Telematics Data in Motor Insurance: Creating Value by Understanding the Impact of Accidents on Vehicle Use

D I S S E R T A T I O N
of the University of St. Gallen
Graduate School of Business Administration,
Economics, Law, and Social Sciences (HSG)
to obtain the title of
Doctor Oeconomiae

submitted by

Tobias Ippisch
from
Germany

Approved on the application of

Prof. Dr. Elgar Fleisch
and
Prof. Dr. Walter Ackermann

Dissertation no. 3829

Lulu Enterprises, Inc. (USA), 2010
The University of St. Gallen, Graduate School of Business Administration, Economics, Law and Social Sciences (HSG) hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St. Gallen, October 26, 2010

The President:

Prof. Ernst Mohr, PhD
Acknowledgements

Nothing we do, we do on our own. There are always people around who support us, provide us a helping hand when needed, challenge everything we do for a good cause, and inspire us so that we achieve our goals in life. I want to acknowledge them and the many individuals who were involved in the development of this work.

First of all, I want to thank my advisor Prof. Dr. Elgar Fleisch for his support and encouragement during the development of this thesis. The freedom and stimulating working environment he provided contributed considerably to the success of this research. I also want to thank Prof. Dr. Walter Ackermann, who kindly agreed to be co-advisor for this thesis.

From August 2009 to March 2010, I stayed at the University of California, Berkeley as a visiting research scholar. This time has been a highly productive phase of my dissertation. I want to thank the Swiss National Science Foundation for their financial support and express my gratitude to J.D. Margulici from the California Center of Innovative Transportation and Prof. Ikhlaq Sidhu from the Center for Entrepreneurship & Technology for making this research stay possible.

This dissertation has benefited substantially from the exchange with several of my colleagues at the University of St. Gallen and the ETH Zurich. I want to thank Albrecht Bereuter for providing valuable insights into the insurance sector and running the I-Lab on numerous exciting industry projects; Prof. Dr. Frédéric Thiesse for his academic support and structural guidance during the different stages of my research; Luca Rabbeni and Harald Trautsch from Octo Telematics for providing the data for this research and being available whenever questions arose; Tobias Kowatsch for his advice on data analysis and programming issues; Johannes Paefgen and Alexander Skorna for their thorough comments and feedback that noticeably improved this dissertation; Dinah Ximena Ortiz from Berkeley for her stylistic edits on the text; and Elisabeth Vetsch-Keller for her outstanding organizational support. I also want to thank all my other colleagues in the research group of Prof. Dr. Elgar Fleisch and the I-Lab in particular for their excellent teamwork and the great time we shared at work and beyond.

Finally, I am very grateful to my parents Rosa-Maria and Alfons Ippisch and my sister Angela Ippisch for their continuous support and encouragement over the years. Thank you for being with me every step along the way.

November, 2010                   Tobias Ippisch
Summary

This dissertation examines how telematics technology and corresponding data can create value in motor insurance. On an operational level, I demonstrate its value by studying the accident impact on travel and driving behavior. Further analyses demonstrate its application outside of insurance and indicate its strategic importance. The dissertation is guided by the research question: How can the analysis of telematics data benefit insurance companies both on a strategic and operational level?

I motivate this dissertation by detailing current business challenges in motor insurance that have made insurers consider telematics-based concepts of vehicle insurance in more detail. I name common expectations regarding these novel policy schemes and indicate research deficits regarding their business value. A review of existing literature together with an overview of available telematics-based policies provides necessary background information.

The Principal-Agent Theory, which examines the general role of information and information asymmetries in insurance markets, provides this dissertation’s theoretical foundation. The Theory of Planned Behavior frames the specific analysis of post-accident travel and driving. I use it to explain habitual behavior in vehicle use and the role of interventions in breaking such habits. Based on this theoretical framework, I develop several hypotheses on the general accident impact on travel and driving, its persistency, the influence of moderating factors, and the selectivity of trip reductions. I test the hypotheses on telematics-based data from 1598 Italian motorists.

My findings indicate that accident involvement significantly impacts motorists’ travel and driving behavior. In the first post-accident month, accident-involved drivers reduce monthly trips by 11.4%, monthly mileage by 13.9%, and average driving speeds by 7.5%. In contrast to the findings of related literature, accident effects do not fade over time; in fact, they are still present 5 months on follow-up. Group differences become discernible for high- and low-mileage drivers. Trip reductions as a response to accident involvement are done selectively by trip purpose, but not in particular to avoid the place or time of the accident. Regarding the strategic value of telematics data, as an example I show its use for location planning in retail and name other possible application scenarios as well.

Finally, I discuss these findings and show their theoretical contribution to transportation research. Moreover, I detail their implications for insurance practitioners on both an operational and strategic level. I conclude by commenting on the limitations of this work and giving suggestions for further research.
Zusammenfassung


Die Auswertungen zeigen, dass Fahrer ihr Verhalten nach einem Unfall signifikant verändern. So geht im Unfallfolgemonat die Anzahl der monatlichen Fahrten um 11.4% zurück, die der gefahrenen Kilometer um 13.9%, und die Durchschnittsgeschwindigkeit um 7.5%. Entsprechende Effekte sind auch 5 Monate nach dem Unfall noch nachweisbar, Gruppenunterschiede jedoch nur zwischen Viel- und Wenigfahrern erkennbar. Strategische Bedeutung besitzen Telematikdaten aufgrund ihres Werts für branchenfremde Unternehmen. Dieser wird exemplarisch für die Standortplanung eines Retail-Unternehmens aufgezeigt.

In einem abschliessenden Kapitel wird der theoretische Beitrag der vorliegenden Forschungsergebnisse diskutiert sowie deren Implikationen für die Versicherungswirtschaft auf operativer und strategischer Ebene abgeleitet. Die Erörterung der Einschränkungen diese Forschungsarbeit sowie die Nennung möglicher weiterer Forschungsfelder beschliessen diese Dissertation.
Table of Contents

List of Tables ........................................................................................................ XIII

List of Figures ........................................................................................................ XV

List of Acronyms ................................................................................................ XVII

1. Introduction ................................................................................................... 1
   1.1 Problem Statement ..................................................................................... 1
       1.1.1 The Market Situation in Motor Insurance .................................... 2
       1.1.2 Explaining the Status Quo ............................................................ 5
   1.2 Expectations towards Telematics-based Vehicle Insurance ...................... 7
       1.2.1 Mileage Reduction ....................................................................... 7
       1.2.2 Improvement of the Customer Risk Portfolio .............................. 8
       1.2.3 Improvement of Horizontal Equity .............................................. 9
   1.3 Research Gaps .......................................................................................... 10
       1.3.1 Research Gaps in the Insurance Context .................................... 10
       1.3.2 Research Gaps in the Transportation Research Context ............ 13
   1.4 Research Questions .................................................................................. 15
   1.5 Methodology and Structure ...................................................................... 16
   1.6 Scientific Context of the Thesis ............................................................... 19
   1.7 Terms and Definitions .............................................................................. 19
       1.7.1 Telematics-based Vehicle Insurance .......................................... 20
       1.7.2 Travel Behavior and Driving Behavior ...................................... 21

2. Background .................................................................................................. 23
   2.1 A Primer on Telematics Technology ....................................................... 23
   2.2 Literature Review ..................................................................................... 25
       2.2.1 The Perception of Distance-based Vehicle Insurance over Time ........................................................................ 25
       2.2.2 Telematics-based Concepts of Vehicle Insurance ..................... 29
   2.3 Practical Applications .............................................................................. 33
   2.4 E-Call ....................................................................................................... 38

3. Theoretical Framework .............................................................................. 41
   3.1 Principal-Agent Theory ........................................................................... 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.1</td>
<td>An Introduction to Principal-Agent Theory</td>
<td>41</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Solutions to Principal-Agent Problems and the Role of Telematics Technology</td>
<td>43</td>
</tr>
<tr>
<td>3.2</td>
<td>The Theory of Planned Behavior</td>
<td>44</td>
</tr>
<tr>
<td>3.2.1</td>
<td>An Introduction to the Theory of Planned Behavior</td>
<td>44</td>
</tr>
<tr>
<td>3.2.2</td>
<td>The Role of Regret</td>
<td>45</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Habitual Behavior and the Role of New Information</td>
<td>45</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Relationship Moderators</td>
<td>48</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Environmental Influences</td>
<td>48</td>
</tr>
<tr>
<td>4.</td>
<td>Hypotheses Development</td>
<td>51</td>
</tr>
<tr>
<td>4.1</td>
<td>The Accident Impact on Travel and Driving Behavior</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>The Persistency of Effects on Travel and Driving Behavior</td>
<td>54</td>
</tr>
<tr>
<td>4.3</td>
<td>The Impact on Male, Female, and Company Car Drivers</td>
<td>55</td>
</tr>
<tr>
<td>4.4</td>
<td>The Impact on Low- and High-Mileage Drivers</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>The Impact of Accident Severity</td>
<td>59</td>
</tr>
<tr>
<td>4.6</td>
<td>The Impact on Trips of Different Purpose</td>
<td>60</td>
</tr>
<tr>
<td>4.7</td>
<td>Situational and Spatial Avoidance</td>
<td>62</td>
</tr>
<tr>
<td>5.</td>
<td>Data and Methodology</td>
<td>63</td>
</tr>
<tr>
<td>5.1</td>
<td>Experimental Setting</td>
<td>63</td>
</tr>
<tr>
<td>5.2</td>
<td>Unit of Analysis</td>
<td>63</td>
</tr>
<tr>
<td>5.3</td>
<td>Data</td>
<td>66</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Travel Data</td>
<td>66</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Accident Data</td>
<td>67</td>
</tr>
<tr>
<td>5.4</td>
<td>Research Methodology</td>
<td>68</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Sampling</td>
<td>69</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Data Preprocessing</td>
<td>72</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Trip Aggregation</td>
<td>73</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Trip / Driver Filtering</td>
<td>75</td>
</tr>
<tr>
<td>5.4.5</td>
<td>Compilation of Individual Driver Information</td>
<td>79</td>
</tr>
<tr>
<td>5.4.6</td>
<td>Endpoint Clustering</td>
<td>80</td>
</tr>
<tr>
<td>5.4.7</td>
<td>Trip Classification</td>
<td>84</td>
</tr>
<tr>
<td>5.4.8</td>
<td>Regression Analysis</td>
<td>89</td>
</tr>
<tr>
<td>5.4.9</td>
<td>Non-Parametric Testing</td>
<td>91</td>
</tr>
<tr>
<td>5.5</td>
<td>Statistical Software</td>
<td>91</td>
</tr>
</tbody>
</table>
6. **Findings** ........................................................................................................ 93
   6.1 Descriptive Statistics ................................................................................ 93
       6.1.1 Trips per Month .......................................................................... 93
       6.1.2 Distance traveled per Month ...................................................... 95
       6.1.3 Average Driving Speed per Month ............................................ 97
       6.1.4 Cluster and Classification Results .............................................. 99
       6.1.5 Accident Times and Locations ................................................. 103
   6.2 Inference Statistics ................................................................................. 105
       6.2.1 The Accident Impact on Travel and Driving Behavior ............ 105
       6.2.2 The Persistency of Effects on Travel and Driving Behavior ... 106
       6.2.3 The Impact on Male, Female, and Company Car Drivers ...... 108
       6.2.4 The Impact on Low- and High-Mileage Drivers ..................... 110
       6.2.5 The Impact of Accident Severity ............................................. 112
       6.2.6 The Impact on Trips of Different Purpose ............................... 113
       6.2.7 Situational and Spatial Avoidance ........................................... 114
       6.2.8 Summary of Hypothesis Testing .............................................. 115
   6.3 Additional Knowledge derived from GPS Data .................................... 116
       6.3.1 Individual Information ............................................................. 117
       6.3.2 Group Dynamics ...................................................................... 118

7. **Discussion and Implications** ..................................................................... 121
   7.1 Theoretical Contribution ........................................................................ 121
   7.2 Practical Implications for the Insurance Industry ................................. 124
       7.2.1 Operational Implications of the Analysis of Post-Accident
            Vehicle Use .............................................................................. 124
       7.2.2 Strategic Implications ............................................................... 126
   7.3 Limitations and Suggestions for Further Research .............................. 127
   7.4 Concluding Remarks .............................................................................. 131

**Appendix A: The DBScan Algorithm** ................................................................. 133

**Appendix B: The C4.5 Algorithm** ................................................................. 141

**References** .............................................................................................................. 147

**CURRICULUM VITAE** ........................................................................................ 169
**List of Tables**

Table 1: Overview of telematics-based vehicle insurance products ......................... 37

Table 2: Travel data attributes .................................................................................. 67

Table 3: Accident data attributes .............................................................................. 68

Table 4: Trip aggregation and filtering ..................................................................... 75

Table 5: Summary of the driver filtering process ..................................................... 78

Table 6: Used trip classes ......................................................................................... 86

Table 7: Confusion matrix of cross-validated training data ..................................... 88

Table 8: Descriptive statistics of trips per month ..................................................... 94

Table 9: Correlations of trip per month ................................................................... 95

Table 10: Descriptive statistics of kilometers traveled per month ......................... 96

Table 11: Correlations of kilometers traveled per month ....................................... 97

Table 12: Descriptive statistics of average driving speeds per month .................... 98

Table 13: Correlations of average driving speeds per month .................................. 99

Table 14: Accident impact on trips per month ....................................................... 105

Table 15: Accident impact on kilometers traveled per month ............................... 106

Table 16: Accident impact on average driving speed per month ............................ 106

Table 17: Accident impact on trips in the 5 follow-up months ............................... 107

Table 18: Accident impact on kilometers traveled in the 5 follow-up months ....... 107

Table 19: Accident impact on average driving speed in the 5 follow-up months .. 108

Table 20: Differences in trips per month for the 3 driver subgroups ..................... 109

Table 21: Differences in kilometers traveled for the 3 driver subgroups ............... 109

Table 22: Differences in average driving speed for the 3 driver subgroups .......... 110

Table 23: Differences in trips per month for the 2 driver subgroups ..................... 111
Table 24: Differences in kilometers traveled for the 2 driver subgroups.............. 111
Table 25: Differences in average driving speed for the 2 driver subgroups ........... 111
Table 26: The impact of accident severity on trips per month .................................. 112
Table 27: The impact of accident severity on kilometers traveled per month ...... 112
Table 28: The impact of accident severity on average driving speed per month ... 113
Table 29: The accident impact on the frequency of trips with different purpose .. 114
Table 30: The accident impact on trips at accident time ........................................... 115
Table 31: The accident impact on trips that pass by the accident location .......... 115
Table 32: Summary of hypothesis testing .............................................................. 116
List of Figures

Figure 1: Inflation-adjusted growth rates of motor insurance premiums in Europe .. 2
Figure 2: Gross premiums of German motor insurance 1998-2008 ......................... 3
Figure 3: Compensation of premium discounts by client acquisition ..................... 4
Figure 4: Claims ratios in German motor insurance 1993-2008 ............................ 5
Figure 5: Profit contribution among the insurer’s customer base ........................... 9
Figure 6: Research questions and methodology .................................................... 17
Figure 7: Telematics-based vs. distance-based vehicle insurance ........................... 21
Figure 8: The eCall system .................................................................................... 39
Figure 9: Principal-Agent Theory ......................................................................... 42
Figure 10: The Theory of Planned Behavior with feedback on control beliefs ...... 47
Figure 11: The telematics systems and services of Octo Telematics ....................... 65
Figure 12: The methodological steps for GPS data analysis .................................... 69
Figure 13: Seasonal matching of data ..................................................................... 71
Figure 14: Seasonal matching based on accident / reference month ..................... 72
Figure 15: The relationship between $\epsilon$ and minimal cluster distance ............... 83
Figure 16: Notation of the experimental design ................................................... 89
Figure 17: Histogram of trips per month ............................................................... 93
Figure 18: Histogram of kilometers traveled per month ....................................... 95
Figure 19: Histogram of average driving speeds per month ................................. 97
Figure 20: Histogram of average driving speeds per road type ............................ 98
Figure 21: Histogram of the number clusters per driver ....................................... 99
Figure 22: Shares of different trip types by driver class ...................................... 100
Figure 23: Distribution of arrival times by trip purpose ...................................... 101
Figure 24: Distribution of arrival days by trip purpose ........................................... 102
Figure 25: Distribution of daily accident and trip times ......................................... 103
Figure 26: Road types of accidents ........................................................................ 104
Figure 27: Percentage of trips that pass by the accident location ........................... 104
Figure 28: Transition probabilities of a single selected company car driver ........... 117
Figure 29: Zone of attraction of a shopping mall ................................................... 120
Figure 30: Overview of implications for the insurance industry ............................. 124
Figure 31: Suggestions for further research .......................................................... 127
Figure 32: A density based notion of clusters ....................................................... 133
Figure 33: The density notions of the DBScan algorithm ..................................... 135
Figure 34: The distorting effect of data normalization on $\varepsilon$......................... 137
Figure 35: Data normalization without the distorting effect on $\varepsilon$ ................... 138
Figure 36: General decision tree ........................................................................... 141
# List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANCOVA</td>
<td>Analysis of Covariance</td>
</tr>
<tr>
<td>am</td>
<td>(lat.) ante meridiem</td>
</tr>
<tr>
<td>BaFin</td>
<td>Bundesanstalt für Finanzdienstleistungsaufsicht</td>
</tr>
<tr>
<td>BBC</td>
<td>British Broadcasting Corporation</td>
</tr>
<tr>
<td>CAN</td>
<td>Controller Area Network</td>
</tr>
<tr>
<td>cf.</td>
<td>(lat.) confer</td>
</tr>
<tr>
<td>CO2</td>
<td>Carbon Dioxide</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-separated Values</td>
</tr>
<tr>
<td>DSRC</td>
<td>Dedicated Short Range Communication</td>
</tr>
<tr>
<td>e.g.</td>
<td>(lat.) exempli gratia</td>
</tr>
<tr>
<td>EWG</td>
<td>Europäische Wirtschaftsgemeinschaft</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>EU-27</td>
<td>European Union (since 2007)</td>
</tr>
<tr>
<td>ERF</td>
<td>European Union Road Federation</td>
</tr>
<tr>
<td>ESA</td>
<td>European Space Agency</td>
</tr>
<tr>
<td>et al.</td>
<td>(lat.) et alii</td>
</tr>
<tr>
<td>GDV</td>
<td>Gesamtverband der Deutschen Versicherungswirtschaft e.V.</td>
</tr>
<tr>
<td>GMAC</td>
<td>General Motors Acceptance Corporation</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>GRPS</td>
<td>General Packet Radio Service</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communications</td>
</tr>
<tr>
<td>ID</td>
<td>Identification</td>
</tr>
<tr>
<td>i.e.</td>
<td>(lat.) id est</td>
</tr>
<tr>
<td>km</td>
<td>Kilometers</td>
</tr>
<tr>
<td>km/h</td>
<td>Kilometers per hour</td>
</tr>
<tr>
<td>m</td>
<td>Meters</td>
</tr>
<tr>
<td>m/s</td>
<td>Meters per second</td>
</tr>
<tr>
<td>PC</td>
<td>Personal Computer</td>
</tr>
<tr>
<td>pm</td>
<td>(lat.) post meridiem</td>
</tr>
<tr>
<td>R²</td>
<td>Coefficient of Determination</td>
</tr>
<tr>
<td>SE</td>
<td>Standard Error</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>U.S.</td>
<td>United States</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>USB</td>
<td>Universal Serial Bus</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Problem Statement

Telematics-based vehicle insurance policies have attracted considerable attention from insurance officials in past years. Enabling novel rating schemes and safety-related add-on services, they form a completely new class of insurance coverage and bring an unprecedented level of innovation to the motor insurance business. Practitioners and academics alike have attributed manifold benefits to these policies, including the attraction of new, safety-conscious drivers, a downturn in accidents, and a reduction in claims costs (Litman, 1997). At the same time, they consider distance-based premiums as an answer to social and environmental problems, that is, lowering CO₂ emissions by inducing reduced vehicle use (Bordoff & Noel, 2008) and improving horizontal equity amongst motorists (Litman, 2008).

The initial hype, however, was followed by disillusionment in recent years, as companies faced several challenges regarding the new coverage. First of all, it became clear that telematics-based insurance does not come at low cost. Getting such a system operational requires considerable investments both in infrastructure and in-vehicle technology. Moreover, high expenses for wireless data transmission add another cost block to the business case calculation. A second pitfall is consumer reticence, as motorists may only anticipate low premium savings not worth a provider change, fear increased payments in case of inappropriate driving, or are scared off by privacy issues.¹ Finally, insurance managers have started to doubt the sustainability of current investments in telematics-based insurance, as required in-vehicle technology will become standard in the near future. Especially the European Commission’s eCall project, which intends to bring a system for automated emergency calls to newly sold vehicles by 2014, acts as a driving force here (Europa.eu, 2009). Insurers, however, are reluctant to support this solution (cf. GDV, 2006), even though its technical infrastructure would also suit insurance services. They fear that eCall, run by telematics providers and car manufacturers,

¹ A well-documented example of consumer reticence that eventually led to the cancellation of a telematics-based insurance product is the case of the “Pay-as-you-drive” policy of the British insurer Norwich Union (cf. Kirk, 2008).
might weaken established insurance-customer relationships and take away considerable business.

As these considerations show, telematics technology leaves insurers with several unsolved problems today. How should companies act in markets where active technology investments might not pay off and passive behavior might threaten established business models? What are sound business models for telematics-based vehicle insurance? Finding appropriate answers to these questions poses a major organizational challenge to many motor insurance companies today.

1.1.1 The Market Situation in Motor Insurance

Insurers’ interest in telematics technology can best be explained by a look at the current market environment for motor insurance sales, which exhibits harsh competition in many European countries. Stagnating market growth, premium erosions, and persistently high claims costs render automobile insurance, especially third-party liability, a business less and less attractive for insurance companies. In Europe, this especially holds true for mature insurance markets, although this sector is still growing in the new member states (cf. Figure 1 for 2006/2007 growth rates).

---

Figure 1: Inflation-adjusted growth rates of motor insurance premiums in Europe (CEA, 2007)

---

2 Growth rates by country from 2006 to 2007; AT = Austria, BE = Belgium, BG = Bulgaria, CH = Switzerland, CY = Cyprus, CZ = Czech Republic, DE = Germany, DK = Denmark, EE = Estonia, ES = Spain, FI = Finland, FR = France, GB = United Kingdom, GR = Greece, HR = Croatia, HU = Hungary, IE = Ireland, IS = Iceland, IT = Italy, LU = Luxembourg, MT = Malta, NL = The
1. Introduction

The decline in earned premiums is not just a temporary effect; data from Germany exhibits a downturn in all motor-related premiums now for the 4th year in a row since 2005, with the recovery in third-party liability between 2000 and 2004 being due to the compensation of excess claims costs at that time (cf. Figure 2).

![Gross premiums of German motor insurance 1998-2008 (GDV, 2008)](image)

*Figure 2: Gross premiums of German motor insurance 1998-2008 (GDV, 2008)*

In an environment with premiums under constant pressure, price competition is nevertheless an apt strategy just for subset companies: while for some it conflicts with quality leadership thought and agent-based distribution modes, for most it runs against growth ambitions and eludes rational market mechanisms. A calculational example by Lauszus & Schmidt-Gallas (2004) illustrates this (cf. Figure 3): Assuming only a slight premium cut by 5%, an expansion of the customer base by 25% would be necessary if profits are to be kept constant. With the market for motor vehicle coverage being saturated and rather inert in general, such increase in number of contracts has to be deemed odd and unrealistic.

---

Netherlands, NO = Norway, PL = Poland, PT = Portugal, SE = Sweden, SI = Slovenia, SK = Slovakia, TR = Turkey

3 “Gross” in this context denotes claims expenses and earned premiums before reinsurance-related cash flows.
A second argument that likewise militates against premium discounts are persistently high claims ratios (i.e., a measure expressing the relationship between gross claims expenses and gross earned premiums multiplied by 100). This measure is high and rising, although absolute claims-related expenses are dropping (primarily due to the overall trend in declining road accidents and the increased safety of passenger vehicles). Data from Germany indicates that written premiums in both third-party liability and collision today hardly cover loss-related costs (cf. Figure 4). While in the past this grievance affected only third-party liability, which allowed for cross-subsidizing by earnings from collision insurance, claims ratios now have risen also in the latter, reaching new maxima for partial and comprehensive coverage in 2008. When taking into consideration that earned premiums also have to cover administrative costs and agent rewards, the low attractiveness of the motor business to insurance companies becomes understandable and its profitability doubtful. Consequently, many insurance managers are currently rethinking their engagement in this domain and are looking for ways to flee from the cutthroat price competition this market exhibits.
1.1.2 Explaining the Status Quo

The state of today's motor insurance business is the outcome of an interaction of market characteristics, product properties, and societal developments of the past 10-15 years. In the following paragraphs, I show how insurance deregulation, a high degree of market saturation, the commodity-like character of insurance, and increased transparency as a result of the Internet all have contributed to the market environment insurers are in today. Other factors might be regarded as equally important, yet I limit myself to explicating the aforementioned.

Deregulation
In the mid 1990s, insurance markets were deregulated on a European level. With issuing the so-called Third-Generation Insurance directives (e.g., directive 90/618/EWG as of 11/08/1990, second directive 90/619/EWG as of 11/08/1990), the European Union for the first time established true price and product competition.

---


4 A comprehensive outline of deregulatory activities is provided by Wein (2001).
in European retail insurance markets (Cummins & Weiss, 2004). These directives, which were transposed into national law by the EU member states in 1994, initiated a total remodeling of the insurance business and brought unprecedented dynamics to the market: premium rates no longer necessitated concession by regulatory authorities or had to be based upon standardized, industry-wide calculation schemes. Beyond that, insurers could now use more and different input variables for premium calculation, leveraging both customer segmentation and price differentiation. At the same time, the establishment of a single European market for financial services allowed insurers to sell coverage in all signing member states. Cross-border transactions, which traditionally had been limited to reinsurance and selected commercial coverage, became possible also in retail markets. This made strong domestic players expand beyond their home country delimitations and increased competition within the different regional markets.

**Market saturation**
Third-party liability car insurance is compulsory in all European countries, so its market volume depends on the number of registered vehicles. This figure, however, features only marginal growth rates in Europe’s most developed countries (e.g., 0.3% in Germany 2008 (Kraftfahrt-Bundesamt, 2009) or 0.7% in Great Britain 2008 (Department of Transport, 2009)), with newly licensed vehicles predominantly serving as spare procurements. As a consequence, insurance companies face a high degree of saturation in established motor business markets today, where growth in written premiums can only be achieved at the cost of competitors.

**Commodity-like insurance products**
The possibilities for competing on motor vehicle coverage are often limited, as the intangible character of insurance products impedes product differentiation. Hence, insurance clients frequently regard the products of different insurers as much alike in their provided coverage and service spectrum. Even if there were substantial differences, it would be hard to convey them to prospects who hardly have any interest in insurance policies. The only differentiation criterion left for insurers in this case is often just the premium.

**Increased transparency in fragmented markets**
The level of consolidation within the European insurance industry is – compared to other industries – still relatively low, although an unprecedented number of mergers and acquisitions have taken place after the insurance market deregulation in 1994 (Cummins & Weiss, 2004). Besides big pan-European players, many mid-sized insurance companies exist today that do business in their home countries only: in
Germany alone, for example, 623 insurance companies and pension funds (as of January 2010) are authorized for business by the German Federal Financial Supervisory Authority BaFin (Bundesanstalt für Finanzdienstleistungsaufsicht, 2010). While such fragmented markets have per se a high level of competition, increased transparency fostered by new media and distribution channels is fueling the battle for customers even more: the advent of the Internet and the success of brokers have contributed their share to making insurance products increasingly comparable. Thus, insurance clients are more than ever in a position to choose the policies that best meet their coverage needs and budget constraints.

1.2 Expectations towards Telematics-based Vehicle Insurance

Expectations towards telematics-based vehicle insurance concepts have been considerable ever since their emergence. Not only have insurance practitioners regarded them as attractive for setting themselves apart from competitors in rough market environments; politicians and academics have also considered them as an approach to solving social and environmental problems. I provide a review of these effects in the following subsections.

1.2.1 Mileage Reduction

Most prominently, academics have ascribed telematics-based vehicle insurance the potential to reduce annual mileage if distance is reflected in insurance payments (Vickrey, 1968; Litman, 1997; Edlin, 2003; Bordoff & Noel, 2008). Once usage-based premiums replace lump-sum insurance costs, an incentive for lower vehicle use is created. Edlin (2003) expects a motorist’s annual mileage to decline by about 10% after switching to per-mile insurance schemes. Parry (2005) achieves similar results and calculates a mileage cut of 9.1%. The benefits of reduced mileage for society are straightforward: a downturn in fuel consumption and carbon emissions as well as lowered traffic congestion, roadway maintenance, and fuel dependence (Litman, 1997; Edlin, 2003; Parry, 2005; Bordoff & Noel, 2008). For insurance companies, less traffic translates into reduced road accidents and claims costs, even though accident rates may not decrease linearly along with mileage reductions.5

5 As Janke (1991) shows, this effect of nonlinearity is due to several factors, including that a) high-risk drivers tend to drive less than low-risk drivers on their own accord; b) high-mileage drivers...
1.2 Expectations towards Telematics-based Vehicle Insurance

Nonetheless, more than 40,000 road fatalities (as of 2007; ERF, 2009) and accident related expenses of more than €180 billion (WHO, 2004) in the EU-27 each year justify considerable effort to curtail these numbers.

1.2.2 Improvement of the Customer Risk Portfolio

Several contributions have attributed telematics-based vehicle insurance a positive effect on insurers’ customer portfolios. Litman (1997) ascribes the new premium scheme the potential to attract significant new business, as it “increases consumer choices and offers motorists a new opportunity to save money.” Arvidsson (2008) as well as Lindberg et al. (2007) argue that usage-based premiums foster self-selection amongst motorists, which positively affects an insurer’s risk portfolio by attracting low-risk customers. Once offered, usage-based policies are assumed to cause three distinct effects on the insurer’s risk portfolio:

- **Good risks enter the insurance pool.** Motorists that regard themselves as responsible and provident drivers will enter the insurance pool in expectation of reduced premiums. They may also be attracted by offered add-on services.

- **Bad risks transform into good risks.** Bad risks from the current insurance pool face rising insurance premiums owing to their reckless driving behavior. They decide to change their driving habits in order to cut down on premiums.

- **Bad risks leave the insurance pool.** Some bad risks may not be willing to give way to a more considerate car use and will leave the risk pool to get covered by another insurance company.

Figure 5 provides a graphical representation of these effects and puts them into the context of profit contribution per customer. Good risks add positively to a company’s profit, while undesirable, higher risks with frequent insurance claims have a negative impact. The described effects can help to thin out the right end of the customer portfolio, thus eventually influencing profits positively.

---

are likely to have newer and hence safer vehicles; and c) low-mileage drivers tend to commute more in urban areas which requires higher engagement in the driving process, while high-mileage drivers tend to use motorways, autobahns, or grade-separated highways more frequently which require less vehicle control and attention.
Figure 5: Profit contribution among the insurer’s customer base
(adopted from Hippner & Wilde, 2003)

Fincham (1996) identifies similar effects on the insurance customer portfolio by studying the impact of telematics technology on accident rates apart from distance-based premium schemes. In his work, he finds evidence that the sole presence of event-data-recorders, which record vehicle acceleration data in accident situations, correlates to reduced accident frequency. Though this phenomenon may be based upon ex-ante selection effects as well as ex-post driving behavior changes, the author does not elaborate on this aspect in more detail.

1.2.3 Improvement of Horizontal Equity

Another positive effect of per-mile premium schemes is the increase of horizontal equity amongst motorists. Nobel-prize winner William Vickrey was among the first to discuss the inequalities caused by lump-sum insurance schemes. As he puts it, the “manner in which [auto insurance] premiums are computed and paid fails

---

6 Horizontal equity in the insurance context means that people of similar risk should pay similar premiums.
miserably to bring home to the automobile user the costs he imposes in a manner that will appropriately influence his decisions” (Vickrey, 1968). Under conventional coverage schemes with hidden information on the clients’ annual mileage, insurers offer an average rate to both low- and high-volume drivers. Such coverage tends to harm infrequent drivers, who get overcharged and have to compensate the premiums of their high-mileage counterparts. This may cause adverse selection (cf. Akerlof, 1970) in the insurance market (Rothschild & Stiglitz, 1976). Distance-based car insurance alleviates this problem, with telematics-based insurance schemes being the best approach in terms of actuarial accuracy. According to Litman (2008), they are superior in this context to other forms of mileage consideration like self-estimates upon policy sign-up or regular odometer reads. In prior work, Litman (1997) argues that improved horizontal equity especially benefits low-income motorists, to whom insurance becomes more affordable and who thereby are discouraged from uninsured driving.

1.3 Research Gaps

Although the debate on distance-based schemes of insurance pricing has been up and running for a couple of years now, some research gaps exist that academics have not addressed so far. Namely, few contributions today view telematics-based vehicle insurance from the insurers’ perspective. This grievance pertains to both strategic and operational levels, as I show in the first part of this subsection. In the second part, I further depict existent deficits in transportation research, which are in understanding post-accident travel and driving behavior. In the course of this dissertation, I align the analysis of behavioral changes after accident involvement with creating value for insurance companies on the operational level.

