset misses one important information, speed limits. Thus, we need to infer the speed limit for every location visited during the study period to be

Table 2.3: Summary statistics: Number of the available daily observations per policyholder in 2016

Quantile	Min	10%	25%	50%	75%	90%	95%	97.5%	99.5%
No. of days per policy number	1	45	92	200	277	311	324	332	344
No. of observations per day per policy number	1	45	81	135	213	313	392	481	744

able to construct variables for speed violations. The required information on speed limits is available from the *Open Street Map* (OSM) database.[9] We wrote a query to extract information about roads and corresponding speed limits via Overpass API in Switzerland and in some neighbouring countries. The speed limit data was stored offline. Storing the speed limit data offline had two important advantages: better query execution time and low risk of exposing policyholder's location information to unauthorized third parties. However, we needed to specify the area of interest for which the information should be extracted. The choice was guided by several factors: The algorithm's running time and the algorithm's complexity were traded-off against the estimated number of untagged points. Country codes attached to some observations suggest that clients were driving in more than 44 countries in Europe and Africa. We decided to load road data for the region between longitude 4.864 - 11.744 and latitude 43.135 - 48.905, since this area contained over 92% of relevant locations and had highest average observation density. The result was a table with suggested speed limits for 5'106'076 locations. We developed an algorithm to select a speed limit from this table for a given data point depending on the relative proximity. Distances are computed using the haversine formula and the distance to the closest point with available speed tag is saved to assess the precision.

The precision of speed limits for a representative subset of observations is summarized in Table 2.4. Table 2.4 shows that for an average location information about speed limit is taken from a tagged point that is approximately 17-18 meters away and for 75% of observation the distance does not exceed 68 meters. In extreme cases for observations outside densely populated areas and away from important roads and highways, the closest speed tags can be as far as 1 - 2 km away. This however only applies to a very small share of data points.

Around 12% of data points has speed limits not in line with official speed regulations. The discrepancies were frequently very small (i.e. 34 instead of presumably 30). More strongly inconsistent speed limits were replaced by the closest official speed

[9] OSM is a crowdsourced spatial database, started in 2004 with an objective of creating a free, detailed and editable map of the world. According to Mooney et al. (2017), OSM compares favorably with other spatial databases in terms of data quality.

Table 2.4: Summary statistics: Precision of inferred speed limits

Quantile	min	10%	25%	50%	75%	95%	97.5%
Precision (km)	0	0.004	0.007	0.017	0.062	0.678	1.204

[1] Based on subsample of 47464854 observations

value.

One problem in the data that could not be adressed is the potential accelerometer recalibration on some vehicles: the directions of axes in the accelerometer can be perturbed in the course of time, which might have caused some incorrect tagging of elevated g-force events (e.g. acceleration will be recorded as harsh braking or cornering).

Variables created from the Telematics Data

In the current subsection we describe and summarize how the logs from the telematics device are used to quantify a policyholder's driving characteristics. The goal is to create a set of variables that capture all important dimensions of driving characteristics affecting risk.

As a starting point, we consider variables frequently used in telematics-based insurance policies. In "Pay-as-you-drive" policies, the insurance price typically depends on distances driven, sometimes with different weighting factors for the time of the day and the road type. In "Pay-how-you -drive" contracts, a higher emphasis is put on the ‘quality of driving’ reflected by average speed, speeding and number of elevated g-force events. To these variables used for pricing in existing telematics-based policies, we add variables from the literature featuring telematics data. Ayuso et al. (2016) use *average distance per day*, *percentage of urban and night driving* and *percentage of km driven above the speed limit*. The authors conclude that speed violations, urban and nighttime driving have a significant effect on reducing the time to the first accident at fault. Kremslehner and Muermann (2016) extract information on the *number of journeys per day, the percentage of weekend driving and average speeding*. They find that the first two are positively associated with a higher probability of a Bonus Malus reclassification.

Table 2.5: Information from driving logs used for distance calculation

Notation	Variable	Observation frequency	Available for
$lat_i, \quad lon_i$	latitude and longitude of the observation	available for all data points	$i \in \{0, N^{tot}\}$
$timestamp_i$	time of the observation ($hour : min : sec$)	available for all data points	$i \in \{0, N^{tot}\}$
mt_i	meters traveled	irregular, approximately every 2 km	$i \in \mathcal{I}_1$
s_i^{point}	point speed	irregular, measured with $s^{\text{point end}}$	$i \in \mathcal{I}_1$
$s_i^{\text{point end}}$	point speed at the end of the event	irregular, measured with s^{point}	$i \in \mathcal{I}_1$
$s_i^{average}$	average speed of the last point	irregular, missing when s^{point} is provided	$i \in \mathcal{I}_2$
$t_i^{elapsed}$	time elapsed	irregular, missing when s^{point} is provided	$i \in \mathcal{I}_2$
N^{tot}	total number of observations during the day		

The variables we create from the telematics dataset describe 5 important aspects of driving characteristics: distances driven, average speed, number of journeys, elevated g-force events and speed violations. We additionally consider variables that specify the above driving behavior for certain locations (e.g. speeding in the urban areas) or times of the day (e.g. distance driven during the night).

Calculating precise distances driven based on the driving logs is not straightforward. The telematics dataset has several groups of measurements that can be used to approximate distances driven. Some of these measurements are available for every data point, others were measured irregularly (see summary in Table 2.5). Most notably information on the *meters traveled* was recorded on average every 2 km, which could result in over- or underestimation of distances driven on a specific road type or during certain time periods. To address this problem we use four different approaches, summarized in Table 2.6. The purpose is twofold: First, the average of the created distance variables is likely to be more precise, and second, the four measures can be used to cross check each other. Table 2.7 shows summary statistics for the distance variables obtained using the four different approaches. *Algorithm 1, 3, 4* result in similar estimates whereas *Algorithm 2* yields higher values. The discrepancy might be caused by the properties of the haversine distance formula used to compute distances between observations. In the subsequent analysis, to evaluate the distance we average results of *Algorithm 1, 3, 4*.

Assessing average speed posed several challenges as well. First, Table 2.5 suggests

Table 2.6: Different approaches to compute distance driven

Distance formula	Comments
$Dist_1 = \sum_{i \in \mathcal{I}_1} \frac{mt_i}{1000}$	imprecise since mt_i is measured irregularly
$Dist_2 = \sum_{i=1}^{N^{tot}} CoordDist(lat_i, lon_i, lat_{i-1}, lon_{i-1})$	results in the lower bound of distance driven by the vehicle
$Dist_3 = \sum_{i \in \mathcal{I}_2} s_i^{point}(timestamp_i - timestamp_{i-1})$	
$Dist_4 = \sum_{i \in \mathcal{I}_1} s_i^{average} t_i^{elapsed}$	imprecise due to speed variations during aggregation period

there are multiple variables that reflect different aspects of vehicle's speed: s_i^{point} is the speed of the vehicle at the location (lat_i, lon_i), whereas $s_i^{average}$ is the average speed between two subsequent observations. Second, since measurements are taken at unequal frequency, to get unbiased estimates of average speed during certain timespan we need to apply weights to individual observations. Data points can be weighted with respect to distance traveled (measured in 2 different ways) or time elapsed between subsequent measurements. The general formula for average speed computation, alongside with further details is given in Table 2.8.

Speed violations are important predictors of accident involvement in previous studies (Kremslehner and Muermann (2016), Ayuso et al. (2016)). To evaluate speeding behavior we have extracted speed limits from *Open Street Map Database*, as described in the previous subsection. We describe it by two types of variables. Similar to Ayuso et al. (2016), we compute *percentage of driving above the speed limit*. Kremslehner and Muermann (2016) suggests another indicator: *average speeding* that summarizes by how much speed limits are exceeded. We slightly modify this predictor by taking weighted average of speeding with respect to distance driven.

Originally, the number of journeys per day was calculated as a number of distinct trip id's per day that coincided with the number of times the engine was switched on. The approach resulted in an extremely high number of journeys for certain days, where either the length of a journey or the length of a stop were less than a minute. This could result from a start-stop system implemented in some vehicles: the engine is automatically shut down when the car stops to reduce the fuel consumption and emission. Consequently stops at a stop light or in a traffic jam could inflate the number

Table 2.7: Predictors aggregated on the annual basis, obtained using different computation procedures

	Quantile					
	5%	**10%**	**25%**	**50%**	**75%**	**97.5%**
Distance calculation						
Algorithm 1	1051.04	1859.75	4337.54	8664.77	13792.03	25717.65
Algorithm 2	1167.63	2092.79	4826.63	9493.42	14944.94	28560.24
Algorithm 3	1051.41	1860.17	4336.51	8664.97	13792.99	25717.90
Algorithm 4	1035.93	1840.52	4359.39	8700.72	14196.41	27505.90
Speed calculation						
Algorithm 1	35.36	39.42	45.62	52.56	60.94	80.04
Algorithm 2	11.96	13.78	16.76	20.36	24.72	34.19
Algorithm 3	26.59	29.01	33.05	37.54	41.65	49.31
Algorithm 4	30.97	34.92	41.90	49.56	58.93	79.24
Algorithm 5	19.00	21.52	25.99	31.48	37.45	51.71
Number of journeys per day						
Trip duration 20 stop duration 120	89.00	157.60	339.00	693.00	1048.00	1743.00
Trip duration 20 stop duration 60	91.00	163.00	351.00	714.00	1078.00	1809.60
Trip duration 30 stop duration 120	89.00	157.60	339.00	693.00	1048.00	1741.20
Trip duration 30 stop duration 60	91.00	163.00	351.00	714.00	1078.00	1809.00
Number of accelerations						
Threshold = 0	86.80	178.00	482.00	1211.00	2560.00	7044.80
Threshold = 1	86.00	178.00	479.00	1202.00	2551.00	7021.60
Threshold = 2	85.80	176.00	473.00	1186.00	2518.00	6909.20
Threshold = 10	78.00	161.60	428.00	1066.00	2172.00	5798.00
Number of brakings						
Threshold = 0	214.80	453.00	1132.00	2456.00	4321.00	9071.20
Threshold = 1	213.60	450.00	1127.00	2436.00	4299.00	9030.60
Threshold = 2	210.60	442.00	1113.00	2400.00	4241.00	8895.80
Threshold = 10	195.80	419.60	1051.00	2263.00	3959.00	8130.60
Number of cornerings						
Threshold = 0	976.60	1765.20	4055.00	8255.00	13288.00	26536.60
Threshold = 1	957.40	1751.00	4033.00	8235.00	13224.00	26500.40
Threshold = 2	950.40	1728.20	3991.00	8156.00	13058.00	26169.80
Threshold = 10	726.40	1345.80	3017.00	6058.00	9653.00	17557.80

Table 2.8: Summary of approaches to assess average speed of the vehicle

Speed	**Type**	**Speed Variable** (sv_i)	**Weight** (w_i)	**Observations**
$Speed^{type} = \frac{\sum sv_j^{type} w_j^{type}}{\sum_j w_j^{type}}$	1	$s_j^{average}$	mt_j	$j \in \mathcal{I}_1$
	2	s_j^{point}	$timestamp_j - timestamp_{j-1}$	$i \in \mathcal{I}_1$
	3	s_j^{point}	$CoordDist(lat_j, lon_j, lat_{j-1}, lon_{j-1})$	$i \in \mathcal{I}_1$

journeys per day. To account for this, we have introduced two thresholds: the stop duration threshold and the trip duration threshold. Only journeys that exceeded the trip duration threshold and with the previous stop exceeding stop duration threshold were evaluated as separate journeys, whereas journeys that did not fulfill these criteria were merged with preceding journeys. We considered different combinations of trip and stop duration thresholds mentioned in the literature (e.g. Ippisch (2010)). Table 2.7 suggests that in our dataset this choice does not have a strong impact on the resulting predictor values. In our subsequent analysis, we use the trip duration threshold equal to 30 seconds and the stop duration threshold of 120 seconds.

In contrast to previous studies with drive recorder data (Ippisch (2010), Ayuso et al. (2016), Kremslehner and Muermann (2016) our dataset contains information about a vehicle's accelerations. Existing literature suggests that elevated g-force events are positively related to accident occurrence, as they decrease the time available to respond to hazards (Dingus et al. (2006)) and make vehicle motion less predictable for other road users, reducing the safety margins (Simons-Morton et al. (2011)). We will therefore add variables based on the accelerometer data to the analysis. First, we consider the frequency of elevated g-force events. Again, this requires a careful adjustment of the data. Some drive recorders registered several cornering events within a second, several accelerations and brakings within 2-3 seconds. Consequently, a higher number of elevated g-force events might be partly caused by properties of the device and not by the policyholder's driving behavior. We have decided to count all events of a certain type within the time span of several seconds as a single event. The total number of elevated g-force events with respect to different time span thresholds are summarized in Table 2.7. Acceleration and cornering counts are more sensitive to the choice of threshold. The driving logs contain additional information about elevated g-force events. The device records point speed at the start and at the end of an event, which can be used to evaluate its severity. Furthermore, accelerations and harsh braking at a higher speed might be more dangerous, thus we create a separate variable to capture this information.

In the regression analysis, all variables are aggregated over 2016 (see also discussion

Table 2.9: Summary statistics: Driving indices

	mean	median	std			Sample quantiles			
				min	5%	25%	75%	95%	97.5%
General characteristics									
Average distance per day (active)	53.03	49.25	23.70	2.60	22.65	36.32	65.09	95.70	109.09
No. of journeys per day (active)	3.96	3.80	1.06	1.26	2.60	3.23	4.47	5.91	6.47
Journey time and location									
Weekend driving percentage	0.30	0.29	0.12	0.00	0.12	0.22	0.36	0.52	0.58
Urban driving percentage (speed=50)	0.31	0.29	0.14	0.01	0.12	0.21	0.39	0.56	0.63
Night driving percentage	0.06	0.04	0.06	0.00	0.00	0.01	0.08	0.18	0.22
Speed and Speeding									
Average speed (type 1)	62.25	61.84	12.95	2.61	42.18	54.30	70.61	83.62	88.02
Percentage of driving above speed limit	0.22	0.21	0.11	0.00	0.06	0.14	0.29	0.41	0.45
Average speeding (weighted)	12.31	11.79	3.44	3.77	7.86	9.93	14.04	18.58	20.59
Percentage of driving above speed limit in urban areas (type 1 adjusted	0.33	0.31	0.13	0.03	0.14	0.23	0.40	0.57	0.63
Average speeding in urban areas (relative)	1.23	1.16	0.31	0.18	0.93	1.05	1.33	1.80	2.02
Characteristics of elevated g-force events									
No. of accelerations per km (tr=2)	0.20	0.16	0.18	0.00	0.03	0.08	0.27	0.55	0.69
No. of braking per km (tr=2)	0.33	0.29	0.20	0.00	0.09	0.20	0.42	0.69	0.81
Average speed difference after acceleration	10.97	10.49	2.55	0.00	7.87	9.32	12.09	16.13	17.59
Average speed at the beginning of accelerations	14.34	13.82	6.47	0.00	4.41	9.74	18.43	25.82	28.31
Average speed difference after braking	-14.39	-14.37	2.01	-23.22	-17.61	-15.67	-13.11	-11.38	-10.76
Average speed at the beginning of braking	37.00	37.03	5.74	0.00	28.18	33.41	40.67	46.17	47.91

in Section 2.2.4). The total distance driven and the total number of journeys are divided by number of days when according to the drive recorder the vehicle was used. Speeds and speedings are averaged over all observations, the number of events is divided by total distance driven. Table 3.1 shows the summary statistics for the created telematics-based driving variables.

2.2.4 Aggregation of Information from all Datasets

The original datasets contain multiple observations for every policyholder corresponding to different time periods. Furthermore, for individuals with available driving logs, observation periods, marked by the dates of the earliest and the latest drive recorder entries, vary significantly. To account for these differences, we have looked at several alternatives. First, similar to Kremslehner and Muermann (2016), we could select a fixed study period and analyze only policyholders with driving logs available throughout this whole time span. This approach introduces the following trade-off: A shorter period increases the number of policyholders that satisfy the selection criteria (number of observations), but the number of submitted liability claims decreases. The second possibility is to aggregate independent variables in such a way that the results are unaffected by the length of the observation period per se (i.e. use *average no. of km per day* instead of *total distance driven*). This entails the problem that variables obtained by processing different numbers of original observations might have non-equal variances. However, once more than a certain number of days are observed, the quality of information about driving characteristics should not improve significantly. If this holds, aggregated driving patterns for policyholders with more than N days of driving logs have the same variance and excluding other policyholders from our analysis should be sufficient to avoid biases. We we will use this approach.

The observation period duration likewise affects the distribution of the dependent variable. The risk realization within a pre-specified timespan is determined by both its length and hazard rate. Since we are interested in the hazard rate, to ensure that the impact of independent variables is not attenuated, the risk exposure periods should not vary too much.

In light of these considerations, we consider data from the year 2016 for policyholders with at least 10 days of driving logs. To decrease potential noise, we also consider only liability claims exceeding 10 CHF. Robustness of our results with respect to the choices of number of available telematics observations adn liability claim threshold is discussed in Section 2.3.3.

Among the remaining policyholders with available contract data and driving logs corresponding to 2016, 212 were older than 26, which on the first glance appears strange considering that this type of contract was offered exclusively to drivers younger than 26. The presence of drivers older than 25 years old can be explained by two factors:

first, in order to increase the degree of anonymity exact birth dates were replaced by some random dates within the same year, second - the device was used by some employees of the company. Consequently we can conclude that all employees older than 26 years old work for the insurer and we exclude them from the analysis to decrease the potential for selection bias. The filtering steps applied to our dataset are summarized in Table 5.6 in the Appendix. Table 2.2.4 provides an overview with the descriptions of variables created from the datasets. To improve readability, variables with self-explanatory labels are not included.

2.2.5 Empirical Approach

The main focus of our analysis is whether adding the telematics data can improve our understanding of risk factors, and in particular which information might be sensibly used in insurance design. We will also conduct the tests for asymmetric information.

For the analysis of risk factors, liability claims both imply culpability of the driver and are observed for all policyholders.[10] Thus in the subsequent analysis a liability claim submission will be used to model risk realization. As only a very small fraction of policyholders submitted multiple liability claims, ex post risk during the considered time period is modeled as a binary variable. Furthermore, following the discussion in Section 2.2.2, we have decided not to count claims less or equal 10 CHF in our main analysis.

To test for asymmetric information, following Chiappori and Salanie (2000) and Finkelstein and Poterba (2014), we use the *positive correlation test* and the *unused observables test* respectively. The *positive correlation test* and the *unused observables test* require fitting a system of regression equations that model *ex post* risk realization and choice of insurance cover. Given our data, the *level of the deductible for liability claims* cannot be used for the choice of insurance coverage as there is not enough variation. For the choice of insurance coverage, we therefore use *Purchase of collision cover* (CollisCover) and *Purchase of bonus protection for TPL claims*(BM TPL). Both decisions are represented by binary variables, whereby 1 stands for the presence of

[10] Contrary to that, i.e. collision claims are only observed for policyholders with collision coverage

insurance coverage.

We can now introduce our econometric model. The dependent variables are binary and thus can be modeled by the following system of probit equations:

$$\text{LiabClaim} = \mathcal{I}(X\alpha_1 + Y\beta_1 + \epsilon_1 > 0) \tag{2.1}$$

$$\text{BM TPL} = \mathcal{I}(X\alpha_3 + Y\beta_3 + \epsilon_3 > 0) \tag{2.2}$$

$$\text{CollisCover} = \mathcal{I}(X\alpha_2 + Y\beta_2 + \epsilon_2 > 0) \tag{2.3}$$

The variables in the regression model can be divided into two groups. The first group (X) comprises variables created based on traditional information collected for risk classification and stored in the contract and claims datasets. The set of telematics based predictors is denoted by Y.

Regarding the question of asymmetric information, the logic of the *positive correlation test* is that the data provides evidence of asymmetric information between policyholer and insurer if there is statistically significant correlation between either ϵ_1 and ϵ_2 or ϵ_1 and ϵ_3. The test however might not be able to detect hidden information under multiple dimensions of asymmetric information, most notably if both risk and risk preferences are heterogeneous and private information.[11] To address some of the test's limitations, Finkelstein and Poterba (2014) introduced the *unused observables* test. The logic of the test is that finding a variable that a) is not used by the insurer for risk classification ("unused observable") and b) is statistically significant for predicting both insurance cover choice and ex-post risk realization provides evidence of an information asymmetry. In our setting, it amounts to testing the null hypothesis $\beta_1 = \beta_2 = 0$ or $\beta_1 = \beta_3 = 0$.

For the *unused observables test* there need to exist variables that are not used for the premium calculation. In the insurance contracts at hand, the telematics data is used to calculate annual driving scores to rank policyholders. The annual ranking determines the premium discount in the subsequent year, which has three possible levels. The scoring algorithm is proposed and implemented by an independent telematics company.

[11] See e.g. Finkelstein and Poterba (2014) for a detailed discussion.

The scores are largely determined by the number of the elevated g-force events (e.g. accelerations, harsh brakings, cornerings) per km driven.[12]

Availability of the raw telematics data to the insurance company, as noted by Kremslehner and Muermann (2016), could lead to biases in our analysis in case when this information is used in other types of underwriting activities, most notably when policyholders are offered different contracts depending on the observed driving patterns. This seems to be unlikely in our setting due to several factors: first, the telematics usage based insurance contracts were offered quite recently, therefore it is unlikely that the company had time to adjust its trading strategies in view of the new available information. Second, in contrast with Finkelstein and Poterba (2014) and Saito (2006) telematics data is not available to the insurance company prior to signing the policy.

2.3 Results

2.3.1 Main Results

We fit a trivariate probit model given by Equations (2.1) - (2.3).[13] Table 2.11 contains the results of the trivariate probit model. The upper part of the table shows the coefficients of traditional risk classification variables and further variables from the insurance contract dataset.

We start with risk factors for liability claim submission and the role of telematics variables in predicting it. Table 2.11 compares models fitted with and without telematics-based variables. To compare the nested models we use the AIC scores. Wagenmakers and Farrell (2004) transform the AIC scores to AIC weights that can be directly interpreted as conditional probabilities for each model. The resulting AIC weights for the two models suggest that the full model is superior with probability over

[12] We cannot disclose the full scoring algorithm. For the argument, the important point is that the scoring algorithm, and therefore ranking and pricing is to a great extent determined by elevated g-force events per km driven only.

[13] A concern regarding the telematics variables is multicollinearity. Table 5.3 in the Appendix provides the correlation matrix with variance inflation factors of telematics variables. Despite the fact that some pairs of independent variables have weak to moderate correlation coefficients (0.3 - 0.6), variance inflation factors indicate that this multicollinearity does not affect significance tests.

99.9%. Figure 5.1 in Appendix 5 shows the corresponding ROC curves for the liability claim submission models. Thus, the information provided from telematics data is valuable to model risk measured by liability claim occurence.

Looking into the results of the full model in Table 2.11 (left columns), we observe that both *average distance per day (active)* and *no. of journeys per day (active)* have a positive and statistically significant effect on accident involvement. Note that both variables reflect risk exposure, as opposed to driving style. The result for *no. of journeys per day active* mirrors the result in Kremslehner and Muermann (2016).

Regarding the variables that describe driving style only *average speeding (weighted)* and *percentage of driving above speed limit (type 1 adjusted)* are significant. Surprisingly, neither number nor characteristics of elevated g-force events are relevant. Furthermore, our analyses revealed that *average speeding (weighted)* is not statistically significant unless information about speeding behavior in urban areas is included in the model. An explanation for this finding could be the following: under normal circumstances a driver violates speed limits either if he is negligent and prone to risk taking or if the limits are too stringent. Urban areas have more stringent speed restriction, thus a person driving 40 km/h can exceed allowed speed by e.g. 20 km /h. Arguably, this type of driving has different implications for the hazard rate than driving 140 km/h on a motorway. Adding information about speed violations in urban areas could help to separate the two cases. This could also partly explain why *average speeding* is not a significant predictor in the model of Kremslehner and Muermann (2016).[14]

Our results suggest a weakly significant negative effect of *weekend driving percentage*. Weekend driving could be used as a proxy for two important risk factors. On the one hand, people driving during weekends might be less experienced (*Sunday drivers effect*), thus higher weekend driving percentage should be associated with higher risk. On the other hand, traffic intensity during weekends tends to be lower than during the

[14] Kremslehner and Muermann (2016) evaluate speeding by comparing point speed with countrywide speed limits per road type. This approach is imprecise since local speed limits might be lower in certain urban areas, or on more dangerous roads. Consequently predictors used by Kremslehner and Muermann (2016) might underestimate average speeding.

week, therefore people who drive predominantly on weekends drive with less traffic and consequently have a smaller accident probability. Since all policyholders in the dataset are relatively inexperienced, the result suggests that the relevant factor is lower traffic intensity.
Statistical significance does not necessary imply relevance. To illustrate the role of significant telematics variables, we compute claim submission probabilities for their sample quantiles while keeping other variables at the mean. Another way to assess variable's influence is to see by how much predictions vary relative to the accident probability computed from regression model without the drive recorder data ('small mode'). This comparison highlights the extent of the imprecision from the absence of drive recorder data in the regression model. This information can be found as the respective third line in Table 2.12, (% of default probability).

The largest spread of predicted probabilities corresponds to the variable *No. of journeys per day (active)*. Undertaking on average 2.8 journeys per day (10% sample quantile) instead of 5.9 (95% sample quantile) decreases claim submission probability by almost the factor of two from 8.7% to 4.4%. The values vary from 48% to 161.6% relative to probability computed using the 'small regression model' (line 6 in Table 2.12).

Both *average speeding (weighted)* and *average distance per day (active)* have a sizeable impact on liability claim submission. Using different sample quantiles, we obtain values ranging from 3.6% to 8% for *average speeding (weighted)* and from 3.6% to 8.8% for *average distance per day (active)*. The effects of *weekend driving percentage* and *percentage of driving above speed limit* are less pronounced with differences in predicted probabilities amounting to approximately 3%.

With respect to standard risk classification variables, in both models with and without telematics variables, older drivers are associated with a lower probability of liability claim submission. This finding is consistent with older drivers being more experienced drivers. Furthermore, newer vehicles are associated with a lower probability of liability claim submission. We also observe that the probability of a liability claim is significantly higher if there was a recent change of address. A possible explanation of this finding is that following a change in address, the policyholder drives in a less familiar area which adversely affects accident risk. Interestingly, *no. of years without first contract* has a (weakly) significant positive effect. There are two possible explanations that could account for this result: First, it might indicate an inexperienced driver. Second, it could also be the case that the policyholder already had a previous insurance contract with another insurer and recently switched insurers. In the latter case, the result would indicate switching of a higher risk individual. However, we cannot disentangle these two effects. In the absence of telematics variables, the magnitude and, e.g. for recent change of address, the significance level of these variables slightly increase. We furthermore find that the claim history, reflected by the *Bonus Malus Score TPL* and *no. of previous mobility claims*, is not statistically significant in both regression models with and without telematics based variables. At first glance, this finding seems surprising. However, since the sample consists of young drivers, the past claim history does not yet contain a lot of information about the policyholder's risk type.

We now turn to the results for contract choice displayed in the middle and right columns of Table 2.11. With respect to variables from the insurance contract dataset, women are associated with a higher probability to purchase collision cover and BM protection. This finding could reflect a higher risk aversion of women.[15] Note that the left columns of Table 2.11 show that gender is not significantly associated with liability claim occurence. The results furthermore show that vehicle age is negatively and a leasing contract positively associated with collision coverage and BM protection. We also find that age is significantly positively related to the purchase of collision cover. Combined with the result of a negative effect of age on risk, this suggests that the underlying driver is not risk type.

Regarding the telematics variables, a higher *no. of journeys per day* is associated with a significantly higher propensity to purchase collision cover and BM protection.

[15] Several experimental studies have found gender differences in risk preferences with more averse choices by women. However, in surveying the literature Filippin and Crosetto (2019) find that gender differences are less ubiquitous than often discussed and systematically correlate with the features of the elicitation method used.

A higher *no. of journeys per day* also has a significant positive effect on liability claim submission. The number of journeys per day is thus positively related to both risk and coverage, consistent with the presence of asymmetric information given the existing risk classification and contracting. Importantly, the number of journeys does not enter the scoring and thus relative ranking of the policyholder that is the basis for premium adjustments. We find the opposite result for *average distance per day (active)*. The average distance is significantly positively related to risk but negatively to the purchase of insurance coverage. A possible explanation behind this result for the average distance per day could be overconfidence as modelled in Sandroni and Squintani (2013), who suggest that some policyholders might be overconfident and underestimate their risk level. Individuals who drive longer distances (*average distance per day (active)*) could thus be overestimating their driving skill and purchasing less insurance coverage. Overall, we furthermore also observe telematics based variables that are statistically significant in the risk regression, but that do not appear to be systematically associated with the insurance choice, and vice versa.

The results of the *positive correlation test* are shown in Table 2.13 in which the correlation coefficients between residuals from Equation (2.1) and Equation (2.2) or respectively (2.3) alongside confidence intervals for significance tests are shown. The positive correlation test fails to detect asymmetric information, both without and with telematics variables included.

The combination of the above findings indicates that strong caution is required for the interpretation of results: Different dimensions of risk type, preferences, perceptions and information seem to interact. In particular with respect to information about and perception of accident risk factors, an important question is how well the policyholders in our sample—all under the age of 26 years with only a few years of driving experience—understand these risk factors.

2.3.2 Older vs. Younger Drivers

In the previous subsection we did not distinguish between inexperienced, younger drivers and more experienced, older drivers who might have a better understanding of their driving behavior and risk type. The purpose of this section is therefore to stratify the dataset with respect to experience, fit the trivariate probit models on these subsets of data and contrast the results.

To capture experience, the most direct proxy is the driving license issuance date, however we do not possess this information. If the majority of policyholders started driving upon reaching the age of 18, stratification with respect to age is a reasonably good proxy. A second option is to infer experience based on the aggregate policy duration, under the premise that policyholders do not change the insurance company. For this option, low switching costs and the commodity-like nature of automobile insurance however imply caution when interpreting the results.

For the selected proxies for driving experience, a threshold that separates more and less experienced policyholders needs the be selected. This choice is constrained by the following practical considerations: to obtain numerically stable results it is desirable to have sufficient amount of claims and observations for every subgroup. Table 2.14 reports this statistics for various feasible options of durational thresholds. The distribution of aggregate contract duration in our dataset is right skewed and for any meaningful choice of threshold, the subgroup of policyholders that have longer contracts is relatively small, thus we won't analyze them further. For three other subgroups, we perform our inference for all plausible durational threshold choices to ensure that the conclusions are robust.

Table 2.15 compares the means of selected telematics variables reflecting driving patterns by experience. Older policyholder have a slightly higher distance driven per day, and a higher average speed (type 1).

The results of the fitting trivariate probit models on subsamples by age for different separation thresholds are summarized in Table 2.16. For a better readability, it is only reported whether and at what level the variables are statistically significant. Table 5.5 in the Appendix provides the coefficients.

Table 2.16 shows that the *number of journeys (active)* and *average speeding (weighted)* are statistically significant risk factors increasing liability claim occurence for young drivers. For older drivers, additionally the average distance per day (active) and, depending on the threshold, the *percentage of driving above the speed limit* and *night driving percentage* are additional risk factors. The results for insurance demand correspond to those of the pooled sample and are largely the same for both older and younger drivers. In particular, a higher number of journeys is associated with a higher demand for collision cover and BM protection for both younger and older drivers and a higher average distance per day is negatively related to insurance demand. Thus, while the insurance demand patterns do not change across the two subsamples, the analysis picks up more significant risk factors for older drivers. These pertain both to risk exposure (*average distance per day*) and *night driving percentage*, as well as to driving style (*percentage of driving above the speed limit*).

These results also suggest caution to interpret the findings as relevant asymmetric information in the form of the policyholder having superior information about risk type or preferences that affects selection in a systematical manner for insurer profits: While, at least stastistically, more risk factors are highlighted for the subset of older—yet still relatively inexperienced—drivers, we do not observe different patterns in insurance contract choice.

Table 2.10: Description of explanatory variables used in the regression analysis.

Variable	Description
Recent change of address: true	Did policyholder changed his address during insurance year ?
No. of years without first contract	How many years passed between driver turning 18 and getting the first contract with the insurer
No. of previous mobility claims	How many mobility claims did driver submit prior to the beginning of insurance year
Average distance per day (active)	Average distance driven during the days when the vehicle is used
No. of journeys per day (active)	Average no. of journeys during the days when the vehicle is used
Average speed (type 1)	Average speed of vehicle, calculated using *Algorithm 1*
Weekend driving percentage	Ratio of distance driven on a weekend to total distance driven
Urban driving percentage (speed = 50)	Ratio of distance driven in urban area to total distance driven. All areas with speed limit $\leq 50km/h$ are classified as Urban
Night driving percentage	Ratio of distance driven during the night to total distance driven.
Percentage of driving above speed limit (type 1 adjusted)	Ratio of distance driven above speed to total distance driven. Speed limits obtained from OSM database are adjusted to conform with Swiss speed restrictions
Average speeding (weighted)	Average of speeding with respect to adjusted speed limits, weighted by the distances driven
No. of accelerations per km (tr = 2)	Average no. of accelerations per km driven, whereby several accelerations registered within the timespan of 2 seconds are counted as 1
No. of braking per km (tr = 2)	Average no. of braking per km driven, whereby several accelerations registered within the timespan of 2 seconds are counted as 1
Average speed at the beginning of accelerations	Average speed at the beginning of acceleration event
speed difference after accelerations	Difference between the vehicle's speed after and before acceleration

[1] Variables with self-explanatory names are not included.