1.3.1 Research Gaps in the Insurance Context

The strategic value of telematics-based vehicle insurance

In past years, academics have evaluated telematics-based vehicle insurance and especially distance-based pricing schemes from different points of view. The main focus therein has been the analysis of possible outcomes and benefits for society. Few contributions exist, however, that consider the insurer’s stake in telematics-based vehicle coverage or that discuss apt business models. Neither do scientific contributions take into account that premium discounts along with high technology investments run against insurers’ market rationale, making telematics-based
business models unattractive up front, nor are concepts developed that illustrate how insurers could profit from such investment otherwise. Parry (2005), for example, in a rigorous comparison of distance-based premiums to gasoline taxes, does not at all consider insurer benefits. He bases his analysis on a zero profit assumption for fully competitive insurance markets, which raises the question of how insurers should recover ex-ante technology investments. In addition, this rationale ignores that telematics-based insurance schemes are substantially different from conventional ones, which runs against common assumptions of the zero profit condition (i.e., the assumption of homogeneous products).

Oberholzer (2003) views usage-based insurance more from a business standpoint. He analyzes its potential in the Swiss insurance market and identifies low-risk and low-mileage drivers as promising target clients. He suggests joint ventures between car manufacturers and insurance companies in order to market and sell the new insurance scheme to the driver population. Nonetheless, his work exhibits several shortcomings. First, his analysis does not consider all major cost blocks of telematics-based vehicle insurance. He incorporates neither hardware fitting costs (which are primarily non-declining labor costs) nor expenses for regular data transmission into his business models. Second, the cost estimates the author uses often seem incorrect and unsubstantiated. For example, he assumes hardware costs to follow Moore’s Law and projects them to drop to $25 by the end of 2006, a price level that has not been reached as of early 2010. Hence, Oberholzer’s contribution has to be regarded as adding only little to the understanding and design of business models for telematics-based vehicle insurance.

For the most part, scientific literature assumes the business value of telematics-based vehicle insurance to be a reduction in claims costs. Since such pricing schemes predominantly appeal to low-risk drivers, they foster self-selection amongst motorists and help to acquire drivers with fewer insurance claims (cf. Edlin, 2003; Coroama & Höckl, 2004; Bordoff & Noel, 2008). These predictions often lack practical validation, however, as both low consumer interest and high market saturation frequently challenge common assumptions such projections are based on (e.g., complete information, rational deciding insurance clients, absence of
1.3 Research Gaps

Furthermore, the introduction of telematics-based insurance products may not be a sound strategy in markets where pricing models are already substantially differentiated, as the number of clients that will eventually benefit from them might be too small (Thiele et al., 2009). In addition, insurers often are reluctant to offer premium discounts together with technology investments, as they fear cannibalization effects with existing customers. Instead, attracting risk-averse drivers who are willing to pay extra for telematics-enabled safety services seems to be a better option. Sound strategies for pricing such add-ons are frequently missing, however, which I attribute to the fact that literature predominantly has focused on the “distance” aspect of telematics-based insurance, while add-on service marketing has played only a marginal role.

Finally, every development of telematics-based insurance products today has to take the upcoming eCall system into consideration (for details on eCall see subsection 2.4). eCall is intended to facilitate automated emergency calls and will become standard in new vehicles by 2014 (Europa.eu, 2009). Its infrastructure may serve as a value added service platform (Dietz, 2007), allowing insurance-related services, amongst other add-ons. Insurance companies’ own technology investments would then become obsolete. Von Trostprugg (2007) assumes that eCall will benefit authorities, car manufacturers, and insurance companies alike in terms of cost savings, although the author does not factor in the challenge eCall might pose to established insurance business models. Many insurance practitioners actually doubt that their companies will gain from eCall in the long run, as they fear becoming dependent on car manufacturers for client acquisition and data access. Though Oberholzer (2003) suggests that insurers should actively pursue collaboration with car manufacturers, this seems likely to be an apt strategy for large European insurers only, who can provide white-label services in many countries and let car manufacturers sell insurance under their brand. The alternatives available to smaller market participants are unclear, however, and have not been addressed in academic literature so far. Getting an understanding of the potentials and pitfalls of eCall would thus be highly desirable for practicing managers in the area of insurance product development and innovation.

---

7 Evidence is given by British insurer Norwich Union here: Its “Pay-as-you-drive” product was cancelled in 2008 after just two years of operation, as it faced high consumer renitence and managed to attract only 10,000 customers during that time (Kirk, 2008).
Creating value from telematics data analysis on an operational level

Telematics-based vehicle insurance will make information of unprecedented quantity and accuracy available to insurance companies. It allows insurers to follow accident causations more precisely and gain new insights on the crash risks of different driver types, which will eventually permit them to fine-tune premium calculation schemes. However, little thought has been given to which knowledge to derive apart from that. Aside from actuarial science, a significant amount of academic work exists that uses GPS data to learn people’s significant locations (Ashbrook & Starner, 2003), derive trip purposes and activities (Wolf et al., 2001; Axhausen et al., 2003; Liao et al., 2005), study travel modes (Liao et al., 2007), or improve route planning (Edelkamp & Schrödl, 2003). The business value behind these analyses remains vague, however, notwithstanding the methodological rigor with which they are performed.

1.3.2 Research Gaps in the Transportation Research Context

In order to conceptualize the deficits this dissertation addresses in the context of transportation research, I consider two different aspects of analysis within this area. First, I approach the research gap on post-accident vehicle use behavior from the perspective of intervention analysis. Second, I detail literature on accident repeaters. In conjunction, these two streams of literature reveal the research deficit regarding post-accident travel and driving which this dissertation intends to address.

The study of interventions to vehicle use

Increased vehicle traffic creates many negative consequences for society, including air pollution, noise, carbon dioxide emissions, and steady congestion (Bamberg, 2006). Policymakers are aware of these problems and are thus continually striving to reduce individual travel demand. Against this background, a significant body of literature in transportation research studies the effect of interventions on vehicle use. Interventions are necessary as motorists frequently follow habitual patterns of travel that are hard to break once established (Gärling & Axhausen, 2003). As Dahlstrand & Biel (1997) as well as Prochaska et al. (1992) point out, it might take several steps to make motorists switch to environmentally-friendly modes of travel.

Scholars have proposed numerous interventional approaches to reduce vehicle traffic. One group relies on psychological and behavioral strategies, which Fuji & Taniguchi (2005) refer to as travel feedback programs. They furnish participants with tailored information based on their past behavior and analyze to which extent such details modify vehicle use. The information provided commonly points out the
environmental impact of driving, names public transport alternatives, and gives advice for reduced car use. Examples of travel feedback programs include “Individualized Marketing” (Brög, 1998), “Travel Smart” (Western Australia Department of Transport, 2000), “Travel Blending” (Rose & Ampt, 2001), and the “Travel Feedback Program” (Taniguchi et al., 2003).

Structural strategies are another form of intervention to vehicle use. They “change the structure surrounding travel behavior, e.g., the service availability of various travel modes and systems that regulate travel behavior” (Fujii et al., 2001). Structural interventions include the imposition of toll roads, the reduction of public transportation fees, or the set-up of traffic restrictions. The latter can be done purposely or along with other activities, as the authors show on the interventional effect of a temporary maintenance-related road closure. Other structural strategies consider residential relocation or the temporary provision of free public transportation tickets (Fujii & Kitamura, 2003; Bamberg, 2006).

Kearny and DeYoung (1996) provide a thorough overview of 29 studies on different interventions to car use. None of these works considers accidents as a possible interventional cause to travel and driving habits, though. The reasons for this are twofold. First, accidents are not a means of active travel demand management, which is the prime focus of the literature stream at hand. They occur as part of everyday traffic and are only an object of investigation; behavioral change is not the research intent. Second, accidents cannot be administered purposely, both from an ethical and practical perspective. As a result of their low probability of occurrence, large sample sizes and extended observation periods are necessary in order to catch a sufficient number of accident-involved motorists for analysis. This cannot be reasonably done with conventional survey methods, both from the perspective of sample size needed and the accuracy of such investigations. Such analyses require the use of telematics technology, which has not been widely available in motor vehicles until the past few years. Consequently, no scientific literature has analyzed the impact of accidents on vehicle use so far.

The study of accident repeaters
The examination of post-accident behavior is in the focus of the analysis of accident repeaters. It studies the stability of driving records over time and the distribution and prediction of accident frequencies (cf. Burg, 1970; Peck et al., 1971; Stewart & Campbell, 1972). Corresponding work has achieved only poor results though, owing to individual drivers’ generally low accident frequency. More recent contributions intend to infer future accident involvement from past driving records.
1. Introduction

They pick up the hypothesis of “accident proneness” as stated in early work by Greenwood & Yule (1920), which hypothesizes that some motorists exhibit more accidents than can be expected by chance. Research by Hauer et al. (1991) supports this argument; their inclusion of previous crash records in the analysis improves their prediction model for future accident involvement significantly. Chen et al. (1995) show that models based on additional information of prior at-fault accidents identify up to 23% more motorists with at least one at-fault crash involvement within two years than models that utilize prior convictions alone. Elliott et al. (2001) and Daigneault et al. (2002) perform similar analyses but focus solely on old and young drivers, respectively. In a most recent study, Chandraratna et al. (2006) employ multiple logistic regression to model the relationship between demographic factors, prior offenses, crash record, and future crash occurrence. They also provide a sound overview of related work in that area.

As this literature review demonstrates, the study of accident repeaters is also not congruent with the analysis of post-accident travel and driving. They differ both in their object of investigation (chance of subsequent accident involvement versus behavioral changes after an accident) and in their methodological approach (logistic regression versus analysis of covariance for the study of accident-induced behavioral changes). Consequently, the research on accident repeaters cannot contribute to closing the research gap on post-accident vehicle operation.

1.4 Research Questions

Today, many business models of telematics-based vehicle insurance are based on incomplete information or are flawed due to false market expectations. This lack of understanding amongst practitioners is reflected in academic literature, which falls short of providing decision support for implementing corresponding coverage schemes. Academics have viewed telematics-based vehicle insurance predominantly from the perspectives of adverse selection, motorist equity, and traffic reduction, but have often ignored the insurers’ stake in it. As a consequence, insurance managers frequently do not regard telematics-based concepts as a strategic alternative to established motor insurance business models. Also, little attention has been given to the strategic value of geospatial travel data for inter-industry value creation so far.

A similar lack of understanding amongst insurers is prevalent today when it comes to the value of telematics data on the operational level. While academics regularly
view GPS data from the perspectives of travel behavior research and traffic management, it has not addressed the question of what new knowledge insurers can derive from it. Together, these deficits lead me to the formulation of the following research question:

**How can the analysis of telematics data benefit insurance companies both on a strategic and operational level?**

I intend to answer this question separately for each perspective but use similar data and methodological toolsets. From the operational viewpoint, I concretely explore how motorists change their travel and driving behavior after accident involvement. Better knowledge of such effects may trigger actuarial improvements (e.g., advanced bonus-malus systems) or the extension of an insurers’ service portfolio (i.e., vehicle training for at-risk motorists). With this analysis, I also address the research gap on post-accident vehicle use that exists in transportation research today. For the strategic angle, I will illustrate how telematics data can create value outside of the insurance domain and name strategies for insurance managers to reap these benefits. The formulation of five research sub-questions builds on these considerations and gives structure to the thesis:

- How can travel and driving behavior be measured using GPS data?
- How does travel and driving behavior change after an accident?
- How do moderating factors influence the extent of behavioral change?
- Which other information can be derived from GPS data?
- How can insurance managers respond to the challenges telematics technology imposes on the motor insurance market?

**1.5 Methodology and Structure**

This thesis brings together methodologies from different academic areas to answer the research questions at hand. For the operational perspective of this thesis, I build the analysis of post-accident travel and driving behavior on established theory from behavioral science and draw from transportation and accident research to formulate corresponding hypotheses. I test these hypotheses using a confirmatory approach; thus my research methodology is solely quantitative. From the strategic perspective, I use methods of data mining and knowledge discovery in databases to derive...
information valuable for inter-industry collaboration. Figure 6 relates the research questions to the methodological steps I use to answer them.

Figure 6: Research questions and methodology

I structure the remainder of the thesis as follows: Section 2 provides a state-of-the-art review of telematics-based vehicle insurance. It starts with a brief outline of key telematics technologies, which include wireless communications, positioning systems, automotive sensors, and event-data-recorders. Next, relevant academic literature on the topic is reviewed. I illustrate the beginnings of mileage-based vehicle insurance in the late 1960s, summarize the U.S. debate on distance-based insurance in the 1990s, and name the current body of literature on the topic. A market overview of current telematics-based vehicle insurance products follows, where I outline the insurance concepts that different companies pursue, their technical implementation, and the current status of product availability. The section
concludes with a description of eCall, where I present its political motivation, the technical infrastructure it uses, and insurers’ attitude towards it.

Section 3 develops the conceptual framework of this thesis. I begin with a short outline of Principal-Agent Theory to cover basic insurance rationale. Next, I introduce antecedents of general behavior as used in the Theory of Planned Behavior and relate them to the context of vehicle driving and road travel. Special consideration is given to habitual effects and interventions, which eventually alter behavioral patterns. The Theory of Planned Behavior allows me to formulate a comprehensive framework which explains changes in travel and driving behavior after crash exposure.

I continue with the development of my research hypotheses in Section 4. I use the number of monthly trips, monthly vehicle mileage, and driving speed as testing variables to analyze the impact of accidents on vehicle use and the influence of moderating factors. First, the argument for a general accident impact is developed. I then consider effect persistency. Next, I divide up the population of crash-involved motorists by different parameters (i.e., male, female and company driver, high- and low-mileage drivers) to compare the impact on driver subgroups. Further hypotheses consider the influence of accident severity and the impact on different kinds of trips. I conclude this section by providing a rationale that suggests that motorists engage in situational and spatial avoidance strategies after accident involvement.

Section 5 outlines the methodological steps of the analysis. It begins with a description of the experimental setting and the unit of analysis of this research. After describing both travel and accident data in detail, the individual methodological steps are delineated. They include sampling, data preprocessing and filtering activities, trip aggregation, the development of measurement variables, and the actual regression and non-parametric analyses.

Section 6 presents the research findings. It shows descriptive statistics on trip, mileage, and driving speed parameters, presents the cluster and classification results, and provides information on accident times and locations. In the subsection on inference, I present the outcome of the hypothesis testing, which serves as an answer to the main research question on the operational level. Next, I illustrate which additional knowledge can be derived from GPS data using concrete examples. I thereby address the strategic aspect of the postulated research question.
1. Introduction

In Section 7, I summarize the theoretical contribution and the practical implications of my work. Based on the research findings, I give recommendations to insurance practitioners both on the operational and strategic level. I conclude with suggesting areas for further research which can be addressed by applying alternative methodological toolsets, linking GPS data to additional client information, and using such data outside of the insurance context.

1.6 Scientific Context of the Thesis

The general problem statement and the research questions of this thesis emerge from a practical research project in conjunction with the analysis of relevant literature on the topic. Together, they reveal research gaps which I intend to close with this dissertation. This approach follows Ulrich (1981), who argues that business administration research derives its research questions from actual business challenges of affected enterprises. It aims at developing practical concepts, deriving normative conclusions, and assisting managers in their decision making process. In this scientific self-conception, theoretical foundations are seen as a tool rather than as an outcome of the research process.

My dissertation follows this notion and derives its motivation from a practical challenge in the insurance industry. Currently, insurance companies still see telematics technology ambivalently, both as an opportunity to gain significant new business and as a threat to established business models. A full understanding of the potential of telematics technology in the insurance business is necessary for basing management decisions upon it. This dissertation sheds light on a neglected aspect of telematics-based vehicle insurance, that is, the potential that lies within the analysis of telematics data. Exemplarily, I show how new knowledge and value can be created from such analyses. While my results are primarily targeted to benefit the insurance domain, they also contribute to accident research and the GPS data analysis community.

1.7 Terms and Definitions

This dissertation makes frequent use of terms of art. In order to facilitate the understanding of the arguments and rationale of this research, I define the following key terms: telematics-based vehicle insurance, travel behavior, and driving behavior.
1.7.1 Telematics-based Vehicle Insurance

This dissertation analyzes the use of in-vehicle telematics technology in motor insurance markets. This technology opens up a novel category of vehicle insurance that extends conventional policies by per-mile and behavior-based insurance rates, security-related add-on services, and vehicle management and diagnostics options. This variety of possible concepts and services has thus far inhibited the establishment of a general term to unambiguously denominate that type of car insurance. Most commonly, it is referred to as “Pay-as-you-Drive” in both business and academia (cf. Guensler et al., 2003; Parry, 2005). This notion, which stresses premium variability in mileage driven, has two inherent shortcomings. First, not all insurance schemes that apply telematics aim at adjusting premiums for vehicle mileage. Some concepts may employ the technology to enable only safety features, while others exclusively use event-data-recorders to prove a motorist inculpable in case of an accident. This also prohibits the use of other familiar terms like “Usage-based Insurance,” “Distance-based Vehicle Insurance,” “Mileage-based Insurance,” or “Per-Mile Premiums” (Litman, 2008) to denote telematics-based coverage. These policies have been around since the late 1960s and developed without any reference to telematics technology,8 which forbids using them to subsume any technology-based insurance models. Second, “Pay-as-you-Drive” is a trademark of Norwich Union (UK) and describes a distinct insurance product rather than a product category. Using this term for referring to the whole insurance concept would create ambiguity and misunderstanding.

Owing to these drawbacks, I employ the term “telematics-based vehicle insurance” to denominate motor coverage that makes use of telematics technology. It deliberately avoids any inferences on the intended use of the technology regarding the insurance models. It nevertheless is constituted broadly enough to encompass all kinds of insurance schemes that make use of vehicle-related information and communication technology. Figure 7 names common “telematics-based vehicle insurance” concepts and shows their relation to distance-based notions.

---

8 Initial notions of “Mileage-based Insurance” and similar terms described concepts like paying insurance at the gas station or performing mileage verification during smog tests.
In this thesis, I illustrate the importance of telematics data for insurance companies through an analysis of post-crash travel and driving behavior. Travel behavior research investigates people’s spatial patterns. It uses cross-sectional data to model people’s choice of available travel alternatives under given circumstances (e.g., residence, profession, marital status). Variables subject to analysis are diverse and include the number of trips and miles traveled, temporal travel patterns, transportation mode choices, as well route selection and destination analyses (Ortuzar & Willumsen, 1995). The research is thereby not limited to describing the as-is state of travel in different regions. Frequently, it tests the effectiveness of interventions as travel behavior is often found to be habitual and rather inelastic short-term (Pendyala et al., 2000). Especially if such persistency is not the result of deliberation, it may be “difficult to influence with rational arguments (e.g., increased costs), since the person making the choice tends to discount relevant information” (Gärling & Axhausen, 2003).

Driving behavior refers to motorists’ driving performance. Its focus is not on where and when people travel but on how they do so. It is evaluated using criteria like lane keeping, distance to other vehicles, steering wheel movements, rear mirror checking, road sign detection, or speeding. Driving behavior research investigates the influence of circumstances like cell phone use (Brookhuis et al., 1991), driving under the influence (Ramaekers et al., 2000), or sleepiness (Lyznicki et al., 2005) on these measures. They may themselves serve as inputs to accident models in subsequent analyses. In general, driving behavior and driver characteristics serve as
good determinants of accident propensity, even if they fall short of considering situational factors and motives in the analysis (Ranney, 1994). This dissertation will study post-accident driving behavior on the parameter vehicle speed as inferable from the available GPS data.
2. Background

This section provides an overview of available scientific literature and existing practical concepts regarding telematics-based vehicle insurance. Its goal is to show the perception and development of this idea since the 1960s and how insurers and telematics providers have turned it into practice. In subsection 2.1, I begin with a short outline of key telematics technologies. Subsection 2.2 then gives a historical review of pay-per-mile insurance and discusses legislative and technical aspects of telematics-based insurance schemes. Subsection 2.3 continues with a market overview of insurance telematics products and details differences in design and market availability of the individual services. I conclude with describing the eCall platform in subsection 2.4, which is about to become a technology standard for telematics applications in the next years and may thus considerably change the telematics and insurance landscape.

2.1 A Primer on Telematics Technology

“Telematics” is not a distinct technology or technology standard. Originally coined by Nora & Minc (1978) to describe the combination of telecommunication and information processing, it today predominantly refers to information and communications technology within road vehicles (Nijkamp et al., 1996; Van Der Laan et al., 1997). Applications facilitated by telematics depend on the entities that data is exchanged with; these may be businesses and corresponding computer systems (e.g., for fleet management, cargo and vehicle tracking, and remote diagnostics), road-side beacons and traffic signs (e.g., for traffic control), or other vehicles (e.g., for collision warning). Also, in-vehicle services like satellite navigation, driver assistance systems, or safety electronics may very well be covered by the “telematics” notion. In the following paragraphs, I briefly introduce the most important technologies enabling such applications.

Wireless communications

Wireless communication allows steady data exchange between vehicles and stationary systems. Most common is cellular communication, including standards like Global System for Mobile Communications (GSM, c.f. Mouly & Pautet, 1992) and General Packet Radio Service (GPRS, cf. Ghribi & Logrippo, 2000), which come at low cost and ensure high network coverage (Goel, 2007). In the future, Dedicated Short Range Communication (DSRC), which to date is mainly used for electronic toll collection, may gain more importance as a communication channel
Positioning systems

Knowing the position of a vehicle is crucial for many telematics applications. Positioning systems provide motorists with autonomous navigation and enable tracking of equipped vehicles. From exact position data, one can derive information on routes, distances, and driving speeds, which may be stored on-board or transmitted to a remote system for further processing (Goel, 2007). Today, the Global Positioning System (GPS) is the predominant standard for satellite-based positioning. It works using highly accurate timing signals and orbital information emitted by GPS satellites. From this data and its transit time to the GPS units, the receivers calculate geospatial positions using geometric trilateration. Under optimal operating conditions (i.e., the line-of-sight to satellites is not obstructed by mountainous landscape or high buildings), the positioning accuracy of GPS is usually 10 meters or less, which is sufficient for most commercial applications (Kaplan & Hegarty, 1996). Means of Assisted GPS can further contribute to faster and more accurate positioning in adverse signal environments. The GPS receiver thereby utilizes auxiliary location data from GSM cells or stationary reference antennas (Weckström et al., 2003). Another satellite-based positioning system currently under development by the European Space Agency (ESA) is Galileo, which will become available by 2013 (BBC, 2007).

Automotive sensors

Sensors are indispensable in today’s electronic vehicle control systems. They are defined as “devices that transform (or transduce) physical quantities such as pressure or acceleration […] into output signals […] that serve as inputs for control systems” (Norton, 1989). With constant progress in automotive development, sensor technology is also evolving. Fleming (2001) provides a thorough overview of sensors used in the automotive domain, including different types of temperature and pressure sensors, rotational sensors, air flow sensors, and acceleration sensors. Corresponding data becomes available in the vehicle’s Controller Area Network-bus (CAN-bus) and can be accessed for telematics applications or stored on event-data-recorders.

Event-data-recorders

Event-data-recorders, often referred to as “black boxes” or “crash recorders,” are small devices installed in cars for recording accident-related data (Haight, 2001). In their most common form, they continuously store different driving data (e.g.,
longitudinal and lateral acceleration, break appliance, steering angle, seat-belt wear) while overwriting old information. In case of an accident, the data gets frozen on the module and can be recovered to assign crash responsibility to involved motorists. I briefly give examples of insurance companies using event-data-recorder in subsection 2.3.

2.2 Literature Review

The following literature review details relevant contributions on telematics-based vehicle insurance. First, it starts off by illustrating the historical development of distance-based notions of vehicle insurance, which emerged in the late 1960s and are still at the core of many telematics-based insurance applications today. Next, literature on telematics-based insurance deals with its general perception as a means of mileage verification, its effects on mileage and accident reduction, and privacy and regulatory issues. Special consideration is given to scientific work on event-data-recorders in the insurance context, which are assumed to improve accident reconstruction and induce safer driving.

Note that although telematics-based vehicle insurance is relevant to other areas of research (e.g., actuarial science, electrical engineering, accident research, or information science), the following literature review does not consider these research areas in more detail. I also omit an in-depth analysis of telematics use for road pricing or traffic management and ignore literature on particular technical problems that does not relate to vehicle insurance. I nevertheless introduce further relevant work in subsequent sections when appropriate.

2.2.1 The Perception of Distance-based Vehicle Insurance over Time

The beginning

The concept of mileage-based insurance premiums was first discussed by Nobel laureate William Vickrey (1968). In his seminal work, he contrasts his view as an economist with legal and political approaches to insurance and identifies several flaws in motor insurance markets at that time. He argues that premiums neither reflect the external costs of accidents nor provide incentives for adjusting trip and driving habits accordingly. As common insurance policies back then hardly reflected any exposure in terms of mileage, premiums appeared as a fixed annual cost block to insurees, providing them with no incentive to cut this expense by reduced driving. As a solution he proposes to tie insurance payments to gasoline
purchases\(^9\) or tire sales. He rejects regular mileage reporting by insurees as a practical solution, owing to potential odometer fraud and misreporting, not aware of the degree of reporting accuracy and automation that has become possible with telematics today.

**The 1990s debate on Pay-at-the-Pump Insurance**

The immediate academic response to Vickrey’s work was nonetheless rather sparse. In fact, it was a public debate in the U.S. on soaring auto insurance costs and uninsured driving in the 1990s that revived the discussion on novel insurance schemes. The first scientific contributions on mileage-based insurance were published by researchers from California, particularly from the University of California, Berkeley, in the form of working papers, position statements, and reports to the State of California. The first publication in the area is by El-Gasseir (1990), who evaluates the benefits and feasibility of Pay-at-the-Pump insurance as part of the Conservation Report of the State of California Energy Resources Conservation and Development Commission. He introduces a premium structure that consists of two different components: the first block would be directly payable at the pump and would cover claims expenses; the second would make up direct payments to the insurer to compensate overhead and commission costs. Interestingly, El-Gasseir’s proposal also considers financial compensation for owners of fuel-inefficient vehicles in the form of year-end rebates, as the author is principally interested in an equitable and efficient provision of insurance.

In the following years, several other versions of Pay-at-the-Pump have been proposed. Tobias (1993) prefers a solution where only bodily injury costs are paid at the pump and the general liability scheme follows no-fault rather than traditional tort law systems.\(^{10}\) Another proposal comes from Sugarman (1993). While differing in several details (e.g., level of maximum compensation) from El-Gasseir’s (1990) original proposal, it actually modernizes his system of split charges and suggests

---

\(^9\) Litman (2008) refers to this pricing scheme as “Pay-at-the-Pump.”

\(^{10}\) Tort and no-fault compensation regimes provide two different approaches of assuming liability for accident damages. Tort encompasses traditional third-party liability. No-fault insurance is compulsory first-party coverage. The fact that each driver covers his or her own accident-related expenses has been shown to cut down on lawsuits that eventually become necessary for accident victims to get compensation under the tort system, thereby driving up transaction costs (Cummins & Tennyson, 1992).
annual fixed payments upon annual vehicle registration in addition to at-the-pump payments. He proposes this as a means of including other risk factors aside from mileage in the premium and suggests adjusting the annual payments for motorist age, driving history, and the existence of vehicle safety features.

Wenzel (1995) provides a thorough comparison of these and other Pay-at-the-Pump proposals and a review of the political debate at that time. He lists a total of nine different concepts (five in California, four outside of California) and studies them on equity grounds. He concludes that Pay-at-the-Pump increases insurance affordability for low-income households and helps to lower uninsured driving. The author also details the political perception of Pay-at-the-Pump and lists opponents and supporters of the system. On the opposing side are insurance companies, the California Trial Lawyers Association, the petroleum industry, and the California Chamber of Commerce. Supporters of the system include the Union of Concerned Scientists, minority representation groups, and environmental organizations.

In a more academic, quantitative analysis, Bernstein (1994) studies the effects of different auto insurance models on motorists. He assembles travel survey data, insurance, and government information in a simulation model and identifies different age and income groups that are better or worse off under a Pay-at-the-Pump coverage scheme. Finally, work by Khazzoom (2000) summarizes criticisms of this insurance system. From his synoptic perspective, he proposes a hybrid approach to Pay-at-the-Pump that retains elements consistent with free market operation under the existing system while restoring the pricing signal in order to capture the economic efficiency of motor insurance. For this purpose, the author first suggests fuel surcharges at a level that covers liability expenses (i.e., bodily injury and property damage expenditures) for all motorists. The at-the-pump premiums are intended to match the accident risk of “good” motorists and are allocated to insurers in proportion to their market share. Second, supplemental liability, collision, comprehensive, and medical coverage can be acquired directly from insurance companies. They also bill extra charges for worse drivers, certain vehicle characteristics, and administrative costs.

**Contributions after 2000**

Notwithstanding considerable discussion, pay-at-the-pump remained a political non-starter in the United States and has not yet been put into place. Nevertheless, academics sustain their discussion and continue to provide impact assessments and comparisons of different mileage-based insurance schemes. Edlin (2003) contributes a formalized discussion of Vickrey’s initial idea of distance-based
vehicle insurance. He analyzes the external effects of driving and determines external accident costs to be 7.5 U.S. cents per mile, while related insurance and gasoline costs are considerably lower at 4 and 6 U.S. cents per mile. In spite of these high social costs, the author argues that primarily high monitoring costs are an obstacle to a wider application of per-mile premium schemes. Furthermore, he explains the slow adoption of distance-based vehicle insurance with the fact that reduced external costs resulting from fewer accidents may not likewise translate into fewer collisions involving motorists under pay-per-mile schemes, as the accident externalities of other drivers (who commute under standard premium schemes) do not decrease accordingly. In other words, insurance companies do not consider benefits to other insurers in their business case calculations that come with the lowered accident risks of their own policyholders switching to distance-based coverage schemes. To facilitate a wider adoption of pay-per-mile insurance schemes, the author proposes odometer readings as a low cost solution that avoids expensive telematics hardware and can be done regularly during mandatory safety and emission checks. Edlin nevertheless falls short of considering the consumer rationale (i.e., privacy concerns and reluctance due to possibly rising premiums) in his strictly formalized economic approach. Also, the emerging capabilities of GPS technology are only a side node in his contribution.

Parry (2005) provides another formalized model. He compares gasoline taxes and distance-based vehicle insurance as means of reducing fuel consumption. He asserts that taxes are superior for lowering fuel demand “on cost effectiveness grounds as they exploit all behavioral responses for reducing fuel demand.” While they “encourage motorists to drive less, manufacturers to incorporate fuel saving technologies in new vehicles, and consumers to choose smaller, fuel-efficient vehicles,” pay-as-you-drive schemes fall short of improving fuel efficiency. This situation may change when considering the external effects of driving: given a specific fuel demand cut to be achieved, such a goal could be met exclusively through reduced driving under a distance-based regime. With tax increases, it would also be caused by switching to more fuel-efficient vehicles. As mileage-related externalities are generally higher than fuel-related externalities (which comprise only decreased greenhouse gas emissions and oil dependency and not lowered congestion and accidents), though the author supposes the welfare gains under distance-based insurance premiums to be more pronounced. He concludes that a full implementation of per-mile premiums would lower fuel demand by 11.4 billion gallons or 9.1%. Despite its academic rigor, the analysis in the end falls short of
considering insurers’ and consumers’ interests in pay-as-you-drive insurance, which again limits its practical relevance.

Zantema et al. (2008) use a simulation study to test the safety and network accessibility (i.e., congestion) effects of seven different instances of distance-based vehicle insurance. They study optional and mandatory insurance schemes for different combinations of road types, driving times, and driver ages and assess their potential to induce mode change, route change, and a reduction in the number of trips made. Their results indicate that the best model under safety considerations may not be equivalent to the best for increased network accessibility. Under the optimal model for road safety, premiums should vary both in driving time and road type, thus lowering collision incidents by more than 5%.

Bordoff & Noel (2008) provide a valuable, recent contribution to the discussion of telematics-based vehicle insurance. They identify three main barriers that prevent its wider application. First, high monitoring costs make distance-based pricing schemes unattractive for insurers, as potential benefits eventually spill over to other stakeholders (i.e., other motorists, other insurance companies, society) and cannot be recaptured by the offering companies. Benefits of pay-as-you-drive for a providing insurers are thus far less than those for society as a whole. Second, high regulation in U.S. insurance markets hinders the implementation of per-mile premiums schemes on a broader scale. In an example, they refer to the case of California, where it is illegal to charge a client with verified vehicle mileage a lower premium than an equal-risk customer with unverified mileage. Finally, patent issues may threaten the further dissemination of distance-based premium schemes, as key functionalities of distance-based premium schemes are covered by broad patents in the possession of U.S. insurer Progressive. To recapture the benefits of per-mile premiums for society in spite of the aforementioned obstacles, the authors propose a three step strategy: more actively passing state laws that allow distance-based insurance schemes, providing $15 million for funding pilot projects, and handing out tax cuts of $100 to insurance companies for every pay-per-mile customer they sign.

2.2.2 Telematics-based Concepts of Vehicle Insurance

Telematics technology as a means of mileage verification
Todd Litman of the Victoria Transport Policy Institute, Canada was among the first to discuss GPS-technology as a means to calculate insurance premiums. Initial research has contributed to the discussion of Pay-at-the-Pump insurance and has
compared it to Mileage Rate Factors\textsuperscript{11} and Per-Mile Premiums\textsuperscript{12} but has not considered telematics-based concepts (Litman, 1997). Indeed, the author includes these in subsequent analyses (Litman, 2008) and compares all approaches using twelve predefined evaluation criteria (implementation costs, road safety, energy and emission, etc.). The author concludes that per-mile premiums based on periodic odometer audits are superior to the other approaches, since they significantly improve actuarial accuracy and provide notable consumer savings, particularly to lower income households. Owing to high technology costs and privacy constraints, he assumes GPS-based pricing to be unsuitable for mandatory implementation.