Table 2.11: Coefficients of fitted probit models (with and without telematics based predictors)

	liability claim ($> 10CHF$)		BM protection		Collision cover	
Traditional Insurance Variables						
Sex of driver: female	−0.031	0.029	0.106**	0.136**	0.368***	0.361***
	(0.056)	(0.059)	(0.053)	(0.056)	(0.056)	(0.059)
Age of driver	−0.061***	−0.053***	−0.002	0.001	0.122***	0.119***
	(0.018)	(0.019)	(0.020)	(0.020)	(0.018)	(0.018)
Vehicle: age	0.020***	0.020***	−0.009	−0.009	−0.289***	−0.291***
	(0.007)	(0.007)	(0.006)	(0.007)	(0.011)	(0.011)
Vehicle: horsepower	−0.000	−0.000	−0.001	−0.002***	0.001*	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.000	0.000	−0.000**	−0.000	−0.001***	−0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Vehicle: price	0.000	−0.000	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Vehicle: mileage	−0.000	−0.001	−0.001**	−0.001***	−0.003***	−0.004***
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
Leasing contract: true	0.059	−0.027	0.173**	0.194**	1.143***	1.184***
	(0.080)	(0.083)	(0.082)	(0.085)	(0.129)	(0.133)
Recent change of address: true	0.151**	0.119*	−0.114*	−0.082	−0.043	−0.024
	(0.069)	(0.070)	(0.065)	(0.067)	(0.071)	(0.072)
Bonus Malus Score TPL	−0.002	−0.011	0.050**	0.061**	0.016	0.021
	(0.023)	(0.024)	(0.024)	(0.025)	(0.023)	(0.023)
No. of years without first contract	0.036**	0.030*	−0.095***	−0.079***	−0.086***	−0.075***
	(0.017)	(0.017)	(0.020)	(0.019)	(0.016)	(0.017)
No. of previous mobility claims	−0.026	−0.040	0.196**	0.185**	−0.047	−0.038
	(0.069)	(0.071)	(0.083)	(0.084)	(0.062)	(0.063)
Telematics-Based Predictors						
Average distance per day (active)		0.004***		−0.009***		−0.006***
		(0.002)		(0.001)		(0.002)
No. of journeys per day (active)		0.112***		0.125***		0.087***
		(0.027)		(0.030)		(0.030)
Average speed (type 1)		−0.002		0.002		0.002
		(0.003)		(0.003)		(0.003)
Weekend driving percentage		−0.449*		−0.294		−0.347
		(0.241)		(0.208)		(0.220)
Urban driving percentage (speed=50)		−0.124		0.246		−0.174
		(0.300)		(0.294)		(0.290)
Night driving percentage		0.589		1.389***		−0.206
		(0.450)		(0.461)		(0.450)
Percentage of driving above speed limit		0.610*		0.216		0.669*
(type 1 adjusted)		(0.348)		(0.332)		(0.350)
Average speeding (weighted)		0.022**		−0.025***		−0.005
		(0.009)		(0.008)		(0.009)
Percentage of driving above speed limit		−0.382		0.151		−0.025
in urban areas (type 1 adjusted)		(0.292)		(0.272)		(0.283)
Average speeding in urban areas (relative)		−0.107		0.081		−0.006
		(0.109)		(0.098)		(0.100)
No. of accelerations per km (tr=2)		0.147		−0.196		−0.267
		(0.178)		(0.179)		(0.175)
No. of braking per km (tr=2)		0.164		−0.053		−0.062
		(0.179)		(0.177)		(0.181)
Average speed at the beginning of accelerations		−0.001		−0.004		−0.015
		(0.013)		(0.013)		(0.013)
Average speed difference after acceleration		0.001		0.020***		0.020***
		(0.006)		(0.006)		(0.006)
Average speed difference after braking		0.008		−0.014		0.023
		(0.018)		(0.017)		(0.017)
Average speed at the beginning of braking		0.002		0.002		−0.008
		(0.007)		(0.007)		(0.007)
McFadden Pseudo R2	0.012	0.036	0.060	0.093	0.611	0.616
Maximum VIF	2.780	3.230	3.380	3.270	3.890	4.170
AIC weights	0	1	0	1	0	1
Observations	5,690	5,690	5,690	5,690	5,690	5,690
Akaike Inf. Crit.	2,721.309	2,685.674	3,052.881	2,977.527	3,093.275	3,083.956
Notes:						***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 2.12: Predicted probability of an annual liability claim submission for selected variables set to their sample quantiles

		Quantiles						
		min	**10%**	**25%**	**5%**	**75%**	**95%**	**97.5%**
Average distance per day (active)	value of predictor	1.375	26.811	36.081	49.070	65.092	95.900	105.059
	predicted probability	0.036	0.045	0.049	0.055	0.063	0.081	0.088
	% of default probability	61.3	77.6	84.4	94.6	108.6	139.9	150.4
No. of journeys per day (active)	value of predictor	1.000	2.802	3.209	3.784	4.467	5.911	6.302
	predicted probability	0.028	0.044	0.048	0.055	0.064	0.087	0.094
	% of default probability	48.0	75.0	82.6	94.3	109.9	149.3	161.6
Weekend driving percentage	value of predictor	0.001	0.158	0.221	0.287	0.368	0.535	0.585
	predicted probability	0.074	0.065	0.061	0.058	0.054	0.046	0.044
	% of default probability	127.5	111.4	105.3	99.3	92.3	79.0	75.4
Percentage of driving above speed limit (type 1 adjusted)	value of predictor	0.000	0.084	0.134	0.205	0.284	0.409	0.441
	predicted probability	0.043	0.048	0.051	0.056	0.062	0.072	0.074
	% of default probability	74.6	83.0	88.4	96.6	106.3	123.3	127.9
Average speeding (weighted)	value of predictor	2.591	8.526	9.892	11.762	14.029	18.586	20.063
	predicted probability	0.036	0.048	0.051	0.056	0.062	0.075	0.080
	% of default probability	62.2	82.5	87.8	95.6	105.7	128.6	136.8

Table 2.13: Correlation coefficients of residuals

	with	**without**
	telematics predictors	
$\rho_{liab,BM}$	0.0011 $(-0.0248, 0.0272)$	-0.0102 $(-0.0362, 0.0158)$
$\rho_{liab,Col}$	-0.012 $(-0.0379, 0.0139)$	-0.0006 $(-0.0267, 0.0253)$

Table 2.14: Number of available claims and observations for more and less experienced drivers depending on selection criteria and durational threshold.

	Older drivers (Age $\geq$ Age Break)				**Younger drivers (Age $<$ Age Break)**				**Contract duration (<Threshold)**		
Threshold (years)	20	21	22	23	20	21	22	23	1	2	3
No. of policyholders	4750	4040	3229	2377	1182	1892	2703	3555	1679	3615	5103
No. of claims	279	237	182	135	96	138	193	240	103	249	338

Table 2.15: Comparison of driving patterns between older (≥ 22) years of age) and younger (< 22 years of age) drivers

Variable	Younger (< 22)	Older (≥ 22)	contract since less than 2 years
Average distance per day (active)	50.953	54.537	53.634
No. of journeys per day (active)	3.956	3.939	3.974
Average speed (type 1)	59.934	63.697	61.034
Weekend driving percentage	0.306	0.301	0.310
Urban driving percentage (speed=50)	0.329	0.297	0.324
Night driving percentage	0.056	0.057	0.054
Percentage of driving above speed limit (type 1 adjusted)	0.206	0.224	0.208
Average speeding (weighted)	12.145	12.391	12.184
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	0.320	0.329	0.312
Average speeding in urban areas (relative)	1.209	1.254	1.230
No. of accelerations per km (tr=2)	0.217	0.194	0.209
No. of braking per km (tr=2)	0.335	0.327	0.334
Average speed at the beginning of accelerations	10.864	11.038	10.878
Average speed difference after acceleration	14.976	13.815	14.130
Average speed difference after braking	-14.321	-14.376	-14.202
Average speed at the beginning of braking	37.594	36.515	36.853

Table 2.16: Comparison of results for different choice of thresholds.

Group	older drivers												younger drivers												contract duration								
Dependent variable	liability claim (> 10CHF)				BM protection				Collision cover				liability claim (> 10CHF)				BM protection				Collision cover				liability claim (> 10CHF)			BM protection			Collision cover		
Threshold	20	21	22	23	20	21	22	23	20	21	22	23	20	21	22	23	20	21	22	23	20	21	22	23	1	2	3	1	2	3	1	2	3
Sex of driver: female					+++	++	++	+	+++	+++	+++	+++								+	+++	+++	+++	+++				+	+	+	+++	+++	+++
Age of driver	*	*		*					+++	+++	+++	+++				**						+	+	+++	*	*	*				+++	+++	+++
Vehicle: age	+++	+++	++	++	**	**	*	**	***	***	***	***			++	++					***	***	***	***	++	++	++	**	**	**	***	***	***
Vehicle: horsepower					*												**	***	***	***													
Vehicle: weight					*				***	***	***	***																			***	***	***
Vehicle: price					+++	+++	+++	+++	+++	+++	+++	+++					+++	+++	+++	+++	+++	+++	+++	+++				+++	+++	+++	+++	+++	+++
Vehicle: mileage					**	**	**		***	***	***	***					*			**	***	***	***	***							***	***	***
Leasing contract: true				*	++	++	+++	+++	+++	+++	+++	+++		+	+						++	+++	+++	+++	*	*	*	+++	+++	+++	+++	+++	+++
Recent change of address: true				+													**	**							+	+	+						
Bonus Malus Score TPL					++	++	+												+	++													
No. of years without first contract	+	+	++		***	***	***	***	***	***	***	***					**		*	**		**	*	***				***	***	***	***	***	***
No. of previous mobility claims					++	++	++	++																				++	++	++			
Average distance per day (active)	+++	+++	+++	++	***	***	***	***	***	**	***	***				+		***	***	***		***	***	**	++	++	++	***	***	***	***	***	***
No. of journeys per day (active)	+++	+++	++	++	+++	+++	+++	+++	+		+	+	++	+++	+++	+++	++	+++	+++	+++	+++	+++	+++	++	++	++	++	+++	+++	+++	+	+	+
Average speed (type 1)																	*																
Weekend driving percentage		*			*				*																								
Urban driving percentage (speed=50)							+																*										
Night driving percentage			+	++	++	++	++	++									+	+	++	++					++	++	++	++	++	++			
Percentage of driving above speed limit (type 1 adjusted)	+	++							++++														++										
Average speeding (weighted)	+	++	+		***	***	***	***					+	+	++	++												***	***	***			
Percentage of driving above speed limit in Urban areas (type 1 adjusted)																					+												
Average speeding in urban areas (relative)											++	++									*	***	***	**							++	++	++
No. of accelerations per km (tr=2)												**																			**	**	**
No. of braking per km (tr=2)																																	
Average speed at the beginning of accelerations																		*															
Average speed difference after acceleration					++	+			+++	+++	+++	++					+++	+++	+++	+++		++	+++	++							++	++	++
Average speed difference after braking																																	
Average speed at the beginning of braking							+	+													**		**					+	+	+			
No. of observations	4590	3907	3126	2296	4590	3907	3126	2296	4590	3907	3126	2296	1100	1783	2564	3394	1100	1783	2564	3394	1100	1783	2564	3394	1528	3430	4888	1528	3430	4888	1528	3430	4888
No. of positive values	275	234	180	133	4179	3542	2823	2068	2477	2201	1820	1361	93	134	188	235	1046	1683	2402	3157	329	605	986	1445	99	242	331	1371	3109	4446	710	1602	2283

[1] *** - positive effect with 1 % significant level [2] * - negative effect with 1% significance level

2.3.3 Robustness: Risk Factors

For the analysis policyholders and liability claims are filtered based on several criteria. In this subsection we discuss robustness of our conclusions regarding risk factors with respect to these choices. Two parameters are considered: the minimum number of days with driving logs available and the minimum liability claim size. The former is used to for selecting policyholders, the latter for discarding very small liability claims.

Table 2.17 displays the coefficients of fitted probit models for liability claim occurence when all claims less than 0 CHF, 10 CHF and 100 CHF are discarded. The sign and order of magnitude of the stastisticially significant variables remain the same. Riks factors that are significant at least on a 5 % level in one of the models are at least 10% significant in the other models. However, some weakly significant variables, such as *Recent change of address: true*, *Weekend driving percentage* become insignificant when the liability claim threshold changes. One potential explanation of this finding is that different driving patterns and risk factors are associated with a higher likelihood of more and less severe accidents. Due to a small number of observed liability claims, we cannot further investigate this finding. However, the key results regarding distance and no. of journeys remain unchanged. Regarding available driving logs, Table 2.18 displays the results of liability claim submission models fitted on subsets of policyholders with not less than 0, 10, 20, 30 and 50 days of available driving logs in 2016. We observe similar patterns as in Table 2.17. Again, the results for distance and number of journeys is robust. Furthermore, the model fit improves when more stringent criteria are applied.

Table 2.17: Coefficients of fitted probit models for liability claim submission. Comparison of different claim size thresholds. Minimum number of observations (days of driving logs): 10

	liability claim (> 0CHF)	liability claim (> 10CHF)	liability claim (> 100CHF)
Traditional Insurance Variables			
Sex of driver: female	−0.017	0.029	0.034
	(0.056)	(0.059)	(0.059)
Age of driver	−0.037**	−0.053***	−0.049***
	(0.018)	(0.019)	(0.019)
Vehicle: age	0.015**	0.020***	0.020***
	(0.007)	(0.007)	(0.007)
Vehicle: horsepower	0.000	−0.000	−0.000
	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Vehicle: price	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)
Vehicle: mileage	−0.000	−0.001	−0.000
	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.047	−0.027	−0.023
	(0.079)	(0.083)	(0.084)
Recent change of address: true	0.124*	0.119*	0.098
	(0.066)	(0.070)	(0.071)
Bonus Malus Score TPL	−0.022	−0.011	−0.022
	(0.023)	(0.024)	(0.025)
No. of years without first contract	0.010	0.030*	0.029*
	(0.018)	(0.017)	(0.017)
No. of previous mobility claims	−0.035	−0.040	−0.028
	(0.065)	(0.071)	(0.070)
Telematics-Based Predictors			
Average distance per day (active)	0.003**	0.004***	0.004***
	(0.001)	(0.002)	(0.002)
No. of journeys per day (active)	0.122***	0.112***	0.118***
	(0.026)	(0.027)	(0.027)
Average speed (type 1)	−0.000	−0.002	−0.002
	(0.003)	(0.003)	(0.003)
Weekend driving percentage	−0.331	−0.449*	−0.387
	(0.228)	(0.241)	(0.244)
Urban driving percentage (speed=50)	−0.035	−0.124	−0.139
	(0.286)	(0.300)	(0.305)
Night driving percentage	0.650	0.589	0.722
	(0.429)	(0.450)	(0.453)
Percentage of driving above speed limit	0.567*	0.610*	0.704**
(type 1 adjusted)	(0.331)	(0.348)	(0.352)
Average speeding (weighted)	0.021**	0.022**	0.022**
	(0.008)	(0.009)	(0.009)
Percentage of driving above speed limit	−0.252	−0.382	−0.394
in Urban areas (type 1 adjusted)	(0.277)	(0.292)	(0.296)
Average speeding in urban areas (relative)	−0.040	−0.107	−0.117
	(0.101)	(0.109)	(0.111)
No. of accelerations per km (tr=2)	0.258	0.147	0.139
	(0.168)	(0.178)	(0.181)
No. of braking per km (tr=2)	0.074	0.164	0.160
	(0.172)	(0.179)	(0.182)
Average speed at the beginning of accelerations	−0.003	−0.001	−0.002
	(0.012)	(0.013)	(0.013)
Average speed difference after acceleration	−0.002	0.001	0.002
	(0.006)	(0.006)	(0.006)
Average speed difference after braking	−0.003	0.008	0.010
	(0.017)	(0.018)	(0.018)
Average speed at the beginning of braking	0.002	0.002	0.003
	(0.007)	(0.007)	(0.007)
McFadden Pseudo R2	0.033	0.036	0.038
Observations	5,690	5,690	5,690
Akaike Inf. Crit.	3,001.268	2,685.674	2,609.516
Notes:			***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 2.18: Coefficients of fitted probit models for liability claim submission. Comparison of different minimum number of daily driving log observations.

	submission of a liability claim> 10CHF in 2016				
Minimum no. of observations	0	10	20	30	50
Traditional Insurance Variables					
Sex of driver: female	0.026	0.029	0.028	0.037	0.048
	(0.057)	(0.059)	(0.059)	(0.060)	(0.061)
Age of driver	−0.054***	−0.053***	−0.051***	−0.051***	−0.047**
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Vehicle: age	0.020***	0.020***	0.020***	0.020***	0.021***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
Vehicle: horsepower	−0.000	−0.000	−0.000	0.000	−0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Vehicle: price	−0.000	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Vehicle: mileage	−0.001	−0.001	−0.001	−0.001	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.029	−0.027	−0.025	−0.022	−0.013
	(0.081)	(0.083)	(0.083)	(0.084)	(0.085)
Recent change of address: true	0.118*	0.119*	0.120*	0.122*	0.128*
	(0.070)	(0.070)	(0.070)	(0.071)	(0.072)
Bonus Malus Score TPL	−0.013	−0.011	−0.012	−0.012	−0.010
	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)
No. of years without first contract	0.031*	0.030*	0.028	0.028	0.028
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
No. of previous mobility claims	−0.040	−0.040	−0.041	−0.044	−0.040
	(0.071)	(0.071)	(0.071)	(0.072)	(0.072)
Telematics-Based Predictors					
Average distance per day (active)	0.004***	0.004***	0.005***	0.005***	0.004***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.115***	0.112***	0.113***	0.113***	0.125***
	(0.027)	(0.027)	(0.027)	(0.028)	(0.029)
Average speed (type 1)	−0.002	−0.002	−0.002	−0.002	−0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	−0.483**	−0.449*	−0.501**	−0.473*	−0.583**
	(0.235)	(0.241)	(0.246)	(0.252)	(0.269)
Urban driving percentage (speed=50)	−0.114	−0.124	−0.057	−0.144	−0.200
	(0.298)	(0.300)	(0.305)	(0.312)	(0.324)
Night driving percentage	0.586	0.589	0.436	0.561	0.591
	(0.441)	(0.450)	(0.459)	(0.465)	(0.477)
Percentage of driving above speed limit	0.617*	0.610*	0.541	0.575	0.629*
(type 1 adjusted)	(0.346)	(0.348)	(0.352)	(0.358)	(0.369)
Average speeding (weighted)	0.023***	0.022**	0.022**	0.022**	0.024***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit	−0.395	−0.382	−0.376	−0.381	−0.429
in Urban areas (type 1 adjusted)	(0.289)	(0.292)	(0.295)	(0.299)	(0.307)
Average speeding in urban areas (relative)	−0.105	−0.107	−0.082	−0.111	−0.128
	(0.108)	(0.109)	(0.110)	(0.113)	(0.116)
No. of accelerations per km (tr=2)	0.151	0.147	0.101	0.123	0.115
	(0.177)	(0.178)	(0.182)	(0.186)	(0.191)
No. of braking per km (tr=2)	0.144	0.164	0.158	0.136	0.153
	(0.172)	(0.179)	(0.181)	(0.185)	(0.192)
Average speed at the beginning of accelerations	−0.001	−0.001	−0.001	−0.005	0.000
	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
Average speed difference after acceleration	0.001	0.001	0.001	−0.002	−0.001
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.005	0.008	0.011	0.011	0.013
	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
Average speed at the beginning of braking	0.002	0.002	0.003	0.004	0.003
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
McFadden Pseudo R2	0.037	0.036	0.037	0.037	0.037
AIC/No. of obs	0.468	0.472	0.473	0.47	0.474
Observations	5,777	5,690	5,613	5,516	5,229
Akaike Inf. Crit.	2,690.415	2,685.674	2,654.028	2,590.371	2,479.813

Notes:
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

2.4 Conclusion

In the context of auto insurance, telematics datasets are a rich source of information that allow the study of individual driving behavior at a high level of granularity. In this paper, we take advantage of a telematics-based auto insurance contract provided by a Swiss insurer to young drivers to study whether telematics data helps to identify driving pattern related risk factors in a systematic way, and whether selection effects can be identified. Our results regarding risk factors suggest that aggregate measures of risk exposure, such as the number of journeys per day and the average distance per day, are important risk factors, whereas most of the granular driving style related variables such as e.g. average speed difference after acceleration are not found to be significant risk factors.

With respect to risk factors and contract choice, we observe that a higher number of journeys per day is positively correlated with the purchase of collision cover, but that average distance per day is negatively correlated with the purchase of collision cover, whereas both are statistically significant risk factors increasing liability claim submission. These results suggest that multiple dimensions of private information may be interacting. However, the interpretation of results has to be done with extreme caution since the sample consists of drivers that are younger than 26 years of age and thus are relatively inexperienced. In particular, the data does not allow to analyze how well the policyholders actually understand their risk. Thus, while the telematics data is important to identify risk factors, the results for our dataset with young drivers do not allow the interpretation that the telematics data helps to identify relevant asymmetric information in the form of the policyholder having superior information about risk type or preferences that affects selection in a systematical manner for insurer profits.

The design of the telematics contract and the available data does not enable us to analyze whether policyholders react to information about their driving patterns and style. In particular, with the telematics contract policyholders receive primarily information with color-coded relative rankings about the elevated g-force events (e.g. accelerations, harsh brakings, cornerings) per km driven. However, we do not have data on whether this information is actually checked. A very interesting avenue for future research are the incentive effects of providing information gathered with the drive recorder, i.e. to analyze whether and how telematics data and the corresponding provision of information and incentive design affect driving behavior and consequently accident occurence.

3. Into Twilight: on the Perils of Late Visits to Unfamiliar Destinations.

Empirical papers analyzing vehicle accident risk come to the robust conclusion that controlling for various factors such as experience, distance driven and speed, number of journeys per day has a statistically significant impact on the hazard level. I used driving logs from a telematics device to search for possible explanations behind this phenomenon and to explore whether more detailed information on visited locations could improve risk modeling. I discovered that frequent driving on unfamiliar roads increases the likelihood of accident involvement. Controlling for driving behavior, two other potentially important risk factors: bad light conditions and fatigue do not influence accident hazard. These insights can be incorporated into risk classification schemes to promote safe driving.

3.1 Introduction

In 2016, 17'799 accidents occurred on Swiss roads, of which 290 led to casualties (BFS (2016)). Despite a steady decline, motor vehicle accidents remain one of the leading causes of mortality among young adults in Switzerland and in other countries. This motivates an active research into accident causes and efficient intervention measures. Relevant empirical studies traditionally relied on various cross-sectional datasets (accident statistics, socio-economical variables), surveys, interviews and laboratory experiments. These sources frequently have one of the following limitations. First, they do not always provide a sufficient level of detail for a rigorous inquiry. Second, as is the case for surveys and interviews, information therein might be imprecise and unreliable. As more data, both in volume and versatility, becomes available, the field is invigorated since a lot of relevant situational factors can be captured and analyzed.

In the late 90s following the developments in telematics technologies, insurance companies introduced *usage based* insurance policies. Under these policies, the vehicle's motion is monitored by an on-board device, often in exchange for a premium reduction or other value-added services. Currently, a typical observation contains information on coordinates, time stamp, point speed and, under certain conditions, information about the vehicle's acceleration. These driving logs, also referred to as Telematics Data, have been used in various driving behavior studies (Muermann et al. (2019), Ippisch (2010), Ayuso et al. (2016)).

Previous contributions by Muermann et al. (2019) and Sycheva et al. (2019) indicate that controlling for speed and distance driven, additional journeys are correlated with accident occurrence. In this chapter, I analyzed driving logs, aggregated on journey level, to shed more light on this phenomenon. My analysis delved deeper into three different aspects: *where*, *when* and *how long* did the policyholder drive? With regard to the first question, my results suggest that controlling for both traditional risk classification variables and telematics-based predictors, driving on an unknown road represents a higher risk situation.

I used information about journey's temporal patterns to infer light conditions during the trips. Light conditions are an important risk-relevant aspect of the environment, yet commonly used premium calculation schemes employ unreliable proxies to incorporate this information. It is a common practice to define night driving based on fixed hours rather than tie it to sunrise and sunset times. I augmented the information on traveling locations with exact sunrise/sunset times to accurately infer driving time during the night, twilight, following the sunrise/before the sunset. Created predictors do not significantly improve my risk model.

Statistics on the trip duration reflect exposure to another risk factor: fatigue. I aggregated it over the study period and tested whether higher frequency of long journeys translates into higher accident risk. I do not find a statistically significant link, yet this result could be attributed to the fact that policyholders rarely take long trips and consequently are not exposed to this risk factor.

3.2 Related Literature

This paper primarily contributes to the literature studying motor vehicle accident risk. Accidents are caused by an interplay of numerous factors therefore the research effort spans over across variety of disciplines, from engineering and infrastructure design to economics and psychology.

Milton and Mannering (1998) study the impact of road geometry on accident frequency, Brodsky and Hakkert (1988) and Bergel-Hayat et al. (2013) quantify the negative impact of bad weather conditions on accident count, whereas Newstead and D'Elia (2007) establish that certain vehicle characteristics, such as color, increase the likelihood of accident involvement. These factors exacerbate that risk, but the main cause behind the majority of accidents is the driver. Widely accepted driver-related risk factors include age, mobile phone usage (Lipovac et al. (2017)), biases in risk perception (DeJoy (1989)), the presence and behavior of other passengers (Vollrath et al. (2002)) and substance abuse (Peck et al. (2008)).

Most of the previous contributions had to rely on self-reports, proxies or aggregated information to capture risk relevant characteristics. Detailed driving logs represent a qualitative shift in accident research, allowing the scientist to examine the impact of various behavioral patterns. Our results in Sycheva et al. (2019) suggest that a deeper analysis of a policyholder's vehicle traveling patterns might be a worthy endeavor. In this paper, I aggregate the logs on the journey-policyholder level and examine spatio-temporal characteristics of the policyholders' journeys.

Based on the created dataset I can quantify route familiarity. In the previous contributions there is no consensus regarding its impact on the accident risk. Some researchers support the intuitive conclusion that driving on a less familiar road constitutes a higher risk situation. Indeed, drivers face unknown junctions, road conditions and potentially different road systems. The situation is further exacerbated by the fact that drivers frequently rely on a navigator to find their destination and consequently pay less attention to the road and surrounding area. This assumption is supported by results of Yannis et al. (2007). They show that foreign drivers in Greece, who are not permanent residents, face higher risk on the country's roads. On the other hand, data from Chen et al. (2005) suggests that the majority of accidents occurs close to home. Charlton and Starkey (2013) explain this phenomenon by inattention blindness. Familiarity might also breed overconfidence: Intini et al. (2016) finds that study subjects tend to increase their speed and be less compliant with speed limits while driving on familiar roads.

Timestamps and coordinates of starting and end points of journey provide sufficient information to infer light conditions during the trip. Wanvik (2009) shows that improving road lights system leads to a decrease in accident count. However, it might be not sufficient to completely offset the risk, as Uttley and Fotios (2017) discover that the likelihood of accidents involving pedestrians still increases after dark. Even with access to driving logs, the current empirical evidence on the impact of bad light conditions is not conclusive: Muermann et al. (2019) do not find a statistically significant association between night driving and subsequent claim submission, whereas Ayuso et al. (2016) conclude that frequent night driving decreases the time until the first accident.

Journey duration is a reliable proxy for another risk factor: fatigue. Several studies examine the effect of fatigue on driving behavior, and find statistical evidence of task performance degradation. (Lyznicki et al. (1998), Brown (1994)). To be more precise, fatigue increases reaction times, reduces attention and information processing capability (Dinges (1995), Philip et al. (2005)). To mitigate this risk, under Swiss law, commercial drivers are required to take at least a 45 minute break after driving for a maximum 4.5 hours.[1]

Policymakers and economists have contemplated various intervention measures to reduce accident risk and provide incentives for safe driving. Some of their efforts were fruitful. Empirical studies demonstrate that accident fatality rates can be reduced by increasing gas (Ahangari et al. (2014)) and fuel prices (Chi et al. (2010)) or increasing

[1] Verordnung ueber die Arbeits- und Ruhezeit der berufsmaessigen Motorfahrzeugfuuehrer und -fuehrerinnen 3. Abschnitt: Lenkzeiten, Arbeitszeiten, Pausen, Ruhezeiten Art. 8

the punishments for traffic infringements. Insurance risk-based pricing is also a potent mechanism for accident prevention. Dionne et al. (2011) show that experience rating is effective in inducing high effort, whereas results of Hultkrantz and Lindberg (2011) suggest that adjusting the risk premium following speed violations helps to reduce this type of behavior. The results presented in this chapter can also be used to better target such intervention efforts.

The remainder of the chapter is organized as follows. Section 3.3 provides detailed information on insurance and telematics datasets. Section 3.4 details the creation of journey-based driving profile. Regression results are reported in Section 3.6, their practical implications are the topic of Section 3.7. Section 3.8 concludes.

3.3 Background and Data

I extracted information from three datasets introduced in Section 2.2. The data was provided by a large Swiss insurance carrier. The company offered young drivers a *usage-based* vehicle insurance policy. In exchange for a premium reduction, policyholders were obliged to install a telematics device, which collected detailed information about the vehicle's motion. During the first insurance year, the size of the reduction was fixed, in subsequent periods it depended on the driving style. Bad driving scores decreased the size of the reduction. However, all policyholders were still financially better off than without a *usage-based* policy.

Available *usage-based* policies fall into one of two main categories: premiums of *pay-as-you-drive* policies are determined primarily by the quantity of driving, whereas the price of a *pay-how-you-drive* contract aims to reflect driving quality. Policyholders in our dataset held a *pay-how-you-drive* contract. Risk-based scoring proceeded as follows: a telematics device monitored vehicle's acceleration along three principal axes. The device was alerted every time the values exceed a predefined threshold, a situation I refer to as an “elevated g-force event". Depending on the axis and duration, all elevated g-force events were classified into three groups: accelerations, braking and cornering.

Table 3.1: Summary statistics: Policyholder's and vehicle's characteristics.

	mean	median	std	Sample quantiles					
				min	5%	25%	75%	95%	97.5%
Policyholder's characteristics									
Age of driver (years)	22.49	22.0	5.10	17.0	18.0	20.0	24.00	26.00	31.0
Bonus Malus score for liability claims	13.39	13.0	1.19	9.0	12.0	13.0	14.00	15.00	17.0
No. of years without first contract	3.19	3.0	3.41	0.0	0.0	1.0	5.00	7.00	8.0
No. of previous mobility claims	0.12	0.0	0.45	0.0	0.0	0.0	0.00	1.00	1.0
Vehicle's characteristics									
Price (CHF)	31819	28600	17647	8900	17700	22100	40300	62330	72622
Age (years)	6.92	6.0	5.55	0.0	0.0	2.0	11.00	17.00	19.0
Horsepower	131.37	115.0	57.04	45.0	68.0	90.0	150.00	241.00	295.0
Weight (kg)	1286.66	1257.0	236.48	785.0	940.0	1115.0	1420.00	1688.25	1816.5
Mileage	79.83	71.5	68.55	1.0	1.0	20.0	121.25	199.00	220.0

Every event was assigned a weight, depending on its type, severity, vehicle speed, time of the day (night, day, rush hour) and road network type (urban, highway, other). The weighted sum of all elevated g-force events determined the premium reduction in the subsequent insurance year.

The installation of a drive recorder also offered non-financial benefits. First, it could be used for a fast recovery of stolen vehicles. Second, it would alert the company in case of an emergency. Lastly, driving logs prior to an accident can shed light on what happen, who is responsible and consequently greatly simplify claims handling.

Offering *usage-based* insurance policies has the following potential benefits for the insurer. First, telematics data is a promising tool to refine risk classification scheme and improve the profit margin. Second, such contract optimizes the risk pool by a) attracting low-risk drivers b) pushing away high-risk drivers c) motivating current clients to exert more preventive efforts (Litman (1997)).

3.3.1 Data

The analysis is based on three datasets. The policyholders' driving logs are stored in the Telematics Dataset introduced in Section 2.2.3. I combined this information with insurance data using a unique identification number. This data comprises two tables: Contract Data contains all information traditionally used for insurance pricing, Claims Data stores all claims submitted by policyholders. For more detailed information about these datasets I refer the reader to Section 2.2.1 and Section 2.2.2.

In the subsequent risk analysis I used all telematics-based predictors, created in Chapter 2. The main creation steps are summarized in Figure 3.1. The procedure is described Section 2.2.3 and Section 2.2.4.

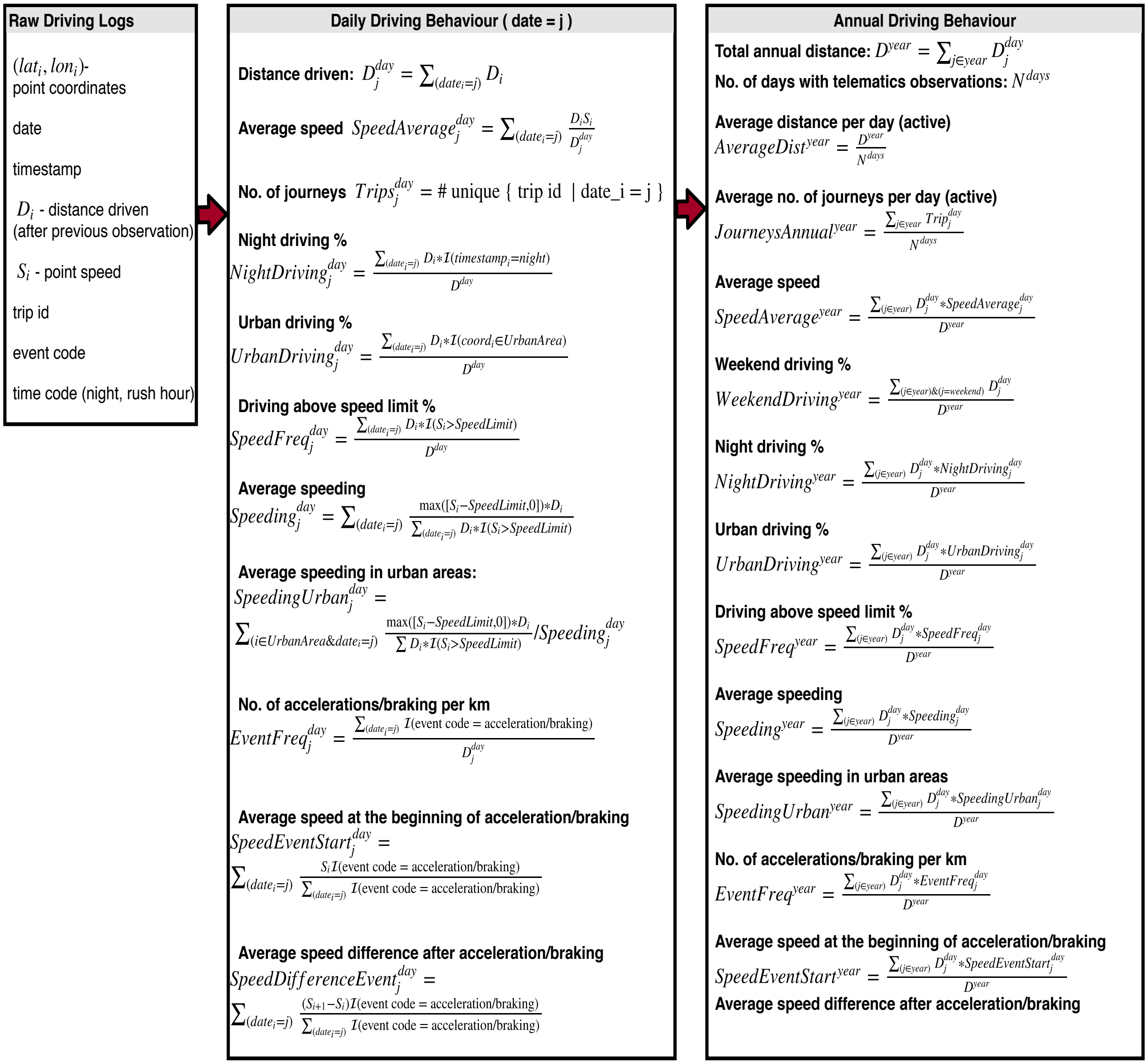
The sample for the risk analysis slightly differs from the one in Chapter 2. First, due to the reasons discussed in Section 3.4.3 I excluded policyholders with less than 20 days of available driving logs in 2016. Second, I did not exclude 213 policyholders older than 26. Table 3.1 and Table 3.2 report the summary statistics of predictors corresponding to policyholders in our sample. Provided that a policyholder used his vehicle, he drove on average 53 km per day[2] and made 3.92 journeys. On average 30% of driving time corresponded to the weekends and 6% of the aggregate distance was driven during the night. Speed violations were both frequent and severe: on average, every fifth kilometer was driven with a certain degree of speed violation, whose relative magnitude is 24% higher is urban areas. Approximately 2 accelerations and 3.3 braking were registered per every 10 km driven.

[2] This value is not directly comparable to the national swiss average of 36.8 km (see Federal statistics office: Mobility and transport) since the number from the Federal Statistics Office is a) an average over entire observational period b) does not include driving abroad.