Oberholzer (2003) proposes solutions to tackle these shortcomings and focuses on the development of business models for usage-based vehicle insurance in Switzerland. In his work, the author identifies low-risk and low-mileage drivers as potential target groups for telematics-based vehicle insurance, proposes premium schemes that account for road type use and time of travel, and calculates distinct per-kilometer rates for different combinations of location and time of travel. In a visionary passage, he suggests joint ventures between car manufacturers and insurance companies to make telematics-based vehicle insurance available to motorists. Nonetheless, the work exhibits several shortcomings. First, the assumed technology costs do not resemble reality and fall short of considering both hardware fitting costs (which are primarily non-declining labor costs) and expenses for data transmission in the analysis. Also the supposed decline of in-vehicle hardware costs is overoptimistic and not veridical (by following Moore’s law, the author presumes costs of in-vehicle hardware to reach $25 by 2006, yet today corresponding prices are still significantly above that level). Second, the described consumer perception of distance-based auto coverage seems somewhat skewed, as neither privacy considerations nor reluctant consumer behavior in fear of increased premiums are discussed. In summary, these drawbacks considerably limit the academic value of this contribution.

\textsuperscript{11} According to the author, Mileage Rate Factors refer to one-time mileage self-assessments by motorists upon policy sign-up without any further proof of the estimate provided.

\textsuperscript{12} According to the author, Per-Mile Premiums consider mileage in annual premium payments by periodic odometer audits. For each renewal period, an initial down-payment is made and lower / higher mileage leads to rebates / additional payments.
From a more technological perspective, Coroama (2006) describes the “Smart Tachograph,” a prototypical platform that enables the calculation of insurance rates based on individual behavior. The system collects sensor information (both GPS data and acceleration data) for personal insurance billing. Therefore, not only is mileage reflected in the insurance premium, but also the extent of speeding, hard breaking, and high lateral acceleration. Quite uniquely, the user gets direct feedback on driving performance through a graphical front-end interface. It shows per-mile performance-based vehicle taxes, insurance premiums, and the real time values upon which these calculations are based. The feedback is intended to sensitize motorists to their driving performance and give financial incentives for behavioral change.

Lindberg et al. (2005) report from two field experiments and expand the concept of distance-based vehicle insurance by monitoring real time driving (i.e., speeding) with electronic data recorders and GPS technology. The authors base their studies on data from the well-documented Borlänge experiments, a series of studies conducted with drivers from Borlänge, Sweden, which intended to answer various questions on travel and driving through the use of telematics technology. The authors show that a system which lowers announced premium discounts in case of speeding eventually deters motorists from going faster than allowed. The analysis shows some methodological flaws, however, and suffers from non-representative case groups, a rather crude statistical analysis, and missing information on the time persistency of the identified effects.

Telematics-based vehicle insurance and privacy
Scientific contributions that primarily concern themselves with privacy issues of telematics-based vehicle insurance have just emerged in recent years. With the ability of GPS-based driver profiling in mind (cf. Greaves & De Gruyter, 2002), several papers have discussed methods to preserve client privacy under telematics-based insurance schemes. Iqbal & Lim (2006) develop a solution where premiums are calculated directly in the vehicle. It does not disclose any position or driving information and reports only aggregated and anonymized data for billing. Payments are done online using “anonymous digital cash.” This approach lacks practical feasibility, however, and ignores the costs of such proprietary, non-standard infrastructure. In a second paper, Iqbal & Lin (2007) study to which extent profiling of individuals is possible by using GPS information. They find that spatial and temporal monitoring provides a good measure of personality traits, though technical limitations may create misleading results.
A similar approach to preserve privacy under telematics-based vehicle insurance is “PriPAYD,” as introduced by Troncosco et al. (2007). It keeps private information on the device and performs premium calculation directly on the on-board unit. Their system does not use any electronic cash; rather, it sends all information necessary for billing to the insurance company. Their system also gives end users full access to their GPS driving information (e.g., via the Internet) to cross check insurance rates.

Hollis & Strauss (2007) and Filipova (2007) view privacy from an economic perspective and consider it in their formalized analyses. The former show that under a pooling insurance scheme, especially consumers who are less concerned about privacy intrusions benefit especially from switching to telematics-based vehicle insurance. Upon changing insurance coverage, that is, dropping out of the pooling contract, the average contract price rises for insurers in perfectly competitive markets. This eventually makes other drivers with higher privacy valuations switch to per-mile-premium schemes as well, even though they are worse off in total utility under the new scheme. However, it is still a better option for them than staying in the pooling contract due to the general rise in premiums. Finally, the authors indicate clients who will be worse off either way (i.e., high risk drivers and motorists with a high privacy valuation). Filipova demonstrates that giving the user the freedom to decide ex-post (i.e., after an accident) whether to reveal GPS-tracking information is not an effective strategy for privacy conservation. This holds true even if the remuneration is based on the quality of the data disclosed, as allowing the consumer this freedom of choice imposes further information asymmetry upon the contract. The author concludes that this approach is not a solution to sustain motorist privacy and will not be profitable in the long run.

**Regulatory issues**

Academic work on regulatory issues regarding telematics-based vehicle insurance comes predominantly from the United States and reflects legal developments overseas. A uniform view cannot be given, however, as insurance regulation is done on the state level and may thus vary regionally. Guensler et al. (2003) examine these discrepancies in legal prerequisites. They report from a survey amongst state insurance commissioners on U.S. state regulatory support for distance-based premium schemes. Having received feedback from 43 states, the authors conclude that 37% of the states do not allow such insurance policies. In most others states, insurers need to prove the equity and transparency of the underlying pricing structures to the regulatory authorities.
A more recent publication on the regulation of pay-per-mile premiums comes from Greenberg (2007), who discusses regulatory incentives to promote a wider spread of distance-based vehicle insurance. For the U.S., he shows that regulatory support is provided neither on state nor on federal level, in spite of the benefits that could arise from its implementation. He details a competency problem, where auto insurance is regulated on the state level with consumer safety in mind, which complicates the promotion of mileage-based schemes motivated by social and environmental considerations on the federal level. The author proposes a solution to this problem, which is to amend legislation on fuel efficiency standards for light trucks as administered by the National Highway Traffic Safety Administration to fit distance-based vehicle insurance.

**Event-data-recorders in telematics-based vehicle insurance**

Previous literature has focused on the use of GPS technology in telematics-based vehicle insurance. Some insurers, however, offer products that solely use event-data-recorders. They value this technology for its ability to facilitate accident reconstruction and ascribe to it the potential to decrease speeding among motorists. Early contributions of interest from an insurance perspective come from Fincham et al. (1995) and Fincham (1996). The authors report from findings of the European “SAMOVAR” (Safety Assessment Monitoring On-Vehicle with Automated Recording) project that assessed different vehicle data recording technologies in a multinational study on 443 equipped vehicles and over 700 controls (i.e., taxis). The results are twofold. First, it shows that recorded accident data improves the quality of accident reconstruction. Second, the findings substantiate the belief that the sheer presence of event-data-recorders within vehicles has a positive effect on driving behavior and accident probability. On average, accident rates have dropped by 28% per vehicle during the trial phase. Wouters and Bos (2000) present similar results and report an accident reduction of 20% amongst the 270 vehicles involved in their study. Interestingly, these results are opposite to work done by Heinzmann & Schade (2003), who find no significant effects of the presence of event-data-recorders on more disciplined and thoughtful driving.

### 2.3 Practical Applications

The following table provides an overview of insurance companies that offer or have offered telematics-based vehicle insurance. For each product, I detail its general concept, the technology platform used, the method of data transmission, and its
current status. The information provided reflects the insurance telematics landscape as of January 2010, although this list may not be exhaustive.13

<table>
<thead>
<tr>
<th>Product</th>
<th>Concept &amp; Status</th>
</tr>
</thead>
</table>
| Admiral Insurance “Pay how you drive” (UK) [http://www.greenroad.com/first_pay_how_you_drive.html](http://www.greenroad.com/first_pay_how_you_drive.html) | **Insurance concept:** Insurance scheme for young drivers, providing a 25% insurance discount upon sign-up and another 10% for every month of safe and eco-friendly driving  
**Technology platform:** GPS-based telematics unit by GreenRoad including accelerometer  
**Data transmission:** Via GPRS  
**Product status:** Available to selected UK councils |
| Aioi “Pay-as-you-drive” (Japan) [http://www.ioi-sonpo.co.jp](http://www.ioi-sonpo.co.jp) | **Insurance concept:** Distance-based vehicle insurance  
**Data transmission:** Via GPRS  
**Product status:** Available to drivers of Toyota and Lexus vehicles |
| Amaguiz.com “Pay-as-you-drive” (France) [http://www.amaguiz.com](http://www.amaguiz.com) | **Insurance concept:** Distance-based vehicle insurance with a low monthly fixed block and a variable, mileage-dependent component; automated emergency calls, stolen vehicle recovery, and driving statistics come as free add-on services  
**Technology platform:** GPS-based telematics unit  
**Data transmission:** Via GPRS  
**Product status:** Available |
| American Family Insurance (USA) [http://www.amfam.com/microsites/teen-safe-driver/events/default.asp](http://www.amfam.com/microsites/teen-safe-driver/events/default.asp) | **Insurance concept:** An audio / video recording device behind the rear view mirror records erratic driving behavior, such as hard breaking and acceleration, or collisions ten seconds before and after an event; both the driver and the road ahead get recorded; videos are analyzed by insurance experts and made available to parents and teen drivers for review  
**Technology platform:** Audio / video recording device with GPRS unit  
**Data transmission:** Via GPRS  
**Product status:** Available |
| Aryeh Insurance (Israel) [http://www.aryeh.co.il](http://www.aryeh.co.il) | **Insurance concept:** Distance-based vehicle insurance  
**Technology platform:** Telematics device retrieves mileage information from vehicle electronics  
**Data transmission:** Wireless transmission at certain petrol stations  
**Product status:** Available |
| Aviva “Autograph” (Canada) [http://www.](http://www.) | **Insurance concept:** Insurees receive a premium discount upon policy renewal if they decide to send their driving data (distances driven, night- and daytime driving hours, driving speeds) to the insurer; data transmission is optional; clients can see the discount they are eligible for prior to transmitting the data to the |

13 Insurance products that do not use telematics technology, though they pursue other mileage verification techniques (e.g., odometer readings, as with Real Insurance, [http://www.payasyoudrive.com.au](http://www.payasyoudrive.com.au), or taking pictures of the odometer, as with MileMeter, [http://www.milemeter.com](http://www.milemeter.com)), are not considered.
<table>
<thead>
<tr>
<th>Insurer</th>
<th>Service Description</th>
<th>Technology Platform</th>
<th>Data Transmission</th>
<th>Product Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AutoGraph Insurance</strong>&lt;br&gt;autographinsurance.com</td>
<td>Insurer; trip logs and driving statistics are available on the client’s local PC&lt;br&gt;<strong>Technology platform:</strong> Device connected to vehicle diagnostics port; mileage is derived from vehicle electronics, not from a GPS signal&lt;br&gt;<strong>Data transmission:</strong> Device connected to PC via USB; data sent to insurer online</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AXA “Autometrica” (Italy)</strong>&lt;br&gt;<a href="http://www.axa-italia.it">http://www.axa-italia.it</a></td>
<td><strong>Insurance concept:</strong> Distance-based vehicle insurance&lt;br&gt;<strong>Technology platform:</strong> GPS-based telematics unit by Octo Telematics&lt;br&gt;<strong>Data transmission:</strong> Via GPRS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AXA-Winterthur “Crash Recorder” (Switzerland)</strong>&lt;br&gt;<a href="http://www.crashrecorder.ch">http://www.crashrecorder.ch</a></td>
<td><strong>Insurance concept:</strong> Event-data-recorder fitted in vehicles, drivers receive a 15% premium discount off the standard insurance premium&lt;br&gt;<strong>Technology platform:</strong> Event-data-recorder&lt;br&gt;<strong>Data transmission:</strong> Accelerometer information retrieved from the data recorder in case of an accident</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cobra Wunelli “Coverbox” (UK)</strong>&lt;br&gt;<a href="http://www.coverbox.co.uk">http://www.coverbox.co.uk</a></td>
<td><strong>Insurance concept:</strong> Distance-based vehicle insurance; stolen vehicle recovery comes as add-on service; the client can choose coverage from multiple insurers (Groupama Insurance, Equity Insurance Group, The Co-operative, Allianz Insurance UK, Markerstudy Insurance)&lt;br&gt;<strong>Technology platform:</strong> Cobra tracking device using GPS technology&lt;br&gt;<strong>Data transmission:</strong> GPRS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Generali “Protezione Satellitare” (Italy)</strong>&lt;br&gt;<a href="http://www.generali.it">http://www.generali.it</a></td>
<td><strong>Insurance concept:</strong> Telematics-based vehicle insurance; motorists receive premium discounts upon policy sign-up; an emergency button for roadside assistance, automated emergency calls, stolen vehicle recovery, and driving statistics come as free add-on services&lt;br&gt;<strong>Technology platform:</strong> GPS-based telematics unit&lt;br&gt;<strong>Data transmission:</strong> Via GPRS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GMAC Insurance (USA)</strong>&lt;br&gt;<a href="http://www.gmacinsurance.com">http://www.gmacinsurance.com</a></td>
<td><strong>Insurance concept:</strong> Distance-based vehicle insurance with a discount for low-mileage drivers (i.e., motorists that travel less than 15,000 miles a year)&lt;br&gt;<strong>Technology platform:</strong> OnStar telematics platform, as available in many General Motors vehicles&lt;br&gt;<strong>Data transmission:</strong> GPRS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hollard “Pay-as-you-drive” (South Africa)</strong>&lt;br&gt;<a href="http://www.payasyoudrive.co.za">http://www.payasyoudrive.co.za</a></td>
<td><strong>Insurance concept:</strong> Distance-based vehicle insurance; drivers can subscribe to insurance plans with varying monthly mileage allowance and pay according premiums; if allowed mileage is higher than actual travel, kilometers are accounted to the next month; excess mileage at the end of the billing period is billed at an individual per-mile rate; stolen vehicle recovery comes as an add-on service&lt;br&gt;<strong>Technology platform:</strong> GPS-based Tracker SkyTrax telematics unit&lt;br&gt;<strong>Data transmission:</strong> Via GPRS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Liberty Mutual “Onboard Advisor” (USA)</strong></td>
<td><strong>MAPFRE “YCAR” (Spain)</strong></td>
<td><strong>Norwich Union “Pay-as-you-drive” (UK)</strong></td>
<td><strong>Progressive “MyRate” (USA)</strong></td>
<td><strong>Progressive “TripSense” (USA)</strong></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------------------------</td>
<td>-----------------------------------------</td>
<td>-------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td><strong>Insurance concept:</strong> Fleet management application for commercial vehicles allowing insurance discounts; 15% premium discount in the first year and 40% on policy renewal; system assigns safety scores based on speeding, hard breaking, and cornering; fleet managers can track vehicles and use the provided information to monitor and coach motorists</td>
<td><strong>Insurance concept:</strong> Telematics-based vehicle insurance for drivers of age 18 to 30; drivers receive a premium discount off the standard insurance premium; automated emergency calls, stolen vehicle recovery, and driving statistics come as free add-on services; depending on mileage, driving performance, and accident involvement, motorists can receive an additional prepaid card for gasoline purchases as a bonus for good driving behavior of up to 600 Euro</td>
<td><strong>Insurance concept:</strong> Distance-based vehicle insurance; for each driver, road-type (motorway, dual carriageway, single lane roads, 20/30/40 mph speed limits) specific per-mile premium rates are calculated (usually a few pence); during pre-defined high-risk hours, a prohibitively high per-mile rate of £1 is charged; peak hours are 11pm - 6am for young drivers and usual rush hours for other drivers; automated emergency calls, stolen vehicle recovery, and speed camera detection come as an add-on services</td>
<td><strong>Insurance concept:</strong> Telematics-based vehicle insurance that bases premium discounts on mileage, driving hours, weekdays of driving, and driving style (excessive acceleration and hard breaking); clients receive premium discounts upon policy renewal; trip logs and driving statistics can be viewed online</td>
<td><strong>Insurance concept:</strong> Distance-based vehicle insurance; trip logs and driving statistics can be viewed online</td>
</tr>
<tr>
<td><strong>Technology platform:</strong> GPS-based telematics unit by General Electric including accelerometer</td>
<td><strong>Technology platform:</strong> GPS-based telematics unit</td>
<td><strong>Technology platform:</strong> GPS-based telematics unit</td>
<td><strong>Technology platform:</strong> Device connected to vehicle diagnostics port; mileage is derived from vehicle electronics, not from a GPS signal</td>
<td><strong>Technology platform:</strong> Device connected to vehicle diagnostics port; mileage is derived from vehicle electronics, not from a GPS signal</td>
</tr>
<tr>
<td><strong>Data transmission:</strong> Via GPRS</td>
<td><strong>Data transmission:</strong> Via GPRS</td>
<td><strong>Data transmission:</strong> Via GPRS</td>
<td><strong>Data transmission:</strong> GPRS unit in device</td>
<td><strong>Data transmission:</strong> Device connected to PC via USB, data sent to insurer online</td>
</tr>
<tr>
<td><strong>Product status:</strong> Available</td>
<td><strong>Product status:</strong> Available</td>
<td><strong>Product status:</strong> Cancelled</td>
<td><strong>Product status:</strong> Available to drivers in selected U.S. states</td>
<td><strong>Product status:</strong> Cancelled, continued as Progressive MyRate</td>
</tr>
</tbody>
</table>
Table 1: Overview of telematics-based vehicle insurance products

<table>
<thead>
<tr>
<th>Product name</th>
<th>Website</th>
<th>Insurance concept</th>
<th>Technology platform</th>
<th>Data transmission</th>
<th>Product status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safeco Teensurance (USA)</td>
<td><a href="https://www.teensurance.com/">https://www.teensurance.com/</a></td>
<td>GPS-based vehicle monitoring device determines current vehicle location; it delivers notifications to parents if a driver is speeding, leaving or arriving at certain locations, or operating the vehicle outside set spatial or temporal boundaries</td>
<td>GPS-based telematics unit</td>
<td>Via GPRS</td>
<td>Available</td>
</tr>
<tr>
<td>Sara “Sarafree” (Italy)</td>
<td><a href="http://www.saraassicurazioni.it">http://www.saraassicurazioni.it</a></td>
<td>Distance-based vehicle insurance, consists of a fixed rate that equals 30% of a conventional premium scheme and a variable rate that reflects mileage (Sarafree km); alternatively, a variable rate is charged for every day the car is driven (Sarafree Day)</td>
<td>GPS-based telematics unit</td>
<td>Via GPRS</td>
<td>Available</td>
</tr>
<tr>
<td>Sara “Sarasat” (Italy)</td>
<td><a href="http://www.saraassicurazioni.it">http://www.saraassicurazioni.it</a></td>
<td>Discount on vehicle theft and fire insurance when installing a tracking device</td>
<td>GPS-based telematics unit by Movitrack, Viasat, or Octo Telematics</td>
<td>Via GPRS</td>
<td>Available</td>
</tr>
<tr>
<td>Unipol “Transat” (Italy)</td>
<td><a href="http://www.unipolonline.it">http://www.unipolonline.it</a></td>
<td>Telematics-based vehicle insurance; premium discount for subscribing motorists; automated emergency calls and stolen vehicle recovery come as an add-on services</td>
<td>GPS-based telematics unit by Octo Telematics</td>
<td>Via GPRS</td>
<td>Available</td>
</tr>
<tr>
<td>Uniqa “Safeline” (Austria)</td>
<td><a href="http://www.uniqa.at">http://www.uniqa.at</a></td>
<td>Telematics-based vehicle insurance; premium discount for subscribing motorists; automated emergency calls and stolen vehicle recovery come as add-on services</td>
<td>GPS-based telematics unit by Octo Telematics</td>
<td>Via GPRS</td>
<td>Available</td>
</tr>
<tr>
<td>WGV-Online “young &amp; safe” (Germany)</td>
<td><a href="http://www.wgv-online.de">http://www.wgv-online.de</a></td>
<td>Telematics-based vehicle insurance that gives driving novices a financial incentive to avoid excessive speeding; telematics hardware continuously compares the current and the allowed speed and gives an acoustic signal in case of speeding; if the vehicle does not slow down within 15 seconds, the speeding incident is recorded on the car’s onboard unit; if 12 speeding occurrences are recorded on the telematics device within a year, the insurance company is informed and a previously issued premium reduction becomes void</td>
<td>GPS-based telematics unit</td>
<td>Via GPRS in case of too many speeding incidents</td>
<td>Discontinued after pilot phase</td>
</tr>
</tbody>
</table>
2.4 E-Call

eCall is a system for in-vehicle emergency calls. It is part of the eSafety initiative, a joint project of public authorities and different industries for improving road safety through new information and communication technologies. eSafety deals with both vehicle-based technologies (i.e., electronic stability control, blind spot monitoring, adaptive headlights, obstacle and collision warning, lane departure warning) and infrastructure-related systems (i.e., eCall, extended environment information, real-time traffic and travel information, dynamic traffic management, local danger warnings, speed alert). It forms the European strategy for the research, development, and deployment of intelligent safety systems for advanced driver assistance.

With the eCall system, emergency calls are generated either manually by motorists or automatically if in-vehicle sensors detect an accident. On activation, eCall establishes a connection to an emergency response service, usually a 112 public safety answering point, where occupants can communicate with an emergency service operator. At the same time, an eCall transmits data on the incident, including time, location, the direction of vehicle travel (important for bringing help to people on lane-separated motorways), the vehicle identification number, and an eCall qualifier, which details whether the call was initiated manually or automatically. eCall is designed to function in automobiles throughout Europe even when traveling outside their home countries. Once fully deployed, the precise information on accident locations will improve the efficiency of the rescue chain. Road accident victims will benefit from faster medical care, which is projected to reduce road fatalities by 2500 every year and lower the severity of injuries for many more (eSafetySupport, 2009).

The original goal of the eCall initiative was to make the system standard in all new vehicles by 2010. Deployment by public authorities, car manufacturers, and mobile network operators was voluntary, but implementation results have been disappointing. Today, eCall is not operational in any EU country (as of March 2010). Cost issues are the primary reason why several countries (Denmark, France, Ireland, Latvia, Malta, and the United Kingdom) have not even committed to the common EU-wide standard, even though the technology is now ready for implementation. In late 2009, the European Commission warned that if no significant progress is made in rolling out eCall, binding regulatory measures will eventually be proposed on an European level to bring eCall to European citizens no later than 2014 (Europa.eu, 2009).
What makes eCall interesting from an insurance perspective is the possibility to run third-party providers’ services on its infrastructure. However, many aspects in this respect are unclear today, including data ownership and who actually should provide such services. While car manufacturers and telecommunications companies are at the forefront of selling such upgrade applications, whether insurers could use this system as well and at which costs is questionable. Although it is clear that satellite positioning and wireless communication systems will become standard in new vehicles in the near future, access to these technologies will probably be exclusive to companies of certain industries, who may sell it to other industries or allow it only in a cooperative way. Insurers fear that eCall might threaten established business models and thus are reluctant to support the system (GDV, 2006).
3. Theoretical Framework

The review of the ascribed benefits of telematics-based vehicle insurance and their lack of practical validity have shown the importance of exploring possible advantages of telematics technology for insurance companies in more detail. As I argued earlier, creating value from the analysis of the new data to better understand insurance clients might be a worthwhile approach to overcome the lacking business rationale of telematics-based insurance.

As an overall theoretical framework, I outline Principal-Agent Theory in subsection 3.1 and describe the general value of information in the insurance context. In settings where information is unevenly distributed between the principal (the insurer) and the agent (the client), several problems occur that might cause inefficiencies and market failure. I show that telematics-based vehicle is a viable technological approach to overcome these problems and is congruent with common solutions proposed in academic literature.

For the operational perspective of this dissertation, my quantitative analysis examines the accident impact on travel and driving behavior. To understand the course of behavioral change and the role of personal traits in it, a conceptualization of this process is necessary. The Theory of Planned Behavior provides a framework to explain such changes in human behavior and their persistency. I develop the theory with bearing upon travel and driving behavior research in subsection 3.2.

3.1 Principal-Agent Theory

3.1.1 An Introduction to Principal-Agent Theory

Principal-Agent Theory investigates the difficulties in contractual and hierarchical relationships that occur under conditions of asymmetric or incomplete information. Having originated from the theory of incomplete contracts (cf. Coase, 1937, 1960), it considers problems that arise when a principal delegates work to an agent in situations where their goals in performing the task may not be congruent (Jensen & Meckling, 1976). Principal-Agent Theory provides a framework for explaining the decision making processes of economic entities under uncertainty; its empirical validity (cf. Eisenhardt, 1989) has fostered its application in areas like accounting (e.g., Demski & Feltham, 1978), marketing (e.g., Basu et al., 1985), organizational
behavior (e.g., Eisenhardt, 1985), sociology (e.g., Eccles, 1985), and insurance (e.g., Spence & Zeckhauser, 1971).

Most generally, Principal-Agent Theory assumes two parties agreeing upon a contract to which the agent has more information than the principal. Depending on the time this information asymmetry arises, either hidden characteristics exist prior to concluding the contract or the problem of hidden action emerges after the parties sign the contract. The corresponding information problems are commonly referred to as adverse selection and moral hazard. The former was first discussed in the seminal work by Nobel laureate George A. Akerlof (1970), who applied it to the example of second-hand vehicle sales. As potential car buyers have incomplete information about the condition of single vehicles and cannot sort out defective vehicles with certainty, they make average-priced offers to potential sellers. These may not be sufficient for owners of good-quality cars, who eventually drop out of the market. Thus, negative selection is taking place, where only “lemons,” that is, cars of substandard quality, remain on sale.

Arrow (1963) identified the latter market failure, moral hazard. He discusses it in the context of health care and argues that full coverage of losses may foster risk-prone behavior. As a consequence, some markets for insurance against rare, uncertain events may have failed to emerge, as optimal market equilibria are non-existent. Pauly (1968) challenges Arrow’s rationale and argues that the absence of certain insurance markets may not necessarily be an indication of non-optimality. It may be the result of market inefficiencies that would exist even if all insurance-seekers behave in a risk-averse fashion. In addition, the author state that “moral hazard” in fact has little to do with immoral behavior, but follows economic market considerations. Rothschild & Stiglitz (1976) contribute to this discussion by challenging common assumptions about equilibria in competitive insurance markets. They show that asserting a single price equilibrium in such market
environments is not viable if imperfect information is present and that equilibria may only be reached under odd market conditions.

3.1.2 Solutions to Principal-Agent Problems and the Role of Telematics Technology

Academics have proposed various mechanisms to overcome information asymmetries and to align the interests of principals and agents. Spence (1973) suggests signaling as a viable solution, where information on one’s type is transferred to the other party. In his job market study, the author regards degrees and certificates as believable signals. In the insurance context, driver records serve a similar purpose. Another alternative comes from Stiglitz (1975), who pioneered the idea of screening: if the less informed party provides a set of (contractual) choices, it can induce the better informed partner to reveal its unobservable characteristics if the contracts are designed in a way that the choice depends on that hidden information. Screening today is common in both life and non-life insurance markets and is done by offering clients contracts with different deductibles or co-payments. The provision of such incomplete coverage may thereby not only decrease the information asymmetry up front; it may also deter clients from engaging in risky behavior later on, that is, lowering moral hazard after contract conclusion (Shavell, 1979). Other alternatives frequently proposed to overcome information asymmetry are hierarchical control and the provision of appropriate incentives.

Telematics technology serves as a new source of information in the motor insurance context. Regardless of whether GPS-based systems or event-data-recorders are used, insurance companies can get new insights into their clients’ driving behavior. The new data reduces information asymmetry, as more accurate data on people’s driving behavior becomes available. Consciousness of the data collection might deter people from engaging in risky driving behavior and reduce moral hazard. Beyond that, telematics-based vehicle insurance can also be used as a screening device. Insurance customers who opt for this coverage scheme instead of a conventional one eventually signal their risk type unconsciously. Telematics technology can thus be regarded as a means to reduce information asymmetry and alleviate the principal-agent problem.
3.2 The Theory of Planned Behavior

3.2.1 An Introduction to the Theory of Planned Behavior

The Theory of Planned Behavior is currently one of the most widely applied models for explaining the intricacy of human social behavior. It employs some of the central concepts of the social and behavioral sciences and defines them in a domain-independent manner, which allows predicting and comprehending human behavior in different contexts. Consequently, scientists have applied the Theory of Planned Behavior to a variety of research domains in order to investigate human behavior, including smoking (Norman et al., 1999), illegal substance abuse (Conner & McMillan, 1999), and career choice (Vincent et al., 1998). It has also been applied in the traffic safety domain and offers a valuable method for organizing and understanding the various influences that affect vehicle usage. In this context, support for its efficacy has been found in studying behaviors like seat belt usage and driving under the influence (Gordon & Hunt, 1998), speeding (Parker et al., 1996), transportation mode choice (Bamberg et al., 2003; Bamberg, 2006), and in intentions to commit driving violations and engage in risky driving behavior (Parker et al., 1992; Parker et al., 1995).

The Theory of Planned Behavior emerges from the Theory of Reasoned Action (cf. Fishbein & Ajzen, 1975; Ajzen & Fishbein, 1980) and postulates that human action is governed by three different kinds of personal characteristics, namely “beliefs about the likely consequences of the behavior (behavioral beliefs), beliefs about the normative expectations of others (normative beliefs), and beliefs about the presence of factors that may further or hinder performance of the behavior (control beliefs)” (Bamberg et al., 2003). In their corresponding aggregates, behavioral beliefs stimulate a positive or negative attitude towards a behavior, normative beliefs give rise to perceived social pressure (often referred to as subjective norm), and control beliefs reflect the perceived easiness or difficulty of performing the behavior. The latter are presumed to reflect both the instrumental resources needed for action and the opportunities for engaging in the actual behavior (Bagozzi & Kimmel, 1995). Together, attitude, subjective norm, and perceived behavioral control determine the formation of behavioral intention, where each factor is positively related to the intention’s favorableness (Bamberg et al., 2003). Behavioral intention is, however, only one determinant of final behavior; especially in situations with limited volitional control of the behavior (i.e., in circumstances where the subject cannot decide at will whether to perform the behavior in question or not), behavioral
control eventually limits the accomplishment of certain activities, regardless of any existing intentions to do so (Ajzen, 1991). Therefore, it is advisable to use behavioral control along with behavioral intention as determinants of actual behavior.

### 3.2.2 The Role of Regret

A limitation frequently attributed to the Theory of Planned Behavior is the lack of an affective processing component (Ajzen, 1991). A substantial body of research contends that people are affected by a post-behavioral feeling known as regret (Loomes & Sugden, 1982; Richard et al., 1995; Richard et al., 1998). Regret can be defined as the consideration of the possibility that one has made a wrong decision (Eagly & Chaiken, 1993). The opposite of to regret is to rejoice; as Newman et al. (2004) put it, “an individual will either experience feelings of regret or rejoice after comparing the outcome of the alternative they chose with the outcome that might have been if they had chosen another alternative.”

In the traffic safety domain, regret can occur if risky driving behavior has caused dangerous situations on the road or has even culminated in a vehicle crash. It might make motorists rethink their behavior, especially if the accident was their fault. This process of deliberation might trigger behavioral changes, that is, driving slower or taking fewer risks while on the road.¹⁴ In this sense, regret is similar to new information in the context of the Theory of Planned Behavior, which is discussed next together with habitual choice.

### 3.2.3 Habitual Behavior and the Role of New Information

The Theory of Planned Behavior clearly assumes human social behavior to be reasonable. While people’s beliefs may be unsubstantiated or biased, their attitudes, perceived social pressure, and perceived behavioral control can be deduced reasonably from these beliefs. Successive behavioral intention and ultimate behavior are thus a direct consequence of these salient beliefs and congruent with them (Bamberg et al., 2003). Scholars frequently challenge this conviction,  

---

¹⁴ As Adams-Guppy & Guppy (1995) show in the context of company car drivers, insurance coverage might actually alleviate some of the negative financial consequences of risky behavior and crashes, thus limiting overall regret consequences.
however, arguing that human behavior may also follow certain habits or automatisms (Triandis, 1977; Bagozzi, 1981; Ronis et al., 1989; Fazio, 1990; Aarts et al., 1998; Ouellette & Wood, 1998; Aarts & Dijksterhuis, 2000). Empirical findings support this rationale, which indicates that human behavior, “rather than being completely reasoned, is at least in part under the direct control of the stimulus situation, that is, that it habituates with repeated performance” (Bamberg et al., 2003). The strong predictive power of past behavior on future actions therefore makes it appropriate to model past activity as an independent predictor of subsequent behavior.15

An incident that might weaken the habitual relationship between past and later behavior is the appearance of new information. As articulated in the Theory of Reasoned Action, new information may alter the cognitive foundations upon which intentions and behaviors are based. Even activities that have become routine over time eventually change at a discernible level once new information becomes available to the individual. As a consequence, the predictive validity of previous actions for future behavior is declining and may even become insignificant (Bamberg et al., 2003).

In the context of this dissertation, the negative experience of collision involvement serves as new information to the models of travel and driving behavior. Road crashes create exceptional conditions most drivers are not familiar with, often causing serious harm or even motorists’ death. Aside from physical injuries, which may lead to hospitalization and permanent convalescence, road victims often face psychiatric consequences, including post-traumatic stress disorder, emotional distress, travel anxiety, and depression (Mayou et al., 1993). A lack of driving confidence immediately after the accident is also an outcome of crash involvement (Mayou et al., 1991). Following Ajzen (1991), this notion of confidence (or perceived self-efficacy), which “is concerned with the judgments of how well one can execute courses of action required to deal with prospective situations” (Bandura, 1982), is identical to the concept of perceived behavioral control. This suggests establishing a direct relationship between actual behavior (or its outcomes, respectively) and control beliefs in modeling travel behavior. It acts as a feedback

15 In the later statistical analyses, I will account for this dependency by an analysis of covariance design, where past behavior serves as a covariate to later behavior.
loop within the model of the Theory of Planned Behavior and leads to a model representation as in Figure 10. The model is capable of catching both new information and past behavior without the integration of any new antecedent constructs (as done for example in the longitudinal study by Bamberg et al., 2003).