Table 3.2: Summary statistics of telematics based predictors

	mean	median	std	Sample quantiles					
				min	5%	25%	75%	95%	97.5%
General characteristics									
Average distance per day (active)	52.94	49.10	23.93	2.60	22.44	36.10	65.10	95.73	109.12
No. of journeys per day (active)	3.97	3.80	1.07	1.26	2.59	3.23	4.49	5.97	6.60
Journey time and location									
Weekend driving percentage	0.30	0.29	0.12	0.00	0.12	0.22	0.37	0.53	0.59
Urban driving percentage (speed=50)	0.31	0.29	0.14	0.00	0.12	0.21	0.39	0.56	0.63
Night driving percentage	0.06	0.04	0.06	0.00	0.00	0.01	0.08	0.18	0.22
Speed and Speeding									
Average speed (type 1)	62.07	61.75	13.18	2.61	41.65	54.18	70.55	83.52	88.04
Percentage of driving above speed limit (type 1 adjusted)	0.21	0.20	0.11	0.00	0.06	0.13	0.28	0.41	0.45
Average speeding (weighted)	12.28	11.75	3.46	2.59	7.83	9.90	14.01	18.56	20.56
Percentage of driving above speed limit in urban areas (type 1 adjusted)	0.32	0.31	0.13	0.00	0.13	0.23	0.40	0.57	0.63
Average speeding in urban areas (relative)	1.24	1.16	0.31	0.18	0.93	1.05	1.33	1.80	2.03
Characteristics of elevated g-force events									
No. of accelerations per km (tr=2)	0.20	0.16	0.18	0.00	0.03	0.08	0.27	0.55	0.69
No. of braking per km (tr=2)	0.33	0.29	0.20	0.00	0.09	0.20	0.42	0.69	0.82
Average speed at the beginning of accelerations	10.98	10.50	2.57	0.00	7.84	9.32	12.13	16.15	17.59
Average speed difference after acceleration	14.26	13.71	6.46	0.00	4.43	9.68	18.32	25.80	28.31
Average speed difference after braking	-14.37	-14.37	2.03	-23.22	-17.56	-15.65	-13.10	-11.29	-10.68
Average speed at the beginning of braking	36.86	36.87	5.82	0.00	27.90	33.27	40.57	46.08	47.88

Figure 3.1: Creation of telematics-based predictors.



Raw Driving Logs

(lat_i, lon_i)- point coordinates

date

timestamp

D_i - distance driven (after previous observation)

S_i - point speed

trip id

event code

time code (night, rush hour)

Daily Driving Behaviour (date = j)

Distance driven: $D_j^{day} = \sum_{(date_i=j)} D_i$

Average speed $SpeedAverage_j^{day} = \sum_{(date_i=j)} \frac{D_i S_i}{D_j^{day}}$

No. of journeys $Trips_j^{day}$ = # unique { trip id | date_i = j }

Night driving %

$$NightDriving_j^{day} = \frac{\sum_{(date_i=j)} D_i * \mathcal{I}(timestamp_i = night)}{D^{day}}$$

Urban driving %

$$UrbanDriving_j^{day} = \frac{\sum_{(date_i=j)} D_i * \mathcal{I}(coord_i \in UrbanArea)}{D^{day}}$$

Driving above speed limit %

$$SpeedFreq_j^{day} = \frac{\sum_{(date_i=j)} D_i * \mathcal{I}(S_i > SpeedLimit)}{D^{day}}$$

Average speeding

$$Speeding_j^{day} = \sum_{(date_i=j)} \frac{\max([S_i - SpeedLimit, 0]) * D_i}{\sum_{(date_i=j)} D_i * \mathcal{I}(S_i > SpeedLimit)}$$

Average speeding in urban areas:

$$SpeedingUrban_j^{day} = \sum_{(i \in UrbanArea \& date_i=j)} \frac{\max([S_i - SpeedLimit, 0]) * D_i}{\sum D_i * \mathcal{I}(S_i > SpeedLimit)} / Speeding_j^{day}$$

No. of accelerations/braking per km

$$EventFreq_j^{day} = \frac{\sum_{(date_i=j)} \mathcal{I}(\text{event code = acceleration/braking})}{D_j^{day}}$$

Average speed at the beginning of acceleration/braking

$$SpeedEventStart_j^{day} = \sum_{(date_i=j)} \frac{S_i \mathcal{I}(\text{event code = acceleration/braking})}{\sum_{(date_i=j)} \mathcal{I}(\text{event code = acceleration/braking})}$$

Average speed difference after acceleration/braking

$$SpeedDifferenceEvent_j^{day} = \sum_{(date_i=j)} \frac{(S_{i+1} - S_i) \mathcal{I}(\text{event code = acceleration/braking})}{\sum_{(date_i=j)} \mathcal{I}(\text{event code = acceleration/braking})}$$

Annual Driving Behaviour

Total annual distance: $D^{year} = \sum_{j \in year} D_j^{day}$

No. of days with telematics observations: N^{days}

Average distance per day (active)

$$AverageDist^{year} = \frac{D^{year}}{N^{days}}$$

Average no. of journeys per day (active)

$$JourneysAnnual^{year} = \frac{\sum_{j \in year} Trip_j^{day}}{N^{days}}$$

Average speed

$$SpeedAverage^{year} = \frac{\sum_{(j \in year)} D_j^{day} * SpeedAverage_j^{day}}{D^{year}}$$

Weekend driving %

$$WeekendDriving^{year} = \frac{\sum_{(j \in year)\&(j = weekend)} D_j^{day}}{D^{year}}$$

Night driving %

$$NightDriving^{year} = \frac{\sum_{(j \in year)} D_j^{day} * NightDriving_j^{day}}{D^{year}}$$

Urban driving %

$$UrbanDriving^{year} = \frac{\sum_{(j \in year)} D_j^{day} * UrbanDriving_j^{day}}{D^{year}}$$

Driving above speed limit %

$$SpeedFreq^{year} = \frac{\sum_{(j \in year)} D_j^{day} * SpeedFreq_j^{day}}{D^{year}}$$

Average speeding

$$Speeding^{year} = \frac{\sum_{(j \in year)} D_j^{day} * Speeding_j^{day}}{D^{year}}$$

Average speeding in urban areas

$$SpeedingUrban^{year} = \frac{\sum_{(j \in year)} D_j^{day} * SpeedingUrban_j^{day}}{D^{year}}$$

No. of accelerations/braking per km

$$EventFreq^{year} = \frac{\sum_{(j \in year)} D_j^{day} * EventFreq_j^{day}}{D^{year}}$$

Average speed at the beginning of acceleration/braking

$$SpeedEventStart^{year} = \frac{\sum_{(j \in year)} D_j^{day} * SpeedEventStart_j^{day}}{D^{year}}$$

Average speed difference after acceleration/braking

3.4 Journey-based Risk Profile

3.4.1 Overview

In Section 3.4.3 I extracte and preprocessed information about individual journeys from the driving logs. I merge it with corresponding spatial cluster labels, aggregate this data over 2016 and mine the resulting dataset for accident risk factors. I focus on three aspects of journey profile.

- *Where did the policyholder drive?* In Section 3.4.3 I evaluate how diverse the set of locations visited by the driver is and use it as a proxy for route familiarity.
- *When did the policyholder drive?* In Section 3.4.4, I use the start and end times of individual journeys to see whether and how long the person was driving under bad light conditions.
- *How long did the policyholder drive?* In Section 3.4.5 compute the frequency of long journeys.

3.4.2 Trip Data

I summarized information about all journeys contained in the telematics data set. Each policyholder's journey was assigned a unique identification number [3] For every trip, I retrieved the following basic information: start and end time, locations, stop duration, distance driven. I excluded 1) policyholders with less than a week of observations 2) journeys shorter than 10 seconds. The final dataset contains information about 8'720'447 journeys for 6'708 policyholder.

[3] Single journey can be extracted based on two indicators: a) every trip is assigned a unique (for each policyholder) identification number b) the original dataset records the times when the engine is switched on and off. Both procedures yield same results.

Table 3.3: Summary statistics of Trip Data

Variable	mean	median	std			Sample quantiles		
				5%	25%	75%	95%	97.5%
No. of journeys per policyholder	1298.37	1081.0	1015.3	117.0	466.0	1867.5	3243.0	3720.75
Timespan (days)	425.44	369.0	282.96	51.8	169.0	641.0	939.0	985.0
Usage frequency	0.72	0.76	0.19	0.34	0.62	0.86	0.96	0.98
No. of days per policyholder	1	298.66	263.0	213.19	111.0	40.0	704.0	770.75
No. of journeys per day (active)	4.35	4.0	2.65	1.0	2.0	6.0	9.0	11.0
Distance per journey	12.37	6.02	20.52	0.39	2.09	14.69	42.81	60.83
Stop duration (minutes)	390.0	105.0	2743.0	1.0	15.0	536.0	1247.0	2177.0

Table 3.3 reports summary statistics of the basic trip characteristics. People in our sample used their vehicle 5 days a week and took 4 trips per day on average . The values are obtained without excluding information about a) policyholders with a small number of journeys b) short trips (5 m) c) journeys following a very short stop.

I performed hierarchical clustering to group together stops, that corresponded to a single destination. Cutoff-values were selected primarily based on heuristics I had developed, following a short literature review. Further details are relegated to Appendix 5. I discovered 970'486 spatial clusters representing locations visited by policyholders. For every destination, I aggregated information over all visits to obtain statistics on *stop duration*, *arriving time arriving date* and to compute *visit frequency (general/weekday/weekend)*.

3.4.3 Route Familiarity

Consider a policyholder that traveled on **N** different routes over the study period. I assessed his average familiarity with the road by answering two complementary questions: a) *What is the aggregate percentage of journeys corresponding to* ***K*** *most popular routes ?* b) *What percentage of unique routes accounts for* ***M*** *percent of journeys?* These two measures are closely related, yet the latter provides more information about the number of times a rarely-visited route is taken. It is not clear a priori what measure better suits our purpose.

Driving logs are sufficient to recreate the vehicle's trips with a very high precision. Analyzing coordinates of every observation in the driving logs is computationally expensive and not guaranteed to yield significant improvements in terms of final premium adjustment. To explore the trade-off between computational burden and precision, I adopted three approaches to represent a route.

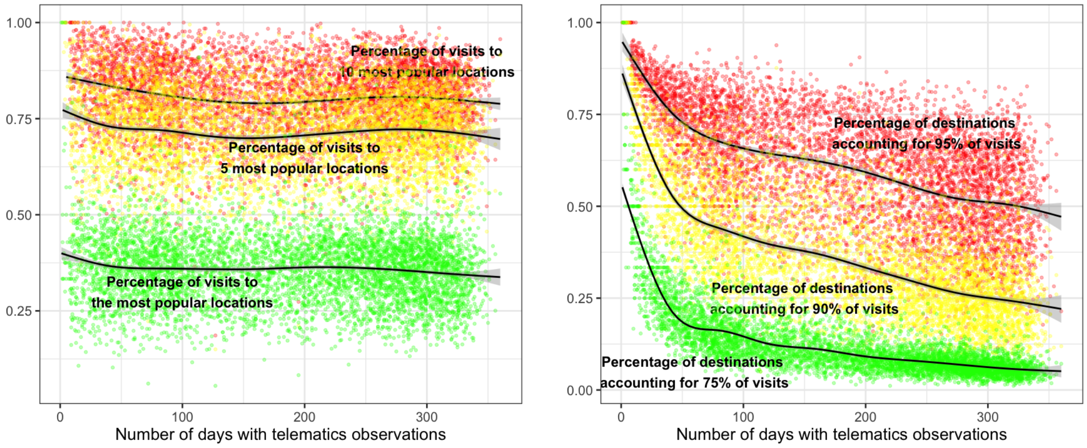
The simplest approach is to proxy a route by its starting point. If a small number of locations accounts for the majority of visits, it is likely that the policyholder was driving on familiar roads most of the time. The reverse is probably the case when person's itinerary featured a lot of rarely-visited destinations. Table 3.4 shows that on average, over one third of the trips started in the most frequently visited location, and in 25% of the cases, the aggregate number of visits to 5 most popular locations accounted for over 79% of all trips.

The second approach is to retain both start and end point of a journey. Resulting predictors will not capture the fact that routes between different destinations might overlap, whereby only a minor section of 'unfamiliar" route is actually unfamiliar. Yet this method offers a reasonable trade-off between accuracy and computational burden. A further modification is to separately count trips in different directions. This would allow to account for the fact that a driver primarily focuses on one part of the road and thus certain important details on the other part, such as side roads, escape his attention. If I did not differentiate between driving directions, according to Table 3.4, approximately 24% of the journeys corresponded to the most popular routes. Policyholders' traveling habits varied within our sample: for 5% of the drivers 20 routes accounted for not more than 44% of visits, whereby for other 5%, 10 routs corresponded to at least 82% of the trips. When I separately counted trips in different directions, most popular route accounted on average for 13% of all journeys. [4]

The created predictors can be divided into six different groups, based on two criteria: 1) what information is used to represent a route 2) how the familiarity with these routes is measured. These variables, however, share a similar limitation: they do not capture

[4] Contrasting sample statistics corresponding to these two types of predictors suggests that some routes are predominantly driven in one direction.

Figure 3.2: Stability of concentration measures with respect to observation period duration



information on the visits, not reflected in the driving logs. There are several plausible scenarios where this is the case. First, prior to drive recorder installation policyholder has been driving for a long time.T thus routes that do not often appear in the telematics dataset are not necessarily unfamiliar. Second, the policyholder might regularly use a second vehicle. Third, the policyholder has visited the destinations by using other modes of transportation or as a passenger. I do not anticipate, however, that any of these scenarios is likely to introduce a systematic bias into our predictors. [5]

As in Section 2.2.4, the length of driving logs corresponding to 2016 varied significantly between policyholders. This raised two related questions: a) Do created predictors depend on the observation timespan b) If yes, how many days are sufficient to obtain reliable estimates of route familiarity. Figure 3.2 provides scatter plots of selected variables against the number of days with available telematics observations. It indicates that the first type of concentration measures are rather stable, however second group exhibit a distinctive downward trend. Its magnitude decreases sharply, once over 50 days of observations are collected. This has implications for our risk

[5] First scenario: a) drivers are young thus it is likely their first vehicle b) even if not, if they did not use the road for a long time, certain details are likely to slip their memory. Second scenario: vehicle usage frequency on is average 80% which means that policyholders in our sample drive between 5 and 6 days per week. Even if they occasionally use another vehicle (for instance of their parents) this is not likely to affect the general picture. Third scenario: arguably a lot of relevant details might escape person's attention unless he is the driver. Previous argument about usage frequency is also valid here

Table 3.4: Summary statistics: Route familiarity measures

	mean	median	std	Sample quantiles					
				min	5%	25%	75%	95%	97.5%
% of short journeys	0.09	0.08	0.04	0.01	0.04	0.06	0.11	0.17	0.20
Percentage of visits to:									
the most popular location	0.36	0.36	0.09	0.06	0.22	0.30	0.41	0.49	0.53
5 most popular locations	0.72	0.72	0.11	0.15	0.54	0.65	0.79	0.88	0.91
10 most popular locations	0.81	0.81	0.09	0.19	0.64	0.75	0.87	0.94	0.96
20 most popular locations	0.87	0.88	0.07	0.23	0.74	0.83	0.93	0.97	0.98
Percentage of clusters accounting for:									
for 75% of visits	0.11	0.09	0.07	0.01	0.04	0.06	0.14	0.26	0.30
for 90% of visits	0.35	0.32	0.15	0.03	0.15	0.23	0.44	0.64	0.68
for 95% of visits	0.60	0.61	0.15	0.09	0.33	0.48	0.72	0.82	0.85
Percentage of journeys on (undirected):									
the most popular route	0.24	0.21	0.13	0.02	0.08	0.14	0.31	0.49	0.56
5 most popular routes	0.48	0.47	0.15	0.06	0.25	0.37	0.58	0.74	0.79
10 most popular routes	0.58	0.58	0.15	0.08	0.34	0.47	0.68	0.82	0.86
20 most popular routes	0.68	0.68	0.14	0.12	0.44	0.58	0.78	0.90	0.93
Percentage of routes (undirected) accounting for:									
75% of journeys	0.25	0.23	0.13	0.01	0.08	0.15	0.32	0.50	0.55
90% of journeys	0.61	0.63	0.14	0.06	0.34	0.52	0.72	0.80	0.82
95% of journeys	0.80	0.82	0.09	0.19	0.62	0.76	0.86	0.90	0.91
Percentage of journeys (directed):									
the most popular	0.13	0.12	0.07	0.01	0.05	0.08	0.17	0.26	0.30
5 most popular routes	0.38	0.36	0.15	0.04	0.17	0.27	0.47	0.65	0.70
10 most popular routes	0.49	0.48	0.15	0.06	0.26	0.38	0.59	0.75	0.80
20 most popular routes	0.59	0.59	0.15	0.08	0.36	0.49	0.70	0.84	0.87
Percentage of routes (directed) accounting for:									
75% of journeys	0.32	0.29	0.15	0.02	0.10	0.20	0.44	0.60	0.63
90% of journeys	0.69	0.71	0.12	0.08	0.43	0.63	0.78	0.84	0.86
95% of journeys	0.84	0.86	0.07	0.24	0.72	0.82	0.89	0.92	0.93

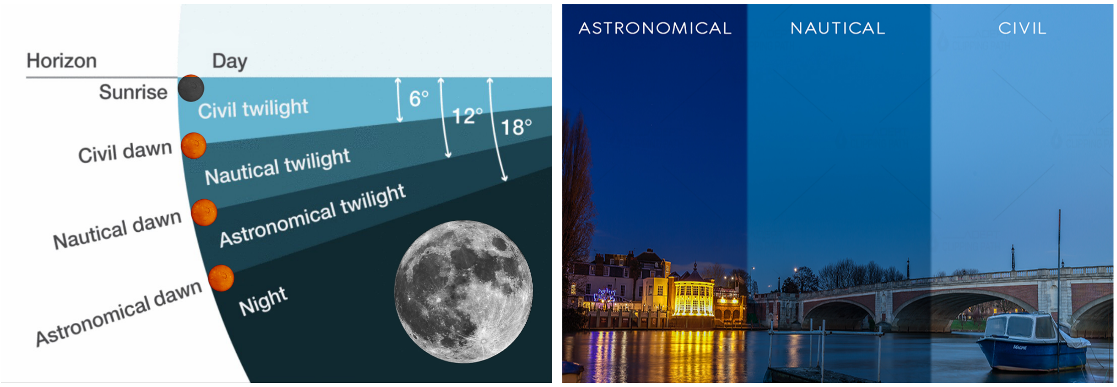
analysis: to reliably estimate the impact of this type of variables, our sample should be restricted to policyholders with sufficient number of daily observations. With that in mind, I decided to fit all subsequent regression models based on data from drivers with at least 20 days of driving records in 2016.

3.4.4 Night and Twilight Driving

To accurately incorporate information about light conditions, I downloaded data on sunrise, sunset and twilight times on a daily basis for 6 locations, scattered around Switzerland [6] and 44 locations abroad (41 in Europe, 2 in Africa and 1 in Asia)[7]. I matched this data to individual journeys in the Trip Data, based on dates and on the coordinates of starting and end points. The distance to the closest locations with known sunrise and sunset times reflects the precision of our predictors. I excluded 0.3% of journeys from subsequent computations, since corresponding information was too imprecise. In the remaining dataset, average distance is 27.33 km and does not exceed 130 km. To put this in perspective, consider two towns 135 km apart on approximately same latitude: Basel and St. Gallen. The time difference between the start of twilight vary between 3 and 5 minutes.

I estimated the fraction of time driven during the night, astronomical, nautical,

Figure 3.3: Twilight Phases and corresponding light conditions.



[6] Zurich, Basel, Geneva, Lugano, St Gallen and Davos

[7] France: Paris, Lyon, Strasbourg, Marseille, Portugal: Lisbon, Proto, Albufeira, Coimbra, Spain: Barcelona, Bilbao, Malaga, Italy: Bari, Milan, Naples, Rome, Venice, Verona, Florence, Slovakia: Ljublijana, Croatia: Dubrovnik, Zagreb, Split, Zadar, Bosnia-Herzegovina: Sarajevo, Kosovo: Prizren, Serbia: Belgrade, Albania: Tirana, Macedonia: Skopje, Austria: Vienna, Graz, Innsbruck, Salzburg, Germany: Frankfurt, Stuttgart, Munich, Berlin, Hannover, Cologne, Netherlands: Amsterdam, Belgium: Brussels, Turkey: Antalya, Benin: Porto-Novo, Guinea: Conakry

Table 3.5: Summary statistics: frequency of driving under bad light conditions

	mean	median	std	Sample quantiles					
				min	5%	25%	75%	95%	97.5%
Percentage of driving time during:									
Night (2)	0.15	0.13	0.07	0.0	0.05	0.09	0.18	0.28	0.32
Twilight	0.16	0.16	0.07	0.0	0.08	0.12	0.20	0.28	0.32
Evening twilight	0.08	0.07	0.04	0.0	0.03	0.05	0.10	0.16	0.18
Morning twilight	0.09	0.08	0.05	0.0	0.01	0.05	0.12	0.18	0.21
After sunrise	0.02	0.01	0.01	0.0	0.00	0.01	0.02	0.05	0.06
Before sunset	0.03	0.03	0.02	0.0	0.01	0.02	0.04	0.06	0.08
Astronomical twilight	0.06	0.05	0.03	0.0	0.02	0.04	0.07	0.11	0.13
Nautical twilight	0.06	0.05	0.03	0.0	0.02	0.04	0.07	0.11	0.13
Civil twilight	0.05	0.05	0.02	0.0	0.02	0.04	0.06	0.09	0.11

civil twilight[8] and fraction of driving right before (30 minutes) sunset and following (30 minutes) sunrise. I aggregated these values to obtain overall driving time during (morning/evening) twilight.

Table 3.5 presents the summary statistics of the new predictors. On average, 16% of driving time corresponded to twilight and this proportion was within a 8% and 28% range for 90% of policyholders. During around 5% of driving time, the sun was at the eye level of the policyholder. Finally, on average 15% of aggregate travel time fell between the end and start of astronomical twilight, simply referred to as *night*.

I constructed an additional set of predictors capturing during what fraction of the time with bad light conditions the driver was in an unknown area. General statistics on the new variables is reported in Table 3.6. It appears, that the conditional probability of driving in an unknown area, provided that light conditions were bad, is quite high: on average 0.39% of night driving, 0.37% of twilight driving took place on routes, with an aggregate number of trip lower than 1% of all trips over the study period.

[8] Twilight are separated into several stages depending on the solar elevation angle. The darkest phase is known as Astronomical twilight: it is the time when the sun is 12-18 degrees below the horizon. During Nautical and Civil twilight sun moves between 6 - 12 degrees and 0-6 degrees respectively. This information is illustrated in Figure 3.3.

Table 3.6: Summary statistics. Frequency of driving under bad light conditions on unfamiliar roads

	mean	median	std	Sample quantiles					
				min	5%	25%	75%	95%	97.5%
Relative proportion of driving on unfamiliar roads (less than 1% of journeys) during:									
Night	0.39	0.38	0.22	0.0	0.00	0.23	0.54	0.76	0.83
Twilight	0.37	0.36	0.19	0.0	0.01	0.23	0.50	0.70	0.77
Morning twilight	0.29	0.23	0.23	0.0	0.00	0.12	0.42	0.76	0.85
Evening twilight	0.49	0.51	0.22	0.0	0.00	0.35	0.64	0.81	0.86
After sunrise	0.46	0.44	0.32	0.0	0.00	0.20	0.68	1.00	1.00
Before sunset	0.48	0.50	0.25	0.0	0.00	0.33	0.67	0.88	1.00
Astronomical twilight	0.34	0.32	0.21	0.0	0.00	0.18	0.48	0.74	0.81
Nautical twilight	0.38	0.37	0.21	0.0	0.00	0.22	0.52	0.74	0.80
Civil twilight	0.43	0.44	0.21	0.0	0.00	0.29	0.58	0.77	0.82

3.4.5 Journey Duration

Drivers are more likely to experience fatigue a) after traveling for a long time without stops and b) during the night. The latter is captured by the predictors created in Section 3.4.4.

The stop duration between journeys can be quite short. Upon examining the data, I decided to sum the travel time if the stop between subsequent journeys did not exceed ten minutes. Philip et al. (2005) demonstrate that if non-professional drivers take three short breaks (two of 15 minutes and one of 30 minutes), their performance does not deteriorate during a 1'000 km journey. Bearing in mind that policyholders in our sample drove much less, a ten minute break is arguably sufficient to recover.

Summary statistics reported in Table 3.7 suggested that journeys exceeding a couple of hours were extremely rare: an average policyholder drove longer than 2 hours less than once per month. This is not surprising, provided that on the one hand, our sample consists of non-professional drivers and on the other, the area of country is not that large.

Table 3.7: Summary statistics: Frequency of long journeys

	mean	median	std	Sample quantiles					
				min	5%	25%	75%	95%	97.5%
Frequency (per day) of journeys exceeding:									
2 hours	0.029	0.014	0.048	0.0	0.0	0.0	0.036	0.112	0.160
3 hours	0.010	0.000	0.022	0.0	0.0	0.0	0.010	0.043	0.068
4 hours	0.004	0.000	0.012	0.0	0.0	0.0	0.003	0.021	0.033
5 hours	0.002	0.000	0.007	0.0	0.0	0.0	0.000	0.011	0.018
6 hours	0.001	0.000	0.005	0.0	0.0	0.0	0.000	0.006	0.011
7 hours	0.001	0.000	0.003	0.0	0.0	0.0	0.000	0.004	0.008
8 hours	0.000	0.000	0.002	0.0	0.0	0.0	0.000	0.000	0.004

3.5 Hypotheses

Variables created in Section 3.4.3 - Section 3.4.5 are potent proxies for several risk factors. I used predictors from Section 3.4.3 to test the following hypothesis:

(H1) *Accident risk increases if policyholder frequently drives on unfamiliar roads.*

Variables created in Section 3.4.4 allowed me to study the link between driving under bad light conditions and subsequent accident involvement. I check whether:

(H2) *Higher frequency of driving under bad light conditions (night / twilight / before sunset and after sunrise) increases accident risk;*

(H2.1) *When driving in unknown area, accident risk is further amplified by bad light conditions.*

Lastly, I incorporated predictors from Section 3.4.5 into the regression model to test whether:

(H3) *Higher frequency of long journeys increases accident risk.*

3.6 Results

I employ a probit regression to study the impact of new predictors on accident risk. The risk is expressed as a binary variable that takes a positive value if the policyholder has submitted at least 1 liability claim, with compensation exceeding 10 CHF in 2016. The base model comprises predictors described in Section **??** and Section **??**. A detailed discussion of the base model can be found in Sycheva et al. (2019).

I add characteristics derived in Section 3.4.3 to Section 3.4.5 and compare fitted models. Variables from Section 3.4.3 are included separately to avoid biases due to multicollinearity. I examine results in Table 3.8, Table 3.9, Table 3.10 and to answer two questions: (1) *Does route familiarity affect accident risk?* (2) *What is the best way to incorporate it into the model?*

Each of the six variable groups created in Section 3.4.3 contains statistically significant predictors of accident involvement as measured by liability claim submission. The sings of corresponding marginal effects indicate that driving on an unfamiliar road is associated with higher accident hazard, thus yielding support for **(H1)**. McFadden Pseudo R2 coefficients suggest representing a route by the start and the end point cluster to achieve a better model fit. Furthermore, predictors that are more informative regarding the number of trips per a rarely visited route have greater explanatory power.

Including new predictors in the model affects the relationship between other significant telematics variables and accident risk. *Weekend driving percentage* becomes statistically significant at a 5% level and the magnitude of the effect increases by 28% on average across the fitted models. A plausible explanation for this shift is as follows: *weekend driving* combines effect of several factors with counteracting effects on the accident hazard. On the one hand, during the weekend, drivers are more likely to visit new destinations, whereby increasing risk. On the other, traffic intensity tends to be much lower during the weekend, which improves the driving situation. Accounting for route familiarity helps to crystallize the latter effect.[9] The effect is most pronounced on the daily distance driven: the magnitude of this effect decreases by the factor of 2 once familiarity measures are incorporated in the model. Interestingly, the impact of *average no. of journeys* does not significantly change after accounting for route familiarity.

Table 3.11 reports results of incorporating information about light conditions in the model. They do not support **(H2)**. Furthermore, ambient light conditions do not exacerbate the risk of driving in unfamiliar area **(H2.1)**. I have added different combinations of variables from Section 3.4.4 and in neither case were these variables statistically significant or considerably improved the model fit (see Table 3.12 and Table 3.13).

I do not find statistically significant relation between long journey frequency and accident risk (Table 3.14), thus **(H3)** is rejected. Bearing in mind that according to Table 3.7 such journeys are extremely rare, driving-induced fatigue is not a prevalent problem for our policyholders.

The external validity of my conclusions could be questioned due to *selection bias*. The dataset comprises young drivers that have opted for a *usage-based* insurance policy. Furthermore they have purchased a *pay-how-you-drive* contract, where the premium depends on the driving style as opposed to distance-based *pay-as-you-drive* policy. In the absence of driving logs from policyholders with other contract types, it is not possible to test whether these results hold for both subsamples. However, none of the risk factors I have tested has a direct impact on the premium discount, which reduces the gravity of the problem.

[9] Kremslehner and Muermann (2016) finds that driving on the weekend decreases the risk and attributes it to *Sunday drivers*.

Table 3.8: Annual probability of liability claim submission (probit) (Average marginal effects)

	Dependent variable: Liability claim submission in 2016 ($> 10CHF$)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Traditional Insurance Variables								
Sex of driver: female	0.003	0.003	0.003	0.004	0.004	0.003	0.004	0.004
	(0.058)	(0.058)	(0.058)	(0.058)	(0.059)	(0.058)	(0.058)	(0.058)
Age of driver	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Vehicle: age	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Vehicle: horsepower	−0.00003	−0.00002	−0.00002	−0.00002	−0.00000	−0.00003	−0.00003	−0.00003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Vehicle: price	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Vehicle: mileage	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00000	0.00000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.0003	−0.0002	−0.0002	−0.0002	−0.0002	−0.0003	−0.001	−0.001
	(0.080)	(0.080)	(0.080)	(0.080)	(0.081)	(0.080)	(0.080)	(0.080)
Recent change of address: true	0.011	0.015	0.010	0.010	0.011	0.011	0.012	0.012
	(0.069)	(0.071)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)
Bonus Malus Score TPL	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.002	−0.002
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
No. of years without first contract	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
No. of previous mobility claims	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.008
	(0.067)	(0.067)	(0.067)	(0.068)	(0.068)	(0.067)	(0.068)	(0.068)
Telematics-Based Predictors								
Average distance per day (active)	0.001^{***}	0.001^{***}	0.0004^{**}	0.0005^{**}	0.0005^{**}	0.001^{***}	0.0005^{**}	0.0004^{**}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.012^{***}	0.012^{***}	0.011^{***}	0.011^{***}	0.011^{***}	0.013^{***}	0.014^{***}	0.014^{***}
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.028)	(0.027)	(0.027)
Average speed (type 1)	−0.0004	−0.0004	−0.0003	−0.0003	−0.0004	−0.0004	−0.0003	−0.0003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	-0.055^{*}	-0.053^{*}	-0.068^{**}	-0.068^{**}	-0.062^{**}	-0.058^{**}	-0.069^{**}	-0.073^{**}
	(0.238)	(0.238)	(0.247)	(0.247)	(0.250)	(0.240)	(0.245)	(0.247)
Urban driving percentage (speed=50)	−0.013	−0.014	−0.011	−0.011	−0.016	−0.013	−0.012	−0.010
	(0.297)	(0.297)	(0.299)	(0.300)	(0.307)	(0.297)	(0.297)	(0.298)
Night driving percentage	0.052	0.050	0.058	0.060	0.042	0.058	0.067	0.068
	(0.449)	(0.449)	(0.452)	(0.455)	(0.470)	(0.452)	(0.453)	(0.452)
Percentage of driving above speed limit (type 1 adjusted)	0.080^{**}	0.082^{**}	0.073^{*}	0.072^{*}	0.079^{*}	0.082^{**}	0.081^{**}	0.077^{*}
	(0.342)	(0.343)	(0.344)	(0.345)	(0.349)	(0.343)	(0.342)	(0.342)
Average speeding (weighted)	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{**}	0.003^{***}	0.003^{***}	0.003^{***}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	-0.061^{*}	-0.061^{*}	-0.060^{*}	-0.060^{*}	-0.068^{*}	-0.061^{*}	-0.059^{*}	-0.059^{*}
	(0.289)	(0.289)	(0.291)	(0.292)	(0.297)	(0.289)	(0.289)	(0.290)
Average speeding in urban areas (relative)	−0.009	−0.009	−0.009	−0.009	−0.007	−0.009	−0.009	−0.009
	(0.107)	(0.107)	(0.108)	(0.108)	(0.109)	(0.107)	(0.107)	(0.107)
No. of accelerations per km (tr=2)	0.008	0.007	0.008	0.009	0.010	0.008	0.009	0.009
	(0.180)	(0.180)	(0.180)	(0.181)	(0.183)	(0.180)	(0.180)	(0.180)
No. of braking per km (tr=2)	0.019	0.019	0.018	0.018	0.019	0.018	0.017	0.017
	(0.176)	(0.176)	(0.177)	(0.177)	(0.180)	(0.176)	(0.176)	(0.176)
Average speed difference after acceleration	−0.0004	−0.0004	−0.0005	−0.0005	−0.001	−0.0004	−0.001	−0.001
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Average speed at the beginning of accelerations	0.0005	0.0004	0.0005	0.0005	0.0003	0.0005	0.0004	0.0004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	0.001	0.001	0.001	0.001	0.001	0.0005	0.0004
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Average speed at the beginning of braking	0.00003	0.00002	0.0001	0.0001	0.0001	0.00003	0.0001	0.0001
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Percentage of visits to:								
the most popular location		0.055						
		(0.323)						
5 most popular locations			-0.069^{**}					
			(0.292)					
10 most popular locations				-0.078^{**}				
				(0.332)				
20 most popular locations					-0.081^{*}			
					(0.394)			
Percentage of clusters accounting for:								
for 75% of visits						0.041		
						(0.413)		
for 90% of visits							0.047^{**}	
							(0.203)	
for 95% of visits								0.058^{**}
								(0.203)
McFadden Pseudo R2	0.0328	0.0335	0.0342	0.0341	0.0336	0.0330	0.0342	0.0349
Log Likelihood	−1,352.928	−1,351.866	−1,350.838	−1,350.146	−1,320.990	−1,352.541	−1,350.953	−1,349.940
Akaike Inf. Crit.	2,763.856	2,763.732	2,761.677	2,760.292	2,701.980	2,765.083	2,761.906	2,759.881
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 3.9: Annual probability of liability claim submission (probit) (Average marginal effects)