Figure 10: The Theory of Planned Behavior with feedback on control beliefs

At this point, one might argue that the experience of a road accident also affects other cognitive considerations besides perceived behavioral control. It might cause changes in attitude towards individual transport, which motorists eventually perceive as more dangerous after an accident. Also, subjective norm might be affected, if perceived social pressure to drive less is imposed by a peer group (e.g., friends, family). In the context of this dissertation, however, I limit myself to assuming only a feedback between behavior outcome and control beliefs; as my dissertation uses GPS-based information rather than survey data, it is not possible to study effects on the different cognitive constructs separately.

In summary, I postulate that accident involvement as a negative outcome of vehicle usage changes motorists’ travel and driving behavior. I assume that such an experience challenges the cognitive foundations of driving, namely by impacting pre-located behavioral control beliefs. The defined feedback cycle clarifies why subsequent accident-free travel helps to restore initial control beliefs, thus restoring driver confidence and eventually making motorists reestablish their travel and driving routines over time. This dissertation will test both the existence of the postulated effects and their persistency; both can be assumed by the proposed conceptual model based upon the Theory of Planned Behavior.
3.2.4 Relationship Moderators

Accidents pose extraordinary situations to all motorists. However, single driver subgroups may perceive accidents differently, with underlying group variables serving as relationship moderators. For example, it might be assumed that accident severity and the question of who was at fault influence the degree to which the collision experience affects driver confidence and post-crash vehicle operation. Likewise, the degree of behavioral change might be moderated by annual driver mileage, past years of driving experience, driver type, driver aggressiveness, or demographic factors like gender and age. Such propositions seem justified as driver characteristics are generally a strong predictor in explaining differences in travel and driving behavior (cf. Gordon et al., 1989; Harré et al., 1996; Stradling, 2000). I test the significance of moderators available within this research (i.e., crash severity, pre-crash annual mileage, type of vehicle ownership (male, female, company)) in subsequent analyses.

3.2.5 Environmental Influences

Living situation

Vehicle operation is not a static behavior and tends to change over time. While short-term travel behavior might be rather fixed and predetermined by the daily routines of one’s personal living situation (i.e., job, children, hobbies, etc.), it usually exhibits much more long-term variability and changes with residence, job, and family situation. To understand changes in travel and driving behavior after an accident, any confounding influences due to living situation change should be eliminated before proceeding with the analyses. This, however, is not possible on grounds of sole GPS data; although one can infer from the data that a change in permanent residency has occurred (i.e., by determining the most frequently visited location of each driver), it is not discernible what caused this change without any additional data from supplementing surveys. In this research, the time period under review for each driver is considerably short though, which allows me to assume that the living situation of an individual stays constant during the study period. I thus do not adjust the data to reflect changes in the drivers’ living situation.

---

16 Most of the analyses involve data from the first pre- and post-accident month, separated by the accident / reference month, which makes a total study period of 3 months. For the analysis of the
Seasonality
Driving season is another strong influence on travel and driving behavior, where people account for changing weather and lighting conditions throughout the year. Especially in countries that have distinct seasons, several driving adjustments may become perceivable: motorists prepare their cars with winter tires and anti-freezing agents; they avoid unnecessary and excursive trips; and some even deregister their vehicles for the winter season, although most drive more slowly and more carefully.\footnote{Note that the seasonal impact on driving is also reflected in the number of accidents, which are highest in the winter months, namely December and January (Brown and Baas, 1997).} However, such seasonal fluctuations confound the analysis of accident-induced changes in vehicle use and thus have to be eliminated, since behavioral changes should only represent the accident influence and should not become prevalent because of seasonal effects. I account for seasonality by sampling accident-involved drivers equally from the 12 calendar months. Pre- and post-accident measures then consider data from all months, thereby cancelling out potential seasonality effects.\footnote{I provide a detailed description of this methodological step in subsection 5.4.1.} Drivers in the control group, that is, motorists who did not have an accident during the study period, get assigned a reference month (i.e., by randomly selecting one of the 12 calendar months), which similarly allows derivation of pre- and post travel and driving statistics in absence of seasonality effects.

\textit{persistency of any accident effects, data from the pre-accident month is compared to the 5 post-accident months, thus the study period stretches over 7 months.}
4. Hypotheses Development

In this section, I develop several hypotheses with which I test the influence of accidents on travel and driving behavior. In particular, I develop the argument that crash involvement has an impact on the number of trips motorists take, their mileage, and the speeds at which they travel. I have organized my propositions as sets of hypotheses, testing assumed influences and group differences on all three variables each time. The hypotheses intend to answer research sub-questions 2 and 3; their outcome serves as the input to research sub-question 5 in the follow-up.

I organize this section as follows: Subsection 4.1 develops the argument that accidents have a constricting influence on travel and driving behavior. Next, I contend that this influence does not persist over time and may fade as road crash victims recuperate from their accident experience. In subsections 4.3 and 4.4, I hypothesize that the accident impact on travel and driving varies for different subsets of motorists. The influence of accident severity on the magnitude of behavioral change is considered in subsection 4.5. A deliberation of the impact of accidents on different trip types follows in subsection 4.6. I conclude with outlining the argument that motorists try to avoid accident-like situations and the crash location in post-accident driving.

4.1 The Accident Impact on Travel and Driving Behavior

In this subsection, I develop the argument that accident involvement impacts motorists’ travel and driving behavior. I develop the underlying rationale independently for travel and driving characteristics, a scheme which I pursue in all of the following subsections except for subsections 4.6 and 4.7.

Travel behavior

Road accidents pose a major problem in both industrialized and developing countries. In the European Union, they account for more than 40,000 fatalities every year (ERF, 2009) and are the main cause of death for people aged 20 to 44 (Eurostat, 2006). They cause both serious physical and mental harm to victims and may result in hospitalization, protracted convalescence, or permanent disabilities. Not surprisingly, the economic losses owing to road crashes are immense, reaching an estimated 180 billion Euro in the European Union every year (WHO, 2004).

From the perspective of travel behavior research, accidents can be regarded as interventions to everyday travel and driving routines. A substantial body of
literature investigates the effects of different interventional measures, primarily with a focus on reducing commuter car use (cf. Kearney & De Young, 1996) and breaking habitual travel patterns in general (Gärling & Axhausen, 2003). In the former review article, the authors report on 29 empirical studies and their effectiveness in deterring work-related vehicle usage. From their synoptical perspective, they conclude that material incentives and disincentives prove particularly effective in changing people’s travel behavior. However, such changes often do not sustain if the incentives are waived. Likewise, the imposition of disincentives tends to cause negative reactions and protests by motorists. Kearney & De Young also find that informing motorists about alternative modes of transportation on a general level is ineffective in changing travel behavior. However, this improves if the information is more specific and considers personal preferences, for example through the establishment of ride sharing coordinators.

The work by Gärling & Axhausen provides a good overview of research on interventional measures on car-use habits in general. Their article, which is an introductory note to a special journal issue on habitual travel choice, summarizes both past and present work on the subject. Interventions under review include temporary road closures (Fujii et al., 2001), residential relocation (Bamberg, 2006), the provision of transitory free admission to public transport (Fujii and Kitamura, 2003), and raising awareness for transportation alternatives (Garvill et al., 2003). Typical measures to quantify the impact of interventions on travel behavior include the number of trips made, distances traveled, trip starting and arrival hours, weekday and weekend travel, the size of the activity space, mode choices, and shares of trips with different purpose (Burnett & Hanson, 1982; Hanson & Huff, 1986; Axhausen et al., 2000; Schlich & Axhausen, 2003; Schönfelder et al., 2006).

Accidents pose another form of intervention to travel behavior, although they differ in two important aspects from the aforementioned measures. First of all, road crashes are not administered purposely and are therefore not constrained to a preselected driver group. They affect all motorists with a low, yet non-equal probability of occurrence, which requires looking at a large driver population as part of a natural experiment (cf. subsection 5.1 for details on the experimental setting of this dissertation) in order to gather sufficient data. Second, the interventional quality

---

19 Bamberg (2006) refers to such incentives as motivational and coercive approaches.
of road accidents differs: They do not primarily defy the foundations of habitual driving behavior, but rather challenge the whole perception of operating a vehicle, especially the confidence in one’s own driving skills. Mayou et al. (1993) report that in the early weeks and months after an accident, a general lack of confidence was present in 38% of accident-involved drivers. Similarly, 22% of road crash victims reported a decrease in driving. Such behavioral consequences can also be inferred from the Theory of Planned Behavior (Ajzen, 1991); accident exposure affects control beliefs and the perceived control over the driving process therefore becomes weakened. Car use is curbed, eventually leading to fewer trips and lower mileage. This allows me to formulate the following research hypotheses:

**Hypothesis 1a:** Motorists’ number of trips decreases after a car accident.

**Hypothesis 1b:** Motorists’ mileage decreases after a car accident.

**Driving behavior**

Driving behavior describes the performance of motorists on the road and focuses on *how* people drive rather than on *where* and *when* they travel. Its analysis intends to identify accident-causing risk factors of driving. Substantial experimental work has been done on the identification of individual differences in basic driving capabilities. Its intention has been to predict accident involvement and identify drivers with above-average accident risks (Ranney, 1994). Criteria for evaluating driving performance are multifaceted, including lane and distance keeping performance (Evans & Wasielewski, 1982), steering performance (Hack et al., 2001), visual behavior (Strayer et al., 2003), and speed control (Reed & Green, 1999). Effects that might impair driving performance range from substance abuse (Ramaekers et al., 2000) and cell phone use (Brookhuis et al., 1991) to sleepiness (Lyznicki et al., 2005) and driving with multiple passengers (Chen et al., 2000).

In addition to travel behavior, accident exposure might also affect one’s driving style. The general lack of driving confidence after an accident makes motorists take some extra safety precautions while operating a vehicle, for example entering intersections more cautiously, double-checking mirrors, or driving at lower speeds. Mayou et al. (1991) support this argumentation: they show that accident-involved people regard themselves as safer drivers thereafter, thus driving more slowly and considerably. However, the focus on GPS data confines the driving behavior parameters to be analyzed in the context of this dissertation, as information on in-vehicle activities (e.g., mirror-checking) and road / car following performance (i.e., lane keeping and distance keeping) is not available. Therefore, I limit myself to
analyzing speed changes as a consequence of accident involvement and look at variations in average driving speeds in particular. The corresponding research hypothesis is:

*Hypothesis 1c: Motorists’ average driving speed decreases after a car accident.*

### 4.2 The Persistency of Effects on Travel and Driving Behavior

In this subsection, I develop the argument that the impact of accident involvement on travel and driving behavior will abate over time, as motorists regain driving confidence. With the passage of time, monthly trips, mileage, and average speeds therefore should align with pre-accident figures.

**Travel behavior**

Accident experience frequently causes post-traumatic stress disorder in motorists. Its symptoms include emotional arousal, anxiety, and intrusive thoughts of the trauma. They are a strong antecedent to changes in vehicle use. As with most post-traumatic syndromes, however, the symptoms tend to improve over time. Mayou et al. (1993) report from a follow-up study on road accident victims in which they analyze the development of psychiatric syndromes and changes in vehicle use. They study car drivers, motorcycle riders, and whiplash injury victims directly after the accident and follow up twice after three and twelve months. Their results show that self-reported anxiety and depression scores improve over time. While directly after the accident 76% of study participants had intrusive thoughts about the accident, this figure decreases to 25% (24%) after 3 (12) months. Car occupants exhibit travel anxiety in 24% of the cases 3 months after the accident, a figure that decreases to 12% at the 12 month follow-up. When it comes to concerns about driving, motorcycle riders display more post-accident distress than do car drivers. Their share of reporting “major” concerns was 76% (55%) after 3 (12) months compared to only 19% (11%) for car drivers.

From this research, I infer that similar accident consequences should become discernible when reviewing GPS-based travel data. The more the cognitive impairment decreases as motorists recapture their perceived control over the driving task, the more trips and mileage they should exhibit over time. For travel behavior, I thus formulate the following research hypotheses:
Hypothesis 2a: The more time has passed after an accident, the more its impact on the number of trips deteriorates.

Hypothesis 2b: The more time has passed after an accident, the more its impact on monthly mileage deteriorates.

Driving behavior
Similarly, I argue that driving style, while usually being more defensive directly after an accident (cf. Mayou et al., 1991), will approximate pre-accident behavior over time. The corresponding research hypothesis is:

Hypothesis 2c: The more time has passed after an accident, the more its impact on the average driving speed deteriorates.

4.3 The Impact on Male, Female, and Company Car Drivers
In the next two subsections, I develop arguments for group differences in accident response. In this subsection, I begin with outlining why collision involvement may cause different reactions in male, female, and company car drivers. Subsection 4.4 outlines a similar rationale for high- and low-mileage drivers.

Travel behavior
Socio-demographic variables are a strong predictor of inter- and intrapersonal variability in daily travel routines and are a valuable input to understand a person’s travel demand (Lu & Pas, 1999). The latter is widely recognized as a derived demand, caused by a need or desire to participate in activities at spatially remote locations (Pas & Koppelman, 1987). These needs often articulate an individual’s socio-demographic context and reflect one’s gender and age, marital status and family size, employment situation, residential location, car ownership, and other variables (Hanson & Huff, 1986).

Substantial research exists that studies gender differences in travel demand. Not only do men travel more, they also do trips for other purposes in many life phases (Collins & Tisdell, 2002). Gustafson (2006) explains this with the dissimilar roles of men and women in working life and differences in family obligations, which make men engage in more work-related travel while women take more child- and shopping-related trips. Following the author, the presence of young children even reduces women’s overall travel activity, which does not consistently hold true for men. Also trip length varies by gender; women in general have shorter work trips
than men, regardless of socio-demographic factors like income, occupation, and family status (Gordon et al., 1989).

Differences in travel demand also exist between ordinary and company car drivers. The latter comprises a wide range of motorists, from “senior executives provided with a company car as a perk of the job, through those who drive non-liveried company owned vehicles both for work and non-work purposes, to those employed to drive fleet cars, vans and other specialist vehicles” (Chapman et al., 2001). Generally speaking, company car drivers tend to feature higher mileage and more trips than ordinary motorists (Linn & Lockwood, 1998). They thus exhibit higher risk exposure, which often makes the general public perceive company cars drivers as a road safety problem (Chapman et al., 2001).

As general travel demand varies for men and women and for ordinary and company car drivers, it seems reasonable to assume such differences in response to road accidents as well. The rationale for gender differences is supported by work from Matthies et al. (2002), who show that women are generally more willing than men to reduce road travel and choose transportation alternatives. Causes of these differences are the higher ecological norms of women and weaker car use habits. When comparing company car and standard drivers, I assume similar differences in car dependency, as time pressure, the remote location of business clients, and transportation requirements may allow no alternatives to car use for business-related travel. Consequently, I argue that the three subgroups under consideration differ in their response to road accident experience in terms of travel behavior. This rationale is expressed in the following research hypotheses:

Hypothesis 3a: Male, female, and company car drivers respond differently to accident involvement with respect to the number of trips made.

Hypothesis 3b: Male, female, and company car drivers respond differently to accident involvement with respect to mileage.

20 Although no information on the vehicles makes of company cars is available within my study, data on average speeds and mileage makes it reasonable to infer that these predominantly consist of ordinary passenger cars rather than trucks, transporters, and slower specialized vehicles.
Driving behavior
Gender differences in driving behavior are evident from personal everyday experience and have been studied by researchers in great detail. Harré et al. (1996) find that men are generally more prone to risky car use behaviors like driving under the influence, speeding, and breaking curfews (as for example for teenage drivers in New Zealand). Men also tend to perform fewer safety-related activities while operating a vehicle, for example using seats belts and observing speed limits less often than do female motorists (Shinar et al., 2001). DeJoy (1992) explains these behavioral differences by concluding that men are generally more optimistic when judging their own driving skills and that they perceive risky driving as generally less likely to cause accidents than women. The higher risk propensity of male drivers is reflected in accident rates: for every mile driven, their risk of being involved in a fatal accident is 55% higher than that of women (Massie et al., 1997).

In addition to men, also company car drivers are presumed to exhibit more risky driving styles (Stradling, 2000). They tend to be less risk averse, especially when it comes to speed choice and decisions about overtaking. Adams-Guppy & Guppy (1995) explain this by strong time demands, where being on time for business appointments is more desirable than obeying speed limits. Also, caring less about fuel consumption and vehicle wear might encourage drivers of employer owned vehicles to travel at higher speeds.

Therefore, as with travel impact, I assume group differences between men, women, and company car drivers in accident response with regard to driving behavior. Women’s driving confidence, which is usually lower than that of men, may be impaired more by accident exposure. Likewise, I expect fewer changes in the driving style of company car motorists due to the obligations and the frequent time pressure they face. I postulate that these differences will become discernible when looking at motorists’ average speeds and thus formulate as a research hypothesis:

*Hypothesis 3c: Male, female, and company car drivers respond differentially to accident involvement with respect to average driving speed.*

4.4 The Impact on Low- and High-Mileage Drivers

Another aspect that may cause group differences in accident response is annual mileage. In this subsection, I develop the argument that accidents cause different reactions in the travel and driving styles of low- and high-mileage travelers.
Travel behavior
Many modern societies today are highly car-dependent. The convenience of getting directly to a desired location and the frequent lack of appropriate public transport alternatives make personal vehicles the mode of choice for many travel activities. Evidence from both the United Kingdom (Cullinane, 1992) and Hong Kong (Cullinane & Cullinane, 2003) suggests that motorists’ views on car necessity and their annual mileage are positively related. Survey data from Hong Kong illustrates that 46% of those motorists who travel more than 8000 miles a year consider their car as completely necessary, compared to 0% of those who drive less than 1000 annual miles. These findings are in line with Begg (1998), Banister (2001), and Dargay (2001), who argue that once a car has been purchased, it turns from a luxury into a necessity, as the owner eventually becomes more and more reliant on it.

The more an individual is dependent on a personal vehicle, however, the more difficult it is to change once established travel behaviors. Fujii et al. (2001), in their investigation of travel mode change during temporary road closures, find that “the frequency of switching to public transport during the closure [is] inversely related to the frequency of automobile commuting before the closure.” Such effects might be attributed to habitual travel choice (Gärling & Axhausen, 2003) or to reservations about transportation alternatives, whose commuting times are often overestimated by high mileage drivers (Fujii et al., 2001).

As the research by Fujii et al. (2001) shows, the impact of interventional effects on road travel varies by car dependency and therefore by mileage driven. This leads me to assume similar differences when investigating accidents as the intervention cause. I incorporate this rationale into the following research hypotheses:

Hypothesis 4a: Low- and high-mileage drivers respond differently to accident involvement with respect to the number of trips made.

Hypothesis 4b: Low- and high-mileage drivers respond differently to accident involvement with respect to mileage.

Driving behavior
As travel behavior varies with mileage, so does driving behavior. In a study on driving aggressiveness, Krahé (2005) surveys female motorists on their engagement in driving behaviors like overtaking on the inside, getting angry at other motorists, and disregarding the speed limit. Her findings suggest that aggressive driving is positively related to kilometers driven per year. The increases of speeding and other
risky behavior in mileage can be explained by a strengthened driving confidence that occurs as a result of higher vehicle use (Marottoli & Richardson, 1998). As the Theory of Planned Behavior would predict, positive behavioral outcomes might strengthen the perceived control over an activity and may thus trigger increased future engagement in that activity. Following this logic, it can be assumed that accident involvement affects the perceived behavioral control and driving self-confidence of high- and low mileage drivers differently: while the former might change their driving style only temporarily, effects might be more pronounced and longer lasting for less confident motorists. Consequently, I argue that the two driver classes respond differently to accident involvement, which leads me to the formulation of the following research hypothesis:

\[ \text{Hypothesis 4c: Low- and high-mileage drivers respond differently to accident involvement with respect to average driving speed.} \]

4.5 The Impact of Accident Severity

Travel behavior

Road accidents can be classified by different parameters. Besides the ones that consider the insurance perspective (driver at-fault and monetary damage), these include accident cause (e.g., speeding, driving under the influence, insufficient safety distance, right of way violations, and turning errors), mechanism (i.e., rear-end and head-on accidents, run-off-road accidents, side collisions, and rollovers), or accident severity. From a technical perspective, the latter is measured by a vehicle’s change in velocity during the collision, commonly labeled Delta-v (Gabler et al., 2004). It denotes the time integral of acceleration during the actual impact, which normally lasts 50 to 150 milliseconds. From a medical point of view, accident severity is assessed by the Injury Severity Score (cf. Baker et al., 1974), an established anatomical scoring system for patients with poly-traumatic injuries. It rates injuries on a scale ranging from 1 (minor) to 6 (unsurvivable) and allocates them to one of six body regions (namely, head and neck, face, chest, abdomen, extremities, external). In their investigation of car occupants with thoracic injury,

\[ \text{Note that (maximum) vehicle acceleration cannot characterize the occupant injury potential of a vehicle collision. In daily life, high acceleration values over very short times regularly occur without causing bodily injury.} \]
Richter et al. (2001) use both Delta-v and the Injury Severity Score to demonstrate the positive correlation of both measures. Similarly, Nance et al. (2006) show the high level of association between both accident parameters in their study on children involved in head-on vehicle crashes.

Mayou et al. (1991) go one step further and relate the Injury Severity Score to the loss of driver confidence. Their findings support the argument that both figures are positively associated. As the authors show, lacking confidence expresses itself in a general driving impairment, nervousness in situations similar to the one that caused the accident, arousal at the place of accident, and avoidance of the accident location. As the rate and probability of performing a task decline in behavioral control (i.e., confidence), I postulate that the degree of post-accident behavior changes rises in accident severity. The more severe a vehicle crash is physically, the more pronounced the decline in trips and mileage thus should be. The corresponding research hypotheses are:

- **Hypothesis 5a:** The more severe an accident is, the more the number of trips decreases.
- **Hypothesis 5b:** The more severe an accident is, the more monthly mileage decreases.

**Driving behavior**

Similar to travel behavior, I assume that driving impairment varies with accident severity. Following the argumentation in subsection 4.1, I state that if vehicle rides cannot be avoided, motorists at least try to apply increased precautions while driving. The more severe an accident has been, the higher the motorist’s need to engage in defensive driving. In the context of this dissertation, this translates into decreased average speeds for each trip and leads me to proposing the following research hypothesis:

- **Hypothesis 5c:** The more severe an accident is, the more the average speed decreases.

**4.6 The Impact on Trips of Different Purpose**

Subsection 4.4 has shown that people can become dependent on cars for their daily travel routines. Goodwin et al. (1995) extend this notion of car-dependent motorists by arguing that car dependency predominantly varies with trip purpose. Data from
the United Kingdom supports their rationale, where 90% of the surveyed motorists use their car for grocery shopping, 89% for visiting friends and family, but only 58% for commuting to work (Lex Services, 1995). Similar data is available from Hong Kong, where 80% of surveyed motorists report always using their car for trips to the countryside, followed by 68% (66%) for chauffeuring children or other family members. Only 52% use their car regularly for commuting to work (Cullinane & Cullinane, 2003). The unavailability of transportation alternatives to certain locations, a lack of parking opportunities, and the necessity to carry and transport things may explain this discrepancy, among others.

Different trip purpose classifications exist in literature which in part reflect differences in the available base data (survey data vs. GPS records from data loggers). Liao et al. (2005) use a rather basic scheme of trip purposes, including commutes, shopping trips, trips for dining, visits, home-bound trips, and others. The scheme used by Axhausen et al. (2003) applies more distinctions; in addition, they differentiate between daily and long term shopping trips, work and work related trips, and define trips for picking up and dropping off. A similar classification scheme is proposed by Wolf (2000). Cullinane & Cullinane (2003) finally use a very detailed taxonomy. It reflects the survey data they use, allowing, for example, differentiation between dining, cinema, and sport trips, a distinction which is not applicable when analyzing GPS data.

Motorists’ response to accident involvement should reflect the manner in which car dependency varies by trip type. Rides that can be avoided (e.g., leisure-related trips) thus should exhibit higher relative reductions, while unavoidable trips (e.g., daily work commutes) are affected less or not at all. As trip frequencies already differ considerably irrespective of accident exposure, it is imperative to consider relative accident-induced trip reductions in the analysis. As I assume journey length and travel speed not to be dependent on trip purpose, I limit myself to the evaluation of interventional changes in trip frequencies. I formulate the corresponding research hypothesis as follows:

Hypothesis 6: The relative trip reduction after an accident varies by trip purpose.
4.7 Situational and Spatial Avoidance

Situational avoidance
As I showed in subsection 4.2, crash involvement frequently causes post-traumatic syndromes in accident victims. In a follow-up study, Mayou et al. (1993) report that intrusive, “horrific” thoughts are present in 30% of motorists directly after the accident experience. After one year, 6% of crash-involved car occupants still suffer from such post-traumatic distress. Nervousness and emotional arousal in situations similar to that of the accident are even more pronounced and longer lasting: as the authors indicate in their 4-6 year follow-up study, nervousness in crash-like situations in still present in 34% of the driver population (Mayou et al., 1991). Consequently, motorists try to avoid corresponding driving conditions, for example by shunning nighttime rides, pulling over to the side in heavy downpours or when getting sleepy, or stopping car use and switching to other modes of transportation during the winter season. With the data at hand, the identification of similar situations is limited, however, and makes me confine the study of situational avoidance to travel hours. They are closely related to other possible accident causes, for example poor lighting conditions or traffic density. For this dissertation, I assume situational avoidance if crash-exposed motorists reduce trips at the accident time to a higher degree than other rides. The corresponding research hypothesis thus is:

Hypothesis 7a: The relative trip reduction is higher for trips that take place at the accident time than for all other trips.

Spatial avoidance
As with situational avoidance, motorists might also try to keep away from the accident location. It tends to create emotional arousal in road accident victims who recall the trauma they were exposed to at that particular location. Mayou et al. (1991) show that 19% of crash-exposed vehicle drivers still get nervous at the place of accident 4-6 years after involvement, with 5% of motorists eventually trying to avoid that particular location. When focusing on the immediate post-accident period, such emotional distress and behavioral consequences should become even more prominent, as accident victims have had only little time for convalescence. Like before, I presume spatial avoidance in motorists if they reduce trips that pass by the accident location to a larger extent than others. The research hypothesis is:

Hypothesis 7b: The relative trip reduction is higher for trips that pass by near the accident location than for all other trips.
5. Data and Methodology

This section gives detail on the data and methodological steps I use for testing the developed research hypotheses developed in section 4. I begin with naming the experimental setting in subsection 5.1, categorizing this research as a natural experiment. Subsection 5.2 continues with a description of the unit of analysis. It provides detail on the source of the GPS data and briefly lists corresponding company and customer information. Next, I illustrate the used data in subsection 5.3. A thorough methodological outline follows in subsection 5.4. It explains the performed data preprocessing and aggregation steps, applied cluster and classification algorithms, and the regression analysis design. Subsection 5.5 names the statistical software used.

5.1 Experimental Setting

In this dissertation, I take travel and driving data from real world entities to analyze the impact of accident involvement on these measures. Such a setting, where “the experimenter simply observes naturally occurring, controlled comparisons of one or more treatments with a baseline” (Harrison & List, 2004) is called a natural experiment. Its major advantage is that individuals pursue their activities in a natural setting, without knowing that their actions are being followed for research purposes, and in which the consequences of their behavior are real and typically substantial. A drawback of natural experiments is the lack of control the researcher has over its procedural course; he cannot determine the individuals that receive the treatment, which kind of intervention (i.e., the accident characteristics) they are exposed to, nor where and when they receive the treatment. In general, natural experiments tend to give up internal validity (i.e., the validity of causal inferences) in exchange for higher external validity (i.e., the generalizability of causal inferences) of the research results. They are commonly used in areas where artificial experimenting is problematic, such as epidemiology, sociology, economics, and cosmology (Shadish et al., 2002).

5.2 Unit of Analysis

The clients of a single Italian insurance company covered by a telematics-based vehicle insurance product form the unit of analysis. The insurer is active both in the life and non-life market, has offered telematics-based vehicle insurance to its clients
since 2005 and has issued more than 400,000 of these policies to date (as of April 2010). The insurer’s telematics product is available only in Italy; thus the data exclusively represents Italian motorists. This allows controlling for national heterogeneity in terms of speed limits, urbanization and regional spread, and the availability of public transport alternatives, which eventually influence travel and driving behavior. Furthermore, the pursued single-firm approach eliminates possible inter-company differences like variance in applied pricing schemes (standard vs. pay-per-mile schemes) or the structure of the customer risk portfolio. While sourcing from just one insurance company reflects to some extent the scarcity of data in both required quality and quantity for the purpose of this research, it is nonetheless in line with other successful contributions in the area of GPS data analysis (e.g., Schuessler & Axhausen, 2009).

For privacy reasons, the telematics data is hosted not by the insurer but externally by Octo Telematics. The Italy-based firm is the global market leader in the research, development, and management of cutting-edge telematics applications for auto insurance. With over 1,000,000 active customers (as of April 2010) and the world’s largest telematics database, its solutions enable mileage-based premiums, automated emergency calls (eCall), stolen vehicle recovery, and fleet management services across Europe and around the world. Octo Telematics manages the provision of telematics devices, their installation at authorized body shops, and the data management for insurance companies. It is a wholly owned subsidiary of the Metasystem Group (Italy), from which Octo sources its telematics hardware.

For the purpose of this research, Octo has provided both trip and accident data. It is collected by telematics boxes that are discretely fitted into the clients’ cars together with GPS and GPRS antennas and connected to the vehicles’ CAN-bus. When a vehicle is operated, the system records the start and end location of every trip and the vehicle’s current position approximately every 2 kilometers. While the engine is turned off, the telematics unit regularly stores the current position and thus enables stolen vehicle tracking if needed. Aside from the position records, the system continuously writes and overwrites longitudinal and lateral acceleration data during driving in one millisecond granularity. In case of an accident, the device freezes the data from 30 seconds before to 15 seconds after the time of the first

---

22 Deviations from this standard value may occur in case of GPS signal loss.
5. Data and Methodology

The GPS data is transferred to Octo’s MultiService Center in regular intervals once the internal memory of the device is full. Accident data is transmitted upon incident occurrence. All information is preprocessed and stored at Octo’s data center, from where it can then be made available to stakeholders like insurers, car makers, public authorities, end consumers, and research institutes. Figure 11 summarizes Octo’s telematics system graphically.

As this research uses highly confidential and sensitive travel and driving information, some measures are necessary to maintain the clients’ privacy. First, I present raw data and research results only on an aggregate level. Where single driver records are shown for illustration purposes, I cipher location information by adding an arbitrary summand to both latitudinal and longitudinal position records. Second, Octo Telematics performed a cross-check of this dissertation and its results to ensure that privacy standards were met. No concerns have been raised on this issue, Octo Telematics approved and authorized the publication of this dissertation.

![Diagram of Octo Telematics telematics systems and services](image)

*Figure 11: The telematics systems and services of Octo Telematics*

---

23 Data storage takes place if acceleration values exceed some preset threshold.
5.3 Data

Octo Telematics provided GPS-based travel data and accelerometer-based crash information for the purpose of this dissertation. All data comes in CSV File Format, which makes it possible to load the data conveniently into various analytical software packages without noticeable preprocessing. A single CarID representing the unique identifier of a vehicle’s black box allows matching travel and accident information. I describe this data in the following subsections.

5.3.1 Travel Data

Travel information is at the core of analysis in this research. It is derived from the GPS signal and represents vehicles’ geographical position during driving. The ignition status is retrieved from the vehicle’s internal electronic network, the CAN-bus, to which the device is connected. The road type is imputed later on after the data has been transmitted to Octo’s backend system. Table 2 describes attributes and format of the travel data provided:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CarID</td>
<td>Car / telematics device ID</td>
</tr>
<tr>
<td>Date</td>
<td>Date on which the dataset was recorded (dd-mm-yyyy)</td>
</tr>
<tr>
<td>Time</td>
<td>Time on which the dataset was recorded (hh:mm:ss)</td>
</tr>
<tr>
<td>Latitude</td>
<td>Latitudinal vehicle position in decimal notation (xxx.dddddddd)</td>
</tr>
<tr>
<td>Longitude</td>
<td>Longitudinal vehicle position in decimal notation (xxx.dddddddd)</td>
</tr>
<tr>
<td>Heading</td>
<td>Vehicle heading at recording time in ° (xxx)</td>
</tr>
<tr>
<td>Speed</td>
<td>Vehicle speed at recording time in km/h (xxx)</td>
</tr>
<tr>
<td>Distance to previous point</td>
<td>Distance traveled since last recording point (usually about 2km); allocated device memory (16 Bit) allows for a maximum recorded distance of 65535m per dataset</td>
</tr>
<tr>
<td>Time since previous point</td>
<td>Time traveled since last recording point</td>
</tr>
<tr>
<td>Panel session</td>
<td>Provides dataset description / dataset purpose:</td>
</tr>
<tr>
<td></td>
<td>0 = Dataset recorded on ignition turn-on</td>
</tr>
<tr>
<td></td>
<td>1 = Dataset recorded during vehicle operation</td>
</tr>
<tr>
<td></td>
<td>2 = Dataset recorded on ignition turn-off</td>
</tr>
<tr>
<td></td>
<td>3 = Dataset recorded in regular intervals while engine is turned off</td>
</tr>
</tbody>
</table>
Table 2: Travel data attributes

<table>
<thead>
<tr>
<th>Road type</th>
<th>Road type at recorded location, imputed by Octo Telematics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Urban</td>
</tr>
<tr>
<td>1</td>
<td>Highway</td>
</tr>
<tr>
<td>2</td>
<td>Extra-urban</td>
</tr>
<tr>
<td>3</td>
<td>Other (not present in the data provided)</td>
</tr>
<tr>
<td>4</td>
<td>Outside of Italy (value is transformed into 3)</td>
</tr>
<tr>
<td>5</td>
<td>Extra-urban in low urban density region (value is transformed into 2)</td>
</tr>
<tr>
<td>6</td>
<td>Urban in low urban density region (value is transformed into 0)</td>
</tr>
<tr>
<td>7</td>
<td>Extra-urban in high urban density region (value is transformed into 2)</td>
</tr>
<tr>
<td>8</td>
<td>Urban in high urban density region (value is transformed into 0)</td>
</tr>
<tr>
<td>-3</td>
<td>Not available</td>
</tr>
</tbody>
</table>

5.3.2 Accident Data

Accident information is derived from acceleration sensors that continuously write in the memory of the telematics unit. It constantly holds accelerometer data from the last 45 seconds of driving and thus continuously overrides older datasets. In case of extraordinary acceleration values that exceed some preset threshold, the telematics box freezes the data from 30 seconds before to 15 seconds after the corresponding incident. The minimum triggering level has to be set carefully and must distinguish accidents from normal vehicle operation. Generally, normal vehicle use can result in longitudinal acceleration / deceleration of up to 0.7g / -1g. Corresponding accident thresholds are thus set slightly above those measures. According to Octo Telematics, they may nonetheless vary by country to some degree for differences in road and pavement quality.

Table 3 shows the available accident data. For the purpose of this research, Octo Telematics preprocessed the raw accident data and provided only aggregated crash

---

24 I infer the acceleration value from supposing a sprint from 0-100km/h in 4s, which is possible for high performance vehicles. Likewise, the deceleration value is derived by assuming a braking distance of 35m when bringing a vehicle traveling at 100km/h to a full stop (1g = 9.81 m/s²).
information. This includes accident date and time along with maximum and average acceleration values; the full body of collision sensor data (which is primarily of use for accident reconstruction) was not supplied. Beyond that, each accident dataset holds information on the registered car owner as imputed by Octo Telematics, which enables the analysis of group differences in accident response (cf. subsections 4.3 and 6.2.3).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CarID</td>
<td>Car / telematics device ID</td>
</tr>
<tr>
<td>Date</td>
<td>Accident date (dd-mm-yyyy)</td>
</tr>
<tr>
<td>Time</td>
<td>Accident time (hh:mm:ss)</td>
</tr>
<tr>
<td>Client Session</td>
<td>Information on registered car owner</td>
</tr>
<tr>
<td></td>
<td>S = Company car (transformed into 1)</td>
</tr>
<tr>
<td></td>
<td>M = Male car owner (transformed into 2)</td>
</tr>
<tr>
<td></td>
<td>F = Female car owner (transformed into 3)</td>
</tr>
<tr>
<td>Acceleration</td>
<td>Resulting vector of the maximum accelerations in X, Y, and Z directions</td>
</tr>
<tr>
<td>Max. Acceleration Y</td>
<td>Maximum lateral acceleration during the accident (in g; 1g = 9.81 m/s²)</td>
</tr>
<tr>
<td>Max. Acceleration X</td>
<td>Maximum longitudinal acceleration during the accident (in g)</td>
</tr>
</tbody>
</table>

*Table 3: Accident data attributes*

5.4 Research Methodology

In the following, I give a detailed outline of the research methodology that I use within this dissertation. Figure 12 provides an overview. I begin by outlining the applied sampling strategy, which intends to eliminate seasonal effects. Next, I give a stepwise description of how the raw GPS data is transformed into single trips, information on travel and driving behavior, and a person’s significant locations. Additionally, I detail the performed filtering steps that are necessary to ensure data integrity and to avoid the inflation of the results by pseudo-effects. I conclude by showing the applied regression and testing procedures.
5.4.1 Sampling

Octo Telematics’s database currently contains GPS information on more than 1,000,000 vehicles (as of April 2010), which equals data on several billion kilometers of driving every year. While the analysis of this amount of data would exceed the computational capabilities available for this dissertation, it is actually not necessary to deal with this amount of data under significance aspects. By a sound sampling strategy, I expect to reduce the number of analyzed drivers considerably while maintaining the statistical power of my research.
Sample size
I perform an introductory power analysis (cf. Cohen, 1977) using GPower 3.1.0 to specify the required sample size. In order to determine whether accident involvement influences travel and driving behavior, I test the significance of a single dichotomous regression coefficient that indicates case / control group membership. The power analysis uses the supposed effect size \( f^2 \), the error probability \( \alpha \), and the power \( 1 - \beta \) as input. \( \alpha \) describes the chance of a Type I error (i.e., accepting a hypothesis \( H_1 \) although no effect is existent) and \( \beta \) represents the chance of a Type II error (i.e., accepting a hypothesis \( H_0 \) even though an effect exists); therefore \( 1 - \beta \) is the power of a test for identifying an effect that is existent.

For the power analysis, I take a conservative approach and intend to achieve significance even for small effect sizes. Following Cohen (1977), I thus set \( f^2 = 0.02 \). The significance will be tested on a 95% confidence level, so \( \alpha = 0.05 \). Following Hinkle et al. (2003), who argue for the ratio of \( \beta \) to \( \alpha \) being 4:1, \( 1 - \beta \) is set to 0.8. For these parameters, GPower 3.1.0 yields a sample size of 311 drivers to identify possible effects on single regression coefficients with sufficient statistical power (one-tailed test). In fact, it was possible to obtain a larger sample from Octo Telematics, which enables the further analysis of sample subgroups.

Controlling seasonality
Travel and driving behavior vary by season, with motorists reducing travel and adjusting speeds during the winter months (cf. subsection 3.2.5). When analyzing the accident impact on both travel and driving, one must account for such seasonal fluctuations. I do this by applying a controlled sampling strategy that allows for seasonal matching of pre- and post-accident data (cf. Barnett & Dobson, 2010).

Figure 13 illustrates this concept. On the left hand side, all sampled motorists are exposed to an accident at the same time of the season. Differences between pre- and

---

25 GPower 3.1.0 is available online from the University of Düsseldorf at: [http://www.psycho.uni-duesseldorf.de/aap/projects/gpower](http://www.psycho.uni-duesseldorf.de/aap/projects/gpower)

26 Seasonal matching is a common approach to control for seasonality in epidemiological studies when treatments or interventions cannot be administered simultaneously. For its application in the health domain, I name work by Rochester et al. (2009) as an example, who study the recovery of the luteal function of obese women after bariatric surgery.
post-accident values therefore primarily reflect seasonality. On the right hand side, motorists are sampled in a way that distributes accidents evenly throughout the year. When aggregating pre- and post-accident scores, seasonality effects are assumed to cancel out. Remaining differences between pre- and post-accident travel therefore reflect only the effect of the intervention. A further inclusion of seasonal dummies in the regression analyses is not necessary.

Upon this consideration, Octo Telematics sampled accident drivers in a way that enables seasonal matching. Their accident dates are spread evenly throughout the year, with the same amount of collisions occurring in each calendar month. Although this does not resemble the true seasonal accident distribution, with more crashes happening in winter months, this can be ignored as the analysis of accident causes (like weather or lighting conditions) is not within the scope of this research.

Seasonal matching must also be applied to non-accident drivers. As these drivers do not experience accidents in the study period, matching is done by assigning each motorist a random reference month (1-12) that determines the pre and post values to be considered. Consequently, data as shown in Figure 14 is used for every driver.
5.4 Research Methodology

Sample characteristics
For this dissertation, I have requested data on 1600 drivers from Octo Telematics, comprising case drivers with an accident in the study period and controls. Cases consist of 600 randomly selected vehicles from Italy that were involved in an accident in 2008, with 50 having had an accident in each calendar month. 13 months of data is available on each accident driver, that is, the six months that preceded the vehicle crash month, the accident month itself, and the six months after the accident. For motorists that were involved in a crash in January 2008 for example, data from July 2007 to July 2008 is available. Study controls consist of 1000 randomly selected vehicles which had not had an accident in 2008. For them, GPS information from July 2007 to June 2009 is available.

5.4.2 Data Preprocessing
Data preprocessing constitutes an intermediary step prior to aggregating the GPS raw data to individual trips. I start by eliminating datasets whose ignition status is neither 0, 1, nor 2. In doing so, I discard erroneous data sets, control and system records, as well as regular position updates while the vehicle is not being operated. Next, I adjust the road type information as stated in Table 2. Next follows the elimination of incomplete trips, that is, trips that lack a dataset with ignition status 0 (indicating departure) or 2 (indicating arrival). Consultation with Octo Telematics revealed that this happens if the GPS signal is not available upon trip start or end, for example when parking underground. In addition, it usually takes GPS devices a short amount of time to compute the current position, which might also explain
some of the missing trip start datasets (cf. Stopher et al., 2005). In the final step, I bring all data into a standardized format and sort it by CarID, date, and time in ascending order, which is a prerequisite for trip aggregation.

As a next step in data preprocessing, literature recommends interposing data smoothing activities (cf. Jun et al., 2006; Schuessler & Axhausen, 2009). Their intent is to augment or reduce erratic data by changing the value of input variables (Hastie et al., 2001). Although most GPS receivers already use proprietary filtering techniques to compensate for data that exceeds known variances (Ogle et al., 2002), additional smoothing measures are often inevitable as multipath signal reflections (also known as urban canyon errors) or signal blockage might compromise the GPS signal (Zito et al., 1995). Nevertheless, I have decided not to perform data smoothing activities in this research. I argue that the data of this dissertation is different from that available to other researchers (e.g., that of Schuessler & Axhausen, 2009); it already contains trip start and end information, whereas this had to be derived in advance from the continuous GPS data streams in other research. In those circumstances, inaccuracies in the GPS signal make further analyses especially cumbersome if preparatory data smoothing is missing. As I do not applied any other plausibility checks due to the amount of data, I directly feed the preprocessed information into trip aggregation.27

5.4.3 Trip Aggregation

Next, I deduce individual trips from the GPS data; the ignition status information indicates which datasets have to be tied together. Although the availability of distinct departure / arrival information makes trip identification straightforward, I consider additional information within this research to further improve the aggregation process. A review of relevant literature yields several measures to identify trip ends; the most common are dwell times for which drivers or pedestrians remain at the same location. (Pearson, 2001; Wolf et al., 2001; Schönfelder & Samaga, 2003). Other heuristics consider heading changes or the distance between the GPS position and the road network (Stopher et al., 2003; Du &

27 Other plausibility checks may include the comparison of the Euclidean distance of two consecutive data points to the recorded distance between two locations or the identification of erroneous road type imputations, for example at intersections of highways and bridged arterial roads.
Aultman-Hall, 2007). With respect to the used data, I follow in part the comprehensive approach of Axhausen et al. (2003) to identify possible trip ends; they designate positions where the engine is turned off as possible trip ends unless the length of stay undershoots some lower threshold deemed too short for a purported stop. In doing so, they intend to rule out short, unintended stops, for example at traffic lights, railroad crossings, or at engine stall. Axhausen’s research group sets the corresponding lower boundary to 5 seconds. I use 20 seconds instead, as I assume not only stops of up to 5 seconds as too short to pursue any meaningful task at the trip end location but also those of up to 20 seconds. Second, the avoidance of other arrival checks as performed by the authors (e.g., circuitry tests) justifies my more conservative approach of using a higher threshold.

Once trip ends are identified by the ignition status and the lower boundary condition, the aggregation of the GPS data to trips is straightforward. For each complete trip, the following attributes become available:

- TripID and CarID
- Date, time, and position of trip start
- Date, time, and position of trip end
- Distance and time traveled per road type
- Length of stay at end location
- Variables indicating weekday and hour of arrival

Table 4 gives an example of how two trips are derived from the available data. Note that datasets with ignition status 3 are not considered. The first trip starts at

Axhausen et al. (2003) also propose an upper threshold to identify clear trip ends, a step which I omit in my research due to a lack of sufficient data granularity. They calculate the time difference between two consecutive datasets and infer a trip end if a preset threshold is exceeded. An inspection of the available data shows that the time gap of two successive position records varies considerably, however; as consecutive data points are recorded only every 2 km, corresponding time intervals may range from 60 seconds (when driving on highways) to several minutes (when doing inner-city rides or being caught in a traffic jam). This makes it challenging to specify a reasonable upper threshold that reliably identifies additional trip ends.

Note that I do not show the attributes CarID, current speed, and current heading due to space considerations.
position 44.904975° North, 8.874011° East at 12:21:08. The motorist travels 5.414km on urban roads, and arrives at 44.939894° North, 8.855678° East at 12:32:06. The trip duration is 10 minutes 58 seconds, and the driver remains at that location for 1 hour, 19 minutes, and 2 seconds.\textsuperscript{30} The second trip can be calculated similarly.

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Latitudinal position</th>
<th>Longitudinal position</th>
<th>Distance to previous point</th>
<th>Time since ignition</th>
<th>Ignition status</th>
<th>Road type</th>
</tr>
</thead>
<tbody>
<tr>
<td>04.07.2007</td>
<td>10:48:04</td>
<td>44.904975</td>
<td>8.874011</td>
<td>0</td>
<td>1845</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>11:19:19</td>
<td>44.904975</td>
<td>8.874011</td>
<td>0</td>
<td>1875</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>11:50:33</td>
<td>44.904975</td>
<td>8.874011</td>
<td>0</td>
<td>1874</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>Trip 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04.07.2007</td>
<td>12:21:08</td>
<td>44.904975</td>
<td>8.874011</td>
<td>0</td>
<td>1835</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>12:25:54</td>
<td>44.913849</td>
<td>8.868916</td>
<td>2020</td>
<td>286</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>12:29:21</td>
<td>44.929331</td>
<td>8.861536</td>
<td>2022</td>
<td>207</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>12:32:06</td>
<td>44.939894</td>
<td>8.855678</td>
<td>1372</td>
<td>165</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>13:05:32</td>
<td>44.939887</td>
<td>8.855668</td>
<td>0</td>
<td>2006</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>13:35:59</td>
<td>44.939887</td>
<td>8.855668</td>
<td>0</td>
<td>1827</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>Trip 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04.07.2007</td>
<td>13:51:08</td>
<td>44.939754</td>
<td>8.855716</td>
<td>0</td>
<td>909</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>13:55:04</td>
<td>44.924259</td>
<td>8.864946</td>
<td>2056</td>
<td>236</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>13:58:59</td>
<td>44.90976</td>
<td>8.875355</td>
<td>2005</td>
<td>235</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:03:04</td>
<td>44.8948</td>
<td>8.864625</td>
<td>2038</td>
<td>245</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:07:03</td>
<td>44.879567</td>
<td>8.852786</td>
<td>2008</td>
<td>239</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:09:35</td>
<td>44.865075</td>
<td>8.837829</td>
<td>2072</td>
<td>152</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:13:06</td>
<td>44.885746</td>
<td>8.81762</td>
<td>2007</td>
<td>211</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:15:03</td>
<td>44.861618</td>
<td>8.795908</td>
<td>2052</td>
<td>117</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:17:36</td>
<td>44.866578</td>
<td>8.773622</td>
<td>2090</td>
<td>153</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:19:47</td>
<td>44.874805</td>
<td>8.751997</td>
<td>2012</td>
<td>131</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:23:45</td>
<td>44.885151</td>
<td>8.744158</td>
<td>1784</td>
<td>238</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>14:55:06</td>
<td>44.885151</td>
<td>8.744158</td>
<td>0</td>
<td>1881</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>15:27:20</td>
<td>44.885151</td>
<td>8.744158</td>
<td>0</td>
<td>1934</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>04.07.2007</td>
<td>15:59:21</td>
<td>44.885151</td>
<td>8.744158</td>
<td>0</td>
<td>1921</td>
<td>3</td>
<td>-3</td>
</tr>
</tbody>
</table>

Table 4: Trip aggregation and filtering

5.4.4 Trip / Driver Filtering

Applied filtering steps
I continue by applying a set of trip and driver filtering steps on both accident and non-accident clients. These filtering steps are necessary because of anomalies in the trip data; they guarantee that possible accident intervention effects are neither distorted nor inflated by pseudo-effects.

\textsuperscript{30} The location information shown has been anonymized for the purpose of this research.
A first inspection of the generated trip data shows that trip information is available for 1598 drivers only. A check-up shows that although data on 600 vehicle crashes is available, there are only 598 exposed motorists, as two drivers experienced two car collisions. I completely eliminate both drivers from the dataset, as the single point of the accident is ambiguous. Next, I found missing data on a non-accident motorist, whose GPS data only consists of records with ignition status 3; all of this data is discarded during trip aggregation. The absence of any concrete travel information may be the result of a malfunctioning GPS device.

In the second step, I review the duration of the individual trips. It appears that some drivers frequently travel non-stop for a very long time, with journeys stretching over three days or more.31 A review suggests that in these cases “ignition off” signals are missing, as the recorded travel time between two consecutive datasets is several hours, while the distance is only 2 kilometers. A reason might be that motorists do not pull out their car key after the engine is shut off. Although drivers may eventually have performed intermediary stops, these cannot be detected based on ignition status information. Consequently, the results become biased, as the absence of trip stops leads to a lower number of detected trips. In addition, lacking arrival information deflates measured driving speeds by considering vehicle idle periods as driving time. My approach towards this problem is twofold. First, all trips that stretch over three or more days are deleted from the dataset. Second, I calculate the sum of days over which a car driver exhibits trips of such extraordinary duration. If the aggregate exceeds 30 days, the motorist is discarded completely. As a result of this filtering step, 14 (82) accident (non-accident) drivers are eliminated from the data. For 92 (95) other accident (non-accident) drivers, single trips of abnormal length are removed.

A third filtering step considers the post-accident travel behavior of crash-involved motorists. As it is evident from the data, 26 motorists hardly do any trips after the

---

31 This filtering step considers only trips that stretch over at least three days. While trips that stretch over two days can be considered usual driving (that is, when arriving past midnight), trips that stretch over three days take at least 24 hours (if assuming, for example, departing on May 5th at 23:59 and arriving on May 7th at 00:00). I deem this too long a trip without any observable stop. Note that my approach to identify such “long” trips considers date information only, so trips that start on May 5th 09:00am and end on May 7th 08:00 are regarded as trips of exceptional duration within my analysis.
accident, frequently showing zero monthly trips. I assume that in these cases vehicles were damaged to an extent that prohibits further operation or repairing.\textsuperscript{32} Motorists may have switched to a new car instead, which prohibits analyzing their post-accident car usage.\textsuperscript{33} For this dissertation, I exclude these 26 accident-involved drivers from the analysis; their reduction in driving must be considered not to reflect behavioral changes primarily, but rather the inoperability of their vehicles after a collision. Keeping these drivers in the analysis would impose pseudo-effects to the measured results and thus must be avoided.

In the final data filtering step, I compare the number of monthly trips to mileage. In 7 (18) accident (non-accident) drivers, discrepancies are noted for frequently showing trips but no mileage. A look at their raw GPS data discloses that the recorded lateral and longitudinal position does not change during vehicle operation, which may be caused by defective GPS receivers.\textsuperscript{34} Consequently, I discard these 25 drivers from the data basis.

After all trip and driver filtering steps have been performed, 549 accident and 899 non-accident drivers remain in the data sample. Those 1448 motorists will be the basis for all subsequent analyses. A summary of the performed filtering steps is provided in Table 5.

\textsuperscript{32} Another explanation might be that drivers were lethally injured in an accident, though corresponding information is not available.

\textsuperscript{33} Consultation with Octo Telematics revealed that telematics devices are not dismounted from wrecked vehicles after an accident. Another telematics box is fitted to the motorist’s new car instead.

\textsuperscript{34} It must be noted that in the case of short trips (e.g., when re-parking a car), driving time may not be sufficient for the GPS device to determine its current location and distances traveled. While this may also cause trips without mileage, these should remain rare events though. The frequency with which such trips happen for the 25 drivers discussed here allows no other conclusion than assuming a defect in the GPS device, though.
5.4 Research Methodology

<table>
<thead>
<tr>
<th></th>
<th>Accident drivers</th>
<th>Non-accident drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial data requested</strong></td>
<td>600</td>
<td>1000</td>
</tr>
<tr>
<td><strong>Initial data provided (two drivers exhibited two accidents in the study period)</strong></td>
<td>598</td>
<td>1000</td>
</tr>
<tr>
<td><strong>Elimination of drivers with more than one accident</strong></td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Elimination of drivers which lack processable data</strong></td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td><strong>Elimination of drivers who exhibit too many long trips (end date of trip - start date of trip &gt; 1)</strong></td>
<td>-14</td>
<td>-82</td>
</tr>
<tr>
<td><strong>Elimination of accident-drivers who do not drive anymore after an accident (totaled vehicles)</strong></td>
<td>-26</td>
<td>0</td>
</tr>
<tr>
<td><strong>Elimination of drivers who frequently exhibit trips but no mileage (defective GPS devices)</strong></td>
<td>-7</td>
<td>-18</td>
</tr>
<tr>
<td><strong>Drivers remaining for analysis</strong></td>
<td>549</td>
<td>899</td>
</tr>
</tbody>
</table>

Table 5: Summary of the driver filtering process

**Testing for filtering bias**

The filtering steps just described can only be regarded as valid if they do not impose any bias to the available data. The effects that require the elimination of drivers must appear randomly, that is, in the absence of any latent variables not present in the data. As a consequence, it is necessary to analyze whether filtered drivers are any different from the retained motorists.

First, I compare the drivers filtered for an unusual amount of long-lasting trips to the 1448 motorists remaining. I use monthly mileage as the evaluation criterion, as the frequent lack of ignition status information prohibits basing the test on the number of trips or average driving speeds. The comparison of filtered and kept motorists shows significant differences in both pre- and post-accident mileage of accident and non-accident drivers ($p < 0.001$).\(^{35}\) However, the accident intervention effect (i.e., the reduction in mileage) in discarded drivers is similar to that of the retained motorists.

---

\(^{35}\) For this purpose, a non-parametric Mann-Whitney-U Test was applied.
5. Data and Methodology

retained drivers (for methodological details on this analysis see the covariance design as described in subsection 5.4.8), with the impact on eliminated motorists being even greater in both relative and absolute terms. Thus, I argue that even though this filtering step may impose bias to the data, it actually narrows the measured effect size. The outcome of my research can therefore be regarded as rather conservative.

Next, I test for filtering bias on vehicles not used anymore after collision. For this purpose, I compare kept and discarded drivers on their trips, mileage, and average speeds before the accident. On all three measures, the tests do not result in significant differences between filtered and retained motorists ($p < 0.001$). An additional comparison of group differences on the maximum acceleration (cf. Table 3) during the course of the accident also yields no significant differences ($p < 0.001$). This outcome suggests that also drivers who had not been caught in a severe accident stopped using their vehicle thereafter; in other words, filtering was done independent of accident severity. This becomes comprehensible when considering that the decision to scrap a car is not solely dependent on crash severity, but also on its current value: if it was already of low value before the accident, even a minor crash can make a motorist switch to a new vehicle if projected repair costs exceed the actual value of the automobile. In summary, these results justify assuming that no bias has been imposed on the data by the second filtering step.

For the driver group that got filtered due to defective GPS units, no statistical tests are performed. Instead, I argue that the device malfunctioning is caused by the rare occurrence of substandard hardware due to random fluctuations during manufacturing. As devices are also installed at random, it is justified to presume that filtering drivers with defective devices does not cause distortions in the underlying data.

5.4.5 Compilation of Individual Driver Information

I next aggregate the filtered data to monthly travel and driving characteristics for each individual motorist. In line with the developed research hypotheses, I determine the number of trips, the distance traveled, and the average speed for each driver month. The calculation of these metrics is straightforward and thus not
elaborated in detail. In addition, I assign each driver a dichotomous variable for distinguishing low- and high-mileage motorists. The cutoff value is set at a projected annual mileage of 30,000km, which is in accordance with the German Traffic Safety Council’s definition (Deutscher Verkehrssicherheitsrat, 2010). As the car owners’ annual travel distance is unknown within this research, I estimate it by adding up the total mileage of the five months that precede the crash / reference month and multiplying it by $\frac{12}{5} = 2.4$.

5.4.6 Endpoint Clustering

In this and the next subsection, I describe two methodological steps that are required for answering research hypothesis 6, which postulates that the accident impact on trip frequency varies by trip purpose. As these are not coded in the available data, I use a combination of cluster and classification algorithms to infer them from the GPS data. The former algorithm identifies a driver’s significant locations and derives different characteristics from them (e.g., frequency of visit, arrival time, weekday of visit, and length of stay). The downstream classification step then uses these properties to determine cluster and trip purposes (i.e., home, workplace, shopping area, etc.). I will also use the findings of these methodological steps apart from accident analysis to answer research sub-question 4 (cf. subsection 6.3).

Clustering with regard to the properties of spatial data

Cluster analysis is a method of unsupervised learning and is applied to detect structures in unclassified data. It forms groups of persons or objects based on the similarity of their attributes. In general, two different clustering approaches exist, namely partitioning and hierarchical algorithms (Kaufman & Rousseeuw, 2005). The former groups a set of objects into a number of predefined clusters $k$. However, such domain knowledge is frequently not available for many applications, as is the case for spatial data; the parameter $k$ depends on the living situation of each individual and thus has to be set for each motorist individually, which prohibits the use of partitioning cluster algorithms such as $k$-means in this context. Nonetheless,

36 Other variables as the dichotomous parameter for accident / non-accident drivers and maximum acceleration and registered vehicle owner are already present in the data and do not have to be derived separately.
some research exists that intends to infer the number of clusters $k$ from the data. Ng & Han (1994), for example, propose the partitioning algorithm CLARANS for the analysis of spatial databases. Although this method computes a good estimator for $k$ given a specific dataset, it is prohibitive for large databases owing to the exponentially growing computational runtime. Similarly, Ashbrook & Starner (2003) use a modification of the $k$-means algorithm but run into the same runtime problem as did Ng & Han. Another factor that militates against using partitioning approaches on spatial data is that the clusters found are convex, although most of a driver’s significant locations may be non-convex; when coming home, drivers may park their cars alongside streets in the neighborhood, so the resulting clusters may be shaped arbitrarily. Likewise, parking lots in front of shopping malls or grocery stores might be non-convex. As a consequence, partitioning algorithms may probably misclassify single trips or find more clusters than there actually are. I therefore do not regard them as applicable for the spatial data at hand.

A second approach is hierarchical clustering, which creates hierarchical decompositions of a given dataset. The creation of these subsets can be done both divisively (top-down) and agglomeratively (bottom-up), starting either from 1 or from $n$ clusters, with $n$ being the number of single data entities. Knowledge on the number of clusters to be generated is not necessary. However, a terminal condition must be set, usually some critical distance when the agglomeration or division process should end. The challenge is to set this parameter small enough to identify all “natural” clusters, while being large enough to avoid a true cluster being split into two separate ones. It becomes even more difficult when considering that single locations require different terminal conditions; street parking in one’s neighborhood for example might require a larger threshold than visits of a grocery store. Further variance may be induced by considering the same for urban versus rural areas. As a consequence, I also regard hierarchical algorithms as of limited use within my research.

The DBScan algorithm
The special properties of geographical information require an algorithm that overcomes the aforementioned pitfalls of conventional clustering schemes. It should require as little domain knowledge as possible (i.e., work without specifying the number of clusters and a cutoff criterion up front) and perform well on large databases. The algorithm “DBScan” developed by Ester et al. (1996) accounts for these considerations and follows a density-based notion of clusters. Today, it is the standard approach in clustering spatial data. Besides the minimum required domain
knowledge, it is capable of finding clusters of arbitrary shape and sorting out noise in the dataset (i.e., locations with infrequent visits), and it works well on large databases. In the following paragraphs, I describe the general concept of the algorithm and show on which parameters it operates. For a detailed outline of DBScan and the necessary adjustments for it to work in the used software package see Appendix A.

DBScan is based upon the thought that clusters can be recognized as the density of points within a cluster is considerably higher than that outside. This density notion is captured by two parameters, $\varepsilon$ and $n_{min}$. $\varepsilon$ describes a threshold up to which two data points are considered as neighbors. In order for an arbitrary point to form a cluster, at least some minimum number of points $n_{min}$ have to be in the $\varepsilon$-neighborhood of that point. Together, both parameters determine the clustering outcome, including the number of identified clusters and the share of unclassified items (i.e., noise). Increasing $\varepsilon$ relaxes the detection of clusters and allows looking for neighbors within a bigger search radius. Increasing $n_{min}$ constrains the process, making it harder for a given point to form a cluster seed. Ester et al. (1996) describe a heuristic that can be used to specify the two parameters. They suggest setting $n_{min} = 4$ and deriving $\varepsilon$ with regard to the properties of noise and cluster points. This recommendation is of limited practicality within my research, however. First, low values of $n_{min}$ eventually result in a plethora of identified clusters, as several thousand trips are available for each driver. Second, some minimum number of cluster members is required in order to derive characteristics such as cluster arrival times, visiting frequencies, and lengths of stay, which are necessary for subsequent trip classification.

Within this research, I have tested several different combinations of $\varepsilon$ and $n_{min}$ on DBScan. Good results were achieved for $n_{min} = 36$, which corresponds to at least three monthly visits in the study period. As $n_{min}$ determines the minimum number of cluster visits, I regard the used parameter as sufficient to derive corresponding cluster characteristics.

Practical considerations guide the determination of $\varepsilon$. An upper boundary for this parameter can rationally be seen in the minimum distance for which people start using their vehicle. Corresponding trip lengths can be small, for example when going grocery shopping. For this dissertation, I assume this minimum trip length to be 300m. Due to the agglomerative nature of the DBScan algorithm, $\varepsilon$ can actually be set smaller than this critical distance (cf. Figure 15). By trying different values for $\varepsilon$, I determine the parameter value that best approximates the average cluster
5. Data and Methodology

For the setting $\varepsilon = 250$ m, the average latitudinal (longitudinal) 95% confidence interval for a cluster is $\approx 286$ m ($\approx 330$ m). I regard these average values as sufficiently close to the proposed minimal cluster diameter. Consequently, I use $\varepsilon = 250$ m for trip end clustering.

![Figure 15: The relationship between $\varepsilon$ and minimal cluster distance](image)

With the specified parameter values, I run DBScan separately for each individual motorist. Note that I consider trip end locations only, thus ignoring departure points. For one, GPS information on trip ends usually is more accurate, as GPS devices require some time to detect the satellite signal and compute the current location upon trip start (Stopher et al., 2005). Second, considering both trip start and end information would cause redundancy, as each trip end point eventually becomes a trip starting point once the task at the visited location has been completed.

As a result, DBScan detects a total of 6580 clusters on the 549 accident-involved motorists. This includes 549 noise “clusters” which contain all the trip end locations that could not be attributed to a spatially delimited cluster.

---

37 For this determination, I consider the two-sided 95% confidence interval of the spread of an average cluster in both latitudinal and longitudinal direction. Using confidence intervals rather than maximal and minimal cluster spreads makes this approach more robust against possible clustered outliers.
**Deriving cluster characteristics**

After the trip end clustering is completed, I derive aggregate information on all of the drivers’ significant locations. These characteristics include:

- Total number of visits, first visit and last visit, number of distinct visiting days
- Arrival times and weekdays of arrival
- Length of stay
- Visiting frequency
- Share of daily and weekly visits
- Latitudinal and longitudinal cluster spread (as 95% confidence intervals around the cluster mean)

Within this process, I treat trips classified as noise as an own cluster and derive these characteristics for them as well, even though they do not form a cluster in the sense of spatially adjacent points. Furthermore, as all noise “clusters” are known once the DBScan algorithm has finished, I can exclude them from the subsequent step of trip classification.

### 5.4.7 Trip Classification

The intent of trip classification is to categorize an individual driver’s trips by purpose. The idea is to infer the function of each cluster from its characteristics and set the purpose of associated trips accordingly. While for a limited set of instances (i.e., clusters) such classification may be done manually by an expert with sufficient domain knowledge, the number of clusters at hand prohibits a complete manual assignment of trip purposes. Instead, I perform manual attribution just for subset instances, while for the rest I apply techniques of supervised machine learning to

---

38 This assignment is based on the assumption that visits to a particular location are all done for the same purpose. While this may be true for clusters of small spatial extension, bigger clusters eventually comprise trips of different purposes.
predict their purpose.\textsuperscript{39} The pre-classified data is used to build the corresponding classification model.

Within this research, I use decision trees to assign cluster functions automatically. Murthy (1998) lists several advantages of decision tree-based classifications:

- they are non-parametric, which allows modeling a plethora of data distributions;
- their hierarchical decompositions make better use of available features;
- they are capable of using uni- and multi-modal data;
- and they feature simple, easy-to-understand semantics, which makes them intuitively appealing.

More specifically, I apply the well-established C4.5 decision tree algorithm for automated trip classification. It was developed initially by Quinlan (1993) and revised by the author in later work to improve performance on continuous attributes (Quinlan, 1996). It is one of the most well-known algorithms for building decision trees in the literature today (Kotsiantis, 2007). A detailed methodological outline of this algorithm is provided in Appendix B. In the following paragraphs, I give detail on the trip classes I pre-specify, the manual assignment of trip purposes to training data, and the performance of the applied classifier.