	Dependent variable: Liability claim submission in 2016 ($> 10CHF$)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Traditional Insurance Variables								
Sex of driver: female	0.003	0.003	0.003	0.003	0.004	0.004	0.004	0.003
	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)
Age of driver	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Vehicle: age	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Vehicle: horsepower	−0.00003	−0.00003	−0.00002	−0.00002	−0.00002	−0.00003	−0.00003	−0.00003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Vehicle: price	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Vehicle: mileage	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.0003	0.00003	−0.0002	−0.0003	−0.0001	−0.001	−0.001	−0.001
	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)
Recent change of address: true	0.011	0.009	0.009	0.009	0.009	0.011	0.011	0.010
	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)
Bonus Malus Score TPL	−0.001	−0.001	−0.001	−0.001	−0.001	−0.002	−0.001	−0.001
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
No. of years without first contract	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
No. of previous mobility claims	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.008	−0.007
	(0.067)	(0.068)	(0.068)	(0.067)	(0.068)	(0.068)	(0.068)	(0.068)
Telematics-Based Predictors								
Average distance per day (active)	0.001^{***}	0.001^{***}	0.0005^{**}	0.0005^{**}	0.0005^{**}	0.0004^{**}	0.0004^{**}	0.0004^{**}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.012^{***}	0.011^{***}	0.011^{***}	0.010^{***}	0.010^{***}	0.013^{***}	0.012^{***}	0.012^{***}
	(0.027)	(0.027)	(0.027)	(0.027)	(0.028)	(0.027)	(0.027)	(0.027)
Average speed (type 1)	−0.0004	−0.0004	−0.0004	−0.0003	−0.0004	−0.0003	−0.0003	−0.0003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	-0.055^{*}	-0.061^{**}	-0.066^{**}	-0.067^{**}	-0.066^{**}	-0.070^{**}	-0.076^{***}	-0.074^{**}
	(0.238)	(0.243)	(0.245)	(0.246)	(0.246)	(0.246)	(0.248)	(0.248)
Urban driving percentage (speed=50)	−0.013	−0.012	−0.011	−0.011	−0.017	−0.011	−0.010	−0.009
	(0.297)	(0.298)	(0.299)	(0.300)	(0.301)	(0.297)	(0.299)	(0.299)
Night driving percentage	0.052	0.057	0.059	0.059	0.059	0.067	0.069	0.067
	(0.449)	(0.452)	(0.453)	(0.453)	(0.458)	(0.452)	(0.453)	(0.452)
Percentage of driving above speed limit (type 1 adjusted)	0.080^{**}	0.075^{*}	0.071^{*}	0.069^{*}	0.070^{*}	0.078^{*}	0.074^{*}	0.073^{*}
	(0.342)	(0.344)	(0.345)	(0.346)	(0.347)	(0.342)	(0.342)	(0.343)
Average speeding (weighted)	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	-0.061^{*}	-0.061^{*}	-0.060^{*}	-0.060^{*}	-0.061^{*}	-0.059^{*}	-0.060^{*}	-0.060^{*}
	(0.289)	(0.290)	(0.291)	(0.291)	(0.292)	(0.290)	(0.291)	(0.290)
Average speeding in urban areas (relative)	−0.009	−0.009	−0.009	−0.009	−0.010	−0.009	−0.008	−0.008
	(0.107)	(0.108)	(0.108)	(0.108)	(0.109)	(0.107)	(0.108)	(0.108)
No. of accelerations per km (tr=2)	0.008	0.009	0.009	0.009	0.005	0.009	0.010	0.010
	(0.180)	(0.180)	(0.180)	(0.181)	(0.182)	(0.180)	(0.180)	(0.180)
No. of braking per km (tr=2)	0.019	0.018	0.017	0.017	0.021	0.017	0.017	0.017
	(0.176)	(0.176)	(0.177)	(0.177)	(0.178)	(0.176)	(0.176)	(0.176)
Average speed difference after acceleration	−0.0004	−0.0005	−0.001	−0.0005	−0.0004	−0.001	−0.001	−0.001
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Average speed at the beginning of accelerations	0.0005	0.0005	0.0005	0.0005	0.001	0.0004	0.0004	0.0004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	0.001	0.001	0.001	0.001	0.0004	0.0003	0.0005
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Average speed at the beginning of braking	0.00003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Percentage of journeys on (undirected):								
the most popular route		−0.036						
		(0.237)						
5 most popular routes			-0.050^{**}					
			(0.212)					
10 most popular routes				-0.054^{**}				
				(0.218)				
20 most popular routes					-0.058^{**}			
					(0.228)			
Percentage of routes (undirected) accounting for:								
for 75% of journeys						0.059^{**}		
						(0.233)		
for 90% of journeys							0.072^{***}	
							(0.222)	
for 95% of journeys								0.103^{**}
								(0.355)
McFadden Pseudo R2	0.0328	0.0334	0.0342	0.0343	0.0342	0.0345	0.0356	0.0351
Log Likelihood	−1,352.928	−1,352.044	−1,350.864	−1,350.268	−1,345.746	−1,350.563	−1,349.016	−1,349.665
Akaike Inf. Crit.	2,763.856	2,764.089	2,761.728	2,760.536	2,751.492	2,761.127	2,758.032	2,759.329
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 3.10: Annual probability of liability claim submission (probit) (Average marginal effects)

	Dependent variable: Liability claim submission in 2016 ($> 10CHF$)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Traditional Insurance Variables								
Sex of driver: female	0.003	0.003	0.003	0.003	0.003	0.004	0.003	0.003
	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)
Age of driver	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Vehicle: age	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Vehicle: horsepower	−0.00003	−0.00002	−0.00002	−0.00002	−0.00002	−0.00003	−0.00003	−0.00003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Vehicle: price	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Vehicle: mileage	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.0003	0.0001	−0.0001	−0.0002	−0.0003	−0.001	−0.001	−0.001
	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)
Recent change of address: true	0.011	0.009	0.009	0.009	0.009	0.011	0.010	0.011
	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)
Bonus Malus Score TPL	−0.001	−0.001	−0.001	−0.001	−0.001	−0.002	−0.001	−0.001
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
No. of years without first contract	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
No. of previous mobility claims	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007
	(0.067)	(0.068)	(0.068)	(0.068)	(0.067)	(0.068)	(0.068)	(0.068)
Telematics-Based Predictors								
Average distance per day (active)	0.001^{***}	0.001^{***}	0.0005^{**}	0.0005^{**}	0.0005^{**}	0.0004^{**}	0.0004^{**}	0.0005^{**}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.012^{***}	0.011^{***}	0.011^{***}	0.011^{***}	0.010^{***}	0.013^{***}	0.012^{***}	0.012^{***}
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Average speed (type 1)	−0.0004	−0.0004	−0.0004	−0.0004	−0.0003	−0.0003	−0.0003	−0.0004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	-0.055^{*}	-0.063^{**}	-0.065^{**}	-0.067^{**}	-0.068^{**}	-0.071^{**}	-0.074^{**}	-0.071^{**}
	(0.238)	(0.244)	(0.245)	(0.246)	(0.247)	(0.247)	(0.248)	(0.248)
Urban driving percentage (speed=50)	−0.013	−0.011	−0.011	−0.011	−0.011	−0.011	−0.010	−0.009
	(0.297)	(0.299)	(0.299)	(0.299)	(0.300)	(0.297)	(0.299)	(0.299)
Night driving percentage	0.052	0.060	0.059	0.060	0.061	0.068	0.068	0.066
	(0.449)	(0.453)	(0.453)	(0.453)	(0.456)	(0.453)	(0.453)	(0.452)
Percentage of driving above speed limit (type 1 adjusted)	0.080^{**}	0.074^{*}	0.073^{*}	0.071^{*}	0.069^{*}	0.078^{*}	0.074^{*}	0.075^{*}
	(0.342)	(0.344)	(0.345)	(0.345)	(0.346)	(0.342)	(0.343)	(0.343)
Average speeding (weighted)	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	-0.061^{*}	-0.061^{*}	-0.060^{*}	-0.060^{*}	-0.061^{*}	-0.059^{*}	-0.060^{*}	-0.060^{*}
	(0.289)	(0.290)	(0.291)	(0.291)	(0.292)	(0.290)	(0.290)	(0.290)
Average speeding in urban areas (relative)	−0.009	−0.009	−0.009	−0.009	−0.009	−0.009	−0.009	−0.009
	(0.107)	(0.108)	(0.108)	(0.108)	(0.108)	(0.107)	(0.108)	(0.108)
No. of accelerations per km (tr=2)	0.008	0.009	0.009	0.009	0.009	0.009	0.010	0.010
	(0.180)	(0.180)	(0.180)	(0.180)	(0.181)	(0.180)	(0.180)	(0.180)
No. of braking per km (tr=2)	0.019	0.018	0.017	0.017	0.018	0.017	0.017	0.018
	(0.176)	(0.176)	(0.177)	(0.177)	(0.177)	(0.176)	(0.176)	(0.176)
Average speed difference after acceleration	−0.0004	−0.0005	−0.0005	−0.001	−0.0005	−0.001	−0.001	−0.001
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Average speed at the beginning of accelerations	0.0005	0.0004	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	0.001	0.001	0.001	0.001	0.0004	0.0004	0.001
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Average speed at the beginning of braking	0.00003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Percentage of journeys on (directed):								
the most popular route		−0.080						
		(0.449)						
5 most popular routes			-0.045^{*}					
			(0.216)					
10 most popular routes				-0.050^{**}				
				(0.212)				
20 most popular routes					-0.053^{**}			
					(0.218)			
Percentage of routes (directed) accounting for:								
for 75% of journeys						0.051^{**}		
						(0.198)		
for 90% of journeys							0.075^{**}	
							(0.260)	
for 95% of journeys								0.110^{**}
								(0.469)
McFadden Pseudo R2	0.0328	0.0336	0.0339	0.0342	0.0342	0.0345	0.0350	0.0343
Log Likelihood	−1,352.928	−1,351.716	−1,351.257	−1,350.790	−1,349.691	−1,350.510	−1,349.745	−1,350.734
Akaike Inf. Crit.	2,763.856	2,763.432	2,762.514	2,761.580	2,759.382	2,761.019	2,759.491	2,761.469
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 3.11: Annual probability of liability claim submission (probit) (Average marginal effects)

	Dependent variable: Liability claim submission in 2016 ($> 10CHF$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Traditional Insurance Variables									
Sex of driver: female	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)
Age of driver	−0.003**	−0.003**	−0.003**	−0.003**	−0.003**	−0.003**	−0.003**	−0.003**	−0.003**
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Vehicle: age	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Vehicle: horsepower	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Vehicle: price	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Vehicle: mileage	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.0003	−0.0003	−0.0005	−0.001	−0.0001	−0.0003	−0.0003	−0.0002	−0.0005
	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)
Recent change of address: true	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)
Bonus Malus Score TPL	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
No. of years without first contract	0.003*	0.003*	0.003*	0.003*	0.003*	0.003*	0.003*	0.003*	0.003*
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
No. of previous mobility claims	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007
	(0.067)	(0.067)	(0.067)	(0.068)	(0.067)	(0.067)	(0.068)	(0.068)	(0.068)
Telematics-Based Predictors									
Average distance per day (active)	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Average speed (type 1)	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	−0.055*	−0.054*	−0.061*	−0.056**	−0.057**	−0.053*	−0.052*	−0.054*	−0.056*
	(0.238)	(0.247)	(0.266)	(0.243)	(0.242)	(0.244)	(0.244)	(0.247)	(0.252)
Urban driving percentage (speed=50)	−0.013	−0.013	−0.013	−0.013	−0.013	−0.013	−0.013	−0.012	−0.013
	(0.297)	(0.297)	(0.297)	(0.297)	(0.297)	(0.297)	(0.297)	(0.297)	(0.297)
Night driving percentage	0.052	0.054	0.055	0.057	0.038	0.056	0.052	0.037	0.042
	(0.449)	(0.470)	(0.470)	(0.490)	(0.498)	(0.471)	(0.450)	(0.507)	(0.509)
Percentage of driving above speed limit (type 1 adjusted)	0.080**	0.080**	0.080**	0.082**	0.080**	0.079**	0.080**	0.080**	0.083**
	(0.342)	(0.342)	(0.342)	(0.343)	(0.343)	(0.342)	(0.343)	(0.343)	(0.343)
Average speeding (weighted)	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	−0.061*	−0.061*	−0.060*	−0.061*	−0.061*	−0.061*	−0.061*	−0.062*	−0.061*
	(0.289)	(0.289)	(0.289)	(0.289)	(0.289)	(0.289)	(0.289)	(0.289)	(0.289)
Average speeding in urban areas (relative)	−0.009	−0.009	−0.009	−0.009	−0.009	−0.009	−0.010	−0.010	−0.010
	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)
No. of accelerations per km (tr=2)	0.008	0.008	0.008	0.007	0.007	0.008	0.008	0.008	0.007
	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)
No. of braking per km (tr=2)	0.019	0.019	0.019	0.018	0.018	0.019	0.019	0.018	0.017
	(0.176)	(0.176)	(0.176)	(0.177)	(0.176)	(0.176)	(0.176)	(0.176)	(0.177)
Average speed difference after acceleration	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Average speed at the beginning of accelerations	0.0005	0.0005	0.0004	0.0004	0.0005	0.0005	0.0004	0.0004	0.0004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Average speed at the beginning of braking	0.00003	0.00003	0.00004	0.00003	0.00004	0.00003	0.00002	0.00001	−0.00000
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Percentage of driving time during:									
Twilight		0.006							
		(0.472)							
Morning twilight			−0.017						
			(0.591)						
Evening twilight			0.045						
			(0.756)						
After sunrise				−0.044					0.394
				(2.029)					(2.878)
Before sunset				0.194					0.533*
				(1.624)					(2.327)
Civil twilight					−0.086			−0.161	−0.576*
					(1.363)			(1.687)	(2.639)
Nautical twilight						0.029		0.078	0.165
						(1.092)		(1.565)	(1.596)
Astronomical twilight							0.052	0.044	−0.001
							(1.041)	(1.270)	(1.331)
McFadden Pseudo R2	0.0328	0.0328	0.0329	0.0332	0.0329	0.0328	0.0329	0.0331	0.0344
Log Likelihood	−1,352.928	−1,352.922	−1,352.774	−1,352.395	−1,352.785	−1,352.902	−1,352.839	−1,352.517	−1,350.684
Akaike Inf. Crit.	2,763.856	2,765.845	2,767.548	2,766.790	2,765.570	2,765.804	2,765.678	2,769.034	2,769.368
Note:									*p<0.1; **p<0.05; ***p<0.01

Table 3.12: Annual probability of liability claim submission (probit) (Average marginal effects)

	Dependent variable: Liability claim submission in 2016 ($> 10CHF$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Traditional Insurance Variables									
Sex of driver: female	0.003	0.004	0.003	0.004	0.004	0.004	0.004	0.004	0.004
	(0.058)	(0.058)	(0.058)	(0.061)	(0.058)	(0.058)	(0.058)	(0.058)	(0.061)
Age of driver	−0.003**	−0.003**	−0.003**	−0.002*	−0.003**	−0.003**	−0.003**	−0.003**	−0.002*
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Vehicle: age	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**
	(0.007)	(0.007)	(0.007)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
Vehicle: horsepower	−0.00003	−0.00002	−0.00004	−0.0001	−0.00002	−0.00004	−0.00002	−0.00004	−0.0001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Vehicle: price	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Vehicle: mileage	0.00001	0.00001	0.00001	0.00000	0.00001	0.00000	0.00001	0.00001	0.00000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.0003	−0.001	−0.001	0.003	−0.001	−0.002	−0.001	−0.002	0.003
	(0.080)	(0.080)	(0.080)	(0.083)	(0.080)	(0.080)	(0.080)	(0.080)	(0.083)
Recent change of address: true	0.011	0.010	0.010	0.012	0.010	0.010	0.010	0.010	0.012
	(0.069)	(0.069)	(0.069)	(0.071)	(0.069)	(0.069)	(0.069)	(0.069)	(0.071)
Bonus Malus Score TPL	−0.001	−0.001	−0.001	−0.0004	−0.001	−0.001	−0.001	−0.001	−0.0004
	(0.023)	(0.023)	(0.023)	(0.024)	(0.023)	(0.023)	(0.023)	(0.023)	(0.024)
No. of years without first contract	0.003*	0.003*	0.003*	0.003	0.003*	0.003*	0.003*	0.003*	0.003
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
No. of previous mobility claims	−0.007	−0.008	−0.007	−0.006	−0.008	−0.007	−0.008	−0.007	−0.006
	(0.067)	(0.068)	(0.068)	(0.068)	(0.068)	(0.067)	(0.067)	(0.067)	(0.068)
Telematics-Based Predictors									
Average distance per day (active)	0.001***	0.0004**	0.0004**	0.0004*	0.0004**	0.0004**	0.0004**	0.0004**	0.0004*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.012***	0.010***	0.010***	0.011***	0.010***	0.010***	0.010***	0.010***	0.010***
	(0.027)	(0.029)	(0.029)	(0.030)	(0.029)	(0.029)	(0.029)	(0.029)	(0.031)
Average speed (type 1)	−0.0004	−0.0003	−0.0003	−0.0001	−0.0003	−0.0003	−0.0003	−0.0003	−0.0001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	−0.055*	−0.078***	−0.076***	−0.077**	−0.077***	−0.076***	−0.078***	−0.076***	−0.078**
	(0.238)	(0.249)	(0.250)	(0.267)	(0.249)	(0.249)	(0.249)	(0.250)	(0.268)
Urban driving percentage (speed=50)	−0.013	−0.009	−0.006	0.003	−0.009	−0.007	−0.009	−0.007	0.004
	(0.297)	(0.299)	(0.300)	(0.324)	(0.299)	(0.300)	(0.299)	(0.300)	(0.324)
Night driving percentage	0.052	0.065	0.070	0.065	0.064	0.070	0.063	0.068	0.062
	(0.449)	(0.455)	(0.456)	(0.500)	(0.458)	(0.454)	(0.456)	(0.462)	(0.506)
Percentage of driving above speed limit (type 1 adjusted)	0.080**	0.068*	0.064	0.044	0.069*	0.065	0.068*	0.065	0.042
	(0.342)	(0.345)	(0.346)	(0.368)	(0.345)	(0.346)	(0.345)	(0.346)	(0.369)
Average speeding (weighted)	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	−0.061*	−0.061*	−0.062*	−0.046	−0.061*	−0.059*	−0.061*	−0.060*	−0.045
	(0.289)	(0.291)	(0.292)	(0.305)	(0.291)	(0.291)	(0.291)	(0.292)	(0.306)
Average speeding in urban areas (relative)	−0.009	−0.008	−0.008	−0.002	−0.008	−0.008	−0.008	−0.007	−0.001
	(0.107)	(0.108)	(0.108)	(0.114)	(0.108)	(0.108)	(0.108)	(0.108)	(0.114)
No. of accelerations per km (tr=2)	0.008	0.010	0.010	0.015	0.010	0.010	0.011	0.011	0.016
	(0.180)	(0.180)	(0.181)	(0.189)	(0.180)	(0.180)	(0.180)	(0.180)	(0.189)
No. of braking per km (tr=2)	0.019	0.018	0.020	0.023	0.018	0.019	0.018	0.020	0.023
	(0.176)	(0.176)	(0.177)	(0.190)	(0.176)	(0.176)	(0.176)	(0.177)	(0.190)
Average speed difference after acceleration	−0.0004	−0.001	−0.0005	0.0005	−0.001	−0.001	−0.001	−0.001	0.0005
	(0.013)	(0.013)	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
Average speed at the beginning of accelerations	0.0005	0.0004	0.001	0.001	0.0004	0.001	0.0004	0.001	0.001
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	0.0004	0.0004	0.001	0.0004	0.0004	0.0004	0.0005	0.001
	(0.018)	(0.018)	(0.018)	(0.019)	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
Average speed at the beginning of braking	0.00003	0.0002	0.0001	−0.0003	0.0002	0.0002	0.0002	0.0002	−0.0003
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Relative proportion of driving on unfamiliar roads (less than 1% of journeys) during:									
Twilight		0.043							
		(0.240)							
Morning twilight			−0.017						
			(0.435)						
Evening twilight			0.037						
			(0.265)						
Before sunset				0.024					−0.020
				(0.237)					(0.459)
After sunrise				0.007					−0.008
				(0.205)					(0.261)
Civil twilight					0.034			−0.004	0.037
					(0.210)			(0.573)	(0.863)
Nautical twilight							0.041	0.033	0.018
							(0.234)	(0.805)	(0.823)
Astronomical twilight						0.043		0.016	0.024
						(0.243)		(0.626)	(0.653)
Percentage of routes (undirected) accounting for for 90% of journeys		0.073***	0.078***	0.073***	0.074***	0.074***	0.074***	0.075***	0.072***
		(0.223)	(0.225)	(0.237)	(0.223)	(0.224)	(0.223)	(0.224)	(0.238)
McFadden Pseudo R2	0.0328	0.0364	0.0359	0.0354	0.0363	0.0359	0.0362	0.0360	0.0358
Log Likelihood	−1,352.928	−1,347.808	−1,343.269	−1,228.046	−1,347.727	−1,345.044	−1,347.604	−1,344.513	−1,227.327
Akaike Inf. Crit.	2,763.856	2,757.617	2,750.537	2,520.091	2,757.454	2,752.089	2,757.207	2,755.027	2,524.655
Note:									*p<0.1; **p<0.05; ***p<0.01

Table 3.13: Annual probability of liability claim submission (probit) (Average marginal effects)

	Dependent variable: Liability claim submission in 2016 ($> 10CHF$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Traditional Insurance Variables									
Sex of driver: female	0.003	0.004	0.003	0.004	0.004	0.004	0.004	0.004	0.004
Age of driver	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{*}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{**}	-0.003^{*}
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Vehicle: age	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}	0.002^{**}
	(0.007)	(0.007)	(0.007)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
Vehicle: horsepower	−0.00003	−0.00002	−0.00004	−0.0001	−0.00002	−0.00004	−0.00002	−0.00004	−0.0001
Vehicle: weight	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
Vehicle: price	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000
Vehicle: mileage	0.00001	0.00001	0.00001	0.00000	0.00001	0.00000	0.00001	0.00000	0.00000
Leasing contract: true	−0.0003	−0.001	−0.002	0.003	−0.001	−0.002	−0.001	−0.002	0.003
Recent change of address: true	0.011	0.010	0.010	0.012	0.010	0.010	0.010	0.010	0.012
Bonus Malus Score TPL	−0.001	−0.001	−0.001	−0.0004	−0.001	−0.001	−0.001	−0.001	−0.0003
No. of years without first contract	0.003^{*}	0.003^{*}	0.003^{*}	0.003	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}	0.003^{*}
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
No. of previous mobility claims	−0.007	−0.008	−0.007	−0.006	−0.008	−0.008	−0.008	−0.008	−0.006
Telematics-Based Predictors									
Average distance per day (active)	0.001^{***}	0.0004^{**}	0.0004^{**}	0.0004^{*}	0.0004^{**}	0.0004^{**}	0.0004^{**}	0.0004^{**}	0.0004^{**}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.012^{***}	0.010^{***}	0.010^{***}	0.011^{***}	0.010^{***}	0.010^{***}	0.010^{***}	0.010^{***}	0.010^{***}
	(0.027)	(0.029)	(0.030)	(0.031)	(0.029)	(0.029)	(0.029)	(0.030)	(0.031)
Average speed (type 1)	−0.0004	−0.0003	−0.0003	−0.0001	−0.0003	−0.0003	−0.0003	−0.0003	−0.0001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	-0.055^{*}	-0.075^{**}	-0.076^{**}	-0.073^{**}	-0.080^{***}	-0.072^{**}	-0.075^{**}	-0.074^{**}	-0.075^{**}
	(0.238)	(0.256)	(0.271)	(0.272)	(0.253)	(0.254)	(0.254)	(0.257)	(0.281)
Urban driving percentage (speed=50)	−0.013	−0.009	−0.006	0.003	−0.009	−0.007	−0.009	−0.006	0.003
	(0.297)	(0.299)	(0.300)	(0.324)	(0.299)	(0.299)	(0.299)	(0.300)	(0.324)
Night driving percentage	0.052	0.070	0.073	0.076	0.052	0.070	0.070	0.055	0.059
	(0.449)	(0.476)	(0.476)	(0.542)	(0.504)	(0.455)	(0.477)	(0.516)	(0.569)
Percentage of driving above speed limit (type 1 adjusted)	0.080^{**}	0.068^{*}	0.064	0.043	0.069^{*}	0.066	0.068^{*}	0.065	0.042
	(0.342)	(0.345)	(0.346)	(0.368)	(0.345)	(0.346)	(0.345)	(0.347)	(0.369)
Average speeding (weighted)	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{***}	0.003^{**}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	-0.061^{*}	-0.061^{*}	-0.062^{*}	−0.046	-0.061^{*}	-0.059^{*}	-0.061^{*}	-0.060^{*}	−0.045
	(0.289)	(0.291)	(0.292)	(0.306)	(0.291)	(0.291)	(0.291)	(0.292)	(0.306)
Average speeding in urban areas (relative)	−0.009	−0.008	−0.008	−0.002	−0.008	−0.008	−0.008	−0.008	−0.001
	(0.107)	(0.108)	(0.108)	(0.114)	(0.108)	(0.108)	(0.108)	(0.108)	(0.114)
No. of accelerations per km (tr=2)	0.008	0.011	0.011	0.016	0.010	0.011	0.011	0.011	0.017
	(0.180)	(0.180)	(0.181)	(0.189)	(0.180)	(0.180)	(0.180)	(0.180)	(0.190)
No. of braking per km (tr=2)	0.019	0.018	0.020	0.023	0.018	0.019	0.018	0.018	0.023
	(0.176)	(0.176)	(0.177)	(0.190)	(0.176)	(0.177)	(0.176)	(0.177)	(0.190)
Average speed difference after acceleration	−0.0004	−0.001	−0.0005	0.0005	−0.001	−0.001	−0.001	−0.001	0.0005
	(0.013)	(0.013)	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
Average speed at the beginning of accelerations	0.0005	0.0004	0.001	0.001	0.0004	0.0005	0.0004	0.0005	0.001
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	0.0004	0.0004	0.001	0.0004	0.0004	0.0004	0.0004	0.001
	(0.018)	(0.018)	(0.018)	(0.019)	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
Average speed at the beginning of braking	0.00003	0.0002	0.0001	−0.0003	0.0002	0.0001	0.0002	0.0002	−0.0003
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Percentage of driving time during:									
Twilight		0.018							
		(0.474)							
Morning twilight			0.004						
			(0.614)						
Evening twilight			0.023						
			(0.781)						
After sunrise				0.150					0.500
				(2.211)					(3.169)
Before sunset				−0.050					0.230
				(1.937)					(2.742)
Civil twilight					−0.082			−0.186	−0.470
					(1.369)			(1.704)	(2.979)
Nautical twilight							0.055	0.114	0.139
							(1.096)	(1.563)	(1.909)
Astronomical twilight						0.088		0.060	−0.011
						(1.050)		(1.280)	(1.543)
Relative proportion of driving on unfamiliar roads (less than 1% of journeys) during:									
Twilight		0.043							
		(0.241)							
Morning twilight			−0.016						
			(0.439)						
Evening twilight			0.038						
			(0.267)						
Before sunset				0.022					−0.017
				(0.238)					(0.462)
After sunrise				0.009					−0.005
				(0.207)					(0.263)
Civil twilight					0.033			−0.012	0.028
					(0.211)			(0.576)	(0.866)
Nautical twilight							0.042	0.037	0.025
							(0.235)	(0.807)	(0.825)
Astronomical twilight						0.045		0.021	0.021
						(0.245)		(0.631)	(0.657)
Percentage of routes (undirected) accounting for for 90% of journeys		0.073^{***}	0.077^{***}	0.074^{***}	0.074^{***}	0.075^{***}	0.074^{***}	0.077^{***}	0.072^{**}
		(0.224)	(0.231)	(0.239)	(0.223)	(0.224)	(0.224)	(0.225)	(0.242)
McFadden Pseudo R2	0.0328	0.0364	0.0360	0.0355	0.0364	0.0361	0.0363	0.0364	0.0367
Log Likelihood	−1,352.928	−1,347.754	−1,343.235	−1,227.851	−1,347.598	−1,344.790	−1,347.512	−1,343.855	−1,226.211
Akaike Inf. Crit.	2,763.856	2,759.507	2,754.469	2,523.703	2,759.196	2,753.579	2,759.023	2,759.710	2,532.422
Note:									*p<0.1; **p<0.05; ***p<0.01

Table 3.14: Annual probability of liability claim submission (probit) (Average marginal effects)

	Dependent variable: Liability claim submission in 2016 ($> 10CHF$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Traditional Insurance Variables									
Sex of driver: female	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)
Age of driver	−0.003**	−0.003**	−0.003**	−0.003**	−0.003**	−0.003*	−0.003*	−0.003*	−0.003*
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Vehicle: age	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**	0.002**
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Vehicle: horsepower	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003	−0.00003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Vehicle: price	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Vehicle: mileage	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00000	0.00000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.001	−0.001	−0.001	−0.0004	−0.0003	−0.001	−0.001	−0.001	−0.001
	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)
Recent change of address: true	0.011	0.011	0.011	0.012	0.012	0.011	0.011	0.011	0.011
	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)
Bonus Malus Score TPL	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.002
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
No. of years without first contract	0.003*	0.003*	0.003*	0.003*	0.003*	0.003	0.003	0.003	0.003
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
No. of previous mobility claims	−0.007	−0.007	−0.007	−0.007	−0.007	−0.008	−0.008	−0.007	−0.007
	(0.068)	(0.068)	(0.068)	(0.067)	(0.067)	(0.068)	(0.068)	(0.068)	(0.068)
Telematics-Based Predictors									
Average distance per day (active)	0.001***	0.001***	0.001***	0.001***	0.001***	0.0004**	0.0004**	0.0004*	0.0004**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***	0.012***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Average speed (type 1)	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.0003	−0.0003	−0.0003	−0.0003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Weekend driving percentage	−0.059**	−0.059**	−0.059**	−0.062**	−0.062**	−0.080***	−0.081***	−0.083***	−0.083***
	(0.234)	(0.237)	(0.236)	(0.235)	(0.235)	(0.247)	(0.246)	(0.246)	(0.246)
Urban driving percentage (speed=50)	−0.012	−0.012	−0.012	−0.013	−0.012	−0.008	−0.009	−0.009	−0.009
	(0.296)	(0.296)	(0.296)	(0.296)	(0.296)	(0.298)	(0.298)	(0.298)	(0.298)
Night driving percentage (new)	0.045	0.045	0.045	0.044	0.042	0.041	0.041	0.039	0.038
	(0.380)	(0.380)	(0.380)	(0.380)	(0.380)	(0.379)	(0.379)	(0.378)	(0.379)
Percentage of driving above speed limit (type 1 adjusted)	0.081**	0.081**	0.081**	0.083**	0.082**	0.075*	0.075*	0.078*	0.077*
	(0.342)	(0.343)	(0.343)	(0.343)	(0.343)	(0.343)	(0.343)	(0.343)	(0.343)
Average speeding (weighted)	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	−0.061*	−0.061*	−0.061*	−0.059*	−0.060*	−0.059*	−0.059*	−0.058*	−0.058*
	(0.289)	(0.289)	(0.289)	(0.289)	(0.289)	(0.290)	(0.290)	(0.290)	(0.290)
Average speeding in urban areas (relative)	−0.009	−0.009	−0.009	−0.009	−0.009	−0.008	−0.008	−0.008	−0.008
	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)
No. of accelerations per km (tr=2)	0.008	0.008	0.008	0.008	0.008	0.010	0.009	0.009	0.009
	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)	(0.180)
No. of braking per km (tr=2)	0.017	0.017	0.017	0.017	0.017	0.014	0.014	0.014	0.014
	(0.175)	(0.175)	(0.175)	(0.175)	(0.175)	(0.175)	(0.175)	(0.175)	(0.175)
Average speed difference after acceleration	−0.0004	−0.0004	−0.0004	−0.0004	−0.0004	−0.001	−0.001	−0.001	−0.001
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Average speed at the beginning of accelerations	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	0.001	0.001	0.001	0.001	0.0002	0.0002	0.0002	0.0002
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.018)
Average speed at the beginning of braking	0.00003	0.00003	0.00003	0.00004	0.00003	0.0001	0.0001	0.0001	0.0001
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Frequency of journeys exceeding									
4 hours		−0.025				−0.038			
		(2.341)				(2.360)			
5 hours			0.028				0.004		
			(3.825)				(3.820)		
6 hours				0.621				0.572	
				(5.573)				(5.571)	
7 hours					0.771				0.732
					(6.938)				(6.915)
Percentage of routes (undirected) accounting for 90% of journeys						0.067***	0.067***	0.067***	0.067***
						(0.221)	(0.221)	(0.221)	(0.221)
McFadden Pseudo R2	0.0328	0.0328	0.0328	0.0331	0.0331	0.0353	0.0353	0.0355	0.0355
Log Likelihood	−1,352.907	−1,352.903	−1,352.905	−1,352.479	−1,352.503	−1,349.420	−1,349.430	−1,349.060	−1,349.057
Akaike Inf. Crit.	2,763.814	2,765.806	2,765.810	2,764.958	2,765.006	2,760.841	2,760.859	2,760.120	2,760.114

Note: *p<0.1; **p<0.05; ***p<0.01

3.7 Discussion

In this section I elaborate on potential practical implications of my results. First, I established that driving on unfamiliar roads increases accident risk. The magnitude of the effect is illustrated in Figure 3.4: (top) shows accident probabilities for different levels of sample quantiles of six variables: *Percentage of clusters (routes directed/undirected) accounting for 90% and 95% of visits (journeys).* Accident probabilities corresponding to 25% sample quantiles of these variables are on average 1.2% lower than the values for 75% sample quantiles.[10]

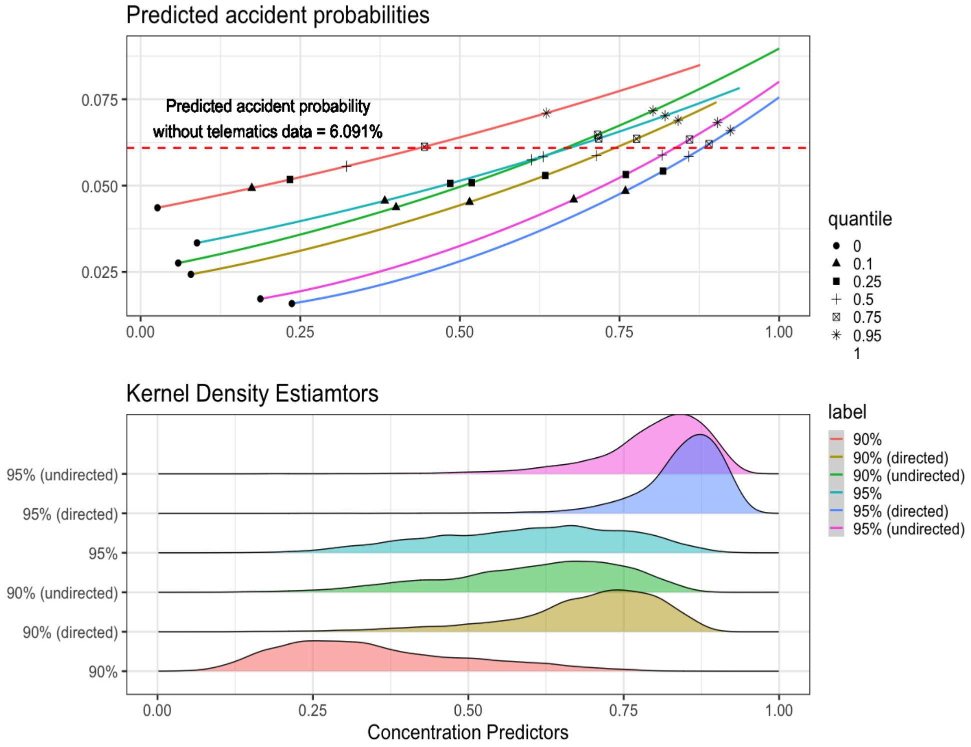
Williams (2006) suggests to counter accident risk by using information about hazardous driving patterns for insurance pricing. After an initial exploitation period, driving logs are sufficient to evaluate the policyholder's familiarity with an undertaken route. Consequently it is possible to incorporate this factor into risk classification. The first approach would be to make the premium proportional to the relative distance driven on rarely visited roads. This however could alienate a lot of low-risk clients. Alternatively, the insurer could introduce stricter penalties for speed violations, harsh braking and accelerations when driving on unfamiliar routes. This practice is widespread to gauge the potential risk stemming from urban and night driving.

It is challenging to distil the effect of light conditions on accident hazard. The most commonly-used proxy, *night driving percentage*, combines the impacts of several different risk factors: on the negative side, drivers are more likely to be tired or under the influence of alcohol, on the positive side, traffic intensity is lower.[11] Two possibilities to crystallize the effect are a) to focus on twilight driving, since the effect of additional risk factors is weaker, b) compare accident rates in the same area at the same time of day between two time periods: in winter after dark and in summer when it's not dark

[10] In our sample, assuming that an accident results in the loss of vehicle and using average vehicle prices reported in Table 3.1, a 1.2% change in the accident probability would require a 380 CHF adjustment of a pure risk premium.