\textbf{Used trip classes}

Before class labels can be assigned to spatial clusters, it is necessary to develop a taxonomy of trip purposes. As shown in subsection 4.6, several trip purpose classifications exist in the literature which vary considerably both in functional breadth and differentiation. For this research, I apply the classification scheme of Axhausen et al. (2003). With respect to the available data, however, some deviations from their scheme are necessary, as extra context information (e.g., motorist age, employment status, exact location of residence and workplace, type of

\textsuperscript{39} Kotsiantis (2007) defines supervised machine learning as “the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances.” “Supervised” thereby refers to the existence of distinct pre-specified classes, as opposed to “unsupervised” approaches such as clustering, where a finite number of instance classes may be found.
vehicle, etc.) that supplements the GPS data is missing. In detail, I subsume the classes pick up / drop off, private business, school, and daily shopping as used in Axhausen’s research under private business. Next, my taxonomy only holds a single work class instead of distinguishing between work and work-related trips. However, I add the extra class “fuel” to the taxonomy to denote stops at gas stations. This trip end category was not considered by the authors but was easily identifiable within my data (primarily due to its small cluster spread, its low but steady frequency of visits, and the very short dwell times). Finally, the “noise” class does not describe any collection of spatially close trip end points. It comprehends the trip destinations that could not be assigned to distinct clusters by the DBScan algorithm. They become discernible directly from the clustering process and thus can be exempt from the trip classification process. I nevertheless include them in the later analysis of trip reductions in different categories. Table 6 summarizes the used trip classes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home (H)</td>
<td>Habitation of motorists; head office of company car drivers</td>
</tr>
<tr>
<td>Work (W)</td>
<td>Workplace or locations visited for work-related purposes</td>
</tr>
<tr>
<td>Recreational (R)</td>
<td>Sport activities; going out; visiting friends; holiday trips and trips to secondary residences; dining</td>
</tr>
<tr>
<td>Private business (P)</td>
<td>Short term shopping; short term errands; drop-offs and pick-ups; mode change</td>
</tr>
<tr>
<td>Shopping (S)</td>
<td>Long term shopping</td>
</tr>
<tr>
<td>Fuel (F)</td>
<td>Gas stations</td>
</tr>
<tr>
<td>Noise (N)</td>
<td>Class of infrequently visited locations (noise as provided by the DBScan algorithm; does not contain trips to spatially close locations and is exempt from the classification analysis as the noise clusters are known up front)</td>
</tr>
</tbody>
</table>

*Table 6: Used trip classes*
Manual trip classification
I continue by manually assigning trip purposes to the significant locations of selected non-accident drivers. I sample the motorists by a stratified design, using car ownership (i.e., male, female, or company car drivers) as the stratification criterion. This is necessary as the different driver subgroups vary in their shares of trip purposes, for example with company car drivers exhibiting a considerably higher share of work-related car use. In total, I manually classify 1061 clusters of 103 sampled motorists, which include 34 male, 14 female, and 55 company car drivers.

For the manual classification process I use two sources of information. First, the cluster center is located on Google Earth. From there, it can be seen whether the cluster lies in a residential area, city centers, industrial zones, or others. This visual inspection also is intended to detect nearby shopping opportunities (i.e., convenience stores, shopping malls), restaurants, and recreational spots. Then, I compare the first notion of the cluster purpose with its characteristics (i.e., number of visits, weekday and daytime of arrival, length of stay). I synopsize both sources of information and then label the cluster with its most likely purpose. Nevertheless, manual classification occasionally turned out to be difficult, especially if cluster sizes were considerably large (which frequently happens if clusters point at city centers with a multitude of opportunities for shopping, dining, recreation, and personal business). In this case, I assign the cluster purpose in a way that best seems to match the task usually pursued at that particular location.

Applying the C4.5 decision tree algorithm for automated trip classification
Pre-classified data instances are next fed into the C4.5 algorithm for training purposes. I keep the preset values in the used implementation of the C4.5 algorithm (cf. subsection 5.5) except for the parameter “minimum number of objects.” It specifies the minimum number of data instances for a tree leaf so that the corresponding subset is not split up any further. By setting this value higher than the preset, proliferating decision trees that lack intuitive interpretation are avoided. In this research, I set the corresponding value to 15. After the tree has been built, I use cross-validation to assess the accuracy of the classification model. (cf. Kotsiantis,

---

40 Note that I perform the cluster and classification analyses on accident drivers only, as answering research question 6 does not require doing the same for non-accident motorists.
For this purpose, the training data is split into twelve distinct, equally-sized subsets (i.e., folds) which are independently tested by the models trained on the union of the other subsets. Table 7 shows the confusion matrix of this cross-validation process; diagonal items indicate correctly classified objects, off-diagonal items stand for misclassifications.

<table>
<thead>
<tr>
<th>True Purpose</th>
<th>Classified as</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>104</td>
<td>13</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>7</td>
<td>357</td>
<td>27</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>4</td>
<td>41</td>
<td>160</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>30</td>
<td>12</td>
<td>57</td>
<td>2</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>15</td>
<td>16</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>18</td>
<td>5</td>
<td>16</td>
<td>11</td>
</tr>
</tbody>
</table>

*Table 7: Confusion matrix of cross-validated training data*

The results show that 72.23% of instances have been classified correctly by the developed decision tree. Home clusters are predicted with especially high accuracy due to unique characteristics like frequent extended stays, similar visiting frequencies for all weekdays (except Sundays), and a relatively high share of arrivals at later times of the day. For the remainder of trip classes, characteristics are less distinctive and thus result in significant misclassifications. Nevertheless, the performance of the algorithm can be deemed satisfactory, as from a lower bound of prediction accuracy of 42.80% (when labeling all instances by the most frequent activity, i.e., work), the gain in predictive power achieved by the model is substantial. While other research, for example from Liao et al. (2005, 2007), generally achieves higher predictive accuracy, the gain in predictive power from a lower bound in those models is nonetheless comparable. Considering the fact that training instances were classified by the experiment participants themselves in the research of Liao et al., I regard the created decision tree model as sufficient to be applied to the unclassified locations of my research.
5. Data and Methodology

5.4.8 Regression Analysis

Upon completion of all previously described methodological steps, I next move towards testing the stated research hypotheses. For the hypothesis sets 1, 2, and 5, I apply regression analysis; for the remainder, I use the non-parametric test procedures as described in subsection 5.4.9.

Experimental design

I carry out regression analysis as a pretest-posttest design with case and control group (cf. Figure 16). The used design equals the pretest-posttest randomized experiment, yet lacks its key feature, namely the random assignment of experiment participants to treatment and non-treatment groups. This absence of arbitrary study participant assignment may be a cause of group value differences; thus it is referred to as “Nonequivalent Groups Design” (Trochim, 2001). It frequently is applied if the random assignment of treatments is either not practicable or not acceptable from an ethical viewpoint. Instead, it uses intact groups for cases and controls which develop beyond the control of the experimenter.

![Figure 16: Notation of the experimental design](image)

Reliability correction

As subsection 6.1 shows, substantial group differences exist in the pre-accident values of all three test variables. These discrepant attributes may, however, bias the regression outcome and result in distorted regression coefficients (Trochim, 2001).41 The impact of accidents on travel and driving behavior may thus not be estimated correctly. The reason is that regression coefficients get attenuated by measurement error on the pretest in combination with group nonequivalence; correcting this attenuation is necessary in order to receive unbiased results (Kenny, 1975). To solve

---

41 A thorough description of the group nonequivalence problem is provided by the author. It may also be accessed from its webpage at [http://www.socialresearchmethods.net/kb/statnegd.php](http://www.socialresearchmethods.net/kb/statnegd.php).
this problem, Trochim (2001) proposes using reliability-adjusted pretest measures in the analysis. In classical test theory, reliability is defined as
\[ \rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \]
and equals true score variance \( \sigma_T^2 \) divided by observed score variance. The latter is the sum of true score and error variance \( \sigma_E^2 \); thus reliability reflects the proportion of measurement error in the analysis. Since the measurement error cannot be determined directly, however, other reliability estimates must be used as a proxy for quantifying measurement error. The author suggests using Cronbach’s Alpha as a reliability estimate\(^{42}\) as given by
\[ \alpha = \frac{K}{K - 1} \left( 1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2} \right). \]
\( K \) is the number of components of the scale (i.e., items) in survey constructs. In this research, \( K = 2 \) and refers to the pre- and post-accident measures used for analysis. \( \sigma_{Y_i}^2 \) denotes the separate variance of pre- and post-accident values, while \( \sigma_X^2 \) is the variance of the observed total test scores. For all three test variables (monthly trips, mileage, and average driving speeds), the reliability-adjusted pretest values are calculated as
\[ X^* = \bar{X}_j + \alpha(X - \bar{X}_j). \]
\( X \) and \( X^* \) are the pretest and adjusted pretest values, respectively. \( \bar{X}_j \) indicates the mean measurement scores of case and control group drivers, as the reliability correction must be done separately for both motorist groups. \( \alpha \) names Cronbach’s Alpha. I will use \( X^* \) instead of \( X \) in the carried out regression analysis in order to rule out the bias caused by group nonequivalence.

\(^{42}\) Trochim (2001) actually proposes to use different reliability measures for reliability correction. He states that the test-retest reliability coefficient (also known as the Spearman-Brown coefficient) functions as a lower-bound estimate for reliability, while Cronbach’s Alpha can be regarded as an upper threshold. Both measures have been applied and compared within this research, yet no difference in calculated reliabilities has been observed. As a consequence, I perform the reliability adjustments only on the basis of Cronbach’s Alpha.
Analysis of Covariance
All performed regression analyses use a covariance design (ANCOVA). It tests whether certain factors impact an outcome variable after eliminating the variance that quantitative predictors (i.e., covariates) account for. In my research, post-accident values serve as dependent variables, which are set in relation to a factor (i.e., crash involvement for hypotheses sets 1 and 2 and accident severity for hypotheses set 5) and the covariate pre-accident behavior. The latter intends to increase the statistical power of the tests by accounting for some of the variance of the dependent variable. The resulting regression coefficient for crash involvement indicates the absolute change in travel and driving performance and documents to which degree accidents influence these behaviors.

5.4.9 Non-Parametric Testing
The hypotheses sets 3, 4, and 6 are validated by non-parametric test procedures. As for example Altman (1991) states, “parametric methods require the observations within each group to have an approximately Normal distribution […] if the raw data do not satisfy these conditions […] a non-parametric method should be used […].” In line with that, I apply Mann-Whitney-U-Tests to evaluate differences in two driver groups and Kruskal-Wallis-Tests to study more than two groups. All tests are carried out on change scores (cf. Vickers, 2005) – that is, the differences between pre- and post-accident driving – which all lack the aforementioned prerequisite of normality.43 Note that all of these tests intend to identify differences in the subgroups of accident-involved drivers; thus considering non-accident motorists is not necessary.

5.5 Statistical Software
All steps for data preprocessing, trip aggregation, filtering, and indicator development are done using Microsoft Visual Basic 6.5, which is a standard component of Microsoft Office packages. End point clustering and trip type imputation are implemented in the Ganymede Eclipse Software Developer Kit 3.4.2

43 Kolmogorov-Smirnov tests have shown non-normality of all change scores of trips, mileage, and travel speeds.
using Java Release 6. The DBScan algorithm and the C4.5 decision tree are both taken from the Weka 3.6.1 Software Package of the University of Waikato, New Zealand.\textsuperscript{44} All statistical tests are carried out in SPSS 16.

\textsuperscript{44} The Weka 3.6.1 software package is available at: \url{http://www.cs.waikato.ac.nz/ml/weka/}. The C4.5 algorithm can be accessed via the classifier “J48”.
6. **Findings**

This section presents the research findings. First, I provide descriptive statistics on the three test variables (trips per month, distance traveled per month, and average driving speed per month), show the outcome of the cluster and classification steps, and detail accident characteristics. Second, I present the results of the hypotheses testing. I conclude by exemplarily showing some other analyses on GPS data that might be of use aside from the insurance industry.

### 6.1 Descriptive Statistics

#### 6.1.1 Trips per Month

Figure 17 shows the histogram of monthly trips made by all drivers over the two year study period. For each of the 549 accident drivers, data on 13 consecutive months is available, while for the 899 non-accident drivers, data from the whole study period is at hand. This results in a total of 28,713 observations (i.e., driver months). In 364 driver months (1.3%), motorists did not use their vehicle. In 26062 driver months (90.8%), the car was used for up to 500 trips, while in 2287 cases (8.0%) car use exceeded that threshold.

![Figure 17: Histogram of trips per month](image-url)
Table 8 provides further descriptive statistics on monthly trips. For both driver groups, it shows the number of trips in the month that preceded the accident / reference month and the five thereafter. Group differences are tested using the Mann-Whitney-U-Test, as the data shows non-normality. The test yields a significant outcome in all pre- and post-accidents months ($p < 0.001$), which indicates that accident drivers engage in considerably more monthly trips than non-accident drivers. Note that the group difference in pre-accident values is the reason for using reliability-corrected covariates in the later regression models as described in subsection 5.4.8.

<table>
<thead>
<tr>
<th></th>
<th>Accident drivers</th>
<th>Non-accident drivers</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>$t = -1$</td>
<td>308.06</td>
<td>175.96</td>
<td>240.93</td>
</tr>
<tr>
<td>$t = 1$</td>
<td>273.60</td>
<td>186.80</td>
<td>238.25</td>
</tr>
<tr>
<td>$t = 2$</td>
<td>282.46</td>
<td>179.43</td>
<td>239.06</td>
</tr>
<tr>
<td>$t = 3$</td>
<td>283.40</td>
<td>193.23</td>
<td>240.99</td>
</tr>
<tr>
<td>$t = 4$</td>
<td>275.19</td>
<td>187.16</td>
<td>240.61</td>
</tr>
<tr>
<td>$t = 5$</td>
<td>263.53</td>
<td>192.04</td>
<td>234.71</td>
</tr>
</tbody>
</table>

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Descriptive statistics of trips per month

Table 9 displays the correlations of the monthly trips over time. All correlations are substantial and of high significance ($p < 0.001$), with highest values being observed in the correlation of post-accident months. The data suggests that correlations tend to decline slightly over time while remaining significant. The high degree of correlation for all combinations of pre- and post accident values supports the use of a covariance design in regression analyses.

---

45 Note that in this and all subsequent tables, $t = -1$ denotes data of the pre-accident month, while $t = 1$ to $t = 5$ refer to data of the five post-accident months.
<table>
<thead>
<tr>
<th>$t = -1$</th>
<th>$t = 1$</th>
<th>$t = 2$</th>
<th>$t = 3$</th>
<th>$t = 4$</th>
<th>$t = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = -1$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t = 1$</td>
<td>0.820***</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t = 2$</td>
<td>0.815***</td>
<td>0.878***</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t = 3$</td>
<td>0.831***</td>
<td>0.855***</td>
<td>0.910***</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$t = 4$</td>
<td>0.813***</td>
<td>0.824***</td>
<td>0.852***</td>
<td>0.905***</td>
<td>1</td>
</tr>
<tr>
<td>$t = 5$</td>
<td>0.785***</td>
<td>0.783***</td>
<td>0.810***</td>
<td>0.852***</td>
<td>0.879***</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 9: Correlations of trip per month

6.1.2 Distance traveled per Month

Figure 18 shows the histogram of monthly mileage for all motorists. As in subsection 6.1.1, only months where data is available were included in the analysis. In 435 driver months (1.5%), no mileage was accrued. This figure differs from the results in subsection 6.1.1, where 364 driver months without trips were identified. I suppose this deviation to be caused by the fact that some trips were too short for detecting the GPS signal, determining the vehicle position, and recording the driven distance (e.g., when reparking the car). As the data shows, motorists generally exhibit a high monthly mileage, with 614 (42.4%) of them having a projected

![Figure 18: Histogram of kilometers traveled per month](image)
annual mileage of 30,000km or higher. I regard this as being due to the high share of company car drivers in the sample and a more extensive car use amongst the Italian population in general.

Next, Table 10 shows the corresponding descriptive statistics. Again, crash-involved drivers show significantly more mileage per month than do their non-crash-involved counterparts (p < 0.001). This result is in line with other research, which shows that accident involvement risks increase with exposure, that is, distance traveled (Litman, 2008). A comparison of the mileage measure over time suggests changes for crash-involved drivers, while for non-accident drivers values remain constant with regard to the reference month. The significance of this argument is tested in subsection 6.2.1.

<table>
<thead>
<tr>
<th>Time</th>
<th>Accident drivers M</th>
<th>Accident drivers SD</th>
<th>Non-accident drivers M</th>
<th>Non-accident drivers SD</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = -1</td>
<td>3847.92</td>
<td>2287.62</td>
<td>2091.45</td>
<td>2099.92</td>
<td>107341.0***</td>
</tr>
<tr>
<td>t = 1</td>
<td>3272.88</td>
<td>2493.44</td>
<td>2029.04</td>
<td>2039.84</td>
<td>149052.5***</td>
</tr>
<tr>
<td>t = 2</td>
<td>3475.52</td>
<td>2467.07</td>
<td>2056.76</td>
<td>2074.49</td>
<td>133500.0***</td>
</tr>
<tr>
<td>t = 3</td>
<td>3404.58</td>
<td>2311.19</td>
<td>2127.70</td>
<td>2168.10</td>
<td>137187.0***</td>
</tr>
<tr>
<td>t = 4</td>
<td>3305.46</td>
<td>2355.13</td>
<td>2087.51</td>
<td>2095.36</td>
<td>149891.0***</td>
</tr>
<tr>
<td>t = 5</td>
<td>3213.21</td>
<td>2427.43</td>
<td>2026.97</td>
<td>2040.48</td>
<td>154779.0***</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 10: Descriptive statistics of kilometers traveled per month

Table 11 concludes with the correlations for the six months under study. Results are similar to those in 6.1.1, with all correlations being highly significant and tending to decline slightly over time.

46 For the calculation of the projected annual mileage see subsection 5.4.5.
6. Findings

<table>
<thead>
<tr>
<th></th>
<th>$t = -1$</th>
<th>$t = 1$</th>
<th>$t = 2$</th>
<th>$t = 3$</th>
<th>$t = 4$</th>
<th>$t = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = -1$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t = 1$</td>
<td>0.779***</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t = 2$</td>
<td>0.768***</td>
<td>0.830***</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t = 3$</td>
<td>0.779***</td>
<td>0.790***</td>
<td>0.830***</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t = 4$</td>
<td>0.778***</td>
<td>0.755***</td>
<td>0.775***</td>
<td>0.824***</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$t = 5$</td>
<td>0.739***</td>
<td>0.740***</td>
<td>0.748***</td>
<td>0.780***</td>
<td>0.845***</td>
<td>1</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 11: Correlations of kilometers traveled per month

6.1.3 Average Driving Speed per Month

In order to study the accident impact on driving behavior, the motorists’ average monthly travel speeds are calculated. Figure 19 shows the corresponding histogram for all driver months that exhibit positive mileage (N = 28277). They can be further broken down by road type, which makes different speed choices at urban roads, extra-urban roads, and highways discernible (cf. Figure 20).

![Figure 19: Histogram of average driving speeds per month](image-url)
Table 12 shows the average driving speeds of accident and non-accident drivers in the six months considered. The groups exhibit significant differences in speed prior to the accident \( (p < 0.001) \), while post-accident values are ambiguous; significant differences in mean values were present only in months 2 and 3 \( (p < 0.001) \). Note that for all six months, data follows a normal distribution \( (p < 0.001) \), which allows group comparisons on parametric \( t \)-tests.

<table>
<thead>
<tr>
<th></th>
<th>Accident drivers</th>
<th>Non-accident drivers</th>
<th>( t )-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( M )</td>
<td>( SD )</td>
<td>( M )</td>
</tr>
<tr>
<td>( t = -1 )</td>
<td>41.99</td>
<td>14.00</td>
<td>38.02</td>
</tr>
<tr>
<td>( t = 1 )</td>
<td>39.00</td>
<td>16.13</td>
<td>37.76</td>
</tr>
<tr>
<td>( t = 2 )</td>
<td>40.97</td>
<td>15.64</td>
<td>37.68</td>
</tr>
<tr>
<td>( t = 3 )</td>
<td>40.98</td>
<td>15.59</td>
<td>38.10</td>
</tr>
<tr>
<td>( t = 4 )</td>
<td>38.96</td>
<td>16.44</td>
<td>37.53</td>
</tr>
<tr>
<td>( t = 5 )</td>
<td>39.03</td>
<td>17.49</td>
<td>37.81</td>
</tr>
</tbody>
</table>

\* \( p < 0.05 \), \** \( p < 0.01 \), \*** \( p < 0.001 \)

Table 12: Descriptive statistics of average driving speeds per month

Correlations follow the same pattern as in 6.1.1 and 6.1.2 and are highly significant \( (p < 0.001) \), which again suggests studying the accident impact on driving behavior by using covariate designs.
### 6.1.4 Cluster and Classification Results

In the following, I show the descriptive results of the applied cluster and classification steps. Note that these methodological steps involve data on accident-involved vehicles only. DBScan finds on average 10.99 (SD 6.77) clusters per accident-involved driver (not including the noise cluster), using the parameters $\varepsilon = 250\text{m}$ and $n_{\text{min}} = 36$ as derived in subsection 5.4.6. For 325 motorists (59.2%), up to 10 clusters become discernible, whereas 11 to 20 clusters are present in 184 cases (33.5%). Only 40 drivers (7.3%) exceed that threshold (cf. Figure 21).

![Histogram of the number of clusters per driver](image)

**Figure 21: Histogram of the number of clusters per driver**

The percentage of trips that can be attributed to clusters and are thus not regarded as noise is 67.0% (SD 14.2%) on average (cf. Figure 22). A majority of the motorists’ travel thus can be regarded as following specific routines, where known, distinct locations are visited on a regular basis.
Trip classification gives further detail on how these locations are composed. When considering the aggregate over all types of vehicle owners, the majority of travel apart from home-bound trips is done for work-related (15.1%) and recreational (13.8%) purposes. Personal business makes up for 7.4% of trips. The low share of shopping related trips (1.0%) may be explained by the fact that only few drivers visit larger shopping malls regularly enough in order to be identified as clusters. As $n_{\text{min}}$ requires a minimum of 36 cluster visits, few shopping locations seem to be visited that regularly during the study period. Furthermore, a lot of shopping is actually done short-term and gets subsumed under “private business,” which may also explain the low frequency of shopping related trips.

Considerable differences in car use exist for company car, male, and female drivers. For the first subgroup, a high degree of business-related car use becomes evident in the data. This not only reflects daily commutes to work, but also visits at other company branches and regular delivery runs. The high percentage of unclustered trips is striking, however, and reflects infrequent visits of business clients in remote locations. If visits are done on a less than weekly basis, the chances are high that corresponding locations are not regarded as clusters by DBScan, thus causing this high share of unclassified trips for this driver subgroup. In comparing men and women, the latter exhibit fewer trips categorized as noise. In other words, women’s travel routines exhibit less variance, as they visit certain locations of interest more regularly. Their trips are classified as recreational (i.e., visiting friends and family) in 22% of the cases and are generally more home bound than that of men. The latter suggests that women’s trip destinations are located more closely to home than that of men, which makes returning home after the task at a specific location has been completed more worthwhile. Note that differences between men and women in
work-related car use are marginal, although varying degrees of employment suggest this deviation to be more pronounced.

Next, I present the distribution of arrival hours and arrival days per trip class, cumulated over all accident drivers (cf. Figure 23 and Figure 24). The intuitive notion these figures express was essential for the manual trip classification performed earlier. Note that the high congruence of these results with intuition also confirms the predictive capability of the classification algorithm to attribute trip purpose correctly.

As expected, pronounced differences exist between the different cluster types. Home clusters show a peak in visits around 5:00pm, often directly after work. Another one becomes noticeable at around 11:00am, which can be regarded as

\[\text{Figure 23: Distribution of arrival times by trip purpose}\]
coming home for lunch. Work places are predominantly visited in the morning, between 7:00am and 10:00am; company car drivers’ client visits account for the relatively high share of work-related trips thereafter. Recreational activities take place primarily in the afternoon and account for most journeys in the later afternoon and at night apart from home-bound trips. Shopping trips accurately predict Italy’s legal shop opening hours, which last until 9:00pm.

Figure 24: Distribution of arrival days by trip purpose

Similar distributions are available for weekdays of arrival. Generally, motorists engage in most travel during the weekdays and on Saturday, with trips of all purposes declining on Sunday except those for recreation. While home is visited most constantly throughout the week, trips for work and shopping related activities decline significantly on the weekends (as legal regulation mandates shopping malls
to be closed on Sundays in Italy). Recreational car use surges on weekends while private business trips, comprising short term shopping and errands, spread evenly over the weekdays.

### 6.1.5 Accident Times and Locations

This subsection provides details on accident times and locations. Figure 25 shows the temporal distribution of accidents in relation to trip patterns. As becomes evident, motorists are especially at risk in three periods where crash probabilities exceed trip frequencies: 7:00am to 8:00am, 4:00pm to 5:00pm, and 2:00am to 3:00am. Horner & Reyner (1995), who identify similar temporal accident patterns, explain that this is the result of an increased probability of driver sleepiness at these times. During morning and evening rush hour traffic, high vehicle density requires considerable driver involvement (i.e., distance keeping, lane changing, and overtaking), which eventually increases crash propensity if sleepiness affects driver attention. For past midnight rides, crash risks may further be amplified by poor lighting conditions and possible substance impaired driving after social and recreational events.

![Figure 25: Distribution of accident and trip times](image)

Figure 26 illustrates accident occurrence by road type. For the available data, collisions occur during inner-city travel in more than 50% of the cases. This again reflects necessary driver involvement, which is highest on urban roads with a plethora of turning and stopping maneuvers.
Figure 26: Road types of accidents

Figure 27 indicates how familiar motorists were with the accident location by detailing the percentage of trips that pass by that particular spot. As the data shows, motorists are much more likely to face a vehicle crash on roads they do not use on a regular basis. Especially for places they just pass by on every 20\textsuperscript{th} trip or less, accident risks are significantly increased compared to those locations they pass by more frequently. For places that are part of 70\% of their trips or more, crash propensity drops to almost zero. As a consequence, it may be inferred that accident probability rises with the variability of spatial trip patterns, with motorists that visit just few distinct locations having a lower chance of experiencing a vehicle collision.

Figure 27: Percentage of trips that pass by the accident location

\footnote{Note that this analysis is not based on trips that pass by the exact accident location. Instead, it considers trips that come as close as 1km to the place of the vehicle crash. This is necessary due to the recording interval of the GPS data, which is approximately 2km.}
6.2 Inference Statistics

Prior to this point, I have used descriptive statistics to explore the effects of vehicle crashes on monthly trips, mileage, and speed choice. The data seems to suggest that accidents have a lowering effect on these measures, a decline which is not reflected in the control group. I now use inference statistics to study the significance of these differences and test the research hypotheses as developed earlier.

6.2.1 The Accident Impact on Travel and Driving Behavior

First, I test the hypotheses for the accident impact on travel and driving behavior. They provide the basis for all subsequent analyses. As stated earlier, I use a covariance design, with accident involvement serving as the factor, pre-accident data as the covariate, and post-accident data as the dependent variable. All analyses are carried out on reliability-adjusted covariates (cf. subsection 5.4.8).

Tables 14 through 16 show the regression results for the accident impact on monthly trips, mileage, and average driving speed. In general, pre-accident measures are highly predictive of post-accident measures, with significant regression coefficients around 1 in all cases. The accident influence is highly significant as well. As the data shows, crash involvement makes motorists reduce their monthly trips by 35.091 units (11.4%) in the first month that follows the accident. Similar results are achieved for monthly kilometers, which decline by 536.606km (13.9%), and for average speeds, falling by 3.135 km/h (7.5%). The results support the postulated hypotheses on the accident impact on travel and driving behavior: as a response to accident experience, motorists limit their car use, while applying lower speed choice to unavoidable trips. The significance of these findings gives full support to the research hypotheses 1a through 1c.

\[
\begin{array}{lcc}
\hline
\text{Predictor} & \text{Coefficient} & \text{SE} & \text{t-Value} \\
\hline
\text{Constant} & -14.567 & 5.960 & -2.444* \\
\text{Trips} t = -1 & 1.049 & 0.019 & 54.521*** \\
\text{Crash} & -35.091 & 6.217 & -5.645*** \\
\hline
\end{array}
\]

* p < 0.05, ** p < 0.01, *** p < 0.001

\[ R^2 = 0.675 \]

\textit{Table 14: Accident impact on trips per month}
### 6.2 Inference Statistics

#### 6.2.1 Predictor Coefficients for Kilometers Traveled Per Month

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-90.946</td>
<td>67.728</td>
<td>-1.343***</td>
</tr>
<tr>
<td>Kilometers $t = -1$</td>
<td>1.014</td>
<td>0.023</td>
<td>44.515***</td>
</tr>
<tr>
<td>Crash</td>
<td>-536.606</td>
<td>87.842</td>
<td>-6.109***</td>
</tr>
</tbody>
</table>

* * p < 0.05, ** p < 0.01, *** p < 0.001

$R^2 = 0.607$

**Table 15: Accident Impact on Kilometers Traveled per Month**

#### 6.2.2 Predictor Coefficients for Average Driving Speed Per Month

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.180</td>
<td>1.284</td>
<td>-3.254**</td>
</tr>
<tr>
<td>Speed $t = -1$</td>
<td>1.103</td>
<td>0.032</td>
<td>34.148***</td>
</tr>
<tr>
<td>Crash</td>
<td>-3.135</td>
<td>0.624</td>
<td>-5.026***</td>
</tr>
</tbody>
</table>

* * p < 0.05, ** p < 0.01, *** p < 0.001

$R^2 = 0.447$

**Table 16: Accident Impact on Average Driving Speed per Month**

### 6.2.2 The Persistency of Effects on Travel and Driving Behavior

In the previous subsection, I showed that accident involvement has a significant influence on motorists’ car use behavior. Theory assumes these effects to deteriorate the more time has passed since the vehicle crash. Consequently, when relating pre-accident data and crash involvement to post-accident data, the accident impact should lower and eventually become insignificant, as motorists regain confidence in vehicle operation and adjust to their old driving style over time.

Tables 17 through 19 show the results of the pursued regression analyses. In each table, pre-accident variables and crash involvement are set in relation to the measures of the five consecutive post-accident months (indicated by $t = 1$ through $t = 5$). For the travel behavior measures (i.e., monthly trips and mileage), the accident influence remains significant during follow-up, with crash-involved drivers not reaching their pre-accident scores in either trips or mileage. A possible explanation for this might be that crash occurrence in part breaks habitual driving routines, making motorists rethink which rides are avoidable and may be cut or substituted by other modes of transportation. Once established, such behavioral changes may become permanent if motorists do not perceive a loss in mobility or mobile convenience. An alternative explanation might be that accidents pose stronger interventions to travel than do others as discussed in the literature, as they affect the cognitive foundations of driving (cf. subsection 4.1). Recovering from such cognitive impairments may actually exceed the five month post-accident time
horizon under review here. In absence of any other follow-up data on post-accident travel, I thus conclude that accident-involved drivers do not recover in terms of pursued travel activities, which leads me to reject research hypotheses 2a and 2b.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t = 1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-14.567</td>
<td>5.960</td>
<td>-2.444***</td>
<td>0.675</td>
</tr>
<tr>
<td>Trips ( t = -1 )</td>
<td>1.049</td>
<td>0.019</td>
<td>54.521***</td>
<td></td>
</tr>
<tr>
<td>Crash</td>
<td>-35.091</td>
<td>6.217</td>
<td>-5.645***</td>
<td></td>
</tr>
<tr>
<td>( t = 2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-5.720</td>
<td>5.923</td>
<td>-0.966</td>
<td></td>
</tr>
<tr>
<td>Trips ( t = -1 )</td>
<td>1.016</td>
<td>0.019</td>
<td>53.113***</td>
<td></td>
</tr>
<tr>
<td>Crash</td>
<td>-24.797</td>
<td>6.179</td>
<td>-4.013***</td>
<td></td>
</tr>
<tr>
<td>( t = 3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-17.594</td>
<td>5.885</td>
<td>-2.990**</td>
<td></td>
</tr>
<tr>
<td>Trips ( t = -1 )</td>
<td>1.073</td>
<td>0.019</td>
<td>56.479***</td>
<td></td>
</tr>
<tr>
<td>Crash</td>
<td>-29.642</td>
<td>6.138</td>
<td>-4.829***</td>
<td></td>
</tr>
<tr>
<td>( t = 4 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-4.770</td>
<td>5.940</td>
<td>-0.803</td>
<td></td>
</tr>
<tr>
<td>Trips ( t = -1 )</td>
<td>1.018</td>
<td>0.019</td>
<td>53.094***</td>
<td></td>
</tr>
<tr>
<td>Crash</td>
<td>-33.793</td>
<td>6.196</td>
<td>-5.454***</td>
<td></td>
</tr>
<tr>
<td>( t = 5 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-4.951</td>
<td>6.365</td>
<td>-0.778</td>
<td></td>
</tr>
<tr>
<td>Trips ( t = -1 )</td>
<td>0.995</td>
<td>0.021</td>
<td>48.394***</td>
<td></td>
</tr>
<tr>
<td>Crash</td>
<td>-37.959</td>
<td>6.639</td>
<td>-5.717***</td>
<td></td>
</tr>
</tbody>
</table>

* \( p < 0.05 \), ** \( p < 0.01 \), *** \( p < 0.001 \)

Table 17: Accident impact on trips in the 5 follow-up months

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t = 1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-90.946</td>
<td>67.728</td>
<td>-1.343</td>
<td></td>
</tr>
<tr>
<td>Kilometers ( t = -1 )</td>
<td>1.014</td>
<td>0.023</td>
<td>44.515***</td>
<td>0.608</td>
</tr>
<tr>
<td>Crash</td>
<td>-536.606</td>
<td>87.842</td>
<td>-6.109***</td>
<td></td>
</tr>
<tr>
<td>( t = 2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-21.021</td>
<td>70.149</td>
<td>-0.300</td>
<td></td>
</tr>
<tr>
<td>Kilometers ( t = -1 )</td>
<td>0.993</td>
<td>0.024</td>
<td>42.123***</td>
<td>0.590</td>
</tr>
<tr>
<td>Crash</td>
<td>-336.230</td>
<td>90.981</td>
<td>-3.586***</td>
<td></td>
</tr>
<tr>
<td>( t = 3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>10.417</td>
<td>67.900</td>
<td>0.153</td>
<td></td>
</tr>
<tr>
<td>Kilometers ( t = -1 )</td>
<td>1.012</td>
<td>0.023</td>
<td>44.345***</td>
<td>0.607</td>
</tr>
<tr>
<td>Crash</td>
<td>-501.302</td>
<td>88.065</td>
<td>-5.692***</td>
<td></td>
</tr>
<tr>
<td>( t = 4 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-6.958</td>
<td>67.022</td>
<td>0.104</td>
<td></td>
</tr>
<tr>
<td>Kilometers ( t = -1 )</td>
<td>1.001</td>
<td>0.023</td>
<td>44.442***</td>
<td>0.606</td>
</tr>
<tr>
<td>Crash</td>
<td>-541.053</td>
<td>86.926</td>
<td>-6.224***</td>
<td></td>
</tr>
<tr>
<td>( t = 5 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>50.262</td>
<td>71.702</td>
<td>0.701</td>
<td></td>
</tr>
<tr>
<td>Kilometers ( t = -1 )</td>
<td>0.945</td>
<td>0.024</td>
<td>39.205***</td>
<td>0.547</td>
</tr>
<tr>
<td>Crash</td>
<td>-473.861</td>
<td>92.996</td>
<td>-5.096***</td>
<td></td>
</tr>
</tbody>
</table>

* \( p < 0.05 \), ** \( p < 0.01 \), *** \( p < 0.001 \)

Table 18: Accident impact on kilometers traveled in the 5 follow-up months
Driving behavior displays an ambiguous result. For post-accident months 1, 4, and 5, again the accident-impact is highly significant ($p < 0.001$). For months 2 and 3, however, the effects are insignificant and of low significance ($p < 0.05$), respectively. In order to give support to the postulated hypothesis, results should be the exact opposite though, that is, the crash involvement variable should become insignificant at later points in time. Consequently, I reject hypothesis 2c as well, although a longer follow-up period might yield different results.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>$t$-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.180</td>
<td>1.284</td>
<td>-3.254***</td>
</tr>
<tr>
<td>Speed $t = -1$</td>
<td>1.103</td>
<td>0.032</td>
<td>34.148*** $R^2 = 0.448$</td>
</tr>
<tr>
<td>Crash</td>
<td>-3.135</td>
<td>0.624</td>
<td>-5.026***</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.654</td>
<td>1.257</td>
<td>-3.704***</td>
</tr>
<tr>
<td>Speed $t = -1$</td>
<td>1.113</td>
<td>0.032</td>
<td>35.233*** $R^2 = 0.468$</td>
</tr>
<tr>
<td>Crash</td>
<td>-1.132</td>
<td>0.610</td>
<td>-1.854</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.328</td>
<td>1.320</td>
<td>-1.764</td>
</tr>
<tr>
<td>Speed $t = -1$</td>
<td>1.063</td>
<td>0.033</td>
<td>32.032*** $R^2 = 0.420$</td>
</tr>
<tr>
<td>Crash</td>
<td>-1.338</td>
<td>0.641</td>
<td>-2.087*</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.843</td>
<td>1.319</td>
<td>-2.913**</td>
</tr>
<tr>
<td>Speed $t = -1$</td>
<td>1.088</td>
<td>0.033</td>
<td>32.796*** $R^2 = 0.428$</td>
</tr>
<tr>
<td>Crash</td>
<td>-2.885</td>
<td>0.641</td>
<td>-4.503***</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.519</td>
<td>1.370</td>
<td>-3.299**</td>
</tr>
<tr>
<td>Speed $t = -1$</td>
<td>1.113</td>
<td>0.034</td>
<td>32.316*** $R^2 = 0.420$</td>
</tr>
<tr>
<td>Crash</td>
<td>-3.199</td>
<td>0.665</td>
<td>-4.808***</td>
</tr>
</tbody>
</table>

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 19: Accident impact on average driving speed in the 5 follow-up months

6.2.3 The Impact on Male, Female, and Company Car Drivers

The accident impact on different driver subgroups is first analyzed for male ($N = 185$), female ($N = 63$), and company car drivers ($N = 297$). Theory suggests variations in travel and driving behavior response due to the varying car dependency among the three groups and differences in the risk perception of operating a vehicle. Tables 20 through 22 show the pre- and post-accident values

---

$^{48}$ Four drivers with missing data on this attribute were discarded.
for all driver subgroups and corresponding differences. Pre-accident data suggests that men exhibit more trips, mileage, and a higher speed choice prior to accident involvement than women, although significant group dissimilarities are only present in mileage \((p < 0.01)\). Company car drivers differ significantly from non-company car drivers in terms of monthly mileage \((p < 0.001)\) and speed choice \((p < 0.01)\). With the number of trips being similar amongst both subgroups, it can be inferred that company cars are used for longer trips in order to reach more remote and / or distant locations. The higher driving speeds of company-owned vehicles suggest that motorists tend to drive more aggressively if they do not use their own cars, which is in line with related literature.

The group comparison of pre-post differences in travel and driving behavior yields only insignificant results, however. The applied Kruskal-Wallis-Tests do not show significant group deviations for either absolute or relative differences.\(^{49}\) This sets accidents apart from other kinds of interventions, where response intensity differs in gender. Consequently, I reject research hypotheses 3a through 3c.

<table>
<thead>
<tr>
<th>Male drivers</th>
<th>Female drivers</th>
<th>Company car drivers</th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M)</td>
<td>(SD)</td>
<td>(M)</td>
<td>(SD)</td>
</tr>
<tr>
<td>(t = -1)</td>
<td>322.86</td>
<td>202.16</td>
<td>272.14</td>
</tr>
<tr>
<td>(t = 1)</td>
<td>282.04</td>
<td>206.47</td>
<td>238.27</td>
</tr>
<tr>
<td>Difference</td>
<td>-40.82</td>
<td>146.76</td>
<td>-33.87</td>
</tr>
</tbody>
</table>

* \(p < 0.05\), ** \(p < 0.01\), *** \(p < 0.001\)

Table 20: Differences in trips per month for the 3 driver subgroups

<table>
<thead>
<tr>
<th>Male drivers</th>
<th>Female drivers</th>
<th>Company car drivers</th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M)</td>
<td>(SD)</td>
<td>(M)</td>
<td>(SD)</td>
</tr>
<tr>
<td>(t = -1)</td>
<td>3476.19</td>
<td>1809.04</td>
<td>2716.15</td>
</tr>
<tr>
<td>(t = 1)</td>
<td>2849.67</td>
<td>1920.03</td>
<td>2137.10</td>
</tr>
<tr>
<td>Difference</td>
<td>-626.52</td>
<td>1798.83</td>
<td>-579.06</td>
</tr>
</tbody>
</table>

* \(p < 0.05\), ** \(p < 0.01\), *** \(p < 0.001\)

Table 21: Differences in kilometers traveled for the 3 driver subgroups

\(^{49}\) The tests on relative differences are not shown here. Furthermore, note that Kolmogorov-Smirnov tests were applied to test for the normality of change values, which does not hold true on all three variables.
6.2 Inference Statistics

### 6.2.4 The Impact on Low- and High-Mileage Drivers

Next, I split the accident population into two subgroups based on projected annual mileage. I apply 30,000km as a threshold and form the subpopulations of low- \( (N = 140) \) and high- \((N = 409)\) mileage drivers. The literature suggests finding discrepancies in behavioral changes as a response to different car dependency and driver self-confidence between the subgroups. Similar to subsection 6.2.3, I test absolute pre-post changes in trips, kilometers, and average speeds.

Tables 23 through 25 show the results for the three test variables. The formed subgroups differ significantly in all their pre- and post-accident values \( (p < 0.001) \). While for the travel characteristics this can be expected from the grouping process, the discrepancy in speed choice is so pronounced that it requires some explanation. The literature anticipates and explains these differences in driving speed by varying degrees of driving self-confidence. Within the available data, however, road usage makes the primary difference, as high-mileage drivers usually do more extended trips and thus spend more time traveling on highways. Therefore, they automatically travel at higher speeds than low-mileage drivers, who use their cars for inner-city travel to a greater extent.

Differences in the interventional effects of accidents between the two subgroups are tested using Mann-Whitney-U tests, as change values are not normal-distributed. The results show that high-mileage drivers reduce both trips \((p < 0.05)\) and mileage \((p < 0.01)\) significantly more than do drivers in the other subgroup.\(^{50}\) This result is

---

\(^{50}\) When testing the relative reductions in trips and mileage, only the former remains significant \((p < 0.01)\). Results are not shown here.

---

<table>
<thead>
<tr>
<th></th>
<th>Male drivers</th>
<th></th>
<th>Female drivers</th>
<th></th>
<th>Company car drivers</th>
<th></th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t = -1)</td>
<td>41.34</td>
<td>13.89</td>
<td>38.63</td>
<td>15.84</td>
<td>44.19</td>
<td>13.34</td>
<td></td>
</tr>
<tr>
<td>(t = 1)</td>
<td>38.48</td>
<td>16.83</td>
<td>35.48</td>
<td>14.79</td>
<td>41.03</td>
<td>15.31</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>-2.86</td>
<td>13.69</td>
<td>-3.15</td>
<td>13.45</td>
<td>-3.16</td>
<td>11.13</td>
<td>0.589</td>
</tr>
</tbody>
</table>

* \( p < 0.05, ** p < 0.01, *** p < 0.001 *

*Table 22: Differences in average driving speed for the 3 driver subgroups*
in line with literature on habitual travel choice, which states that car dependency and the share of habitual travel are positively related to mileage. When an intervention challenges habitual behavior, the higher share of habitual travel amongst high-mileage drivers will lead to a bigger measurable impact on this driver subgroup. These findings consequently confirm the research hypotheses 4a und 4b.

Differences in the impact of accidents on driving are yet nonsignificant, both in absolute and relative terms. This suggests that driving self-confidence is affected similarly in the two driver groups. Therefore, I reject the research hypothesis 4c.

<table>
<thead>
<tr>
<th>Low-mileage drivers</th>
<th>High-mileage drivers</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = -1$</td>
<td>269.55</td>
<td>321.25</td>
</tr>
<tr>
<td>$t = 1$</td>
<td>251.76</td>
<td>281.08</td>
</tr>
<tr>
<td>Difference</td>
<td>-17.79</td>
<td>-40.17</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 23: Differences in trips per month for the 2 driver subgroups

<table>
<thead>
<tr>
<th>Low-mileage drivers</th>
<th>High-mileage drivers</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = -1$</td>
<td>1839.82</td>
<td>4535.29</td>
</tr>
<tr>
<td>$t = 1$</td>
<td>1595.36</td>
<td>3847.09</td>
</tr>
<tr>
<td>Difference</td>
<td>-244.46</td>
<td>-688.20</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 24: Differences in kilometers traveled for the 2 driver subgroups

<table>
<thead>
<tr>
<th>Low-mileage drivers</th>
<th>High-mileage drivers</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = -1$</td>
<td>32.35</td>
<td>45.29</td>
</tr>
<tr>
<td>$t = 1$</td>
<td>30.19</td>
<td>42.02</td>
</tr>
<tr>
<td>Difference</td>
<td>-2.15</td>
<td>-3.27</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 25: Differences in average driving speed for the 2 driver subgroups
6.2.5 The Impact of Accident Severity

Research on traumatized accident victims has shown that accident severity and traumatic disorders are positively correlated (cf. subsection 4.5). This leads to the assumption that accident severity should also correlate with any impact on travel and driving behavior. Tables 26 through 28 show the results of covariance analyses where the dichotomous variable “crash involvement” is replaced by a metric indicating maximum acceleration.\(^\text{51}\) The results show a high impact of accident severity on the three tested variables. With regard to monthly trips, data indicates that the increase in accident severity of 1g leads to a reduction in monthly trips of 9.333 units. Similar results are achieved for traveled kilometers and average speeds and let me infer that accident severity significantly impacts travel and driving behavior changes. The findings of this subsection therefore fully support the hypotheses 5a to 5c.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-14.481</td>
<td>5.881</td>
<td>-2.463*</td>
</tr>
<tr>
<td>Trips ( t = -1 )</td>
<td>1.050</td>
<td>0.019</td>
<td>55.227***</td>
</tr>
<tr>
<td>Max. Acceleration</td>
<td>-9.333</td>
<td>1.335</td>
<td>-6.992***</td>
</tr>
</tbody>
</table>

\( * p < 0.05, ** p < 0.01, *** p < 0.001 \)

\( R^2 = 0.679 \)

Table 26: The impact of accident severity on trips per month

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-88.897</td>
<td>67.720</td>
<td>-1.313 ***</td>
</tr>
<tr>
<td>Kilometers ( t = -1 )</td>
<td>0.999</td>
<td>0.022</td>
<td>45.978***</td>
</tr>
<tr>
<td>Max. Acceleration</td>
<td>-113.001</td>
<td>18.212</td>
<td>-6.205***</td>
</tr>
</tbody>
</table>

\( * p < 0.05, ** p < 0.01, *** p < 0.001 \)

\( R^2 = 0.608 \)

Table 27: The impact of accident severity on kilometers traveled per month

\(^{51}\) Note that this variable holds the maximum value out of the absolute lateral and longitudinal accelerations that occurred during the accident.
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.042</td>
<td>1.278</td>
<td>-3.164***</td>
</tr>
<tr>
<td>Speed t = -1</td>
<td>1.101</td>
<td>0.032</td>
<td>34.581***</td>
</tr>
<tr>
<td>Max. Acceleration</td>
<td>-0.847</td>
<td>0.134</td>
<td>-6.345***</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001

\[ R^2 = 0.453 \]

Table 28: The impact of accident severity on average driving speed per month

6.2.6 The Impact on Trips of Different Purpose

Car use depends on trip purpose (cf. subsection 4.6). It reflects the need for temporarily and spatially flexible transportation and makes motorists frequently prefer personal vehicles to other modes of transportation. Varying degrees of car dependency should also become discernible when reviewing the accident impact on different types of trips: the fewer alternatives to car use exist for trips of a certain category, the lower the accident-induced changes in trip rates should be.

Table 29 shows average trip frequencies in the accident driver population, relative changes attributed to accident experience, and the results of the Mann-Whitney-U Test. The findings indicate that motorists selectively choose which types of trips to reduce after an accident, causing significant differences in the relative reductions of the six trip categories under review (\( p < 0.05 \)). Work-related trips do not get curtailed at all; the necessity to go to work and a desire to work flexible hours seem to leave employees with no alternative to using their car for this purpose. Also, the high share of company car drivers might explain these minor relative changes, as riders who did not have the same accident experience might share the same vehicle. For recreational and shopping related trips, the measured relative trip changes are

\[ \frac{\sum_{i=1}^{N} t_i - \sum_{i=1}^{N} s_i}{\sum_{i=1}^{N} s_i} \]

1. In here, \( i = 1 \ldots N \) designates a single accident-involved driver, with \( t_i \) and \( s_i \) being the number of trips in the pre- and post-accident months, respectively. Due to similar considerations, it is not possible to provide standard deviations for the relative trip differences.

53 Note that the test of research hypothesis 6 does not consider trips for refueling. Trips with this intention are a result of general vehicle use. Practically, they cannot be avoided and are required to keep the vehicle operational. They eventually decline already if other trips are curtailed. Thus, I exclude them from the analysis.
considerably higher. For the former, motorists seem to engage more frequently in alternative modes of recreation that do not require car use or switch to other modes of transportation after the accident. For the latter, shopping trips might become bundled (i.e., purchasing grocery supplies that last for longer time periods) or may in part be substituted by public transportation. Actually, the measured downturn in shopping related trips is surprising, as it contrasts with other research, which states that the need for carrying things makes shopping trips highly car dependent (cf. Cullinane & Cullinane, 2003). In the remaining trip categories (i.e., home-bound trips, trips for private business, and uncategorized trips), medium trip reductions of about 11% are present. These closely resemble the average relative trip change as can be derived from Table 8. In summary, the findings of this subsection support research hypothesis 6.

<table>
<thead>
<tr>
<th>Home</th>
<th>Work</th>
<th>Recreational</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>$t = -1$</td>
<td>82.10</td>
<td>62.93</td>
</tr>
<tr>
<td>$t = 1$</td>
<td>73.05</td>
<td>60.78</td>
</tr>
<tr>
<td>Rel. Difference</td>
<td>-11.02%</td>
<td>-0.62%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Private business</th>
<th>Shopping</th>
<th>Noise</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>$t = -1$</td>
<td>45.52</td>
<td>97.91</td>
<td>13.70</td>
</tr>
<tr>
<td>$t = 1$</td>
<td>40.64</td>
<td>97.26</td>
<td>10.78</td>
</tr>
<tr>
<td>Rel. Difference</td>
<td>-10.74%</td>
<td>-21.35%</td>
<td>-13.29%</td>
</tr>
</tbody>
</table>

*p < 0.05, ** p < 0.01, *** p < 0.001

Table 29: The accident impact on the frequency of trips with different purpose

6.2.7 Situational and Spatial Avoidance

In subsection 4.7, I listed situational and spatial avoidance as strategies of road accident victims to evade emotional arousal and intrusive thoughts of the accident. Trauma research has shown that in some motorists such behaviors are still present 4-6 years after accident occurrence. In Table 30, I present relative trip changes of accident-involved motorists from the pre- to post-accident month. As the data shows, the relative differences between trips at the accident time and all other trips do not vary significantly. Both types decline by about 11% and exhibit little difference from the values presented in Table 8. If motorists engaged in situational avoidance, trips that are done at the accident time should decline considerably more
than all others. From the available data, research hypothesis 7a must therefore be rejected.

<table>
<thead>
<tr>
<th>Trips at accident time</th>
<th>Trips off accident time</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M )</td>
<td>( SD )</td>
<td>( M )</td>
</tr>
<tr>
<td>( t = -1 )</td>
<td>27.34</td>
<td>20.00</td>
</tr>
<tr>
<td>( t = 1 )</td>
<td>24.17</td>
<td>20.39</td>
</tr>
<tr>
<td>Rel. Difference</td>
<td>-11.59%</td>
<td>-11.15%</td>
</tr>
</tbody>
</table>

* \( p < 0.05 \), ** \( p < 0.01 \), *** \( p < 0.001 \)

**Table 30: The accident impact on trips at accident time**

Table 31 shows similar data for accident victims’ spatial avoidance. Again, no significant differences become discernible in the relative reductions of trips that pass and do not pass by the accident location. This suggests that motorists do not specifically choose other travel routes to circumnavigate the accident location. Research hypothesis 7b thus is not confirmed by the data as well.

<table>
<thead>
<tr>
<th>Trips passing by the accident location</th>
<th>Trips not passing by the accident location</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M )</td>
<td>( SD )</td>
<td>( M )</td>
</tr>
<tr>
<td>( t = -1 )</td>
<td>50.12</td>
<td>74.01</td>
</tr>
<tr>
<td>( t = 1 )</td>
<td>43.71</td>
<td>66.04</td>
</tr>
<tr>
<td>Rel. Difference</td>
<td>-12.79%</td>
<td>-10.87%</td>
</tr>
</tbody>
</table>

* \( p < 0.05 \), ** \( p < 0.01 \), *** \( p < 0.001 \)

**Table 31: The accident impact on trips that pass by the accident location**

6.2.8 **Summary of Hypothesis Testing**

In summary, the research findings show that accidents have a significant impact on daily travel and driving behavior (hypotheses 1a through 1c). The corresponding effects do not deteriorate over the concourse of five months and remain persistent during that time horizon (hypotheses 2a through 2c). The effects are nevertheless not strong enough to allow for inferences about male, female, or company car drivers (hypotheses 3a through 3c), although differences in high- and low-mileage motorists become discernible at least for travel behavior (hypotheses 4a through 4c). Furthermore, the research results indicate that accident severity influences the degree to which crash involvement changes travel and driving behavior (hypotheses 5a through 5c). The findings also show that relative trip reductions vary by travel
6.3 Additional Knowledge derived from GPS Data

This subsection answers research sub-question 4 and details other knowledge that can be derived from the analysis of GPS data. The descriptive results I present leave aside actuarial questions, although they may very well be used for that purpose. The
The intention of this subsection is to show further possibilities for analyzing GPS driving data and inferring from it individual driver information as well as knowledge for understanding group dynamics.

### 6.3.1 Individual Information

In this first part, I show further GPS-based information on single motorists by deriving transition routines. Related research frequently uses transition routines to better understand people’s spatial behavior and to assign purpose to different kinds of locations (e.g., Liao et al., 2005; Eagle & Pentland, 2006). As an example, I discuss the transition probabilities of a single company car driver (cf. Figure 28).\(^{54}\) Percentage values are conditional transition probabilities, that is, the likelihood of going to a particular location when at a specific location.

![Transition Probabilities Diagram](image)

*Figure 28: Transition probabilities of a single selected company car driver*

A notable aspect in the transition probabilities of the commercial vehicle is the high share of work to work transitions (94.9%). As can be inferred from the available data in combination with context information from Google Earth, the car is most likely used for the delivery runs of a grocery wholesaler. It stops at regional convenience stores in the early morning hours and brings items requested for restocking. Delivery runs on average include 19 regularly visited business clients before the vehicle goes back to the wholesaler’s warehouse or an unspecified point.

\(^{54}\) I have selected the motorist because (s)he does not visit all possible trips locations, which eases the graphical representation of the transition probabilities.
The share of infrequently visited locations is minor and comprises only quick stops for refueling, short personal shopping, or getting food.55

This example shows only basic aspects of transition probabilities; a plethora of other analyses is possible, for example making transitions conditional on time and weekday distributions or on the length of stay when locations are visited from different origins. These and other analyses could help to improve classification techniques as well as serve actuarial purposes.

6.3.2 Group Dynamics

An aspect that has not been considered in the previous post-accident analyses are group dynamics. They play an important role in traffic management and control (e.g., California Center for Innovative Transportation, 2010). Other work from Eagle & Pentland (2006) analyzes GPS data to infer group relationships in terms of daily proximity and friendship. While such analyses are not feasible with the available data (motorists are not located spatially near each other and thus potential common locations are not visited often enough to make such inferences), it is nonetheless possible to infer knowledge on single locations that get visited more often. As an example, I detail information that can be derived for individual shopping malls or super markets which may hold useful for shop owners. For this purpose, I consider the shopping mall at 43.847° North, 11.1425° East, northwest of Florence, Italy. Reviewing this location on Google Earth gives a clear image of the mall and its parking lot. From the motorists’ trips, it is possible to identify those which ended at that corresponding parking lot, that is, trips that were likely done with the intention of going shopping at that particular mall. 58 of the 549 accident-involved drivers visited the mall at least once during the study period. As the motorists’ home locations are known from earlier analyses, I use this information to identify the mall’s zone of attraction. Figure 29 shows the home locations of shop visitors as an overlay over a Google Earth map. White triangles indicate the customers’ home locations; the white cross points at the mall.

55 Note that this analysis of transition probabilities may be further extended by considering a motorist’s clusters individually and not grouping them by trip purpose. If several work-related locations exist, more trip routines may become discernible.
As the figure shows, visitors to the shopping mall are mainly from the Florence and Prato area, with some customers coming from the surrounding rural environment. 95% of customers are located within a radius of approximately 103km around the mall.\textsuperscript{56} The use of such information can be seen in strategically selecting locations for a particular shopping mall carrier and its competitors. Where should new shops be located in order to reach customers who have not visited a particular company’s stores because they are spatially too remote? From which regions do competitors attract business, and are there undeveloped regional “niches” where the installation of new shopping opportunities could be beneficial? Also more complex questions become answerable with such GPS data, for example, evaluating the effectiveness of different (regional) marketing campaigns. Which new customers are attracted by special marketing events? Where do they come from? To which degree do such marketing measures deter customers from shopping at competitors? How long do such marketing effects last, and thus, how effective are they in the long run? While such questions are not of primary interest for insurance companies today, they might be in the future, if they intensify their collaboration with external industry partners to whom such information may matter.

\textsuperscript{56} This estimate considers all drivers regardless of how often they visited the shopping mall. A more accurate estimate of the mall’s zone of attraction may be found when only considering motorists who visit the mall on a regular basis. In the absence of sufficient data and as the above calculation is primarily done for demonstration purposes, I omit such analysis here.
Figure 29: Zone of attraction of a shopping mall
7. Discussion and Implications

This section summarizes the research findings of this dissertation and translates them into operational and strategic recommendations for insurance practitioners. I begin with an outline of the theoretical contribution of this work in subsection 7.1. Next, subsection 7.2 states practical advice for the insurance industry that can be inferred from this dissertation. It comprises implications for post-accident driver management on an operational level and strategic directives in the face of the growing importance of telematics technology in motor insurance. Subsection 7.3 discusses both methodological and practical limitations of this work and shows areas for further research. Short concluding remarks in subsection 7.4 end this dissertation.

7.1 Theoretical Contribution

The confirmation of the hypotheses in set 1 suggests that accident involvement has a significant impact on vehicle use. Directly after an accident, road victims reduce their overall trips, cut down on mileage, and apply a more defensive driving style. These findings are in line both with Ajzen’s Theory of Planned Behavior as well as with the psychological studies on road accident victims as conducted by Mayou et al. (1993). Both perspectives explain behavioral changes by a psychological impairment as a result of accident experience, which lowers the motorists’ confidence in their vehicle operation skills. My research supplements the latter findings with practical evidence and adds more detail to the extent of behavioral change.

The rejection of the hypotheses in set 2 on the persistency of accident effects demands some explanation. Why, despite the indication of both theory and empirical psychological findings, does people’s car use not recover in the five month period after the accident? I see three different ways to explain this. First, even though accident victims seem to recover on the basis of survey-based driving anxiety ratings, unconscious negative feelings about vehicle use might still be prevalent amongst motorists and make them use their car less often and more cautiously. Second, accident experience might have broken habitual car use to some degree. While people drive less immediately after the accident due to a lack of self-confidence, they might simultaneously start to realize that for some trips, vehicle use is not unavoidable. Even though they regain confidence in their own driving over time, changes in travel and driving habits remain. Third, the five month
follow-up period may not be long enough to detect any normalization in vehicle usage. Mayou et al. (1993) for example used a twelve month observation period to derive their findings, although significant reductions in self-reported anxiety and depression scores already became discernible at the six month follow-up. Further investigations with extended observation periods may thus be advised for this research in order to achieve clear results.

The third set of hypotheses was not confirmed, as no significant group differences between male, female, and company car drivers could be recognized. These findings suggest that variances in car dependency are less pronounced than expected in the three driver subgroups. One possible explanation might be that the three subgroups do not reflect common gender rules associated with them. This might be the case if the underlying insurance product had been marketed to a specific target group or its conditions seemed especially appealing to such a subgroup. The share of high-mileage drivers within all subgroups suggests that the telematics policy is especially attractive to frequent car users. Even amongst female motorists, more than 50% exhibit a projected annual mileage of 30,000km or more. Thus, a high share of work related driving may be assumed for all groups, which eventually evens out expected group differences. Another source of bias may come from the fact that gender / company car information reflects car registration rather than actual vehicle use. Neither the identity of the person driving at the time of the accident nor other people allowed to operate the vehicle are known. The subgroups may be less distinct than the data actually suggests, so differences in the accident response of the subgroups stay covert.

The hypotheses in set 4 are partially confirmed: high- and low-mileage drivers respond differently to accident involvement in terms of travel demand while showing no significant variance in speed adaption. Evidence shows that frequent drivers reduce their car use to a higher degree than more occasional car users do. These results are contrary to the findings of Fujii et al. (2001) in their study of travel mode changes during temporary road closures, which demonstrated that travel behavior changes are more pronounced in infrequent car commuters. This discrepancy may be due to differences in the interventional quality of road closures and accidents: while the former affect road users on a short term basis, the latter have a more sustained impact due to motorists’ cognitive impairment. Short-term flexibility in travel demand as a response to interventions is higher in low-mileage drivers. In contrast, high-mileage drivers exhibit long-term elasticity in travel behavior; thus changes in their vehicle use may only become present if they are
exposed to more long-lasting interventional effects (i.e., accidents). The sustained psychological burden of accident experience may thus explain the discrepancy between the findings of this dissertation and those of related work.

The confirmation of the hypotheses in set 5 underpins the psychological findings of Mayou et al. (1991) on the positive association between accident severity and the impairment of driver confidence. The more severe an accident, the higher post-traumatic disorders, nervousness, and anxiety are, and thus the more pronounced changes in later driving are. This thesis supports the authors’ empirical findings with evidence from real world driving data.

Hypothesis 6 weakly confirms the notion of Goodwin et al. (1995) on varying car dependencies of different trip types. As the results show, trips are not reduced equally after an accident. Instead, the findings suggest that motorists deliberately choose which kinds of trips they reduce. Differences in relative reductions are nonetheless only of low significance, which I primarily attribute to erroneous trip class assignments during the classification process. As a consequence, trip type variances in accident response become blurred, which makes detectable group differences less distinct than they would be if all trips were classified correctly.

The two hypothesis in set 7 are both rejected. First, while other research finds that motorists try to stay away from situations similar to the one of the accident, the results of my research do not confirm this. I assume that time of day is a situational accident characteristic which motorists do not deliberately try to avoid. Other situational circumstances may be regarded as a bigger challenge to driving, for example bad weather conditions or driving in rush hour traffic. Also motorists’ travel routines (especially those of company car drivers) may prohibit traveling at other times of the day; appointments with business clients during regular working hours and fixed shop opening hours often leave few possibilities to avoid driving at a particular time of the day. Second, no significant results were also achieved regarding spatial avoidance. The reason might be that only a low percentage of motorists actively avoid the accident location, with Mayou et al. (1991) reporting that only 5% of study participants engaging in such behavior 4-6 years after the accident. Although I argued that more drivers show spatial avoidance in the first post-accident months, the results do not support this hypothesis. Eventually, the avoidance strategies of single motorists cannot be measured within the aggregate data.
7.2 Practical Implications for the Insurance Industry

The results of this research have several practical inferences for the insurance industry. Figure 30 gives an overview of these implications and categorizes them by their need for collaboration with external partners and the organizational level they address.

Figure 30: Overview of implications for the insurance industry

<table>
<thead>
<tr>
<th>Operational Level</th>
<th>Strategic Level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal Perspective</strong></td>
<td><strong>External Perspective</strong></td>
</tr>
<tr>
<td>Adjust bonus malus systems</td>
<td>Actively manage accident drivers</td>
</tr>
<tr>
<td></td>
<td>Develop tailored add-on services for business and fleet clients</td>
</tr>
<tr>
<td></td>
<td>Build competency in telematics technology and analysis</td>
</tr>
<tr>
<td></td>
<td>Intensify the collaboration with automotive partners</td>
</tr>
</tbody>
</table>

7.2.1 Operational Implications of the Analysis of Post-Accident Vehicle Use

Adjust bonus-malus systems

Bonus-malus systems are a common practice in many European motor insurance markets today. They reward motorists for accident-free years of driving and penalize accident causation by imposing premium surcharges. Claims-free years cause a decline of some degrees on the bonus-malus system; reported claims raise a driver’s level on the rating scheme. However, these adjustments do not reflect post-accident driving behavior and are frequently flat-rate. An understanding of vehicle use after accident involvement could help to fine-tune such bonus-malus schemes by making changes on the bonus level dependent on post-accident driving behavior. The more pronounced and sustained the changes in driving behavior are, the less post-accident premiums might rise. Alternatively, insurers could offer motorists with substantial changes in vehicle use regaining their initial bonus level more quickly. A comparison with other accident-involved drivers allows anticipating expected behavioral changes and thus entails an adjustment of the bonus-malus level directly after the accident. Further adjustments once real driving behavior becomes discernible may also be conducted.
**Actively manage accident drivers**

Information on post-accident driving behavior is also valuable apart from the actuarial perspective and can be used to actively manage accident drivers. For at-fault clients of different demographic backgrounds, insurers might assess whether and how fast they fall back to old, incautious driving styles. If accident experience causes just minor or short-term changes, the chances of repeated accident involvement are higher. Such motorists might eventually cause negative marginal returns that cannot be compensated in full by the earnings in premiums. Canceling such clients’ contracts might then be advisable. It helps to improve an insurer’s customer portfolio, thereby increasing its profitability (cf. Figure 5).

Another aspect of managing accident drivers is providing support apart from pure financial compensation. It might come in form of driver trainings or driving behavior reviews based on the available GPS data. Primarily, I see two target client groups for such measures. The first are at-fault drivers in which accident experience triggers only minor or short-term changes in driving style. Educational trainings and reviews might help to detect risk-prone behavior, reveal driving deficits, and eventually make motorists change their behavior. The second group are not at-fault drivers. Non-financial measures are intended to rebuild this group’s driving confidence, which might have been compromised by the crash involvement. The provision of such add-on services can thus help clients get back to their normal travel and living routines more quickly. To my knowledge, currently no major European insurer assists its clients with vehicle trainings and driving data reviews after accident involvement. They might thus provide a unique selling proposition when it comes to the design and marketing of new motor insurance products.