[11] I fit several models where I exclude speed and speeding related telematics predictors to see whether in their absence *night driving* becomes statistically significant. Contrarily to my expectations this was not the case

Figure 3.4: Predicted annual accident probabilities for different empirical quantiles of route familiarity variables.



yet. The second approach is not feasible with our dataset, since accidents are infrequent and exact times and locations are not known. Following the former approach, I find no evidence that frequent driving under bad light conditions increases the accident hazard.

When discussing practical implications of this result, it is worth noting that premium increase is not the only punishment for driving “misbehavior" and accident involvement. Since the average driver age is 22.5 years in our sample, a considerable share only has a probationary driver’s license. A driver with such a license faces harsher penalties and stricter restrictions on alcohol consumption. This might suffice to offset the risks associated with late driving. Therefore insurers could attract more policyholders without incurring considerable losses by not using information about night driving for premium calculation. This result also suggests that the government could save money by using more energy-efficient technologies for road lighting without compromising traffic safety.

I do not find any statistically significant association between accident involvement and the frequency of long journeys. To summarize, neither of two situations (late driving and long uninterrupted traveling), when driver could be prone to fatigue, translates into higher accident probability. This might challenge the necessity of governmental regulations of the maximum driving time for the professional drivers. The external validity of such a conclusion is under question since our sample comprises drivers who rarely undertake long journeys and thereby is not suitable for exploring the impact of fatigue on accident risk. This is further exacerbated by the fact that due to well-maintained infrastructure on the one hand and a good driving situation on the other driving is less effortful and more time passes before the driver gets tired.

3.8 Conclusion

Some aspects of the Big Data revolution are reminiscent of the so-called Gold Rush: stakeholders greedily and indiscriminately collect data in the hope that one day they can mine it for economic value. Driving logs are good examples for these new types of data. Large businesses collect them to monitor their vehicle fleet, car producers use this information to improve vehicle diagnostics, driving logs enable emergency service operations, thereby decreasing the number of accident fatalities. In this contribution I focused on a particular use case of driving logs: insurance pricing.

I created driving profiles for policyholders of a large Swiss insurance carrier and studied whether this information can predict subsequent accident involvement. Previous contributions suggest that controlling for distance driven, location and frequency of speed violations, the *No. of journeys* is a statistically significant risk factor (Muermann et al. (2019), Sycheva et al. (2019)). To shed light on this phenomenon I extended the driving profiles to reflect *where*, *when* and *how long* did policyholder drive. I found that driving in unfamiliar locations increases accident hazard. Neither driving under ambient light conditions nor undertaking long distance journeys impacts the likelihood of accident involvement.

These insights could be used by the insurance company to reach two closely-aligned objectives: refine risk classification scheme and provide incentives for safe driving. In a field experiment, Hultkrantz and Lindberg (2011) demonstrate that financial penalties are effective in reducing speed violations. The trade-off between these goals arises when deciding how to incorporate risk-relevant information into premium calculation. Machine learning algorithms would yield more accurate premia. In these models, however, the link between driving behavior and the premium is obfuscated. Consequently, the incentives embedded in the pricing scheme are less clear for policyholders and therefore less effective.

Apart from accident risk, driving produces a lot of externalities, including CO2 emissions, traffic congestion and further environmental damages. Driving logs could serve as basis for successful internalization mechanisms. The greatest hurdle to collecting more data are the privacy concerns of data subjects. Thus various stakeholders would need to find a delicate balance between extracting valuable insights from the data without infringing on individual's privacy.

4. A Franc Less for a Pound More: (Price) Discrimination and the Value of Privacy

Price discrimination based on consumers' personal data has become common practice in many markets. We analyze the willingness to share personal data when this data is used for price discrimination in subsequent markets. In a laboratory experiment, participants can sell a bundle of personal data. Participants are categorized based on the content of their personal data and receive category-dependent payoffs in a subsequent stage. The experimental variations modify the category-dependent payoff structure. We find no effect of subsequent financial discrimination on the general willingness to sell personal data. A significant change in the data reservation price is only observed under strong negative discrimination. Furthermore, we observe important gender differences in the reservation price for private information and the role of underlying privacy concerns.

4.1 Introduction

Using customer information to boost sales is an old and time-tested strategy. Private doctors traveling between cities in Ancient Greece charged more to the rich than to the poor. With easier access to private information, price discrimination based on consumers' detailed personal data as well as individual behavior is increasingly becoming common practice in many markets.[1] Airlines and other companies, such as Home

[1] Price discrimination can be beneficial for consumers and retailers. The seller can often generate more revenue by offering services and products at lower costs to groups that tend to be more price sensitive, for example in the case of student and senior discounts.

Depot, offer individual prices for different customers on their websites based on factors such as the time and day of the online activity as well as the customers' zip codes.[2] While price discrimination can be beneficial for consumers, these practices are often intransparent to consumers and violate their privacy.

The use of Big Data enables firms to pursue these practices with greater precision and therefore may lead to negative outcomes for consumers. Ezrachi and Stucke (2016) argue that online behavioral discrimination through big data will likely differ from the brick-and-mortar type of price discrimination in three ways: (1) a shift from third-degree, imperfect price discrimination to near-perfect or first-degree price discrimination by segmenting consumers into smaller groups and identifying their reservation prices more precisely (2) an increase in overall consumption through marketing strategies that target consumers' emotions more effectively and (3) a stronger durability of discrimination stemming from personalization and data-driven network effects. The authors point out that the increase in personalized product offers and individual pricing makes it harder for consumers to evaluate all options and assess general market prices.

A very controversial issue is risk-based pricing in insurance. In various lines of insurance business, such as health insurance, life insurance and automotive insurance, policies are priced based on policyholders' personal data that is used to predict their risk type. Risk-based pricing can incentivize policyholders to behave less risky and create desirable spill-over effects, such as an improvement in road safety and better general health. Risk-based pricing can become problematic though when potential insurance buyers are priced out of the market based on factors beyond their control, such as genetic conditions. A pricing scheme that makes health insurance prohibitively expensive for an individual with such a condition seems unfair and possibly illegal. Genetic tests may even allow insurers to uncover medical conditions that the affected customer might not yet know about.

[2] Access to consumer information may intensify competition. E.g., Choe et al. (2019) show that the ability to gather information and use personalized prices reduces firms' total profits in a model of asymmetric collection of personal information of consumers (e.g. via cookies).

The willingness to share personal information that influences the demand for insurance products with such a pricing scheme differs across individuals. In a data set for pay-as-you-drive contracts, Kremslehner and Muermann (2016) show that such a car insurance policy that involves information sharing is more likely to be chosen by younger, female consumers who live in urban and/or wealthier areas.

Besides the potential direct economic consequences, privacy concerns play an important role for consumers' decision to purchase such products. The value of privacy has been subject to a public debate that has become increasingly relevant with public scandals, such as the Facebook–Cambridge Analytica data scandal.[3]

In this article, we make use of a laboratory experiment to elicit individuals' willingness to share personal data when this data is subsequently used to price discriminate. Thus, we analyze participants' privacy concerns as well as their response to payoff discrimination based on the content of their personal data. The personal data that the participants can sell in the experiment consists of the bundle of their height, weight, bank account balance information as well as a photo of their face. To implement price discrimination in the lab, participants are then categorized based on whether they sold their data to the experimenters, and importantly based on the content of their data, whereby the category-cutoffs depend in particular on a person's weight and bank account balance. These categories entail different payoffs in a subsequent stage, thus implementing data-based price discrimination in a reduced form.

Comparing treatments with and without data-dependent payoff differences, we find no effect of price discrimination per se on the general willingness to sell personal data. Thus, within the experiment, we don't find evidence of a disutility of financial discrimination attached to the content of personal data per se. With respect to the price of personal data demanded by participants, we find a significant change in the data reservation price under strong negative discrimination, i.e. when the subsequent payoff decreases strongly for one data category. Interestingly, this increase in the reservation price is observed also for participants that do not fall into the corresponding category.

[3] See for instance New York Times (2018).

Furthermore, we find important gender differences in how general privacy concerns and trust related to the context of the experiment affect both the general willingness to sell the data as well as the reservation price of the data. The general privacy concerns and trust with respect to the decision context are thereby derived from answers of a comprehensive survey about privacy-related attitudes and behavior. These findings are important in light of the consequences of personal data sharing for subsequent market interaction, not only with respect to price discrimination, but the usage of personal data more generally.

Related Literature

This article mainly relates to two areas of research: (1) Price discrimination based on consumers' personal data and (2) Consumers' valuation of private information.

Academic research on price discrimination based on consumers' personal data mainly focuses on the effects on market allocations, market efficiency, and social welfare. One common example in academic literature is the use of genetic information for the pricing of health insurance and life insurance contracts (Crocker and Snow (2013), Dionne and Rothschild (2014), Crainich (2017)).

Montes et al. (2018) analyze theoretically how price discrimination based on consumers' private information affects prices, profits, and consumer surplus in a consumption goods market. In their framework, firms can acquire consumer data for price discrimination from a third party intermediary and individual consumers can prevent the use of their private information by paying a *privacy cost*. The authors find that higher *privacy costs* decrease competing duopolists' profits and increase consumer surplus. In the monopoly case, the effect on consumer surplus and social welfare is ambiguous.

Belleflamme and Vergote (2016) assume that consumers can react to a monopolist's *tracking technology*, that identifies consumers' willingness to pay with a certain probability, by making use of a *hiding technology*. The authors show that while more accurate price discrimination by the use of such *tracking technologies* decreases consumer surplus, the availability of privacy protecting technologies may imply an even higher reduction of consumer surplus. The rationale is that the availability of privacy protecting technologies incentivizes the monopolist to limit the use of the *tracking technology* and to raise the regular market price of the product or service.

In an experimental setting, Richards et al. (2016) analyze perceptions of price fairness and *self-interested inequity aversion* in the context of price discrimination. Consumers with *self-interested inequity aversion* regard prices as unfair and tend to purchase less, if they perceive other consumers to pay a lower price. They tend to regard prices as more fair, if inequity is in their favor, which results in higher purchases. The authors find that the implications of such inequity aversion can be at least partially reversed if consumers are involved in the price formation.

The literature on consumers' valuation of private information is extensive and covers a wide range of fields. Acquisti et al. (2016) summarize and link various streams of theoretical and empirical economic research that investigates individual and societal trade-offs associated with protecting and disclosing personal information. The authors note that privacy related issues of economic relevance can be observed in diverse contexts and that situations can arise in which the protection of privacy can both enhance and reduce individual and social welfare. Further, they find that imperfect information about the purpose and the consequences of data collection severely hinders consumers' ability to make informed decisions about their privacy in digital economies.

Some experimental studies evaluate individuals' valuation of privacy by asking participants indirectly (Beresford et al. (2012), Regner and Riener (2017)) and directly (Benndorf and Normann (2017)) to sell private information.[4]

Beresford et al. (2012) conduct a field experiment to measure participants' willingness to pay for privacy. Participants are given the opportunity to purchase a DVD from one of two online stores, for which they have to provide personal information, such as last name, postal address, and e-mail address. In addition to those common data items, one store requires information about the date of birth and monthly in-

[4] For a detailed overview, see Kern et al. (2018).

come, whereas the other store asks for year of birth and favorite color. Except that the first store requires more sensitive personal information than the latter, both stores are identical. The authors find that when DVDs are offered for one Euro less at the store asking for more sensitive information, almost all participants choose to buy from the cheaper store. When prices are identical at both stores, however, participants buy equally often from either one.

In a Pay-What-You-Want (PWYW) online music store and a mimicking online experiment, Regner and Riener (2017) analyze the effect of revealing customer information, such as name and e-mail address, to the seller on consumers' purchasing behavior. While for donations and public goods, reduced anonymity can lead to higher PWYW revenues due to self-image motivations, Regner and Riener (2017) find that revealing customers' information in the online store context reduces the number of customers purchasing. Overall, lifting anonymity leads to a revenue loss of 25% (35%) in the online music store (in the online experiment). The authors conclude that the substantial reduction of customers might be explained by privacy concerns.

Benndorf and Normann (2017) conduct laboratory experiments to assess participants' willingness to sell personal data to a telecommunications company. Participants in the experiments can sell five different bundles of personal information that covered participants' information on (1) preferences (2) contact data (3) both preferences and contact data (4) facebook profile and (5) facebook timeline. The authors find considerable heterogeneity in participants' willingness to sell personal information. About one sixth of the participants refuse to sell any personal data while a similar fraction sells for 2.50 € or less. The average price requested is 15 € for contact details and 19 € for Facebook data. The authors also find a gender effect: Female participants' valuation of personal data appears to be more sensitive to the type of data.

Various articles analyze the effect of external factors on the value of privacy. These include among others endowment effects (Acquisti et al. (2013)), pre-existing attitudes or dispositions, limited cognitive resources, and momentary affective states (Kehr et al. (2015b)), data-breach notifications (Feri et al. (2016)), positive or negative information on companies' attitudes towards privacy (Marreiros et al. (2017)), the number of information recipients (Schudy and Utikal (2017)), and implicitly and explicitly stated prices, political orientation, income proxies and membership in loyalty programs (Plesch and Wolff (2018)).

Benndorf and Normann (2017) name three explanatory factors for differences in results between different studies on the value of privacy: incentivized decisions to share personal information, a salient focus on privacy issues, and transparent information with respect to the use of data shared.

In this experiment, we highlight in the experimental instructions for participants that personal information sold is not shared with third parties nor used for other purposes than for the data analysis in this experiment. Subjects are informed about the use of their personal information and data is sold explicitly. Hence, our experimental setting ensures salience, incentivization, and transparency. Apart from the pure value of privacy, we examine the effects of financial discrimination based on the personal information shared.

The remainder of this chapter is organized as follows. In the next section, we present the experimental set-up and our hypotheses. Section 4.3 provides the data analyses and discusses the results. The final section concludes.

4.2 Experiment

4.2.1 Experimental Design

We apply a between-subject design to analyze individuals' valuation of personal data in light of potentially discriminatory use of this data. Specifically, we examine whether and how subjects' willingness to sell their personal data is affected by both inherent privacy concerns and financial discrimination based on their data. The mechanism by which personal data is bought/sold in the experiment is the Becker-DeGroot-Marschak

(BDM) mechanism (Becker et al. (1964)), a standard incentive-compatible method for eliciting private values in laboratory experiments. Only if the participant wants to sell his or her personal data according to the BDM mechanism, the data is collected by the experimenters.

The experiment consists of three parts: In Part 1, participants practice the BDM procedure by having the option to sell back a 5 CHF coin which they receive as initial endowment to the experimenter. In Part 2, subjects can sell their personal data to the experimenters via the BDM mechanism. In the experiment, the personal data is the following bundle of personal information: The participant's height, weight, gender, bank account balance, and a picture of the participant's face. If participants sell their personal data via the BDM mechanism, their personal data is collected subsequently. In Part 3, participants first receive an additional payoff, and after that play a trust game, make decisions with respect to risky payoffs, and answer a post-experimental questionnaire. The payoff at the beginning of Part 3 represents the "payoff discrimination stage" in a reduced form: Depending on the experimental treatment, the payoff differs according to whether data was sold and the content of the data. There are five experimental treatments, which differ precisely in whether and how subjects are categorized based on their personal data, and the associated Part 3 payoffs. Participants have full information about data dependent categorization and associated payoffs prior to selling the data.

At the beginning of the experiment, control questions ensure that subjects have understood the instructions that describe the experiment. After all participants have successfully completed the control questions, and before Part I starts, subjects have the opportunity to self-verify their personal data in the absence of the experimenters. For this purpose, a measurement tape and a scale are located in the entry hall in front of the laboratory. Participants are informed that they can go to this entry hall in order to measure their height and weight and use their own cell phones to self-verify their bank account balance without experimenters observing it.

After the self-verification, participants begin with Part 1. Part 1 precedes the selling of personal data to familiarize participants with the BDM mechanism. Subjects receive a 5 CHF coin which they can sell back to the experimenters via the Becker-DeGroot-Marschak (BDM) mechanism (Becker et al. (1964)). The BDM mechanism by which the 5 CHF coin is bought/sold works as follows. Subjects have to state the minimum amount of money they would accept in exchange for an object they could sell to the experimenters. We refer to this amount as the reservation price for the respective object. The market price is then determined by a random draw. If this market price exceeds the reservation price stated by a participant, the object is sold and the subject receives the randomly determined market price as a payment. If the reservation price claimed exceeds the market price, the participant keeps the object but does not receive any money. Because the BDM procedure can be rather demanding, we make several arrangements to familiarize the participants with the mechanism. Following Grether and Plott (1979), we stress that subjects have an incentive to state their true valuation and that renegotiations are excluded. We also clarify that the random draw is independent of actual choices. Finally, prior to making their actual choices, subjects also have the possibility to conduct several tests with different prices and random draws using a payoff simulator that displays the hypothetical outcomes. Participants can use the simulator as long as they want to before continuing to Part 2.

In Part 2, participants can sell the bundle of personal data to the experimenters via the BDM mechanism. Participants are informed that their market price is randomly drawn from the range $[0 \text{ CHF}, 60 \text{ CHF}]$, where each 10 cent increment is equally likely. Prior to selling their data via the BDM mechanism, subjects receive information about what happens with the data sold to the experimenters, i.e. whether and how the data is used subsequently. The BDM ensures that participants only sell the data if they wish to do so. Their decision screen has two items: A field where they can put the minimum price (*reservation price*) at which they are willing to sell the data, as well as a box they can tick which reads “I don’t want to sell the data in any case”. Participants ticking this box will be classified as not agreeing to sell the data in the results section.[5]

In all parts, the markets are fully internal to the experiment such that no personal data is shared outside of the lab. Participants have this information and are asked a corresponding control question to ensure that there is no ambiguity about data sharing.

Subsequent to the decision to sell their personal data, all participants are brought one by one into a separate part of the laboratory, the *measurement room*, that other participants can neither enter, nor can they hear or see anything that is happening inside this room. If participants have decided to sell their data, their personal data is collected by experimenters in this separate room.[6] If participants have decided to not sell their data, they are brought to the measurement room by the experimenters and are asked to wait there for 2-3 minutes before they are picked up and brought back to the main part of the laboratory again. This serves the purpose that all participants are brought to the separate room by the experimenters and other participants do not observe who sold their personal data. Participants know this process.

The experimental treatment variation is about whether and how personal data sold is used in Part 3 for discriminatory purposes. In the baseline Treatment (I), the personal data is not used, and all subjects receive the same payoff of 20 CHF from this part of the experiment.[7] I.e., the Part 3 payoff is independent of their personal data and of whether personal data is sold to the experimenters. In this treatment, with the BDM mechanism, we elicit the pure privacy value attached to the personal data.

In Treatments (II)-(V), participants are classified into one of three groups according to their data provided. Subjects that do not sell their personal data are classified as category A. For those that did sell their personal data, the classification is based on

[5] If participants tick the box but specify a reservation price lower than 60 CHF in the other field, they are asked whether they want to adjust their answer, as the box tick of not selling in any case will receive priority. Independent of that, all participants are asked whether they are certain about the decision or would like to adjust it before proceeding to the next screen.

[6] A detailed description of how data was collected in the measurement room is provided in the additional Online Appendix.

[7] Additional to this payoff, participants can earn money in the trust game and the lottery choice decision in Part 3. Besides providing information about social and risk preferences, these two decision situations serve the purpose of putting weight on Part 3.

both Body-Mass-Index (BMI)[8] (gender-specific) as well as bank account balance. In particular, subjects with a bank account balance above or equal to 1000 CHF and a BMI below 22 (23.5) for female (male) participants are classified as category B.[9] The remaining subjects, i.e. participants who did sell their data and have a bank account balance below 1000 CHF or a BMI above 22 (23.5) for female (male) participants, are classified as category C. Table 4.1 summarizes how participants' categorization depends on their decision to sell their personal data and the content of the data.

Table 4.1: Categorization based on BMI and Bank Account Balance

Sold Data	BMI	Bank Account (CHF)	Category
No	-	-	**A**
Yes	< 22 female (< 23.5 male)	≥ 1000	**B**
Yes	< 22 female (< 23.5 male)	< 1000	**C**
Yes	< 22 female (< 23.5 male)	≥ 1000	**C**
Yes	≥ 22 female (≥ 23.5 male)	< 1000	**C**

In Treatment (II), subjects are categorized as described above but participants' payoff is independent of their category and of whether they sell their personal information to the experimenter. All subjects receive a payoff of 20 CHF from this part of the experiment as in Treatment (I). We use Treatment (II) additionally to Baseline Treatment (I) to have a second baseline with non-payoff relevant categorization based on personal data, and the corresponding willingness to sell the data.

In Treatments (III)-(V), payoffs in Part 3 vary with the personal data based categorization. This is done in the following way: In Treatment (III), both categories, A and C, receive the baseline payoff of 20 CHF from this part of the experiment. Participants classified as category B, i.e. with a high bank account balance and a low BMI, receive CHF 30 instead. In both Treatments, (IV) and (V), categories A and B receive the baseline payoff of 20 CHF, whereas participants classified as C receive a lower payoff

[8] The BMI is calculated as $BMI = \frac{\textit{Body weight in kilogramm}}{\textit{(Body height in meter)}^2}$.

[9] We chose values in the interval $[18.5, 24.9]$ that defines a *normal* or *healthy* BMI according to the World Health Organization (2018a). Since in Switzerland, the mean BMI for males is slightly higher than for females World Health Organization (2018b), we chose a higher threshold for the BMI categorization for male participants than for female participants.

from this part of the experiment. In Treatment (IV), we employ the same payoff difference between category A and C as we have used in Treatment (III) between category A and B. Therefore, participants that are categorized as C receive 10 CHF as a payoff from this part of the experiment in Treatment (IV). In Treatment (V), we amplify the payoff difference and the payoff for subjects in category C is 0 CHF. Table 4.2 summarizes how the participants' payoffs depend on their data-based categorization in the respective Treatments. Participants receive full information in the experimental instructions about data dependent categorization and associated payoffs prior to selling the data, i.e. they have full knowledge about data-based payoff discrimination.

After payoffs in Part 3, we elicit each participant's tendency to trust others using the loss domain treatment of the trust game from Kvaløy et al. (2017). Further, we elicit participants' risk preferences using the standard (Holt and Laury (2002)) price list. One of the two games is randomly selected to be payoff-relevant. Subsequent to that, participants answer a comprehensive post-experimental questionnaire that collects information on privacy attitudes and behavior. At the end of the experiment, subjects observe a summary screen of their payoffs.

Table 4.2: Payoffs for Different Experimental Treatments

Treatment			Payoff		
Name	Abbreviation	Categorization	A	B	C
Baseline Treatment	(I)	no	20	20	20
Baseline Treatment (Categorization)	(II)	yes	20	20	20
Positive Discrimination	(III)	yes	20	**30**	20
Negative Discrimination	(IV)	yes	20	20	**10**
Strong Negative Discrimination	(V)	yes	20	20	**0**

4.2.2 Experimental Procedure

We make use of a standard laboratory experiment at the ETH Decision Science Lab. The recruitment was performed by the Decision Science Lab using the joint subject pool of University of Zurich and ETH Zurich. Invitations were sent out via email to randomly selected German speaking subjects. The subjects sign up for a specific session, whereas the signing up procedure was limited in order to maintain a gender-balanced pool of subject in each session. Participants are instructed upon arrival at the lab. This includes a short verbal introduction about the organizational procedure as well as written instructions that explain the detailed experimental procedure.[10] Participants receive a show-up fee of 5 CHF. The experiment is partly computerized using the standard z-Tree software (Fischbacher (2007)), and partly performed with pen and paper in the lab (collection of personal data).

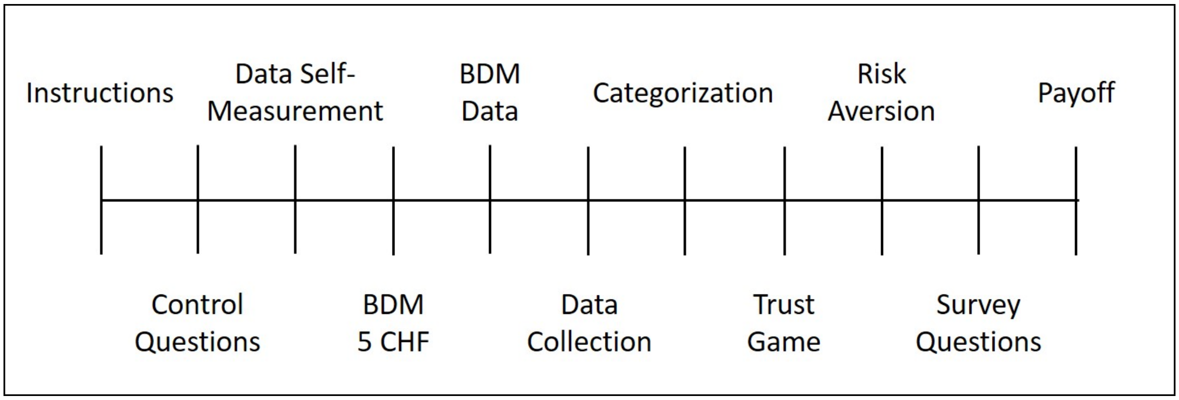
The experimental sessions were conducted from August to October and in December 2018 at the ETH Decision Science Laboratory. 282 subjects participated in 29 experiment sessions, 5 sessions were conducted for Treatment (I) and (II) each, 6 sessions per treatment were conducted for Treatment (IV) and (V) and finally 7 sessions of Treatment (III). Participants were, on average, 22.29 years old; 50.0% of the participants were female. All participants were enrolled students. More than one third of the participants were enrolled for natural sciences (28.36%), roughly one fifth for engineering (22.69%), 9.21% for medicine, 6.38% for humanities, and 9.21% for economics. The remaining 24.11% of participants were enrolled in other subjects. Subjects participated in exactly one session. Sessions lasted on average about 75 minutes. Participants earned 55 CHF (including a 5 CHF show-up fee), on average with some variance depending on participants' own decisions. A comprehensive set of control questions and the BDM practice procedure with the 5 CHF coin as initial endowment ensured that all participants understood the sequence of decisions in the experiment, the payoff consequences, and the BDM procedure.

After the main experiment and the trust game from Kvaløy et al. (2017) and standard (Holt and Laury (2002)) price list decisions, we also launch a comprehensive post-experimental questionnaire. Among other statements, participants can but do not have to specify whether they have self-verified their data, what category they would have been in if they had sold their data (in case they have not sold it), whether they were allocated to the category that they had expected to be allocated to, and why they did

[10] The German written instructions as well as the respective English translation for Treatment (I) and Treatment (IV) can be found in the additional Online Appendix.

not sell their data in case they had not done so. Furthermore, besides including the Falk et al. (2016) preference modules on social preferences and risk attitudes, participants are asked questions about their privacy concerns in other domains, their behavior with respect to private information in social networks and in the communication with various institutions, and about their self-assessment on trust in human beings as well as institutions.[11] In a last step, we ask participants about their rationale when taking decisions in the experiment and record the age, the gender and their field of study. Figure 4.1 illustrates the timeline of the experiment.

Figure 4.1: Experimental timeline



At the end of each session, participants' payoff is calculated as the sum of the payoff from decisions in Parts 1-3. Participants are paid the total amount in private at the very end of the experiment.

Incentives to sell the data

Without selling personal data, with the show-up fee of 5 CHF, the 5 CHF coin in Part 1 that can result in a payoff exceeding 5 CHF, the baseline Part 3 payoff of 20 CHF as well as the two short games at the end of Part 3, participants receive a payoff that corresponds approximately to the average payoff of participating in a laboratory experiment of that length in Zurich. I.e. participants are remunerated for their time

[11] The full German questionnaire as well as the respective English translation can be found in the additional Online Appendix.

cost and according to standard expected payoffs even without selling the data. Thus, the stated reservation price for the personal data does not have the problem of needing to remunerate participation in the experiment per se, but captures the valuation for the personal data.

4.2.3 Predictions

In the experiment, we elicit the valuation of personal data and analyze whether and how data-based financial discrimination is accounted for. In the experiment, subjects have full information about subsequent payoff discrimination. If a subject's utility depends only on final payoffs, then by backward induction the reservation price (RP) for personal data should fully adjust to payoff differences of the data-based categories.[12]

For the Baseline Treatments (I) and (II), the only difference is that participants are categorized based on their personal data, but there are no associated payoff differences. Unless there is a (dis)utility of categorization based on personal data per se, even without payoff consequences, the average RP should not differ. In the Discrimination Treatments (III-V), the RP adjustments should be upward for category C types in Treatments (IV) and (V) compared to both Treatment (III) and the Baseline Treatments. For category B types, the RP adjustments should be downward in Treatment (III) compared to all other treatments. Importantly, if utility depends only on final payoffs, then there should be no difference in category B types' RP between the Baseline Treatments and the Negative Discrimination Treatments (IV) and (V). Given the design, these RP adjustments by category/type are also underlying the overall RP adjustments by treatment. These predictions are summarized below. The average reservation price in treatment j is denoted by $\overline{RP}_j$ and the average reservation price for revealed category k in treatment j by $\overline{RP}_j^k$.

Hypothesis. ***Overall RP***

[12] As the personal data in our experiment pertains to personal attributes of the individual, i.e. type and not behavior, we cannot analyze the impact of financial discrimination in a within-subject design, as the personal attributes (contrary to behaviors) cannot be sold multiple times. In our between-subject design, we therefore need to analyze the impact of discrimination across groups of subjects.

The average reservation price does not differ between Treatment (I) and (II). Furthermore, $\overline{RP}_{III} \leq \overline{RP}_{I+II} \leq \overline{RP}_{IV} \leq \overline{RP}_{V}$.

Hypothesis. ***RP by Revealed Category***
For revealed category B, $\overline{RP}^{B}_{III} \leq \overline{RP}^{B}_{I+II} = \overline{RP}^{B}_{IV} = \overline{RP}^{B}_{V}$.
Furthermore, for revealed category C, $\overline{RP}^{C}_{III} = \overline{RP}^{C}_{I+II} \leq \overline{RP}^{C}_{IV} \leq \overline{RP}^{C}_{V}$.

4.3 Results

This section presents the main results of the experiment: Subsection 4.3.1 provides an overview of the results on the reservation price for personal data in light of subsequent financial discrimination. In section 4.3.2, we use the data from the post-experimental survey to uncover privacy attitudes. Section 4.3.3 uses these in regression analyses of the reservation price for personal data under financial discrimination.

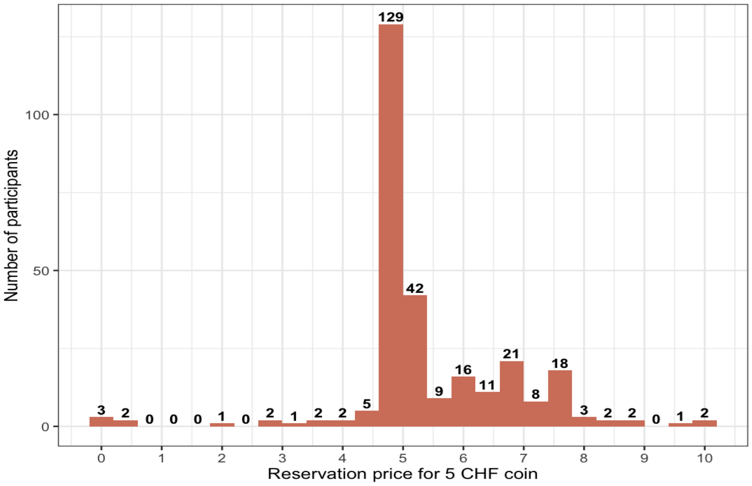
4.3.1 The Willingness to Sell Personal Data

Before presenting the results on the valuation of personal data, we look at the results from Part 1—the selling of a 5 CHF coin using the BDM mechanism—to check the understanding of the BDM method among participants. In the following, we will refer to the minimum price for the good (either the 5 CHF coin or the personal data bundle) stated by the participant as the (participant's) 'reservation price' and to the random draw in the BDM as the 'market price'.

The participants' reservation prices for the 5 CHF coins are reported in Figure 4.2. Values exceeding 5 CHF could be attributed to an endowment effect.[13] Reservation prices below 5 CHF are less easy to explain and suggest that the corresponding participant might not have understood the workings of the BDM or how to set their selling price. For this reason, we chose to exclude the 11 observations (participants) stating

[13] Furthermore, a gambling motive might explain high stated reservation prices.

Figure 4.2: Histogram of reservation prices for the 5 CHF coin



a reservation price below 4 CHF in the subsequent analysis of the willingness to sell personal data.[14]

A summary for the 271 remaining observations is provided in Table 4.3. Female and male participants are represented equally. Whether we observe the personal data (type) depends on the willingness to sell the data as well as on the market price, i.e. the results from the random number generator. There are three scenarios:

1. Participant refuse to sell the personal data in any case and signal this by checking the box for "*I don't want to sell the data in any case*". In this case, no data is sold and consequently the data cannot be used in the analyses. In the following, these observations are classified as “**not agreeing to sell the personal data”**.

2. Participants are willing to sell the personal data for a certain price but the stated reservation price exceeds the market price. In this case, no data is sold and the participant's type is not observed. The reservation price is available.

3. Participants are willing to sell the personal data and the market price exceeds the reservation price. In this case, the personal data is collected and can be used in the analyses.

[14] These 11 observations are not excluded from the survey analyses.

Table 4.3: Number of observations, grouped by experimental treatment, gender and reservation prices

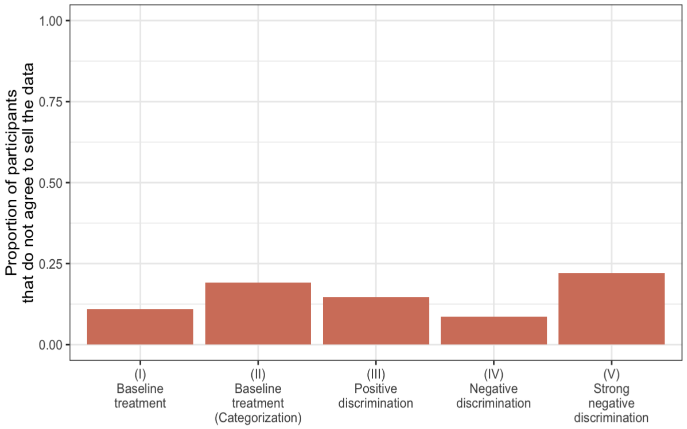
	Total no. of subjects			No. of subjects that did						No. of subjects that		
				not agree to sell the data			not sell the data			sold the data		
		female	male		female	male		female	male		female	male
Baseline Treatment (I)	46	21	25	5	2	3	22	13	9	24	8	16
Baseline Treatment with Categorization (II)	47	22	25	9	6	3	25	13	12	22	9	13
Positive Discrimination (III)	61	31	29	9	8	1	27	16	10	34	15	19
Negative Discrimination (IV)	58	30	28	5	2	3	29	14	15	29	16	13
Strong Negative Discrimination (V)	59	31	28	13	7	6	32	20	12	27	11	16
Total	271	135	135	41	25	16	135	76	58	136	59	77

Columns (5) to (7) in Table 4.3 show the participants not agreeing to sell by treatment and gender, (8) to (10) in Table 4.3 summarize the numbers of observation for which no personal data was sold, that is all observations in the first two scenarios. Columns (11) to (13) in Table 4.3 show the numbers of observation where the data was sold.[15]

Across all treatments, 15.1% (41/271) subjects refuse to sell the data at all, even though the data stay within the experiment. To compare, the rates of refusing to sell facebook timeline data and a combination of preference and contact data to a telecommunications company in Benndorf and Normann (2017) are at roughly 20%. Figure 4.3 displays the shares of participants not agreeing to sell the data by treatment. These shares do not differ significantly across treatments, in particular, the shares for the Treatments (III)-(V) do not differ significantly from the shares in the Baseline Treatments (I) and (II). Thus, a first observation is that financial discrimination based on personal data per se does not significantly impact the decision on agreeing to sell data overall, as measured by ticking the box "*I don't want to sell the data in any case*".

[15] In one Positive Discrimination Treatment (III) session, one of the participants did not indicate his or her gender in the post-experimental questionnaire. This participant agreed to sell private information for a certain price but did not sell the data (scenario 2). As a result, the observations for male and female subjects do not add up to the overall number of subjects for subjects who agreed to sell data (columns (11) to (13)), for subjects who did not sell the data (columns (8) to (10)) and for the total number of subjects (columns (2) to (4)).

Figure 4.3: Share of subjects that do not agree to sell the data by experimental treatment

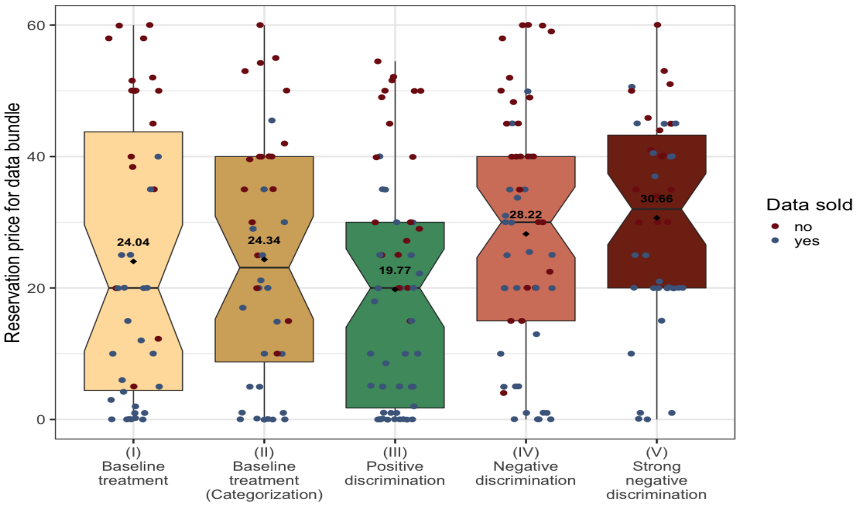


We now turn to the valuation of personal data as measured by the participants' stated reservation price for those who indicate one. Figure 4.4 displays the boxplots with quantiles, means and individual observations differentiated by whether the data was actually sold per treatment. The data presented is based on all subjects that indicate a RS, e.g. it is excluding the subjects that do not agree to sell the data. In both baseline treatments, in which the personal data has no impact on subsequent payoffs, the mean reservation price for participants who are willing to sell their data is roughly 24 CHF. This suggests that participants on average place a substantial monetary value on the privacy of their personal data, even when the data remains fully internal to the experiment. The share of subjects that indicate a reservation price of 0 is 8.7% and 8.5% in Treatments (I) and (II) respectively.

For a first analysis of the role of financial discrimination on the value of privacy, we group the non-discrimination treatments (I) and (II) together and compare these to the Discrimination Treatments (III)-(V).[16] In line with predictions, the mean reservation price is lower compared to the Baseline (I+II) when data-based discrimination is only positive in Treatment (III)(19.77 CHF), whereas it is higher when data-based discrimination is negative (28.22 CHF in (IV) and 30.66 CHF in (V) respectively). For

[16] As the Baseline Treatments (I) and (II) do not differ in important ways in the distribution of reservation prices, they are pooled together for most parts of the analysis.

Figure 4.4: Reservation price of data by treatment.



†Points represent individual observations and their color indicates, whether the bundle was sold

this average reservation price, the differences are statistically significant for (V) versus (I+II) (MWU: one-sided $p < 0.01$), as well as comparing the Negative Discrimination Treatments against the Positive Discrimination Treatments: (V) vs. (III) (MWU: one-sided $p < 0.01$) and (IV) vs. (III) (MWU: one-sided $p = 0.02$).[17]

Figure 4.5 displays the average reservation price stratified by gender. A first observation is that the mean reservation price of women is higher than that of men in each treatment. This difference in reservation prices between women and men is statistically significant in Baseline Treatment (I + II) (MWU: $p < 0.01$) and Treatment (V) (MWU: $p = 0.012$).

The higher mean reservation price also translates into lower shares of personal data revealed, as illustrated in Figure 4.6.[18] Interestingly, when stratifying by gender, we find that the differences in average reservation prices across experimental treatments for women are not statistically significant at the 5% level, except for (V) vs. (III). Thus, the above found treatment differences, in particular of Treatment (V) versus

[17] Table 5.13 in Appendix 5 shows nonparametric test results for all treatment comparisons and by gender.

[18] These differences are not statistically significant.

Baseline Treatment (I+II), are more strongly driven by changes in men's reservation prices.

Payoff discrimination is based on a subject's category. The experimental variation allows us to compare reservation price adjustments conditional on the category, and compare it to the payoff differences stemming from data-based discrimination. Table 4.4 displays the mean reservation prices by treatments, category, and gender as well as the corresponding numbers of observations. In the following, we will analyze reservation prices by category, but refrain from a detailed discussion of gender differences by category due to our small sample sizes for gender-category groups.

Figure 4.7 displays the mean reservation prices by revealed category B and C. A first observation is that category C types have a higher average reservation price than category B types in Treatments (I) and (II), despite facing the same payoff consequences.

We find that the mean reservation price of category B types is significantly lower in (III) than in (V) (MWU: $p < 0.01$), as predicted based on simple payoff differences. The mean reservation price in (III) is also lower than that in the Baseline (I+II) as

Figure 4.5: Mean reservation prices for data by treatment and gender

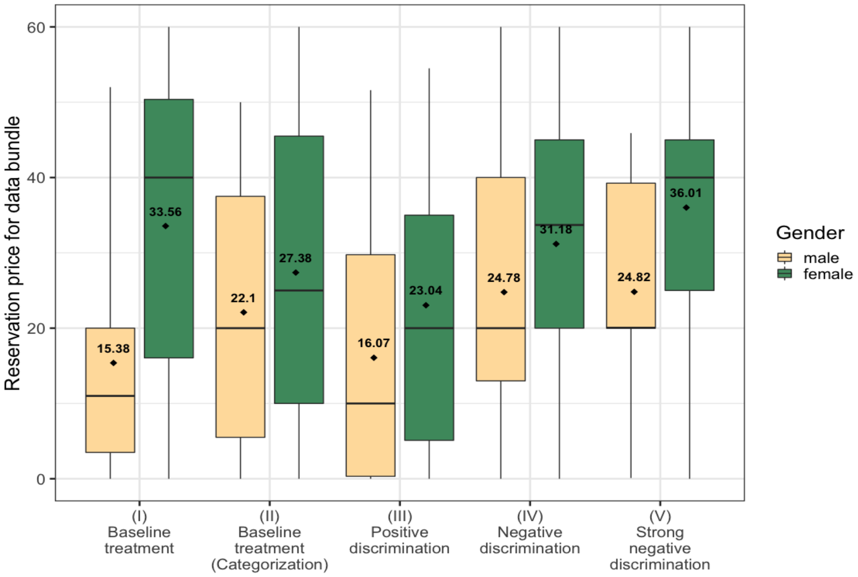


Figure 4.6: Proportion of participants selling personal data by gender

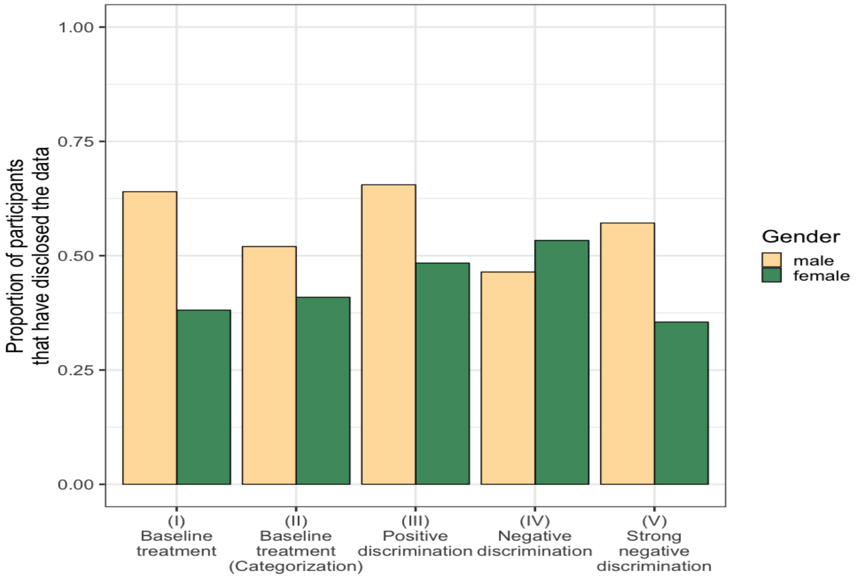
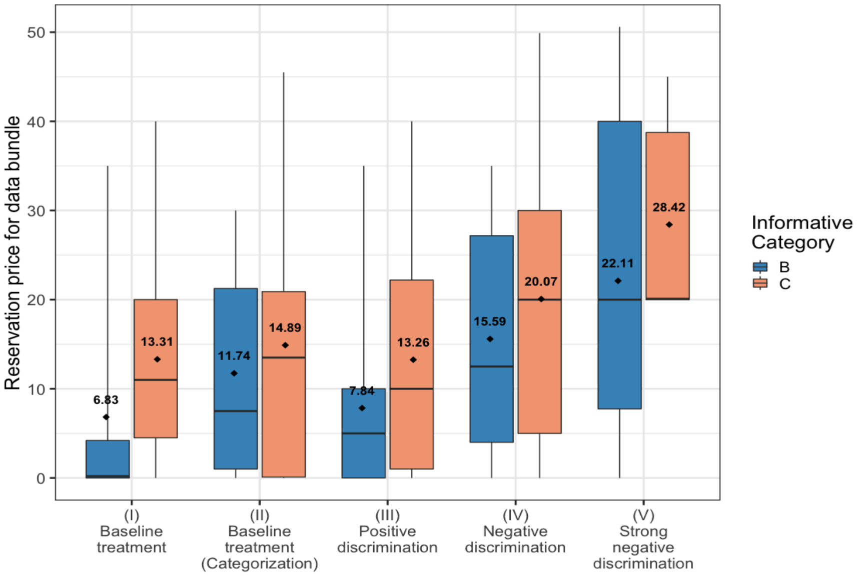


Table 4.4: Mean reservation price of personal data, grouped by experimental treatment, gender and category

			Overall			Category B			Category C	
		all	female	male	all	female	male	all	female	male
(I)	mean	24.62	35.32	15.38	7.66	2.10	9.52	13.31	17.67	10.70
	(no. obs)	41	19	22	8	2	6	16	6	10
(II)	mean	23.69	26.81	21.28	11.74	9.15	14.33	14.89	27.90	9.31
	(no. obs)	38	16	22	12	6	6	10	3	7
(III)	mean	20.12	22.77	16.91	7.80	8.33	7.67	13.34	16.02	8.73
	(no. obs)	52	23	28	15	3	12	19	12	7
(IV)	mean	27.81	30.51	24.78	15.59	14.94	16.67	20.07	26.85	12.60
	(no. obs)	53	28	25	8	5	3	21	11	10
(V)	mean	31.29	36.01	26.35	22.11	32.23	14.23	28.42	32.62	26.01
	(no. obs)	46	24	22	16	7	9	11	4	7

predicted, however the difference is not statistically significant. Furthermore, we find that mean reservation prices in Treatments (IV) and (V) are higher than in Treatments (I) and (II). Participants with revealed category B demand on average 7.4 CHF more in Treatment (IV) and 11.94 CHF in Treatment (V) compared to the pooled Baseline Treatment (I + II) without financial consequences. The difference between Treatment

Figure 4.7: Mean reservation prices for data by treatment and revealed category



(V) and (I+II) is statistically significant (MWU: p=0.04).[19] This finding is curious, as category B types face the same payoff consequences in Treatments (V), (I) and (II). One interpretation of this result is that, while there is full information on how the personal data is categorized, subjects perceive their data or in which category they fall as uncertain and therefore increase their reservation price. Another potential interpretation is that while category B types are not discriminated against, they experience a disutility similar to advantageous inequality aversion from the data-based financial discrimination and therefore increase the reservation price.

For category C, changes in subsequent payoff translate in statistically significant shifts of reservation prices in line with predictions: The mean reservation price is significantly lower in (III) vs. (V) (MWU: $p < 0.01$) and significantly higher in (V) vs. (I + II) (MWU: $p < 0.01$). The higher mean reservation price in (IV) compared to (III) is not statistically significant. Looking at the magnitudes, we observe that the average reservation price of category C types in (V) is 14.50 CHF higher than that in (I+II), compared to a payoff difference of 20 CHF.

[19] Table 5.14 and Table 5.15 in Appendix 5 report the effect sizes alongside with p-values from several non-parametric tests by category.

4.3.2 Survey Analysis

Besides subsequent data-based financial discrimination, other characteristics may influence the decision to sell personal data. To minimize the risk of omitted variables in our analysis, in this section we will discern and quantify other relevant factors of the *privacy calculus* with the help of a post-experimental survey. In Section 4.3.2, we introduce the survey questions that participants are asked to answer, and in Section 4.3.2, we discuss the role of two latent variables that can affect participants' decision whether to sell their private information: trust towards the experimenters and general attitudes towards privacy related issues. Sections 5 and 5 in the Appendix describe the construction of these latent variables via exploratory and confirmatory factor analyses.

Survey creation

Table 4.5 presents parts of the post-experimental survey questions that are grouped into six different categories: (1) Risk aversion (2) Value of privacy (3) General privacy concerns (4) Privacy related behavior (5) Social network usage and (6) Lack of trust towards the experimenters. Henceforth, we refer to the individual questions using the corresponding abbreviations listed in the table.

Malhotra et al. (2004), Kehr et al. (2015b) and Smith et al. (1996) have proposed and tested instruments to infer individuals' attitudes towards privacy. Following Kehr et al. (2015b), we assess the general importance of privacy for the participants by slightly adapting 3 out of 5 questions from the construct *Global informational privacy concerns* of Malhotra et al. (2004). These questions are designated as VP.1 - VP.3.

Since, the motivation for our experiment partially comes from the secondary usage of personal information by companies, measuring attitudes towards privacy in this context could potentially be a strong predictor in our regression analyses. Therefore, we include questions GP.1 - GP.6 from Smith et al. (1996) in order to assess subjects' perception of information privacy practices in organizations.

Expressed general attitudes are not always reliable predictors for behavior. Recent research on online user behavior has discovered great discrepancies between users' stated attitudes and their actual behavior with respect to the value of their privacy. Specifically, users tend to claim a high level of concern about their privacy, while they engage in little action to protect their private information. This phenomenon is known as the privacy paradox.[20] Potential reasons behind this discrepancy are numerous: situation specific factors, affect, perceived behavioral control, or social desirability bias. To assess whether individual's concerns actually translate into concrete actions, we collect more information about participants' behavior in situations related to protecting personal data (PB.1 - PB.8).

Voluntary sharing of personal data in social networks is a frequently observed and studied phenomenon in the context of the privacy paradox. Active users of online social network seem to be either unaware or unconcerned about the consequences of their data sharing. To capture this potentially informative behavior, we develop and add questions SN.1 - SN.10 to our survey.

In the context of our experiment, it is of great importance whether study participants believe in the integrity of the experimenters and hence whether subjects believe that all information stated in the instructions is true. To capture this, we include questions IT.1 - IT.3. Acquisti et al. (2016) consider control over information flow integral to the very definition of privacy. This further motivates question IT.3. that clarifies whether participants believe they can influence the subsequent usage of information provided during the experiment.

[20] For a literature review on the theories regarding the privacy paradox, see Barth and de Jong (2017).

Table 4.5: Excerpt from the post experimental survey (english translation)

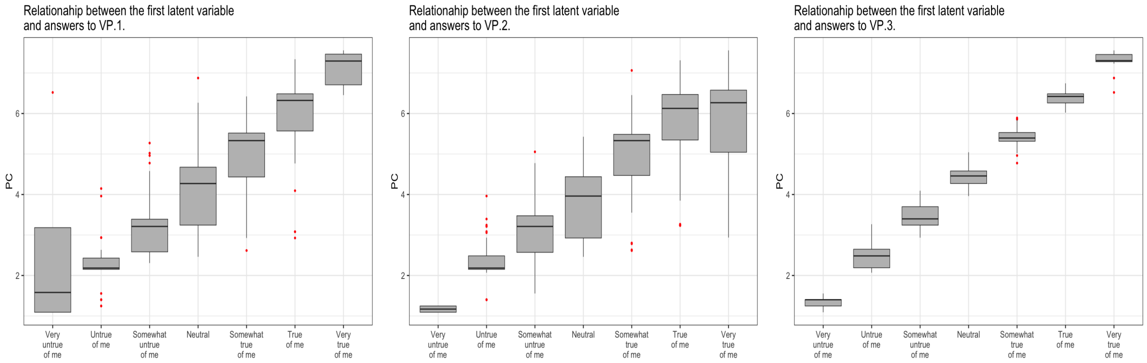
Abbreviation	Question
Risk aversion	
RA	How do you see yourself: as a person who is generally willing to take risks, or as someone who prefers to avoid them?
Value of privacy	
VP.1	Compared to others, I am more sensitive/cautious with respect to how companies handle my personal information
VP.2	To me, keeping my data private is of highest importance
VP.3	Compared to others, I tend to be more concerned about threats to my data privacy
General privacy concerns	
GP.1	It usually bothers me when companies ask me for personal information
GP.2	Companies should not use personal information unless it has been authorized by the respective person
GP.3	Companies should invest more time, effort, and costs in preventing unauthorized access to personal information
GP.4	I (sometimes) think twice before sharing private information with companies
GP.5	Companies should never share personal information with other companies without authorization by the respective person
GP.6	I am concerned that companies collect too much personal information about me
Privacy related behavior	
PB.1	Do you use a mobile app to execute transactions from your bank account or to check your account balance?
PB.2	Do you hide your bank card's PIN number when using an ATM or making purchases?
PB.3	Do you read the privacy policy before registering on a website?
PB.4	Do you remove cookies?
PB.5	Do you check your computer for spy ware?
PB.6	Do you use the private browser mode?
PB.7	How often do you change your passwords?
PB.8	Do you use the same password for different websites and services?
Social networks usage	
SN.1	How often are you active or online on social networks (e.g. Facebook, Twitter, Instagram, ...)?
SN.2	I share my general contact information (name, hometown, age, occupation)
SN.3	I share my online contact information (email, Skype, MSN)
SN.4	I share my physical contact information (phone number, address)
SN.5	I use my own photograph in my profile
SN.6.	I am honest with respect to the information about myself in my profile and my posts
SN.7	I post information on my current mood
SN.8	I share content and engage in activities that reveal my lifestyle
SN.9	Do you use the privacy settings to control who can see which piece of your information in social networks?
SN.10	Do you delete anything that you have posted in the past?
Lack of trust towards experimenters	
IT.1	I am concerned that the information I share during this experiment could be misused
IT.2	I am concerned about providing personal information during this experiment because it could be used in a way I did not foresee
IT.3	I believe I am in control over how my personal information is used by the experimenters

Latent variables

In order to identify latent characteristics of the participants, we first conduct an exploratory factor analysis to group the questions presented in the previous subsections into seven groups, of which each measures one latent characteristics. Via a confirmatory factor analysis, we identify two groups that satisfy our reliability criteria and that provide us with a good model fit. The detailed process of constructing and selecting the latent variables is described in Section 5 and Section 5 in the Appendix. In the current subsection, we introduce and interpret the two latent variables estimated by our measurement model (Figure 5.4 in Section 5 of the Appendix). The plausibility of conclusions is checked by the means of data plots on the one hand and using correlation analysis on the other.

The first latent factor, denoted as PC, is measured through the responses to questions VP.1 - VP.3 using a 7-point Likert-scale anchored by the response options "*very untrue for me*" and "*very true for me*". The relationships between survey items and the estimated value of latent characteristics are depicted in Figure 4.8. This visualization suggests that subjects with higher values of PC put more value on their privacy and tend to be more concerned about it.

Figure 4.8: Values of estimated variable PC grouped by survey responses



To cross-check our interpretation, we compute correlations between PC and various related characteristics. The results reported in Table 4.6 confirm our hypothesis that the latent variable PC captures general attitudes towards privacy related issues, as

Table 4.6: Correlation between PC and other variables

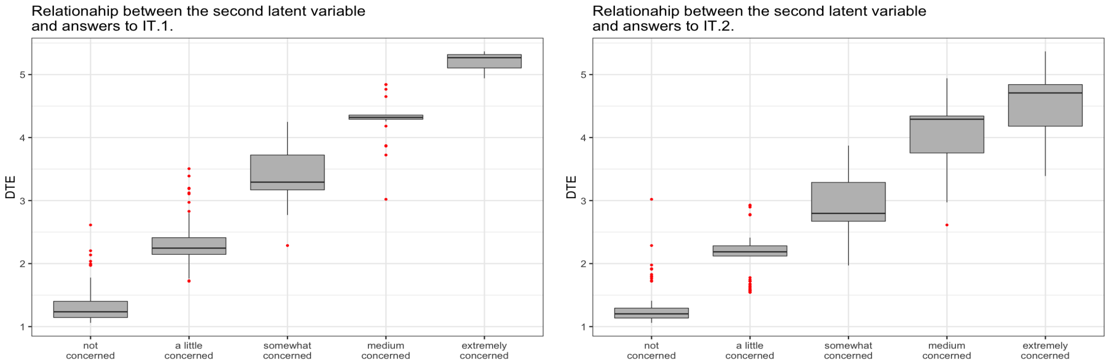
	General privacy concerns						Privacy related behavior							
	GP.1	GP.2	GP.3	GP.4	GP.5	GP.6	PB.1	PB.2	PB.3	PB.4	PB.5	PB.6	PB.7	PB.8
tau	0.346	0.106	0.219	0.372	0.147	0.454	-0.193	0.139	0.235	0.256	0.294	0.19	0.207	-0.274
pvalue	0.000	0.026	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.000	0.00	0.000	0.000

	Social Network usage									
	SN.1	SN.2	SN.3	SN.4	SN.5	SN.6	SN.7	SN.8	SN.9	SN.10
tau	0.138	0.155	0.125	0.081	0.162	0.159	0.020	-0.091	0.195	0.062
p-value	0.003	0.002	0.011	0.102	0.001	0.001	0.671	0.045	0.000	0.165

we find a significant correlation between PC and responses to the questions related to general attitudes towards privacy and privacy related behavior.[21]

We denote the second latent variable that is determined by questions IT.1 and IT.2 as *Distrust towards the experimenter* (DTE).[22] Figure 4.9 illustrates the relationship between the survey items and the estimated value of DTE. The boxplots show that high values of the latent variable DTE indicate that a participant is concerned about the protection and the use of private information that is sold to the experimenters.

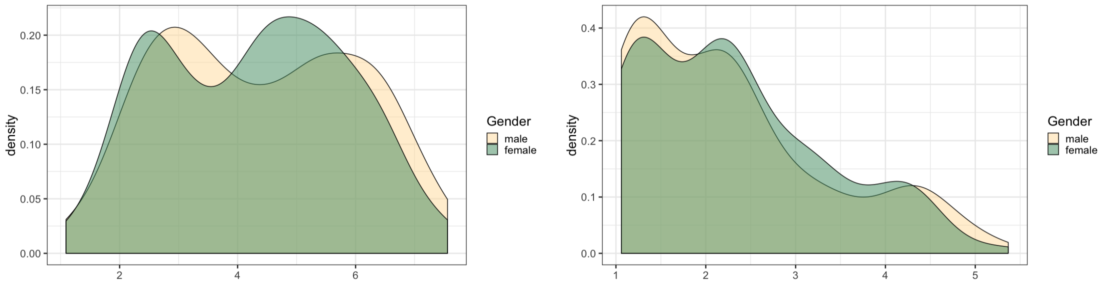
Figure 4.9: Values of estimated variable DTE grouped by survey questions



[21] To measure the relation between the continuous latent variable PC and ordinal answers to post-experimental survey questions, we use kendall's tau correlation coefficient. An association with the binary decision whether to sell private information is captured using point biserial correlation. For the relationship between PC and the reservation price for the personal data bundle, we use the Pearson correlation coefficient.

[22] The answer pattern suggest that responses to IT.1 and IT.2 are affected by a different unobserved factor as the answers to IT.3. Indeed, the first two questions assess trust towards the experimenters, whereas the last question refers more to the perceived control overs subsequent usage. It is plausible that subjects do not trust the experimenters, but are still convinced they have control over the situation and vice versa. More details can be found in Section 5.

Figure 4.10: Distribution of created variables: PC (left) and DTE (right) grouped by gender



Again, we cross-check our interpretation using other variables generated from the survey questions. DTE is not significantly correlated with answers to question I.16 and 10% correlated with answers to I.17. Furthermore, there is no association between the second latent variable DTE and the amount insured in the trust game. These discrepancies can be attributed, however, to the importance of context and situational factors. This suggests that DTE captures the specific trust towards the experiment with respect to the personal data rather than more general trust as it is captured in the trust game.

Our results from the previous subsection point at gender differences with respect to privacy related decision making. One questions is whether the variables DTE and PC could shed light on the sources of gender differences. To address this, we first compare the variables' distribution by gender. Figure 4.10 presents population kernel density estimators of PC and DTE. Neither visual inspection nor statistical tests (Table 4.7) indicate that male or female subjects tend to be more concerned about their privacy or more distrustful towards experimenters.

For a first indication of whether these two latent characteristics are related to participants' willingness to sell personal data, we determine the latent variables' correlation

Table 4.7: Comparison of distributions of latent characteristics by gender

Variable	mean difference	effect size	MWU	KS test
PC	-0.087	-0.054	0.698	0.798
DTE	-0.083	-0.073	0.664	0.723

Table 4.8: Correlation between latent variables and decision variables by gender.

	Female				Male			
	Reservation price for data		Agree to sell the data		Reservation price for data		Agree to sell the data	
	full	partial	full	partial	full	partial	full	partial
PC	0.365 (0.00***)	0.337 (0.00***)	-0.468 (0.00***)	-0.309 (0.00***)	0.141 (0.125)	-0.005 (0.961)	-0.165 (0.241)	0.001 (0.988)
DTE	0.157 (0.101)	-0.046 (0.636)	-0.173 (0.167)	0.068 0.435	0.308 (0.00***)	0.300 (0.00***)	-0.350 (0.012**)	-0.191 (0.027**)

[1] Partial correlation: measures association between two variables controlling for the effect of the third.

with subjects' decision to agree to sell personal data, as well as with participants' reservation price for the data bundle. Table 4.8 presents these correlations separately for both gender and both latent variables. Since latent variables are also correlated, to account for potential confounding, we compute partial correlation coefficients. All but one full correlation coefficients are significant, yet the magnitude of this association differs. General privacy concerns have a stronger influence on women's decision to agree to sell and on the reservation price they demand in return. For male participants, DTE exhibits a significant and positive correlation with the reservation price and a significant and negative correlation with the binary decision variable whether to agree to sell the data. Gender differences become more pronounced when we examine partial correlation coefficients: controlling for PC, there is no statistically significant association between DTE and the main decision variables for female participants, the opposite holds for males. Thus, both latent variables could increase the explanatory power of our regression models and are therefore included in the subsequent analyses.

4.3.3 Regression Analysis

We can now extend our analysis of the willingness to sell personal data under financial discrimination using the two latent variables for privacy concerns and trust derived from survey responses. We perform regression analyses to model two decisions made by the participants: whether to agree to sell the data at all and which reservation price (RP) to chose for the data.

As the first stage, we examine what factors determine whether an individual agrees to sell the data. We fit a probit regression model where the dependent variable equals 1 if the subject does not tick the box "I don't want to sell the data in any case", i.e. the subject indicates a reservation price. Since both treatments do not feature any financial discrimination, we pool together Treatment (I) and Treatment (II), as discussed in Section 4.3.1.

Besides controlling for underlying privacy concerns and trust, we include further covariates, which are either derived from responses in the survey or participants' decisions in other parts of the experiment. For the covariates from survey responses, first, we control for e-banking usage, since it is related to effort costs of information disclosure. Furthermore, we include whether participants self-checked their personal data. The variable 'perceived control over data' is a binary variable derived from the answers to the survey question of whether the subjects agree that they control over how the personal data is used in the experiment.[23] The data nonsensitivity score measures how sensitive the personal data bundle is for the participant. For each personal data item, weight, height and bank account statement[24], the participants are asked how sensitive the item is on a scale of 1 (most sensitive) to 5 (least sensitive). The data nonsensitivity score corresponds to the sum of the answers, i.e. a higher value indicates a lower overall sensitivity for the data bundle. Figure 4.11 plots the nonsensitivity by gender. From Figure 4.11, it is apparent that the sensitivity of the data bundle does not differ importantly across gender.

We furthermore control for risk aversion[25], age (self-reported), family income[26], 5 CHF coin stage income and the net reservation price (RP) for the 5 CHF coin, which is calculated as the participant's reservation price for the 5 CHF coin - 5 CHF.

Table 4.9 shows the estimation results of probit models for the propensity to agreeing to sell the data. The regression results, showing no significant effect of the Treat-

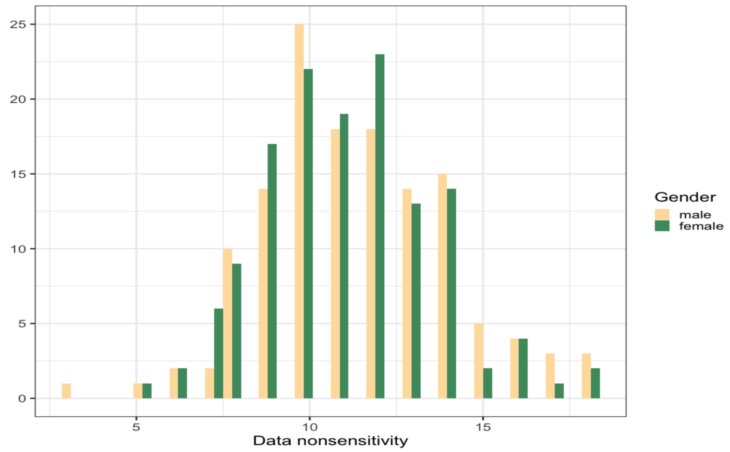
[23] The variable takes the value one if participants agree strongly, moderately or rather agree that they have control over how the personal data is used in the experiment.

[24] These are the three personal data items which are used for the categorization.

[25] Risk aversion is measured by the answer to the corresponding Falk et al. (2016) preference module.

[26] The family income is measured by the self-reported allocation into one of four income brackets in the post-experimental questionnaire.

Figure 4.11: Nonsensitivity score by gender.



ments (III)-(V), confirm the result from the previous sections that the general willingness to sell the data is not affected by data-based payoff discrimination. Model (2) shows a significant negative impact of the privacy concerns PC on the probability of agreeing to sell the data. However, when including gender and the interaction effects of gender with PC and DTE (Model 5), we can observe that the effect is moderated by gender: while PC is not statistically significant any more when including the interaction effects, we observe a significant negative interaction term of female and PC, suggesting that higher privacy concerns for females lead to a willingness to agree to sell the data. Interestingly, regarding trust towards the experimenter, the interaction term with female is positive, suggesting that for men a higher distrust translates into a lower willingness to agree to sell the data. We also find a significant positive impact of the perceived control over data and the data nonsensitivity, consistent with intuition. Furthermore, self-checking the data significantly increases the probability to agree to sell. This can be interpreted as an effect of risk/ambiguity about the content of the personal data on agreeing to sell. Interestingly, general risk aversion, as measured by the response to the Falk et al. (2016) preference module question on general risk preferences, is not significantly associated to the probability of agreeing to sell the personal data. Furthermore, we do not find an effect of prior experimental income (5 CHF coin income).

We now turn to the second-stage regression. Table 4.10 shows the OLS estimation results for the reservation price. Model (2), which includes the treatments and gender, confirms our results from Section 4.3.1: The coefficient of female is positive and statistically significant. We also find that the RP is weakly significantly lower in the Strong Negative Discrimination treatment (V) compared to the no-discrimination treatments (I+II). As seen in Section 4.3.1, this is due not only to the higher RP of revealed and unrevealed category C types, but also due to a higher RP of revealed category B types. Furthermore, we find a significant positive effect of DTE on the RP for men. General privacy concerns as captured by PC are the driver of an increase the RP for women. Thus, the gender-specific impact of privacy concerns and distrust towards the experimenter not only affect the general willingness to agree to sell the data, as seen in Table 4.9, but also the reservation price for the data. Interestingly, both perceived control and data nonsensitivity, which had a significant association with agreeing to sell the data at all, are not significant determinants of the reservation price for the data. The net RP 5 CHF, but not the first stage income, however is significantly positively related to the data reservation price. One interpretation of the net RP for the 5 CHF is that of a gambling motive: Although it is a dominant strategy to state the true valuation, participants in the 5 CHF selling and the data selling might try to state a high price to receive a higher selling gain. Importantly, all the effects discussed above, most importantly that of Treatment (V), are unaffected by controlling for this possible gambling motive.

The results from the regression analyses including privacy-related attitudes derived from the survey answers confirm the overview results from Section 4.3.1. In particular, subsequent data-based payoff discrimination does not affect the general willingness to agree to sell personal data, and a (weak) significant change in the reservation price is only observed when there is a strong negative payoff adjustment for one category. The regression analysis using the survey answers furthermore highlights the important gender-differentiated effect of privacy concerns—which are relevant for females— and distrust towards the experimenter—more relevant for males—for both the willingness to agree to sell the data as well as the reservation price.

Table 4.9: Probability of Agreeing to Sell the Personal Data (Probit) (Average marginal effects)

	Dependent variable: Agree to sell the data						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(III)	0.003	0.007	−0.011	−0.008	−0.024	−0.018	−0.020
	(0.253)	(0.256)	(0.260)	(0.263)	(0.268)	(0.286)	(0.294)
(IV)	0.069	0.070	0.067	0.068	0.060	0.035	0.024
	(0.283)	(0.284)	(0.297)	(0.298)	(0.305)	(0.334)	(0.342)
(V)	−0.065	−0.061	−0.059	−0.057	−0.074	−0.078	−0.048
	(0.242)	(0.243)	(0.253)	(0.254)	(0.260)	(0.284)	(0.304)
PC			−0.043***	−0.044***	−0.005	0.0001	−0.014
			(0.071)	(0.072)	(0.102)	(0.107)	(0.114)
DTE			−0.012	−0.012	−0.048	−0.057**	−0.042
			(0.101)	(0.102)	(0.142)	(0.153)	(0.160)
Female		−0.066		−0.065	0.110	0.111	0.086
		(0.190)		(0.197)	(0.669)	(0.708)	(0.759)
Female * PC					−0.080***	−0.064**	−0.043
					(0.151)	(0.160)	(0.169)
Female * DTE					0.083**	0.070*	0.038
					(0.205)	(0.220)	(0.234)
Self check: any						0.125***	0.114***
						(0.259)	(0.264)
Perceived control over data						0.131**	0.153**
						(0.251)	(0.266)
Data nonsensitivity						0.019**	0.019**
						(0.044)	(0.047)
E-banking usage: false						−0.048	−0.055
						(0.223)	(0.231)
Risk Aversion							−0.001
							(0.058)
Net RP 5CHF							0.006
							(0.107)
First stage (5 CHF coin) income							−0.011
							(0.061)
Age							−0.006
							(0.037)
Family income							0.018
							(0.112)
McFadden Pseudo R2	0.0182	0.0282	0.0760	0.0873	0.1193	0.2193	0.2352
Observations	271	270	271	270	270	267	262
Log Likelihood	−113.068	−111.756	−106.407	−104.959	−101.272	−89.392	−84.304
Akaike Inf. Crit.	234.136	233.512	224.815	223.918	220.543	204.784	204.609

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.10: Reservation Price for Personal Data (OLS)

	Dependent variable: Reservation price of Personal Data				
	(1)	(2)	(3)	(4)	(5)
(III)	−3.971	−4.611	−4.326	−3.936	−4.333
	(3.253)	(3.157)	(3.136)	(2.992)	(3.005)
(IV)	3.714	2.916	4.267	3.735	2.885
	(3.234)	(3.126)	(3.117)	(2.951)	(3.001)
(V)	6.567*	5.831*	5.269	6.129*	5.421*
	(3.379)	(3.264)	(3.295)	(3.145)	(3.195)
PC			1.566*	−0.030	−0.172
			(0.847)	(1.047)	(1.048)
DTE			3.443**	5.405***	5.593***
			(1.368)	(1.702)	(1.703)
Female		9.359***		3.382	2.096
		(2.331)		(6.567)	(6.576)
Female*PC				4.929***	4.308***
				(1.650)	(1.632)
Female *DTE				−6.393**	−5.332**
				(2.593)	(2.593)
Self check: any					−1.247
					(2.331)
E-banking usage: false					−3.003
					(2.385)
Perceived control over data					−0.367
					(2.857)
Data nonsensitivity					−0.129
					(0.434)
Risk aversion					−0.213
					(0.577)
Net RP 5CHF					3.855***
					(1.090)
First stage (5 CHF coin) income					−0.473
					(0.644)
Age					0.308
					(0.384)
Constant	24.094***	19.947***	10.107***	7.918*	7.468
	(2.050)	(2.231)	(3.805)	(4.709)	(12.333)
Observations	230	229	230	229	226
R^2	0.040	0.109	0.118	0.223	0.284
Adjusted R^2	0.028	0.093	0.099	0.194	0.229
Note:			*p<0.1; **p<0.05; ***p<0.01		

4.4 Conclusion

We provide the results of an experiment on the value of personal data in light of data-based price discrimination. We find that in our experiment the general willingness to sell personal data is not significantly affected by subsequent data-based payoff discrimination. Furthermore, we only find a significant change in the reservation price of the data when one data-based category implies a strong decrease of the subsequent payoff. Interestingly, the change in the reservation price is not only driven by participants who fall into this category, but a general increase of the reservation price under strong negative price discrimination. We observe this effect even though the data-based categorization is fully known and risk/ambiguity-free.