**Develop tailored add-on services for business and fleet clients**

The results of this dissertation show that company car drivers are especially at risk for both high mileage and more aggressive driving. While the latter makes this type of client rather undesirable from an actuarial perspective, corporate motor business is too big a market from which to retreat. Safety-related add-on services of telematics-based vehicle insurance can make this market more attractive for insurers and help them differentiate from competitors. For example, the GPS data allows insurers to discover at-risk motorists in a fleet of company vehicles. These motorists may be assisted with vehicle trainings and driving reviews to reduce potential accident risks actively. While insurers may benefit from a downturn in claims expenses by such measures, business clients will eventually gain higher workforce mobility and a lowered rate of illness-related absence. Moreover, data analysis can
also focus on efficiency aspects of driving. It can be used to identify motorists who are not driving providently, that is, traveling at higher speeds or accelerating heavily. Those drivers can be coached on ecological and fuel-efficiency aspects of driving. When administered to all drivers in a commercial fleet, the decrease in fuel expenses for that company can be substantial. In conclusion, if meaningful add-on services are designed and bundled to meet the needs of individual business clients, insurance companies can establish themselves as trusted partners on safety issues for business clients, which may result in extended collaboration and presumably higher earnings in premiums.

### 7.2.2 Strategic Implications

**Intensify the collaboration with automotive partners**

Telematics technology will become a standard vehicle feature within the next few years. Insurance companies have to prepare today for the changes this development holds for them. However, what do feasible strategies look like, and which recommendations can be given to different kinds of insurance companies in the market?

A general recommendation for insurers is to intensify their collaboration with automotive and telematics partners. Once eCall is rolled out on a broad basis, they are the ones who provide and install the required equipment and probably are in control of the surrounding infrastructure. Likewise, the telematics data might also be at their disposal. However, the way in which such collaboration might function differs for bigger and smaller insurance companies. For larger, pan-European players in the field, the strategic goal should be to become the provider of telematics-based insurance services for car manufacturers. They should position themselves as an insurance carrier for automobile companies, who can then sell white-labeled insurance products and services under their own brand, as for example Volkswagen is already doing today. Only large insurers can offer such services in several countries simultaneously, which comes as a benefit for car manufacturers when they want their insurance services to be bundled and handled by a single carrier. Insurers that are strong just in their home countries might thus not qualify as white-labeling partners of car companies.

**Build competency in telematics technology and analysis**

Smaller insurance companies need to pursue a different strategy. Their intra-industry cooperation should focus on telematics providers like Octo Telematics to develop competency in the area of telematics and corresponding data. This requires
engaging in their own trials, pilot projects, or even rolling out new telematics-based insurance products. Developing such competency and analyzing the telematics data for the benefit of the customer may open up profitable niches, for example in business clients or in special subsets of non-commercial customers (e.g., low-mileage drivers). Once eCall gets rolled out on a broader scale, they then benefit from not having to equip vehicles with telematics hardware themselves but from building upon an existing infrastructure. This may even make it possible to attract new business amongst clients who did not consider telematics-based vehicle insurance before. As a consequence, it will be crucial for small insurance players to develop competency in the analysis of telematics data and to develop new services upon it. Only this way they can target-market and attract specific driver subgroups and differentiate themselves from their competitors. Significant competency in the analysis of this data may also make them an attractive partner for automotive companies.

7.3 Limitations and Suggestions for Further Research

This subsection discusses the limitations of this research. It shows how additional context information enables new kinds of analyses and enhances the quality of existing ones. I also detail methodological alternatives for trip classification that may improve the accuracy of trip purpose assignment. Finally, I point out other areas of research aside from insurance where the existing data and methods may be applied.

Figure 31: Suggestions for further research
Making Use of Driver Context Information
All analyses within this research have exclusively used telematics data. No demographic client information or policy details have been evaluated. However, linking GPS data and driver context information allows answering existing research questions more precisely and opens up the opportunity to approach new ones. Different subgroups can be formed on demographic and policy data and compared on their accident response. Corresponding variables include gender, motorist age, marital status, subsequent years of accident-free driving, additional purchase of comprehensive coverage, and others. Note that while gender / company car information is already included in the accident data, it reflects car ownership rather than the accident-involved person. Access to this parameter could help to improve the already carried out group comparison. The analysis of gender / company car variances is likely to benefit as well from information on other allowed car users (e.g., spouses, children).

A second use of driver context information is the comparison of standard and telematics customers. Examining the demographic differences between both populations may allow identifying driver subgroups to whom telematics-based vehicle insurance especially appeals. This would facilitate more precise target-marketing of those products. Furthermore, additional information on non-telematics customers enables a comparison of the accident propensity in both populations. It allows inferring whether telematics-based coverage schemes succeed in attracting low-risk customers, as is contemplated in literature.

Finally, further research may consider group evaluations of at-fault and not at-fault drivers. It may be valuable for understanding behavioral changes as a cause of accident involvement in more depth. While it can be assumed that guiltiness predominantly challenges a motorist’s confidence in vehicle operation, non-fault accident involvement eventually sensitizes the motorist to the perils of road traffic more generally. Further analyses of at-fault and not at-fault motorists may also consider differences in the persistency of accident-induced behavioral changes.

Making Use of Road Type Information
The data provided by Octo Telematics includes road type information for every position record. I have not considered this data in my analyses, however, which opens up areas for further research. A first possibility is to study the accident impact on road type use. Questions of interest include whether motorists try to avoid particular road types as a result of the accident. Second, detailed information on the
road use patterns of different driver subgroups may improve premium models, as accident propensity is dependent on road type (cf. Figure 26).

**Linking GPS Data and Spatial Context Information**

In this dissertation, all analyses were performed without considering spatial context information. Such data is available in Geographic Information Systems, which store, analyze, and present data linked to locations. Of interest in this respect may be land use patterns, which can be incorporated to refine spatial clustering algorithms. Land use patterns geographically code industrial zones, residential areas, and other land use types, and could be applied to impose “natural” boundaries to cluster dimensions. By preventing clusters from stretching over several land use categories, trips of different purposes may be kept apart and the number of misattributed trip categories reduced. Likewise, classification algorithms can benefit from land use data if no further context information is available. They can aid a domain expert in manually identifying trip purposes and add extra information to automated trip classification procedures. Both measures will eventually improve the accuracy of trip purpose assignments and thus help to provide a more accurate data basis for subsequent analyses.

**Advanced Trip Classification**

Aside from using spatial context information, trip purpose assignment may benefit from more methodologically advanced classification schemes. In this dissertation, I have applied basic decision trees for this task, although other classifiers may be more accurate for determining trip purposes. Good results have been achieved by Liao et al. (2005, 2007), for example, who use Bayesian Networks and Conditional Random Fields to determine trip categories. They also employ transition probabilities as input to their analysis to improve their classification models further. Similar work is done by Eagle & Pentland (2006), who have worked with transition probabilities to determine location purpose and people’s relationship network. As corresponding information can be easily derived from the available data (cf. subsection 6.3), subsequent analyses may use them for classification issues.

**Testing for Moral Hazard in Insurance Markets**

With the combined availability of GPS driving data and client policy information, open questions can also be addressed in other research areas. From an actuarial perspective, it enables the analysis of moral hazard in insurance markets. Linking both types of data allows investigating whether motorists are good at estimating their own driving risk and thus choosing proper deductibles or co-payments. In other words, do motorists who frequently break the speed limit and engage in risky
driving behavior (e.g., night time rides) anticipate these risks when they purchase insurance? Relevant literature suggests that high risk drivers buy more insurance with lower amounts of out of pocket payments. Other analyses might focus on the accuracy of client self-assessments upon policy sign-up. On this occasion, motorists are frequently asked to name the primary vehicle use (e.g., commuting to work, recreational, company car). This information can be compared to real travel after signing the contract.

**Measuring the Effect of Fuel Price Changes on Travel Behavior**

Further research outside of the insurance domain can be seen in the influence of fuel price fluctuations on vehicle use. Corresponding analyses are possible with the available data, which is sufficient both in sample size and time period covered. Actually, the time between mid-2007 and mid-2009 shows substantial fluctuations in crude oil prices, which surged from 80 U.S. dollar to more than 140 U.S. dollar from late 2007 to mid 2008 and then recovered to between 60 and 80 U.S. dollar in 2009.\(^{57}\) It might be very interesting to see how people respond to such price fluctuations in both the short and long term. Do motorists curtail car use and drive more slowly in times of high gas prices? Do such behavioral changes persist once the price of fuel eventually declines? Moreover, inferring stops at gas stations from the available data (as they have already been identified by the trip classification algorithms) can provide further insights into whether personal experience of high fuel prices at the pump causes short term behavioral changes.

**Mode Choice Effects of Accident Involvement**

This research has shown that accident involvement leads to travel behavior changes, both in trips made and actual miles driven. The question of whether trips to locations of interest are completely avoided or just done by other modes of transportation still remains. As accident experience is not likely to change personal lifestyles, hobbies, or consumer preferences, it might be interesting to analyze how people keep up their daily business even when using their car less. Do people rely

---

\(^{57}\) The provided values denote trading prices for one barrel of Light Sweet Crude Oil. They are taken from comdirect.de at: http://isht.comdirect.de/html/detail/main.html?hist=5y&sSym=CLT0.NYM&DEBUG=0&Xsearch=UKN&bFirstTime=1&ind0=VOLUME&more=ROH%d6L&overview_hist=1d&reason=positive&sCat=FUT&sPageType=extended&sTab=chart&sWpType=UKN&type=CONNECTLINE
more heavily on other modes of transportation, do they engage in ride shares with friends and family, or are trips getting bundled, that is, do they visit more destinations in a row without going home in between? Answering these questions is not easy and requires additional surveying of accident drivers. Privacy considerations might complicate such endeavor, as insurers may be reluctant to let their clients be approached for research purposes. Alternatively, GPS data might be collected from clients’ mobile phones upon their approval to compare actual travel and car usage.

7.4 Concluding Remarks

On the example of post-accident travel and driving analysis, this research has shown how new knowledge and value can be created in the insurance domain from the analysis of GPS data. Aside from closing the corresponding research gaps existent in travel behavior research today, a primary intention of this dissertation has been to sensitize insurance practitioners to the potential that comes with telematics-based vehicle insurance and the affiliated data. Not only can it be a threat to insurance providers, applied deliberately, it can also create new business opportunities. This, however, requires considerable rethinking of the way in which business is traditionally done in this sector. One cornerstone for success will be the intensified collaboration with outside industry partners. On the one hand, they will gain more and more importance as a sales channel for insurance services. The impact of this development is observable already today. On the other hand, access to telematics data and competency in its analysis may open up business opportunities also in the other direction, that is, by insurers offering data-based services to external partners. Needless to say, such a collaboration must account for privacy considerations.

In conclusion, a general answer on how to respond to the changes telematics technology imposes onto the insurance landscape cannot be given today. A look into the near future of motor insurance, however, shows one thing clearly: telematics technology will become a commodity within road vehicles in the next couple of years. The eCall Initiative of the European Commission will ensure that a standard telematics platform will be created and that its in-vehicle implementation will be backed by authorities’ support. Insurers have to prepare themselves today for the full scale implementation of this technology by developing competency in the areas of telematics and GPS data analysis. Only by taking proactive measures they can become beneficiaries of the upcoming changes in the motor insurance business.
Appendix A: The DBScan Algorithm

Appendix A describes the DBScan clustering algorithm as developed by Ester et al. (1996). I start with introducing its basic concept and the density notions it uses. Next, the actual algorithm is presented. An outline of required adjustments to the Weka implementation of DBScan for it to work properly on the GPS data makes up the remainder of this appendix.

Basic idea and density notions
The DBScan algorithm by Ester et al. (1996) is one of today’s most common and frequently cited clustering algorithms. Its inherent benefits (i.e., minimum required domain knowledge, capability to find arbitrary shaped clusters, ability to identify noise; insensibility to the ordering of instances in the database) make it a first choice for clustering spatial data.

DBScan follows a density based notion of clusters, that is, clusters can be recognized for the density of points within a cluster being considerably higher than that outside (cf. Figure 32). DBScan formalizes this intuitive understanding and measures density by the number of instances that adjoin a single data point within a given radius. To be considered a cluster point, an instance has to have at least some minimum number of data points within that neighborhood; the distance function
$\text{dist}(p, q)$ and the distance threshold $\varepsilon$ determine its scope.\textsuperscript{58} Ester et al. define this $\varepsilon$-neighborhood of a point $p$ as

$$N_\varepsilon(p) = \{ q \in D | \text{dist}(p, q) \leq \varepsilon \},$$

with $D$ denoting the geographical locations in the database. A simple aggregation approach could mandate at least some minimum number of points $n_{\text{min}}$ to be in the $\varepsilon$-neighborhood of $p$. However, this ignores that there are two types of cluster points, that is, core and border points, which differ in their respective $\varepsilon$-neighborhood. For grouping all instances belonging to the same cluster $\varepsilon$ thus has to be set at a very low level, which eventually misclassifies noise. Therefore, the authors define the concept of direct density-reachability, which requires that “for every point $p$ in a cluster $C$ there is a point $q$ in $C$ so that $p$ is inside of the $\varepsilon$-neighborhood of $q$ and $N_\varepsilon(q)$ contains at least” some minimum number of points $n_{\text{min}}$ (Ester et al., 1996). More formalized, $p$ can be considered directly density-reachable from $q$ with respect to $\varepsilon$ and $n_{\text{min}}$ if $p \in N_\varepsilon(p)$ and $|N_\varepsilon(q)| \geq n_{\text{min}}$.

This non-symmetric condition can be canonically extended to density-reachability. Following the authors, $p$ is density-reachable from $q$ if “there is a chain of points $p_1, \ldots, p_n, p_1 = q, p_n = p$ such that $p_{i+1}$ is directly density-reachable from $p_i$.” The relation of $p$ and $q$ is now transitive, but only symmetric if two core points are involved. To extend density-reachability to the case of two border points, Ester et al. next define the notion of density-connectivity. With respect to $\varepsilon$ and $n_{\text{min}},$ $p$ and $q$ are density-connected if a point $o$ exists from which both points are density-reachable. This concept of density-connectivity now describes a symmetric relationship of cluster points.

Figure 33 summarizes the density notions of the DBScan algorithm. They lay the foundation for the authors’ final definitions of clusters and noise points. With respect to $\varepsilon$ and $n_{\text{min}},$ a cluster $C$ is a non-empty subset of $D$ whose points fulfill the following conditions:

\[ \forall p, q: \text{if } p \in C \text{ and } q \text{ is density-reachable from } p, \text{ then } q \in C. \]

\[ \forall p, q \in C: p \text{ and } q \text{ are density-connected.} \]

\textsuperscript{58} The DBScan algorithm works with any distance function for two points $p$ and $q$. For clustering purposes, I use the Euclidian distance metric, as it is most appropriate for spatial data.
The former criterion guarantees the maximality of the cluster and describes the cluster points as a set of mutually density-connected data instances. The latter ensures the connectivity of the associated items. Data points are regarded as noise if they cannot be associated with any of the clusters, thus

\[
\text{Noise} = \{p \in D | \forall i: p \notin C_i\}.
\]

**The algorithm**

With the parameters \(\varepsilon\) and \(n_{\text{min}}\) at hand, a cluster can be found by two consecutive steps: First, select an arbitrary data point that satisfies the core point condition. It serves as a seed to the algorithm. Second, determine all instances that are density-reachable from it. If sufficiently many points are found, a new cluster is created. Otherwise, the seed is regarded as noise. Note that it may nonetheless be assigned to another cluster later on if it shapes up as density-reachable from one of its core points.

Once an instance is assigned to a cluster, the points in its \(\varepsilon\)-neighborhood are attributed to the cluster as well. This recursive process continues until all points to the cluster are found. Then, the algorithm selects a new seed from the unvisited
points and the process starts all over again. The robustness of the algorithm against arbitrary seed choice ensures the identification of stable clusters regardless of the ordering of instances in the database.

Following Ester et al. (1996), the formal representation of the DBScan algorithm is:

\[ \text{DBSCAN (SetOfPoints, Eps, MinPts)} \]
\[ // \text{SetOfPoints is UNCLASSIFIED} \]
\[ \text{ClusterId := nextId(NOISE);} \]
\[ \text{FOR i FROM 1 TO SetOfPoints.size DO} \]
\[ \text{Point := SetOfPoints.get(i);} \]
\[ \text{IF Point.ClId = UNCLASSIFIED THEN} \]
\[ \text{IF ExpandCluster(SetOfPoints, Point, ClusterId, Eps, MinPts) THEN} \]
\[ \text{ClusterId := nextId(ClusterId)} \]
\[ \text{END IF} \]
\[ \text{END IF} \]
\[ \text{END FOR} \]
\[ \text{END;} // \text{DBSCAN} \]

The function ExpandCluster used therein is:

\[ \text{ExpandCluster(SetOfPoints, Point, ClId, Eps, MinPts) : Boolean;} \]
\[ \text{seeds := SetOfPoints.regionQuery(Point, Eps);} \]
\[ \text{IF seeds.size<MinPts THEN // no core point} \]
\[ \text{SetOfPoint.changeClId(Point, NOISE);} \]
\[ \text{RETURN False;} \]
\[ \text{ELSE // all points in seeds are density-reachable from Point} \]
\[ \text{SetOfPoints.changeClIds(seeds, ClId);} \]
\[ \text{seeds.delete(Point);} \]
\[ \text{WHILE seeds <> Empty DO} \]
\[ \text{currentP := seeds.first();} \]
\[ \text{result := SetOfPoints.regionQuery(currentP, Eps);} \]
\[ \text{IF result.size >= MinPts THEN} \]
\[ \text{FOR i FROM 1 TO result.size DO} \]
\[ \text{resultP := result.get(i);} \]
\[ \text{IF resultP.ClId IN (UNCLASSIFIED, NOISE)} \]
\[ \text{THEN} \]
\[ \text{IF resultP.ClId = UNCLASSIFIED THEN} \]
\[ \text{seeds.append(resultP);} \]
\[ \text{END IF;} \]
\[ \text{SetOfPoints.changeClId(resultP, ClId);} \]
\[ \text{END IF;} // \text{UNCLASSIFIED or NOISE} \]
\[ \text{END FOR;} \]
\[ \text{END IF;} // \text{result.size >= MinPts} \]
\[ \text{seeds.delete(currentP);} \]
\[ \text{END WHILE;} // \text{seeds <> Empty} \]
\[ \text{RETURN True;} \]
\[ \text{END IF} \]
\[ \text{END;} // \text{ExpandCluster} \]
Running DBScan in the Weka environment

The DBScan algorithm is executed in Java Release 6 using the Ganymede Eclipse Software Developer Kit 3.4.2. It is implemented by the open-source WEKA 3.6.1 data mining library. A review of the source code shows that the program operates on normalized data, that is, spatial input data is rescaled to the codomain \([0; 1]\) in latitudinal and longitudinal direction. It sets minimum and maximum coordinate values to 0 and 1, respectively, and recalculates all others accordingly. However, initial codomains for longitudinal and lateral data are eventually not equal, so the adjustments in both dimensions are different. As a consequence, the radius \(\varepsilon\) in which DBScan seeks neighboring points deforms to an ellipsis and cannot be regarded as reflecting Euclidian distance measures anymore, eventually resulting in unpredictable clustering results (cf. Figure 34). As the algorithm is run for each driver individually, the search radius \(\varepsilon\) may furthermore vary from driver to driver if set as a global parameter. While distorting effects may be minor in cases where \(\text{Lat}_{\text{max}} - \text{Lat}_{\text{min}} \approx \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}}\), they increase the more the latitudinal and longitudinal spreads differ.

As a solution, I add an extra point \(P_{\text{max}}^*\) and to the database so that \(\text{Lat}_{\text{max}} - \text{Lat}_{\text{min}} = \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}}\). It holds the maximum coordinate in that dimension which exhibits the larger spread and a deducted auxiliary value in the other. It is considered in the cluster runs of each individual driver together with \(P_{\text{min}}\), which holds the global minima \(\text{Lat}_{\text{min}}\) and \(\text{Lon}_{\text{min}}\) over all motorists’ trip destinations. The coordinates of \(P_{\text{max}}^*\) are

![Figure 34: The distorting effect of data normalization on \(\varepsilon\)](image-url)
\[ \text{Lat}_{\text{max}} \text{ and} \]
\[ \text{Lon}_{\text{max}}^* = \frac{((\text{Lat}_{\text{max}} - \text{Lat}_{\text{min}}) - (\text{Lon}_{\text{max}} - \text{Lon}_{\text{min}}))}{\cos \left( \frac{(\text{Lat}_{\text{max}} - \text{Lat}_{\text{min}}) \cdot \pi}{2 \cdot 180} \right) \cdot \frac{2 \cdot \pi \cdot 6378000}{360}} + \text{Lon}_{\text{max}} \]

if \( \text{Lat}_{\text{max}} - \text{Lat}_{\text{min}} \geq \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}} \) and

\[ \text{Lat}_{\text{max}}^* = \frac{((\text{Lat}_{\text{max}} - \text{Lat}_{\text{min}}) - (\text{Lon}_{\text{max}} - \text{Lon}_{\text{min}}))}{111320} + \text{Lat}_{\text{max}} \]

\[ \text{Lon}_{\text{max}} \]

if \( \text{Lat}_{\text{max}} - \text{Lat}_{\text{min}} < \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}} \).\(^{59}\) Note that \( \text{Lon}_{\text{max}}^* \) takes into consideration that the geographical length covered by 1° of longitude depends on underlying latitude. I set it at the arithmetic mean of \( \text{Lat}_{\text{max}} \) and \( \text{Lat}_{\text{min}} \). As latitudes are equidistant at 111.320km in vertical direction, I use a simpler formula for \( \text{Lat}_{\text{max}}^* \). With \( P_{\text{max}}^* \) being spatially remote from any other trip end points, it will not interfere with the identification of any clusters and gets discarded from the database along with \( P_{\text{min}} \) once the clustering process is completed. Figure 35 summarizes the use of \( P_{\text{max}}^* \) graphically.

\[ \text{Figure 35: Data normalization without the distorting effect on } \varepsilon \]

In a final adjustment step, I translate the parameter \( \varepsilon \), which is specified in meters, into a normalized value \( \varepsilon^* \). It is the geometric mean over its relative size in the two geographical dimensions:

\(^{59}\) All calculations are done in meters. 6’378’000 is the earth radius in meters, 111’320 is 1° latitude in meters.
Appendix A: The DBScan Algorithm

\[ \varepsilon^* = \sqrt{\frac{\varepsilon}{\text{Lat}_{\text{max}} - \text{Lat}_{\text{min}}}} \cdot \frac{\varepsilon}{\text{Lon}_{\text{max}} - \text{Lon}_{\text{min}}} \text{ for } \text{Lat}_{\text{max}} - \text{Lat}_{\text{min}} \geq \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}} \]

\[ \varepsilon^* = \sqrt{\frac{\varepsilon}{\text{Lat}^*_\text{max} - \text{Lat}_{\text{min}}}} \cdot \frac{\varepsilon}{\text{Lon}_{\text{max}} - \text{Lon}_{\text{min}}} \text{ for } \text{Lat}_{\text{max}} - \text{Lat}_{\text{min}} < \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}} \]

Concludingly, note that even after data normalization \( \text{Lat}^*_\text{max} - \text{Lat}_{\text{min}} \neq \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}} \). The small remaining discrepancy that is still on hand stems from the fact that again the distance in 1° longitude is a function of the underlying latitude. If \( \text{Lat}^*_\text{max} \) is calculated, that is, if \( \text{Lat}_{\text{max}} - \text{Lat}_{\text{min}} < \text{Lon}_{\text{max}} - \text{Lon}_{\text{min}} \), the latitudinal spread increases, and so does the arithmetic mean of \( \text{Lat}^*_\text{max} \) and \( \text{Lat}_{\text{min}} \). The bias is yet minor and partially eliminated by the calculation of \( \varepsilon^* \).
Appendix B: The C4.5 Algorithm

Appendix B describes the C4.5 classification algorithm as developed by Quinlan (1993), which I used to automatically assign purpose to individual trip end locations. I start with a general introduction to decision trees and name basic concepts and terminology. Next, I illustrate the C4.5 heuristic. It follows a thorough derivation of the evaluation criteria used to compare the building of possible instance subsets. I conclude with a representation of the decision tree that was obtained from the training data and applied within my research.

A primer on decision trees

Decision trees describe a process of multistage decision making (Safavian & Landgrebe, 1991). They follow a divide-and-conquer approach (cf. Hunt et al., 1966) by consecutively breaking complex decision problems into subsets of more simple decisions. Decision trees are a form of supervised learning where a domain expert presets the classes to be found. The latter also has to provide training data, that is, instances for which class labels have already been assigned. The decision tree is developed upon them and then used to predict the class membership of uncategorized entities in a subsequent step.

Some basic terminology exists to describe decision trees (cf. Figure 36). On the top of a tree, a root node initially holds all instances. On each node, a decision rule is applied which assigns the instances to 2...n subsets based on their attribute values. Subsets that are not split up any further are called leaves, which is the case if some

Figure 36: General decision tree (adopted from Safavian & Landgrebe, 1991)
stopping criterion is met or the leaf holds just one item. All instances on a leaf are then tagged with the label of the most frequent class. This process iterates until the decision tree is completely specified.

The algorithm
C4.5 is an entropy based classifier that extends the author’s earlier ID3 algorithm (cf. Quinlan, 1986) by the use of continuous input variables. Given a set of pre-classified training instances \( S \), the heuristic is:

- If \( S \) meets a stopping criterion, cancel further tree building on that leaf and attribute \( S \) the most frequent class in \( S \).
- Apply a test \( T \) with mutually exclusive outcomes \( T_1, T_2, \ldots, T_k \) to split \( S \) into subgroups \( S_1, S_2, \ldots, S_k \). Each \( S_i \) contains those instances associated with \( T_i \). \( T \) forms the root of the tree, the outcomes \( T_i \) its subtrees.
- Apply the same procedure for the subsets \( S_i \) recursively.

Several stopping conditions exist to prohibit further tree building. Such criteria include that all instances in \( S \) are the same class, some minimum number of instances per leaf are required, or the information gain by splitting \( S \) has to exceed some threshold. For a more thorough discussion of stopping criteria for C4.5, the interested reader is referred to Quinlan (1996).

C4.5 can perform tests on both discrete and continuous attributes \( A \). These are \( A = ? \) and \( A \leq t \), respectively. In the continuous case, possible outcomes are \( true \) and \( false \); threshold \( t \) is set to maximize the applied splitting criterion. Possible values for \( t \) are given by \( t = \frac{1}{2} \cdot (v_i + v_{i+1}) \), where \( v_i \) and \( v_{i+1} \) are pairs of ordered, adjacent attribute values.

Evaluating information gain
C4.5 uses a entropy-based gain ratio as splitting criterion. It is an extension of the gain criterion that was initially applied in Quinlan’s ID3 algorithm. In the following paragraphs, I gradually develop both criteria to provide an understanding of how the algorithm builds a decision tree by selecting distinct tests from a plethora of testing alternatives. My description closely follows the author’s own description of the test evaluation process (cf. Quinlan, 1993).

Imagine a set of cases \( S \) that holds instances from \( k \) different classes. Let \( freq(C_j, S) \) furthermore denote the number of cases in \( S \) that belong to class \( C_j \), then
Appendix B: The C4.5 Algorithm

\[ \frac{freq(C_j, S)}{|S|} \]

indicates the proportion of cases in \( S \) belonging to this class. Following Shannon (1948), the information a corresponding message conveys is

\[ -\log_2 \left( \frac{freq(C_j, S)}{|S|} \right) \]

bits. The expectancy value of information from a message on class affiliation then is the sum over all single message information weighted by the proportions of single class attributes in \( S \), given by

\[ info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \cdot \log_2 \left( \frac{freq(C_j, S)}{|S|} \right) \text{ bits.} \]

It corresponds to the average amount of information required to determine the class membership of a particular instance in \( S \). Let \( S \) now be partitioned into \( n \) subsets \( S_i \) by an arbitrary testing procedure \( X \). The expected information requirement over all subsets then is

\[ info_X(S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot info(S_i), \]

which is the weighted sum of the subsets’ information requirements. The information gain by applying test \( X \) can thus be expressed as

\[ gain(X) = info(S) - info_X(S). \]

The algorithm selects that test which maximizes the gain in information. Unfortunately, this gain ratio is prone to large numbers of outcomes (which are present if testing on a discrete variable) and yields the biggest value when each subset \( S_i \) holds just one entity. Quinlan therefore uses a split criterion in C4.5 to overcome this problem. He normalizes the gain criterion by the split information

\[ split\ info(X) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot \log_2 \left( \frac{|S_i|}{|S|} \right), \]

which represents “the potential information generated by diving \([S]\) into \( n \) subsets, whereas the information gain measures the information relevant to classification.
that arises from the same division” (Quinlan, 1993). The resulting gain ratio of test $X$ then is

$$gain\text{ ratio}(X) = \frac{gain(X)}{split\ info(X)}.$$ 

**Decision tree results**

I finally present the decision tree as derived from the 958 training instances (i.e., trip end clusters). It is the result of a 12-fold cross validation process with the number of minimum instances per leaf being set to 15. Its nodes show the used split attributes and their values, the leaves display the associated trip classes, following the taxonomy of Table 6. The values in parentheses indicate the amount of total and misclassified instances associated with every leaf.

---

60 This value is the difference of the total number of clusters in the accident driver population (1061) minus the number of clusters to which the trip purpose was assigned manually (103).
J48 pruned tree

---------------
Length_of_Stay(12-24h) <= 0.016949
 | Arrival_Day(Sunday) <= 0.02
 |   | Client_Session <= 1: W (306.0/43.0)
 |   | Client_Session > 1
 |   |   | Length_of_Stay(0-0.25h) <= 0.542373
 |   |   |   | Arrival_Day(Wednesday) <= 0.11: R (16.0/6.0)
 |   |   |   | Arrival_Day(Wednesday) > 0.11
 |   |   |   |   | Arrival_Time(06:00-07:00) <= 0.058
 |   |   |   |   |   | 95%_Longitudinal_CI <= 235.228039: S (16.0/10.0)
 |   |   |   |   |   | 95%_Longitudinal_CI > 235.228039: R (15.0/7.0)
 |   |   |   |   |   | Arrival_Time(06:00-07:00) > 0.058: W (42.0/3.0)
 |   |   |   | Length_of_Stay(0-0.25h)> 0.542373
 |   |   |   |   | Days_from_first_to_last_Visit <= 311: W (23.0/13.0)
 |   |   |   |   | Days_from_first_to_last_Visit > 311
 |   |   |   |   |   | Arrival_Time(14:00-15:00) <= 0.064: P (24.0/5.0)
 |   |   |   |   |   | Arrival_Time(14:00-15:00) > 0.064: F (18.0/6.0)
 |   |   | Arrival_Day(Sunday) > 0.02
 |   |   |   | Length_of_Stay(1-2h) <= 0.049231
 |   |   |   |   | 95% Latitudinal_CI <= 136.653407: F (15.0/4.0)
 |   |   |   |   | 95% Latitudinal_CI > 136.653407
 |   |   |   |   | Length_of_Stay(4-8h) <= 0.002208: P (57.0/22.0)
 |   |   |   |   | Length_of_Stay(4-8h) > 0.002208: R (15.0/8.0)
 |   |   |   | Length_of_Stay(1-2h) > 0.049231
 |   |   | Arrival_Day(Sunday) <= 0.08
 |   |   |   | ClientSession <= 1: W (35.0/14.0)
 |   |   |   | ClientSession > 1: R (31.0/16.0)
 |   |   | Arrival_Day(Sunday) > 0.08: R (76.0/12.0)
 | Length_of_Stay(12-24h) > 0.016949
 | Number_of_Days_the_Cluster_was_visited <= 112
 |   | Arrival_Day(Sunday) <= 0.09
 |   |   | Arrival_Time(21:00-22:00) <= 0.004: W (63.0/23.0)
 |   |   | Arrival_Time(21:00-22:00) > 0.004: R (19.0/8.0)
 |   |   | Arrival_Day(Sunday) > 0.09: R (62.0/1.0)
 | Number_of_Days_the_Cluster_was_visited > 112: H (125.0/13.0)
References


Filipova, L. (2007). Monitoring and Privacy in Automobile Insurance Markets with Moral Hazard. *Volkswirtschaftliche Diskussionsreihe No. 293, Faculty of Business Administration and Economics, University of Augsburg (Germany)*. Available at: http://webdoc.sub.gwdg.de/ebook/serien/lm/Augsburger_vwl_diskussionsreih/293.pdf (October 6th, 2009).


Ng, R. T., & Han, J. (1994). *Efficient and Effective Clustering Methods for Spatial Data Mining*. Paper presented at the 20th International Conference on Very Large Data Bases.


Spaces. Paper presented at the 85th Annual Meeting of the Transportation Research Board.


Stradling, S. G. (2000). Driving as part of your work may damage your health. In G. B. Grayson (Ed.), *Behavioural Research in Road Safety IX* (pp. 1-9). Berkshire (UK): Transport Research Laboratory.


Wolf, J. (2000). *Using GPS Data Loggers To Replace Travel Diaries In the Collection of Travel Data*. Georgia Institute of Technology, Atlanta (USA).


CURRICULUM VITAE

Personal Details

Nationality    German

Date of birth  September 29, 1980 in Cham (Bavaria, Germany)

Education

2006-2010  University of St. Gallen (HSG), St. Gallen, Switzerland
Ph.D. student and research associate at the Institute of Technology Management

2009-2010  University of California, Berkeley, USA
Visiting research scholar at the California Center for Innovative Transportation (CCIT); research fellowship of the Swiss National Science Foundation (SNSF)

2002-2006  University of Eichstätt-Ingolstadt, Ingolstadt, Germany
Diploma in Business Administration, specialization in logistic management, computer science, and statistics

2004-2005  University of Memphis, Memphis, USA
Study abroad

2001-2002  University of Bayreuth, Bayreuth, Germany
Economics and Philosophy