The bundle of personal data that the participants can sell consists of their height, weight, gender, information on their bank account balance and a photo of their face. A comparison of the self-reported sensitivity of this personal data bundle shows that there are no important differences across gender how sensitive this data is perceived to be. Nevertheless, we find important gender differences in how general privacy concerns and trust related to the context of the experiment affect both the general willingness to sell the data as well as the reservation price of the data. These findings are important in light of the consequences of personal data sharing for subsequent market interaction, not only with respect to price discrimination, but the usage of personal data more generally.

In our setting, the rules of how the personal data is used were fully known and transparent: Both the data-based categorization was known to participants and exogenous, as well as the attached payoff consequences. When sharing personal data in online markets, however both how this data is interpreted, e.g. by algorithms, as well as the (financial) consequences are much less transparent, and often fully ambiguous. It is an interest avenue for future research to analyze how ambiguity about the informational content of personal data and the attached consequences affect the willingness to share this data.

5. Concluding remarks

> *Not everything that counts can be counted, and not everything that can be counted counts.*
>
> William Bruce Cameron / Albert Einstein

This thesis comprises three empirical studies pertaining to Big Data technologies. In Chapter 2 and Chapter 3, I explored a telematics dataset to determine what new insights can we get from this data and what practical implications they could have. Chapter 4 switched the focus to data subjects. We conducted a laboratory experiment to explore the impact of financial incentives on the decision to share personal information. Today, there is little doubt that Big Data technologies will be widely adopted in different industries, yet their role and impact are yet unclear. This thesis contributes to the analysis of the role of Big Data technologies and shows promising avenues for future research in this area.

The experimental setting of Table 4 can be extended in several ways to yield further insights into privacy-related decision making. First, the study participants knew the effect of disclosing personal information. In practice, useful insights from large datasets are extracted by Machine Learning algorithms. These algorithms are well described by the “black box" metaphor: neither the person who has created, nor the one who has implemented an algorithm can predict with certainty which output will be obtained from the input. With that in mind, it would be instructive to incorporate risk, ambiguity, or even uncertainty regarding financial consequences into the setting. Second, one could modify the payoff structure. The magnitude of price discrimination did not exceed 20 CHF. As the gap between category-dependent payoffs increases, I expect financial incentives to play a more prominent role.

Our results suggest that non-financial factors play a significant role in privacy-related decisions. Bearing in mind that information is the main unit of currency in a data-driven economy, it is interesting to further explore what could prompt individuals to share it. Kehr et al. (2015a) show that positive affect, elicited by a well-designed user interface, influences privacy assessment. Examining the role of general mental attitudes, such as optimism and pessimism, and the Big Five personality traits is a promising endeavor.

A unifying objective of Big Data technologies is to extract insights and economic value from a variety of data sources. If a market participant prevails in inferring relevant information, this can lead to a power shift on the market. Markets with informational problems are expected to be more affected by the change. In this vein the following questions arise: (1) whether and if yes how much information is disclosed in the market equilibrium (2) how will privacy concerns affect the outcomes and (3) whether these markets should be regulated.

On the markets, where collecting more personal data yields significant competitive advantage, privacy concerns of data subjects are likely to affect not only the market equilibrium but also the market structure. There is a perceived link between a company's size and the probability of a data breach. Consequently, privacy-sensitive individuals are more likely to disclose pertinent information to well-established market players. Campbell et al. (2015) develop a theoretical model to demonstrate that requiring companies to obtain an individual's consent for using his information disproportionately affects smaller companies and new entrants. Designing optimal policies that are able to balance the interests of various stakeholders is an important open challenge.

The premise of the discussion above is that market players can realize the potential of the new data sources. However, many companies lack the time or the intellectual capacity to do it. Outsourcing data analytics is a potential remedy but not a solution. Another big challenge is the *selection bias*: if the decision to disclose the data is correlated with its content, the out-of sample predictive power of the algorithm is damaged. Consider as an example *usage based* insurance: risk scores from driving logs were developed by a third-party telematics company. The algorithm punished drivers for elevated g-force events and night driving. Yet we do not find a statistically significant relation between these characteristics and subsequent accident involvement. We cannot establish whether this result is attributed to the *selection bias* or to deficiencies of risk modeling. In this regard, due to a rich talent pool and lower potential for conflict of interest, the academic community might be in a better position to explore various use cases of new datasets.

The discussion about the future of Big Data has moved from *whether* to *when*. Optimistic scenarios promise a higher life standard and a decrease in the welfare gap between developing and developed countries (Schwab (2017)). Sceptics predict a further increase in social inequality (O'Neil (2017))[1] and a power shift towards big corporations and governments. Yet, a concerted effort of researchers with backgrounds in economics, computer science and psychology could gear the society towards better outcomes.

[1] As an example: the errors, and imprecisions in ML predictions are much more likely to affect low level employees and lower income citizens. For more privileged individuals and valuable specialist, a human is more likely to enter the decision making loop and correct for algorithmic errors.

Bibliography

Alessandro Acquisti, Leslie K John, and George Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, 2013.

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–92, 2016.

Charu C Aggarwal. *Data Mining: the Textbook*. Springer, 2015.

Hamed Ahangari, Jason Outlaw, Carol Atkinson-Palombo, and Norman W. Garrick. Investigation into impact of fluctuations in gasoline prices and macroeconomic conditions on road safety in developed countries. *Transportation Research Record*, 2465 (1):48–56, 2014. doi: 10.3141/2465-07.

Mercedes Ayuso, Montserrat Guillén, and Manuela Alcañiz. The impact of traffic violations on the estimated cost of traffic accidents with victims. *Accident Analysis & Prevention*, 42(2):709–717, 2010.

Mercedes Ayuso, Montserrat Guillén, and Ana María Pérez-Marín. Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention*, 73:125–131, 2014.

Mercedes Ayuso, Montserrat Guillen, and Ana María Pérez-Marín. Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2):10, 2016.

Susanne Barth and Menno DT de Jong. The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and Informatics*, 2017.

Gordon M Becker, Morris H DeGroot, and Jacob Marschak. Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232, 1964.

Paul Belleflamme and Wouter Vergote. Monopoly price discrimination and privacy: The hidden cost of hiding. *Economics Letters*, 149:141–144, 2016.

Itzhak Benenson, Karel Martens, and Slava Birfir. Parkagent: An agent-based model of parking in the city. *Computers, Environment and Urban Systems*, 32(6):431–439, 2008.

Volker Benndorf and Hans-Theo Normann. The willingness to sell personal data. *The Scandinavian Journal of Economics*, 2017.

Alastair R Beresford, Dorothea Kübler, and Sören Preibusch. Unwillingness to pay for privacy: A field experiment. *Economics Letters*, 117(1):25–27, 2012.

Ruth Bergel-Hayat, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. Explaining the road accident risk: Weather effects. *Accident Analysis & Prevention*, 60:456–465, 2013.

BFS. Death and its main causes in switzerland, 2016, 2016.

Jason Bordoff and Pascal Noel. Pay-as-you-drive auto insurance: A simple way to reduce driving-related harms and increase equity. *Hamilton Project Discussion Paper*, 2008.

Harold Brodsky and A Shalom Hakkert. Risk of a road accident in rainy weather. *Accident Analysis & Prevention*, 20(3):161–176, 1988.

Ivan D Brown. Driver fatigue. *Human Factors*, 36(2):298–314, 1994.

Bridget RD Burdett, Samuel G Charlton, and Nicola J Starkey. Mind wandering during everyday driving: An on-road study. *Accident Analysis & Prevention*, 122: 76–84, 2019.

James Campbell, Avi Goldfarb, and Catherine Tucker. Privacy regulation and market structure. *Journal of Economics & Management Strategy*, 24(1):47–73, 2015.

Samuel G Charlton and Nicola J Starkey. Driving on familiar roads: Automaticity and inattention blindness. *Transportation Research Part F: Traffic Psychology and Behaviour*, 19:121–133, 2013.

Irene G Chen, Dennis R Durbin, Michael R Elliott, Michael J Kallan, and Flaura Koplin Winston. Trip characteristics of vehicle crashes involving child passengers. *Injury Prevention*, 11(4):219–224, 2005.

Guangqing Chi, Arthur G Cosby, Mohammed A Quddus, Paul A Gilbert, and David Levinson. Gasoline prices and traffic safety in mississippi. *Journal of Safety Research*, 41(6):493–500, 2010.

Pierre-Andre Chiappori and Bernard Salanie. Testing for asymmetric information in insurance markets. *Journal of Political Economy*, 108(1):56–78, 2000.

Chongwoo Choe, Stephen King, and Noriaki Matsushima. Pricing with cookies: Behavior-based price discrimination and spatial competition. *Management Science*, 64(12):5669–5687, March 2019. ISSN 0025-1909. doi: 10.1287/mnsc.2017.2873.

Alma Cohen. Asymmetric information and learning: Evidence from the automobile insurance market. *Review of Economics and Statistics*, 87(2):197–207, 2005.

Alma Cohen and Liran Einav. Estimating risk preferences from deductible choice. *The American Economic Review*, 97(3):745–788, 2007.

Anna B Costello and Jason W Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7):1–9, 2005.

David Crainich. Self-insurance with genetic testing tools. *Journal of Risk and Insurance*, 84(1):73–94, 2017.

Keith J Crocker and Arthur Snow. The theory of risk classification. In *Handbook of Insurance*, pages 281–313. Springer, 2013.

David De Meza and David C Webb. Advantageous selection in insurance markets. *RAND Journal of Economics*, pages 249–262, 2001.

David M DeJoy. The optimism bias and traffic accident risk perception. *Accident Analysis & Prevention*, 21(4):333–340, 1989.

David F Dinges. An overview of sleepiness and accidents. *Journal of Sleep Research*, 4:4–14, 1995.

Thomas A Dingus, Sheila G Klauer, Vicki L Neale, A Petersen, Suzanne E Lee, JD Sudweeks, MA Perez, J Hankey, DJ Ramsey, S Gupta, et al. The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment. Technical report, 2006.

Georges Dionne and Casey Rothschild. Economic effects of risk classification bans. *The Geneva Risk and Insurance Review*, 39(2):184–221, 2014.

Georges Dionne, Christian Gouriéroux, and Charles Vanasse. Testing for evidence of adverse selection in the automobile insurance market: A comment. *Journal of Political Economy*, 109(2):444–453, 2001.

Georges Dionne, Jean Pinquet, Mathieu Maurice, and Charles Vanasse. Incentive mechanisms for safe driving: a comparative analysis with dynamic data. *The Review of Economics and Statistics*, 93(1):218–227, 2011.

Aaron S Edlin. Per-mile premiums for auto insurance. Technical report, National Bureau of Economic Research, 1999.

Ariel Ezrachi and Maurice E. Stucke. The rise of behavioural discrimination. *European Competition Law Review*, 37(12):484–491, 2016.

Armin Falk, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. The preference survey module: A validated instrument for measuring risk, time, and social preferences. 2016.

Francesco Feri, Caterina Giannetti, and Nicola Jentzsch. Disclosure of personal information under risk of privacy shocks. *Journal of Economic Behavior & Organization*, 123:138–148, 2016.

Antonio Filippin and Paolo Crosetto. A reconsideration of gender differences in risk attitudes. *Management Science*, 62(11):3138–3160, May 2019. ISSN 0025-1909. doi: 10.1287/mnsc.2015.2294.

Amy Finkelstein and James Poterba. Testing for asymmetric information using “unused observables" in insurance markets: Evidence from the uk annuity market. *Journal of Risk and Insurance*, 81(4):709–734, 2014.

Urs Fischbacher. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178, 2007.

David M Grether and Charles R Plott. Economic theory of choice and the preference reversal phenomenon. *The American Economic Review*, 69(4):623–638, 1979.

Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. The age of analytics: Competing in a data-driven world. *McKinsey Global Institute*, 4, 2016.

Charles A Holt and Susan K Laury. Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655, 2002.

James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

Lars Hultkrantz and Gunnar Lindberg. Pay-as-you-speed an economic field experiment. *Journal of Transport Economics and Policy (JTEP)*, 45(3):415–436, 2011.

Paolo Intini, Pasquale Colonna, Nicola Berloco, and Vittorio Ranieri. The impact of route familiarity on drivers’ speeds, trajectories and risk perception. In *17th International Conference Road Safety On Five Continents (RS5C 2016), Rio de Janeiro, Brazil, 17-19 May 2016*, page 12. Statens väg-och transportforskningsinstitut, 2016.

Tobias Ippisch. *Telematics Data in Motor Insurance: Creating Value by Understanding the Impact of Accidents on Vehicle Use*. PhD thesis, Citeseer, 2010.

Flavius Kehr, Tobias Kowatsch, Daniel Wentzel, and Elgar Fleisch. Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal*, 25(6):607–635, 2015a.

Flavius Kehr, Tobias Kowatsch, Daniel Wentzel, and Elgar Fleisch. Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal*, 25(6):607–635, 2015b.

Jana Kern, Benjamin Fabian, and Tatiana Ermakova. Experimental privacy studies-an initial review of the literature. 2018.

Daniela Kremslehner and Alexander Muermann. Asymmetric information in automobile insurance: Evidence from driving behavior. 2016.

Ola Kvaløy, Miguel Luzuriaga, and Trond E Olsen. A trust game in loss domain. *Experimental Economics*, 20(4):860–877, 2017.

Jean Lemaire. *Automobile Insurance: Actuarial Models*, volume 4. Springer Science & Business Media, 2013.

Krsto Lipovac, Miroslav Deric, Milan Tešić, Zoran Andrić, and Bojan Marić. Mobile phone use while driving-literary review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 47:132–142, 2017.

Todd Litman. Distance-based vehicle insurance as a tdm strategy. *Transportation Quarterly*, 51:119–137, 1997.

James M Lyznicki, Theodore C Doege, Ronald M Davis, Michael A Williams, et al. Sleepiness, driving, and motor vehicle crashes. *Jama*, 279(23):1908–1913, 1998.

Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

Helia Marreiros, Mirco Tonin, Michael Vlassopoulos, and MC Schraefel. "now that you mention it": A survey experiment on information, inattention and online privacy. *Journal of Economic Behavior & Organization*, 140:1–17, 2017.

John Milton and Fred Mannering. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation*, 25(4):395–413, 1998.

Rodrigo Montes, Wilfried Sand-Zantman, and Tommaso Valletti. The value of personal information in online markets with endogenous privacy. *Management Science*, March 2018. ISSN 0025-1909. doi: 10.1287/mnsc.2017.2989. URL https://doi.org/10.1287/mnsc.2017.2989.

Peter Mooney, Marco Minghini, et al. A review of openstreetmap data. 2017.

Alexander Muermann, Alois Geyer, and Daniela Kremslehner. Asymmetric information in automobile insurance: Evidence from driving behavior. *Journal of Risk and Insurance*, pages 1–27, 2019.

New York Times. How Trump consultants exploited the Facebook data of millions. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html, 2018. Last accessed on Sep 30, 2018.

Stuart Newstead and Angelo D'Elia. An investigation into the relationship between vehicle colour and crash risk. *Prevention*, 17(1):47–56, 2007.

Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

Raymond C Peck, Michael A Gebers, Robert B Voas, and Eduardo Romano. The relationship between blood alcohol concentration (bac), age, and crash risk. *Journal of Safety Research*, 39(3):311–319, 2008.

Pierre Philip, Patricia Sagaspe, Nicholas Moore, Jacques Taillard, André Charles, Christian Guilleminault, and Bernard Bioulac. Fatigue, sleep restriction and driving performance. *Accident Analysis & Prevention*, 37(3):473–478, 2005.

Joachim Plesch and Irenaeus Wolff. Personal-data disclosure in a field experiment: Evidence on explicit prices, political attitudes, and privacy preferences. *Games*, 9 (2), 2018.

Michael I Posner, Charles R Snyder, and R Solso. Attention and cognitive control. *Cognitive Psychology: Key Readings*, 205, 2004.

Tobias Regner and Gerhard Riener. Privacy is precious: On the attempt to lift anonymity on the internet to increase revenue. *Journal of Economics & Management Strategy*, 26(2):318–336, 2017.

Timothy J Richards, Jura Liaukonyte, and Nadia A Streletskaya. Personalized pricing and price fairness. *International Journal of Industrial Organization*, 44:138–153, 2016.

Kuniyoshi Saito. Testing for asymmetric information in the automobile insurance market under rate regulation. *Journal of Risk and Insurance*, 73(2):335–356, 2006.

Alvaro Sandroni and Francesco Squintani. Overconfidence and asymmetric information: The case of insurance. *Journal of Economic Behavior & Organization*, 93:149–165, 2013.

Simeon Schudy and Verena Utikal. 'you must not know about me'—on the willingness to share personal data. *Journal of Economic Behavior & Organization*, 141:1–13, 2017.

Klaus Schwab. *The Fourth Industrial Revolution*. Currency, 2017.

Bruce G Simons-Morton, Marie Claude Ouimet, Zhiwei Zhang, Sheila E Klauer, Suzanne E Lee, Jing Wang, Paul S Albert, and Thomas A Dingus. Crash and risky driving involvement among novice adolescent drivers and their parents. *American Journal of Public Health*, 101(12):2362–2367, 2011.

H Jeff Smith, Sandra J Milberg, and Sandra J Burke. Information privacy: measuring individuals' concerns about organizational practices. *MIS quarterly*, pages 167–196, 1996.

Anastasia Sycheva, Wanda Mimra, and Christian Waibel. Driving behavior and risk classification in auto insurance: Evidence from telematics data. 2019.

Jim Uttley and Steve Fotios. The effect of ambient light condition on road traffic collisions involving pedestrians on pedestrian crossings. *Accident Analysis & Prevention*, 108:189–200, 2017.

Peter van der Waerden, Harry Timmermans, and Marloes de Bruin-Verhoeven. Car drivers' characteristics and the maximum walking distance between parking facility and final destination. *Journal of Transport and Land Use*, 10(1):1–11, 2017.

William Vickrey. Automobile accidents, tort law, externalities, and insurance: an economist's critique. *Law and Contemporary Problems*, 33(3):464–487, 1968.

Mark Vollrath, Tobias Meilinger, and Hans-Peter Krüger. How the presence of passengers influences the risk of a collision with another vehicle. *Accident Analysis & Prevention*, 34(5):649–654, 2002.

Eric-Jan Wagenmakers and Simon Farrell. Aic model selection using akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, 2004.

Per Ole Wanvik. Effects of road lighting: an analysis based on dutch accident statistics 1987–2006. *Accident Analysis & Prevention*, 41(1):123–128, 2009.

Allan F Williams. Young driver risk factors: successful and unsuccessful approaches for dealing with them and an agenda for the future. *Injury Prevention*, 12(suppl 1): i4–i8, 2006.

World Health Organization. Global Health Observatory (GHO) data. http://www.who.int/gho/ncd/risk_factors/bmi_text/en/, 2018a. Last accessed on Oct 24, 2018.

World Health Organization. Mean body mass index trends among adults, crude Estimates by country. http://apps.who.int/gho/data/view.main.BMIMEANADULTCv?lang=en, 2018b. Last accessed on Oct 24, 2018.

Matthew R Yanko and Thomas M Spalek. Route familiarity breeds inattention: A driving simulator study. *Accident Analysis & Prevention*, 57:80–86, 2013.

George Yannis, John Golias, and Eleonora Papadimitriou. Accident risk of foreign drivers in various road environments. *Journal of Safety Research*, 38(4):471–480, 2007.

Appendix

Appendix to Chapter 2

Additional Figures and Tables

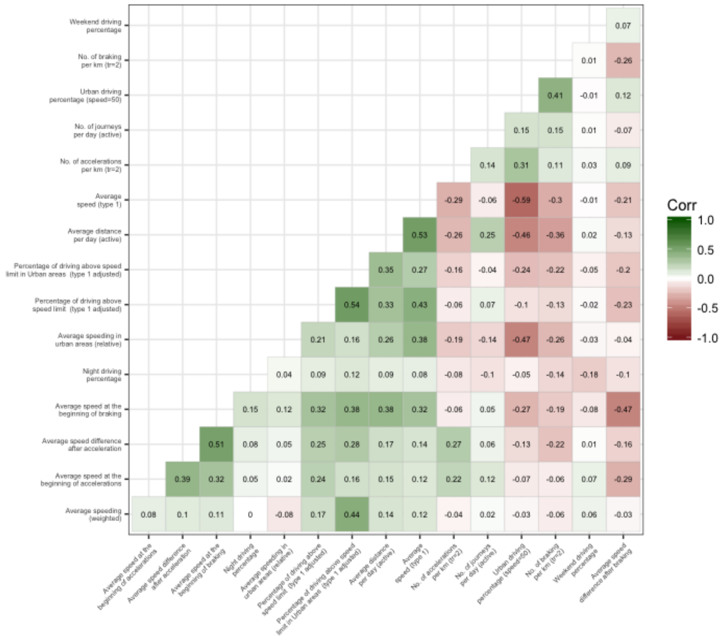
Table 5.1: Notation

Notation	Explanation
$type$	type of elevated g-force event (acceleration, harsh braking, cornering, speeding)
d	time of event $d \in \{Rush, Night, Other\}$
r	road type on which the event is registered $r \in \{Urban, Motorway, Other\}$
$dist_{t,r}$	total weekly distance driven on road type r during daytime d
v_t	speed at the time of the event
sev	severity of the event $sev \in \{Red, Yellow\}$
$\#Ev^{type}_{r,d,v_t,sev}$	number of events with certain type and severity that took place during ceratin daytime and on certain road type
w_l	$l \in \{d, r, v_t, sev\}$ weight of the event due to the factors above
$w^{type}_{d,r,v_t,sev}$	resulting weight of the event
$AbsSc(w)$	Absolute weekly driving score
$NormSc(w)$	Normalized weekly driving score

Table 5.2: Summary of the driver / liability claim filtering process

	Policyholders	**Liability claims**
Original datasets	**9244**	**1799**
Data from 2016	-1775	-1095
Driving logs available ?	-1479	-223
Multiple liability claims during 2016	NA	-10
Claim size > 10 CHF	NA	-77
Younger than 26	-213	-18
No. of daily observations in 2016 $\geq$ 10 ?	- 87	-5
Data used in the regression analysis	**5690**	**371**

Table 5.3: Correlation table (left) and variance inflation factors (right) for the regression models



		VIF	
Equation	**(1)**	**(2)**	**(3)**
Average distance per day (active)	2.207	2.233	2.154
No. of journeys per day (active)	1.421	1.370	1.350
Average speed (type 1)	2.436	2.331	2.319
Weekend driving percentage	1.113	1.123	1.126
Urban driving percentage (speed=50)	2.461	2.428	2.386
Night driving percentage	1.133	1.115	1.136
Percentage of driving above speed limit (type 1 adjusted)	2.017	1.913	2.021
Average speeding (weighted)	1.396	1.340	1.361
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	2.172	2.050	2.159
Average speeding in urban areas (relative)	1.415	1.433	1.442
No. of accelerations per km (tr=2)	1.522	1.536	1.522
No. of braking per km (tr=2)	1.812	1.798	1.785
Average speed at the beginning of accelerations	1.544	1.551	1.415
Average speed difference after acceleration	2.182	2.132	2.183
Average speed difference after braking	1.791	1.849	1.777
Average speed at the beginning of braking	2.327	2.279	2.237

Figure 5.1: Receiver Operating Characteristics Curve (ROC) illustrating predictive power of probit models of annual submission of liability claims exceeding 10 CHF with and without telematics based predictors.

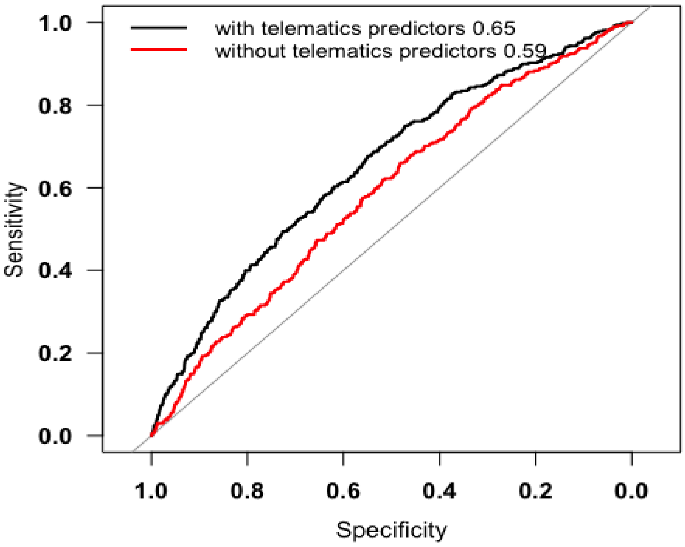


Table 5.4: Average marginal effects computed from the trivariate model.

	liability claim ($> 10CHF$)	BM protection	Collision cover
Traditional Insurance Variables			
Sex of driver: female	0.004	0.019^{**}	0.054^{***}
	(0.059)	(0.056)	(0.059)
Age of driver	-0.006^{***}	0.000	0.018^{***}
	(0.019)	(0.020)	(0.018)
Vehicle: age	0.002^{***}	-0.001	-0.043^{***}
	(0.007)	(0.007)	(0.011)
Vehicle: horsepower	-0.000	-0.000^{***}	0.000
	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.000	-0.000	-0.000^{***}
	(0.000)	(0.000)	(0.000)
Vehicle: price	-0.000	0.000^{***}	0.000^{***}
	(0.000)	(0.000)	(0.000)
Vehicle: mileage	-0.000	-0.000^{***}	-0.001^{***}
	(0.001)	(0.000)	(0.001)
Leasing contract: true	-0.003	0.024^{**}	0.179^{***}
	(0.083)	(0.085)	(0.133)
Recent change of address: true	0.015	-0.012	-0.004
	(0.070)	(0.067)	(0.072)
Bonus Malus Score TPL	-0.001	0.008^{**}	0.003
	(0.024)	(0.025)	(0.023)
No. of years without first contract	0.004^{*}	-0.011^{***}	-0.011^{***}
	(0.017)	(0.019)	(0.017)
No. of previous mobility claims	-0.005	0.025^{**}	-0.006
	(0.071)	(0.084)	(0.063)
Telematics-Based Predictors			
Average distance per day (active)	0.001^{***}	-0.001^{***}	-0.001^{***}
	(0.002)	(0.001)	(0.002)
No. of journeys per day (active)	0.014^{***}	0.017^{***}	0.013^{***}
	(0.027)	(0.030)	(0.030)
Average speed (type 1)	-0.000	0.000	0.000
	(0.003)	(0.003)	(0.003)
Weekend driving percentage	-0.055^{*}	-0.041	-0.052
	(0.241)	(0.208)	(0.220)
Urban driving percentage (speed=50)	-0.015	0.034	-0.026
	(0.300)	(0.294)	(0.290)
Night driving percentage	0.072	0.191^{***}	-0.031
	(0.450)	(0.461)	(0.450)
Percentage of driving above speed limit (type 1 adjusted)	0.074^{*}	0.030	0.100^{*}
	(0.348)	(0.332)	(0.350)
Average speeding (weighted)	0.003^{**}	-0.003^{***}	-0.001
	(0.009)	(0.008)	(0.009)
Percentage of driving above speed limit in urban areas (type 1 adjusted)	-0.047	0.021	-0.004
	(0.292)	(0.272)	(0.283)
Average speeding in urban areas (relative)	-0.013	0.011	-0.001
	(0.109)	(0.098)	(0.100)
No. of accelerations per km (tr=2)	0.018	-0.027	-0.040
	(0.178)	(0.179)	(0.175)
No. of braking per km (tr=2)	0.020	-0.007	-0.009
	(0.179)	(0.177)	(0.181)
Average speed at the beginning of accelerations	-0.000	-0.001	-0.002
	(0.013)	(0.013)	(0.013)
Average speed difference after acceleration	0.000	0.003^{***}	0.003^{***}
	(0.006)	(0.006)	(0.006)
Average speed difference after braking	0.001	-0.002	0.003
	(0.018)	(0.017)	(0.017)
Average speed at the beginning of braking	0.000	0.000	-0.001
	(0.007)	(0.007)	(0.007)
McFadden Pseudo R2	0.036	0.093	0.616
Maximum VIF	3.23	3.27	4.17
Observations	5,690	5,690	5,690
Akaike Inf. Crit.	2,685.674	2,977.527	3,083.956

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

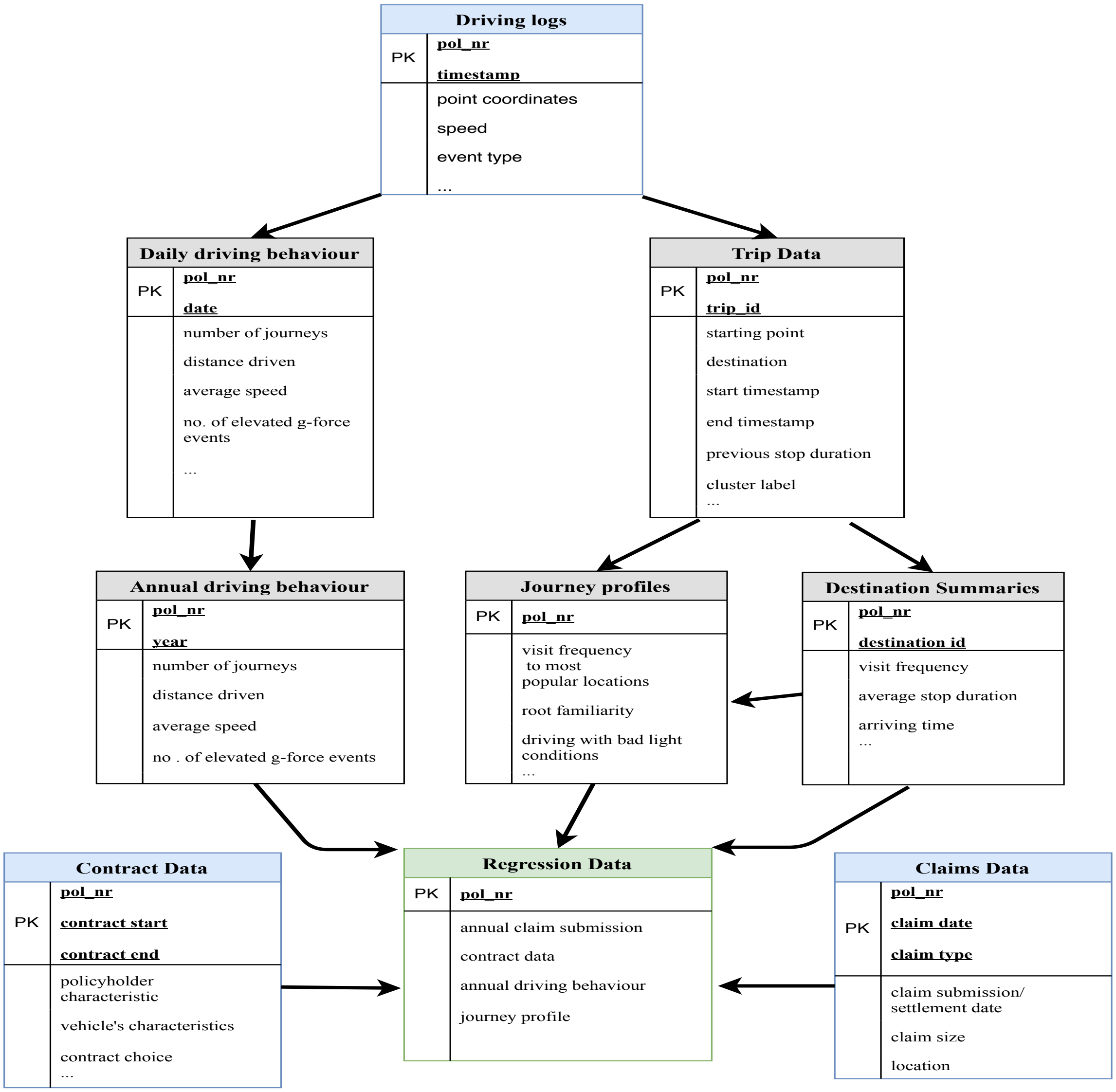
Table 5.5: Coefficients of fitted probit models

	Drivers older than 22			Drivers younger than 22			Drivers with contract since less than 2 years		
	liability claim (> 10CHF)	BM protection	Collision cover	liability claim (> 10CHF)	BM protection	Collision cover	liability claim (> 10CHF)	BM protection	Collision cover
Traditional Insurance Variables									
Sex of driver: female	0.104	0.160**	0.274***	−0.046	0.127	0.498***	−0.047	0.101	0.368***
	(0.082)	(0.072)	(0.077)	(0.086)	(0.092)	(0.093)	(0.073)	(0.070)	(0.077)
Age of driver	−0.053	0.021	0.119***	−0.039	−0.022	0.080*	−0.008	−0.006	0.052
	(0.034)	(0.032)	(0.031)	(0.048)	(0.055)	(0.049)	(0.056)	(0.053)	(0.057)
Vehicle: age	0.021**	−0.017*	−0.284***	0.022**	0.000	−0.299***	0.023**	−0.007	−0.309***
	(0.011)	(0.009)	(0.015)	(0.010)	(0.010)	(0.017)	(0.009)	(0.008)	(0.016)
Vehicle: horsepower	−0.000	−0.001	0.001	−0.001	−0.004***	−0.000	−0.000	−0.002*	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vehicle: weight	0.000	−0.000	−0.001***	−0.000	−0.000	−0.000	0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Vehicle: price	−0.000	0.000***	0.000***	−0.000	0.000***	0.000***	−0.000	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Vehicle: mileage	−0.000	−0.002**	−0.003***	−0.001	−0.001	−0.006***	−0.001	−0.002***	−0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leasing contract: true	−0.173	0.272***	1.260***	0.253*	−0.013	1.140***	0.043	0.242**	1.283***
	(0.108)	(0.104)	(0.173)	(0.132)	(0.150)	(0.219)	(0.108)	(0.112)	(0.213)
Recent change of address: true	0.113	−0.034	−0.006	0.105	−0.149	−0.025	0.015	−0.006	0.006
	(0.090)	(0.083)	(0.089)	(0.113)	(0.117)	(0.126)	(0.095)	(0.088)	(0.102)
Bonus Malus Score TPL	0.001	0.051*	0.032	−0.016	0.082*	0.016	−0.047	0.056	−0.079
	(0.031)	(0.031)	(0.029)	(0.039)	(0.048)	(0.040)	(0.056)	(0.053)	(0.058)
No. of years without first contract	0.038**	−0.079***	−0.074***	−0.023	−0.096*	−0.094*	−0.011	−0.094*	−0.026
	(0.018)	(0.021)	(0.018)	(0.049)	(0.052)	(0.048)	(0.057)	(0.054)	(0.058)
No. of previous mobility claims	−0.054	0.220**	−0.019	−0.062	0.096	−0.094	−0.127	−0.039	0.328*
	(0.084)	(0.099)	(0.071)	(0.133)	(0.161)	(0.153)	(0.179)	(0.160)	(0.180)
Telematics-Based Predictors									
Average distance per day (active)	0.005***	−0.009***	−0.006***	0.004	−0.009***	−0.007***	0.004**	−0.009***	−0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
No. of journeys per day (active)	0.087**	0.126***	0.070*	0.138***	0.127***	0.121***	0.125***	0.119***	0.119***
	(0.038)	(0.039)	(0.040)	(0.039)	(0.048)	(0.046)	(0.033)	(0.036)	(0.039)
Average speed (type 1)	−0.005	0.005	0.002	0.002	−0.002	0.002	0.002	0.003	0.004
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.003)	(0.004)
Weekend driving percentage	−0.533	−0.165	−0.437	−0.362	−0.405	−0.232	−0.569*	−0.029	−0.367
	(0.334)	(0.265)	(0.291)	(0.355)	(0.346)	(0.348)	(0.298)	(0.258)	(0.283)
Urban driving percentage (speed=50)	−0.179	0.687*	0.139	−0.084	−0.339	−0.760*	0.002	0.278	−0.089
	(0.425)	(0.388)	(0.403)	(0.435)	(0.462)	(0.429)	(0.392)	(0.377)	(0.390)
Night driving percentage	1.074*	1.129**	−0.291	−0.002	1.890**	−0.069	0.753	1.661***	−0.446
	(0.601)	(0.574)	(0.587)	(0.690)	(0.799)	(0.712)	(0.558)	(0.584)	(0.581)
Percentage of driving above speed limit (type 1 adjusted)	0.731	0.388	0.390	0.479	−0.072	1.136**	0.318	−0.284	0.607
	(0.475)	(0.420)	(0.466)	(0.525)	(0.561)	(0.551)	(0.439)	(0.411)	(0.461)
Average speeding (weighted)	0.021*	−0.041***	−0.002	0.026**	−0.001	−0.008	0.009	−0.022**	−0.017
	(0.012)	(0.011)	(0.012)	(0.013)	(0.014)	(0.013)	(0.011)	(0.010)	(0.012)
Percentage of driving above speed limit in Urban areas (type 1 adjusted)	−0.120	0.099	−0.219	−0.617	0.225	0.117	−0.080	0.244	−0.085
	(0.399)	(0.347)	(0.376)	(0.439)	(0.460)	(0.434)	(0.372)	(0.347)	(0.386)
Average speeding in urban areas (relative)	−0.142	0.179	0.326**	−0.074	−0.101	−0.556***	−0.038	0.069	−0.090
	(0.146)	(0.126)	(0.132)	(0.167)	(0.158)	(0.175)	(0.132)	(0.120)	(0.128)
No. of accelerations per km (tr=2)	0.182	−0.102	−0.372	0.148	−0.285	−0.238	0.279	−0.015	−0.462**
	(0.256)	(0.242)	(0.254)	(0.253)	(0.271)	(0.247)	(0.220)	(0.233)	(0.229)
No. of braking per km (tr=2)	0.269	−0.032	0.051	0.070	−0.068	−0.199	0.131	−0.102	−0.078
	(0.244)	(0.231)	(0.249)	(0.267)	(0.283)	(0.270)	(0.220)	(0.217)	(0.230)
Average speed at the beginning of accelerations	0.014	−0.012	−0.009	−0.020	0.008	−0.023	0.000	−0.036**	−0.010
	(0.017)	(0.016)	(0.017)	(0.020)	(0.021)	(0.021)	(0.016)	(0.016)	(0.018)
Average speed difference after acceleration	0.002	0.008	0.021***	0.003	0.035***	0.024***	0.000	0.026***	0.013
	(0.008)	(0.008)	(0.008)	(0.009)	(0.010)	(0.009)	(0.008)	(0.007)	(0.008)
Average speed difference after braking	−0.001	−0.003	0.023	0.016	−0.031	0.017	0.014	−0.020	0.042*
	(0.026)	(0.022)	(0.024)	(0.025)	(0.028)	(0.025)	(0.023)	(0.022)	(0.023)
Average speed at the beginning of braking	−0.010	0.014*	−0.000	0.013	−0.015	−0.021**	−0.002	−0.003	−0.009
	(0.010)	(0.008)	(0.009)	(0.010)	(0.011)	(0.011)	(0.009)	(0.008)	(0.009)
McFadden Pseudo R2	0.043	0.113	0.608	0.043	0.079	0.612	0.038	0.109	0.631
Maximum VIF	3.31	3.05	4.3	3.13	3.14	3.91	14.82	14.67	15.01
Observations	3,126	3,126	3,126	2,564	2,564	2,564	3,430	3,430	3,430
Akaike Inf. Crit.	1,375.359	1,823.578	1,722.426	1,345.147	1,171.406	1,384.372	1,740.860	1,957.241	1,807.897
Notes:									***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

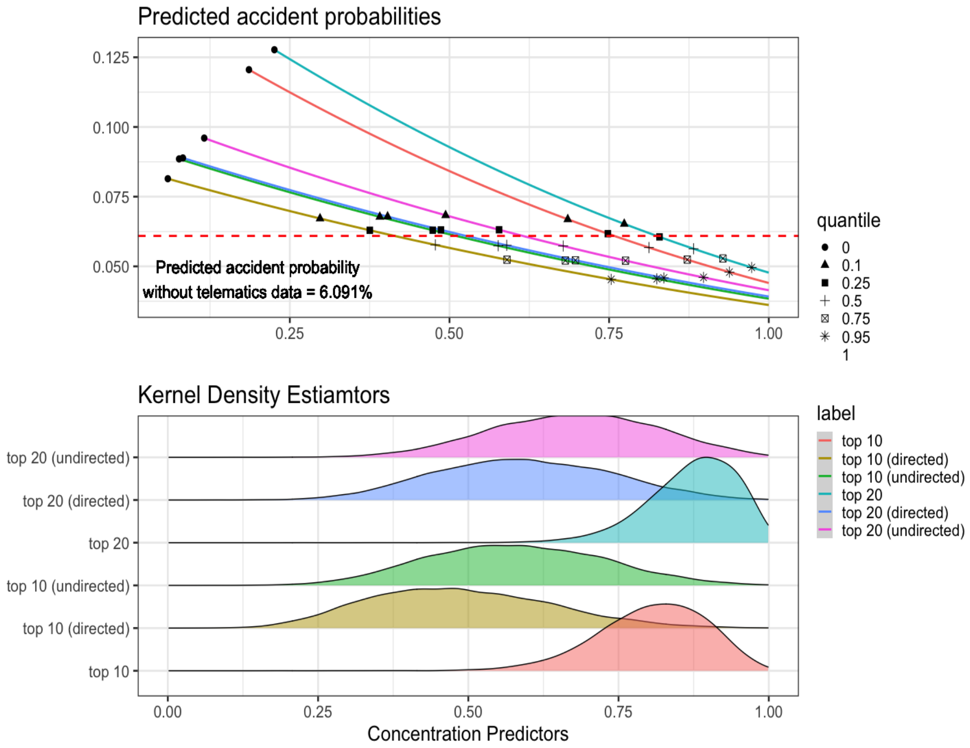
Appendix to Chapter 3

Appendix 1: Additional figures

Figure 5.2: Dataflow diagram.



†Blue headers denote original datasets at our disposal, gray - intermediate results of our computations, green - final regression tables



Appendix 2: End point clustering

Cluster analysis is an unsupervised learning technique that poses a lot of practical challenges, most prominently algorithm selection and subsequent parameter tuning. The choice can be guided by general domain knowledge on the one hand and by various cluster validation techniques on the other.

Spatial clusters oftentimes have non-convex shapes, the number of clusters is not known in advance which suggests using either hierarchical clustering with single linkage or density based algorithms (Aggarwal (2015)). Domain of feasible input parameter values is be restricted based on the insights from the literature. We are interested in distances a driver is willing to walk between parking facilities and final destinations. Benenson et al. (2008) and van der Waerden et al. (2017)[2] assume that in general it does not exceed 700 meters. Statistics on the maximum and average walking distance is used for (a) restricting the set algorithm input parameters (eps for DBSCAN and

[2] write more about partition

cut-off point for hierarchical clustering) (b) result validation.

Input parameters tuning based on internal cluster validation criteria turned out to be misleading [3] The potential reasons behind are discussed in Chapter 6.9 of Aggarwal (2015). We conceive of two approaches to overcome this issue: constructing more appropriate (for the task domain) clustering scores and using external validation criteria. Having labeled stops that are known to correspond to single destination provides a more reliable instrument for cluster validation. With this objective in mind, in the next subsection we try to get reliable cluster labels for a subset of destination.

Cluster analysis of overnight journeys.

We focus our attention on overnight stops. This decision has the following benefits: first, we have certain intuitive beliefs about some characteristics and number of such clusters, second problem size decreases significantly. Cluster labels from this stage, combined with results from previous subsection can be used to validate candidates for final location clustering.

For every policyholder we have extracted coordinates of the earliest journey of every day and filtered out locations with previous stop duration shorter than 4 hours [4] This gave us 1884120 journeys for 6706 policyholders.

The locations were analyzed separately for each policyholder using hierarchical clustering algorithm where clusters are merged using single linkage criteria. This allows the algorithm to discover clusters of arbitrary non-convex shapes. The main challenge is the choice of cut-off value for the dendogram, that determines the number of clusters. The physical interpretation of this parameter is elusive and depends on the linkage method. In single linkage agglomerative clustering, two groups of points will be merged in one cluster if the minimum distance between the points is less than cut-off value.

[3] We ran a slightly updated version of DBSCAN and selected minEps based on silhouette coefficient. In the resulting partitions some of the points belonging to different clusters were 5 - 10 meters apart

[4] If the driver travels overnight, the start of his earliest journey would not necessary correspond to the place where he has spent the night. 4 hours restriction allows to discard these locations from our analysis.

Conversely, the minimum distance of each resulting cluster to its neighbor clusters is guaranteed to exceed cut-off distance. Following suggestions from Benenson et al. (2008) and van der Waerden et al. (2017) we have restricted the set of possible cut-off values to {50m, 100m, 200m, 300m, 400m, 500m, 700m } and obtained 7 versions of location cluster labels for each policyholder.

In the absence of any information regarding the "ground truth" labeling of the locations, we first resorted to internal cluster validation criteria. 7 different solutions were evaluated using Silhouette coefficient and Dunn index, the results are reported in Table 5.7. It suggests that both measures tend to favor larger cutoff values. Aggarwal (2015) cautions however against relying too much on any metrics, therefore we have decided to check the basic physical characteristics of results. Some of the clusters have diameter larger than 5 km respectively 7 km and contain points that are more than 3 km respectively 5 km away from the cluster centroid for solutions with best values of silhouette and Dunn coefficient. Due to that we refrain from using these metrics in subsequent decision making and construct our own measures of clustering results.

For every set of clusters for a policyholder we calculate the following aggregate characteristics: minimum distance to other clusters, various quantiles of inter-cluster distances (50 %, 75%, 95%) and quantiles of distances to cluster centroids. Since we consider only overnight stops, common sense suggests that preference should be given to solutions with less clusters and consequently larger cutoffs, provided cluster characteristics satisfy certain[5] criteria.

We select the dendogram cutoff value for each policyholder using the following iterative procedure: on every step for all unassigned policies restrict the set of possible solutions using criteria, described in Table 5.6. Set the cutoff value to the highest cut off value available, or apply other less stringent criteria. Prior to that we deal separately with policies, that had stable solutions across medium parameter values (200, 300, 400). Cutoff values for the last 3 policies were assigned manually using visual inspection.[6] As we can see from Table 5.7, resulting cut-off values quite often

[5] they are not known in advanced and established only once we take a look at the data

[6] I was looking for a cutoff point, where characteristics of clustering solutions abruptly changed

Table 5.6: Parameter selection for overnight clusters

Step	Variable	Threshold (m)	No. of assigned policies
Step 0			
	Minimal distance to other clusters	400	
	95% distance to centroid	300	
	75% distance to centroid	200	
	Median distance to centroid	100	
			2451
Step 1			
	Maximum distance to centroid	700	
	Minimal distance to other clusters	300	
	95% distance to centroid	300	
	75% distance to centroid	200	
	Median distance to centroid	100	
			3632
Step 2			
	Maximum distance to centroid	700	
	Minimal distance to other clusters	200	
	95% distance to centroid	400	
	75% distance to centroid	200	
	Median distance to centroid	100	
			454
Step 3			
	Maximum distance to centroid	800	
	Minimal distance to other clusters	100	
	95% distance to centroid	400	
	75% distance to centroid	200	
	Median distance to centroid	100	
			134
Step 4			
	Maximum distance to centroid	800	
	Minimal distance to other clusters	100	
	95% distance to centroid	500	
	75% distance to centroid	400	
	Median distance to centroid	150	
			32

Table 5.7: Propositions of silhouette coefficient and Dunn's index.

cutoff value	Silhouette	Dunn	Our
0.70	2205	2637	1703
0.50	974	950	984
0.40	781	723	2982
0.30	760	652	566
0.20	956	799	361
0.10	606	536	110
0.05	352	409	

differ from the ones indicated by the internal validation criteria.

Table 5.8: Parameter selection for all clusters

Step	Variable	Threshold (m)	No. of assigned policies
Step 0			
	Minimal distance to other clusters	200	
	95% distance to centroid	400	
	75% distance to centroid	200	
	Median distance to centroid	50	
			2999
Step 1			
	Maximum distance to centroid	700	
	Minimal distance to other clusters	200	
	95% distance to centroid	300	
	75% distance to centroid	200	
	Median distance to centroid	50	
			1093
Step 2			
	Maximum distance to centroid	800	
	Minimal distance to other clusters	100	
	95% distance to centroid	300	
	75% distance to centroid	200	
	Median distance to centroid	50	
			1992
Step 3			
	Maximum distance to centroid	800	
	Minimal distance to other clusters	90	
	95% distance to centroid	400	
	75% distance to centroid	200	
	Median distance to centroid	100	
			311
Step 4			
	Maximum distance to centroid	800	
	Minimal distance to other clusters	50	
	95% distance to centroid	500	
	75% distance to centroid	200	
	Median distance to centroid	100	
			305

Spatial clustering of all stops of a policyholders

We use single linkage hierarchical clustering with different dendogram cutoff values to group policyholder's stops based on their spatial proximity. In previous subsection, we have preformed spatial clustering on a subset of overnight stops. Resulting labels are used to contrast possible solutions. More precisely, for every policyholder and for all 7 different sets of cluster labels we compute the following characteristics: cluster purity, recall, entropy and gini index.

Optimal cut off selection procedure is similar to the one used in Section5. Restriction criteria applied on every step are listed in Table 5.8. The main difference is that on

Step 0 we focus on clustering solutions that maximize all external validation criteria. During **Step 1** and **Step 2** we select among options that maximize at least one of external validation measures. Cut off values for 6 last policies are assigned manually, based on visual inspection.[7]

[7] I have examined aggregated cluster characteristics and selected cutoffs corresponding to "jumps".

Appendix to Chapter 4

Exploratory factor analysis

The latent traits that we try to measure are inextricably linked and boundaries between them can be quite elusive. For many items in our survey it is difficult to provide a definite answer what latent trait it is related to. For instance, positive answers to question GP.6.[8] might be affected by both, lack of trust towards the companies and /or high level of privacy concerns. Instead of entering a lengthy discourse on the nature and relevance of various survey items for individual's latent characteristics, we take a more pragmatic approach and perform an exploratory factor analysis whereby letting the data speak for itself. Our general knowledge can be used later to check the validity of the conclusions and select the most promising measurement model. To perform an exploratory factor analysis, we need to provide a correlation matrix, select the factor analysis method and determine the number of factors to extract. The following paragraphs elaborate on each of these aspects.

Almost all survey items are measured using different types of Likert scales. These scales are conventional practice in social sciences and psychology, however they are prone to several behavioral biases that should be kept in mind.[9] We treat answers from the Likert scale as ordinary variables[10] and compute Spearman's rho and Kendall's tau correlation coefficients.

We extract common factors using the *weighted least squares* (WLS) procedure, since it relies on few distributional assumptions and has good numerical convergence

[8] "I'm concerned that companies collect too much information about me"

[9] The composition of a Likert scale makes the answers susceptible to an acquiescence bias, when participants tend to agree with most of the statements. To counter this effect, several reversed items are included. Furthermore, responses might exhibit a central tendency bias with which individuals avoid using extreme answers in fear of not confirming to the general norm. This is a subcase of the social desirability bias. Social desirability biases could further manifest themselves with subjects providing answers that rather reflect society's expectations than their own views.

[10] There is no general consensus on whether answers to a Likert scale should be treated as ordinal or interval variables, however the majority view tends to favor the former option.

Table 5.9: Optimal number of factors based on different criteria for Spearman and Kendall correlation matrices.

correlation coefficient	**parallel**	**vss1**	**vss2**	**eBIC**	**SABIC**
Spearman	9	2	9	9	10
Kendall	9	2	7	6	8

properties.[11] The last required input parameter is the number of factors to be extracted. Costello and Osborne (2005) list various procedures, that could facilitate the choice.[12] Recommendations based on various criteria are summarized in Table 5.9.

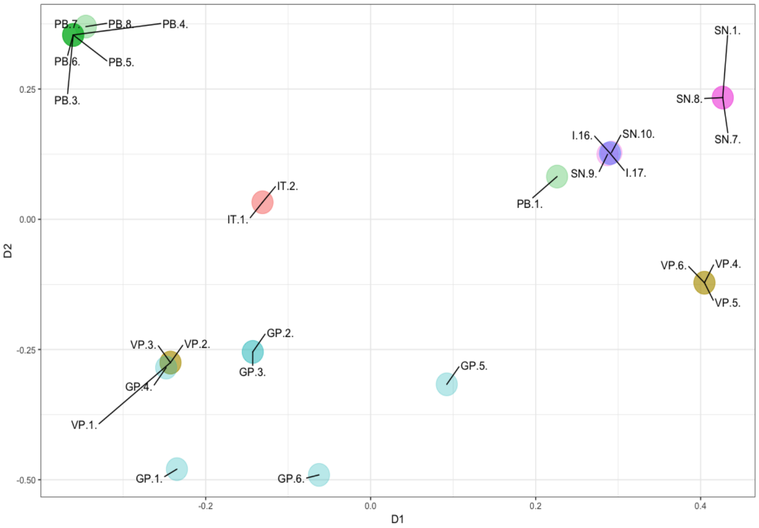
The optimal number of factors varies significantly depending on the type of correlation coefficient and the measure chosen. Instead of selecting a single approach, we have decided to perform an exploratory factor analysis using all suggested parameters and aggregate the results. Our objective is to see, which survey items are more likely to assess the same latent trait. This is achieved using the following steps: for every model we compute a binary matrix, where rows and columns correspond to observed variables. An entry equals 1 if both variables have an absolute factor loading exceeding 0.3 for the same latent variable. All matrices are summed together, rows are divided by the size of the diagonal element, representing for how many latent variables the questions was used, and then substracted from a matrix of ones. The resulting object could be viewed as a distance matrix, where variables measuring similar characteristics are closer to one another.

Visual representation of this information is obtained using multidimensional scaling.

[11] Sometimes the estimation algorithm might yield item commonalities equal or exceeding 1. This is the proportion of each variable's variance that can be explained by the factors (e.g., the underlying latent continua). These situations are referred to as Heywood and ultra-Heywood cases respectively and indicate that the resulting factor solution is invalid. Such anomalies could be caused by low sample size, too many factors extracted or other inadequacies in the common factor model. After trying different factoring method options available in the *psych* package in R, we have decided to extract factors using the weighted least squares approach, since it did not produce ultra-Heywood cases despite our low sample size. Note that the results of the inference will differ depending on the type of correlation coefficient used.

[12] An important pragmatic consideration is, that due to a limited sample size, number of independent variables in the regression equations should not be too high, furthermore having too few observed variables per latent factor might lead to unstable solutions. Due to these considerations we have decided to restrict our search to models with no more than 10 latent variables

Figure 5.3: Multidimensional scaling of observed variables.



The results are depicted in Figure 5.3 from which we see that certain groups of variables, for instance, PB.4 - PB.8, SN.1, SN.8, and SN.7, are almost always used together to assess a latent factor. Some of the variable clusters confirm our earlier intuition: items VP.1 - VP.3 measure the same latent characteristics - how much an individual values his privacy. Other suggested combinations, such as SN.9, SN.10, I.16, and I.17 are not anticipated. Post factum, the logic behind grouping them together becomes more apparent. Certain questions seem to be only loosely related to the common factors in the data set, this however does not render them invaluable for the subsequent analysis. Question PB.1, for instance, is directly related to the effort cost of providing information about the bank account.

Confirmatory factor analysis

Based on the insights from Figure 5.3 and a literature review, we can start developing our measurement model and study its psychometric properties. Table 5.10 lists various groupings of variables suggested by the exploratory factor analysis, whereby each group of questions measures one latent characteristics. For each instrument proposed we first ascertain its reliability by computing Cronbach's alpha and split half reliability. Values

Table 5.10: Grouping of the survey questions

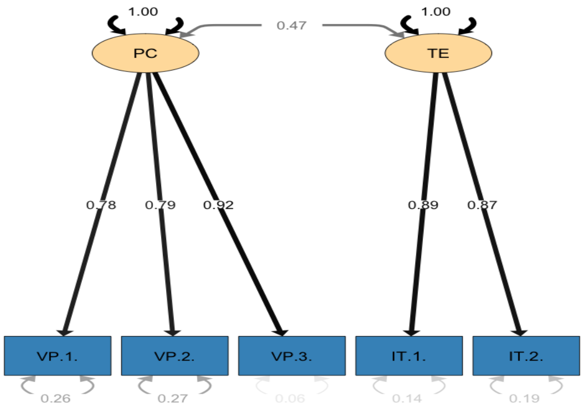
Group	Variables	Reliability		
		alpha	best alpha	split half
group 1	VP.1, VP.2, VP.3, GP.4	0.838	0.871	0.850
group 2	IT.1, IT.2	0.877	0.787	0.880
group 3	VP.4, VP.5, VP.6	0.542	0.565	0.532
group 4	PB.3, PB.4, PB.5, PB.6, PB.7, PB.8	0.743	0.736	0.567
group 5	SN.9, SN.10, I.16, I.17	0.419	0.429	0.453
group 6	SN.1, SN.7, SN.8	0.538	0.549	-0.285
group 7	GP.1, GP.2, GP.3, GP.5, GP.6	0.749	0.736	0.768
group 7 (full)	GP.1, GP.2, GP.3, GP.4, GP.5, GP.6	0.785	0.777	0.803

exceeding 0.7 are considered acceptable.

It reveals that group 3, group 5 and group 6 do not possess desirable psychometric properties and consequently won't be considered in further analyses. The reliability of group 1 can be slightly improved from 0.835 to 0.863 if the variable GP.4 is dropped. The confirmatory factor analysis is based on a correlation matrix with Spearmann's correlation coefficients. To ensure that the model is identified, we fix the variances of latent variables to 1. Fitting a model with 3 latent variables results in a poor model fit, with Chi = 101.054 on 39 df suggesting significant differences between model-based and observed correlation matrices. To achieve a better model fit, we drop the third latent variable. This decision should not have a strong negative impact on the explanatory power of our regression models since group 1 and group 4 measure similar latent characteristics (which is further confirmed by an estimated covariance of 0.536). The model with two latent variables has a good fit, reflected by a non-significant Chi-square p-value of 0.132, and value of RMSEA = 0.053. Path coefficients and relevant statistics are reported in Table 5.11. Figure 5.4 visualizes the final measurement model.

Based on the model fit, we can check other psychometric properties of our scales. Table 5.12 contains all necessary information to check the composite reliability and convergent/discriminant validity of survey items. Both latent factors have a composite reliability exceeding 0.9. Convergent validity can be established if the average variance extracted (AVE) is higher than 0.5. Discriminant validity is checked using the Fornell–Larcker criterium that compares the square root of the average variance extracted with correlations between the latent factors. If the former is higher, discriminant va-

Figure 5.4: Diagram of fitted CFA model



lidity can be established.

The objective of our analysis is to estimate the values of latent characteristics for all study participants. However, several subjects did not provide answers to all survey questions as this part of the experiment is neither mandatory nor financially compensated. Out of 282 people, 4 did not provide answers to one of the questions

Table 5.11: Coefficients of fitted CFA model

	Model			
	Estimate	Std. Err.	z	p
		Factor Loadings		
PC				
VP.1.	0.779	0.047	16.676	0.000
VP.2.	0.787	0.047	16.703	0.000
VP.3.	0.923	0.044	20.903	0.000
TE				
IT.1.	0.89	0.06	15.96	0.00
IT.2.	0.87	0.06	15.48	0.00
		Residual Variances		
VP.1.	0.26	0.03	9.27	0.00
VP.2.	0.27	0.03	9.25	0.00
VP.3.	0.06	0.02	2.63	0.01
IT.1.	0.14	0.06	2.27	0.02
IT.2.	0.19	0.06	3.07	0.00
		Latent Variances		
PC	1.00^{+}			
DTE	1.00^{+}			
		Latent Covariances		
PC w/TE	0.47	0.05	9.10	0.00
		Fit Indices		
χ^2	1.04(df=4)			0.132
CFI	0.997			
TLI	0.992			
RMSEA	0.053			

$^{+}$Fixed parameter

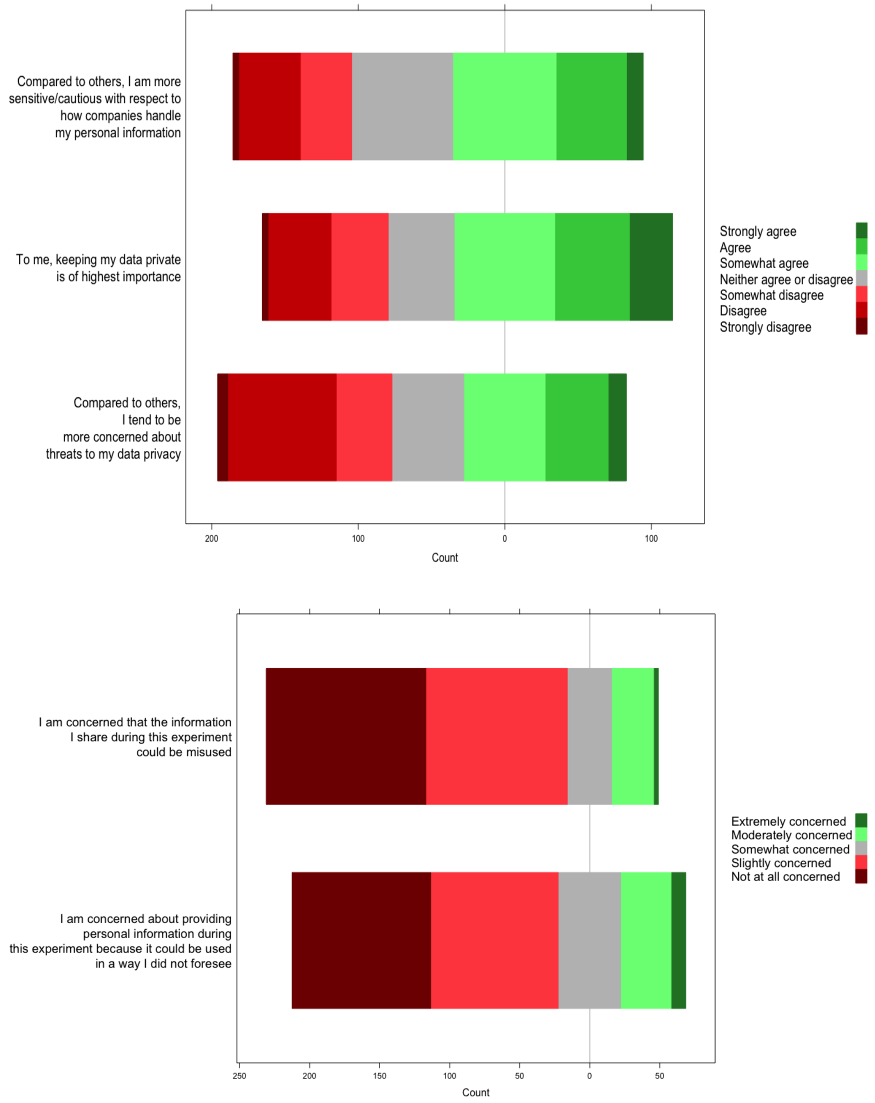
Table 5.12: Check for composite reliability, discriminant validity and Convergent Validity

	CR	AVE	PC	DTE
PC	0.912	0.778	0.882	0.472
DTE	0.905	0.827	0.472	0.909

VP.1 - VP.3 and 2 other participants did not answer one of the questions IT.1 - IT.2. Due to the small sample size, we want to avoid loosing observations and impute missing answers based on the information from other questions in the post experimental survey. This task is accomplished using a combination of bootstrap and the EM algorithm (Honaker et al. (2011)) implemented in the R package *Amelia*.

Answers to post-experimental survey

Figure 5.5: Answers to privacy-related questions in the post-experimental survey



Results of directional non parametric tests

Table 5.13: Results of directional non parametric tests. Dependent variable: price of informational bundle (not adjusted by 60)

Hypothesis		all observations		
			female	male
(V) $\geq$ (I + II)	mean difference	7.12	4.71	8.02
	effect size	0.36	0.22	0.49
	MWU	0.01***	0.27	0.02**
	KS test	0.01***	0.13	0.06*
	Permutation Test	0.02**	0.17	0.03**
	Bootstrap	0.01***	0.15	0.04**
(IV) $\geq$ (I + II)	mean difference	3.64	-0.79	6.45
	effect size	0.19	-0.04	0.40
	MWU	0.13	0.58	0.07*
	KS test	0.24	0.79	0.17
	Permutation Test	0.14	0.56	0.07*
	Bootstrap	0.14	0.57	0.07*
(I + II) $\geq$ (III)	mean difference	4.04	8.53	1.42
	effect size	0.23	0.47	0.08
	MWU	0.11	0.06*	0.26
	KS test	0.44	0.22	0.66
	Permutation Test	0.12	0.06*	0.36
	Bootstrap	0.11	0.06*	0.38
(V) $\geq$ (III)	mean difference	11.16	13.24	9.44
	effect size	0.62	0.74	0.55
	MWU	0.00***	0.00***	0.02***
	KS test	0.00***	0.02**	0.02**
	Permutation Test	0.00***	0.00***	0.03**
	Bootstrap	0.00***	0.00***	0.03**
(IV) $\geq$ (III)	mean difference	7.68	7.74	7.87
	effect size	0.43	0.43	0.45
	MWU	0.02**	0.07*	0.05**
	KS test	0.13	0.11	0.10*
	Permutation Test	0.02**	0.07*	0.05**
	Bootstrap	0.02**	0.07*	0.05**
(V) $\geq$ (IV)	mean difference	3.48	5.50	1.57
	effect size	0.19	0.29	0.09
	MWU	0.13	0.13	0.35
	KS test	0.30	0.33	0.76
	Permutation Test	0.16	0.12	0.37
	Bootstrap	0.16	0.13	0.38

Table 5.14: Results of directional non parametric tests. Dependent variable: price of informational bundle. Category B

Hypothesis		Category B		
			female	male
(III) $\leq$ (I + II)	mean difference	2.31	-0.95	4.26
	effect size	0.21	-0.09	0.38
	MWU	0.28	NA	0.23
	KS test	0.68	NA	0.47
	Permutation Test	0.28	NA	0.20
	Bootstrap	0.24	NA	0.18
(I + II) vs (IV)	mean difference	5.48	7.55	4.74
	effect size	0.46	0.76	0.36
	MWU	0.28	0.55	NA
	KS test	0.87	0.71	NA
	Permutation Test	0.16	0.17	NA
	Bootstrap	0.14	0.15	NA
(I + II) vs (V)	mean difference	12.00	24.84	2.31
	effect size	1.00	2.49	0.17
	MWU	0.04**	0.03**	0.54
	KS test	0.14	0.04**	0.90
	Permutation Test	0.01***	0.00***	0.34
	Bootstrap	0.01***	0.00***	0.34
(III) $\leq$ (V)	mean difference	14.31	23.90	6.57
	effect size	1.33	2.30	0.58
	MWU	0.01***	NA	0.06*
	KS test	0.03**	NA	0.17
	Permutation Test	0.01***	NA	0.11
	Bootstrap	0.01***	NA	0.10*
(III) $\leq$ (IV)	mean difference	7.79	6.61	9.00
	effect size	0.72	0.63	0.80
	MWU	0.09*	NA	NA
	KS test	0.39	NA	NA
	Permutation Test	0.09*	NA	NA
	Bootstrap	0.08*	NA	NA
(IV) vs (V)	mean difference	6.52	17.29	-2.43
	effect size	0.45	0.97	-0.23
	MWU	0.46	0.17	NA
	KS test	0.67	0.30	NA
	Permutation Test	0.18	0.07	NA
	Bootstrap	0.19	0.06	NA

Table 5.15: Results of directional non parametric tests. Category C

Hypothesis		Category C		
			female	male
(I + II) ≤ (V)	mean difference	14.50	11.55	15.88
	effect size	1.12	0.76	1.57
	MWU	0.00***	NA	0.00***
	KS test	0.00***	NA	0.00***
	Permutation Test	0.00***	NA	0.00***
	Bootstrap	0.00***	NA	0.00***
(I + II) ≤ (IV)	mean difference	6.15	5.78	2.47
	effect size	0.48	0.38	0.24
	MWU	0.09*	0.17	0.36
	KS test	0.19	0.36	0.59
	Permutation Test	0.07*	0.21	0.29
	Bootstrap	0.0*7	0.22	0.26
(III) vs (I + II)	mean difference	0.58	5.05	1.40
	effect size	0.04	0.35	0.13
	MWU	NA	NA	NA
	KS test	NA	NA	NA
	Permutation Test	0.44	0.22	0.40
	Bootstrap	0.42	0.21	0.40
(III) ≤ (IV)	mean difference	6.73	10.83	3.87
	effect size	0.50	0.75	0.35
	MWU	0.10*	0.05**	0.29
	KS test	0.29	0.18	0.58
	Permutation Test	0.08*	0.05**	0.26
	Bootstrap	0.07*	0.04**	0.23
(III) ≤ (V)	mean difference	15.08	16.60	17.29
	effect size	1.12	1.15	1.58
	MWU	0.00***	NA	0.01***
	KS test	0.00***	NA	0.01***
	Permutation Test	0.00***	NA	0.01***
	Bootstrap	0.00***	NA	0.02**
(IV) ≤ (V)	mean difference	8.35	5.77	13.41
	effect size	0.54	0.37	1.10
	MWU	0.08*	NA	0.02**
	KS test	0.12	NA	0.05**
	Permutation Test	0.06*	NA	0.02**
	Bootstrap	0.06*	NA	0.02**